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

Caleb Dahlke
University of Arizona

Sue Zheng
Analog Devices

Jason Pacheco
University of Arizona

Abstract

Computing mutual information (MI) of random variables lacks a closed-form in nontrivial models. Variational MI approximations are widely used as flexible estimators for this purpose, but computing them typically requires solving a costly nonconvex optimization. We prove that a widely used class of variational MI estimators can be solved via moment matching operations in place of the numerical optimization methods that are typically required. We show that the same moment matching solution yields variational estimates for so-called "implicit" models that lack a closed form likelihood function. Furthermore, we demonstrate that this moment matching solution has multiple orders of magnitude computational speedup compared to the standard optimization-based solutions. We show that theoretical results are supported by numerical evaluation in fully parameterized Gaussian mixture models and a generalized linear model with implicit likelihood due to nuisance variables. We also demonstrate on the implicit simulation-based likelihood SIR epidemiology model, where we avoid costly likelihood free inference and observe many orders of magnitude speedup.

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) (Lindley, 1956; Blackwell, 1950; Bernardo, 1979) MI is used to measure the amount of information provided by each hypothesized experiment. Additionally, MI plays a key role in measuring and optimizing the amount of information that can be transmitted along noisy communication channels (Cover and Thomas, 2006; MacKay et al., 2003). MI is essential in optimizing sensor configurations (Krause and Guestrin, 2005), sensor selection (Williams, 2007), active learning (Settles, 2012), representation learning (Tishby et al., 2000), 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 (Zheng et al., 2018; Rainforth et al., 2018). 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 ratio estimation (LFIRE) (Thomas et al., 2022), which can be slow due to repeated fitting of generalized linear models (GLMs) in an inner-loop (Kleinegesse et al., 2021).

Recent variational approaches provide an appealing alternative to MI estimation by recasting the calculation as an optimization problem (Poole et al., 2019). Such methods provide convenient bounds (Barber and Agakov, 2004) and approximations that apply even in the setting of implicit likelihood models (Foster et al., 2019). These approaches have proven successful in a range of sequential decision making tasks (Pacheco and Fisher III, 2019; Foster et al., 2020). Yet, despite their computational benefits, computing such estimators can still be prohibitive due to the underlying nonconvex optimization. In this work we will demonstrate that such estimators can be calculated efficiently for a class of variational approximations in the exponential family.

Contributions We provide an overview of our primary contributions below:

- We prove conditions on exponential family variational approximations where MI estimates can be solved via fast moment matching projection. The moment matching projection can be solved more efficiently than standard gradient-based optimization approaches.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

- For a Gaussian variational approximation we show a single moment matching projection of the joint is sufficient to optimize three variational MI estimates: an upper-bound, lower-bound, and a non-bound (implicit likelihood) approximation.
- We characterize bias of the empirical estimators when variational MI quantities cannot be computed analytically.
- We demonstrate that our approach flexibly adapts to a variety of models including Gaussian Mixture Models (Huber et al., 2008), a generalized linear model "Extrapolation Experiment", and the SIR epidemiology model. The latter involves a simulation-based implicit likelihood.

The focus of this work is to provide fast calculation of existing estimators. We do not provide any claims on improving accuracy of said estimators. In all experiments we demonstrate orders of magnitude speedup over existing gradient- and simulation-based estimation techniques.

2 COMPUTING & 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) (Cover and Thomas, 2006; MacKay et al., 2003):

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

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. 8.3. 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.

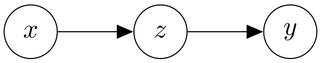


Figure 1: **Implicit Likelihood via Nuisance Variables** Likelihood $p(y \mid x)$ marginalizes z.

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 (Zheng et al., 2018; Rainforth et al., 2018). 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. 8.3.

3 VARIATIONAL MI ESTIMATION

Variational MI estimators (Poole et al., 2019) 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.

Entropy Notation. We use several notations for entropy. $H_p(X)$ is the entropy w.r.t. $p(x)$. When the distribution is clear from context use the shorthand $H(X)$. Cross-entropy between distributions p and q is denoted $H_p(q)$ and $H_p(q(X))$ when the random variable must be explicit.

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 (Barber and Agakov, 2004),

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

which we call the *variational posterior lower bound*. Observe that calculation of the lower bound I_{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*,

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

Observe that evaluation of the upper bound I_{marg} requires evaluation of the conditional entropy $H(X \mid Y)$ under the model p. For this reason, both bounds (I_{post} and I_{marg}) apply only when the model entropy terms can be calculated or ignored–typically true for BOED (Pacheco and Fisher III, 2019; Foster et al., 2019, 2020; Barber and Agakov, 2004).

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 (Foster et al., 2019):

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

where the variational distributions are $q_m(x)$ (marginal) and $q_p(x \mid y)$ (posterior). Reversing the entropy terms yields an analogous estimator: $I_{m+\ell} \equiv H_p(q_m(Y)) - H_p(q_\ell(Y \mid X))$ Both estimators avoid evaluation of model probabilities, and thus are useful for implicit likelihood models. We focus on I_{m+p} for consistency, but note that our results in Sec. 4 apply equally to $I_{m+\ell}$, 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 MI ESTIMATORS

In general, computing the optimal variational estimators (e.g. I_{marg}, I_{post}, and $I_{\text{m+p}}$) 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. For the special case of 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 optimally bounded) variational distributions for all three estimators. 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 distribution $q(x)$ is a member of the exponential family if the PDF / PMF is of the following form,

$$q(x) = h(x) \exp\left[\eta^T T(x) - A(\eta)\right]. \quad (6)$$

where η are the *natural parameters*, $h(x)$ is the *base measure*, $T(x)$ are the *sufficient statistics*, and $A(\eta)$ is the *log-partition function*. In addition to the natural parameters η each exponential family has an alternate set of *mean parameters* μ, defined as the expected sufficient statistics: $\mu = \mathbb{E}_q[T(x)]$. 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 $p(x)$ and exponential family $q(x)$ whose support includes that of p the minimum Kullback-Leibler projection:*

$$q^* = \underset{q}{\text{argmin}} \ \text{KL}(p \,\|\, q)$$

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

The interested reader can consult the texts Bishop (2006); Murphy (2012) for a proof of Lemma 4.1 and more details on the exponential family. 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, I_{marg}, we minimize the marginal cross-entropy,

$$I(X;Y) \leq \min_{q_m} H_p(q_m(X)) - \underbrace{H_p(X \mid Y)}_{\text{const.}}, \quad (7)$$

where the conditional entropy $H_p(X \mid Y)$ is constant w.r.t. $q_m(x)$ and can be ignored during optimization. If $q_m(x)$ is in the exponential family, then the minimization is found by moment matching as stated next.

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

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

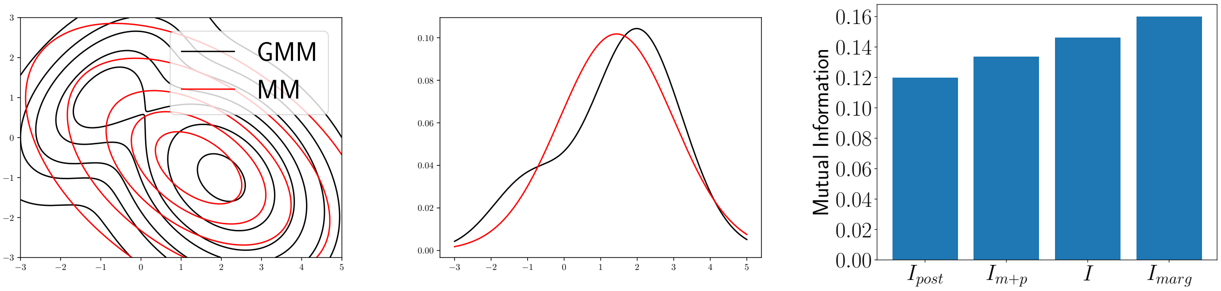
Theorem 4.2 is straightforward and is a direct result the moment matching property in Lemma 4.1. We included it as a separate statement for later results.

4.3 Variational Posterior (lower bound)

To optimize the variational posterior lower bound, I_{post}, we minimize the conditional entropy,

$$I(X;Y) \geq \underbrace{H_p(X)}_{\text{const.}} - \min_{q_p} H_p(q_p(X \mid Y)) \quad (8)$$

The $H_p(X)$ term is constant in q_p and can be ignored for optimization. The optimum of Eqn. (8) is occurs at the condition specified in the following Lemma.



