Fast Variational Estimation of Mutual Information for Implicit and Explicit Likelihood Models

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Computing the mutual information (MI) between sets of random variables lacks a closed-form solution in nontrivial models. Variational MI approximations are widely used as flexible estimators for this purpose, but computing these estimators typically requires solving a costly non-convex optimization. In this paper we demonstrate a class of variational estimators whose solution is not only convex, but efficiently solved via moment matching conditions. We show that the moment-matched variational estimator provides optimal upper and lower bounds on MI in models with explicit forward models. Furthermore, we show that the same moment matching solution yields accurate MI approximations in so-called "implicit likelihood models", where the observation likelihood lacks an analytic representation. In further analysis on implicit likelihood models we prove that our moment matching solution is equivalent to existing gradient based methods but at a significantly reduced computational cost. Our theoretical results are supported by numerical evaluation, showing that the proposed approach elegantly applies to fully parameterized (explicit likelihood) Gaussian mixture models, as well as implicit models arising from marginalization of nuisance variables. Finally, using the SIR model in epidemiology, we show that our approach easily applies to implicit simulation-based likelihood models, while avoiding costly Likelihood-Free Inference Ratio Estimation (LFIRE) common to such models.

1 Introduction

In this paper we address a fundamental problem of measuring the information shared between random quantities. The focus of this work is the *mutual information* (MI), which is key in a diverse range of applications. For example, in Bayesian optimal experiment design (BOED) [13, 5, 3] MI is used to measure the amount of information provided by each hypothesized experiment. Additionally, MI plays a key roll in measuring and optimizing the amount of information that can be transmitted along noisy communication channels [6, 14]. MI plays a key role in optimizing sensor configurations [12], sensor selection [22], active learning [19], representation learning [21], and many other applications.

Despite its broad use, exact calculation of MI is typically not possible. Sample-based estimates of MI can be inefficient both in terms of computation and sample complexity. Such sample-based estimators require Nested Monte Carlo (NMC) estimation, which exhibits large finite sample bias that decays slowly [23, 18]. Additionally, straightforward Monte Carlo estimation cannot be applied in so-called *implicit likelihood* models that lack a closed-form data generating distribution. Such models typically require likelihood-free inference by ration estimation (LFIRE) [20], which can be slow due to repeated fitting of generalized linear models (GLMs) in an inner-loop [11].

Recent variational approaches provide an appealing alternative to MI estimation by recasting the calculation as an optimization problem [17]. Such methods provide convenient lower bounds [2], upper bounds, and even apply in the setting of implicit likelihood models [7]. These approaches have proven successful in a range of sequential decision making tasks [16, 8]. Yet, despite their computational benefits, computing such estimators can still be prohibitive due to the underlying non-convex optimization.

In this paper we improve the computational properties of several existing variational bounds and approximations to MI. We show that for a class of variational distributions the optimal upper and lower bounds on MI can be computed by matching moments of the model. The resulting estimates are equivalent to existing estimators in previous work [17, 2] but we show that they can be computed with a fraction of the computation. Furthermore, we consider the case of MI approximation for implicit likelihood models. We show that the same moment matching solution yields equivalent estimates to existing work [7] and minimizes an upper bound on absolute approximation error. In doing so we unify the solution of all three estimators (upper / lower bounds and implicit approximation) at a fraction of the computational cost of existing methods.

We show the accuracy and speed of our approach compared to the variational and sample-based estimates in a variety of experiments of both *explicit* and *implicit* likelihood models. We begin begin with estimating MI in Gaussian mixtures, for which entropy is notoriously difficult to compute [9]. Our proposed approach yields accurate estimates with computation that is capable of scaling to very high-dimensional GMMs. We then consider two implicit likelihood models: one, a variation of the generalized linear model ("Extrapolation Experiment") from Foster et al. (2019) and the other a simulation-based SIR Epidemiological model as explored in [11]. In all cases we find that our method offers substantial speedup while still producing high-quality MI approximations and (when possible) bounds.

2 Computing and Approximating MI

Consider an arbitrary joint distribution $p(x, y)$ with latent variable x and observable variable y. The shared information between these can be computed via the *mutual information* (MI) [6, 14]:

$$I(X;Y) = H(Y) - H(Y \mid X). \tag{1}$$

The *marginal entropy* is given by $H(Y) = \mathbb{E}[-\log p(Y)]$ while the *conditional entropy* is $H(Y \mid X) = \mathbb{E}[-\log p(Y \mid X)]$. Entropy expectations are taken with respect to the joint $p(x, y)$.

2.1 Calculating MI : Explicit and Implicit Models

Despite its simple definition (Eqn. (1)) calculating MI is difficult in practice since entropy terms require exact evaluation of the probabilities. For example, calculating the marginal entropy $H(Y)$ requires evaluation of $\log p(y)$, which often lacks a closed-form. Similarly, $\log p(y \mid x)$ may lack a closed-form in so-called *implicit likelihood models* that require marginalization of nuisance variables (c.f. Fig. 1) or are defined by simulation as in the SIR model of Sec. 6.3. Another option is to use the symmetric form $I(X;Y) = H(X) - H(X \mid Y)$. But this approach requires evaluation of the posterior $\log p(x \mid y)$, which is also not generally closed-form. For these reasons approximations must be considered, such as the commonly employed sample-based estimators discussed next.

