Tight Mutual Information Estimation With Contrastive Fenchel-Legendre Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Successful applications of InfoNCE (Information Noise-Contrastive Estimation) and its variants have popularized the use of contrastive variational mutual information (MI) estimators in machine learning . While featuring superior stability, these estimators crucially depend on costly large-batch training, and they sacrifice bound tightness for variance reduction. To overcome these limitations, we revisit the mathematics of popular variational MI bounds from the lens of unnormalized statistical modeling and convex optimization. Our investigation yields a new unified theoretical framework encompassing popular variational MI bounds, and leads to a new simple and powerful contrastive MI estimator we name FLO. Theoretically, we show that the FLO estimator is tight, and it converges under stochastic gradient descent. Empirically, the FLO estimator overcomes the limitations of its predecessors and learns more efficiently. The utility of FLO is verified using extensive benchmarks, and we further inspire the community with novel applications in meta-learning. Our presentation underscores the foundational importance of variational MI estimation in data-efficient learning.

1 Introduction

Assessing the dependence between pairs of variables is integral to many scientific and engineering endeavors [58, 59]. *Mutual information* (MI) is a popular metric to quantify generic associations [42], and its empirical estimators have been widely used in applications such as independent component analysis [3], fair learning [27], neuroscience [50], Bayesian optimization [37], among others. Notably, the recent advances in deep *self-supervised learning* (SSL) heavily rely on nonparametric MI optimization [65, 49, 29, 13, 25]. In this study we investigate the likelihood-free variational approximation of MI using only paired samples, and improve the data-efficiency of current machine learning practices.

MI estimation has been extensively studied [6, 43, 42, 51, 53, 66]. While most classical estimators work reasonably well for low-dimensional cases, they scale poorly to big datasets: naïve density-based estimator(s) and k-nearest neighbor estimators [39, 52, 20] struggle with high-dimensional inputs, while kernel estimators are slow, memory demanding and sensitive to hyperparameters [24, 23]. Moreover, these estimators are usually either non-differentiable or need to hold all data in memory. Consequently, they are not well suited for emerging applications where the data representation needs to be differentiably optimized based on small-batch estimation of MI [33]. Alternatively, one can approach MI estimation through an estimated likelihood ratio [62, 33], but the associated numerical instability has raised concerns [2].

To scale MI estimation to the growing size and complexity of modern datasets, and to accommodate the need for representation optimization [8], variational objectives have been widely utilized recently [49]. Instead of directly estimating data likelihoods, density ratios, or the corresponding gradients [69], variational approaches appeal to mathematical inequalities to construct tractable lower or

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

upper bounds of the mutual information [54], facilitated by the use of auxiliary critic functions[1]. This practice turns MI estimation into an optimization problem. Prominent examples include the *Barber-Agakov* (BA) estimator [4], the *Donsker-Varadhan* (DV) estimator [15], and the *Nguyen-Wainwright-Jordan* (NWJ) estimator [47]. These variational estimators are closely connected to the variational objectives for likelihood inference [1].

Despite reported successes, these variational estimator have a major limitation:: their estimation variance grows exponentially to the ground-truth MI [44]. This is especially harmful to applications involving deep neural nets, as it largely destabilizes training [61]. An effective fix is to leverage multi-sample contrastive estimators, pioneered by the work of `InfoNCE` [49]. However, the massive reduction in the variance comes at a price: the performance of the `InfoNCE` estimator is upper bounded by $\log K$, where K is the number of *negative* samples used [54]. For a large MI, K needs to be sufficiently large to allow for an adequate estimate, consequently placing a significant burden on computation and memory. While variants of `InfoNCE` have been motivated to achieve more controllable bias and variance tradeoffs [54, 61], little research has been conducted on the cost-benefit aspect of contrastive learning.

A critical insight enabled by `InfoNCE` is that mutual information closely connects to contrastive learning [28, 49]. Paralleled by the empirical successes of instance discrimination-based self-supervision [45, 71, 13, 29] and multi-view supervision [64, 55], `InfoNCE` offers an *InfoMax* explanation to why the ability to discriminate naturally paired *positive* instances from the randomly paired *negative* instances leads to universal performance gains in these applications [40, 60, 54]. Despite these encouraging developments, the big picture of MI optimization and contrastive learning is not yet complete: (i) There is an ongoing debate about to what extent MI optimization helps to learn [67]; (ii) how does the contrastive view reconcile with those non-contrastive MI estimators; crucial for practical applications, (iii) are the empirical tradeoffs made by estimators such as `InfoNCE` absolutely necessary? And theoretically, (iv) formal guarantees on the statistical convergence of popular variational non-parametric MI estimation are missing currently.

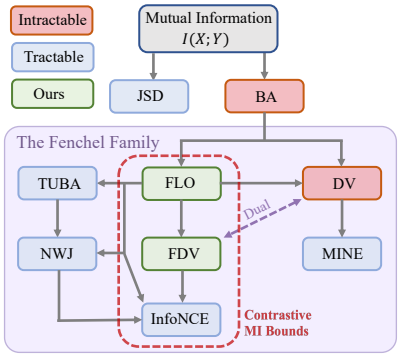

Figure 1: Schematic of variational lower bounds of mutual information. `FLO` provides a novel unified framework to analyze contrastive MI bounds.

In this work we seek to bridge the above gaps by approaching the MI estimation from the novel perspective of energy modeling. While this subject has recently been studied extensively using information-theoretic and variational inequalities, we embrace a new view from the lens of unnormalized statistical modeling. Our main contributions are four-fold:

- Unifying popular variational MI bounds under unnormalized statistical modeling;
- Deriving a simple but powerful novel contrastive variational bound called `FLO`;
- Providing theoretical justification of the `FLO` bound (tightness and convergence);
- Demonstrating strong empirical evidence of the superiority of `FLO` over its predecessors.
- Highlighting the importance of MI in data-efficient learning with novel applications

We contribute in-depth discussion to bridge the gaps between contrastive learning and MI estimation, along with principled practical guidelines informed by theoretical insights.

2 Fenchel-Legendre Optimization for Mutual Information Estimation

2.1 Preliminaries

This section briefly reviews the mathematical background needed for our subsequent developments.

Unnormalized statistical modeling defines a rich class of models of general interest. Specifically, we are interested in problems for which the system is characterized by an energy function $\tilde{p}_\theta(x) : \mathcal{X} \to \mathbb{R} : \exp(-\psi_\theta(x))$, where θ is the system parameters and $\psi_\theta(x)$ is known as the *potential function*. The goal is to find a solution that is defined by a normalized version of $\tilde{p}_\theta(x)$, *i.e.*,

[1] When estimates are sharp, these critic functions usually recover some transformation of the likelihood ratio.