(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 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 I_{marg}, I_{post}, and $I_{\text{m}+p}$ are all plotted. Notice that $I_{\text{post}} \leq I_{\text{m}+p} \approx I \leq I_{\text{marg}}$, this ordering is discussed in Lemma 6.1.

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

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

The equation 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 I_{post} via the following Lemma.

Theorem 4.4. *Let $q(x,y)$ be in the exponential family with sufficient statistics, $T(x,y) = [\tau(x), \tau(y), \tau(x,y)]^T$ where $\tau(x)$ are the sufficient statistics dependent only on x, $\tau(y)$ only on y, and $\tau(x,y)$ on both. Further, let the posterior expected statistics be a linear combination of marginal statistics as in,*

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

where $\tau_i(y)$ is the i^{th} component of $\tau(y)$ and $g_i(\eta)$ are functions of only the parameter η. Then, the optimal variational distribution, q_p, for I_{post} is defined by joint moment matching: $\mathbb{E}_{p(x,y)}[T(X,Y)] = \mathbb{E}_{q(x,y)}[T(X,Y)]$.

The linearity conditions of the sufficient statistics presented in Theorem 4.4 show that moment matching is equivalent to the optimality condition presented in Lemma 4.3 for maximizing the lower bound on I_{post} in Eqn. (8). Any distribution in the exponential family that satisfies these conditions can be optimized for I_{post} lower bound by simply moment matching the sufficient statistics instead of computing the optimal via gradient descent.

4.4 Variational Marginal & Posterior (approximation)

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

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

which is nonconvex in general and involves our target MI $I(X;Y)$. We take the approach proposed in Foster et al. (2019) and instead minimize the following upper bound, which is further convex in our case.

Lemma 4.5. *For any model $p(x,y)$ and distributions $q_m(x)$, $q_p(x \mid y)$, the following bound holds:*

$$|I_{m+p} - I| \leq H_p(q_m(X)) + H_p(q_p(X \mid Y)) + C$$

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 approaches (Foster et al., 2019) minimize this upper bound via (stochastic) gradient descent. The objective decomposes so that marginal and posterior variational distributions can be separately optimized:

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

The following theorem states that for appropriately chosen exponential family approximations the solution to Eqn. (13) can be solved via moment matching.

Theorem 4.6. *Moment Matching = Optimization*
Let $q_m(x)$ and $q(x, y)$ be exponential family distributions. Further, let $q(x, y)$ satisfy the linear conditional expectations property in Eqn. (11).

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

Then, moment matching the joint $q(x, y)$ and marginal $q_m(x)$

$$\mathbb{E}_{p(x,y)}[T(X,Y)] = \mathbb{E}_{q(x,y)}[T(X,Y)]$$
$$\mathbb{E}_{p(x)}[T(X)] = \mathbb{E}_{q(x)}[T(X)]$$

yield optimal $q_p(x \mid y) \propto q(x, y)$ and $q_m(x)$ that minimize the bound on I_{m+p} in Lemma 4.5.

The proof of Theorem 4.6 follows immediately from Theorem 4.2 and Theorem 4.4.

4.5 Variational Gaussian Distribution

Results of the preceding sections show the general conditions of exponential families such that moment matching yields optimal variational MI approximations. The scenario further simplifies when $q(x, y)$ is chosen as a joint Gaussian distribution. In this case all three variational MI approximations (I_{marg}, I_{post}, $I_{\text{m}+p}$) can be determined by moment matching the joint distribution. We begin by stating a simple property of Gaussians, namely moment matching the joint implies moment matching of the marginals.

Lemma 4.7. *Let $q(x, y) = \mathcal{N}(m, \Sigma)$ be a Gaussian and $q_m(x) = \int q(x, y)dy$, then*

$$\mathbb{E}_{p(x,y)}[T(x,y)] = \mathbb{E}_{q(x,y)}[T(x,y)] \tag{14}$$

implies

$$\mathbb{E}_{p(x)}[T(x)] = \mathbb{E}_{q_m(x)}[T(x)] \tag{15}$$

By Lemma 4.7, we see that moment matching a joint Gaussian will be the optimal variational distribution for I_{marg} as discussed in Theorem 4.2. For I_{post}, we show that the conditions of Lemma 4.3 are satisfied.

Theorem 4.8. *Let $q(x, y) = \mathcal{N}(m, \Sigma)$ be a multivariate Gaussian distribution. Then $q_p(x \mid y)$ is also Gaussian and satisfies conditions of both Lemma 4.3 and Theorem 4.4. Furthermore, the optimal I_{post} is obtained by joint Gaussian moment matching conditions,*

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

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

The proof of Theorem 4.8 shows that moment matching a joint Gaussian will optimize I_{post}. Finally, we show that the general result of moment matching the marginal and joint to optimize $I_{\text{m}+p}$ in Theorem 4.6 is satisfied by simply moment matching a joint, $q(x, y) = \mathcal{N}(\mu, \Sigma)$.

Corollary 4.9. *Let $q(x, y) = \mathcal{N}(\mu, \Sigma)$, $q_m(x) = \int q(x, y)dy$, and $q_p(x \mid y) = \frac{q(x,y)}{q(y)}$. Then by Theorem 4.6 moment matching the joint $q(x, y)$ yields optimal Gaussian q_p and q_m that minimize the bound on I_{m+p} in Lemma 4.5.*

In summary of these results, for $q(x, y)$ Gaussian moment matching the joint distribution yields optimal Gaussian approximations for all three of the variational MI methods I_{marg}, I_{post}, and $I_{\text{m}+p}$.

5 EMPIRICAL ESTIMATORS OF VARIATIONAL MI

The results in Sec. 4 assume that moments of the target distribution are available in closed-form. However, we generally need to estimate these moments via samples. In this section we discuss the details of computing variational MI using empirical estimators. Given samples $\{x_i\}_{i=1}^N$ we define the Monte Carlo entropy estimators as,

$$H_p(p(x)) \approx -\frac{1}{N}\sum_i^N \log(p(x_i)) = \hat{H}_p(p(x)) \tag{16}$$

$$H_p(q(x)) \approx -\frac{1}{N}\sum_i^N \log(q(x_i)) = \hat{H}_p(q(x)) \tag{17}$$

The method of moment matching to find optimal variational estimators requires computing the expectation of sufficient statistics, which must also be approximated. The notation for the analytic moment matched variational distribution, $q(x, y)$, will be defined by the parameters

$$\mu^* = \mathbb{E}_{p(x,y)}[T(x,y)] \tag{18}$$

whereas the empirical moment matched variational distribution, $\hat{q}(x, y)$, is defined by the parameters

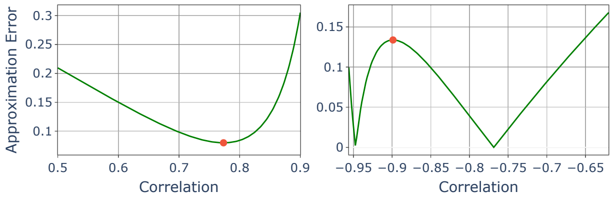
$$\hat{\mu} = \frac{1}{N}\sum_i^N T(x_i, y_i) \tag{19}$$

where $x_i, y_i \sim p(x, y)$. The use of a plug-in variational distribution $\hat{q}$ based on empirical moments induces bias in the estimated cross-entropy. The following Lemma provides an ordering of each of these estimators according to their biases.

Lemma 5.1. *Entropy Ordering*
Let $x \in \mathbb{R}^d$, $p(x)$ be an arbitrary distribution, and $q(x) = \mathcal{N}(\mu^*, \Sigma^*)$ where μ^*, Σ^* analytically moment matched to $p(x)$. Furthermore, let $\hat{q}(x) = \mathcal{N}(\hat{\mu}, \hat{\Sigma})$ be the empirically moment matched variational distribution. Then,

$$H_p(p(x)) \overset{\text{(a)}}{=} \mathbb{E}\left[\hat{H}_p(p(x))\right] \overset{\text{(b)}}{\leq} \mathbb{E}\left[\hat{H}_p(\hat{q}(x))\right] \overset{\text{(c)}}{\leq} H_p(q(x))$$

To summarize the implication of Lemma 5.1 it shows that the plug-in empirical cross-entropy estimator is negatively



(a) Global Minimum (b) Local Maximum

Figure 3: **Moment Matched Optimum** (a) The moment matching solution (red) is the global 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.

biased w.r.t. to the true cross-entropy (c) (by Jensen's inequality). This estimator remains an upper bound on the corresponding empirical estimate of the target entropy (b) which itself is unbiased (a) by the law of large numbers (though impractical to compute unless $p(x)$ is known). The bias induced by (c) can result in bound violations at small sample sizes, as observed in Fig. 4. Note that when moment matching conditions are satisfied the cross-entropy is equivalent to the entropy of the exponential family, as established in the next Lemma.