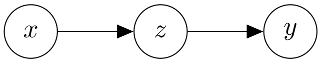


Figure 1: **Implicit Likelihood via Nuisance Variables** Likelihood $p(y \mid x)$ marginalizes z.

2.2 Nested Monte Carlo (NMC) Estimation

Given samples $\{(x^i, y^i)\}_{i=1}^N \sim p$ one may use a simple Monte Carlo procedure to estimate MI,

$$\hat{I}_{NMC} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(y^i \mid x^i)}{\frac{1}{N} \sum_{j=1}^{N} p(y^i \mid x^j)} \tag{2}$$

The use of a plug-in estimator for the marginal $p(y^i) \approx \frac{1}{N} \sum_j p(y^i \mid x^j)$ makes this a *nested* Monte Carlo (NMC) estimator. The NMC is consistent, but exhibits considerable finite sample bias, as can be shown by Jensen's inequality [23, 18]. Due to its bias NMC is often used as a probabilistic bound on MI, but the bound gap can be significant as bias decays slowly. A bigger limitation is that the NMC estimator (Eqn. (2)) requires pointwise evaluation of the conditional probability $p(y \mid x)$, which may be impossible for simulation-based implicit likelihood models, such as the SIR Epidemiology model in Sec. 6.3.

3 Variational MI Estimation

Variational MI estimators [17] address the computational and sample complexity issues of NMC estimators by recasting MI calculation as an optimization problem. In some cases we can obtain MI bounds using Gibbs' inequality. The proof is a result of non-negativity of the Kullback-Leibler divergence, briefly: $\mathrm{KL}(p \,\|\, q) = H_p(q) - H(p) \geq 0$, and so we can bound entropy as $H_p(q) \geq H(p)$. In other cases we desire an approximation, rather than a bound. We discuss both cases.

3.1 Variational MI Bounds

Applying Gibbs' inequality to the conditional entropy $H(X \mid Y) \leq H_p(q(X \mid Y))$ we have the lower bound [2],

$$I(X;Y) \geq \max_q \; H(X) - H_p(q(X \mid Y)) \equiv \hat{I}_{\text{post}}. \tag{3}$$

which we call the *variational posterior lower bound*. Observe that calculation of the lower bound $\hat{I}_{\text{post}}$ requires evaluation of the marginal entropy $H(X)$ under the model p, which may be prohibitive. Applying Gibbs' inequality, instead, to the marginal entropy $H(X) \leq H_p(q(X))$ we obtain the *variational marginal upper bound*,

$$I(X;Y) \leq \min_q \; H_p(q(X)) - H(X \mid Y) \equiv \hat{I}_{\text{marg}} \tag{4}$$

Observe that evaluation of the upper bound $\hat{I}_{\text{marg}}$ requires evaluation of the conditional entropy $H(X \mid Y)$ under the model p. For this reason, both bounds ($\hat{I}_{\text{post}}$ and $\hat{I}_{\text{marg}}$) apply only when the model entropy terms can be calculated or ignored–typically true for sequential decision making [16, 7, 8, 2].

3.2 Variational MI Approximation : Implicit Likelihood Models

In many cases, the model entropy terms in Eqns. (3) and (4) cannot be calculated and so we cannot obtain MI bounds. By replacing both entropy terms with their cross-entropies we have the following approximation [7]:

$$I(X;Y) \approx H_p(q_m(X)) - H_p(q_p(X \mid Y)) \equiv \hat{I}_{\text{m}+p} \tag{5}$$

where the variational distributions are $q_m(x)$ (marginal) and $q_p(x \mid y)$ (posterior). Reversing the entropy terms yields an analogous estimator: $\hat{I}_{m+\ell} \equiv H_p(q_m(Y)) - H_p(q_\ell(Y \mid X))$ Both estimators avoid evaluation of model probabilities, and thus are useful for implicit likelihood models. We focus on $\hat{I}_{\text{m}+p}$ for consistency, but note that our results in Sec. 4 apply equally to $\hat{I}_{m+\ell}$, which is the form discussed in Foster et al. (2019). In Sec. 4 we will discuss how to find the best such approximation.

4 Moment Matching Variational MI Estimators

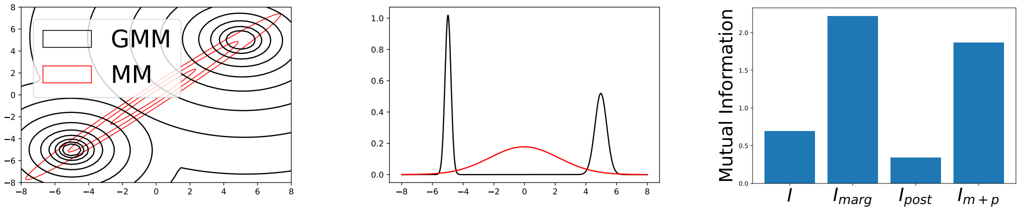
In general, computing the optimal variational estimators (e.g. $\hat{I}_{\text{marg}}$, $\hat{I}_{\text{post}}$, and $\hat{I}_{\text{m}+p}$) requires solving nonlinear–and often nonconvex–optimization problems. In the following sections we demonstrate a class of variational distributions in the exponential family that correspond to an efficient convex optimization. Specifically, for Gaussian variational distributions the optimal estimators can be solved in closed-form by matching expected sufficient statistics (means and variances). The same (efficient) moment calculation yields optimal (or optimal bounded) estimators for all three cases. Unless provided, all proofs can be found in the Appendix.