$\min_\theta \left\{ \mathcal{L}\left(\frac{\tilde{p}_\theta}{\int \tilde{p}_\theta(x')\,\mathrm{d}\mu(x')} \right) \right\}$, where $\mathcal{L}(\cdot)$ is the loss function, μ is the base measure on $\mathcal{X}$ and $Z(\theta) \triangleq \int \tilde{p}_\theta(x')\,\mathrm{d}\mu(x')$ is called the *partition function* for $\tilde{p}_\theta(x)$. Problems in the above form arise naturally in statistical physics [57], Bayesian analysis [9], and maximal likelihood estimation [63]. A major difficulty with unnormalized statistical modeling is that the partition function $Z(\theta)$ is generally intractable for complex energy functions [2], and in many applications $Z(\theta)$ is further composed by $\log Z(\theta)$, whose concavity implies any finite sample estimate Monte-Carlo of $Z(\theta)$ will render the loss function biased. Bypassing the difficulties caused by the intractable partition function is central to unnormalized statistical modeling [22, 46, 31, 34, 28].

Mutual information and unnormalized statistical models. As a generic score assessing the dependency between two random variables (X, Y), *mutual information* is formally defined as the *Kullback-Leibler divergence* (KL) between the joint distribution $p(x, y)$ and product of the respective marginals $p(x)p(y)$ [59], *i.e.*, $I(X;Y) \triangleq \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right]$. The integrand $\log \frac{p(x,y)}{p(x)p(y)}$ is often known as the *point-wise mutual information* (PMI) in the literature. Mutual information has a few appealing properties: (i) it is invariant wrt invertible transformations of x and y, and (ii) it has the intuitive interpretation of reduced uncertainty of one variable given another variable[3].

To connect MI to unnormalized statistical modeling, we consider the classical *Barber-Agakov* (BA) estimator of MI [5]. To lower bound MI, BA introduces a variational approximation $q(y|x)$ for the posterior $p(y|x)$, and by rearranging the terms we obtain an inequality

$$\begin{aligned} I(X;Y) &= \mathbb{E}_{p(x,y)}\left[\log \frac{p(y|x)}{p(y)}\right] = \mathbb{E}_{p(x,y)}\left[\log \frac{q(y|x)}{p(y)}\right] + \mathbb{E}_{p(x)}[\mathrm{KL}(p(y|x) \parallel q(y|x))] \\ &\geq \mathbb{E}_{p(x,y)}\left[\log \frac{q(y|x)}{p(y)}\right] \triangleq I_{\mathtt{BA}}(X;Y|q). \end{aligned} \tag{1}$$

Here we have used notation $I_{\mathtt{BA}}(X;Y|q)$ to highlight the dependence on $q(y|x)$, and when $q(y|x) = p(y|x)$ this bound is sharp. Unfortunately, this naïve BA bound is not useful for sample-based MI estimation, as we do not know the ground-truth $p(y)$. But we can bypass this difficulty by setting $q_\theta(y|x) = \frac{p(y)}{Z_\theta(x)} e^{g_\theta(x,y)}$, where we call $e^{g_\theta(x,y)}$ the *tilting function* and recognize $Z_\theta(x) = \mathbb{E}_{p(y)}[e^{g_\theta(x,y)}]$ as the associated partition function. Substituting this $q_\theta(x|y)$ into (1) gives the following *unnormalized BA* bound (UBA) that pertains to unnormalized statistical modeling [54]

$$I_{\mathtt{UBA}}(X;Y|g_\theta) \triangleq \mathbb{E}_{p(x,y)}[g_\theta(x,y) - \log Z_\theta(x)] = \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}\left[\log \frac{e^{g_\theta(x,y)}}{Z_\theta(x)}\right]\right]. \tag{2}$$

While this UBA bound remains intractable, now with $Z_\theta(x)$ instead of $p(y)$ we can apply different techniques for empirical estimates of $Z_\theta(x)$ to render a tractable surrogate target. This has led to various popular MI bounds listed in Table 1, please refer to the Appendix for details.

InfoNCE and noise contrastive estimation. InfoNCE is a multi-sample mutual information estimator proposed in [49], built on the idea of *noise contrastive estimation* (NCE) [28]. NCE learns statistical properties of a target distribution by comparing the *positive* samples from the target distribution to the "*negative*" samples from a carefully crafted noise distribution, and this technique is also known as *negative sampling* in some contexts [45, 26]. The InfoNCE estimator implements this contrastive estimation idea via using the naïve empirical estimate of $Z_\theta(x)$ in UBA[4], *i.e.*

$$I^K_{\mathtt{InfoNCE}}(X;Y|g_\theta) \triangleq \mathbb{E}_{p^K(x,y)}\left[\log \frac{e^{g_\theta(x_1,y_1)}}{\frac{1}{K}\sum_j e^{g_\theta(x_1,y_j)}}\right], I^K_{\mathtt{InfoNCE}}(X;Y) \triangleq \max_{g_\theta \in \mathcal{F}} \{I^K_{\mathtt{InfoNCE}}(X;Y|g_\theta)\}, \tag{3}$$

where g_θ is known as the *critic* in the nomenclature of contrastive learning, and we have used $p^K(x, y)$ to denote K independent draws from the joint density $p(x, y)$, and $\{(x_k, y_k)\}_{k=1}^K$ for each pair of samples. Here the positive and negative samples are respectively drawn from the joint $p(x, y)$ and product of marginals $p(x)p(y)$. Intuitively, InfoNCE tries to accurately classify the positive samples when they are mixed with negative samples, and the Proposition below formally characterize InfoNCE's statistical properties as a MI estimator.

[2] In the sense that they do not render closed-from expressions.

[3] Formally, $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$, where $H(X)$ (resp. $H(X|Y)$) denotes the Shannon entropy (resp. conditional Shannon entropy) of a random variable.

[4] This estimator is technically equivalent to the original definition due to the symmetry of K samples.

Fenchel-Legendre duality. Our key idea is to exploit the convex duality for MI estimation. Let $f(t)$ be a proper convex, lower-semicontinuous function; then its convex conjugate function is defined as $f^*(v) \triangleq \sup_{t\in\mathcal{D}(f)}\{tv - f(t)\}$, where $\mathcal{D}(f)$ is the domain of function f [32]. We call $f^*(v)$ the *Fenchel conjugate* of $f(t)$, which is also known as the *Legendre transform* in physics. The Fenchel conjugate pair (f, f^*) are dual to each other, in the sense that $f^{**} = f$, *i.e.*, $f(t) = \sup_{v\in\mathcal{D}(f^*)}\{vt - f^*(v)\}$. For $f(t) = -\log(t)$ and its Fenchel conjugate $f^*(t) = -1 - \log(-v)$, we have inequality

$$-\log(t) \geq -u - e^{-u}t + 1, \quad \text{for } u \in \mathbb{R} \tag{4}$$

with the equality holds when $u = \log(t)$.

