FMMI: Flow Matching Mutual Information Estimation

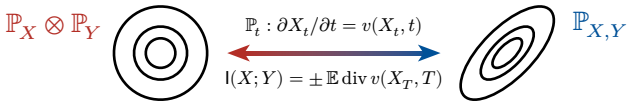
Ivan Butakov[1 2 3] **Alexander Semenenko**[1 2] **Valeriya Kirova**[4] **Ivan Oseledets**[5 3]
Alexey Frolov[1]

[1]Applied AI Institute, Moscow, Russia
[2]Moscow Independent Research Institute of Artificial Intelligence, Moscow, Russia
[3]Institute of Numerical Mathematics RAS, Moscow, Russia
[4]National Research Nuclear University MEPhI, Moscow, Russia [5]AXXX, Moscow, Russia

ivan.butakov@applied-ai.ru semenenko@applied-ai.ru

Abstract

We introduce a novel Mutual Information (MI) estimator that fundamentally reframes the discriminative approach. Instead of training a classifier to discriminate between joint and marginal distributions, we learn a normalizing flow that transforms one into the other. This technique produces a computationally efficient and precise MI estimate that scales well to high dimensions and across a wide range of ground-truth MI values.



1 Introduction

Mutual Information (MI) is a fundamental measure of non-linear statistical dependence between two random vectors, defined as the Kullback-Leibler divergence between the joint distribution and the product of marginals [1]: $\mathsf{I}(X;Y) = \mathsf{KL}\left[\mathbb{P}_{X,Y} \,\middle\|\, \mathbb{P}_X \otimes \mathbb{P}_Y\right]$. It is well-defined for most joint distributions, non-negative and zero if and only if X and Y are independent, invariant to bijections, and possesses many other useful properties.

For these reasons, MI is used extensively for theoretical analysis of overfitting [2, 3], hypothesis testing [4], feature selection [5, 6, 7] representation learning [8, 9, 10, 11, 12], and studying mechanisms behind generalization in deep neural networks (DNNs) [13, 14, 15, 16].

In practical scenarios, $\mathbb{P}_{X,Y}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$ are unknown, requiring MI to be estimated from finite samples. This reliance on empirical estimates leads to the curse of dimensionality: the sample complexity of MI grows exponentially with the number of dimensions [17, 18]. Long-tailed distributions and large values of MI further complicate the estimation [18, 19]. These problems considerably limit the applications of information theory to real-scale problems. However, recent advances in the neural estimation methods show that complex parametric estimators achieve relative practical success in the cases where classical MI estimation techniques fail.

Contemporary parametric MI estimators fall into two broad categories: *discriminative* and *generative* [20]. While the latter learn $\mathbb{P}_{X,Y}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$ from scratch, the former employ a classifier to discern between these two distributions. Despite the discriminative approach being more elegant, modern state-of-the-art MI estimators are mostly generative [21, 22, 23]. This reflects the inherent limitations of current discriminative methods: they are essentially *generalized energy-based models* (GEBMs) [24], which are notoriously suboptimal for significantly different prior and posterior distributions.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In this work, we take a leap forward and leverage **continuous-time Normalizing Flows (CNFs)** and **Flow Matching (FM)** to advance the discriminative approach beyond GEBMs. Contrary to the traditional generative approach, we learn a function which transforms a product marginal distributions into a joint distribution (*couples* the data). It can be shown that, in continuous case, the expected log-Jacobian of this transform is precisely the Mutual Information between the coupled data. The flow matching technique is used to avoid the reliance on unknown data distribution. This approach can also be generalized to any multi-variable extension of MI.

Overall, our contribution is as follows:
1. We propose a *novel* and *universal* flow-based estimator of information quantities — **FMDoE**.
2. We provide theoretical guarantees for our method and amortize it to reduce the cost.
3. The estimator is evaluated across a diverse set of benchmarks, including high-dimensional, high-MI and other challenging setups. The results indicate the superiority of our method.

The remainder of the paper is structured as follows: in Section 2, the necessary background in information theory and flow models is provided; the related works are discussed in Section 3; Section 4 introduces a general approach that is then reduced to the MI and O-information estimation in Section 5 and the method is evaluated across numerous setups and competitors in Section 7; finally, we discuss the results in Section 8.

2 Background

Information Theory. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with sample space Ω, σ-algebra $\mathcal{F}$, and probability measure $\mathbb{P}$ defined on $\mathcal{F}$. For another probability measure $\mathbb{Q}$ that is continuous with respect to (w.r.t.) $\mathbb{P}$ (denoted $\mathbb{Q} \ll \mathbb{P}$), the Kullback-Leibler (KL) divergence between $\mathbb{P}$ and $\mathbb{Q}$ is $\mathsf{KL}[\mathbb{Q} \parallel \mathbb{P}] = \mathbb{E}_{\mathbb{Q}}\left[\log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right]$, which is non-negative and vanishes if and only if (iff) $\mathbb{P} = \mathbb{Q}$.

Consider random vectors $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ with joint distribution $\mathbb{P}_{X,Y}$ and marginals $\mathbb{P}_X$ and $\mathbb{P}_Y$, respectively. We denote product measure by $\mathbb{P}_X \otimes \mathbb{P}_Y$. Wherever needed, we assume the relevant Radon-Nikodym derivatives exist. The mutual information (MI) between X and Y quantifies the divergence between the joint distribution and the product of marginals:

$$\mathsf{I}(X;Y) = \mathbb{E} \log \frac{\mathrm{d}\,\mathbb{P}_{X,Y}}{\mathrm{d}\,\mathbb{P}_X \otimes \mathbb{P}_Y} = \mathsf{KL}\left[\mathbb{P}_{X,Y} \;\middle\|\; \mathbb{P}_X \otimes \mathbb{P}_Y\right] \tag{1}$$

MI posses several outstanding properties. Namely, $\mathsf{I}(X;Y) = 0$ iff $X \perp\!\!\!\perp Y$ and $\mathsf{I}(X;Y) = \mathsf{I}(X; g(Y))$ for an invertible and measurable g [1].

When $\mathbb{P}_X$ admits a probability density function (PDF) p_X w.r.t. the Lebesgue measure, the differential entropy is defined as $\mathsf{h}(X) = -\mathbb{E}[\log p_X(X)]$, where $\log(\,\cdot\,)$ denotes the natural logarithm. Likewise, the joint entropy $\mathsf{h}(X,Y)$ is defined via the joint density $p_{X,Y}(x,y)$, and conditional entropy is $\mathsf{h}(X \mid Y) = -\mathbb{E}\left[\log p_{X|Y}(X \mid Y)\right]$. Under the existence of PDFs, MI satisfies

$$\mathsf{I}(X;Y) = \mathsf{h}(X) - \mathsf{h}(X \mid Y) = \mathsf{h}(Y) - \mathsf{h}(Y \mid X) = \mathsf{h}(X) + \mathsf{h}(Y) - \mathsf{h}(X,Y). \tag{2}$$

Optimal Transport. Consider a Polish metric space (M, d). Let $\Gamma(\mathbb{P}_X, \mathbb{P}_Y)$ be a set of joint measures $\mathbb{P}_{X,Y}$ (*couplings*) whose marginals are $\mathbb{P}_X$ and $\mathbb{P}_Y$, defined on M. For $p \geq 1$, the Wasserstein p-distance between $\mathbb{P}_X, \mathbb{P}_Y$ is $\mathsf{W}_p(\mathbb{P}_X, \mathbb{P}_Y) = \inf_{\mathbb{P}_{X,Y} \in \Gamma(\mathbb{P}_X, \mathbb{P}_Y)} \sqrt[p]{\mathbb{E}\, d(X,Y)^p}$. Following the idea behind (1), one can define the *Wasserstein dependency measure* (WMI) [25]:

$$\mathsf{WI}_p(X;Y) = \mathsf{W}_p\left(\mathbb{P}_{X,Y}, \mathbb{P}_X \otimes \mathbb{P}_Y\right)$$

While this measure maintains some key properties of conventional MI (namely, nullification iff $X \perp\!\!\!\perp Y$), it also loses many of them (e.g., the invariance to bijections and DPI). Hence it is not as widely adopted as MI.

Normalizing Flows. Consider two absolutely continuous distributions $\mathbb{P}$ and $\mathbb{Q}$ defined on $\mathbb{R}^d$ with PDFs p and q correspondingly. Suppose we want to learn an invertible transform $f : \mathbb{R}^d \to \mathbb{R}^d$ such that $\mathbb{Q}$ is a push-forward of $\mathbb{P}$:

$$\mathbb{P} = \mathbb{Q} \circ f, \quad p(x) = q(f(x)) \cdot \left|\det \frac{\partial f}{\partial x}\right|(x) \tag{3}$$

If $q(x)$ is tractable, f can be learnt via likelihood maximization [26].

Most contemporary normalizing flows are *continuous* [27]. That is, they are parametrized by a velocity field through the following ordinary differential equation (ODE):

$$\frac{\partial x_t}{\partial t} = v(x_t, t), \quad t \in [0; 1], \quad f(x_0) \stackrel{\text{def}}{=} x_1(x_0) \tag{4}$$

In this case, the change of variables formula becomes[1]

$$\frac{\mathrm{d}}{\mathrm{d}t}[\log p_t(x_t)] = -\operatorname{div} v(x_t, t), \qquad \begin{matrix} p_0 = p \\ p_1 = q \end{matrix} \tag{5}$$

While calculating the divergence naïvely requires $O(d)$ differentiations, an unbiased Hutchinson trace estimator [28] can be used to reduce the cost down to $O(1)$ [29]:

$$\operatorname{div} v(x, t) = \operatorname{tr} \frac{\partial v}{\partial x} = \mathbb{E}\, a^\intercal \frac{\partial v}{\partial x} a, \quad \text{where} \quad \begin{matrix} \mathbb{E}\, a = 0 \\ \operatorname{cov}(a) = \mathrm{I} \end{matrix}$$

Flow Matching. Vanilla CNFs possess *two severe limitations*: the maximum-likelihood training requires (a) q to be tractable and (b) x_t to be simulated through a numerical ODE integration. Framing velocity field learning as a regression (*matching*) task alleviates these issues. Consider any tractable conditional velocity field $v(x, t \mid x_1)$ that transforms p_0 into delta-measure δ_{x_1} through a conditional probability path $p_{X_t|x_1}$ with tractable expectations. One can then learn $v(x, t \mid x_1)$ via

$$\mathbb{E}\left\|v\big(X_{T|x_1}, T \;\big|\; x_1\big) - \hat{v}\big(X_{T|x_1}, T\big)\right\|_2 \to \min_{\hat{v}}, \tag{6}$$

where $T \sim \mathrm{U}[0; 1]$ and $X_{t|x_1} \sim p_{X_t|x_1}$. However, the goal is to learn *unconditional* velocity field. Thankfully, averaging (6) over $x_1 \sim p_1$ produces a valid objective for matching $v(x, t)$ [30]:

$$\mathbb{E}\left\|v(X_T, T \mid X_1) - \hat{v}(X_T, T)\right\|_2 \to \min_{\hat{v}}, \tag{7}$$

where $X_1 \sim p_1$ and $X_t = X_{t|X_1}$. If (7) attains its minima, then $\hat{v}$ transforms p_0 to p_t through (5) for any $t \in [0; 1]$. A common choice for p_t and $v(x, t \mid x_1)$ is *conditional optimal transport* [30]:

$$X_t = (1 - t)X_0 + t\, X_1 \quad v(x, t \mid x_1) = \frac{x_1 - x}{1 - t}, \tag{8}$$

where pairs (X_0, X_1) are sampled from a coupling distribution (typically independent) with marginals p_0 and p_1. However, other options also exist [31].

3 Related Works

Non-parametric methods pioneered the field by providing cheap MI estimates through binning [32], kernel [33] and k-nearest neighbors [34, 35] density estimation. However, recent studies show that these approaches fail horribly on high-dimensional and high-MI tasks [19, 22]. While random slicing scales these methods to higher dimensions [36], it also exhibits severe inherent limitations [37].

In contrast, complex parametric estimators have achieved relative practical success in dealing with difficult distributions. This family consist of *generative* and *discriminative* methods. While the generative estimators approximate $\mathbb{P}_{X,Y}$, $\mathbb{P}_X$ and $\mathbb{P}_Y$ from scratch, thus providing a plug-in estimate of MI (usually through Equation (2)), the discriminative approaches focus on the direct estimation of $\frac{\mathrm{d}\,\mathbb{P}_{X,Y}}{\mathrm{d}\,\mathbb{P}_X \otimes \mathbb{P}_Y}$ [20, 38].

Generative methods. Naïve application of generative models to approximate densities typically produce highly biased MI estimates due to misalignment of estimated $\hat{\mathbb{P}}_{X,Y}$, $\hat{\mathbb{P}}_X$ and $\hat{\mathbb{P}}_Y$ [20]. Subse-

[1] Hereinafter div and ∇ are with respect to x, if not stated otherwise.

quent research has addressed this issue by using a single model to estimate all three distributions simultaneously [22, 39, 40, 41, 42]. Among these methods, we highlight the **RFMI** estimator [43], which uses flow matching with a Gaussian prior to estimate $\mathbb{P}_X$ and $\mathbb{P}_{X|Y}$.

A separate subfamily of generative methods has recently emerged that avoids density estimation altogether, instead employing *diffusion models* and Girsanov's theorem [21, 23]. While these methods provide better estimates, they demand orders of magnitude more samples and compute. These estimators also typically fail when faced with long-tailed distributions. Overall, both families approximate the data distribution from scratch, thus performing unnecessary extra work.

Discriminative methods. While the direct estimation of density ratios is the most elegant approach, modern discriminative estimators face severe limitations. These include high demands on batch and sampling sizes [18, 44], as well as constraints on the level of mutual information they can reliably measure [20, 45]. Since all contemporary discriminative methods are classifier-based, these issues mirror the limitations of generalized energy-based models (GEBMs) [24]. For instance, GEBMs struggle to estimate the density ratio between two highly dissimilar distributions [46], a scenario that corresponds to high MI in (1).

The literature primarily addresses these problems in two ways. One line of research involves alternative classification objectives [47, 48], which offer modest practical improvements but fail to overcome the fundamental theoretical limitations. The other focuses on telescoping density ratio estimation [46, 49]. This approach achieves significantly higher accuracy by training multiple (potentially infinite, as in **DRE-∞** by K. Choi, C. Meng, Y. Song, and S. Ermon [49]) discriminators; however, it explicitly requires a path of distributions $\mathbb{P}_t$ with well-defined density ratios that interpolates between $\mathbb{P}_X \otimes \mathbb{P}_Y$ and $\mathbb{P}_{X,Y}$, which is not generally available out-of-the-box.

Relation to our approach. Our method is directly inspired by the **RFMI** and **DRE-∞**, taking the best from the generative and discriminative approaches. Similarly to RFMI, we leverage flow matching to learn $\mathbb{P}_{X,Y}$ (or $\mathbb{P}_{X|Y}$), but using $\mathbb{P}_X \otimes \mathbb{P}_Y$ (or $\mathbb{P}_X$) instead of the Gaussian prior.

This connects us to DRE-∞, which estimates the density ratio between two distributions in a continuous, bridge-like manner. However, instead of learning the density ratio evolution at any fixed point (which is often an ill-posed problem), we "follow" samples along the probability path, thus avoiding degenerate density ratios. For additional discussion, please refer to Section D.

4 Core Method

Consider an absolutely continuous X_0 and a smooth bijective mapping f. Define $X_1 = f(X_0)$. Due to (3) and, if f is parametrized by a velocity field v, due to (5)

$$\mathsf{h}(X_1) = \mathsf{h}(X_0) + \mathbb{E}\log\left|\det \frac{\partial f}{\partial x}(X_0)\right| \quad \overset{(5)}{\Longrightarrow} \quad \mathsf{h}(X_1) - \mathsf{h}(X_0) = \mathbb{E}\int_0^1 \operatorname{div} v(x_t(X_0), t)\,\mathrm{d}t$$

This equation requires solving ODE to simulate x_t. However, if expectations over $\mathbb{P}_{X_t}$ are tractable, a simulation-free result can be derived:

Lemma 4.1. For any $t \in [0; 1]$, let X_t satisfy $\partial X_t/\partial t = v(X_t, t)$. Let $T \sim \mathrm{U}[0; 1]$. Then

$$\mathsf{h}(X_1) - \mathsf{h}(X_0) = \mathbb{E}\operatorname{div} v(X_T, T)$$

In practice, however, f and v are unknown, but independent samples from p_0 and p_1 are available. Since none of the true distributions are tractable, we propose using flow matching to learn v from samples via (7) — see Algorithm 1. The estimation is Monte-Carlo-based — see Algorithm 2.