4.1 Exponential Families

Our results rely heavily on properties of the exponential family, which we briefly review here. A joint distribution $q(x, y)$ is a member of the exponential family if the PDF / PMF is of the following form,

$$q(x, y) = h(x, y) \exp\left[\eta^T T(x, y) - A(\eta)\right]. \tag{6}$$

where η are the *natural parameters*, $h(x, y)$ is the *base measure*, $T(x, y)$ the *sufficient statistics*, and $A(\eta)$ is the *log-partition function*. The exponential family includes many well-known distributions: Bernoulli, Categorical, Poisson, Gamma, Gaussian, etc. In addition to the natural parameters η each



(a) Joint pdf contour (b) Marginal pdf (c) MI approximations

Figure 2: **Moment Matched Gaussian Mixture Model** (a) A bimodal GMM p overlaid with the moment matched Gaussian q has its .5, 1, and 1.5 standard deviation level curves plotted on top in red. (b) The marginal PDF is plotted for the Gaussian mixture model and the moment matched Gaussian. (c) The true $I(X, Y)$ is shown with estimates $\hat{I}_{\text{marg}}$, $\hat{I}_{\text{post}}$, and $\hat{I}_{\text{m}+p}$ are all plotted. Notice that $\hat{I}_{\text{marg}} \geq \hat{I}_{\text{m}+p} \geq \hat{I}_{\text{post}}$.

exponential family has an alternate set of *mean parameters* μ, defined as the expected sufficient statistics: $\mu = \mathbb{E}_q[T(x, y)]$. Mean parameters play a key role in finding projections onto the exponential family, as shown in Lemma 4.1.

Lemma 4.1 (Moment Matching Projection)**.** *For any distribution $p(x, y)$ and exponential family $q(x, y)$ whose support includes that of p the minimum Kullback-Leibler projection:*

$$q^* = \operatorname*{argmin}_q \text{ KL}(p(X, Y) \,\|\, q(X, Y)) = \operatorname*{argmin}_q \mathbb{E}_p \left[\log \frac{p(X, Y)}{q(X, Y)} \right]$$

is convex and the solution given by moment matching *conditions:* $\mathbb{E}_p[T(X, Y)] = \mathbb{E}_q[T(X, Y)] = \mu^*$

The interested reader can consult the texts [4, 15] for a proof of Lemma 4.1 and more details on the exponential family. In this paper, we will focus on Gaussian $q(x, y)$. Fig. 2 shows an example of a GMM, $p(x, y)$, with a moment matched variational Gaussian, $q(x, y)$ (left), corresponding marginal projection (center), and resulting variational estimators (right).

4.2 Variational Marginal (upper bound)

To optimize the variational marginal upper bound, $\hat{I}_{\text{marg}}$, we must minimze the bound gap.

$$I(X; Y) \leq \min_{q_m} H_p(q_m(X)) - H_p(X \mid Y) = \min_{q_m(X)} H_p(q_m(X)) - \text{const.} \tag{7}$$

where the model entropy $H_p(X \mid Y)$ is constant w.r.t. $q_m(x)$ and can be ignored. If $q_m(x)$ is a Gaussian, then the minimization is found by moment matching:

Lemma 4.2. *Let $q_m(x)$ be in the exponential family with statistics $T(x)$, then for any $p(x, y)$, the optimal $\hat{I}^*_{marg}$ is given by moment matching:*

$$\mathbb{E}_{q_m(x)}\left[T(X)\right] = \mathbb{E}_{p(x)}\left[T(X)\right]$$

Proof. Since $H_p(X)$ is constant in q_m we have,

$$\operatorname*{argmin}_{q_m} H_p(q_m(X)) = \operatorname*{argmin}_{q_m} H_p(q_m(X)) - H_p(X) = \operatorname*{argmin}_{q_m} \text{KL}(p(X) \,\|\, q(Y))$$

By Lemma 4.1, $\mathbb{E}_{q_m(x)}\left[T(X)\right] = \mathbb{E}_{p(x)}\left[T(X)\right]$ minimizes $\text{KL}(p(X) \,\|\, q(Y))$. □

Now consider the Gaussian case where $q_m(x) = \mathcal{N}(m, S)$ and a joint Gaussian $q(x, y) = \mathcal{N}(\mu, \Sigma)$. Moment matching the joint distribution, $\mathbb{E}_{q(x,y)}\left[T(X, Y)\right] = \mathbb{E}_{p(x,y)}\left[T(X, Y)\right]$, provides the marginal moment matching condition. To see this, note that the marginal moments of the joint are $q(x) = \mathcal{N}(m_x, \Sigma_{xx})$ where $m_x = m$ is the x component of the mean and $\Sigma_{xx} = S$ is the block variance of x in Σ. Thus, matching the joint Gaussian statistics satisfies Lemma 4.2 and yields the optimal Gaussian q_m and corresponding optimal $\hat{I}_{\text{marg}}$.