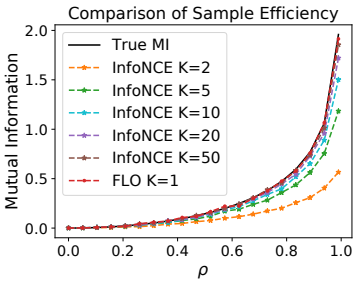

Figure 2: K-sample `InfoNCE` and single-sample `FLO`. Note `FLO` is tight regardless of sample-size.

Proposition 2.1 ([54]). `InfoNCE` is an asymptotically tight mutual information lower bound, *i.e.* $I^K_{\texttt{InfoNCE}}(X;Y|g_\theta) \leq I(X;Y)$, $\lim_{K\to\infty} I^K_{\texttt{InfoNCE}}(X;Y) \to I(X;Y)$.

2.2 Fenchel-Legendre Optimization for tight mutual information estimation

With the above mathematical tools, we are ready to present the main result of this paper: a tight, data-efficient variational MI lower bound that can be efficiently implemented.

Lower bounding MI with Fenchel-Legendre Optimization. Our key insight is that MI estimation is essentially an unnormalized statistical model, which can be efficiently handled by the Fenchel-Legendre transform technique. Take the integrand from `UBA` in (2) and we can rewrite it as

$$\log \frac{\exp(g_\theta(x,y))}{Z_\theta(x)} = -\log\left\{\mathbb{E}_{p(y')}[\exp(g(x,y') - g(x,y))]\right\}, \tag{5}$$

where $p(y')$ is the same probability density as $p(y)$ (*i.e.*, Y' is an independent copy of Y). Now let us use the Fenchel inequality of $-\log(t)$ from (4), plugging it into the above equation and then we have

$$\log \frac{\exp(g_\theta(x,y))}{Z_\theta(x)} \geq \left\{-u - e^{-u}\mathbb{E}_{p(y')}[\exp(g(x,y') - g(x,y))]\right\} + 1. \tag{6}$$

for all $u \in \mathbb{R}$. This implies for any function $u_\phi(x,y) : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, the following inequality holds

$$\log \frac{\exp(g_\theta(x,y))}{Z_\theta(x)} \geq -\{u_\phi(x,y) + e^{-u_\phi(x,y)}\mathbb{E}_{p(y')}[\exp(g(x,y') - g(x,y))]\} + 1. \tag{7}$$

By putting (7) back to (2), we obtain our new Fenchel-Legendre Optimization (`FLO`) MI lower bound

$$I_{\texttt{FLO}}(X;Y|g_\theta,u_\phi) \triangleq \mathbb{E}_{p(x,y)}\left[-\{u_\phi(x,y) + e^{-u_\phi(x,y)}\mathbb{E}_{p(y')}[e^{g_\theta(x,y')-g_\theta(x,y)}]\}\right] + 1, \tag{8}$$

and concludes the proof for the following Proposition.

Proposition 2.2. $I_{\texttt{FLO}}(X;Y|g_\theta,u_\phi) \leq I_{\texttt{UBA}}(X;Y|g_\theta) \leq I(X;Y)$.

In practice, `FLO` can be estimated with the following naïve empirical K-sample estimator

$$\hat{I}^K_{\texttt{FLO}}(X;Y|g_\theta,u_\phi) \triangleq -\left\{u_\phi(x_i,y_i) + e^{-u_\phi(x_i,y_i)}\frac{1}{K-1}\sum_{j\neq i} e^{g_\theta(x_i,y_j)-g_\theta(x_i,y_i)}\right\} + 1. \tag{9}$$

Since the summation in $\hat{I}^K_{\texttt{FLO}}$ is not encapsulated by a convex log transformation, $I^K_{\texttt{FLO}} \triangleq \mathbb{E}_{p^K}[\hat{I}^K_{\texttt{FLO}}]$ is an unbiased estimator for $I_{\texttt{FLO}}(X;Y|g_\theta,u_\phi)$ independent of the batch size K (see Figure 2).

Why is the `FLO` bound more appealing? At first sight, it may appear counter-intuitive that $I_{\texttt{FLO}}$ is a better MI bound compared to prior arts such as `NWJ` or `InfoNCE`: it seems to be more complicated as an extra variational function $u_\phi(x,y)$ has been introduced. To answer this question, we next explain the statistical meaning of the newly introduced $u_\phi(x,y)$, and establish some important statistical properties of `FLO` that makes it more favorable: that $I_{\texttt{FLO}}$ is tight, meaning the ground-truth MI can be recovered for some specific choice of $g_\theta(x,y)$ and $u_\phi(x,y)$; and that $I^K_{\texttt{FLO}}$ for any batch size K is effectively optimizing `InfoNCE` with an infinite batch size. And in Sec 2.4, we further justify `FLO`'s advantages from optimization perspectives.

Given the close connection between FLO and UBA, we first recall UBA's optimal critic that gives the tight MI estimate is $g_*(x,y) = \log p(x|y) + c(x)$, where this $c(x)$ can be any function of x [41]. This $g_*(x,y)$ is not directly meaningful in a statistical sense, however, by integrating out y', we have

$$\mathbb{E}_{p(y')}\left[e^{g_*(x,y')-g_*(x,y)}\right] = \mathbb{E}_{p(y')}\left[\frac{p(x|y')}{p(x|y)}\right] = \frac{p(x)}{p(x|y)} = \frac{p(x)p(y)}{p(x,y)}, \tag{10}$$

which is the likelihood ratio between the marginals and joint. On the other hand, based on the Fenchel-Legendre inequality (4), we know for fixed $g(x,y)$ our FLO bound in (8) can be maximized with $u_g(x,y) = \log \mathbb{E}_{p(y')}\left[e^{g(x,y')-g(x,y)}\right]$. Putting these all together we have $u_{g_*}(x,y) = -\log \frac{p(x,y)}{p(x)p(y)}$. This shows the $u_\phi(x,y)$ introduced in FLO actually tries to recover the negative PMI. Comparing to the competing MI bounds that only optimizes for g_θ, eliminating the drift term $c(x)$ reveals FLO enjoys the appealing *self-normalizing* property [28] that helps stabilize training. Plugging (g_*, u_{g_*}) into (8), we readily see $I_{\texttt{FLO}}(X;Y|u_{g_*}, g_*) = I(X;Y)$, proving FLO is a tight MI bound.