Lemma 5.2. *Analytic Entropy*
Let $p(x)$ be any distribution and $q(x)$ be in the exponential family with constant base measure, $h(x) = C$, which is analytically moment matched to $p(x)$ and $\hat{q}(x)$ is empirically moment matched, then

$$H_p(q(x)) = H_q(q(x)) \quad \hat{H}_p(\hat{q}(x)) = H_{\hat{q}}(\hat{q}(x)) \tag{20}$$

The above result is convenient in practice as the exponential family often has closed form solutions for their entropy. This allows us to avoid the empirical expectation w.r.t. p in the cross-entropy.

6 ACCURACY OF VARIATIONAL MI ESTIMATORS

Previous sections have addressed the efficient computation of variational MI approximations. However, it remains to consider the accuracy of each estimator, and under what conditions one may be preferable over another. This section provides some preliminary analysis that addresses this issue. We begin by showing that the MI estimators satisfy a total ordering.

Lemma 6.1. *For any $q_m(x)$ and $q_p(x \mid y)$,*

$$I_{post} \leq I_{m+p} \approx I \leq I_{marg}$$

where $I \equiv I(X;Y)$ is the true MI and $(I_{marg}, I_{post}, I_{m+p})$ are computed from q_m and q_p.

The above Lemma states that approximation $I_{\text{m}+p}$ is *never the least accurate* out of all three methods. When $I_{\text{m}+p}$ is an over approximation it is a tighter upper bound than I_{marg}. The converse holds if it is an under approximation (it is a tighter than I_{post}). Ideally we would be able to determine which of the three estimators is most accurate for any instance. The following Lemma provides these conditions.

Lemma 6.2. *For any $q_m(x)$ and $q_p(x \mid y)$ the following statements hold:*

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

Unfortunately, the conditions in Lemma 6.2 cannot be evaluated in practice as they involve KL w.r.t. the true posterior. However, these inequalities offer some insight into how one may determine the most accurate estimator in an online fashion. 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 $I_{\text{m}+p}$ is the best approximation to use.

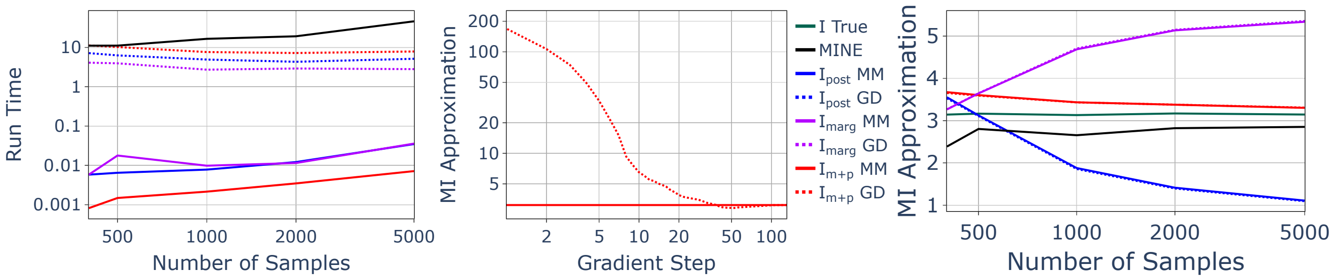
We conclude with an observation on the optimality of the moment matching estimator for $I_{\text{m}+p}$. Recall that optimizing $I_{\text{m}+p}$ exactly requires minimizing the nonconvex error $|I(X;Y) - I_{\text{m}+p}|$ in Eqn. (12). We instead show that moment matching minimizes a convex upper bound on this error in Lemma 4.5. It is then natural to ask under what conditions is this bound tight. Consider a two-component Gaussian mixture with moments,

$$\mu_x = \mathbb{E}_p\left[X\right] \qquad \mu_y = \mathbb{E}_p\left[Y\right]$$
$$\sigma_x^2 = \mathbb{E}_p\left[XX^T\right] - \mu_x^2 \qquad \sigma_y^2 = \mathbb{E}_p\left[YY^T\right] - \mu_y^2$$
$$\rho = \left(\mathbb{E}_p\left[XY^T\right] - \mu_x\mu_y\right)/\sigma_x\sigma_y.$$

Fig. 3 shows the relationship between the moment matching solution of Gaussian $I_{\text{m}+p}$ with the above parameters and the absolute error. We shows this for two mixture models where we plot the absolute error $|I(X;Y) - I_{\text{m}+p}|$ as a function of only the correlation parameter ρ. In Fig. 3(a) the moment matching solution yields the global minimum error. However in Fig. 3(b) a local maximum is found. We further observe in this latter case that there exists two values of ρ that yield the exact MI value (zero error). In conclusion, there are cases where the moment matching solution of $I_{\text{m}+p}$ may yield a global minimum but other cases where it yields a local maximum despite the existence of good solutions. Exploring these conditions in more detail is a topic of future work.

7 PREVIOUS WORK

In this paper, we focused on the variational methods, I_{marg}, I_{post} (Barber and Agakov, 2004), and $I_{\text{m}+p}$ (Foster et al.,



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

Figure 4: **High-dimensional GMM.** (a). As per our theoretical analysis, moment matching for all variational methods achieves orders of magnitude computational speedup. Mine is the slowest of all the methods and computation time grows the fastest with respect to the samples. (b) We see the gradient descent approaches the same value as moment matching but takes around 200 gradient steps to converge which is where the moment matching solution achieves its drastic computational speed-up. (c) The MI approximation vs samples illustrates the same ordering of $I_{\text{post}} \leq I \approx I_{\text{m}+p} \leq I_{\text{marg}}$ except for low samples where the bias discussed in Lemma 5.1 adds noticeable error. "True MI" is calculated via Monte Carlo estimation with exact evaluation of the model probabilities.

2019). The focus of each of these methods was computation speedups for computing the optimal distribution. We also briefly discussed the Nested Monte Carlo estimator in Sec. 2 and some of the challenges it faced. Foster et al. (2019) provides an in-depth analysis of convergence rate and run time for the variational estimators discussed in this paper. For an alternative implicit likelihood approximator, we also consider the likelihood-free inference by ratio (LFIRE) used by Kleinegesse et al. (2021) as a baseline for comparison purposes. Poole et al. (2019) provides an investigation and comparison of a variety of variational MI estimators–see also Foster et al. (2020). Additional estimators are based on flexible artificial neural network approximators, such as Mutual Information Neural Estimator (MINE) (Belghazi et al., 2018). This approach optimizes the dual representation of KL divergence introduced by Donsker and Varadhan (1983), providing a lower bound. In most cases such density-free expressions cannot be computed analytically and empirical estimators are typically biased. McAllester and Stratos (2020) give an analysis of the fundamental bias of such high-confidence distribution-free MI lower bounds.

8 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. 8.1). We then evaluate two implicit likelihood models: one arising from the non-closed-form marginalization of nuisance variables in a GLM (Sec. 8.2), and the other is a simulation-based SIR epidemiology model (Sec. 8.3). 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.

We also compute MINE (Belghazi et al., 2018) in Sec. 8.1 and Sec. 8.3 as well as LFIRE (Thomas et al., 2022) in Sec. 8.3 for a comparison of our methods to show the computation speed and accuracy trade off against these other common MI estimates.