4.3 Variational Posterior (lower bound)

To optimize the variational posterior lower bound, $\hat{I}_{\text{post}}$, we must minimize the bound gap.

$$I(X;Y) \geq \max_{q_p} H_p(X) - H_p(q_p(X \mid Y)) = \min_{q_p} H_p(q_p(X \mid Y)) + \text{const} \tag{8}$$

The $H_p(X)$ term is constant in q_p and can be ignored for optimization. The maximization on the left is turned into a minimization on the right by the negative leading the conditional entropy. Eqn. (8) is optimized by satisfying the following condition,

Lemma 4.3. *If $q_p(x \mid y)$ takes the form of Eqn. 10, then the minimization of Eqn. (8) is when*

$$\mathbb{E}_{p(y)}\left[\mathbb{E}_{q_p(x|y)}\left[T(X,Y)\right]\right] = \mathbb{E}_{p(x,y)}\left[T(X,Y)\right] \tag{9}$$

This result was also shown previously by Pacheco and Fisher (2019). The condition in Lemma 4.3 is seemingly a moment matching condition. However, the l.h.s. of Eqn. 9 is an expectation w.r.t. mixed distributions $p(y)q_p(x \mid y)$, and is difficult to satisfy in general. To simplify we consider the joint exponential family distribution $q(x,y;\eta)$ with natural parameters η and conditional given by,

$$q_p(x \mid y) = q(x \mid y;\eta) = \frac{q(x,y;\eta)}{q(y;\eta)} \tag{10}$$

Note that $q(y;\eta) = \int q(x,y;\eta)\,dx$ is not necessarily in the exponential family. We can now show that moment matching joint statistics of $q(x,y;\eta)$ yields the optimal $\hat{I}_{\text{post}}$ via the following Lemma.

Lemma 4.4. *Let $q_p(x \mid y)$ takes the form of Eqn. 10. Further, let the posterior expected statistics be a linear combination of marginal statistics as in,*

$$\mathbb{E}_{q_p(x|y)}\left[T(X,Y)\right] = \sum_i^k g_i(\eta)T_i(Y) \tag{11}$$

where $T_i(y)$ is the i^{th} component of the joint statistics only depending on y and $g_i(\eta)$ are arbitrary functions. Then, the optimal $\hat{I}_{post}$ is given by joint moment matching: $\mathbb{E}_{p(x,y)}[T(X,Y)] = \mathbb{E}_{q(x,y)}[T(X,Y)]$.

Thus both of these conditions imply that moment matching the joint is optimal for Eqn. (8). Each of these lemmas are written in term of an exponential family distribution with some conditions, since the multivariate Gaussian is the focus of this paper we need to show that it satisfies these conditions

Corollary 4.5. *Let $q(x,y) = \mathcal{N}(m,\Sigma)$ be a Gaussian. Then $q_p(x \mid y)$ is also Gaussian and satisfies conditions of both lemma 4.3 and lemma 4.4. Furthermore, the optimal $\hat{I}_{post}$ is obtained by joint Gaussian moment matching conditions,*

$$m^* = \mathbb{E}_{p(x,y)}\left[(X,Y)^T\right], \qquad \Sigma^* = Cov_{p(x,y)}\left((X,Y)^T\right)$$

And moments of $q_p(x \mid y)$ are the corresponding Gaussian conditional moments of m^* and Σ^*.

4.4 Variational Approximation

Since $\hat{I}_{\text{m}+p}$ is neither an upper nor lower bound, we must minimize the absolute error

$$\hat{I}^*_{\text{m}+p} = \underset{q_m,q_p}{\text{argmin}} \left| I(X;Y) - \hat{I}_{\text{m}+p}(q_m,q_p) \right| \tag{12}$$

which is non-convex in general. We instead minimize the upper bound as in Foster et al. (2019),

Lemma 4.6. *For any model $p(x,y)$ and distributions $q_m(x)$, $q_p(x \mid y)$, the following bound holds:*

$$\left|\hat{I}_{m+p}(X,Y) - I(X,Y)\right| \leq -\mathbb{E}_{p(x,y)}\left[\log q_m(X) + \log q_p(X \mid Y)\right] + C$$

where $C = -H_p(p(X)) - H_p(p(X \mid Y))$ does not depend on q_m or q_p. Further, the RHS is 0 iff $q_m(x) = p(x)$ and $q_p(x \mid y) = p(x \mid y)$ almost surely.

Previous optimization approaches [7] minimize this upper bound using (stochastic) gradient descent:

$$q^*_m = \underset{q_m}{\text{argmax}}\, \mathbb{E}_{p(x,y)}[\log(q_m(X))] \qquad q^*_p = \underset{q_p}{\text{argmax}}\, \mathbb{E}_{p(x,y)}[\log(q_p(X \mid Y))] \tag{13}$$

Note that Eqn. (13) does not assume that $q_m(x)$ and $q_p(x \mid y)$ share a joint distribution $q(x,y)$. We show that under Gaussianity conditions, not only are optimal q_m and q_p the marginal and posterior of a joint Gaussian, but that the optimal joint is found via moment matching.