Proposition 2.3. The FLO estimator is tight.

Corollary 2.4. Let (g_*, u_{g_*}) be the maximizers for (8), then $I(X;Y) = \mathbb{E}_{p(x,y)}[-u_{g_*}(x,y)]$.

Finally, we give a simple asymptotic argument showing FLO essentially optimizes InfoNCE with an infinite batch size. In virtue of the law of large numbers, we have the denominator in InfoNCE converging to $\lim_{K\to\infty} \frac{1}{K}\sum_{j=1}^{K} e^{g_\theta(x_i,y_j)} \to \mathbb{E}_{p(y')}[e^{g_\theta(x_i,y')}] = Z_\theta(x_i)$, and consequently it recovers the UBA bound. Since FLO is derived from UBA, we can view FLO as using the optimization of $u_\phi(x,y)$ to amortize the difficulty of evaluating infinite number of $e^{g_\theta(x_i,y_j)}$ with InfoNCE.

Efficient implementations of FLO. A lingering concern is that the newly introduced $u_\phi(x,y)$ can incur extra computation overhead. This is not true, as we can maximally encourage parameter sharing by jointly model $u_\phi(x,y)$ and $g_\theta(x,y)$ with a single neural network $f_\Psi(x,y) : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^2$ with two output heads, *i.e.*, $[u_i, g_i] = f_\Psi(x_i, y_i)$. Consequently, while FLO adopts a dual critics design, it does not actually invoke extra modeling cost compared to its single-critic counterparts (*e.g.*, InfoNCE). Experiments show this shared parameterization in fact promotes synergies and speeds up learning (see our ablation studies in Appendix).

To further enhance the computation efficiency, we consider a massively parallelized *bi-linear* critic design that uses all in-batch samples as negatives. Let $g_\theta(x,y) = \tau\cdot\langle h_\theta(x), \tilde{h}(y)\rangle$, where $h : \mathcal{X} \to \mathbb{S}^p$ and $\tilde{h} : \mathcal{Y} \to \mathbb{S}^p$ are respectively encoders that map data to unit sphere $\mathbb{S}^p$ embedded in $\mathbb{R}^{p+1}$, $\langle a, b\rangle = a^T b$ is the inner product operation, and $\tau > 0$ is the inverse temperature parameter. Thus the evaluation of the *Gram* matrix $G = \tau \cdot h(\mathbb{X})^T \tilde{h}(\mathbb{Y})$, where $[\mathbb{X}, \mathbb{Y}] \in \mathbb{R}^{K\times(d_x+d_y)}$ is a mini-batch of K-paired samples and $g_\theta(x_i, y_j) = G_{ij}$, can be parallelized via matrix multiplication. In this setup, the diagonal terms of G are the positive scores while the off-diagonal terms negative scores. A similar strategy has been widely employed in the contrastive representation learning literature (*e.g.*, [13])[5]. We can simply model the PMI critic as $u(x,y) = \texttt{MLP}(h(x), \tilde{h}(y))$, whose computation cost is almost neglectable in practice, where feature encoders $h, \tilde{h}$ dominate computing.

2.3 Connections to the existing MI bounds

Due to space limitations, we elaborate the connections to the existing MI bounds here, and have relegated an extended related work discussion in a broader context to the Appendix.

From $\log$-partition approximation to MI bounds. To embrace a more holistic understanding, we list popular variational MI bounds together with our FLO in Table 1, and visualize their connections in Figure 1. With the exception of JSD, these bounds can be viewed from the perspective of unnormalized statistical modeling, as they differ in how the $\log$ partition function $\log Z(x)$ is estimated. We broadly categorize these estimators into two families: the $\log$-family (DV, MINE, InfoNCE) and the exponential-family (NWJ, TUBA, FLO). In the $\log$-family, DV and InfoNCE are multi-sample estimators that leverage direct Monte-Carlo estimates $\hat{Z}$ for $\log Z(x)$, and these two differ in whether to include the positive sample in the denominator or not. To avoid the excessive in-batch computation of the normalizer

[5] As an important note to the community, most open source implementations for the bilinear contrastive loss have mechanically implemented $\frac{1}{T}\langle\cdot,\cdot\rangle$ following the practice from pioneering contrastive learning studies, which is numerically unstable compared to our parameterization $\tau\langle\cdot,\cdot\rangle$ proposed here.

Table 1: Comparison of popular variational MI estimators. Here $g(x, y), u(x, y)$ and $u(x)$ are variational functions to be optimized, $\sigma(u) = \frac{1}{1+\exp(-u)}$ is the Sigmoid function, $\mathcal{E}[f(u), \eta]$ denotes exponential average of function $f(u)$ with decay parameter $\eta \in (0, 1)$, and $\alpha \in [0, 1]$ is the balancing parameter used by α-`InfoNCE` trading off bias and variance between `InfoNCE` and `TUBA`. we use (x_i, y_i) to denote positive samples from the joint density $p(x, y)$, and (x_i, y_j) or (x'_k, y'_k) to denote negative samples drawn from the product of marginal $p(x)p(y)$. In context, $y_\oplus$ and $y_\ominus$ have the intuitive interpretation of positive and negative samples. We exclude variational upper bounds here because their computations typically involve the explicit knowledge of conditional likelihoods.

Name	Objective	Bias	Var.	Converge
$(x_i, y_i) \overset{iid}{\sim} p(x, y),\ (x'_k, y'_k) \overset{iid}{\sim} p(x)p(y),\ m_{\alpha,u}(x, y_{1:K}) \triangleq \alpha \frac{1}{K} \left\{ \sum_{k=1}^{K} \exp(g(x, y_k)) \right\} + (1-\alpha) \exp(u(x))$				
`DV` [15]	$g(x_i, y_i) - \log(\sum_{k=1}^{K} \exp(g(x'_k, y'_k))/K)$	high	high	no
`MINE` [7]	$g(x_i, y_i) - \log(\mathcal{E}[\exp(g(x_i, y_j)), \eta])$	low	high	no
`NWJ` [47]	$g(x_i, y_i) - \exp(g(x_i, y_j) - 1)$	low	high	no
`JSD` [33]	$g^*(x_i, y_i) - \exp(g^*(x_i, y_j) - 1)$	low	high	no
	$g^* \xleftarrow{\arg\max} \{\log \sigma(g(x_i, y_i)) + \log \sigma(-g(x_i, y_j))\}$			
`TUBA` [54]	$g(x_i, y_i) + u(x_i) + 1 - \exp(g(x_i, y_j) - u(x_i))$	low	high	no
`InfoNCE` [49]	$g(x_i, y_i) - \log(\sum_j \exp(g(x_i, y_j))/K)$	high	low	no
α-`InfoNCE` [54]	$g(x_i, y_i) - g(x_i, y_j) - \log(m_{\alpha,u}(x, y_{1:K})) + \log(m_{\alpha,u}(x'_k, y'_k))$			no
α-`InfoNCE` interpolates between low-bias high-var ($\alpha \to 1$, `NWJ`) to high-bias low-var ($\alpha \to 0$, `InfoNCE`)				
`FLO` (ours)	$-u(x_i, y_i) - \exp(-u(x_i, y_i) + g(x_i, y_j) - g(x_i, y_i))$	**low**	**moderate**	**yes**

