Estimators of Entropy and Information via Inference in Probabilistic Models

Feras A. Saad
MIT

Marco Cusumano-Towner
MIT

Vikash K. Mansinghka
MIT

Abstract

Estimating information-theoretic quantities such as entropy and mutual information is central to many problems in statistics and machine learning, but challenging in high dimensions. This paper presents *estimators of entropy via inference* (EEVI), which deliver upper and lower bounds on many information quantities for arbitrary variables in a probabilistic generative model. These estimators use importance sampling with proposal distribution families that include amortized variational inference and sequential Monte Carlo, which can be tailored to the target model and used to squeeze true information values with high accuracy. We present several theoretical properties of EEVI and demonstrate scalability and efficacy on two problems from the medical domain: (i) in an expert system for diagnosing liver disorders, we rank medical tests according to how informative they are about latent diseases, given a pattern of observed symptoms and patient attributes; and (ii) in a differential equation model of carbohydrate metabolism, we find optimal times to take blood glucose measurements that maximize information about a diabetic patient's insulin sensitivity, given their meal and medication schedule.

1 INTRODUCTION

This paper studies the fundamental problem of estimating the Shannon entropy $H(Y) \coloneqq -\mathbb{E}[\log p(Y)]$ of a random element Y, in situations where its marginal distribution involves an intractable multidimensional integral over a known joint probability distribution:

$$p(y) = \int_{\mathcal{X}} p(x, y)\,\mathrm{d}x \qquad (y \in \mathcal{Y}). \tag{1}$$

In (1), the term $p(x, y)$ refers to a probabilistic generative model that can be sampled from and whose joint density can be computed pointwise, as is common in a broad class of probabilistic systems that includes Bayesian networks (Pearl, 1988), deep generative models (Kingma and Welling, 2019), and generative probabilistic programs (Wingate et al., 2011). In this setting, a key challenge is that the expression

$$H(Y) = -\int_{\mathcal{Y}} \log\left[\int_{\mathcal{X}} p(x, y)\,\mathrm{d}x\right] p(y)\,\mathrm{d}y \tag{2}$$

contains an intractable integral inside the logarithm, which rules out the unbiased simple Monte Carlo estimator $-1/n\sum_{i=1}^{n}\log p(Y_i)$ (for $Y_i \sim p(y)$, $1 \le i \le n$).

To address these challenges, we develop a new class of *estimators of entropy via inference* (EEVI) that return interval estimates of doubly intractable entropies as in (2). EEVI uses auxiliary-variable importance sampling constructs similar to those from pseudo-marginal methods (Andrieu and Roberts, 2009) to first compute unbiased estimates of the intractable quantities $p(y)$ and $1/p(y)$ for the inner integral. Under the log transform, these estimates become lower and upper bounds of $\log p(y)$, which are then embedded in a simple Monte Carlo estimator for the outer integral to form an interval estimate of $H(Y)$. In the limit of computation, the interval width can be driven to zero, squeezing the true entropy value at a rate that depends on the quality of the importance sampling proposal.

Our contribution is a family of entropy estimators that

(C1) Apply to arbitrary random variables in any probability distribution that can be sampled from and whose full joint density is tractable; no marginals or conditionals need to be tractable.

(C2) Guarantee upper and lower bounds in expectation, which can be composed (Fig. 1) to squeeze many information quantities, e.g., (11)–(15).

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

(C3) Leverage a broad family of proposal distributions that includes both variational and Monte Carlo inference for increasing accuracy as a function of computational effort.

The rest of the paper is organized as follows: Sec. 2 gives an overview of EEVI and explains how interval estimates of entropy can be composed to form interval estimates of several other information quantities, such as conditional mutual information and interaction information. Sec. 3 presents theoretical properties of importance sampling-based estimators of log marginal probabilities of the form given in (1), and gives examples of inference-based variational and Monte Carlo auxiliary-variable proposals to deliver accurate upper and lower bounds. Sec. 4 illustrates the scalability and efficacy of EEVI for two optimal design tasks in a probabilistic expert system for diagnosing liver disorders and a dynamic model of carbohydrate metabolism in diabetic patients. Sec. 5 discusses related work.

2 OVERVIEW OF EEVI

Suppose that $p(z_1, \dots, z_d)$ is a d-dimensional probability density (with respect to an appropriate σ-finite measure) such that it is possible to sample $(Z_1, \dots, Z_d) \sim p(z_1, \dots, z_d)$ and evaluate density values pointwise. Let $A \subset \{1, \dots, d\}$ be a subset of indexes and let $Y := \{Z_i, i \in A\}$ and $X := \{Z_i, i \notin A\}$ be the corresponding partition of variables in Z. We aim to estimate the marginal entropy $H(Y)$ as defined in (2), where $\mathcal{Y}$ and $\mathcal{X}$ are the sets in which Y and X take values, respectively. As the partition A is arbitrary, neither the marginal densities $p(x)$ and $p(y)$ nor conditional densities $p(x \mid y)$ and $p(y \mid x)$ are necessarily tractable, posing a key challenge for estimating $H(Y)$.