Theorem 4.7. *Equivalence of Moment Matching and Stochastic Gradient Descent* *Let $q_m(x)$ and $q_p(x \mid y)$ be exponential family. Further, let $q_p(x \mid y)$ satisfy the form of Eqn.* (10) *and the linear conditional expectations property (Eqn.* (11)*). Then, moment matching the joint distribution $q(x, y)$ yields optimal q_m and q_p that minimize the bound on $\hat{I}_{m+p}$ in Lemma 4.6.*

The proof of Theorem 4.7 (Appendix) follows immediately from Lemma 4.2 and Lemma 4.4. Finally, we show that the general result in Theorem 4.7 is satisfied for Gaussian $q_m(x) = \mathcal{N}(m, S)$ and a linear conditional Gaussian $q_p(x \mid y) = \mathcal{N}(Ay + b, \Sigma_p)$ satisfy.

Corollary 4.8. *Let $q_m(x) = \mathcal{N}(m, S)$ and $q_p(x \mid y) = \mathcal{N}(Ay + b, \Sigma_p)$. Then theorem 4.7 is satsified and thus moment matching a Joint Gaussian $q(x, y) = \mathcal{N}(\mu, \Sigma)$ will minimize lemma 4.6.*

4.5 Properties of Variational Estimators

We pause here to reflect on the implication of our results in Sec. 4.2 - 4.4. Namely, for joint Gaussian variational $q(x, y)$ all optimality conditions are satisfied by joint moment matching. Thus, the same moment-matched joint Gaussian is optimal for all three variational estimators. For the approximation $\hat{I}_{\text{m}+p}$, joint moment matching is equivalent to optimizing an upper bound on error, but is not globally optimal in general. We conclude this section with additional properties of these estimators. For example, it is trivial to show that they obey the following ordering:

Lemma 4.9. *For any $q_m(x)$ and $q_p(x \mid y)$, $\hat{I}_{post} \leq \hat{I}_{m+p} \leq \hat{I}_{marg}$.*

Thus, $\hat{I}_{\text{m}+p}$ is never the least accurate out of all three methods. If $\hat{I}_{\text{m}+p}$ is an over approximation, it is a tighter upper bound than $\hat{I}_{\text{marg}}$ and the converse holds if it is an under approximation (it is a tighter than $\hat{I}_{\text{post}}$). We can also ask when $\hat{I}_{\text{m}+p}$ is closer in absolute error than either of the bounds

Lemma 4.10. *For a variational $q_m(x)$ and $q_p(x \mid y)$, if*

1. *If* $\mathrm{KL}(p(X \mid Y) \,\|\, q(X \mid Y)) \geq \frac{1}{2}\mathrm{KL}(p(X) \,\|\, q(X))$ *then $\hat{I}_{m+p}$ has lower error than $\hat{I}_{post}$*
2. *If* $\mathrm{KL}(p(X) \,\|\, q(X)) \geq \frac{1}{2}\mathrm{KL}(p(X \mid Y) \,\|\, q(X \mid Y))$ *then $\hat{I}_{m+p}$ has lower error than $\hat{I}_{marg}$*

While these conditions cannot be checked in practice they do offer some insight. For example, if $q_m(x)$ approximates $p(x)$ about as well as $q_p(x \mid y)$ approximates $p(x \mid y)$ (in KL) then $\hat{I}_{\text{m}+p}$ is the best approximation to use.

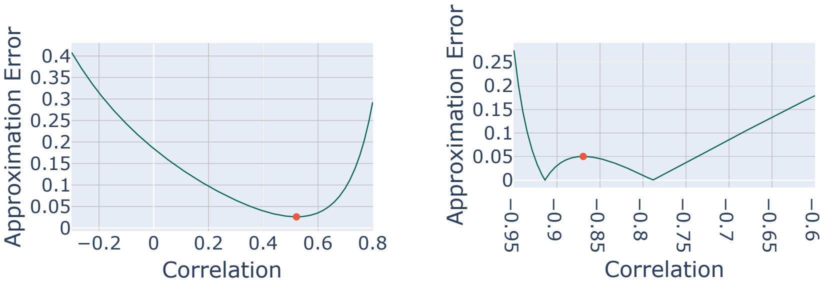
Moment matching stationary point is not always a local minimum to the optimization of $\hat{I}_{\text{m}+p}$. Consider Fig. 3 where a two dimensional Gaussian is moment matched to Gaussian mixture models. All of the parameters of the moment matched Gaussian are held constant but the correlation parameter, ρ, is varied an the absolute error of $\hat{I}_{\text{m}+p}$ is plotted. In one case, the minimum is found and any changed value of ρ results in a worse approximation of MI. However is the second case, a local maximum is found, and we notice that there is a range of values for ρ that result in not only better approximations of MI, but sometimes exact. Exploring this property is a topic of future work.

5 Previous Work

In this paper, we focused on the variational methods, $\hat{I}_{\text{marg}}$, $\hat{I}_{\text{post}}$ [2], and $\hat{I}_{\text{m}+p}$. The focus of each of these methods was computation speed ups for computing the optimal distribution. We also breifly discussed the Nested Monte Carlo estimator in Sec. 2 and some of the challenges it faced. For each of these methods, Foster et al. [7] does a much more thorough analysis of convergence rate and run time. For an alternative implicit likelihood approximator, we also consider the likelihood-free inference by ratio (LFIRE) used by Kleinegesse and Gutmann [11] as a baseline for comparison purposes. For a general discussion of a variety of variational methods, a good resource is Poole et al [17] or Foster et al [8].