and the associated memory drain, `MINE` further employed an *exponential moving average* (EMA) to aggregate the normalizer across batches. Note for the log-family estimators, their variational gaps are partly caused by the log-transformation on finite-sample average due to Jensen's inequality (*i.e.*, $\log Z = \log \mathbb{E}[\hat{Z}] \leq \mathbb{E}[\log \hat{Z}]$). In contrast, the objective of exponential-family estimators do not involve such log-transformation, since they can all be derived from the Fenchel-Legendre inequality: `NWJ` directly applies the Fenchel dual of f-divergence for MI [48], while `TUBA` exploits this inequality to compute the log partition $\log Z(x) = \log \mathbb{E}_{p(y')}[\exp(g(x, y'))]$. Motivated from a contrastive view, our `FLO` applies the Fenchel-Legendre inequality to the log-partition of contrast scores.

A contrastive view for MI estimation. The MI estimators can also be categorized based on how they contrast the samples. For instance, `NWJ` and `TUBA` are generally considered to be non-contrastive estimators, as their objectives do not compare positive samples against negative samples on the same scale (*i.e.*, log versus exp), and this might explain their lack of effectiveness in representation learning applications. For `JSD`, it depends on a two-stage estimation procedure similar to that in adversarial training to assess the MI, by explicitly contrasting positive and negative samples to estimate the likelihood ratio. This strategy has been reported to be unstable in many empirical settings. The log-family estimators can be considered as a multi-sample, single-stage generalization of `JSD`. However, the `DV` objective can go unbounded thus resulting in a large variance, and the contrastive signal is decoupled by the EMA operation in `MINE`. Designed from contrastive perspectives, `InfoNCE` trades bound tightness for a lower estimation variance, which is found to be crucial in representation learning applications. Our `FLO` formalizes the contrastive view for exponential-family MI estimation, and bridges existing bounds: the PMI normalizer $\exp(-u(x, y))$ is a more principled treatment than the EMA in `MINE`, and compared to `DV` the positive and negative samples are explicitly contrasted and adaptively normalized.

Important FLO variants. We now demonstrate that `FLO` is a flexible framework that not only recovers existing bounds, but also derives novel bounds such as

$$I_{\texttt{FDV}} \triangleq \texttt{StopGrad}[I_{\texttt{DV}}(\{(x_i, y_i)\})] + \frac{\sum_j \exp(c_\theta(x_i, y_i, y_j))}{\texttt{StopGrad}[\sum_j \exp(c_\theta(x_i, y_i, y_j))]} - 1. \tag{11}$$

Recall the optimal $u^*(x, y)$ given $g_\theta(x, y)$ is in the form of $-g_\theta(x, y) + s_\psi(x)$, and parameterizing $u(x, y)$ in this way recovers the `TUBA` bound. Additionally, we note that (i) fixing either of u and g, and optimizing the other also gives a valid lower bound to MI; and (ii) a carefully chosen multi-input $u(\{(x_i, y_i)\})$ can be computationally appealing. As a concrete example, if we set u_ϕ to $\mathtt{u}_\theta(\{(x_i, y_i)\}) \leftarrow \log \left(\frac{1}{K} \sum_j e^{c(x_i, y_i, y_j; g_\theta)} \right)$ and update $u_\theta(x, y)$ while artificially keeping the critic

$g_\theta(x, y)$ fixed [6], then FLO falls back to DV. Alternatively, we can consider the Fenchel dual version of it: using the same multi-input $\mathfrak{u}_\theta(\{(x_i, y_i)\})$ above, treat u_ϕ as fixed and only update g_θ, and this gives us the novel MI objective in (11), we call it *Fenchel-Donsker-Varadhan* (FDV) estimator.

2.4 Gradient and convergence analysis of FLO

In this section, we will establish that FLO better optimizes the MI because its gradient is more accurate than competing variational bounds such as NWJ and TUBA; also, we provide the first convergence analysis for variational MI estimation by showing FLO converges under SGD.

First, recall most tractable variational MI bounds are derived from and upper bounded by the intractable UBA bound [54]. For instance, with the same critic g_θ we have $I_{\texttt{NWJ}} \leq I_{\texttt{TUBA}} \leq I_{\texttt{UBA}}$. So if we can show $\nabla_\theta I_{\texttt{FLO}} \approx \nabla_\theta I_{\texttt{UBA}}$ then FLO is better optimized. To simplify notations, we denote $c_\theta(x, y, y') \triangleq g_\theta(x, y') - g_\theta(x, y)$ and $\mathcal{E}_\theta(x, y) \triangleq 1/\mathbb{E}_{p(y')}[e^{c_\theta(x,y,y')}]$, and we can easily verify

$$\mathbb{E}_{p(y')}\left[\nabla_\theta \left\{e^{c_\theta(x,y,y')}\right\}\right] = \nabla_\theta \left\{\frac{1}{\mathcal{E}_\theta(x, y)}\right\} = -\frac{\nabla \mathcal{E}_\theta(x, y)}{(\mathcal{E}_\theta(x, y))^2} = -\frac{\nabla_\theta \log \mathcal{E}_\theta(x, y)}{\mathcal{E}_\theta(x, y)}. \tag{12}$$