Suppose momentarily that we can compute two measurable real functions $w, w' : \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ such that for some random variables U, U' taking values in a set $\mathcal{U}$ and all $y \in \mathcal{Y}$ except for a p-measure zero set, we have

$$\mathbb{E}\left[w(U, y)\right] = p(y) \qquad \mathbb{E}\left[w'(U', y)\right] = 1/p(y). \tag{3}$$

If w and w' are nonnegative almost everywhere, then concavity of log and Jensen's inequality gives bounds

$$\mathbb{E}\left[\log w(U, y)\right] \leq \log \mathbb{E}\left[w(U, y)\right] = \log p(y), \tag{4}$$

$$\mathbb{E}\left[\log w'(U', y)\right] \leq \log \mathbb{E}\left[w'(U', y)\right] = -\log p(y), \tag{5}$$

which together imply that

$$\mathbb{E}\left[\log w(U, y)\right] \leq \log p(y) \leq \mathbb{E}\left[-\log w'(U', y)\right]. \tag{6}$$

If the real functions $y \mapsto \mathbb{E}\left[\log w(U, y)\right]$ and $y \mapsto -\mathbb{E}\left[\log w'(U', y)\right]$ defined on $\mathcal{Y}$ are themselves both measurable then monotonicity of expectation gives

$$\mathbb{E}\left[\log w(U, Y)\right] \leq \mathbb{E}\left[\log p(Y)\right] \leq \mathbb{E}\left[-\log w'(U', Y)\right] \tag{7}$$

The two expectations in (7) that squeeze $\mathbb{E}\left[\log p(Y)\right]$ can now be estimated via unbiased Monte Carlo

$$\mathcal{L}_{n,m} := \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{m} \sum_{j=1}^{m} \log w(U_{ij}, Y_i) \right], \tag{8}$$

$$\mathcal{T}_{n,m} := \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{m} \sum_{j=1}^{m} -\log w'(U'_{ij}, Y'_i) \right], \tag{9}$$

where Y_i, Y'_i are identically distributed to Y; U_{ij} identically to U; and U'_{ij} identically to U' ($i = 1, \dots, n; j = 1, \dots, m$). Letting $\check{H}_Y := -\mathcal{T}_{n,m}$ and $\hat{H}_Y := -\mathcal{L}_{n,m}$, (7) implies that the means of $\check{H}_Y$ and $\hat{H}_Y$ satisfy

$$\mathbb{E}[\check{H}_Y] \leq H(Y) \leq \mathbb{E}[\hat{H}_Y]. \tag{10}$$

If using i.i.d. samples in (8) and (9), under mild conditions the central limit theorem and (10) imply the interval estimator $[\check{H}_Y, \hat{H}_Y]$ has coverage probability

$$\Pr[\check{H}_Y \leq H(Y) \leq \hat{H}_Y] \approx \Phi\left(\sqrt{t}\check{B}/\check{\sigma}\right) \Phi\left(\sqrt{t}\hat{B}/\hat{\sigma}\right),$$

where Φ is the standard normal CDF; $t = nm$; $\check{B} := \mathbb{E}\left[-\log w'(U', Y)\right] - \mathbb{E}\left[\log p(Y)\right]$ and $\hat{B} := \mathbb{E}\left[\log p(Y)\right] - \mathbb{E}\left[\log w(U, Y)\right]$ are the biases in (10); and $\check{\sigma}$ and $\hat{\sigma}$ are the standard deviations of $\log w'(U', Y)$ and $\log w(U, Y)$. So far we have assumed access to functions w and w' that satisfy (3): Sec. 3 shows how they can be constructed via importance sampling.

2.1 Extending entropy bounds to additional information-theoretic quantities

The lower and upper bounds on entropy in (10) can be composed to bound several derived information-theoretic quantities that measure the degree of relationship between arbitrary subcollections of variables in a model, possibly conditioned on others (Fig. 1). Letting $H(A) := H(\{Z_i, i \in A\})$ for $A \subset [d]$, by adding and subtracting upper and lower bounds on $H(A)$ we can also build interval estimators of:

• *conditional entropy* (Shannon, 1948)

$$H(A_1 \mid A_2) := H(A_1 \cup A_2) - H(A_2) \tag{11}$$

• *conditional mutual information* (Shannon, 1948)

$$I(A_1 : A_2 \mid A_0) := H(A_1 \mid A_0) - H(A_1 \mid A_2, A_0) \tag{12}$$

• *conditional total correlation* (Watanabe, 1960)

$$C(\{A_i\}_{i=1}^{n} | A_0) := \sum_{i=1}^{n} H(A_i | A_0) - H(\bigcup_{i=1}^{n} A_i \mid A_0) \tag{13}$$

• *conditional interaction information* (Ting, 1962)

$$T(\{A_i\}_{i=1}^{n} \mid A_0) := \textstyle\sum_{S \subset [n]} -1^{|S|} H(\cup_{i \in S} A_i \mid A_0) \tag{14}$$

• *conditional dual correlation* (Han, 1978)

$$D(\{A_i\}_{i=1}^{n} \mid A_0) := H(\cup_{i=1}^{n} A_i \mid A_0) - \textstyle\sum_{i=1}^{n} H(A_i \mid \cup_{j=0, j \neq i}^{n} A_i) \tag{15}$$