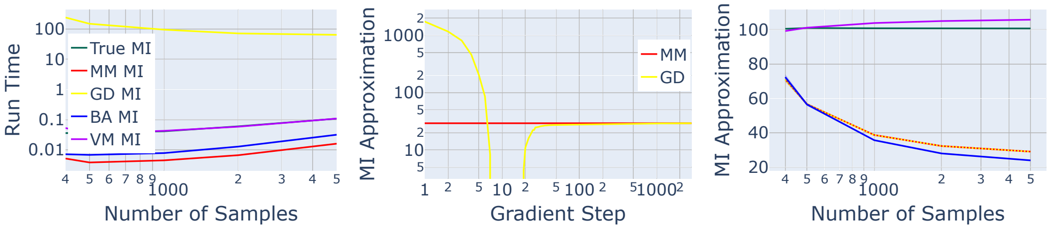
6 Experiments

We demonstrate efficacy and efficiency of our moment matching variational MI estimators in a range of experiments beginning with a Gaussian mixture model (Sec. 6.1). We then evaluate two implicit



(a) Local minimum (b) Local maximum

Figure 3: **Moment Matched Optimum** (a) The moment matching solution (red) is the local minimum as a variational apprximation to a GMM. (b) For a seperate GMM, the moment matched solution is a local maximum and there is a range of values for ρ that result in better approximation, and two that result in exact values of MI.



(a) Computation Time (b) GD Convergence (c) MI Convergence

Figure 4: **High-dimensional Bimodal GMM.** Moment matching estimates have orders of magnitude lower computation as a function of sample size (a). As per our theoretical analysis, our moment-matching solution to $\hat{I}_{\text{m+}p}$ achieves the same estimate while avoiding gradient iterations (b). In this model we see that $\hat{I}_{\text{marg}}$ yields the most accurate estimates. “True MI” is calculated via Monte Carlo estimation with exact evaluation of the model probabilities.

likelihood models: one arising from the non-closed-form marginalization of nuisance variables in a GLM (Sec. 6.2), and the other is a simulation-based SIR epidemiology model (Sec. 6.3). In all cases we find that the proposed moment matching estimators offer substantial computational speedups while achieving identical MI bounds and approximations to existing methods.

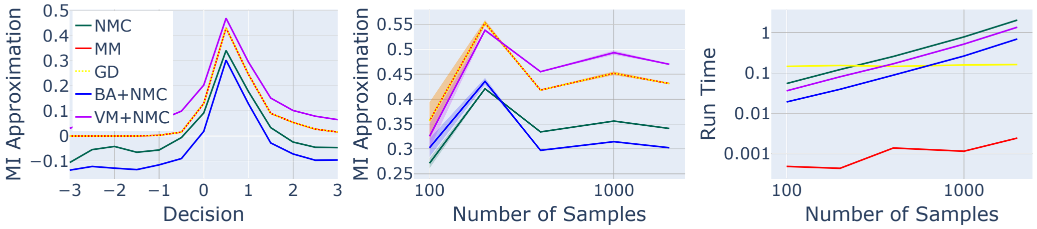
6.1 Multivariate Gaussian Mixture Model

GMMs are pervasive in statistics due to their universal approximation properties, yet calculating MI for a GMM is notoriously challenging [9]. In this section we extend the two-dimensional example (Fig. 2) to high-dimensional GMMs. We simulate a bivariate GMM,

$$p(x, y) = \omega \mathcal{N}(m_0, \Sigma_0) + (1 - \omega)\mathcal{N}(m_1, \Sigma_1) \tag{14}$$

with $\omega \in [0, 1]$ and dimensions $X \in \mathbb{R}^{100}$ and $Y \in \mathbb{R}^{200}$. We use this setting to demonstrate efficient MI estimation even in very high-dimensional distributions.

Fig. 4 shows substantial speedups in runtime (left) for all methods as compared to gradient optimizatoin (center). Notice that GD takes approximately 2000 gradient steps to converge for $5,000$ samples whereas moment matching found this solution immediately, independ of any gradient steps. As per our theoretical results we find that $\hat{I}_{\text{m+}p}$ always lies between the MI upper bound $\hat{I}_{\text{marg}}$ and lower bound $\hat{I}_{\text{post}}$ with $\hat{I}_{\text{marg}}$ being most accurate estimator in this model (right).



(a) MI Convergence (b) Computation Time (c) Samples

Figure 5: **Extrapolation** (a) The MI for decisions $d \in [-3, 3]$ with variances $\sigma_x^2 = 3$, and $\sigma_y^2 = 1$ for each approximation with 4000 samples is plotted with $d = .5$ being the maximum. Note the negative bias resulting from NMC. (b) The convergence rate versus samples is plotted at maximum decision ($d = .5$) (c) The run time for each method is plotted with moment matching being orders of magnitude faster than any other method.