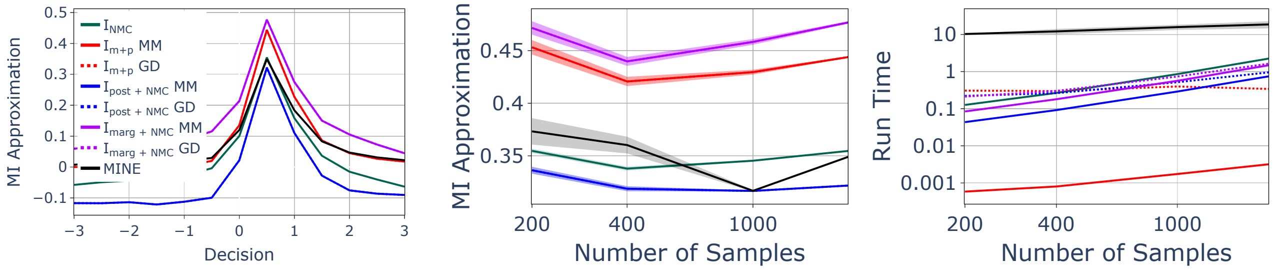
8.1 Multivariate Gaussian Mixture Model

GMMs are pervasive in statistics due to their universal approximation properties, yet calculating MI for a GMM is notoriously challenging (Huber et al., 2008). In this section we extend the two-dimensional example (Fig. 2) to high-dimensional GMMs. We simulate a 5 component GMM, $p(x, y) = \sum_i^5 \omega_i \mathcal{N}(m_i, \Sigma_i)$, with $\sum \omega_i = 1$ and dimensions $X \in \mathbb{R}^{60}$ and $Y \in \mathbb{R}^5$. We use this setting to demonstrate efficient MI estimation even in high-dimensional distributions.

Fig. 4 shows substantial speedups in runtime (left) for all methods as compared to gradient optimization. Notice that GD takes approximately 200 gradient steps to converge for $5,000$ samples whereas moment matching found this solution immediately, independent of any gradient steps (center). As per our theoretical results we find that $I_{\text{m}+p}$ lies between the MI upper bound I_{marg} and lower bound I_{post} for large enough of a sample size, however for low samples the finite sample bias of the empirical estimators discussed in Lemma 5.1 changes the ordering of the estimators. (right). For MINE, we utilize code provided by Kleinegesse and Gutmann (2020) to see that it performs similarly to $I_{\text{m}+p}$ in accuracy however we note at least four orders of computation speed up by our method with the computation time of MINE growing the fastest of all the methods with respect to the number of samples. In this example, we see that $I_{\text{m}+p}$ is the most accurate estimator however this does not hold in general (Appendix D).

8.2 Extrapolation

We adapt the following experiment from Foster et al. (2019) intended to evaluate the implicit likelihood MI estimator



(a) Total MI

(b) MI Convergence

(c) Run Time

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 2000 samples is plotted with $d = .5$ being the maximum decision. Note the negative bias resulting from NMC estimate of the entropy terms. (b) The convergence rate versus samples is plotted at maximum decision ($d = .5$). (c) We again observe a substantial computational savings from moment matching. This is seen for smaller values of samples with $\hat{I}_{marg}$ and $\hat{I}_{post}$ until around 500 samples when NMC becomes the dominating computation time. $\hat{I}_{m+p}$ however maintains the fast computation speedup as it does not need any NMC evaluations. MINE is orders of magnitude slower than all the other methods for all ranges of sample size taken.

$I_{\text{m}+p}$ (or $I_{m+\ell}$). Labeled information, y, from a subset of the design space is used to predict labels at a location x that can't be directly observed. The model is as follows,

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

where $X_\theta = (1, -\frac{1}{2})$ 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 directly. As a baseline we draw N samples from the joint and use the NMC estimator to compute entropies as:

$$H(\theta) = -\int p(\theta) \log(p(\theta)) dx$$
$$\approx -\frac{1}{N} \sum_i \log \left(\frac{1}{N-1} \sum_{j \neq i} p(\theta_i | \psi_j) \right) \tag{21}$$

Fig. 5 summarizes the proposed estimators and runtime. We emphasize that I_{marg} and I_{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. (21)), and refer to them as variational marginal (posterior) plus NMC, denoted as $I_{marg+NMC}$ ($I_{post+NMC}$). It is important to note that this is not the same as *variational NMC* in the literature (Foster et al., 2019) which is a consistent variational estimator. For $I_{post+NMC}$ and $I_{marg+NMC}$, 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. As the theory suggests our moment matching estimators provide substantial speedup. We notice that for small number of samples $I_{post+NMC}$ and $I_{marg+NMC}$ are computed substantially faster with moment matching compared to their gradient descent counterparts. Once larger samples are taken, the cost of the NMC term dominates the computation cost. In $I_{\text{m}+p}$ we notice that the multiple orders of magnitude computation time speed up is maintained across all number of samples as it can avoid the NMC term and benefits drastically from the speed-up of moment matching instead of gradient descent. Again, for MINE we have some increased accuracy at higher number of samples however see that computation time is one to two order of magnitude slower than I_{NMC}, $I_{post+NMC}$ and $I_{marg+NMC}$ while $I_{\text{m}+p}$ is upwards of 5 orders of magnitude faster.

8.3 SIR Epidemiology Model

The SIR model describes the time-evolution of infection in a fixed population (Kermack and McKendrick, 1927; Allen, 2008). 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,

$$S(t + \Delta_t) = S(t) - \Delta I(t) \tag{22}$$
$$I(t + \Delta_t) = I(t) + \Delta I(t) - \Delta R(t) \tag{23}$$
$$R(t + \Delta_t) = R(t) + \Delta R(t) \tag{24}$$

At each time the distribution of change in infected is $\Delta I(t) \sim \text{Binomial}(S(t), \frac{\beta I(t)}{N})$ and for recovered is $\Delta R(t) \sim \text{Binomial}(I(t), \gamma)$, 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.

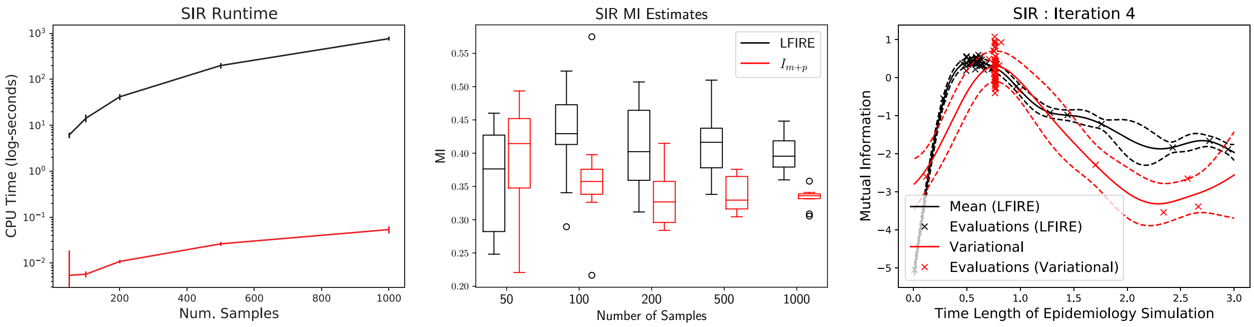


Figure 6: **SIR Sequetial Design.** The top plots show benchmark time and utility evaluation between LFIRE and the Variational estimator for a fixed design ($d = 1.0s$) over 10 runs each at a range of sample sizes. The variational estimator is orders of magnitude more efficient (*left*) and shows lower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

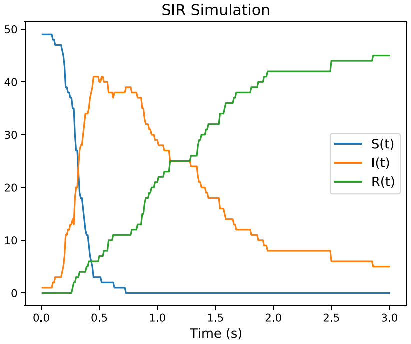


Figure 7: **SIR Simulation.** An example of a single SIR model simulation (*far-right*) for $\beta = 0.14, \gamma = 0.01$.

Sequential Design We select a time $t > 0$ with the 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,

$$t_K^* = \underset{t>0}{\text{argmax}} \; I(\{\beta,\gamma\}; \{S(t), I(t)\} \mid \mathcal{H}_K). \qquad (25)$$

Optimizing Eqn. (25) 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. (22)-(24). Existing design approaches to sequential design in this model (Kleinegesse et al., 2021) rely on LFIRE (Thomas et al., 2022) estimates of the ratio $\frac{p(S,I,R|\beta,\gamma)}{p(S,I,R)}$ for MI in Eqn. (25). Furthermore, we have a continuous decision domain $t \in [0, 3]$ of the duration of the epidemiology simulation. To find the maximal decision, Bayesian optimization is performed using a Gaussian process where a limited number of evaluations are used to find the maximum. We compare our moment matched $I_{\text{m}+p}$ 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.