Since for fixed $g_\theta(x, y)$ the corresponding optimal $u^*_\theta(x, y)$ maximizing $I_{\texttt{FLO}}(u_\phi, g_\theta) \triangleq 1 - \left\{u_\phi(x, y) + \mathbb{E}_{p(y')}[e^{-u_\phi(x,y)+c(x,y,y';g_\theta)}]\right\}$ is given by $u^*_\theta(x, y) = \log \mathbb{E}_{p(y')}[e^{c_\theta(x,y,y')}] = -\log \mathcal{E}_\theta(x, y)$ (using (4)), we see that the term $e^{-u_\phi(x,y)}$ is essentially optimized to approximate $\mathcal{E}_\theta(x, y)$. To emphasize this point, we now write $\hat{\mathcal{E}}_\theta(x, y) \triangleq e^{-u_\phi(x,y)}$. When this approximation is sufficiently accurate (*i.e.*, $\mathcal{E}_\theta \approx \hat{\mathcal{E}}_\theta$), we can see that $\nabla I_{\texttt{FLO}}$ approximates $\nabla I_{\texttt{UBA}}$ as follows

$$\begin{aligned}\nabla_\theta\{I_{\texttt{FLO}}(u_\phi, g_\theta)\} &= -\mathbb{E}_{xy}\left[e^{-u_\phi(x,y)}\mathbb{E}_{y'}[\nabla_\theta e^{c_\theta(x,y,y')}]\right] = \mathbb{E}_{xy}\left[\tfrac{\hat{\mathcal{E}}_\theta(x,y)}{\mathcal{E}_\theta(x,y)}\nabla_\theta \log \mathcal{E}_\theta(x, y)\right] \\ &\approx \mathbb{E}_{xy}\left[\nabla_\theta \log \mathcal{E}_\theta(x, y)\right] = \nabla_\theta\left\{\mathbb{E}_{p(x,y)}[\log \mathcal{E}_\theta(x, y)]\right\} = \nabla_\theta\{I_{\texttt{UBA}}(g_\theta)\}.\end{aligned} \tag{13}$$

We can prove FLO will converge under much weaker conditions, even when this approximation $\hat{u}(x, y)$ is rough. The intuition is as follows: in (13), the term $\frac{\hat{\mathcal{E}}_{\theta_t}}{\mathcal{E}_{\theta_t}}$ only rescales the gradient, so the optimizer is still proceeding in the same direction as UBA in SGD. The informal version of our result is summarized in the Proposition below (see the Appendix for the formal version and proof).

Proposition 2.5 (Convergence of FLO, informal version)**.** Let $\{\eta_t\}_{t=1}^\infty$ be the stochastic *Robbins-Monro* sequence of learning rates: $\sum_t \mathbb{E}[\tilde{\eta}_t] = \infty$ and $\sum_t \mathbb{E}[\tilde{\eta}_t^2] < \infty$. If $\frac{\hat{\mathcal{E}}_{\theta_t}}{\mathcal{E}_{\theta_t}}$ is bounded between $[a, b]$ ($0 < a < b < \infty$), then under the stochastic gradient descent scheme described in Algorithm 1, θ_t converges to a stationary point of $I_{\texttt{UBA}}(g_\theta)$ with probability 1, *i.e.*, $\lim_{t\to\infty} \|\nabla I_{\texttt{UBA}}(g_{\theta_t})\| = 0$. Additionally assume $I_{\texttt{UBA}}$ is convex with respect to θ, then FLO converges with probability 1 to the global optimum θ^* of $I_{\texttt{UBA}}$ from any initial point θ_0.

Algorithm 1 FLO

Empirical data $\hat{p}_d = \{(x_i, y_i)\}_{i=1}^n$
Model parameters $\Psi = (\theta, \phi)$
for $t = 1, 2, \cdots$ **do**
  Sample $i, j \overset{iid}{\sim} [n]$
  $u_{ii} = u_\phi(x_i, y_i), g_{ii} = g_\theta(x_i, y_i)$,
  $g_{ij} = g_\theta(x_i, y_j)$
  $\mathcal{F} = u_{ii} + \exp(-u_{ii} + g_{ij} - g_{ii})$
  $\Psi_t = \Psi_t - \eta_t \nabla_\Psi \mathcal{F}$
end for

Importantly, this marks the first convergence result for variational MI estimators. The convergence analyses for MI estimation is non-trivial and scarce even for those standard statistical estimators [51, 20, 56]. The lack of convergence guarantees has led to a proliferation of unstable MI-estimators used in practice (in particular, DV, JSD, and MINE) that critically rely on various empirical hacks to work well (see discussions in [61]). Our work establishes a family of variational MI estimators that provably converges, a contribution we consider significant as it fills an important gap in current literature on both theoretical and practical notes.

3 Experiments

We consider an extensive range of tasks to validate FLO and benchmark it against state-of-the-art solutions. Limited by space, we present only the key results in the main text, and defer ablation studies and details of our experimental setups to the Appendix. Our code is available from `https://github.com/author_name/FLO`. All experiments are implemented with `PyTorch`.

[6] That is to say g_θ in u_ϕ is an independent copy of g_θ.

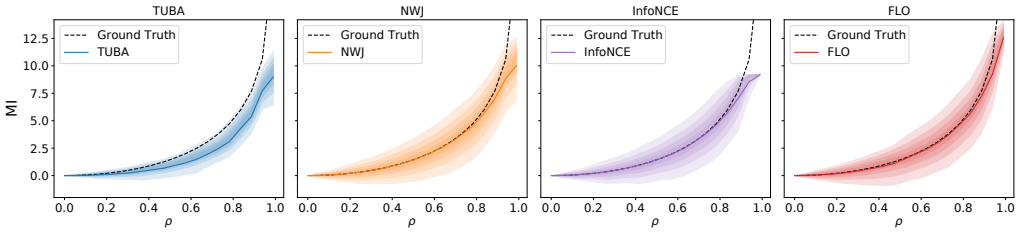

Figure 3: Bias-variance plot for popular variational MI bounds with the 10-D Gaussians. Estimators that are more concentrated around the dashed line is considered better (low-bias, low-variance). In the more challenging high-MI regime, FLO shows a clear advantage over competing alternatives.

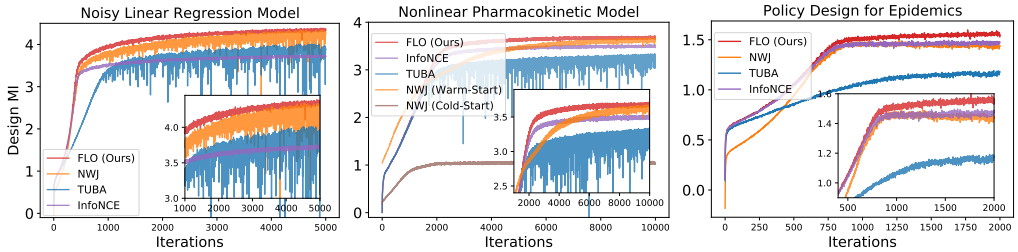

Figure 4: Bayesian Optimal Experiment Design results. FLO consistently performs best, demonstrating superior strength in learning efficiency and robustness. NWJ takes the runner-up, but it has larger variance and is sensitive to network initializations. InfoNCE is less competitive due to low sample inefficiency, but its smaller variance helps in the more challenging dynamic case.