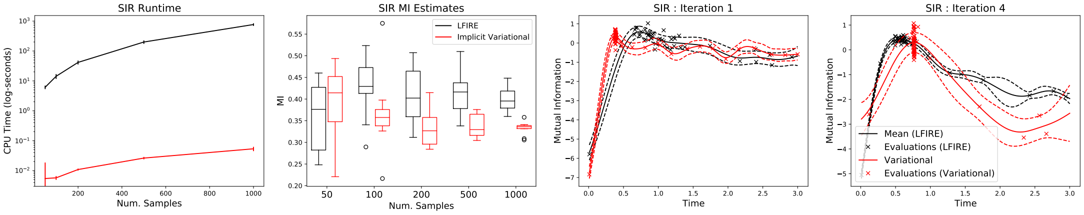


Figure 6: **SIR Sequetial Design.** Left plots show benchmark time and utility evaluation between LFIRE and the Variational estimator for a fixed design ($d = 1.0s$) over 10 runs each at a range of sample sizes. The variational estimator is orders of magnitude more efficient (*left*) and shows lower variance at each sample (*center-left*). The first (*center-right*) and fourth (*right*) sequential BED iterations yield comparable designs between both methods (GP posterior MI shown).

6.2 Extrapolation

We adapt the following experiment from Foster et al. (2019) intended to evaluate the implicit likelihood MI estimator $\hat{I}_{\text{m}+p}$ (or $\hat{I}_{m+\ell}$). Labeled information, y, from a subset of the design space is be used to predict labels at a location x that can't be directly observed. The model is as follows,

$$\psi \sim \mathcal{N}(\mu_\psi, \Sigma_\psi), \quad \theta \mid \psi \sim \mathcal{N}\left((X_\theta^T \psi)^2, \sigma_x^2\right) \quad y \mid \psi, d \sim \mathcal{N}\left((X_d^T \psi)^2, \sigma_y^2\right)$$

where $X_x = (1, -\frac{1}{2})$ and $X_d = (-1, d)$. The aim is to choose a design $d \in \mathbb{R}$ that maximizes $I(\theta; Y \mid d)$. Thus, ψ is a nuisance variable that must be marginalized. This marginalization lacks a closed-form and so the likelihood $p(y \mid \theta)$ is implicit–it cannot be evaluated nor efficiently sampled. As a baseline we draw N samples from the joint and use the NMC estimator to compute entropies as:

$$H(\theta) = -\int p(\theta) \log(p(\theta)) dx \approx -\frac{1}{N} \sum_i \log \left(\frac{1}{N-1} \sum_{j \neq i} p(\theta_i | \psi_j) \right) \tag{15}$$

Fig. 5 summarizes the proposed estimators and runtime. As the theory suggests our moment matching estimators provide substantial speedup, and we observe accurate estimates with $\hat{I}_{\text{m}+p}$ in this model. We emphasize that $\hat{I}_{\text{marg}}$ and $\hat{I}_{\text{post}}$ are infeasible due to the need to estimate model entropies in this implicit likelihood model. We instead augment these methods with the NMC estimator (e.g. Eqn. (15)), which is referred to as *variational NMC* in the literature [7]. However, the finite sample bias of NMC violates expected bound properties for few samples–see Fig. 5 (center). We include these estimators to highlight the difficulty of estimating MI in implicit likelihood models and to emphasize their practical limitations.

6.3 SIR Epidemic Model

The SIR model describes the time-evolution of infection in a fixed population [10, 1]. At each time t the population is divided into three components: *susceptible* $S(t)$, *infected* $I(t)$, and *recovered* $R(t)$ according to the time-series,

$$S(t + \Delta_t) = S(t) - \Delta I(t) \tag{16}$$

$$I(t + \Delta_t) = I(t) + \Delta I(t) - \Delta R(t) \tag{17}$$

$$R(t + \Delta_t) = R(t) + \Delta R(t) \tag{18}$$

At each time the change in infected $\Delta I(t)$ and recovered $\Delta R(t)$ are Binomially distributed,

$$\Delta I(t) \sim \text{Binomial}\left(S(t), \frac{\beta I(t)}{N}\right), \quad \Delta R(t) \sim \text{Binomial}(I(t), \gamma)$$

with unknown random parameters $\beta, \gamma \sim \text{Uniform}(0, 0.5)$. Our simulations use a fixed discrete time interval $\Delta_t = 0.01$ with a population $N = 50$ and boundary conditions $S(t = 0) = N - 1$, $I(t = 0) = 1$, and $R(t = 0) = 0$. See Fig. 7 for an example of the SIR simulation.

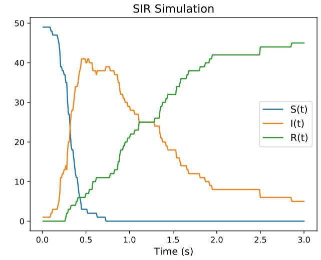


Figure 7: SIR model simulation for $\beta = 0.14, \gamma = 0.01$.