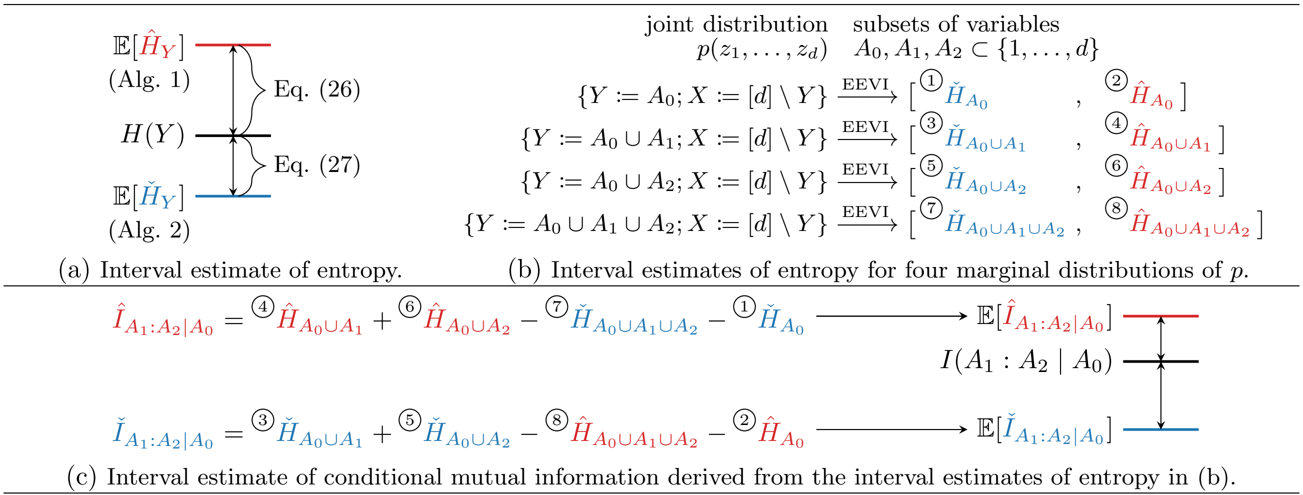


Figure 1: Composing interval estimators of entropy to obtain bounds on derived information measures.

3 SAMPLING BOUNDS ON LOG MARGINAL PROBABILITIES

Importance sampling in log space. Implementing the estimators $\hat{H}_Y, \check{H}_Y$ in (10) requires functions w, w' that satisfy (3). Our starting point is an identity from importance sampling. Let h and g be two probability densities on a common set $\mathcal{X}$ such that h is absolutely continuous with respect to g (written $h \ll g$); i.e., $\int_A g(x)\,\mathrm{d}x = 0 \implies \int_A h(x)\,\mathrm{d}x = 0$ for all measurable A. Suppose $h(x) = \tilde{h}(x)/Z_h$, $g(x) = \tilde{g}(x)/Z_g$ are only known up to normalizing constants. Then, for $X \sim g$,

$$\mathbb{E}\left[\tilde{h}(X)/\tilde{g}(X)\right] = Z_h/Z_g. \tag{16}$$

(All proofs in Appx. B.) Under log transform, the ratio in (16) is a lower bound on $\log(Z_h/Z_g)$ in expectation with a gap equal to the KL divergence from h to g:

$$\mathbb{E}\left[\log\left(\tilde{h}(X)/\tilde{g}(X)\right)\right] = \log\left(Z_h/Z_g\right) - \mathrm{D_{KL}}\left[g||h\right]. \tag{17}$$

Eq. (17) does not require $h \ll g$. However, the expectation is well-defined only if $g \ll h$ and is finite only if $\mathrm{D_{KL}}\left[g||h\right] < \infty$. Moreover, the variance

$$\begin{aligned}&\mathrm{Var}\left[\log(\tilde{h}(X)/\tilde{g}(X))\right] \\ &\quad = \mathbb{E}\left[\log^2(h(X)/g(X))\right] - (\mathrm{D_{KL}}\left[g||h\right])^2\end{aligned} \tag{18}$$

is finite only if $\mathrm{D_{KL}}\left[g||h\right] < \infty$ and $\log^2(h(X)/g(X))$ has finite expectation. Applying Markov's inequality to (16) gives a right tail bound for $\log \tilde{h}(X)/\tilde{g}(X)$:

$$\Pr\left[\tilde{h}(X)/\tilde{g}(X) \geq e^t(Z_h/Z_g)\right] \leq e^{-t} \tag{19}$$
$$\implies \Pr\left[\log(\tilde{h}(X)/\tilde{g}(X)) \geq t + \log(Z_h/Z_g)\right] \leq e^{-t},$$

for any $t > 0$. The mean absolute deviation satisfies

$$\mathbb{E}\left[\left|\log(\tilde{h}(X)/\tilde{g}(X)) - \mu\right|\right] \leq 2 + 2\,\mathrm{D_{KL}}\left[g||h\right], \tag{20}$$

where $\mu := \mathbb{E}[\log(\tilde{h}(X)/\tilde{g}(X))]$; i.e., it is upper bounded by two plus twice the bias in (17), which decreases as g more closely matches h.

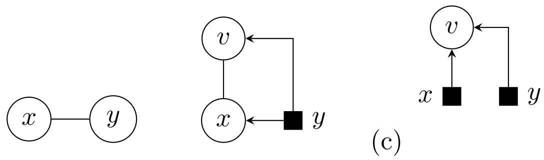
Interval estimators of entropy. Recalling the distribution $p(x, y)$ from Sec. 2, suppose that $q(x; y)$ and $q'(x; y)$ are normalized proposal densities over $\mathcal{X}$ parameterized by $\mathcal{Y}$. From (16), for fixed $y \in \mathcal{Y}$, setting $\tilde{h}(x) = p(x, y)$, $g(x) = q(x; y)$ gives an unbiased estimate of $Z_h \equiv p(y)$; and setting $h(x) = q'(x; y)$, $\tilde{g}(x) = p(x, y)$ gives an unbiased estimate of $1/Z_g \equiv 1/p(y)$:

$$\mathbb{E}\left[\frac{p(X, y)}{q(X; y)}\right] = p(y), \quad \mathbb{E}\left[\frac{q'(X'; y)}{p(X', y)}\right] = 1/p(y), \tag{21}$$

as needed for (3), by defining $w(x, y) := p(x, y)/q(x; y)$ and $w'(x, y) := q'(x; y)/p(x, y)$ and letting $X \sim q(x; y)$ be U and $X' \sim p(x \mid y)$ be U'. Then (8) yields the Monte Carlo *upper* bound $\hat{H}_Y$ in (10); and (9) yields the Monte Carlo *lower* bound $\check{H}_Y$ in (10). While sampling $X' \sim p(x \mid y)$ given a fixed value y is typically intractable under our assumptions on p, since $H(Y)$ is the expectation of random values $-\log p(Y)$ for $Y \sim p$, it suffices to use joint samples $(X', Y') \sim p(x, y)$ to obtain the lower bound, since

$$\begin{aligned}&\int_{\mathcal{X}\times\mathcal{Y}} \log\left[q'(x; y)/p(x, y)\right] p(x, y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathcal{Y}} \left[\int_{\mathcal{X}} \log\left[q'(x; y)/p(x, y)\right] p(x \mid y)\,\mathrm{d}x\right] p(y)\,\mathrm{d}y \\ &\leq \int_{\mathcal{Y}} \left[-\log p(y)\right] p(y)\,\mathrm{d}y = H(Y),\end{aligned}$$

where the second line follows from Schervish (1995, Thms. B.46, B.52) and third line from (17). Given an initial sample $(X'_1, Y) \sim p(x, y)$, an additional $m - 1$ samples $\{X'_2, \dots, X'_m\}$ from $p(x \mid Y')$ to use for $\mathcal{T}_{n,m}$ in (9) can be obtained by simulating a Markov chain initialized at X'_1 that leaves $p(x \mid Y')$ invariant, which will reduce $\mathrm{Var}\left[\mathcal{T}_{n,m}\right]$ iff $\Pr[X'_i \neq X'_1] > 0$ for some i.



(a) Target $p(x, y)$ (b) Proposal $q(v, x; y)$ (c) Aux. Proposal $r(v; x, y)$

Figure 2: Target, proposal, and auxiliary proposal distributions used for interval estimators (Algs. 1 and 2) of the entropy $H(Y) = -\mathbb{E}[\log p(Y)]$, $Y \sim p(y)$.

Algorithm 1 Monte Carlo upper bound $\hat{H}_Y$ on $H(Y)$

for $i = 1 \dots n$ **do**
  $(\widetilde{X}, Y) \sim p(x, y)$
  for $j = 1 \dots m$ **do**
    $(V, X) \sim q(v, x; Y)$
    $t_{ij} \leftarrow \log \dfrac{p(X, Y) r(V; X, Y)}{q(V, X; y)}$
return $-\sum_{i=1}^{n} \sum_{j=1}^{m} t_{ij}/nm$

Algorithm 2 Monte Carlo lower bound $\check{H}_Y$ on $H(Y)$

for $i = 1 \dots n$ **do**
  $(X'_1, Y) \sim p(x, y)$
  $(X'_{2:m}) \sim$ Markov chain targeting $p(x \mid Y)$ starting at X'_1 (optional step)
  for $j = 1 \dots m$ **do**
    $V \sim r'(v; X'_j, Y)$
    $t_{ij} \leftarrow -\log \dfrac{q'(V, X'_j; Y)}{p(X'_j, Y) r'(V; X'_j, Y)}$
return $-\sum_{i=1}^{n} \sum_{j=1}^{m} t_{ij}/nm$

3.1 Constructing accurate proposals

Estimators of normalizing constants and their inverses in direct space as in (21) can suffer from notoriously high variance, especially when using proposals that do not closely match the target (Neal, 2008). However, (19) suggests that log space estimators can be more stable and, by (17) and (20), the quality of the entropy bounds obtained via importance sampling depends on constructing proposal distributions $q(x; y)$ and $q'(x; y)$ that have small biases in expectation (over $Y \sim p(y)$):

$$\mathbb{E}[\hat{H}_Y] - H(Y) = \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left[q(x; Y) || p(x \mid Y)\right]\right], \quad (22)$$

$$H(Y) - \mathbb{E}[\check{H}_Y] = \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left[p(x \mid Y) || q'(x; Y)\right]\right]. \quad (23)$$

We next discuss two approaches to constructing accurate proposals using probabilistic inference algorithms.

Amortized variational inference. One approach to constructing accurate proposals $q(x; y)$ and $q'(x; y)$ in (21) is to use a dataset $\{(x_i, y_i)\}_{i=1}^{n}$ simulated from p to variationally train recognition networks $q_\phi(x; y)$ and $q'_\varphi(x; y)$ that each specify a family of distributions over $\mathcal{X}$, parametrized by $y \in \mathcal{Y}$ and ϕ, φ, respectively. Training q via "exclusive" amortized variational inference, as in variational autoencoders (Kingma and Welling, 2014), minimizes both the bias and an upper bound on the mean absolute deviation (MAD) of the $\check{H}_Y$ in Alg. 1. Similarly, training q' via "inclusive" amortized variational inference, as in the "sleep" phase of the wake-sleep algorithm (Hinton et al., 1995), minimizes both the bias and an upper bound on the MAD of $\hat{H}_Y$ in Alg. 2. Thus, EEVI can leverage advances in training neural networks via stochastic gradient descent to improve estimation accuracy (refer to Fig. 3).