Comparison to baseline MI bounds. We start by comparing FLO to the following popular competing variational estimators: NWJ, TUBA, and InfoNCE. We use the bilinear critic implementation for all models which maximally encourages both sample efficiency and code simplicity, and this strategy does perform best based on our observations. We consider the synthetic benchmark from [54], where $(X \in \mathbb{R}^d, Y \in \mathbb{R}^d)$ is jointly standard Gaussian with diagonal cross-correlation parameterized by $\rho \in [0, 1)$. We report $d = 10$ and $\rho \in [0, 0.99]$ here (other studies only report ρ up to 0.9, which is less challenging.), providing a reasonable coverage of the range of MI one may encounter in empirical settings.

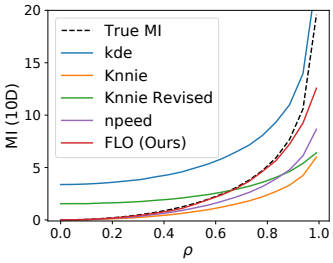

Figure 5: FLO compares favorably to classical MI estimators.

To focus on the bias-variance trade-off, we plot the decimal quantiles in addition to the estimated MI in Figure 3, where FLO significantly outperformed its variational counterparts in the more challenging high-MI regime. In Figure 5, we show FLO also beats classical MI estimators [39, 68, 21]. In the Appendix I, we further demonstrate FLO compare favorably to recent entropy-based estimator MIND, and discuss recent works on parametric estimators [14, 10] and alternative information metrics [73].

Bayesian optimal experiment design (BOED). We next direct out attention to BOED, a topic of significant interest shared by the statistical and machine learning communities [11, 70, 30, 19]. The performance of machine learning models crucially relies on the quality of data supplied, and BOED is a principled framework that optimizes the data collection procedure (in statistical parlance, conducting *experiments*) [18]. Mathematically, let x be the data to be collected, θ be the parameters to be inferred, and d be the experiment parameters the investigator can manipulate (*a.k.a*, the *design parameters*), BOED solves $\arg\max_d I(x; \theta; d)$. We focus on the more generic scenario where explicit likelihoods are not available [37, 38].

We consider three carefully-selected models from recent literature for their progressive practical significance and the challenges involved [17, 35, 36]: static designs of (i) simple linear regression model and (ii) complex nonlinear pharmacokinetic model for drug development; and the dynamic policy design for (iii) epidemic disease surveillance and intervention (*e.g.*, for Covid-19 modeling). In Figure 4 we compare design optimization curves using different MI optimization strategies, where FLO consistently leads. Popular NWJ and InfoNCE reports different tradeoffs that are less susceptible to FLO. We examine the FLO predicted posteriors and confirm they are consistent with the ground-truth parameters (Figure 6). For the dynamic policy optimization, we also manually inspect the design

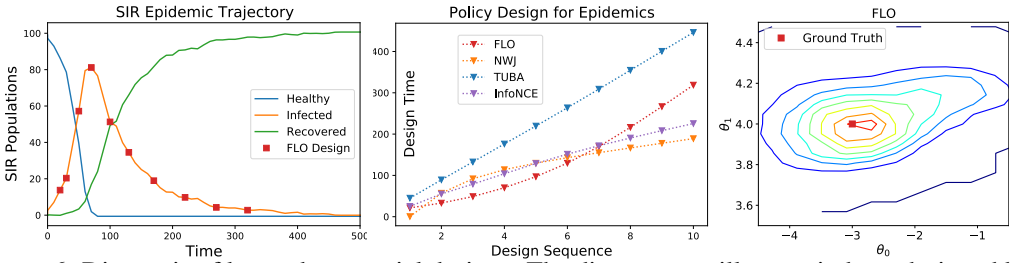

Figure 6: Diagnosis of learned sequential designs. The disease surveillance windows designed by `FLO` makes more sense: measures more frequently as infection spikes, and more sparsely when the pandemic slowly fades. The estimated parameter posterior (right) is consistent with the ground truth.

strategies reported by different models (Figure 6). Consistent with human judgement, `FLO` policy better assigns budgeted surveillance resources at different stages of pandemic progression.

A novel meta-learning framework. A second application of our work is to meta-learning, an area attracting substantial recent interest. In meta-learning, we are concerned with scenarios that at training time, there are abundant different labelled tasks, while upon deployment, only a handful of labeled instances are available to adapt the learner to a new task. Briefly, for an arbitrary loss $\ell_t(\hat{y}, y)$, where t is the task identifier and $\hat{y} = f(x)$ is the prediction made by the model, we denote the risk by $R_t(f) = \mathbb{E}_{p_t(x,y)}[\ell_t(f(x), y)]$. Denote $R(f) \triangleq \mathbb{E}_{t\sim p(t)}[R_t(f)]$ as the expected risk for all tasks and $\hat{R}(f)$ for the mean of empirical risks computed from all training tasks. Inspired by recent information-theoretic generalization theories [72], we derived a novel, principled objective

$$\mathcal{L}_{\texttt{Meta-FLO}}(f) = \hat{R}(f) + \lambda\sqrt{I_{\texttt{FLO}}(\hat{\mathcal{D}}_t; \hat{E}_t)}, \tag{14}$$

where λ is known given the data size and loss function, $(\hat{\mathcal{D}}_t, \hat{E}_t)$ are respectively data and task embeddings for training data, which for the first time lifts contrastive learning to the task and data distribution level. Our reasoning is that $\mathcal{L}_{\texttt{Meta-FLO}}(f)$ theoretically bounds $R(f)$ from above, and it is relatively sharp for being data-dependent. We give more information on this in the Appendix and defer a full exposition to a dedicated paper due to independent interest and space limits here.

Table 2: Multi-view representation learning on `Cifar`

Model	`InfoNCE`	`SpecNCE`	`FLO`	`FDV`
MI	$5.73 \pm .07$	$4.76 \pm .08$	$5.83 \pm .08$	$\mathbf{5.93 \pm .08}$

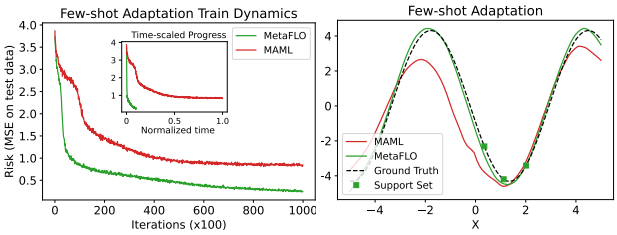

Figure 7: Few-shot adaptation with `Meta-FLO`.