Note that estimated $\hat{v}$ typically does not produce the same probability path $\mathbb{P}_{X_t}$. To address this, we provide the following approximation and convergence analysis:

Theorem 4.2. Let $\mathbb{P}_{X,T}$ be such that $p_t(x) \overset{\text{def}}{=} p(x \mid t)$ and $K^2 \overset{\text{def}}{=} \mathbb{E}\left\|\nabla \log p_T(X)\right\|_2^2$ exist. Then, for any smooth vector field $\epsilon(x, t)$ satisfying $p_t(x)\|\epsilon(x, t)\| \in o\big(\|x\|^{1-d}\big)$ for any t,

$$|\mathbb{E}\operatorname{div}\epsilon(X, T)| \leq K\sqrt{\mathbb{E}\left\|\epsilon(X, T)\right\|_2^2}$$

Corollary 4.3. (FMDoE approximation error) Under the setup from Lemma 4.1, define $\mathrm{DoE}_u = \mathbb{E}\operatorname{div} u(X_T, T)$ for any $u(x,t)$. Let p_t and $(v - \hat{v})$ satisfy Theorem 4.2. Then

$$|\mathrm{DoE}_v - \mathrm{DoE}_{\hat{v}}| \leq K\sqrt{\mathbb{E}\,\|v(X_T,T) - \hat{v}(X_T,T)\|_2^2}$$

Proof. Substitute $\epsilon = v - \hat{v}$ in Theorem 4.2. □

Theorem 4.4. (FMDoE convergence rate) Let $\hat{v}_N$ be the trained velocity obtained from the minimization problem (9) using N i.i.d. copies of (X_0, X_1, T). Let $\widehat{\mathrm{DoE}}_{\hat{v}_N} = \frac{1}{M}\sum_{i=1}^{M} \operatorname{div} \hat{v}_N(X_i, T_i)$, where $\{(X_i, T_i)\}_{i=1}^{M}$ are i.i.d. copies of (X_T, T), independent of the training samples. Set $v^*(x,t) = \mathbb{E}[X_1 - X_0 \mid X_T = x, T = t]$ and $\mathrm{DoE}_{v^*} = \mathbb{E}[\operatorname{div} v^*(X_T, T)]$. Let $K^2 = \mathbb{E}\,\|\nabla \log p_T(X_T)\|^2 < \infty$, and assume $\operatorname{var}[\operatorname{div} \hat{v}_N(X_T, T) \mid \hat{v}_N] \leq s^2$. Then, under the assumptions of Theorem A.27 and Theorem 4.2, for sufficiently large N holds

$$\mathbb{E}\left|\mathrm{DoE}_{v^*} - \widehat{\mathrm{DoE}}_{\hat{v}_N}\right| \leq C_{\alpha,d} A^2 K (\log N)^{\alpha+3} N^{-\frac{1}{d+3}} + sM^{-\frac{1}{2}},$$

where $A = (1 + L_0)(1 + \sigma)^{\alpha+1}$, with $L_0 > 0$, $\alpha \geq 0$ and $\sigma > 0$, and $C_{\alpha,d}$ depends only on α and d.

A similar theory can be derived for $\mathsf{W}_p(\mathbb{P}_0, \mathbb{P}_1)$ between the marginals of X_0 and X_1:

Lemma 4.5. [50, Proposition 1.1] Let $\mathbb{P}_0, \mathbb{P}_1$ be absolutely continuous, $T \sim \mathrm{U}[0;1]$. Define $V(\mathbb{P}_0, \mathbb{P}_1) = \{v \mid v \text{ transforms } \mathbb{P}_0 \text{ into } \mathbb{P}_1 \text{ via (5)}\}$. Then,

$$\mathsf{W}_p(\mathbb{P}_0, \mathbb{P}_1) = \inf_{v \in V(\mathbb{P}_0, \mathbb{P}_1)} \sqrt[p]{\mathbb{E}\,\|v(X_T,T)\|_p^p} \qquad \text{where } \partial X_t/\partial t = v(X_t, t),\ \ X_0 \sim \mathbb{P}_0$$

Notice the similarity between the expressions in Theorem 4.2 and Lemma 4.5. Since $\mathsf{h}(X_1) - \mathsf{h}(X_0)$ is invariant to the choice of $v \in V(\mathbb{P}_0, \mathbb{P}_1)$, one can select v to attain the infimum in Lemma 4.5:

Corollary 4.6. Suppose Theorem 4.2 holds. Then $|\mathsf{h}(X_1) - \mathsf{h}(X_0)| \leq K \cdot \mathsf{W}_2(\mathbb{P}_0, \mathbb{P}_1)$.

Therefore, we also suggest estimating $\mathbb{E}\,\|v(X_T,T)\|_p^p$ as a cheap surrogate that is connected to both $\mathsf{W}_p(\mathbb{P}_0, \mathbb{P}_1)$ and $\mathsf{h}(X_1) - \mathsf{h}(X_0)$ (see line 7 in Algorithm 2).

Algorithm 1 FMDoE, training

Input: samples from $\mathbb{P}_{X_0,X_1}$, sampler from $\mathbb{P}_{X_t|X_1}$, initial velocity network $\hat{v}$.
Output: learned velocity network $\hat{v}$.
while not converged **do**
  Get $\{x_0^n, x_1^n, t^n\}_{n=1}^N \sim \mathbb{P}_{X_0,X_1} \otimes \mathrm{U}[0;1]$
  Using $\{x_0^n, x_1^n\}_{n=1}^N$, get $\{x_t^n\}_{n=1}^N \sim \mathbb{P}_{X_t|X_1}$
  $\mathcal{L}(\hat{v}) \leftarrow \frac{1}{N}\sum_{n=1}^{N} \|\hat{v}(x_t^n, t) - v(x_t^n, t \mid x_1^n)\|_2^2$
  Update $\hat{v}$ using $\partial\mathcal{L}/\partial\hat{v}$
end

Algorithm 2 FMDoE, estimation

Input: samples from $\mathbb{P}_{X_0,X_1}$, sampler from $\mathbb{P}_{X_t|X_1}$, velocity network $\hat{v}$.
Output: estimated $\mathsf{h}(X_1) - \mathsf{h}(X_0)$ and $\mathsf{W}_p(\mathbb{P}_0, \mathbb{P}_1)$.
Get $\{x_0^n, x_1^n, t^n\}_{n=1}^N \sim \mathbb{P}_{X_0,X_1} \otimes \mathrm{U}[0;1]$
Using $\{x_0^n, x_1^n\}_{n=1}^N$, get $\{x_t^n\}_{n=1}^N \sim \mathbb{P}_{X_t|X_1}$
$\widehat{\mathrm{DoE}} \leftarrow \frac{1}{N}\sum_{n=1}^{N} \operatorname{div} \hat{v}(x_t^n, t)$
$\widehat{\mathsf{W}}_p \leftarrow \sqrt[p]{\frac{1}{N}\sum_{n=1}^{N} \|\hat{v}(x_t^n, t)\|_p^p}$

5 Mutual Information

Our method can be readily extended to Mutual Information estimation. Recall that, in continuous case, MI can be expressed through the difference of entropies via (2). Each formula in (2) can be approximated through the FMDoE estimator by constructing flows either between the joint distribution and the product of the marginals, or between the marginal and conditional distributions.

jFMMI. Setting $\mathbb{P}_0 = \mathbb{P}_{X,Y}$ and $\mathbb{P}_1 = \mathbb{P}_X \otimes \mathbb{P}_Y$, we learn a velocity field $v : \mathcal{X} \times \mathcal{Y} \times [0;1] \to \mathcal{X} \times \mathcal{Y}$ that performs *forward* transformation $\mathbb{P}_0$ to $\mathbb{P}_1$. Alternatively, one may consider the *reverse* transformation by setting $\mathbb{P}_0 = \mathbb{P}_X \otimes \mathbb{P}_Y$ and $\mathbb{P}_1 = \mathbb{P}_{X,Y}$ to be the joint distribution. These two formulations are theoretically equivalent, though their practical behavior may differ slightly, as observed in our experiments. By Lemma 4.1, for $Z_0 \sim \mathbb{P}_0$, $Z_1 \sim \mathbb{P}_1$, and $T \sim \mathrm{U}[0;1]$, we have $\mathsf{h}(Z_1) - \mathsf{h}(Z_0) = \mathbb{E}\operatorname{div} v(Z_T, T)$, which is equivalent to

$$\mathsf{I}(X;Y) = \mathsf{h}(X) + \mathsf{h}(Y) - \mathsf{h}(X,Y) = \mathbb{E}\,\mathrm{div}\, v(Z_T, T).$$

In practice, a batch from $\mathbb{P}_{X,Y}$ is converted into samples from $\mathbb{P}_X \otimes \mathbb{P}_Y$ by independently permuting the x's and y's coordinates. Training and estimation then follow Algorithm 1 and Algorithm 2.

cFMMI. In the conditional approach, for each fixed y we set $\mathbb{P}_0 = \mathbb{P}_{X|Y=y}$ and $\mathbb{P}_1 = \mathbb{P}_X$ and learn $v : \mathcal{X} \times \mathcal{Y} \times [0,1] \to \mathcal{X}$ that transforms $\mathbb{P}_X$ to $\mathbb{P}_{X|Y=y}$. Applying Lemma 4.1 for a given y, we have $\mathsf{h}(X) - \mathsf{h}(X \mid Y) = \mathbb{E}_Y\, \mathbb{E}[\mathrm{div}\, v(X_T, Y, T) \mid Y] = \mathbb{E}[\mathrm{div}\, v(X_T, Y, T)]$. Thus,

$$\mathsf{I}(X;Y) = \mathsf{h}(X) - \mathsf{h}(X \mid Y) = \mathbb{E}\,\mathrm{div}\, v(X_T, Y, T).$$

In practice, unconditional samples are obtained by shuffling only the x's samples while keeping the corresponding y's fixed. Training and estimation similarly use algorithms from the previous section.

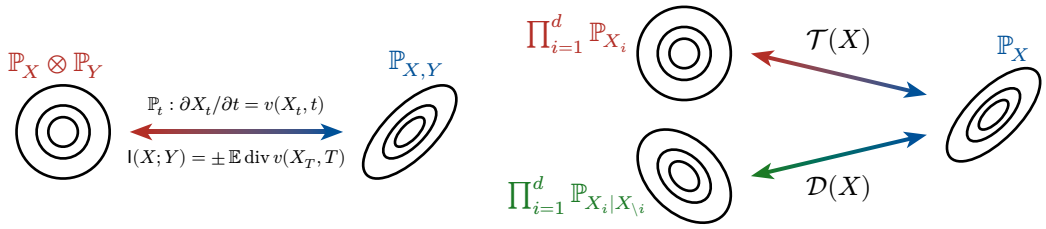


Figure 3: Flow Matching Mutual Information and O-Information estimation

Analysis. Since jFMMI operates in $(d_X + d_Y)$-dimensional space, while cFMMI works in d_X (or d_Y) dimensions, the latter can be more efficient in high-dimensional setups. These estimators also benefit from the Hutchinson trace estimator for efficient divergence computation. Under the assumptions of Corollary 4.3 and Theorem 4.2, both estimators are consistent, with approximation error bounded by the flow matching error and the Lipschitz constants of the true distributions.

6 O-Information

Although MI serves as a fundamental measure for quantifying non-linear dependence between two random variables, its applicability becomes limited when analyzing systems comprising more than two variables, which is a common scenario e.g., in neuroscience [51] and climatology [52]. A natural multivariate extension of MI is the interaction information, or co-information, defined via an inclusion-exclusion formula. While co-information can reflect the interplay between synergy and redundancy, its scalability and interpretability are limited for large systems. A more computable and interpretable generalization is provided by O-information [53, 54], defined as

$$\Omega(X) = \mathcal{T}(X) - \mathcal{D}(X),$$

where the redundancy term $\mathcal{T}(X)$ and the synergy term $\mathcal{D}(X)$ are given by

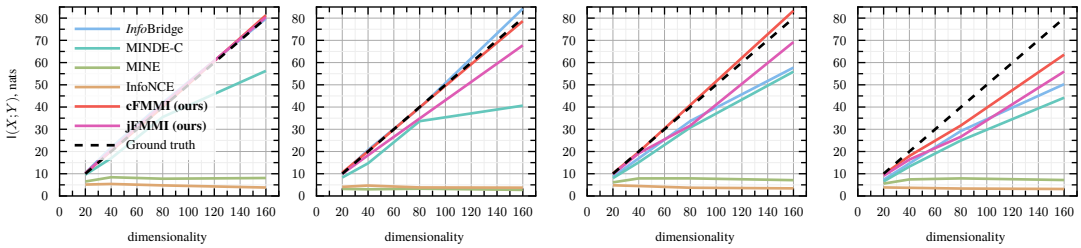
$$\mathcal{T}(X) = \sum_{i=1}^{d} \mathsf{h}(X_i) - \mathsf{h}(X) \qquad \mathcal{D}(X) = \mathsf{h}(X) - \sum_{i=1}^{d} \mathsf{h}\Big(X_i \;\Big|\; X_{\setminus i}\Big).$$

Here $X_{\setminus i}$ denotes $(d-1)$-dimensional vector obtained by removing the i-th component from X. Intuitively, O-Information measures the overall strength and nature of interactions in a multivariate system. A positive $\Omega(X)$ indicates redundancy dominance, meaning that similar information is duplicated across different variables. Conversely, a negative $\Omega(X)$ corresponds to synergy, reflecting information that emerges only from the joint consideration of all variables, beyond what can be captured by pairwise interactions.

Although less straightforward, O-information also admits a DoE representation, enabling its estimation via our framework. Consider a $d \times d$ matrix M of i.i.d. rows from $\mathbb{P}_X$. By construction,

$$\mathsf{h}(\mathrm{diag}\,\mathrm{M}) = \sum_{i=1}^{d} \mathsf{h}(X_i) \qquad \mathsf{h}(\mathrm{diag}\,\mathrm{M} \mid \mathrm{offdiag}\,\mathrm{M}) = \sum_{i=1}^{d} \mathsf{h}\Big(X_i \;\Big|\; X_{\setminus i}\Big),$$

Therefore, $\mathcal{T}(X) = \mathsf{h}(\mathrm{diag}\,\mathrm{M}) - \mathsf{h}(X)$ and $\mathcal{D}(X) = \mathsf{h}(X) - \mathsf{h}(\mathrm{diag}\,\mathrm{M} \mid \mathrm{offdiag}\,\mathrm{M})$.



(a) Corr. $\mathcal{N}$ (b) H.-C. Corr. $\mathcal{N}$ (c) Corr. U (d) Sm. U

Figure 4: Comparison of MI estimates across dimensions and MI for high mutual information.

While one can apply FMDoE to approximate each term individually, we suggest a more artful approach. Instead of learning two velocity fields, we employ a single $\hat{v}$ that takes two-dimensional $t \in \mathbb{R}^2$ as a time input, combined with one-hot encoding s of the the path. In this amortized setup, $t = (0, 0)$ corresponds to $\mathbb{P}_X$, $t = (1, 0)$ — to $\prod_{i=1}^{d} \mathbb{P}_{X_i}$, and $t = (0, 1)$ — to $\prod_{i=1}^{d} \mathbb{P}_{X_i|X_{\setminus i}}$. The regression objective (7) and affine paths (8) are extended trivially to this multidimensional time approach. We argue that this technique is also applicable to other information-theoretic quantities, e.g., to Transfer Entropy [55].

7 Experiments

7.1 Mutual Information

To evaluate our estimator, we follow the experimental protocol from Section 5.4 in [23]. We approximate v via a two-layer MLP with hidden dimensionality 512. AdamW optimizer [56] with learning rate 10^{-3} is employed, 10^4 gradient steps are used. Training set size is 10^5 samples, test size is 10^4. For details, please, refer to the supplementary source code.

The benchmark features *Correlated Normal*, *Half-cube Correlated Normal*, *Correlated Uniform* and *Smoothed Uniform* distributions from the `mutinfo` package [57]. The results are provided in Figure 4. Note that, despite using $\times$ 20 fewer gradient steps, cFMMI consistently outperforms MINDE and *Info*Bridge — two state-of-the-art estimators for high-dimensional data. Compared to these methods, FMMI estimates were also more stable and did not require averaging over the last n epochs.