Auxiliary-variable Monte Carlo. Another approach to constructing accurate proposals, which can be composed with variational learning (Salimans et al., 2015), is Monte Carlo methods such as annealed importance sampling (AIS; Neal, 2001) and sequential Monte Carlo (SMC; Del Moral et al., 2006) that define proposal distributions on extended state-spaces and yield state-of-the-art estimates of normalizing constants. More specifically, these proposals $q(v, x; y)$ are defined over an extended space $\mathcal{V} \times \mathcal{X}$, such that the marginal $q(x; y) = \int_{\mathcal{V}} q(v, x; y)\,\mathrm{d}v$ is an integral over all auxiliary random variables sampled by q. As the ratios in (21) can no longer be evaluated, we define a tractable "auxiliary proposal distribution" $r(v; x, y)$ over $\mathcal{V}$ parameterized by $\mathcal{X} \times \mathcal{Y}$ (Fig. 2) such that

$$w(v, x, y) \coloneqq \left[p(x, y) r(v; x, y)\right] / q(v, x; y), \quad (24)$$

$$w'(v, x, y) \coloneqq q'(v, x; y) / \left[p(x, y) r'(v; x, y)\right]. \quad (25)$$

By (16), these weight functions satisfy (3) by letting $(V, X) \sim q(v, x; y)$ serve as U and $(X', V') \sim p(x \mid y) r(v; x, y)$ as U'. From (17), the gap when lower bounding $\log p(y)$ using these extended-space weights w and w' now accounts for the accuracy of the auxiliary proposals r and r' (illustrated in Appx. A, Fig. 8):

$$\begin{aligned}\mathbb{E}[\hat{H}_Y] - H(Y) = \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left[q(x; Y) || p(x|Y)\right]\right] \quad (26)\\ + \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left[q(v \mid X; Y) || r(v; X, Y)\right]\right],\end{aligned}$$

$$\begin{aligned}H(Y) - \mathbb{E}[\check{H}_Y] = \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left[p(x \mid Y) || q'(x; Y)\right]\right] \quad (27)\\ + \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left[r(v'; X', Y) || q'(v \mid X'; Y)\right]\right].\end{aligned}$$

Algs. 1 and 2 show interval estimators $[\check{H}_Y, \hat{H}_Y]$ that implement (8) and (9) using the extended weights w and w' in (24) and (25). Proposals without auxiliary variables are a special case, where $\mathcal{V} = \{\omega\}$ is a singleton and $r(v; x, y) = \delta(v; \omega)$ (Appx. A, Algs. 7 and 8).

Example 3.1 (Sampling-Importance Resampling)**.** To fix ideas, consider a base proposal $q_0(x; y)$ that has no auxiliary variables, which may have been hand constructed or trained variationally. The proposal q_0 can be embedded in a sampling-importance-resampling (SIR) scheme that generates P variables $x_{1:P}$ i.i.d. from q_0 and a selection index k taking value i with relative probability $p(x_i, y)/q_0(x_i; y)$ (i.e., the auxiliary

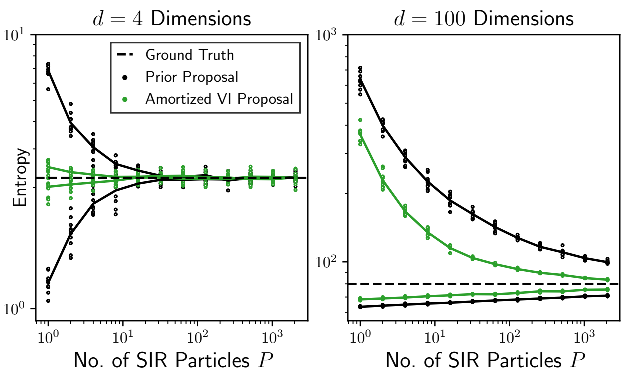


Figure 3: Lower and upper bounds on the entropy of $d/2$ dimensions Y of a d-dimensional Gaussian (X, Y) using Algs. 1 and 2 with the SIR scheme from Example 3.1. The base proposals $q_0(x; y)$ are the prior and an amortized variational approximation to the posterior that specifies a separate regression for each dimension of X given Y. While the bounds converge to the ground truth (known in closed form for Gaussians) as the number of SIR particles P increases using both proposals, the variational proposal is closer in "exclusive" and "inclusive" KL to the posterior, resulting in a higher accuracy at each P. At $d = 100$, the lower bounds exhibit much lower bias and variance as compared to the upper bounds, especially for small P.

variables $v \coloneqq (x_{1:P}, k))$, then sets $x \leftarrow x_k$:

$$q((x_{1:P}, k), x; y) = \prod_{j=1}^{P} q_0(x_j) \left[\frac{\frac{p(x_k, y)}{q_0(x_k; y)}}{\sum_{i=1}^{P} \frac{p(x_i, y)}{q_0(x_i; y)}} \right] \delta(x; x_k).$$

The task of the auxiliary proposal $r((x_{1:P}, k); x, y)$ is to infer v for an (x, y) pair as follows:

$$r((x_{1:P}, k); x, y) = \prod_{\substack{j=1 \\ j \neq k}}^{P} q_0(x_j; y) \left[\frac{1}{P} \right] \delta(x_k; x).$$

The weight (24) is then precisely the usual SIR estimate of the marginal density of y,

$$\frac{p(x, y) r(v; x, y)}{q(v, x; y)} = \frac{1}{P} \sum_{j=1}^{P} \frac{p(x_j, y)}{q_0(x_j; y)}. \tag{28}$$