Note other MI bounds are not suitable for this task due to resource and variance concerns. In Figure 7 we show `Meta-FLO` wins big over the state-of-the-art *model agnostic meta-learning* (MAML) model on the regression benchmark from [16].

Converging evidence from more applications. In addition to the tasks above, we also examined `FLO` with extensive ablations and validated its utility of tasks such as representation learning (Table 2) and fair classification. Due to space limits, results and analyses are delegated to the Appendix. We also point readers to other researches inspired by our work, which shows significant boosts in both performance and sample efficiency compared to `InfoNCE`-based SOTA solutions on `ImageNet`-scale datasets using the `FDV` bound [12].

4 Conclusion

We have described a new framework for the contrastive estimation of mutual information from energy modeling perspectives. Our work not only encapsulates popular variational MI bounds but also inspires novel objectives such as `FLO` and `FDV`, which comes with strong theoretical guarantees. In future work, we will leverage our theoretical insights to improve practical applications involving MI estimation, such as representation learning, fairness, and in particular, data efficient learning.

References

[1] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken ELBO. In *ICML*, pages 159–168, 2018.

[2] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017.

[3] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(Jul):1–48, 2002.

[4] David Barber and Felix Agakov. The IM algorithm: a variational approach to information maximization. *NIPS*, 16:201, 2004.

[5] David Barber and Felix V Agakov. Information maximization in noisy channels: A variational approach. *NIPS*, 16, 2003.

[6] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 5(4):537–550, 1994.

[7] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML*. PMLR, 2018.

[8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[9] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[10] Rob Brekelmans, Sicong Huang, Marzyeh Ghassemi, Greg Ver Steeg, Roger Baker Grosse, and Alireza Makhzani. Improving mutual information estimation with annealed and energy-based bounds. In *ICLR*, 2021.

[11] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[12] Junya Chen, Zhe Gan, Xuan Li, Qing Guo, Liqun Chen, Shuyang Gao, Tagyoung Chung, Yi Xu, Belinda Zeng, Wenlian Lu, et al. Simpler, faster, stronger: Breaking the log-k curse on contrastive learners with flatnce. *arXiv preprint arXiv:2107.01152*, 2021.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[14] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In *ICML*, 2020.

[15] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135. PMLR, 2017.

[17] Adam Foster, Desi R Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. In *ICML*, 2021.

[18] Adam Foster, Martin Jankowiak, Eli Bingham, Paul Horsfall, Yee Whye Teh, Tom Rainforth, and Noah Goodman. Variational bayesian optimal experimental design. In *NeurIPS*, 2019.

[19] Adam Foster, Martin Jankowiak, Matthew O'Meara, Yee Whye Teh, and Tom Rainforth. A unified stochastic gradient approach to designing bayesian-optimal experiments. In *AISTATS*, pages 2959–2969. PMLR, 2020.

[20] Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *AISTATS*. PMLR, 2015.

[21] Weihao Gao, Sewoong Oh, and Pramod Viswanath. Demystifying fixed k-nearest neighbor information estimators. *IEEE Transactions on Information Theory*, 64(8):5629–5661, 2018.

[22] Charles J Geyer. On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):261–274, 1994.

[23] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, Bernhard Schölkopf, et al. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 2005.

[24] Arthur Gretton, Ralf Herbrich, and Alexander J Smola. The kernel mutual information. In *ICASSP*, 2003.

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

[26] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *SIGKDD*, 2016.

[27] Umang Gupta, Aaron Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. *arXiv preprint arXiv:2101.04108*, 2021.

[28] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[30] José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *NIPS*, 2014.

[31] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[32] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[33] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

[34] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

[35] Desislava Ivanova, Adam Foster, Steven Kleinegesse, Michael U Gutmann, and Thomas Rainforth. Implicit deep adaptive design: Policy-based experimental design without likelihoods. *NeurIPS*, 34, 2021.

[36] Steven Kleinegesse, Christopher Drovandi, and Michael U Gutmann. Sequential bayesian experimental design for implicit models via mutual information. *Bayesian Analysis*, 1(1):1–30, 2021.

[37] Steven Kleinegesse and Michael U Gutmann. Bayesian experimental design for implicit models by mutual information neural estimation. In *ICML*. PMLR, 2020.

[38] Steven Kleinegesse and Michael U Gutmann. Gradient-based bayesian experimental design for implicit models using mutual information lower bounds. *arXiv preprint arXiv:2105.04379*, 2021.

[39] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[40] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[41] Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. *arXiv preprint arXiv:1809.01812*, 2018.

[42] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[43] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

[44] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. *arXiv preprint arXiv:1811.04251*, 2018.

[45] Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *NIPS*, 2013.

[46] Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

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

[48] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *NIPS*, 2016.

[49] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[50] Stephanie E Palmer, Olivier Marre, Michael J Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015.

[51] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

[52] Fernando Pérez-Cruz. Estimation of information theoretic measures for continuous random variables. In *NIPS*, 2008.

[53] Josien PW Pluim, JB Antoine Maintz, and Max A Viergever. Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging*, 22(8):986–1004, 2003.

[54] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*. PMLR, 2019.

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[56] Tom Rainforth, Rob Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting monte carlo estimators. In *International Conference on Machine Learning*, pages 4267–4276. PMLR, 2018.

[57] Linda E Reichl. *A modern course in statistical physics*. John Wiley & Sons, 2016.

[58] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[59] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[60] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[61] Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *ICLR*, 2020.

[62] Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, pages 5–20. PMLR, 2008.

[63] Chenyang Tao, Liqun Chen, Shuyang Dai, Junya Chen, Ke Bai, Dong Wang, Jianfeng Feng, Wenlian Lu, Georgiy Bobashev, and Lawrence Carin. On fenchel mini-max learning. In *NeurIPS*, 2019.

[64] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[65] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

[66] Kari Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of machine learning research*, 2003.

[67] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *ICLR*, 2020.

[68] Greg Ver Steeg and Aram Galstyan. Information-theoretic measures of influence based on content dynamics. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 3–12, 2013.

[69] Liangjian Wen, Yiji Zhou, Lirong He, Mingyuan Zhou, and Zenglin Xu. Mutual information gradient estimation for representation learning. In *ICLR*, 2020.

[70] CF Jeff Wu and Michael S Hamada. *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons, 2011.

[71] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[72] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *NIPS*, 2017.

[73] Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *ICLR*, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...

(a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
(b) Did you describe the limitations of your work? [Yes]
(c) Did you discuss any potential negative societal impacts of your work? [N/A]
(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] They are in the Supplementary Material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] They are summarized in the Supplementary Material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Smaller variance is the highlight of this paper.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]