To quantify the behavior of FMMI in extreme dimensionalities, we perform a $d \in \{32, 64, ..., 512\}$ and $N \in \{10^2, ..., 10^5\}$ sweep using Correlated Normal and Smoothed Uniform distributions and $\mathsf{I}(X;Y) = d \times 2$ nats. These results are presented in Table 1 and 2.

Finally, we conduct additional experiments and compare our estimator to other baselines in Section E.

Table 1: FMMI estimates for the Correlated Normal synthetic test. The GT value of MI is $d \times 2$ nats.

Method	d	$N = 100$	316	1000	3162	10000	31622	100000
cFMMI	32	91.7 ± 7.3	89.63 ± 0.83	72.88 ± 0.37	65.99 ± 0.15	63.698 ± 0.069	64.25 ± 0.16	63.92 ± 0.12
	64	117.9 ± 5.9	139.0 ± 8.9	136.7 ± 1.2	128.41 ± 0.38	128.99 ± 0.22	128.46 ± 0.27	128.31 ± 0.56
	128	274 ± 11	287 ± 10	275.9 ± 5.8	261.56 ± 0.47	252.12 ± 0.61	256.35 ± 0.51	254.80 ± 0.42
	256	420 ± 19	487 ± 20	548 ± 11	528.4 ± 1.6	509.0 ± 1.2	514.71 ± 0.89	507.43 ± 0.84
	512	365 ± 22	1072 ± 51	1348 ± 30	1059.0 ± 3.8	1011.0 ± 1.8	975.8 ± 1.8	967.7 ± 3.6
jFMMI	32	30.48 ± 0.97	63.6 ± 1.3	62.95 ± 0.46	64.32 ± 0.24	64.487 ± 0.094	64.66 ± 0.17	64.76 ± 0.27
	64	14.8 ± 2.1	79.0 ± 1.9	117.8 ± 1.1	122.61 ± 0.63	127.63 ± 0.28	128.97 ± 0.36	128.87 ± 0.42
	128	6.8 ± 9.7	70.5 ± 2.8	163.0 ± 1.7	210.31 ± 0.77	248.80 ± 0.76	254.47 ± 0.47	255.06 ± 0.44
	256	33 ± 45	0.0 ± 0.0	172.3 ± 2.6	361.7 ± 1.8	444 ± 13	500.5 ± 7.7	481.2 ± 1.1
	512	1130 ± 31	0.0 ± 0.0	0.0 ± 0.0	332.7 ± 3.2	553 ± 22	675 ± 51	1008 ± 22

Table 2: FMMI results for the Smoothed Uniform synthetic test. The GT value of MI is $d \times 2$ nats.

Method	d	$N = 100$	316	1000	3162	10000	31622	100000
cFMMI	32	100.4 ± 2.1	89.99 ± 0.67	71.52 ± 0.48	63.79 ± 0.23	63.258 ± 0.091	64.77 ± 0.17	62.445 ± 0.067
	64	128.7 ± 6.4	152.4 ± 5.1	134.9 ± 1.7	125.25 ± 0.37	125.08 ± 0.24	129.52 ± 0.20	124.70 ± 0.19
	128	260.6 ± 8.7	315 ± 11	286.3 ± 4.9	256.13 ± 0.45	258.25 ± 0.67	253.37 ± 0.45	246.92 ± 0.41
	256	455 ± 19	466 ± 26	612 ± 13	525.4 ± 1.4	498.0 ± 1.1	500.1 ± 1.5	491.7 ± 1.0
	512	510 ± 37	1076 ± 54	1171 ± 44	1092.6 ± 3.0	999.8 ± 1.6	958.8 ± 3.1	941.9 ± 4.8
jFMMI	32	30.7 ± 1.1	65.16 ± 0.87	61.42 ± 0.59	63.77 ± 0.24	61.39 ± 0.37	60.30 ± 0.19	60.28 ± 0.31
	64	19.2 ± 2.2	76.7 ± 1.1	115.6 ± 1.3	119.79 ± 0.58	122.99 ± 0.33	122.80 ± 0.28	122.70 ± 0.25
	128	40 ± 23	70.2 ± 2.5	159.0 ± 1.7	208.01 ± 0.54	241.99 ± 0.24	247.47 ± 0.57	248.5 ± 4.1
	256	244 ± 140	0.0 ± 0.0	186.7 ± 1.9	359.2 ± 2.2	458.4 ± 2.0	467.65 ± 0.84	469.76 ± 0.76
	512	1210 ± 84	0.0 ± 0.0	0.0 ± 0.0	317.7 ± 6.8	558.8 ± 4.7	719 ± 49	919 ± 15

Table 3: FMOI results for the AR(1) synthetic test with $d = 32$ and varying ρ. GT denotes the ground truth O-information computed via Corollary 7.2.

Ω, nats	$\rho = 0.1$	0.5	0.9	0.99	0.999	0.9999
GT	0.0015	0.968	16.01	48.51	82.84	117.36
Estimate	0.4 ± 1.0	1.3 ± 1.1	15.97 ± 0.45	51.46 ± 0.26	95.33 ± 0.95	126.6 ± 3.2

7.2 O-Information

In this section, we validate our **FMOI** estimator on a set of synthetic benchmarks and real data. For all experiments, we reuse the network from the previous section, but reduce the hidden dimensionality to 128. Other settings (optimizer, learning rate, etc.) are also the same. For more details, please, refer to the supplementary source code.

Synthetic tests. For the synthetic tests, we adopt a Gaussian-based methodology from [54] but employ a different family of covariance matrices. Specifically, we use the order-one Gaussian autoregressive process, where the covariance matrix is $(\Sigma)_{ij} = \rho^{|i-j|}$ for $\rho \in [0, 1)$ and dimension $d \in \mathbb{N}$. This construction allows us to control the strength of redundancy and synergy via the parameter ρ. For this distribution, the total correlation and dual total correlation admit closed-form expressions.

Proposition 7.1. For a d-dimensional Gaussian vector $X \sim \mathcal{N}(m, \Sigma)$, the O-information is given by

$$\Omega(X) = \frac{1}{2}\left(\sum_{i=1}^{d} \log \sigma_i^2 + (d-2)\log\det\Sigma - \sum_{i=1}^{d} \log\det\Sigma_{\setminus i}\right),$$

where $\Sigma_{\setminus i}$ is Σ without its i-th row and column.

Corollary 7.2. For the order-one Gaussian autoregressive process with covariance matrix $(\Sigma)_{ij} = \rho^{|i-j|}$ for any $\rho \in [0, 1)$ and $d \in \mathbb{N}$, the O-information is given by $\Omega(X) = \left(1 - \frac{d}{2}\right)\ln(1 - \rho^4)$.

We evaluate FMOI on this benchmark with $d = 32$ and varying values of ρ. For each ρ, we generate a single covariance matrix Σ and sample 10^3 training and test points from $\mathcal{N}(0, \Sigma)$. FMOI is then trained for 10^4 gradient steps in a single-batch regime. The results are summarized in Table 3.

Embedding quality evaluation. Contemporary SSRL methods often produce good compressed representations [58], but their assessment is typically based on downstream task performance. Currently, only simple linear metrics are used to assess their quality in an unsupervised manner [59]. We expect that O-information could advance this methodology, as it directly measures redundancy and synergy in a random vector.

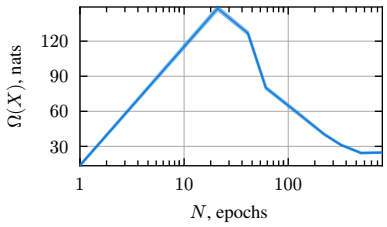


Figure 5: O-information dynamics during representation learning.

In our experiment, we train ResNet-18 on CIFAR-10 as in [58] and measure the O-information (10 runs) of embeddings after N epochs. The results are reported in Figure 5, where the low O-information at the epoch zero is explained by random initialization, while the initial spike and subsequent decaying trend indicate that embeddings become more synergistic during training.

fMRI data. We use fMRI data from the Human Connectome Project (HCP) preprocessed by [60], consisting of 581 healthy participants performing seven tasks that results in 14 brain states and 8134 fMRI data units. fMRI measures BOLD signals as a proxy for neural activity, yielding high-dimensional, noisy spatiotemporal data. For more details, please, refer to Section F. The complex dependence structure across brain regions, subjects, and experimental conditions makes fMRI a natural application domain for information-theoretic analysis.

Kirova et al. [60] identified unique sets of brain regions whose activity is most strongly associated with cognitive conditions, achieving high classification accuracy. Authors represented each fMRI data unit by the averaged activity values across brain regions. By focusing on the mean activity values, they aim to demonstrate that each brain state is associated with a unique network of regional activations, which can serve as a reliable signature for classification. In this work, we compute the O-Information specifically on these vectors of mean regional activity for each brain state. This allows us to quantify higher-order statistical dependencies among the selected brain regions and to characterize whether each cognitive state is encoded in a predominantly independent, redundant, or synergistic manner at the level of regional mean activations.

The results presented in [60] suggest that for brain states with high classification accuracy (above 0.9), it is possible to identify a small set of unique brain regions whose mean activity is sufficient for reliable classification. In contrast, for the remaining states, the number of contributing regions is substantially larger and, in some cases, approaches coverage of most regions in the brain atlas, indicating more distributed neural representations. A detailed breakdown of the number of identified regions is provided in Section F.

We compute the O-information for all brain states, including both those characterized by a small number of identified regions and those involving a large, distributed set of regions. For each setup from the fMRI dataset, we perform 10 independent runs with different random seeds. In each run, the dataset of 581 samples is split into training (70%) and test sets (30%). Due to the limited dataset size, training employs a full-batch update and thus one gradient step per each of 10^4 epochs. For each run, we compute the median O-information over the last 100 training epochs on the test set. The results are summarized in Figure 6 using box plots per region, where each box represents the distribution of the 10 seed-wise median values.

For states with a small number of significant regions, $\Omega(X) \approx 0$. This behavior indicates that the selected regions encode largely independent or weakly interacting information, suggesting that the classifier relies on a compact set of complementary features with minimal redundancy or synergy. In this regime, near-zero O-information supports the interpretation that the identified regions are well localized and capture distinct aspects of the underlying neural processes.

In contrast, for states associated with a large number of significant regions, we observe $\Omega(X) \gg 0$. This pattern reflects strong higher-order statistical dependencies and redundancy across brain areas, indicating that information about these cognitive states is represented in a highly distributed and overlapping manner. Such elevated O-information is consistent with more global and integrative brain states, where multiple regions contribute correlated or redundant signals to the overall representation.

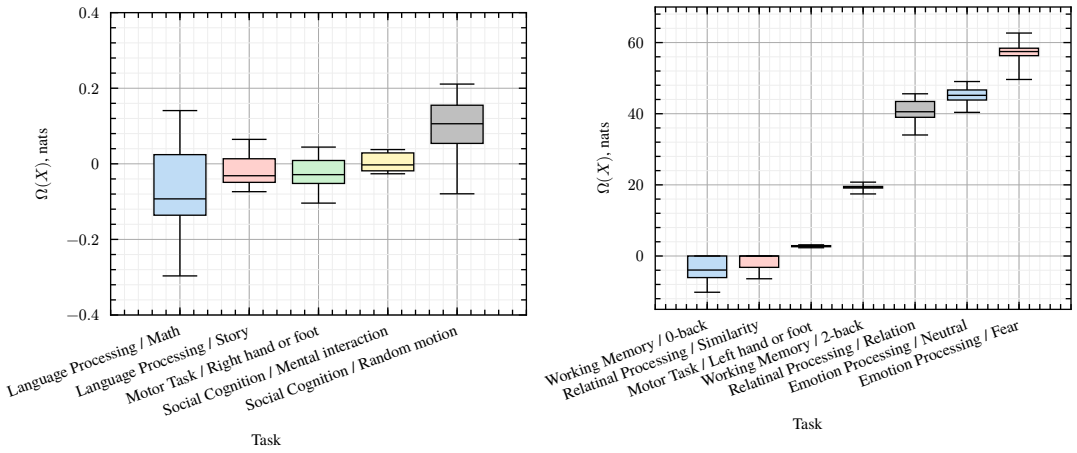


Figure 6: O-Information estimates for fMRI data (10 seeds): setups with low **(left)** and high **(right)** absolute value of O-Information. Low or even negative values indicate a successful selection of the most relevant regions; high positive values suggest that selected regions encode similar information.

8 Discussion

To date, entropy and mutual information estimation remains extremely challenging, particularly for high-dimensional random vectors with complex dependencies. To address this, we introduce novel estimators based on flow matching — **FMDoE** and **FMMI** respectively. FMMI is a special case of FMDoE and comes in two variants: a conditional (**cFMMI**) and a joint (**jFMMI**) formulation.

In contrast to traditional generative MI estimators, our method constructs a flow between two data distributions, making it more akin to discriminative approaches. Evaluation on synthetic benchmarks demonstrates the superiority of the proposed method: while its computational load lies between that of discriminative and generative families, FMMI outperforms both in terms of accuracy, especially on high-dimensional and long-tailed data.

We provide theoretical guarantees proving our method's consistency. Furthermore, we propose Wasserstein-based variants of our estimators and establish a connection to FMDoE and FMMI.

Finally, a novel O-Information estimator (**FMOI**) is proposed. This method combines the FMDoE approach with an efficient multidimensional time amortization scheme, enabling fast training with only one network and two forward passes per batch. Applying FMOI to functional MRI data reveals redundancy in some groups of brain regions and a synergistic nature in others.

Limitations. Although FMMI is more lightweight than state-of-the-art diffusion-based MI estimators, it remains relatively demanding in terms of both computational and sample complexity.

Reproducibility statement. To ensure reproducibility of our results, we provide complete proofs in Section B and experimental details in Section 7. We also provide our source code in the supplementary material.

References

[1] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. Cambridge University Press, 2024. [Online]. Available: https://books.google.ru/books?id=CySo0AEACAAJ

[2] A. Asadi, E. Abbe, and S. Verdu, "Chaining Mutual Information and Tightening Generalization Bounds," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Associates, Inc., 2018, p. . [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/8d7628dd7a710c8638dbd22d4421ee46-Paper.pdf

[3] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, "Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, p. . [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/05ae14d7ae387b93370d142d82220f1b-Paper.pdf

[4] B. Duong and T. Nguyen, "Conditional Independence Testing via Latent Representation Learning," in *2022 IEEE International Conference on Data Mining (ICDM)*, Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2022, pp. 121–130. doi: 10.1109/ICDM54844.2022.00022.

[5] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994, doi: 10.1109/72.298224.

[6] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005, doi: 10.1109/TPAMI.2005.159.

[7] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, Jan. 2014.

[8] R. D. Hjelm *et al.*, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bklr3j0cKX

[9] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning Representations by Maximizing Mutual Information Across Views," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, p. . [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/ddf354219aac374f1d40b7e760ee5bb7-Paper.pdf

[10] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep Graph Infomax," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rklz9iAcKQ

[11] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, "On Mutual Information Maximization for Representation Learning," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rkxoh24FPH

[12] I. Butakov, A. Semenenko, A. Tolmachev, A. Gladkov, M. Munkhoeva, and A. Frolov, "Efficient Distribution Matching of Representations via Noise-Injected Deep InfoMax," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=mAmCdASmJ5

[13] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2015.

[14] R. Shwartz-Ziv and N. Tishby, "Opening the Black Box of Deep Neural Networks via Information." 2017.

[15] Z. Goldfeld *et al.*, "Estimating Information Flow in Deep Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., in Proceedings of Machine Learning Research, vol. 97. PMLR, Sept. 2019, pp. 2299–2308. [Online]. Available: https://proceedings.mlr.press/v97/goldfeld19a.html

[16] I. Butakov, A. Tolmachev, S. Malanchuk, A. Neopryatnaya, A. Frolov, and K. Andreev, "Information Bottleneck Analysis of Deep Neural Networks via Lossy Compression," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=huGECz8dPp

[17] Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy, "Convergence of Smoothed Empirical Measures With Applications to Entropy Estimation," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4368–4391, 2020, doi: 10.1109/TIT.2020.2975480.