Fast and accurate variational estimates. Fig. 6 shows that our moment matching estimates achieve several orders of magnitude speedup (*left*). We only use 4 design stages to match Kleinegesse et al. since the code is prohibitively slow for further designs with LFIRE. Using our estimator it is possible to conduct many more design iterations in a fraction of the time. Furthermore, we notice a significant reduction in variance (*center*) at each sample. The evaluation points chosen for Bayesian optimization are chosen using different estimators (LFIRE and $I_{\text{m}+p}$) hence the differnce in evaluation locations in the two methods (*right*). The goal however is not to match MI estimation across the entire continuous design domain, but instead to approximate the decision with maximal MI. The two methods have comparable maximum decisions with LFIRE choosing around $t \approx .5$ and $I_{\text{m}+p}$ around $t \approx .75$. We note that the true maximal decision point is unknown and compare only the relative accuracy of each method to each other and the computation speed up achieved by moment matched $I_{\text{m}+p}$.

9 DISCUSSION

In this paper, we prove conditions that allow for fast moment matching projections to obtain optimal variational distributions in the exponential family. This substantially reduces the computation time compared to previous methods. For the Gaussian case, we show the result simplifies for all three variational methods, I_{marg}, I_{post}, and $I_{\text{m}+p}$, to be moment matching the same joint Gaussian distribution. We demonstrate the substantial computational speed up, relative accuracy, and wide use-case of $I_{\text{m}+p}$. For future work, we would like to explore other exponential family distributions besides the Gaussian case that satisfy the necessary conditions.

References

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

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

M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm. Mine: Mutual information neural estimation. 2018.

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

C. M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

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

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.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

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.

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.

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.

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.

S. Kleinegesse and M. U. Gutmann. Bayesian experimental design for implicit models by mutual information neural estimation. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5316–5326. PMLR, 13–18 Jul 2020.

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.

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.

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.

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

D. McAllester and K. Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR, 2020.

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

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

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.

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

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

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.

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

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

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.

Appendix : Fast Variational Estimation of MI for Implicit and Explicit Models

A PROOFS FOR RESULTS IN SECTION 4

The sections below provide all proofs for results in the main text.

A.1 Section 4.2 Proofs

Theorem 4.2. *Let $q_m(x)$ be in the exponential family with statistics $T(x)$, then for any $p(x)$, the optimal 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,

$$\underset{q_m}{\operatorname{argmin}}\ H_p(q_m(X)) = \underset{q_m}{\operatorname{argmin}}\ H_p(q_m(X)) - H_p(X) = \underset{q_m}{\operatorname{argmin}}\ \mathrm{KL}(p(X) \,\|\, q(X)) \tag{26}$$

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

A.2 Section 4.3 Proofs

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

$$\mathbb{E}_{p(y)}\left[\mathbb{E}_{q_p(x|y)}\left[T(X,Y)\right]\right] = \mathbb{E}_{p(x,y)}\left[T(X,Y)\right] \tag{27}$$

Proof. The goal is to minimize $H_p(q_p(X|Y))$ where $q_p(x|y)$ is generated from $q(x,y;\eta)$ in the exponential family. We will find the minimizing parameters of this distributions. We appeal to the property of exponential families that $\frac{\partial}{\partial\eta}A(\eta) = \mathbb{E}_{q(x,y)}\left[T(x,y)\right]$

$$\frac{\partial}{\partial\eta}\left(H_p(q_p(X|Y))\right) = -\frac{\partial}{\partial\eta}\int p(x,y)\log\left(q_p(X|Y)\right) = -\int p(x,y)\frac{\partial}{\partial\eta}\log\left(\frac{q(X,Y;\eta)}{q(Y;\eta)}\right) \tag{28}$$

$$= -\int p(x,y)\frac{\partial}{\partial\eta}\left(\log(h(x,y)) + \eta^T T(x,y) - A(\eta) - \log\left(q(y;\eta)\right)\right)dxdy \tag{29}$$

$$= -\int p(x,y)\left(T(x,y) - \frac{\partial}{\partial\eta}A(\eta) - \frac{\partial}{\partial\eta}\log\left(q(y;\eta)\right)\right)dxdy \tag{30}$$

$$\begin{aligned} = &-\mathbb{E}_{p(x,y)}\left[T(x,y)\right] + \mathbb{E}_{q(x,y)}\left[T(x,y)\right] + \\ &\int p(x,y)\frac{1}{q(y;\eta)}\frac{\partial}{\partial\eta}\left(\int h(x',y)\exp\left(\eta^T T(x',y) - A(\eta)\right)dx'\right)dxdy \end{aligned} \tag{31}$$

$$\begin{aligned} = &-\mathbb{E}_{p(x,y)}\left[T(x,y)\right] + \mathbb{E}_{q(x,y)}\left[T(x,y)\right] + \\ &\int p(x,y)\frac{1}{q(y;\eta)}\left(\int q(x',y;\eta)\left(T(x',y) - \frac{\partial}{\partial\eta}A(\eta)\right)dx'\right)dxdy \end{aligned} \tag{32}$$

$$\begin{aligned} = &-\mathbb{E}_{p(x,y)}\left[T(x,y)\right] + \mathbb{E}_{q(x,y)}\left[T(x,y)\right] + \\ &\int p(x,y)\left(\int q(x'|y)\left(T(x',y) - \mathbb{E}_{q(x,y)}\left[T(x,y)\right]dx'\right)dxdy\right) \end{aligned} \tag{33}$$

$$= -\mathbb{E}_{p(x,y)}\left[T(x,y)\right] + \mathbb{E}_{p(y)}\left[\mathbb{E}_{q_p(x|y)}\left[T(x,y)\right]\right] \tag{34}$$

The zero derivative yields the stationary condition $\mathbb{E}_{p(x,y)}\left[T(x,y)\right] = \mathbb{E}_{p(y)}\left[\mathbb{E}_{q_p(x|y)}\left[T(x,y)\right]\right]$. It now remains to show that the objective is convex in η. Expanding the form of $H_p(q(X \mid Y))$ we have the objective,

$$\min_{\eta} -\mathbb{E}_p\left[\log(h(X,Y)) + \eta^T T(X,Y) - A(\eta) - \log(q(Y;\eta))\right] \tag{35}$$

The term $\eta^T T(X,Y)$ is linear in η. Convexity of $A(\eta)$ in η is a standard property of the exponential family, however we will show a constructive proof that $A(\eta) + \log(q(y;\eta))$ is convex using Hölder's inequality. Let $\eta = \lambda\eta_1 + (1-\lambda)\eta_2$ where $\lambda \in [0,1]$ and η_1, η_2 in the convex set of valid exponential family parameters of q then:

$$A(\eta) + \log(q(y;\eta)) = A(\eta) + \log\left(\int h(x,y)\exp(\eta^T T(x,y) - A(\eta))\,dx\right) \tag{36}$$

$$= A(\eta) + \log\left(\exp(-A(\eta))\int h(x,y)\exp(\eta^T T(x,y))\,dx\right) \tag{37}$$

$$= \log\left(\int h(x,y)\exp(\eta^T T(x,y))\,dx\right) \tag{38}$$

$$= \log\left(\int (h(x,y)\exp(\eta_1^T T(x,y)))^{\lambda}\,(h(x,y)\exp(\eta_2^T T(x,y)))^{(1-\lambda)}\,dx\right) \tag{39}$$

$$\leq \lambda\log\left(\int h(x,y)\exp(\eta_1^T T(x,y))\,dx\right) + (1-\lambda)\log\left(\int h(x,y)\exp(\eta_2^T T(x,y))\,dx\right) \tag{40}$$

$$= \lambda(A(\eta_1) + \log q(y;\eta_1)) + (1-\lambda)(A(\eta_2) + \log q(y;\eta_2)) \tag{41}$$

Thus convexity holds in η and the stationary conditions $\mathbb{E}_{p(x,y)}\left[T(x,y)\right] = \mathbb{E}_{p(y)}\left[\mathbb{E}_{q_p(x|y)}\left[T(x,y)\right]\right]$ are globally optimal. □

Theorem 4.4. *Let $q(x,y)$ be in the exponential family with sufficient statistics, $T(x,y) = [\tau(x), \tau(y), \tau(x,y)]^T$ where $\tau(x)$ are the sufficient statistics dependent only on x, $\tau(y)$ only on y, and $\tau(x,y)$ on both. Further, let the posterior expected statistics be a linear combination of marginal statistics as in,*

$$\mathbb{E}_{q_p(x|y)}\left[T(X,Y)\right] = \sum_i^k g_i(\eta)\tau_i(Y) \tag{42}$$

where $\tau_i(y)$ is the i^{th} component of $\tau(y)$ and $g_i(\eta)$ are functions of only the parameter η. Then, the optimal variational distribution, q_p, for I_{post} is defined by joint moment matching: $\mathbb{E}_{p(x,y)}[T(X,Y)] = \mathbb{E}_{q(x,y)}[T(X,Y)]$.