By Burda et al. (2016, Thm. 1), if $p(x, y)/q_0(x; y)$ is bounded then $D_{\mathrm{KL}} \left[q(v, x; y) || p(x \mid y) r(x) \right] \to 0$ (bias in (17)) as $P \to \infty$. Refer to Fig. 3 for an illustration and Appx. A, Algs. 9 and 10 for EEVI with SIR. ◁

Example 3.2 (Generalized Sequential Monte Carlo)**.** The SIR proposal from Example 3.1 can be generalized to the setting of a sequence $\{\tilde{p}_t(x; y)\}_{t=0}^{T}$ with T intermediate (unnormalized) target densities such that $\tilde{p}_T(x, y) = p(x, y)$, which may represent partial posteriors in particle filtering for temporal models (Doucet and Johansen, 2011) or tempered distributions as in AIS (Neal, 2001) and sequential Bayesian

Algorithm 3 SMC Proposal $q(v, x; y)$

Require: Observation y
Ensure: Approximate sample x from $p_T(x; y)$ and record v of all sampled auxiliary random variables.

1: $x_0^i \sim q_0(-; y)$ $(i = 1 \dots P)$
2: $w_0^i \leftarrow \tilde{p}_0(x_0^i; y)/q_0(x_0^i; y)$ $(i = 1 \dots P)$
3: **for** $t = 1 \dots T$ **do**
4: $\quad a_t^i \leftarrow \text{Categorical}(w_{t-1}^{1:P})$ $(i = 1 \dots P)$
5: $\quad x_t^i \sim q_t(-; x_{t-1}^{a_t^i}, y)$ $(i = 1 \dots P)$
6: $\quad w_t^i \leftarrow \dfrac{\tilde{p}_t(x_t^i; y) l_{t-1}(x_{t-1}^{a_t^i}; x_t^i, y)}{\tilde{p}_{t-1}(x_t^{a^i}; y) q_t(x_t^i; x_{t-1}^{a_t^i}, y)}$ $(i = 1 \dots P)$
7: $I_T \sim \text{Categorical}(w_T^{1:P})$
8: **return** $(v, x) \coloneqq ((I_T, a_{1:T}^{1:P}, x_{0:T}^{1:P}), x^{I_T})$

Algorithm 4 Auxiliary SMC Proposal $r(v; x, y)$

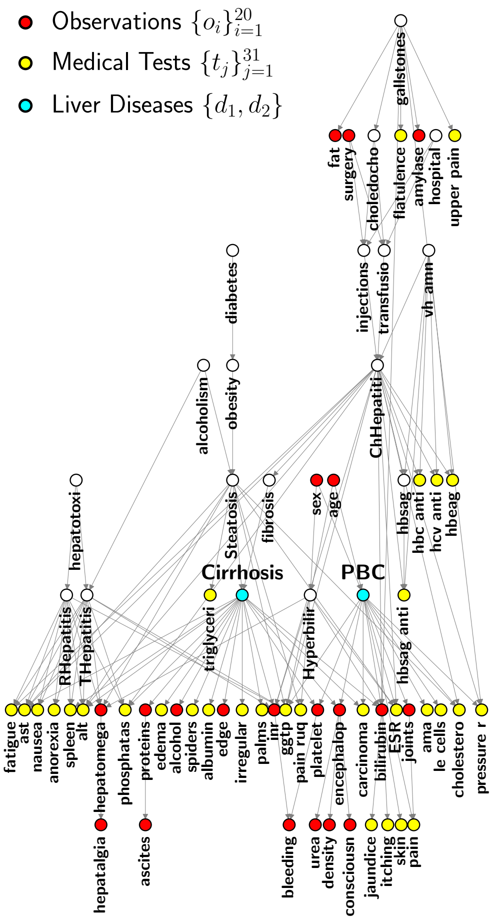
Require: Observation (x, y)
Ensure: Approximate sample v of auxiliary variables generated by a run of Alg. 3 that returned x.

1: $I_T \sim \text{Uniform}(1 \dots P)$
2: $x_T^{I_T} \leftarrow x$
3: **for** $t = T - 1 \dots 0$ **do**
4: $\quad I_t \sim \text{Uniform}(1 \dots P)$
5: $\quad x_t^{I_t} \sim l_t(-; x_{t+1}^{I_{t+1}}, y)$
6: $\quad a_{t+1}^{I_{t+1}} \leftarrow I_t$
7: $x_0^i \sim q_0(-; y)$ $(i = 1 \dots P; i \neq I_0)$
8: $w_0^i \leftarrow \tilde{p}_0(x_0^i; y)/q_0(x_0^i; y)$ $(i = 1 \dots P)$
9: **for** $t = 1 \dots T$ **do**
10: $\quad a_t^i \leftarrow \text{Categorical}(w_{t-1}^{1:P})$ $(i = 1 \dots P, i \neq I_t)$
11: $\quad x_t^i \sim q_t(-; x_{t-1}^{a_t^i})(i = 1 \dots P; i \neq I_t)$
12: $\quad w_t^i \leftarrow \dfrac{\tilde{p}_t(x_t^i; y) l_{t-1}(x_{t-1}^{a_t^i}; x_t^i, y)}{\tilde{p}_{t-1}(x_t^{a^i}; y) q_t(x_t^i; x_{t-1}^{a_t^i}, y)}$ $(i = 1 \dots P)$
13: **return** $v \coloneqq (I_T, a_{1:T}^{1:P}, x_{0:T}^{1:P})$