[18] D. McAllester and K. Stratos, "Formal Limitations on the Measurement of Mutual Information," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, S. Chiappa and R. Calandra, Eds., in Proceedings of Machine Learning Research, vol. 108. PMLR, Aug. 2020, pp. 875–884. [Online]. Available: https://proceedings.mlr.press/v108/mcallester20a.html

[19] F. Czyż Paweł and Grabowski, J. Vogt, N. Beerenwinkel, and A. Marx, "Beyond Normal: On the Evaluation of Mutual Information Estimators," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 16957–16990. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/36b80eae70ff629d667f210e13497edf-Paper-Conference.pdf

[20] J. Song and S. Ermon, "Understanding the Limitations of Variational Mutual Information Estimators," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1x62TNtDS

[21] G. Franzese, M. BOUNOUA, and P. Michiardi, "MINDE: Mutual Information Neural Diffusion Estimation," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=0kWd8SJq8d

[22] I. Butakov, A. Tolmachev, S. Malanchuk, A. Neopryatnaya, and A. Frolov, "Mutual Information Estimation via Normalizing Flows," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=JiQXsLvDls

[23] S. Kholkin, I. Butakov, E. Burnaev, N. Gushchin, and A. Korotin, "InfoBridge: Mutual Information estimation via Bridge Matching," in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=y8Kzu9SKpv

[24] M. Arbel, L. Zhou, and A. Gretton, "Generalized Energy Based Models," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=0PtUPB9z6qK

[25] S. Ozair, C. Lynch, Y. Bengio, A. van den Oord, S. Levine, and P. Sermanet, "Wasserstein Dependency Measure for Representation Learning," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, p. . [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/f9209b7866c9f69823201c1732cc8645-Paper.pdf

[26] E. G. Tabak and C. V. Turner, "A Family of Nonparametric Density Estimation Algorithms," *Communications on Pure and Applied Mathematics*, vol. 66, no. 2, pp. 145–164, 2013, doi: https://doi.org/10.1002/cpa.21423.

[27] E. G. Tabak and E. Vanden-Eijnden, "Density estimation by dual ascent of the log-likelihood," *Commun. Math. Sci.*, vol. 8, no. 1, pp. 217–233, 2010.

[28] M. Hutchinson, "A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines," *Communication in Statistics- Simulation and Computation*, vol. 18, pp. 1059–1076, Jan. 1989, doi: 10.1080/03610919008812866.

[29] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, and D. Duvenaud, "Scalable Reversible Generative Models with Free-form Continuous Dynamics," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rJxgknCcK7

[30] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow Matching for Generative Modeling." [Online]. Available: https://arxiv.org/abs/2210.02747

[31] Y. Lipman *et al.*, "Flow Matching Guide and Code." [Online]. Available: https://arxiv.org/abs/2412.06264

[32] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, no. 3, pp. 233–248, 1989, doi: https://doi.org/10.1016/0165-1684(89)90132-1.

[33] Y.-I. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Phys. Rev. E*, vol. 52, no. 3, pp. 2318–2321, Sept. 1995, doi: 10.1103/PhysRevE.52.2318.

[34] L. F. Kozachenko and N. N. Leonenko, "Sample Estimate of the Entropy of a Random Vector," *Problems Inform. Transmission*, vol. 23, no. 2, pp. 95–101, 1987.

[35] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 6, p. 66138, June 2004, doi: 10.1103/PhysRevE.69.066138.

[36] Z. Goldfeld and K. Greenewald, "Sliced Mutual Information: A Scalable Measure of Statistical Dependence," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=27qon5Ut4PSl

[37] A. Semenenko, I. Butakov, A. Frolov, and I. Oseledets, "Curse of Slicing: Why Sliced Mutual Information is a Deceptive Measure of Statistical Dependence," in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=KxeBgh1zWr

[38] M. Federici, D. Ruhe, and P. Forré, "On the Effectiveness of Hybrid Mutual Information Estimation." [Online]. Available: https://arxiv.org/abs/2306.00608

[39] B. Duong and T. Nguyen, "Diffeomorphic Information Neural Estimation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, pp. 7468–7475, June 2023, doi: 10.1609/aaai.v37i6.25908.

[40] C. Dahlke and J. Pacheco, "Flow-based Variational Mutual Information: Fast and Flexible Approximations," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=spDUv05cEq

[41] Y. Chen, Z. Ou, A. Weller, and M. Gutmann, "Neural Mutual Information Estimation with Vector Copulas," in *The Thirty-ningth Annual Conference on Neural Information Processing Systems*, 2025.

[42] H. Ni and M. Lotz, "A Neural Difference-of-Entropies Estimator for Mutual Information." [Online]. Available: https://arxiv.org/abs/2502.13085

[43] C. Wang, G. Franzese, A. Finamore, and P. Michiardi, "RFMI: Estimating Mutual Information on Rectified Flow for Text-to-Image Alignment," in *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025. [Online]. Available: https://openreview.net/forum?id=8CogNH4vBp

[44] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding." [Online]. Available: https://arxiv.org/abs/1807.03748

[45] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On Variational Bounds of Mutual Information," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., in Proceedings of Machine Learning Research, vol. 97. PMLR, Sept. 2019, pp. 5171–5180. [Online]. Available: https://proceedings.mlr.press/v97/poole19a.html

[46] B. Rhodes, K. Xu, and M. U. Gutmann, "Telescoping Density-Ratio Estimation," in *The Thirty-fourth Annual Conference on Neural Information Processing Systems*, 2020. [Online]. Available: https://arxiv.org/abs/2006.12204

[47] R. Liao, D. Moyer, P. Golland, and W. M. Wells, "DEMI: Discriminative Estimator of Mutual Information." [Online]. Available: https://arxiv.org/abs/2010.01766

[48] N. A. Letizia, N. Novello, and A. M. Tonello, "Mutual Information Estimation via f-Divergence and Data Derangements," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=PThi9hf9UT

[49] K. Choi, C. Meng, Y. Song, and S. Ermon, " Density Ratio Estimation via Infinitesimal Classification ," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, G. Camps-Valls, F. J. R. Ruiz, and I. Valera, Eds., in Proceedings of Machine Learning Research, vol. 151. PMLR, 28 2022, pp. 2552–2573. [Online]. Available: https://proceedings.mlr.press/v151/choi22a.html

[50] J.-D. Benamou and Y. Brenier, "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem," *Numer. Math. (Heidelb.)*, vol. 84, no. 3, pp. 375–393, Jan. 2000.

[51] E. Ganmor, R. Segev, and E. Schneidman, "Sparse low-order interaction network underlies a highly correlated and learnable neural population code," *Proceedings of the National Academy of Sciences*, vol. 108, no. 23, pp. 9679–9684, 2011, doi: 10.1073/pnas.1019641108.

[52] J. Runge *et al.*, "Inferring causation from time series in Earth system sciences," *Nature communications*, vol. 10, no. 1, p. 2553, 2019.

[53] F. E. Rosas, P. A. M. Mediano, M. Gastpar, and H. J. Jensen, "Quantifying high-order interdependencies via multivariate extensions of the mutual information," *Phys. Rev. E*, vol. 100, no. 3, p. 32305, Sept. 2019, doi: 10.1103/PhysRevE.100.032305.

[54] M. Bounoua, G. Franzese, and P. Michiardi, "SΩI: Score-based O-Information Estimation," in *ICML 2024, 41st International Conference on Machine Learning*, 2024.

[55] S. P. G. Munoz, M. Bounoua, G. Franzese, P. Michiardi, and M. Filippone, "TENDE: Transfer Entropy Neural Diffusion Estimation." [Online]. Available: https://arxiv.org/abs/2510.14096

[56] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[57] I. Butakov *et al.*, "MUTINFO." [Online]. Available: https://github.com/VanessB/mutinfo

[58] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, 2020, pp. 1597–1607.

[59] A. Tsitsulin, M. Munkhoeva, and B. Perozzi, "Unsupervised embedding quality evaluation," in *Topological, Algebraic and Geometric Learning Workshops 2023*, 2023, pp. 169–188.

[60] V. Kirova, D. Kadieva, D. Vlasenko, I. B. Blank, and F. Ratnikov, "Dynamic Functional Connectivity Features for Brain State Classification: Insights from the Human Connectome Project." [Online]. Available: https://arxiv.org/abs/2510.05325

[61] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks," *Journal of Machine Learning Research*, vol. 20, no. 63, pp. 1–17, 2019.

[62] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A distribution-free theory of nonparametric regression*. Springer, 2002.

[63] U. Grenander, *Abstract Inference*. New York: John Wiley & Sons, 1981.

[64] P. Rigollet, "18. s997: High dimensional statistics," *Lecture Notes), Cambridge, MA, USA: MIT OpenCourseWare*, vol. 44, 2015.

[65] D. Yarotsky, "Error bounds for approximations with deep ReLU networks," *Neural networks*, vol. 94, pp. 103–114, 2017.

[66] I. Covert, S. M. Lundberg, and S.-I. Lee, "Understanding Global Feature Contributions With Additive Importance Measures," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Curran Associates, Inc., 2020, pp. 17212–17223. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/c7bf0b7c1a86d5eb3be2c722cf2cf746-Paper.pdf

[67] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006.

[68] "1200 Subjects Data Release Reference Manual." Accessed: Jan. 06, 2026. [Online]. Available: https://www.humanconnectome.org/study/hcp-young-adult

[69] D. M. Barch *et al.*, "Function in the human connectome: task-fMRI and individual differences in behavior," *Neuroimage*, vol. 80, pp. 169–189, Oct. 2013.

A Convergence Analysis

***Assumption* A.1.** (Independence) Denote $X_t = (1-t)X_0 + tX_1,\ Z = (X_T, T),\ Y = X_1 - X_0$, where Z has law μ on $\mathbb{R}^D$ with $D = d+1$, and let

$$T \sim \mathrm{U}[0;1], \quad T \perp\!\!\!\perp (X_0, X_1), \quad X_0 \perp\!\!\!\perp X_1.$$

***Assumption* A.2.** (Sub-Gaussian endpoints) There exists $\sigma > 0$ such that for all $s \in \mathbb{R}$, all $v \in \mathbb{R}^d$ and each $k \in \{0,1\}$,

$$\mathbb{E}\exp(s\langle X_k, v\rangle) \le \exp((\sigma^2 s^2 \|v\|^2)/2).$$

***Assumption* A.3.** (Lipschitz velocity) Denote $v^*(z) = \mathbb{E}[Y \mid Z = z]$. For every $R > 1$, the restriction of v^* to $Q_R = [-R, R]^d \times [0,1]$ satisfies

$$\|v^*(z) - v^*(z')\| \le L_R \|z - z'\|, \quad L_R = L_0(1+R)^\alpha,$$

for some $L_0 > 0$ and $\alpha \ge 0$.

Throughout this section $\|\cdot\|$ is the Euclidean norm and $\|\cdot\|_\infty$ is the coordinate-wise maximum norm.

***Definition* A.4.** (ReLU network) We call an *architecture* a directed acyclic graph whose nodes are partitioned into an input layer (layer 0, with D nodes), $\ell - 1$ hidden layers and an output layer (layer ℓ), with edges directed from any earlier layer to any later one, so that skip connections are allowed. Every hidden node applies $\sigma(u) = \max(u, 0)$ and the output node is affine. Its *depth* is ℓ and its *parameters* are one weight per edge and one bias per non-input node. A *network* is an architecture together with a real value for every parameter, computing a function by the standard forward pass.

***Definition* A.5.** For integers $P \ge 2$, $\ell \ge 2$ let $\mathcal{V}(P, \ell)$ be the set of scalar functions $v : \mathbb{R}^D \to \mathbb{R}$ computed by some network with one output, at most P parameters and at most ℓ layers, and let

$$\mathcal{V}^d(P, \ell) = \{v = (v_1, ..., v_d) : v_k \in \mathcal{V}(P, \ell) \text{ for each } k\}$$

be the d-fold product of d scalar networks with at most P parameters per coordinate.

***Remark* A.6.** Note that the classes are nested: $\mathcal{V}(P, \ell) \subseteq \mathcal{V}(P', \ell')$ and $\mathcal{V}^d(P, \ell) \subseteq \mathcal{V}^d(P', \ell')$ whenever $P' \ge P$ and $\ell' \ge \ell$.

***Remark* A.7.** The model size of $\mathcal{V}^d$ is imposed per coordinate, which allows reduction to d scalar problems (since $\hat{\mathcal{R}}(v) = \sum_{k=1}^d \hat{\mathcal{R}}^{(k)}(v_k)$ as will be shown in what follows, and an empirical risk minimizer over $\mathcal{V}^d$ is obtained coordinatewise).

Theorem A.8. [61, Theorem 7] There is $C > 0$ such that for every $P \ge 2$ and $\ell \ge 1$ holds

$$V_{\mathcal{V}(P,\ell)^+} \le CP\ell \log P.$$

Proof of Theorem A.8. Recall that $V_{\mathcal{V}(P,\ell)^+}$ is the VC dimension of the subgraphs $\{(z, c) : c \le v(z)\}$. This is also known as pseudo-dimension $\mathrm{Pdim}(\mathcal{V})$ [61, Definition 2]. By [61, Theorem 7] there is $C > 0$ such that $V_{\mathcal{V}(P,\ell)^+} = \mathrm{Pdim}(\mathcal{V}(P, \ell)) \le CP\ell \log P$ for $P \ge 2$ and $\ell \ge 1$. □

***Definition* A.9.** For $\beta > 0$ the truncation operator on scalar functions is

$$\mathrm{clip}_\beta u = \begin{cases} u & \text{if } |u| \le \beta, \\ \beta\,\mathrm{sign}(u) & \text{otherwise,} \end{cases}$$

and acts coordinatewise on vector-valued ones. Write

$$\mathcal{W}_N = \mathrm{clip}_\beta \mathcal{V}_N, \quad \mathcal{V}_N = \mathcal{V}^d(P_N, \ell_N).$$

The next proposition shows that truncation keeps us inside the ReLU family at the cost of a constant number of parameters and two layers:

Proposition A.10. $\mathrm{clip}_\beta \mathcal{V}(P, \ell) \supset \mathcal{V}(P, \ell)$ and $\mathrm{clip}_\beta \mathcal{V}(P, \ell) \subseteq \mathcal{V}(P + o(1), \ell + 2)$.

Proof of Proposition A.10. First note that $\mathrm{clip}_\beta u = u - \sigma(u - \beta) + \sigma(-u - \beta)$. Thus, following the definition of the architecture, one can realize output truncation by appending two layers to a network

computing u. Namely, the layer $\ell+1$ has two ReLU units, $\sigma(u-\beta)$ with weight 1 and bias $-\beta$ and $\sigma(-u-\beta)$ with weight -1 and bias $-\beta$. Next, the layer $\ell+2$ is an affine unit combining u (with weight 1 via skip connection), the two ReLU units (with weights -1 and 1) and a bias. Hence the network $v \in \mathrm{clip}_\beta \mathcal{V}(P,\ell)$ has at most $P+7$ parameters and $\ell+2$ layers. □

Note that minimizing the empirical risk $\hat{\mathcal{R}}$ over a single fixed class leads to an irreducible approximation error. Thus, consistency requires to relate the model size to the sample size N. There are several ways to do it. One can use regularization [62, Ch. 12], constrain a norm of the weights, or restricting the class. We adopt the last and simplest of them, the method of sieves [63], following the treatment of [62, Ch. 10]. In contrast to penalty or a weight-norm constraints, it does not change the optimization problem and leaves the estimator untouched.

Following the method of sieves, enlarging the class lowers the approximation error and inflates the estimation error Thus, conducting the consistency analysis we let the class grow, but slowly enough. Specifically, we let the number of parameters P_N, the depth ℓ_N and the truncation level β_N be non-decreasing in the sample size and consider a sequence of classes $\{\mathcal{V}(P_N,\ell_N)\}_{N=1}^{\infty}$ with increasing complexity:

$$\mathcal{V}(P_N,\ell_N) \subseteq \mathcal{V}(P_{N+1},\ell_{N+1}).$$

The more data we have, the more complex the space of models we can entertain.