Proof. From Lemma 4.3, we know that $\mathbb{E}_{p(x,y)}\left[T(x,y)\right] = \mathbb{E}_{p(y)}\left[\mathbb{E}_{q_p(x|y)}\left[T(x,y)\right]\right]$ is the optimality condition. Let us now show that the condition in Eqn. (11) implies that joint moment matching satisfies the optimality condition of Eqn. (9)

$$\mathbb{E}_{p(x,y)}\left[T(x,y)\right] = \mathbb{E}_{q(x,y)}\left[T(x,y)\right] \tag{43}$$

$$= E_{q(y)}\left[\mathbb{E}_{q_p(x|y)}\left[T(x,y)\right]\right] = \mathbb{E}_{q(y)}\left[\sum_i^k g_i(\eta)\tau_i(y)\right] \tag{44}$$

$$= \sum_i^k g_i(\eta)\mathbb{E}_{q(y)}\left[\tau_i(y)\right] = \sum_i^k g_i(\eta)\mathbb{E}_{p(y)}\left[\tau_i(y)\right] \tag{45}$$

$$= \mathbb{E}_{p(y)}\left[\sum_i^k g_i(\eta)T_i(y)\right] = \mathbb{E}_{p(y)}\left[\mathbb{E}_{q_p(x|y)}\left[T(x,y)\right]\right] \tag{46}$$

So with Lemma 4.3 and the assumption of the posterior expected statistics being a linear combination of joint statistics (Eqn. (11)) results in $\mathbb{E}_{p(x,y)}[T(X,Y)] = \mathbb{E}_{q(x,y)}[T(X,Y)]$ being the optimal conditions. □

A.3 Section 4.4 Proofs

Lemma 4.5. *For any model $p(x,y)$ and distributions $q_m(x)$, $q_p(x \mid y)$, the following bound holds:*

$$|I_{m+p} - I| \leq H_p(q_m(X)) + H_p(q_p(X \mid Y)) + C$$

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.

Proof. We reproduce the proof from Foster et al. (2019).

$$\begin{aligned}
|I_{\mathrm{m}+p}(X,Y) - I(X,Y)| &= |H_p(q_m(X)) - H_p(q_p(X|Y)) - H_p(p(X)) + H_p(p(X|Y))| && (47)\\
&= |-H_p(p(X)) + H_p(q_m(X)) + H_p(p(X|Y)) - H_p(q_p(X|Y))| && (48)\\
&= |\mathrm{KL}(p(X) \,\|\, q_m(X)) - \mathrm{KL}(p(X|Y) \,\|\, q_p(X|Y))| && (49)\\
&\leq |\mathrm{KL}(p(X) \,\|\, q_m(X))| + |\mathrm{KL}(p(X|Y) \,\|\, q_p(X|Y))| && (50)\\
&= -\, H_p(p(X)) + H_p(q_m(X)) - H_p(p(X|Y)) + H_p(q_p(X|Y)) && (51)\\
&= H_p(q_m(X)) + H_p(q_p(X|Y)) + C && (52)
\end{aligned}$$

Where $C = -H_p(p(X)) - H_p(p(X \mid Y))$. □

Theorem 4.6. *Let $q_m(x)$ and $q(x,y)$ be exponential family distributions. Further, let $q(x,y)$ satisfy the linear conditional expectations property in Eqn. (11).*

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

Then, moment matching the joint $q(x,y)$ and marginal $q_m(x)$

$$\mathbb{E}_{p(x,y)}[T(X,Y)] = \mathbb{E}_{q(x,y)}[T(X,Y)]$$
$$\mathbb{E}_{p(x)}[T(X)] = \mathbb{E}_{q(x)}[T(X)]$$

yield optimal $q_p(x \mid y) \propto q(x,y)$ and $q_m(x)$ that minimize the bound on I_{m+p} in Lemma 4.5.

Proof. We break this down into the two cases of Theorem 4.2 and Theorem 4.4. Notice that the variational distributions $q_m(x)$ and $q_p(x \mid y)$ need not share a common joint $q(x,y)$. So, let us use different natural parameters, η_1 and η_2, for each (i.e. $q_m(x) = q(x;\eta_1)$ and $q_p(x|y) = q(x|y;\eta_2)$). We optimize the bound in Lemma 4.5 with respect to both natural parameters, beginning with η_1:

$$\frac{\partial}{\partial \eta_1}\left(-\mathbb{E}_{p(x,y)}\left[\log q(x;\eta_1) + \log q(x \mid y;\eta_2)\right] + C\right) = -\frac{\partial}{\partial \eta_1} E_{p(x,y)}\left[\log q(x;\eta_1)\right]$$

This is exactly the condition in Theorem 4.2 which we know is solved by moment matching the marginal. Likewise, for η_2:

$$\frac{\partial}{\partial \eta_2}\left(-\mathbb{E}_{p(x,y)}\left[\log q(x;\eta_1) + \log q(x \mid y;\eta_2)\right] + C\right) = -\frac{\partial}{\partial \eta_2} E_{p(x,y)}\left[\log q(x \mid y;\eta_2)\right]$$

The above is the start of the proof for Lemma 4.3 in Eqn. (28) and along with Eqn. (11) in Theorem 4.4, we get that moment matching the joint finds the optimal q_p. Therefore, the optimization of Lemma 4.5 simply reduces to moment matching the marginal and the joint. □

A.4 Section 4.5 Proofs

Lemma 4.7. *Let $q(x,y) = \mathcal{N}(m,\Sigma)$ be a Gaussian and $q_m(x) = \int q(x,y)dy$, then*

$$\mathbb{E}_{p(x,y)}[T(x,y)] = \mathbb{E}_{q(x,y)}[T(x,y)] \tag{53}$$

implies

$$\mathbb{E}_{p(x)}[T(x)] = \mathbb{E}_{q_m(x)}[T(x)] \tag{54}$$

Proof. We can break down the moment matching operation of the joint into its components, $\tau(x)$, $\tau(x,y)$ and $\tau(y)$.

$$\mathbb{E}_{p(x,y)}\left[T(x,y)\right] = \mathbb{E}_{q(x,y)}\left[T(x,y)\right] \quad \Longrightarrow \quad \begin{aligned}
\mathbb{E}_{p(x)}\left[\tau(x)\right] &= \mathbb{E}_{q(x)}\left[\tau(x)\right] && (55)\\
\mathbb{E}_{p(x,y)}\left[\tau(x,y)\right] &= \mathbb{E}_{q(x,y)}\left[\tau(x,y)\right] && (56)\\
\mathbb{E}_{p(y)}\left[\tau(y)\right] &= \mathbb{E}_{q(y)}\left[\tau(y)\right] && (57)
\end{aligned}$$

The Gaussian has the property that the sufficient statistics of a marginal are $\tau(x)$, the components of the joint sufficient statistics, $T(x)$, dependent on only x. We see that Eqn. (55) is exactly the condition for marginal moment matching, so

$$\mathbb{E}_{p(x,y)}[T(x,y)] = \mathbb{E}_{q(x,y)}[T(x,y)] \quad \Longrightarrow \quad \mathbb{E}_{p(x)}[T(x)] = \mathbb{E}_{q_m(x)}[T(x)] \tag{58}$$

□

Theorem 4.8. *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 Theorem 4.4. Furthermore, the optimal I_{post} is obtained by joint Gaussian moment matching conditions,*

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

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

Proof. We first see the conditions of Lemma 4.3 are satisfied by the setup of the problem as

$$q_p(x \mid y) = \frac{q(x, y)}{q(y)}$$

It suffices to verify the assumption of the posterior expected statistics being a linear combination of joint statistics (Eqn. (11)) is satisfied. Recall the sufficient statistics of a multivariate Gaussian

$$T(x, y) = \begin{bmatrix} x \\ y \\ \text{vec}(xx^T) \\ \text{vec}(xy^T) \\ \text{vec}(yy^T) \end{bmatrix}$$