updating (Del Moral et al., 2006) for static models. Alg. 3 shows the proposal $q(v, x; y)$ from a run of SMC with initial kernel $q_0(x_0; y)$; forward kernels $q_t(x_t; x_{t-1}, y)$ $(t = 1 \dots T)$; backward kernels $l_t(x_t; x_{t+1}, y)$ $(t = 0 \dots T-1)$; and P particles. Here, $v \coloneqq (I_T, a_{1:T}^{1:P}, x_{0:T}^{1:P})$ contains all auxiliary variables and $x \sim \delta(x_T^{I_T})$ is the selected final particle. Alg. 4 shows the auxiliary proposal $r(v; x, y)$, which infers v given (x, y) using generalized "conditional SMC" (Andrieu et al., 2010; Cusumano-Towner and Mansinghka, 2017). Simplifying the weights (24) and (25) gives

$$w(v, x, y) = \prod_{t=0}^{T} \left[\frac{1}{P} \sum_{j=1}^{P} w_t^j \right], \; w'(v, x, y) = \frac{1}{w(v, x, y)},$$

where w_t^j terms are defined in Alg. 3 line 6 for $w(v, x, y)$ and Alg. 4 line 12 for $w'(v, x, y)$. It is also possible to compose SMC with SIR (Appx. A, Algs. 11 and 12) and learn q_t variationally (Maddison et al., 2017). ◁



(a) HEPAR Liver Disease Model (Oniśko, 2003)

(b) Top 10 medical tests for liver disease $d = \text{PBC}$

Decreasing Information Value ↓

Medical Test t	$H(d \mid t, \{o_i\}_{i=1}^{20})$	Prediction Error for d
ama	0.274	0.030
ESR	0.346	0.094
cholesterol	0.385	0.126
ggtp	0.397	0.131
carcinoma	0.402	0.144
pain	0.404	0.148
pressure ruq	0.410	0.174
le cells	0.411	0.172
irregular liver	0.413	0.173
edge	0.415	0.175
no test	0.418	0.175
no test or obs	0.663	0.486

(c) Top 10 tests for liver disease $d = \text{Cirrhosis}$

Decreasing Information Value ↓

Medical Test t	$H(d \mid t, \{o_i\}_{i=1}^{20})$	Prediction Error for d
irregular liver	0.225	0.035
edge	0.247	0.036
spiders	0.253	0.037
spleen	0.256	0.043
palms	0.261	0.043
carcinoma	0.261	0.049
edema	0.262	0.056
triglycerides	0.264	0.056
albumin	0.268	0.068
phosphatase	0.272	0.068
no test	0.290	0.070
no test or obs	0.320	0.080

Figure 4: Using EEVI to rank diagnostic medical tests (yellow) in the HEPAR liver disease network by how informative they are about diseases (blue) given a pattern of observations (red). For both the PBC disease in (b) and cirrhosis diseases in (c), conducting tests that give lower conditional entropy $H(d \mid t, \{o_i\}_{i=1}^{20})$ of the disease (i.e., higher conditional mutual information) results in lower prediction errors about its presence or absence.

4 APPLICATIONS

We applied EEVI to two data acquisition problems: In Sec. 4.1, we rank medical tests in an expert system for diagnosing liver disorders according to their conditional mutual information with diseases of interest, given a pattern of symptoms and patient attributes. In Sec. 4.2, we analyze a dynamic insulin model to compute optimal times to take blood glucose measurements that maximize information about a patient's insulin sensitivity, given their insulin and meal schedule. Experiments were conducted with the Gen (Cusumano-Towner et al., 2019) probabilistic programming system (see code supplement).

4.1 HEPAR Liver Disease Network

HEPAR (Lucas et al., 1989) is a medical expert system that helps physicians diagnose complex disorders in the liver and biliary tract. We analyze the Bayesian network variant of HEPAR from Oniśko (2003) shown in Fig. 4a, which contains nodes representing a patient's (i) attributes, such as age and obesity; (ii) latent liver diseases, such as PBC and cirrhosis; and (iii) symptoms, such as nausea and blood pressure. We consider the following inference problem: *Given a patient with a set $\{o_i\}$ of observed attributes and symptoms, which medical tests $\{t_j\}$ for symptoms should the physician conduct to maximize information about the absence or presence of a disease d?* We formalize the problem as ranking the tests by decreasing conditional mutual information (CMI) with d:

$$I(d : t \mid \{o_i\}) = H(d \mid \{o_i\}) - H(d \mid t, \{o_i\}). \quad (29)$$

Since $H(d \mid \{o_i\})$ is constant, it is more computationally efficient (and equivalent) to rank tests $\{t_j\}$ by increasing conditional entropy $H(d \mid t_j, \{o_i\})$ as in (11).

Results. Fig. 4a shows a setting of 20 observed nodes (red), 31 medical tests for symptoms (yellow), and two