***Definition* A.11.** The expected and empirical risks of a velocity field v are

$$\mathcal{R}(v) = \mathbb{E}[\|v(Z)-Y\|^2], \quad \hat{\mathcal{R}}(v) = \frac{1}{N}\sum_{i=1}^{N}\|v(Z_i)-Y_i\|^2$$

on N i.i.d. samples $\{(Z_i,Y_i)\}_{i=1}^N$. Note that the empirical risk $\hat{\mathcal{R}}$ is a random function on functions. The minimizer of $\mathcal{R}$ over $L^2(\mu)$ is the function $v^*(z)=\mathbb{E}[Y \mid Z=z]$, and the excess risk of v is

$$\mathcal{E}(v) = \mathcal{R}(v) - \mathcal{R}(v^*).$$

We denote by $v_N \in \arg\min_{v\in\mathcal{V}_N}\hat{\mathcal{R}}(v)$ an empirical risk minimizer over the sieve.

***Remark* A.12.** We stress that the expected risk is a function of v and we average over the joint distribution of (Z,Y), so the random v_N naturally yields random risk $\mathcal{R}(v_N)$.

Using the introduced notation, the excess risk admits the following alternative representation.

Proposition A.13. For any measurable v with $\mathbb{E}\,\|v(Z)\|^2<\infty$ holds

$$\mathcal{E}(v) = \mathcal{R}(v) - \mathcal{R}(v^*) = \mathbb{E}\Big[\|v(Z)-v^*(Z)\|^2\Big].$$

Proof of Proposition A.13. By the definition of the risk,

$$\begin{aligned}\mathcal{E}(v) = \mathcal{R}(v) - \mathcal{R}(v^*) &= \mathbb{E}\Big[\|v-Y+v^*-v^*\|^2\Big] - \mathbb{E}\Big[\|v^*-Y\|^2\Big] \\ &= \mathbb{E}\Big[\|v-v^*\|^2\Big] + \mathbb{E}\Big[\|v^*-Y\|^2\Big] - \mathbb{E}\Big[\|v^*-Y\|^2\Big] = \mathbb{E}\Big[\|v-v^*\|^2\Big].\end{aligned}$$

□

We now establish necessary tools for the convergence rate for the clipped-target estimator. Specifically, we replace the flow-matching target by its coordinate-wise clip at level β. Besides, the same clip is applied to the empirical risk minimizer:

***Definition* A.14.** For $\beta \geq 1$ the clipped-target estimator is

$$\hat{v}_{N,\beta} = \mathrm{clip}_\beta v_{N,\beta}, \quad v_{N,\beta} \in \arg\min_{v\in\mathcal{V}_N}\left\{\hat{\mathcal{R}}_\beta(v) = \frac{1}{N}\sum_{i=1}^{N}\big\|v(Z_i)-\mathrm{clip}_\beta Y_i\big\|^2\right\}, \tag{9}$$

where clip_β acts coordinatewise. Write $v^*_\beta = \mathbb{E}\big[\mathrm{clip}_\beta Y \mid Z\big]$. Once $\beta=\beta_N$ is allowed to grow, we simplify notation by writing $\hat{v}_N \equiv \hat{v}_{N,\beta}$.

On the event $\{\|Y\|_\infty \le \beta\}$ the clipped target coincides with Y, so the estimator $\hat{v}_{N,\beta}$ becomes the one like considered in Algorithm 1 with truncated output. The level β and the domain radius R are allowed to grow with N. In particular, in Theorem A.27 we take both of order $\sqrt{\log N}$.

Proposition A.15. If $\|Y\| \le \beta$ a.s. for some $\beta > 0$, then $\|v^*(z)\| \le \beta$ for μ-almost every z.

Proof of Proposition A.15. By Jensen inequality, $\|v^*(Z)\| = \|\mathbb{E}[Y \mid Z]\| \le \mathbb{E}[\|Y\| \mid Z] \le \beta$ a.s. □

The next result lies at the heart of the convergence analysis.

Theorem A.16. ([62, Theorem 11.5]) Let $N \in \mathbb{N}$ and $1 \le \beta < \infty$. Assume that Y is scalar and $|Y| \le \beta$ a.s. Let $m^* = \mathbb{E}[Y \mid Z]$ and $\hat{m}_N = \mathrm{clip}_\beta \tilde{m}_N$, where $\tilde{m}_N \in \arg\min_{m \in \mathcal{F}_N} \frac{1}{N} \sum_{i=1}^N (m(Z_i) - Y_i)^2$. Let $Z \sim \mu$ be a drawn independent of the empirical samples $\{Z_i\}_{i=1}^N$. Then,

$$\mathbb{E}\Big[(\hat{m}_N(Z) - m^*(Z))^2\Big] \le \Big(c_1 + (c_2 + c_3 \log N) V_{\mathcal{F}_N^+}\Big) N^{-1} + 2 \inf_{f \in \mathcal{F}_N} \mathbb{E}\Big[(f(Z) - m^*(Z))^2\Big],$$

where[2] $c_1 = O(\beta^4), c_2 = O(\beta^4 \log \beta^2)$, and $c_3 = O(\beta^4)$.

Corollary A.17. Let $N \in \mathbb{N}$, $P_N \ge 2$, $\ell_N \ge 1$, and $1 \le \beta \le N$. Let $\hat{v}_{N,\beta}$ be the estimator from Definition A.14 and $v^*_\beta = \mathbb{E}\big[\mathrm{clip}_\beta Y \mid Z\big]$. Then

$$\begin{aligned} \mathbb{E}\Big[\big\|\hat{v}_{N,\beta}(Z) - v^*_\beta(Z)\big\|^2\Big] &\le dC\beta^4 P_N \ell_N (\log P_N)(\log N) N^{-1} \\ &\quad + 2 \inf_{v \in \mathcal{V}_N} \mathbb{E}\Big[\big\|v(Z) - v^*_\beta(Z)\big\|^2\Big], \end{aligned} \tag{10}$$

where Z is independent of empirical samples $\{Z_i\}_{i=1}^N$ and C does not depend on N, P_N, ℓ_N, or β.

Proof. Since $\mathcal{V}_N = \mathcal{V}^d(P_N, \ell_N)$ is a product class and

$$\big\|v(Z) - \mathrm{clip}_\beta Y\big\|^2 = \sum_{k=1}^d \Big(v^{(k)}(Z) - \big(\mathrm{clip}_\beta Y\big)^{(k)}\Big)^2 \quad \text{with } \Big|\big(\mathrm{clip}_\beta Y\big)^{(k)}\Big| \le \beta,$$

the estimator splits into d scalar problems. Apply Theorem A.16 to each coordinate with class $\mathcal{V}(P_N, \ell_N)$. By Theorem A.8, $V_{\mathcal{V}(P_N,\ell_N)^+} \le C_0 P_N \ell_N \log P_N$. For $\beta \in [1, N]$ one has $\log \beta \le \log N$, so the factor $\log \beta$ in c_2 is of order $\log N$, thus there is $C > 0$:

$$c_1 + (c_2 + c_3 \log N) V_{\mathcal{V}(P_N,\ell_N)^+} \le C\beta^4 (\log N) P_N \ell_N \log P_N.$$

It remains to sum up the inequality from Theorem A.16 over all d components noting that the sum of the scalar approximation errors is the approximation error on the product $\mathcal{V}^d(P_N, \ell_N)$. □

***Remark* A.18.** We stress that the left-hand side of (10) is the squared L^2 error about v^*_β, not about $v^* = \mathbb{E}[Y \mid Z]$. They coincide when $\mathrm{clip}_\beta Y = Y$ a.s. and there is no clipping bias, but in general we have the gap $\mathbb{E}\,\big\|v^*_\beta - v^*\big\|^2$, which is bounded in Theorem A.21.

***Definition* A.19.** Let $D = d + 1$ and $R > 1$. Set $Q_R = [-R, R] \times [0, 1] \subset \mathbb{R}^D$.

Corollary A.20. Under Assumption A.1 and Assumption A.2 for any $t \ge 0$ holds

1. $\mathbb{P}(\|Y\|_\infty > t) \le 2d \exp(-t^2/4\sigma^2), \quad \mathbb{P}\Big(\|X_T\|_\infty > t\Big) \le 2d \exp(-t^2/2\sigma^2),$
2. $\mathbb{E}\,\|Y\|^4 \le C_d^2 \sigma^4, \quad \mathbb{E}\,\|X_T\|^4 \le C_d^2 \sigma^4, \quad \mathbb{E}\,\|v^*\|^4 \le C_d^2 \sigma^4$, where $C_d > 1$.

Proof. Note that Y is sub-Gaussian with constant $2\sigma^2$. Indeed, under Assumption A.2 for a unit vector v:

$$\mathbb{E}\, e^{s\langle Y, v\rangle} = \mathbb{E}\, e^{s\langle X_1 - X_0, v\rangle} = \mathbb{E} \exp(s\langle X_1, v\rangle)\, \mathbb{E} \exp(-s\langle X_0, v\rangle) \le \exp\big((2\sigma^2 s^2)/2\big).$$

Similarly, one shows that X_t, where $t \in [0, 1]$, is sub-Gaussian with constant σ^2:

[2] More precisely, $c_1 = 24 \cdot 214\beta^4(1 + \log 42), \quad c_2 = 48 \cdot 214\beta^4 \log(480e\beta^2), \quad c_3 = 48 \cdot 214\beta^4$.

$$\mathbb{E}\, e^{s\langle X_t, v\rangle} = \mathbb{E}\, e^{s\langle tX_1 + (1-t)X_0, v\rangle} = \mathbb{E}\exp(st\langle X_1, v\rangle)\,\mathbb{E}\exp(s(1-t)\langle X_0, v\rangle)$$
$$\leq \exp((\sigma^2 s^2(t^2 + (1-t)^2))/2) \leq \exp((\sigma^2 s^2)/2).$$

Averaging over time T preserves the bound, so X_T is sub-Gaussian. Now we are ready to prove each item. Note that $Y^{(k)}$ and $X_T^{(k)}$ are also sub-Gaussian (take $v = e_k$ be the standard basis vector).

1. By using Markov inequality and sub-Gaussianity of $Y^{(k)}$, for any $1 \leq k \leq d$ and $s > 0$ holds

$$\mathbb{P}\big(|Y^{(k)}| > t\big) \leq \mathbb{P}\big(Y^{(k)} > t\big) + \mathbb{P}\big(-Y^{(k)} > t\big) \leq \mathbb{E}\, e^{sY^{(k)}} e^{-st} + \mathbb{E}\, e^{-sY^{(k)}} e^{-st} \leq 2\exp(\sigma^2 s^2 - st).$$

   Minimizing over the free parameter $s \in \mathbb{R}$, gives $\mathbb{P}\big(|Y^{(k)}| > t\big) \leq 2\exp(-t^2/(4\sigma^2))$, implying

$$\mathbb{P}(\|Y\|_\infty > t) = \mathbb{P}\left(\bigcup_{k=1}^{d}\big\{|Y^{(k)}| > t\big\}\right) \leq \sum_{k=1}^{d}\mathbb{P}\big(|Y^{(k)}| > t\big) \leq 2d\exp(-t^2/(4\sigma^2)).$$

   Analogously, one derives the bound for X_T.

2. By [64, Lemma 1.4] there exists $c > 2$ such that for any $1 \leq k \leq d$ holds

$$\mathbb{E}\left(Y^{(k)}\right)^4 \leq c^2\sigma^4, \quad \mathbb{E}\left(Y^{(k)}\right)^2 \leq c^2\sigma^2. \tag{11}$$

   Therefore, applying Cauchy–Schwarz inequality gives

$$\mathbb{E}\,\|Y\|^4 \leq \mathbb{E}\left(\sum_{k=1}^{d}\left(Y^{(k)}\right)^2\right)^2 \leq \mathbb{E}\, d\sum_{k=1}^{d}\left(Y^{(k)}\right)^4 \leq C_d^2\sigma^4,$$

   where $C_d^2 = c^2d^2 > 1$. Analogously, one derives the bound for X_T. From conditional Jensen's inequality, we deduce

$$\mathbb{E}\,\|v^*(Z)\|^4 = \mathbb{E}\,\|\mathbb{E}[Y \mid Z]\|^4 \leq \mathbb{E}\,\mathbb{E}[\|Y\|^4 \mid Z] \leq C_d^2\sigma^4.$$

□

Theorem A.21. Under Assumption A.1 and Assumption A.2 with $v_\beta^* = \mathbb{E}\big[\mathrm{clip}_\beta Y \mid Z\big]$,

$$\mathbb{E}\left\|v_\beta^* - v^*\right\|^2 \leq C_d^2\sigma^2\exp(-\beta^2/(8\sigma^2)). \tag{12}$$

Proof. Since $v_\beta^* - v^* = \mathbb{E}\big[\mathrm{clip}_\beta Y - Y \mid Z\big]$, the conditional Jensen's inequality gives

$$\mathbb{E}\left\|v_\beta^* - v^*\right\|^2 \leq \mathbb{E}\left\|\mathrm{clip}_\beta Y - Y\right\|^2 = \sum_{k=1}^{d}\mathbb{E}\left|Y^{(k)} - \mathrm{clip}_\beta Y^{(k)}\right|^2 = \sum_{k=1}^{d}\mathbb{E}\left(|Y^{(k)}| - \beta\right)_+^2$$
$$\leq \sum_{k=1}^{d}\mathbb{E}\left|Y^{(k)}\right|^2\mathbb{1}_{\{|Y^{(k)}|>\beta\}} \leq \sum_{k=1}^{d}\sqrt{\mathbb{E}\left|Y^{(k)}\right|^4}\sqrt{\mathbb{P}\big(|Y^{(k)}| > \beta\big)}$$
$$\overset{(11)}{\leq} dc\sigma^2\sqrt{\mathbb{P}(\|Y\|_\infty > \beta)} \leq \sqrt{2d}cd\sigma^2\exp(-\beta^2/(8\sigma^2)),$$

where the last inequality follows from Corollary A.20 and bounding $\sqrt{d}$ with d. □

***Remark* A.22.** We additionally allow L_R to grow with the domain "radius" R to cover the case when the Lipschitz constant depends on x, which is the case for Gaussian endpoints, for example.

To bound the approximation error $\inf_v \mathbb{E}\left\|v - v_\beta^*\right\|^2$ in (10), we invoke [65, Theorem 1] about the the approximation error of ReLU architecture on the unit ball of Sobolev space $W^{1,\infty}([0,1]^D)$:

$$\mathcal{B}_W = \big\{f : [0,1]^D \to \mathbb{R} : f \text{ is } L_f - \text{Lipschitz}, \max\big\{\|f\|_{L^\infty}, L_f\big\} \leq 1\big\}.$$

We denote by $\mathcal{B}_W^d := \underbrace{\mathcal{B}_W \times \ldots \times \mathcal{B}_W}_{d \text{ times}}$ the Cartesian product of d balls.

Theorem A.23. ([65, Theorem 1]) For any $\varepsilon \in (0,1)$ there exists $c = c(D) > 0$ such that, for every $\ell \geq c(\log(1/\varepsilon) + 1)$, $P \geq c\varepsilon^{-D}(\log(1/\varepsilon) + 1)$, and every $f \in \mathcal{B}_W^d$, there is a network $h \in \mathcal{V}^d(P, \ell)$:

$$\sup_{s\in[0,1]^D} \|f(s)-h(s)\|^2 \le d\varepsilon^2.$$

Proof. Fix $\varepsilon \in (0,1)$ and $f = \left(f^{(1)},...,f^{(d)}\right) \in \mathcal{B}_W^d$. Now, apply [65, Theorem 1] to each coordinate: there exist $c = c(D)$ such that, for every $\ell \ge c(\log(1/\varepsilon)+1)$, $P \ge c\varepsilon^{-D}(\log(1/\varepsilon)+1)$, and every k, there is $h^{(k)} \in \mathcal{V}(P,\ell)$ with $\sup_{s\in[0,1]^D}\left|f^{(k)}(s)-h^{(k)}(s)\right| \le \varepsilon$. Note that the network $h = \left(h^{(1)},...,h^{(d)}\right) \in \mathcal{V}^d(P,\ell)$, since P is counted per coordinate, and

$$\sup_{s\in[0,1]^D} \|f(s)-h(s)\|^2 \le \sum_{k=1}^{d} \sup_{s\in[0,1]^D} \left|f^{(k)}(s)-h^{(k)}(s)\right|^2 \le d\varepsilon^2.$$

□

Lemma A.24. Let Assumption A.3 holds and $\sup_{z\in Q_R}\|v^*(z)\|_\infty \le \beta$ for some $\beta > 0$. Denote $\lambda(\varepsilon) = \log(1/\varepsilon)+1$, and let $c = c(D)$ be the constant of Theorem A.23. Then, for any $\varepsilon \in (0,1)$ and all $P \ge c\varepsilon^{-D}\lambda(\varepsilon)+8$ and $\ell \ge c\lambda(\varepsilon)+2$, there exists $w \in \mathcal{V}^d(P,\ell)$ such that $\|w\|_\infty \le \beta$ and

$$\inf_{v\in\mathcal{V}^d(P,\ell)} \mathbb{E}\|v-v^*\|^2 \le dB^2\varepsilon^2 + \mathbb{E}\|w-v^*\|^2 \mathbb{1}_{Q_R^c}, \tag{13}$$

where $B = \max\{\beta, 2RL_R\}$.