In this case $\tau_1(y) = y$ and $\tau_2(y) = \text{vec}(yy^T)$. We now verify that the expected value under $q(x|y)$ of each term in the sufficient statistic is a linear function of $\tau_1(y)$ and $\tau_2(y)$

1. x

$$\mathbb{E}_{q(x|y)}[x] = \mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$$

2. y

$$\mathbb{E}_{q(x|y)}[y] = y$$

3. xx^T

$$\begin{aligned}\mathbb{E}_{q(x|y)}\left[xx^T\right] =& \Sigma_{x|y} + \mu_{x|y}\mu_{x|y}^T \\ =& \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}\Sigma_{xy}^T + (\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y))(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y))^T \\ =& \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}\Sigma_{xy}^T + \mu_x\mu_x^T + \dots \\ & \mu_x(y - \mu_y^T)\Sigma_{yy}^{-1}\Sigma_{xy}^T + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)\mu_x + \dots \\ & \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)(y - \mu_y)^T\Sigma_{yy}^{-1}\Sigma_{xy}^T\end{aligned}$$

4. xy^T

$$\begin{aligned}\mathbb{E}_{q(x|y)}\left[xy^T\right] =& (\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y))y^T \\ =& (\mu_x - \Sigma_{xy}\Sigma_{yy}^{-1}\mu_y)y^T + \Sigma_{xy}\Sigma_{yy}^{-1}yy^T\end{aligned}$$

5. yy^T

$$\mathbb{E}_{q(x|y)}\left[yy^T\right] = yy^T$$

So the statistics are linear functions of $\tau_1(y) = y$ and $\tau_2(y) = yy^T$ so $p(x|y)$ satisfies the conditions of Theorem 4.4 and moment matching the joint $q(x, y) = \mathcal{N}(m, \Sigma)$ yields the optimal I_{post}. □

Corollary 4.9. *Let $q(x, y) = \mathcal{N}(\mu, \Sigma)$, $q_m(x) = \int q(x, y)dy$, and $q_p(x \mid y) = \frac{q(x,y)}{q(y)}$. Then by Theorem 4.6 moment matching the joint $q(x, y)$ yields optimal Gaussian q_p and q_m that minimize the bound on I_{m+p} in Lemma 4.5.*

Proof. This Corollary holds by Theorem 4.8 to show $q(x, y)$ satisfies the linear conditional expectation property, so moment matching the joint yields the optimal $q_p(x \mid y)$. To see that moment matching the joint implies moment matching the marginal, note that the sufficient statistics of the joint Guassian are

$$T(x, y) = \left[x, y, \text{vec}(xx^T), \text{vec}(xy^T), \text{vec}(yy^T)\right]^T$$

and the sufficient statistics of a marginal distribution are

$$T(x) = \left[x, \text{vec}(xx^T)\right]^T$$

which are simply the first and third sufficient statistic from the joint. Therefore moment matching the joint Gaussian trivially moment matches the marginal and Theorem 4.6 applies. □

B PROOFS FOR RESULTS IN SECTION 5

Lemma 5.1. *Entropy Ordering*
Let $x \in \mathbb{R}^d$, $p(x)$ be an arbitrary distribution, and $q(x) = \mathcal{N}(\mu^*, \Sigma^*)$ where μ^*, Σ^* analytically moment matched to $p(x)$. Furthermore, let $\hat{q}(x) = \mathcal{N}(\hat{\mu}, \hat{\Sigma})$ be the empirically moment matched variational distribution. Then,

$$H_p(p(x)) \overset{(a)}{=} \mathbb{E}_p\left[\hat{H}_p(p(x))\right] \overset{(b)}{\leq} \mathbb{E}_p\left[\hat{H}_p(\hat{q}(x))\right] \overset{(c)}{\leq} H_p(q(x))$$

Proof. (a) $\hat{H}_p(p(x))$ is an empirical mean estimator so by law of large numbers is an unbiased estimator of $H_p(p(x))$. (b) Gibbs' Inequality. (c) Uses the fact that $\log(\det(\cdot))$ is a concave function

$$\begin{aligned}
&\mathbb{E}_p\left[\hat{H}_p(\hat{q}(x))\right] = \mathbb{E}_p\left[H_{\hat{q}}(\hat{q}(x))\right] \\
=&\mathbb{E}_p\left[\frac{1}{2}\log(\det(2\pi e\hat{\Sigma}))\right] \\
=&\frac{1}{2}\left(d\log(2\pi e) + \mathbb{E}_p\left[\log(\det(\hat{\Sigma}))\right]\right) \\
\leq&\frac{1}{2}\left(d\log(2\pi e) + \log(\det(\mathbb{E}_p\left[\hat{\Sigma}\right]))\right) \\
=&\frac{1}{2}\left(d\log(2\pi e) + \log(\det(\Sigma^*))\right) = H_p(q(x))
\end{aligned}$$

We see that $\hat{H}_p(\hat{q}(x))$ is a biased estimator from below of $H_p(q(x))$. □

Lemma 5.2. *Analytic Entropy*
Let $p(x)$ be any distribution and $q(x)$ be in the exponential family with constant base measure, $h(x) = C$, which is analytically moment matched to $p(x)$ and $\hat{q}(x)$ is empirically moment matched, then

$$H_p(q(x)) = H_q(q(x)) \quad \hat{H}_p(\hat{q}(x)) = H_{\hat{q}}(\hat{q}(x)) \tag{59}$$

Proof.

- $H_p(q(x)) = H_q(q(x))$

$$\begin{aligned}
H_p(q(x)) &= -\mathbb{E}_{p(x)}\left[\log(q(x))\right] = -\mathbb{E}_{p(x)}\left[\log(h(x)) + \eta^T T(x) - A(\eta)\right] \\
&= -\left(C + \eta^T \mathbb{E}_{p(x)}\left[T(x)\right] - A(\eta)\right) = -\left(C + \eta^T \mathbb{E}_{q(x)}\left[T(x)\right] - A(\eta)\right) \\
&= -\mathbb{E}_{q(x)}\left[\log(h(x)) + \eta^T T(x) - A(\eta)\right] = -\mathbb{E}_{q(x)}\left[\log(q(x))\right] = H_q(q(x))
\end{aligned}$$

- $\hat{H}_p(\hat{q}(x)) = H_{\hat{q}}(\hat{q}(x))$

$$\begin{aligned}
\hat{H}_p(\hat{q}(x)) &= -\frac{1}{N}\sum_i^N \log(\hat{q}(x_i)) = -\frac{1}{N}\sum_i^N \left(\log(h(x_i)) + \hat{\eta}^T T(x_i) - A(\hat{\eta})\right) \\
&= -\left(C + \hat{\eta}^T \frac{1}{N}\sum_i^N (T(x_i)) - A(\hat{\eta})\right) = -\left(C + \hat{\eta}^T \mathbb{E}_{\hat{q}(x)}\left[T(x)\right] - A(\hat{\eta})\right) \\
&= -\mathbb{E}_{\hat{q}(x)}\left[\log(h(x)) + \hat{\eta}^T T(x) - A(\hat{\eta})\right] = -\mathbb{E}_{\hat{q}(x)}\left[\log(\hat{q}(x))\right] = H_{\hat{q}}(q\widehat{(x)})
\end{aligned}$$

note that $x_i \sim p(x)$.

□

C PROOFS FOR RESULTS IN SECTION 6

Lemma 6.1. *For any $q_m(x)$ and $q_p(x \mid y)$,*

$$I_{post} \leq I_{m+p} \approx I \leq I_{marg}$$

where $I \equiv I(X;Y)$ is the true MI and (I_{marg}, I_{post}, I_{m+p}) are computed from q_m and q_p.