Sequential Design We select a time $t > 0$ with maximal information about the parameters β, γ as in, $\text{argmax}_t\ I(\{\beta, \gamma\}; \{S(t), I(t)\})$. We ignore R in the MI quantity since it is deterministic via: $R(t) = N - I(t) - S(t)$. After choosing a time t^* we observe $S(t^*) = s$, $I(t^*) = \iota$, and $R(t^*) = r$. In stage K of sequential (greedy) design we condition on $K - 1$ previously-chosen times $t_1^*, \ldots, t_{K-1}^*$ and their resulting observations $\{s_1^{K-1}, \iota_1^{K-1}\}$, denoted by the "history" set $\mathcal{H}_{K-1}$. The K^{th} time is chosen to maximize,

$$t_K^* = \underset{t>0}{\text{argmax}}\ I(\{\beta, \gamma\}; \{S(t), I(t)\} \mid \mathcal{H}_K)\,. \tag{19}$$

Optimizing Eqn. (19) is complicated since the SIR lacks an explicit likelihood $p(S(t), I(t), R(t) \mid \beta, \gamma)$–it is defined implicitly through simulation of Eqns. (16)-(18). Existing design approaches to sequential design in this model [11] rely on LFIRE [20] estimates of the ratio $\frac{p(S,I,R|\beta,\gamma)}{p(S,I,R)}$ for calculating MI in Eqn. (19).

Fast and accurate variational estimates. We compare our moment-matched $\hat{I}_{m+\ell}$ MI estimator to the LFIRE estimator. For sequential design we use the implementation of Kleinegesse et al. (2021) which estimates MI based on importance weighted expectations of the LFIRE ratio estimator. Fig. 6 shows that our moment matching estimates achieve several orders of magnitude speedup (left) with comparable estimates and reduced variance (left-center). We note that our estimates are based on the same samples as those used in LFIRE. In sequential greedy design we observe comparable time-point selections across iterations as compared to that of Kleinegesse et al. (center-right and right). Note that Kleinegesse et al. report only 4 design stages since the code is prohibitively slow for further designs. Using our estimator it is possible to conduct many more design iterations in a fraction of the time.

7 Discussion

In this paper, we introduce moment matching for computing optimal variational distributions in the exponential family which substantially reduces the computation time compared to previous methods. For the Gaussian case, the result simplifies for all three variational methods, $\hat{I}_{\text{marg}}$, $\hat{I}_{\text{post}}$, and $\hat{I}_{\text{m}+p}$, to be moment matching the same joint Gaussian distribution. We demonstrate the substantial computational speed up, relative accuracy, and wide use-case of $\hat{I}_{\text{m}+p}$. For future work, we would like to explore conditions to classify when the moment matched solution is a global minimum as well as generalizing the method to other exponential family distributions besides the Gaussian case.

References

[1] L. J. Allen. An introduction to stochastic epidemic models. In *Mathematical epidemiology*, pages 81–130. Springer, 2008.

[2] D. Barber and F. Agakov. The IM algorithm: a variational approach to information maximization. *NIPS*, 16:201, 2004.

[3] J. M. Bernardo. Expected Information as Expected Utility. *Ann. Stat.*, 7(3):686–690, May 1979.

[4] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. Springer New York, 2006.

[5] D. Blackwell. Comparison of experiments. In J. Neyman, editor, *2nd BSMSP*, pages 93–102, Berkeley, CA, August 1950. UC Berkeley.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006. ISBN 0471241954.

[7] A. Foster, M. Jankowiak, E. Bingham, P. Horsfall, Y. W. Teh, T. Rainforth, and N. Goodman. Variational bayesian optimal experimental design. In *Advances in Neural Information Processing Systems 32*, pages 14036–14047. 2019.

[8] A. Foster, M. Jankowiak, M. O'Meara, Y. W. Teh, and T. Rainforth. A unified stochastic gradient approach to designing bayesian-optimal experiments. In *International Conference on Artificial Intelligence and Statistics*, pages 2959–2969. PMLR, 2020.

[9] M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck. On entropy approximation for gaussian mixture random vectors. In *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 181–188. IEEE, 2008.

[10] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[11] S. Kleinegesse, C. Drovandi, and M. U. Gutmann. Sequential bayesian experimental design for implicit models via mutual information. *Bayesian Analysis*, 16(3):773–802, 2021.

[12] A. Krause and C. Guestrin. Optimal Nonmyopic Value of Information in Graphical Models – Efficient Algorithms And Theoretical Limits. In *IJCAI*, pages 1339–1345, July 2005.

[13] D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, December 1956. ISSN 0003-4851.

[14] D. J. MacKay, D. J. Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[15] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[16] J. Pacheco and J. Fisher III. Variational information planning for sequential decision making. In *AISTATS*, pages 2028–2036, 2019.

[17] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

[18] T. Rainforth, R. Cornish, H. Yang, and A. Warrington. On nesting monte carlo estimators. In *International Conference on Machine Learning*, pages 4264—4273, 2018.

[19] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[20] O. Thomas, R. Dutta, J. Corander, S. Kaski, and M. U. Gutmann. Likelihood-free inference by ratio estimation. *Bayesian Analysis*, 17(1):1–31, 2022.

[21] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[22] J. L. Williams. *Information Theoretic Sensor Management*. PhD thesis, MIT, Cambridge, MA, USA, 2007.

[23] S. Zheng, J. Pacheco, and J. Fisher. A robust approach to sequential information theoretic planning. In *International Conference on Machine Learning*, pages 5936–5944, 2018.

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] Will be included in Apendix.
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] Will be included in Apendix
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (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? [N/A]
   (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? [No]
   (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]

A Appendix

Optionally include extra information (complete proofs, additional experiments and plots) in the appendix. This section will often be part of the supplemental material.