Proof. Splitting the integral we have for some $w \in \mathcal{V}^d(P,\ell)$:

$$\inf_{v\in\mathcal{V}^d(P,\ell)} \mathbb{E}\|v-v^*\|^2 \le \mathbb{E}\|w-v^*\|^2 \mathbb{1}_{Q_R} + \mathbb{E}\|w-v^*\|^2 \mathbb{1}_{Q_R^c}. \tag{14}$$

In the rest of the proof we show that w can be selected such that $\|w\|_\infty \le \beta$ and the first term on the RHS can be made arbitrary small by using Theorem A.23. We start with the later.

Perform the change of coordinates $z = \varphi(s)$, where $\varphi : [0,1]^D \to Q_R$ is the invertible affine map $\varphi(s) = (R(2s_1-1),...,R(2s_d-1),s_{d+1})$. By Assumption A.3, for any $s, s' \in [0,1]^D$:

$$\|v^*(\varphi(s))-v^*(\varphi(s'))\| \le L_R\|\varphi(s)-\varphi(s')\| \le L_R\|\varphi\|_{\text{op}}\|s-s'\| = 2RL_R\|s-s'\|.$$

Together with $\sup_{z\in Q_R}\|v^*(z)\|_\infty \le \beta$, $B := \max\{\beta, 2RL_R\}$ it gives $f(s) := B^{-1}v^*(\varphi(s)) \in \mathcal{B}_W^d$.

Set $P_0 = P-8$ and $\ell_0 = \ell-2$ so that $P_0 \ge c\varepsilon^{-D}\lambda(\varepsilon)$ and $\ell_0 \ge c\lambda(\varepsilon)$. By Theorem A.23 there is $h \in \mathcal{V}^d(P_0,\ell_0)$ with $\sup_{s\in[0,1]^D}\|f(s)-h(s)\|^2 \le d\varepsilon^2$. Define a network $\psi(z) := B\,h(\varphi^{-1}(z))$. Since φ^{-1} is affine it can realized by adjusting weights and biases of the first layer of h, so $\psi \in \mathcal{V}^d(P_0,\ell_0)$ and

$$\sup_{z\in Q_R}\|\psi(z)-v^*(z)\|^2 = B^2 \sup_{s\in[0,1]^D}\|h(s)-f(s)\|^2 \le dB^2\varepsilon^2. \tag{15}$$

By Proposition A.10, $w := \text{clip}_\beta\psi \in \mathcal{V}^d(P,\ell)$, and $\|w\|_\infty \le \beta$ by construction.

$$\mathbb{E}\|w-v^*\|^2\mathbb{1}_{Q_R} \le \sup_{z\in Q_R}\|w(z)-v^*(z)\|^2\,\mathbb{P}(Q_R) \le \sup_{z\in Q_R}\|\psi(z)-v^*(z)\|^2 \overset{(15)}{\le} dB^2\varepsilon^2.$$

Combining this bound with (14) completes the proof. □

Theorem A.25. Let Assumption A.1, Assumption A.2 and Assumption A.3 hold, and $\sup_{z\in Q_R}\|v^*(z)\|_\infty \le \beta$, where $1 \le \beta \le N$. Let $\varepsilon \in (0,1)$ and let $P_N = P \ge 2$, $\ell_N = \ell \ge 2$ are chosen as in Lemma A.24. Then,

$$\begin{aligned}\mathbb{E}\,\mathcal{E}(\hat{v}_{N,\beta}) \le\ & \underbrace{2E_{N,\beta}}_{\text{estimation error}} + \underbrace{8dB^2\varepsilon^2}_{\text{approximation error}} \\ & + \underbrace{32dC_d(\beta^2+\sigma^2)\exp(-R^2/4\sigma^2) + 10C_d^2\sigma^2\exp(-\beta^2/(8\sigma^2))}_{\text{truncation error}}.\end{aligned} \tag{16}$$

where $E_{N,\beta} = dC\beta^4 P\ell(\log P)(\log N)N^{-1}$ is from Corollary A.17, C_d is the constant from Corollary A.20, and $B = \max\{\beta, 2RL\}$ is the constant from Lemma A.24.

Proof of Theorem A.25. For brevity, let $\hat{v} = \hat{v}_{N,\beta}$, then by Proposition A.13

$$\mathbb{E}\,\mathcal{E}(\hat{v}) = \mathbb{E}\,\|\hat{v} - v^*\|^2 \le 2\,\mathbb{E}\,\|\hat{v} - v^*_\beta\|^2 + 2\,\mathbb{E}\,\|v^*_\beta - v^*\|^2 \tag{17}$$

By Corollary A.17, the first term obeys

$$\begin{aligned}\mathbb{E}\,\|\hat{v} - v^*_\beta\|^2 &\le E_{N,\beta} + 2\inf_{v\in\mathcal{V}^d(P,\ell)} \mathbb{E}\,\|v - v^*_\beta\|^2 \\ &\le E_{N,\beta} + 4\inf_{v\in\mathcal{V}^d(P,\ell)} \mathbb{E}\,\|v - v^*\|^2 + 4\,\mathbb{E}\,\|v^* - v^*_\beta\|^2,\end{aligned} \tag{18}$$

where $E_{N,\beta} = dC\beta^4 P\ell(\log P)(\log N)N^{-1}$. Substituting (18) into (17) and using Lemma A.24 yields

$$\begin{aligned}\mathbb{E}\,\mathcal{E}(\hat{v}) &\le 2E_{N,\beta} + 8\inf_{v\in\mathcal{V}^d(P,\ell)} \mathbb{E}\,\|v - v^*\|^2 + 10\,\mathbb{E}\,\|v^* - v^*_\beta\|^2 \\ &\overset{(13)}{\le} 2E_{N,\beta} + 8dB^2\varepsilon^2 + 8\,\mathbb{E}\,\|w - v^*\|^2\mathbb{1}_{Q_R^c} + 10\,\mathbb{E}\,\|v^* - v^*_\beta\|^2,\end{aligned} \tag{19}$$

where $\|w\|_\infty \le \beta$. Note that from Corollary A.20 follows

$$\begin{aligned}\mathbb{E}\,\|w - v^*\|^2\mathbb{1}_{Q_R^c} &\le 2\,\mathbb{E}\,\|w\|^2\mathbb{1}_{Q_R^c} + 2\,\mathbb{E}\,\|v^*\|^2\mathbb{1}_{Q_R^c} \\ &\le 2d\beta^2\,\mathbb{P}(Z\notin Q_R) + 2\sqrt{\mathbb{E}\,\|v^*\|^4\,\mathbb{P}(Z\notin Q_R)} \\ &\le 2d\beta^2\sqrt{\mathbb{P}(Z\notin Q_R)} + 2C_d\sigma^2\sqrt{\mathbb{P}(Z\notin Q_R)} \\ &\le 2C_d(\beta^2+\sigma^2)\sqrt{\mathbb{P}(Z\notin Q_R)} = 2C_d(\beta^2+\sigma^2)\sqrt{\mathbb{P}\big(\|X_T\|_\infty > R\big)} \\ &\le 4dC_d(\beta^2+\sigma^2)\exp(-R^2/4\sigma^2),\end{aligned} \tag{20}$$

because $Z = (X_T, T)$ and $T \in [0,1]$ always, so the events $\{Z \notin Q_R\}$ and $\big\{\|X_T\|_\infty > R\big\}$ coincide. Substituting (12), (20) into (19), we get

$$\mathbb{E}\,\mathcal{E}(\hat{v}) \le 2E_{N,\beta} + 8dB^2\varepsilon^2 + 32dC_d(\beta^2+\sigma^2)\exp(-R^2/4\sigma^2) + 10C_d^2\sigma^2\exp(-\beta^2/(8\sigma^2)).$$

□

It remains to choose R, β and ε as functions of N.

Lemma A.26. Under Assumption A.2 and Assumption A.3, let $R_N = \sqrt{8\sigma^2\log N}$. Then, one can choose β_N as follows

$$R_N \le \beta_N \lesssim A(\log N)^{\frac{\alpha+1}{2}}, \quad A = (1+L_0)(1+\sigma)^{\alpha+1}. \tag{21}$$

Moreover, the approximation error in (16) can be similarly bounded as

$$B_N := \max\big\{\beta_N, 2R_N L_{R_N}\big\} \lesssim A(\log N)^{\frac{\alpha+1}{2}} \tag{22}$$

We write $A \lesssim B$ iff $A \le C_{\alpha,d}B$, where $C_{\alpha,d} > 0$ is a constant that can depend only on d and a, and may take a different value at each occurrence.

Proof. For any $z \in Q_R = [-R,R]^d \times [0,1]$ holds $\|z\| \le \sqrt{dR_N^2 + 1} \le \sqrt{d+1}R_N$, so by Assumption A.3 one can write

$$\|v^*(z)\| \le \|v^*(0)\| + L_{R_N}\|z\| \le M + \sqrt{d+1}L_{R_N}R_N,$$

where $\|v^*(0)\| \le M$ for some $M > 0$. The bound (16) from Theorem A.25 holds for $\sup_{z\in Q_R}\|v^*(z)\|_\infty \le \beta$, where $1 \le \beta \le N$, so with additional requirement $\beta_N \ge R_N$ one can define

$$\beta_N := \max\big\{R_N, M + \sqrt{d+1}L_{R_N}R_N\big\}.$$

Before we prove the claimed bound for β_N, let's examine the Lipschitz constant L_R:

$$L_{R_N} \le L_0(1+R_N)^\alpha = L_0\Big(1+\sqrt{8\sigma^2\log N}\Big)^\alpha \lesssim L_0(1+\sigma)^\alpha(\log N)^{\frac{\alpha}{2}}.$$

Now, by taking into account the asymptotics of R_N and L_{R_N}:

$$\begin{aligned}\beta_N &\le R_N + M + \sqrt{d+1}L_{R_N}R_N \lesssim L_{R_N}R_N \\ &\lesssim L_0(1+\sigma)^\alpha(\log N)^{\frac{\alpha}{2}}\sqrt{8\sigma^2\log N} \\ &\lesssim L_0(1+\sigma)^{\alpha+1}(\log N)^{\frac{\alpha+1}{2}} \le A(\log N)^{\frac{\alpha+1}{2}}.\end{aligned}$$

The bound for B_N follows immediately by noting that $B_N \le \beta_N + 2R_N L_{R_N} \lesssim R_N L_{R_N}$ using the above derivation. □

Theorem A.27. Let Assumption A.2, Assumption A.3 fulfil. Set $R_N = \sqrt{8\sigma^2\log N}$ and denote $A = (1+L_0)(1+\sigma)^{\alpha+1}$. Let $\hat{v}_N = \hat{v}_{N,\beta_N}$ be the estimator Definition A.14. Then, for sufficiently large N holds

$$\mathbb{E}\,\mathcal{E}(\hat{v}_N) \le C_{a,d}A^4(\log N)^{2\alpha+6}N^{-\frac{2}{d+3}},$$

where $A = (1+L_0)(1+\sigma)^{\alpha+1}$ and $C_{\alpha,d}$ is a constant that depends on d and α exclusively.

Proof. To get the final rate we examine each term in (16) from Theorem A.25. We start with two terms corresponding to truncation error. Setting $R_N = \sqrt{8\sigma\log N}$ and bounding $\beta_N \ge R_N$, as Lemma A.26 suggests, yields

$$\begin{aligned}\text{Truncation error} &= 32dC_d(\beta_N^2+\sigma^2)\exp(-R_N^2/4\sigma^2) + 10C_d^2\sigma^2\exp(-\beta_N^2/(8\sigma^2)) \\ &\lesssim (\beta_N^2+\sigma^2)\exp(-R_N^2/4\sigma^2) + \sigma^2\exp(-\beta_N^2/(8\sigma^2)) \\ &= (\beta_N^2+\sigma^2)N^{-2} + \sigma^2 N^{-1} \lesssim \sigma^2 N^{-1},\end{aligned} \tag{23}$$

where in the last line we used that $(\beta_N^2+\sigma^2)N^{-2} \le \sigma^2 N^{-1}$ for sufficiently large N as β_N grows logarithmically.

Consider now estimation error. Denote $\lambda(\varepsilon) = \log(1/\varepsilon)+1$. Choose $\ell = c\lambda(\varepsilon)$ and $P = c\varepsilon^{-D}\lambda(\varepsilon)$ that satisfy conditions of the theorem. Clearly, $P\ell\log P\lambda(\varepsilon)) \le c\varepsilon^{-D}\lambda^3(\varepsilon)$. In order to simplify the minimization of the upper bound, we now restrict $\varepsilon \in [1/N, 1)$, giving $\lambda^3(\varepsilon) \le (\log N)^3$. Thus,

$$P_N\ell_N\log P_N \le c\varepsilon^{-D}(\log N)^3.$$

Therefore, the estimation term in (16) simplifies as follows:

$$\begin{aligned}\text{Estimation error} &= dC\beta_N^4 P_N\ell_N(\log P_N)(\log N)N^{-1} \\ &\lesssim \beta_N^4\varepsilon^{-D}(\log N)^4N^{-1} \overset{(21)}{\lesssim} A^4(\log N)^{2\alpha+6}N^{-1}\varepsilon^{-D}.\end{aligned} \tag{24}$$

The approximation term in (16) can be upper bounded using (22) from Lemma A.26:

$$\text{Approximation error} = 8dB_N^2\varepsilon^2 \lesssim A^2(\log N)^{\alpha+1}\varepsilon^2 \tag{25}$$

Here, ε can be chosen up to logarithmic factors by equating sample-size powers in the estimation and approximation terms, i.e., $N^{-1}\varepsilon_N^{-D} = \varepsilon_N^2$, which gives $\varepsilon_N = N^{-1/(D+2)}$. Setting $\varepsilon = \varepsilon_N$ in (24) and (25), and after that combining with (23), the RHS of (16) from Theorem A.25 takes the following form

$$\begin{aligned}\mathbb{E}\,\mathcal{E}(\hat{v}_N) &\lesssim A^4(\log N)^{2\alpha+6}N^{-1}\varepsilon_N^{-D} + A^2(\log N)^{\alpha+1}\varepsilon_N^2 + \sigma^2N^{-1} \\ &= A^4(\log N)^{2\alpha+6}N^{-\frac{2}{D+2}} + A^2(\log N)^{\alpha+1}N^{-\frac{2}{D+2}} + \sigma^2N^{-1} \\ &\le C_{\alpha,d}A^4(\log N)^{2\alpha+6}N^{-\frac{2}{d+3}},\end{aligned}$$

where we used that $D = d+1$. □

B Complete Proofs

Lemma 4.1. For any $t \in [0;1]$, let X_t satisfy $\partial X_t / \partial t = v(X_t, t)$. Let $T \sim \mathrm{U}[0;1]$. Then

$$\mathsf{h}(X_1) - \mathsf{h}(X_0) = \mathbb{E}\,\mathrm{div}\, v(X_T, T)$$

Proof of Lemma 4.1. Let $x_t(x_0)$ be the solution to (4) corresponding to an initial condition x_0. Then

$$\begin{aligned}\mathsf{h}(X_1) - \mathsf{h}(X_0) &= \mathbb{E}\int_0^1 \mathrm{div}\, v(x_t(X_0), t)\,\mathrm{d}t = \int_0^1 \mathbb{E}\,\mathrm{div}\, v(x_t(X_0), t)\,\mathrm{d}t \\ &= \int_0^1 \mathbb{E}_{X_t \sim \mathbb{P}_{X_t}} \mathrm{div}\, v(X_t, t)\,\mathrm{d}t = \mathbb{E}_{T\sim \mathrm{U}[0;1], X_T \sim \mathbb{P}_{X_T}} \mathrm{div}\, v(X_T, T)\end{aligned}$$