Proof. We prove the lower bound on $I_{\mathrm{m}+p}$ first and then the upper
1) $I_{\mathrm{post}} \leq I_{\mathrm{m}+p}$

$$\begin{aligned} I_{\mathrm{post}} =& H_p(p(x)) - H_p(q(x \mid y)) \leq H_p(p(x)) - H_p(q(x \mid y)) + \mathrm{KL}(p(x) \,\|\, q(x)) \\ =& H_p(p(x)) - H_p(q(x \mid y)) - H_p(p(x)) + H_p(q(x)) = H_p(q(x)) - H_p(q(x \mid y)) = I_{\mathrm{m}+p} \end{aligned}$$

2) $I_{\mathrm{marg}} \geq I_{\mathrm{m}+p}$

$$\begin{aligned} I_{\mathrm{marg}} =& H_p(q(x)) - H_p(p(x \mid y)) \geq H_p(q(x)) - H_p(p(x \mid y)) - \mathrm{KL}(p(x \mid y) \,\|\, q(x \mid y)) \\ =& H_p(q(x)) - H_p(p(x \mid y)) + H_p(p(x \mid y)) - H_p(q(x \mid y)) = H_p(q(x)) - H_p(q(x \mid y)) = I_{\mathrm{m}+p} \end{aligned}$$

In both of these, we simply appeal to $\mathrm{KL}(p \,\|\, q) \geq 0$ and $\mathrm{KL}(p \,\|\, q) = -H_p(p) + H_p(q)$. □

Lemma 6.2. *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 I_{m+p} has lower error than 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 I_{m+p} has lower error than I_{marg}*

Proof. We will look at the error of each of the statements

1. $|I_{\mathrm{m}+p} - I| \leq |I_{\mathrm{post}} - I|$

$$\begin{aligned} |I_{\mathrm{m}+p} - I| \leq& |I_{\mathrm{post}} - I| \\ |\mathrm{KL}(p(x) \,\|\, q_m(x)) - \mathrm{KL}(p(x \mid y) \,\|\, q_p(x \mid y))| \leq& |\mathrm{KL}(p(x \mid y) \,\|\, q_p(x \mid y))| \\ \mathrm{KL}(p(x) \,\|\, q_m(x)) \leq& 2\mathrm{KL}(p(x \mid y) \,\|\, q_p(x \mid y)) \\ \frac{1}{2}\mathrm{KL}(p(x) \,\|\, q_m(x)) \leq& \mathrm{KL}(p(x \mid y) \,\|\, q_p(x \mid y)) \end{aligned}$$

2. $|I_{\mathrm{m}+p} - I| \leq |I_{\mathrm{marg}} - I|$

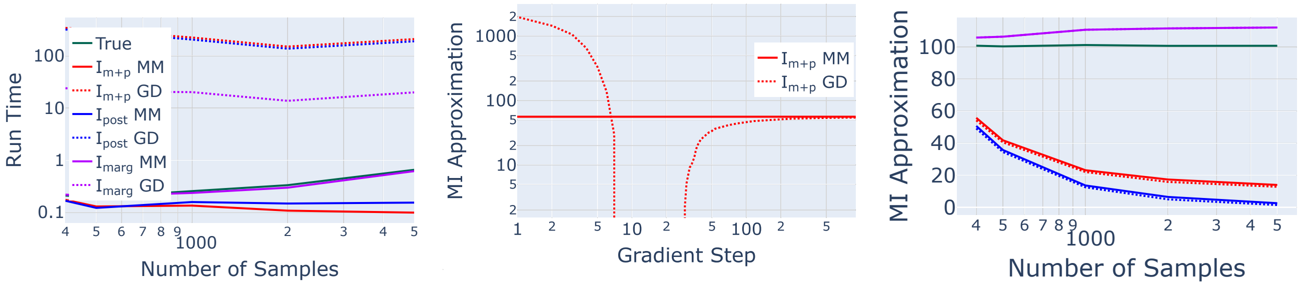
$$\begin{aligned} |I_{\mathrm{m}+p} - I| \leq& |I_{\mathrm{marg}} - I| \\ |\mathrm{KL}(p(x) \,\|\, q_m(x)) - \mathrm{KL}(p(x \mid y) \,\|\, q_p(x \mid y))| \leq& |\mathrm{KL}(p(x) \,\|\, q_m(x))| \\ \mathrm{KL}(p(x \mid y) \,\|\, q_p(x \mid y)) \leq& 2\mathrm{KL}(p(x) \,\|\, q_m(x)) \\ \frac{1}{2}\mathrm{KL}(p(x \mid y) \,\|\, q_p(x \mid y)) \leq& \mathrm{KL}(p(x) \,\|\, q_m(x)) \end{aligned}$$

□

D ADDITIONAL EXPERIMENTS FOR GAUSSIAN MIXTURE MODEL

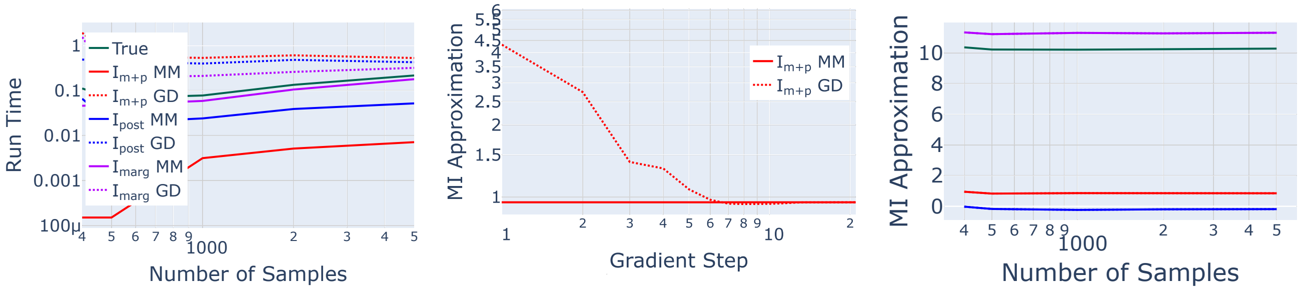
Our focus of the GMM experiment in Sec. 8.1 was to demonstrate improved computation of estimators computed via moment matching relative to standard optimization-based approaches. We demonstrated this speedup in a relatively high-dimensional setting ($X \in \mathbb{R}^{60}$, $Y \in \mathbb{R}^5$). See Fig. 4 for details. We further noted that while $I_{\mathrm{m}+p}$ yielded lowest approximation error, this conclusion does not hold in general.

To emphasize the above observation we conduct a variety of additional GMM experiments for various settings here. We conduct two alternative forms of the GMM experiment; where we increase the dimension of the scenario to $X \in \mathbb{R}^{100}$ and



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

Figure 8: **High-dimensional** 10 **Component GMM.** (a). We again see multiple orders of magnitude time save by moment matching I_{m+p}, I_{marg}, and I_{post} instead of gradient descent. (b) Again, the gradient descent approach to I_{m+p} converges to the solution found by moment matching. (c) We see the same ordering and behavior of I_{m+p}, I_{marg}, and I_{post} in comparison to the true mutual information as we did in Sec. 8.1.



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

Figure 9: **Low-dimensional** 100 **Component GMM.** (a). We again see a computational speed up by moment matching over gradient descent by about an order of magnitude. It is less of a speed up in this case because gradient descent can handel the lower dimensional case better than It did the large. (b) Gradient descent takes a more direct approach to the MM solution in this case due to lower dimensionality. (c) We again see the same ordering of I_{m+p}, I_{marg}, and I_{post} in comparison to the true mutual information as we did in Sec. 8.1, this time however, the 400 samples is more than adequate fully approximate the integrals compared to the high dimensional case which needed a few thousand to adequately approximate the integrals.

$Y \in \mathbb{R}^{200}$ (using 10 components in the GMM), and another experiment where we increase the number of components to 100 done in a low dimensional setting ($X \in \mathbb{R}^5$ and $Y \in \mathbb{R}^{10}$).

For the high dimensional 10-component GMM we again observe multiple orders of magnitude speed up by moment matching, relative to gradient optimnization–see Fig. 8(a). Yet we find in Fig. 8(c) that I_{post} and I_{m+p} have much higher finite sample bias (for less than a 1000 samples) as discussed from Lemma 5.1 and the I_{marg} is the most accurate estimator overall.

For the 100-component case we again see a computational improvement from moment matching–see Fig. 9(a). As would be expected, the time savings is less dramatic due to the lower dimensional setting, resulting in faster convergence of gradient optimization (Fig. 9(b)). Yet, We notice almost no finite sample bias in any of the estimators, suggesting that bias is heavily driven by dimension of the model. Fig. 9(c) shows that in this scenario that I_{marg} is again the most accurate estimator.

In practice, we do not know which will be the most accurate estimator among $I_{\text{marg}}, I_{\text{post}}, I_{m+p}$. However, Lemma 6.2 provides conditions for identifying the most accurate estimate. As we note in the postscript to Lemma 6.2, these conditions cannot practically be computed in general. It is a topic of ongoing work to determine surrogate conditions that approximate those of Lemma 6.2 and thus enable selection of the most accurate estimate in an online manner.