□

Theorem 4.2. Let $\mathbb{P}_{X,T}$ be such that $p_t(x) \stackrel{\text{def}}{=} p(x \mid t)$ and $K^2 \stackrel{\text{def}}{=} \mathbb{E}\left\|\nabla \log p_T(X)\right\|_2^2$ exist. Then, for any smooth vector field $\epsilon(x,t)$ satisfying $p_t(x)\|\epsilon(x,t)\| \in o\big(\|x\|^{1-d}\big)$ for any t,

$$|\mathbb{E}\,\mathrm{div}\,\epsilon(X,T)| \leq K\sqrt{\mathbb{E}\left\|\epsilon(X,T)\right\|_2^2}$$

Proof of Theorem 4.2. For any fixed $t \in [0,1]$ one has

$$\begin{aligned}\mathbb{E}[\mathrm{div}\,\varepsilon(X,T) \mid T = t] &= \int_{\mathbb{R}^d} p_t(x)\,\mathrm{div}\,\epsilon(x,t)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \mathrm{div}(p_t(x)\,\epsilon(x,t))\,\mathrm{d}x - \int_{\mathbb{R}^d} \langle \nabla p_t(x), \epsilon(x,t)\rangle\,\mathrm{d}x.\end{aligned}$$

Now, for a ball $B_R = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$, we have by the divergence theorem:

$$\left|\int_{B_R} \mathrm{div}(p_t(x)\,\epsilon(x,t))\,\mathrm{d}x\right| = \left|\int_{\partial B_R} p_t(x)\langle \varepsilon, n\rangle\,\mathrm{d}S\right| \leq \max_{\|x\|_2 = R} p_t(x)\,|\varepsilon(x,t)|\,\mu(\partial B_R).$$

By the assumption, $\forall \delta > 0\, \exists R_0 > 0 : \forall x \in \mathbb{R}^d : \|x\|_2 \geq R_0$ holds $p_t(x)|\varepsilon(x,t)| \leq \delta \|x\|_2^{1-d}$, so the integral is upper bounded by $\delta R^{1-d} C_d R^{d-1}$, using the formula for the surface area of the ball with radius R. Thus, we deduce

$$\mathbb{E}[\mathrm{div}\,\varepsilon(X,T) \mid T = t] = -\int_{\mathbb{R}^d} \langle p_t(x)\nabla \log p_t(x), \varepsilon(x,t)\rangle\,\mathrm{d}x.$$

It suffices to average over $T \sim \mathrm{U}[0;1]$ and apply Cauchy-Schwarz inequality:

$$|\mathbb{E}\,\mathrm{div}\,\varepsilon(X,T)| \leq \sqrt{\mathbb{E}\left\|\nabla \log p_T(X)\right\|_2^2}\sqrt{\mathbb{E}\left\|\varepsilon(X,T)\right\|_2^2} \leq K\sqrt{\mathbb{E}\left\|\varepsilon(X,T)\right\|_2^2}$$

□

Theorem 4.4. (FMDoE convergence rate) Let $\hat{v}_N$ be the trained velocity obtained from the minimization problem (9) using N i.i.d. copies of (X_0, X_1, T). Let $\widehat{\mathrm{DoE}}_{\hat{v}_N} = \frac{1}{M}\sum_{i=1}^M \mathrm{div}\,\hat{v}_N(X_i, T_i)$, where $\{(X_i, T_i)\}_{i=1}^M$ are i.i.d. copies of (X_T, T), independent of the training samples. Set $v^*(x,t) = \mathbb{E}[X_1 - X_0 \mid X_T = x, T = t]$ and $\mathrm{DoE}_{v^*} = \mathbb{E}[\mathrm{div}\, v^*(X_T, T)]$. Let $K^2 = \mathbb{E}\left\|\nabla \log p_T(X_T)\right\|^2 < \infty$, and assume $\mathrm{var}[\mathrm{div}\,\hat{v}_N(X_T, T) \mid \hat{v}_N] \leq s^2$. Then, under the assumptions of Theorem A.27 and Theorem 4.2, for sufficiently large N holds

$$\mathbb{E}\left|\mathrm{DoE}_{v^*} - \widehat{\mathrm{DoE}}_{\hat{v}_N}\right| \leq C_{\alpha,d} A^2 K (\log N)^{\alpha+3} N^{-\frac{1}{d+3}} + sM^{-\frac{1}{2}},$$

where $A = (1 + L_0)(1+\sigma)^{\alpha+1}$, with $L_0 > 0$, $\alpha \geq 0$ and $\sigma > 0$, and $C_{\alpha,d}$ depends only on α and d.

Proof of Theorem 4.4. By the triangle inequality,

$$\mathbb{E}\left|\mathrm{DoE}_{v^*} - \widehat{\mathrm{DoE}}_{\hat{v}_N}\right| \leq \mathbb{E}\left|\mathrm{DoE}_{v^*} - \mathrm{DoE}_{\hat{v}_N}\right| + \mathbb{E}\left|\mathrm{DoE}_{\hat{v}_N} - \widehat{\mathrm{DoE}}_{\hat{v}_N}\right|. \tag{26}$$

The first term can be bounded via Corollary 4.3:

$$\mathbb{E}\Big[\big|\mathrm{DoE}_{v^*} - \mathrm{DoE}_{\hat{v}_N}\big| \;\big|\; \hat{v}_N\Big] \le K\sqrt{\mathbb{E}\Big[\|v^*(Z) - \hat{v}_N(Z)\|_2^2 \;\big|\; \hat{v}_N\Big]}.$$

Taking expectation and applying Jensen's inequality gives

$$\mathbb{E}\big|\mathrm{DoE}_{v^*} - \mathrm{DoE}_{\hat{v}_N}\big| \le K\sqrt{\mathbb{E}\,\mathcal{E}(\hat{v}_N)} \le C_{\alpha,d} K\theta^2 (\log N)^{\alpha+3} N^{-\frac{1}{d+3}},$$

where the last inequality follows from Theorem A.27.

The second term in (26) is simply a MC error, so

$$\mathbb{E}\Big[\big|\mathrm{DoE}_{\hat{v}_N} - \widehat{\mathrm{DoE}}_{\hat{v}_N}\big| \;\big|\; \hat{v}_N\Big] \le \sqrt{\mathrm{var}\Big[\widehat{\mathrm{DoE}}_{\hat{v}_N} \;\big|\; \hat{v}_N\Big]} \le sM^{-\frac{1}{2}}.$$

Taking expectations and substituting into (26) proves the theorem. □

Corollary 4.6. Suppose Theorem 4.2 holds. Then $|\mathsf{h}(X_1) - \mathsf{h}(X_0)| \le K \cdot \mathsf{W}_2(\mathbb{P}_0, \mathbb{P}_1)$.

Proof of Corollary 4.6. By Lemma 4.1 and Theorem 4.2, one can note

$$|\mathsf{h}(X_1) - \mathsf{h}(X_0)| = |\mathbb{E}\,\mathrm{div}\, v(X_T, T)| \le K\sqrt{\mathbb{E}\,\|v(X_T, T)\|_2^2},$$

Taking the infimum over $v \in V(\mathbb{P}_0, \mathbb{P}_1)$ and applying Lemma 4.5 for $p = 2$ completes the proof. □

Proposition 7.1. For a d-dimensional Gaussian vector $X \sim \mathcal{N}(m, \Sigma)$, the O-information is given by

$$\Omega(X) = \frac{1}{2}\left(\sum_{i=1}^{d} \log \sigma_i^2 + (d-2)\log\det\Sigma - \sum_{i=1}^{d} \log\det\Sigma_{\setminus i}\right),$$

where $\Sigma_{\setminus i}$ is Σ without its i-th row and column.

Proof of Proposition 7.1. By the chain rule, $\mathsf{h}(X_i \mid X_{\setminus i}) = \mathsf{h}(X) - \mathsf{h}(X_{\setminus i})$, so the definition of O-information can be written as

$$\Omega(X) = \sum_{i=1}^{d} \mathsf{h}(X_i) + (d-2)\;\mathsf{h}(X) - \sum_{i=1}^{d} \mathsf{h}\big(X_{\setminus i}\big).$$

Substituting the entropies for a Gaussian vector yields

$$\Omega(X) = \frac{1}{2}\sum_{i=1}^{d} \log(2\pi e\,\sigma_i^2) + \frac{d-2}{2}\log\big((2\pi e)^d \det\Sigma\big) - \frac{1}{2}\sum_{i=1}^{d} \log\big((2\pi e)^{d-1}\det\Sigma_{\setminus i}\big),$$

which simplifies to the stated expression. □

Corollary 7.2. For the order-one Gaussian autoregressive process with covariance matrix $(\Sigma)_{ij} = \rho^{|i-j|}$ for any $\rho \in [0, 1)$ and $d \in \mathbb{N}$, the O-information is given by $\Omega(X) = \big(1 - \frac{d}{2}\big)\ln(1-\rho^4)$.

Proof of Corollary 7.2. For the AR(1) covariance matrix Σ with entries $(\Sigma)_{ij} = \rho^{|i-j|}$, the total correlation and dual total correlation are

$$\mathcal{T}(X) = -\frac{d-1}{2}\ln(1-\rho^2) \qquad \mathcal{D}(X) = \frac{d-2}{2}\ln(1+\rho^2) - \frac{1}{2}\ln(1-\rho^2).$$

Using the identity $\Omega(X) = \mathcal{T}(X) - \mathcal{D}(X)$, we obtain

$$
\begin{aligned}
\Omega(X) &= -\frac{d-1}{2}\ln(1-\rho^2) - \frac{d-2}{2}\ln(1+\rho^2) + \frac{1}{2}\ln(1-\rho^2) \\
&= -\frac{d-2}{2}\ln(1-\rho^2) - \frac{d-2}{2}\ln(1+\rho^2) \\
&= -\frac{d-2}{2}\ln((1-\rho^2)(1+\rho^2)) \\
&= \left(1-\frac{d}{2}\right)\ln(1-\rho^4).
\end{aligned}
$$

□

C High MI Regime

It is known that any distribution-free high-confidence lower bound on MI computed on N samples cannot exceed $O(\log N)$ [18]. On the other hand, the variance of MINE or NWJ for example at the optimal critic can grow exponentially with the true mutual information [20]. FMMI is obtained from a velocity regression and provides an approximation of MI, rather than a lower or upper bound, so those limitations do not apply to it directly.

Theorem 4.2 implies that FMMI estimation error is the product of two factors. The regression error, controlled by Theorem A.27, and a Fisher-like factor K. The example of Gaussian endpoints below shows that K can grow exponentially with MI. Once that growth dominates the regression rate, the reliable MI estimation is infeasible.

Proposition C.1. Under Assumption A.1, let $X_0 \sim \mathcal{N}(m_0, \Sigma_0)$ and $X_1 \sim \mathcal{N}(m_1, \Sigma_1)$ with Σ_0 and Σ_1 positive definite. Let $X_t = (1-t)X_0 + tX_1$ has density p_t. Then $X_t \sim \mathcal{N}(m_t, \Sigma_t)$ with $m_t = (1-t)m_0 + tm_1$, $\Sigma_t = (1-t)^2\Sigma_0 + t^2\Sigma_1$, and

$$
K^2 = \mathbb{E}\,\|\nabla \log p_T(X_T)\|^2 = \int_0^1 \mathrm{Tr}(\Sigma_t^{-1})\,\mathrm{d}t.
$$

Proof. Fix $t \in [0,1]$. Since $X_0 \perp\!\!\!\perp X_1$ are both are Gaussian, $X_t \sim \mathcal{N}(m_t, \Sigma_t)$ with the mean and covariance above, and Σ_t is positive definite. Since $\nabla \log p_t(x) = -\Sigma_t^{-1}(x - m_t)$, we have

$$
\begin{aligned}
K_t^2 &= \mathbb{E}\,\|\nabla \log p_t(X_t)\|^2 = \mathbb{E}\,\left\|\Sigma_t^{-1}(X_t - m_t)\right\|^2 \\
&= \mathrm{Tr}\Big(\Sigma_t^{-2}\,\mathbb{E}(X_t - m_t)(X_t - m_t)^{\mathsf{T}}\Big) = \mathrm{Tr}(\Sigma_t^{-1}).
\end{aligned}
$$

Averaging over $T \sim \mathrm{U}[0;1]$ completes the proof. □

Corollary C.2. Consider the following pair of jointly Gaussian d-dimensional random vectors:

$$
(\xi, \eta) \sim \mathcal{N}(m, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \otimes \mathrm{I}_d, \quad \rho \in (-1;1).
$$

Under this setup, the factor K of Proposition C.1 satisfies

$$
K^2 = \Theta\left(d\exp\left(\frac{\mathrm{I}(\xi;\eta)}{d}\right)\right).
$$

Proof. In the flow-matching interpolation path $X_t = (1-t)X_0 + tX_1$, we have $X_0 \sim \mathbb{P}_{\xi,\eta}$ which is Gaussian with covariance matrix Σ and $X_1 \sim \mathbb{P}_\xi \otimes \mathbb{P}_\eta$ that is also Gaussian with uncorrelated unit-variance components. Therefore, $X_t \sim \mathcal{N}(m_t, \Sigma_t)$ with

$$
\Sigma_t = (1-t)^2\Sigma + t^2\mathrm{I}_{2d} = \begin{pmatrix} \tau^2 & (1-t)^2\rho \\ (1-t)^2\rho & \tau^2 \end{pmatrix} \otimes \mathrm{I}_d,
$$

where $\tau^2 = (1-t)^2 + t^2$. Computing the inverse of 2 by 2 matrix and taking the trace yields

$$\operatorname{Tr}(\Sigma_t^{-1}) = \frac{2\tau^2 d}{\tau^4 - (1-t)^4\rho^2} = \frac{d}{\tau^2 - (1-t)^2\rho} + \frac{d}{\tau^2 + (1-t)^2\rho}. \tag{27}$$

We now examine the first fraction on the RHS of the equality above. Simplify the denominator setting $s = 1 - t$ and $a = 2 - \rho$

$$\tau^2 - (1-t)^2\rho = s^2 + (1-s)^2 - s^2\rho = (2-\rho)s^2 - 2s + 1 = a\left(s - \frac{1}{a}\right)^2 + \frac{a-1}{a}$$

and substitute into the first fraction averaging the latter over t:

$$\begin{aligned}
&\frac{d}{a}\int_0^1 \frac{\mathrm{d}s}{\left(s - \frac{1}{a}\right)^2 + \left(\frac{\sqrt{a-1}}{a}\right)^2} s = \frac{d}{a}\frac{a}{\sqrt{a-1}}\arctan\left(\frac{a}{\sqrt{a-1}}s\right)\Bigg|_{-\frac{1}{a}}^{1-\frac{1}{a}} \\
&= \frac{d}{\sqrt{a-1}}\left[\arctan\left(\sqrt{a-1}\right) - \arctan\left(-\frac{1}{\sqrt{a-1}}\right)\right] = \frac{\pi d}{2\sqrt{a-1}}.
\end{aligned}$$

Due to symmetry the integral of the second fraction in (27) has a similar expression, but with $a = 2 + \rho$. Thus, we deduce

$$K^2 = \int_0^1 \operatorname{Tr}(\Sigma_t^{-1})\,\mathrm{d}t = \frac{\pi d}{2\sqrt{1-\rho}} + \frac{\pi d}{2\sqrt{1+\rho}} = \frac{\pi d}{2}\frac{\sqrt{1-\rho} + \sqrt{1+\rho}}{\sqrt{1-\rho^2}},$$

where $\sqrt{2} \le \sqrt{1-\rho} + \sqrt{1+\rho} \le 2$. For this Gaussian pair, $\mathsf{I}(\xi;\eta) = -\frac{d}{2}\log(1-\rho^2)$, so $\exp\left(\frac{\mathsf{I}(\xi;\eta)}{d}\right) = \frac{1}{\sqrt{1-\rho^2}}$ and

$$\frac{\pi d\sqrt{2}}{2}\exp\left(\frac{\mathsf{I}(\xi;\eta)}{d}\right) \le K^2 \le \pi d \exp\left(\frac{\mathsf{I}(\xi;\eta)}{d}\right).$$

□

D Relation to Other Estimators

In this section, we provide broader discussion on the connection of our method to the most relevant existing approaches.

RFMI [43]. Flow-based mutual information estimators are numerous [20, 22, 39, 40]. However, only RFMI [43] and VCE [41] use flow matching as their backbone method.

Both techniques employ a Gaussian prior and two flow models to learn the data distribution: VCE estimates $\mathbb{P}_X$ and $\mathbb{P}_Y$, while RFMI estimates $\mathbb{P}_{X|Y}$ and $\mathbb{P}_X$. Since VCE uses the same copula trick as MIENF [22], it is inherently less flexible then RFMI.

In contrast to these methods, we propose learning a single flow from $\mathbb{P}_X \otimes \mathbb{P}_Y$ (or $\mathbb{P}_X$) to $\mathbb{P}_{X,Y}$ (or $\mathbb{P}_{X|Y}$). This approach avoids generative modeling altogether, focusing only on the coupling transform, which is sufficient for MI estimation [41].

DRE-∞ [49]. Traditional discriminative MI estimators often fail in high-MI scenarios, a problem attributed to the "density chasm." This issue can be partially mitigated by using a telescoping chain of density ratio estimators [46]. DRE-∞ advances this approach by learning a continuous chain of infinitesimal classifiers [49].

Similar to our method, DRE-∞ requires a probability path $\mathbb{P}_t$ between two distributions. However, it also demands that

$$\lim_{\Delta t \to 0} \frac{\mathrm{d}\,\mathbb{P}_{t+\Delta t}}{\mathrm{d}\,\mathbb{P}_t}(x)$$

is non-degenerate for any x and t, a condition that is difficult to ensure in practice. Because our method is not a density ratio estimator, it is free from this limitation.

"Loss Comparison" [66]. The connection between MI and traditional losses (e.g., MSE, MAE, accuracy) is well-established: if X enables an accurate prediction of Y, the mutual information $\mathsf{I}(X;Y)$ is typically high [67]. However, this approach only provides crude bounds tied to a specific predictive model $\mathbb{P}_{Y|X}$ [66].

In contrast, our method does not merely predict Y from X; it fully reconstructs the conditional distribution $\mathbb{P}_{Y|X}$, thereby achieving a far more accurate MI estimate.

E Additional synthetic tests

High-MI settings. We present additional experimental results for low-dimensional ($d \in \{1, 2, 3\}$), but high-MI (up to 5 nats) settings to further highlight superiority of the proposed method in challenging regimes (see Table 4), reporting the mean absolute error (MAE). Here, 5000 samples are used for training, 5000 for evaluation, and the results are averaged across 10 runs and all three values of d.

Table 4: MAE of MI estimates for each estimator (rows) at each true MI value from 0 to 5 nats (columns). Results are shown for three distributions: Smoothed Uniform, Correlated Uniform, and Log-Gamma-Exponential.

	Smoothed Uniform						Correlated Uniform						Log-Gamma-Exponential					
Estimator	MI=0	1	2	3	4	5	MI=0	1	2	3	4	5	MI=0	1	2	3	4	5
cFMMI (ours)	0.02	**0.03**	**0.07**	**0.08**	**0.06**	**0.13**	0.02	**0.01**	**0.01**	**0.04**	**0.17**	**0.47**	0.01	0.03	**0.01**	**0.03**	**0.02**	**0.03**
KSG	**0.00**	0.11	0.18	0.33	0.60	0.98	**0.00**	0.03	0.11	0.29	0.60	0.99	**0.00**	**0.01**	0.02	0.04	0.06	0.11
WKL	**0.00**	0.10	0.13	0.21	0.44	0.79	**0.00**	0.05	0.07	0.19	0.44	0.79	**0.00**	0.02	0.02	**0.03**	0.04	0.08
MINE	**0.00**	0.07	0.08	0.12	0.19	0.38	**0.00**	0.07	0.12	0.25	0.46	0.82	**0.00**	0.02	0.07	0.11	0.24	0.50
NWJ	**0.00**	0.07	0.09	0.15	0.26	0.52	**0.00**	0.07	0.14	0.28	0.53	0.96	**0.00**	0.03	0.10	0.17	0.35	0.67
InfoNCE	**0.00**	0.07	0.07	0.09	0.14	0.26	**0.00**	0.07	0.14	0.22	0.35	0.58	**0.00**	0.03	0.05	0.09	0.18	0.37
MINDE-c	0.06	0.10	0.15	0.30	0.47	0.96	0.06	0.16	0.33	0.50	0.68	0.72	1.17	0.06	0.29	0.31	0.40	0.85
MINDE-j	0.11	0.53	1.90	1.28	1.43	1.64	0.11	0.97	1.89	2.37	2.50	2.53	0.16	0.81	1.63	1.93	1.79	1.44

Long tails settings. Our evaluation protocol already includes a heavy-tailed example (the "half-cube" distribution, obtained from the normal distribution by a tail-elongation mapping $x \to x\sqrt{|x|}$). To further validate our method in complex, long-tailed, outlier-driven scenarios, we provide additional experiments with the Student-t distribution with 1 degree of freedom, for which no finite moments exist – a clear indicator of heavy tails.

As in the high-MI experiments, we estimate MI across 10 runs using 5000 training and 5000 evaluation points, reporting MAE.

Table 5: MAE of MI estimates for each estimator (rows) at each true MI value from 1 to 7 nats (columns). Results are shown for Student-t distribution with 1 degree of freedom.

Estimator	MI=1	2	3	4	5	6	7
cFMMI (ours)	0.11	0.10	0.11	**0.09**	**0.09**	**0.24**	**0.67**
KSG	0.08	0.08	0.18	0.42	0.78	1.21	1.70
WKL	**0.03**	**0.03**	**0.07**	0.26	0.49	0.66	1.19
MINE	0.12	0.18	0.24	0.34	0.60	1.12	1.90
NWJ	0.13	0.19	0.28	0.47	0.85	1.43	2.26
InfoNCE	0.11	0.12	0.16	0.24	0.42	0.75	1.29
MINDE-c	0.13	0.18	0.31	0.51	0.49	0.64	1.03
MINDE-j	0.67	1.26	1.75	1.58	1.42	1.19	1.33

As shown in Table 5, our method surpasses all NN-based alternatives and only slightly falls behind non-parametric methods when the true MI is low.

F Application to fMRI data

Functional magnetic resonance imaging (fMRI) is a non-invasive neuroimaging modality that measures brain activity indirectly via blood-oxygen-level-dependent (BOLD) signals, reflecting changes in local hemodynamics associated with neural activation. fMRI data are inherently high-dimensional, spatiotemporal, and noisy, consisting of time series of voxel-wise signals across the brain. The complex dependence structure across brain regions, subjects, and experimental conditions makes fMRI a natural application domain for information-theoretic analysis.

The fMRI data was obtained from the Human Connectome Project (HCP) 1200 Subjects Release [68]. In this study, we rely on the preprocessed and curated dataset provided by [60]. The sample included 581 healthy, right-handed participants. Participants completed a series of task-based fMRI paradigms [69] designed to assess seven major cognitive domains, sampling the diversity of large-scale human brain networks: visual, motion, somatosensory, and motor systems; category specific representations; working memory/cognitive control systems; language processing (semantic and phonological); social cognition (Theory of Mind); relational processing; and emotion processing. A detailed description of all tasks is provided in Section F. Each participant performed seven tasks with two conditions, resulting in 14 distinct brain states (Table 6) and 8134 fMRI data units in total.

Each fMRI data unit is represented as an activity matrix $\boldsymbol{X}$ of size $r \times l$, where each line $\boldsymbol{x}_i = (x_{i1}, ..., x_{il})$ corresponds to the activity of the i-th brain region over a time interval of length l. Each element $x_{i,j}$ of the activity matrix $\boldsymbol{X}$ corresponds to the activity level of the i-th brain region at time point j.

In [60], participants were selected using three inclusion criteria to ensure high data quality. First, only individuals who successfully completed all seven HCP task paradigms were retained. Second, all cases flagged with quality issues during the HCP quality control process were excluded, retaining only participants with no reported imaging problems. Third, only adults aged 22 to 35 years at the time of scanning were included.

The final sample included 581 healthy participants, all right-handed. Participants were met with a series of tasks [69] designed to assess seven main cognitive domains sampling the diversity of human neural networks, including: 1) visual, motion, somatosensory, and motor systems; 2) category specific representations; 3) working memory/cognitive control systems; 4) language processing (semantic and phonological processing); 5) social cognition (Theory of Mind); 6) relational processing; and 7) emotion processing. Three of these tasks were completed in one session, while the remaining four were performed in another session. Below, we provide a brief description of each task.

Working Memory. This task was designed to study short-term memory and information retention processes. Participants were shown blocks of images depicting places, tools, faces, and body parts. Each run included blocks from four types of stimuli, with half of the blocks performed with a 2-back task requiring memory retention, and the other half with a 0-back task for comparison. The 2-back task required participants to remember the sequence of images and determine whether the current image matched the one shown two steps earlier. The 0-back task only required participants to identify whether the current image matched a given target. Task blocks alternated with fixation blocks, where participants looked at a cross on the screen, allowing researchers to track brain activity changes in the absence of cognitive load.

Gambling Task. This task simulated decision-making processes under uncertainty and risk. Participants played a card game in which they had to guess whether the number on a hidden card was higher or lower than 5. Depending on their answer, they could win (green upward arrow with "1"), lose (red downward arrow with "0.50"), or receive no reward (neutral outcome). The task was divided into blocks dominated by either winning or losing outcomes, enabling researchers to examine how the brain responds to reward anticipation and loss.

Motor Task. This task was aimed at mapping motor areas of the brain. Participants were presented with visual cues instructing them to perform movements such as tapping their fingers on the left or

right hand, squeezing their toes, or moving their tongue. Movement blocks lasted 12 seconds, and each of the two runs contained 13 blocks, including tongue, hand, and foot movements, as well as fixation blocks. These tasks help identify motor cortex areas activated in response to different types of movements.

Language Processing. This task aimed to study language comprehension and arithmetic processing. In the story blocks, participants listened to short narratives, such as adaptations of Aesop's fables, and then answered questions about the story content. In the math blocks, participants solved verbal arithmetic problems. The tasks alternated, and the block durations were adjusted to ensure equal completion times, allowing comparison of brain activity during language processing and numerical operations.

Social Cognition (Theory of Mind). Participants were shown short video clips where geometric shapes (squares, circles, triangles) either interacted with each other or moved randomly. After each video, participants evaluated whether the shapes had intentions and were interacting or whether their movements were random. This task allowed researchers to study brain areas associated with understanding social interactions and recognizing intentions.

Relational Processing. This task involved analyzing relationships between objects on the screen. In one condition, participants determined which characteristic (shape or texture) differentiated object pairs and assessed whether this distinction applied to another pair. In the control condition, they simply identified whether the lower object matched one of the upper objects based on a given criterion. This task helped researchers study the brain's ability to perform comparative analysis and identify relationships between objects.

Emotion Processing. Participants were shown faces displaying fear or anger, as well as neutral figures, and had to choose one of two faces or figures that matched the presented stimulus. Task blocks alternated with fixation blocks, allowing researchers to assess brain activity during emotional stimulus processing compared to neutral conditions.

In the work [60], authors represented each activity matrix $\boldsymbol{X}$ by the averaged activity values across brain regions. For each brain region $i = (1, ..., r)$ they calculated the mean value $\overline{\boldsymbol{x}}_i$ of the time series $\boldsymbol{x}_i$. Thus, for each activity matrix $\boldsymbol{X}$, we obtain a vector of mean activity values across all brain regions, which will be used in further analysis: $\boldsymbol{x}_{\text{mean}} = (\overline{\boldsymbol{x}}_1, ..., \overline{\boldsymbol{x}}_r)$.

The authors placed particular emphasis on the high classification accuracy achieved for the following brain states: *left hand or foot, right hand or foot, math, story, mental interaction, and random motion*. For these states, a relatively small number of identified brain regions is sufficient to achieve high classification performance based on their mean activity values. Moreover, the sets of regions are state-specific, forming compact and largely non-overlapping signatures for each cognitive condition. The HCP indices of the brain regions that are most significant for classification, together with their anatomical labels, are reported in Table 7.

As also demonstrated by the authors, for the remaining cognitive states the number of regions contributing to classification is substantially larger, approaching coverage of most regions in the brain atlas. This indicates more distributed and less localized neural representations, consistent with more globally integrated cognitive processes.

Table 6: Two brain states for each cognitive task, between which classification is made

Cognitive Task	State name 1	State name 2
Working Memory	0-back	2-back
Gambling	Win	Loss
Motor Task	Left hand or foot	Right hand or foot
Language Processing	Story	Math
Social Cognition	Random motion	Mental interaction
Relational Processing	Relation	Similarity
Emotion Processing	Neutral	Fear

Table 7: The HCP indices and anatomical labels of significant brain regions

Index	Label
Language Processing: Story	
80	IFJa_L, IFJa, Fr, Inferior Frontal
117	AIP_L, AIP, Par, Superior Parietal
150	PGi_L, PGi, Par, Inferior Parietal
Language Processing: Math	
44	6ma_L, 6ma, Fr, Paracentral Lobular and Mid Cingulate
128	STSda_L, STSda, Temp, Auditory Association
135	TF_L, TF, Temp, Medial Temporal
137	PHT_L, PHT, Temp, Lateral Temporal
146	IP0_L, IP1, Par, Inferior Parietal
186	V4_R, V4, Occ, Early Visual
313	TE1p_R, TE1p, Temp, Lateral Temporal
337	FST_R, FST, Occ, MT+Complex and Neighboring Visual Areas
367	LP_L, Thalamus: Lateral Posterior
Social Cognition: Mental Interaction	
4	V2_L, V2, Occ, Dorsal Stream Visual
338	V3CD_R, V3CD, Occ, MT+Complex and Neighboring Visual Areas
Social Cognition: Random Motion	
13	V3A_L, V3A, Occ, Dorsal Stream Visual
128	STSda_R, STSda, Temp, Auditory Association
146	IP0_L, IP0, Par, Inferior Parietal
185	V3_R, V3, Occ, Early Visual
203	MT_R, MT, Occ, MT+Complex and Neighboring Visual Areas
309	STSdp_R, STSdp, Temp, Auditory Association
Motor Task: Right Hand or Foot	
39	5L_L, 5L, Par, Paracentral Lobular and Mid Cingulate
55	6mp_L, 6mp, Fr, Paracentral Lobular and Mid Cingulate
231	1_R, 1, Par, Somatosensory and Motor
Motor Task: Left Hand or Foot	
8	4_L, 4, Fr, Somatosensory and Motor
36	5m_L, 5m, Par, Paracentral Lobular and Mid Cingulate
39	5L_L, 5L, Par, Paracentral Lobular and Mid Cingulate
54	6d_L, 6d, Fr, Premotor
55	6mp_L, 6mp, Fr, Paracentral Lobular and Mid Cingulate
105	PFcm_L, PFcm, Par, Early Auditory
148	PF_L, PF, Par, Inferior Parietal
220	24dd_R, 24dd, Fr, Paracentral Lobular and Mid Cingulate
231	1_R, 1, Par, Somatosensory and Motor
232	2_R, 2, Par, Somatosensory and Motor
281	OP1_R, OP1, Par, Posterior Opercular

Motor Task: Left Hand or Foot	
285	PFcm_R, PFcm, Par, Early Auditory
296	PFt_R, PFt, Par, Inferior Parietal

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We provide theoretical and practical proofs for our contribution.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Please, refer to the dedicated paragraph.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

   Justification: See Section B.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: We provide experimental details in the main text and Appendix.

   Guidelines:

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

the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open Access to Data and Code**

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

   Answer: [Yes]

   Justification: We provide source code in the supplementary material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: We provide hyperparameters in the main text and Appendix

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

Justification: All plots report confidence intervals. Not all confidence intervals are visible due to small magnitude of errors.

Guidelines:

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

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: We specify our resources in the appendix.

   Guidelines:

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

9. **Code of Ethics**

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

   Answer: [Yes]

   Justification: This research uses only synthetic data and does not involve human subjects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

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

    Answer: [NA]

    Justification: This paper presents a foundational method for MI estimation, there is no clear path to negative societal impacts.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: No models or datasets with misuse risk are released.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for Existing Assets**

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

    Answer: [Yes]

    Justification: We provide correct attribution to every material we used.

    Guidelines:

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

13. **New Assets**

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

    Answer: [No]

    Justification: All experiments are based on synthetic data and existing libraries.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

16. **Declaration of LLM Usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLM was used solely for polishing the text (writing, editing, and formatting).

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.