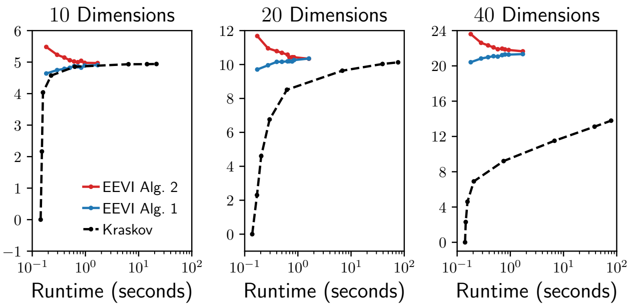
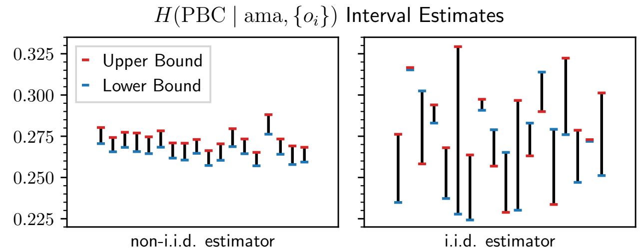
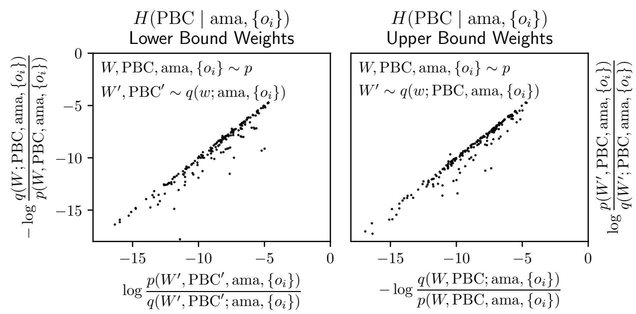


Figure 5: Runtime for estimating entropies in the HEPAR network using EEVI (Algs. 1 and 2) and the baseline nonparametric estimator of Kraskov et al. (2004) for varying dimensionality. Runtime increases with number of SIR samples P for EEVI (see (28)) and number of simulations from $p(x, y)$ for Kraskov et al.



(a) 36 Realizations of estimators of conditional entropy.



(b) 200 random samples of log importance weights

Figure 6: (a) Reducing the variance of interval estimators of $H(\mathrm{PBC} \mid \mathrm{ama}, \{o_i\})$ (Fig. 4b, row 1) by non-i.i.d. sampling. (b) As conditional entropy bounds (30) and (31) are average differences (x-axes minus y-axes) of random weights that here are positively correlated, computing differences using non-i.i.d. samples reduces variance by twice the covariance (the variable W refers to all other variables in the HEPAR network Fig. 4a).

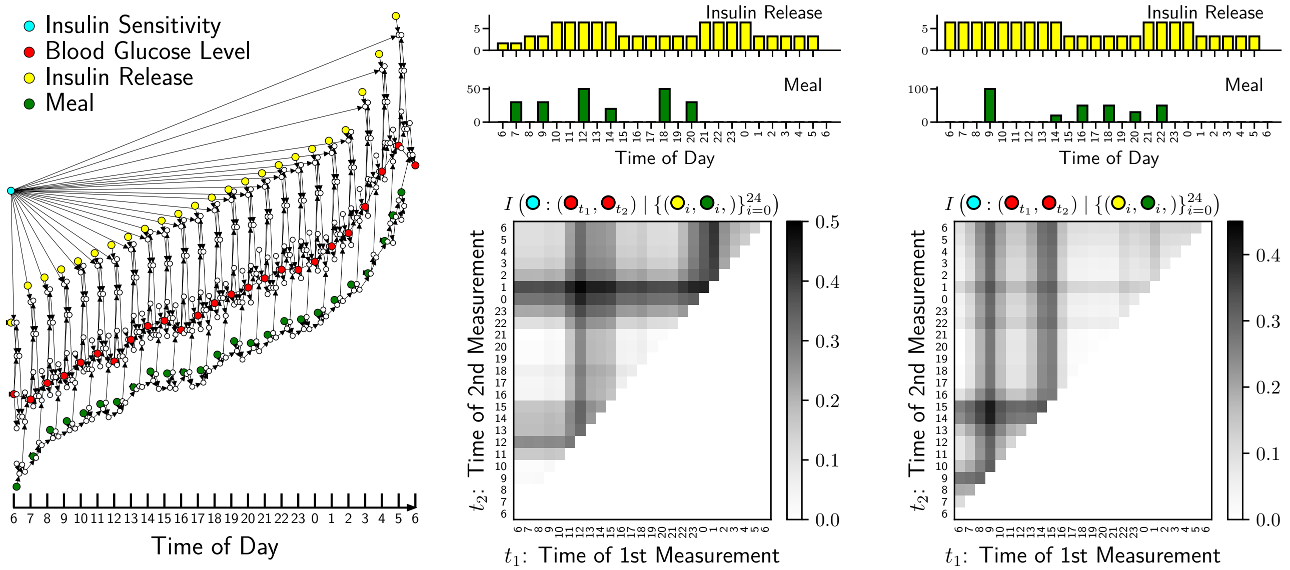
liver diseases (blue). Fig. 4c shows the top 10 tests ranked by $H(d \mid t_j, \{o_i\})$ In Figs. 4b and 4c the first and second columns show the top 10 most informative tests and conditional entropy values for PBC and cirrhosis diseases, respectively. The conditional entropies are computed as the midpoint of interval estimates from EEVI (Algs. 1 and 2), using SIR auxiliary variable proposal (Example 3.1) with ancestral sampling base proposal and number of particles P that drives the interval width to below 10^{-3} nats. To assess how useful these rankings might be to a physician, the third columns in Figs. 4b and 4c shows median prediction errors for each disease in 5000 random patients obtained by conditioning on a given test. For each test t_k, prediction error is defined as the log loss between the posterior $p(d \mid t_k, \{o_i\})$ and the ground-truth label; for reference, prediction errors using $p(d \mid \{o_i\})$ (i.e., no test) and $p(d)$ (i.e., no test or obs) are also shown. The results confirm that tests with higher information values correlate with lower errors and that conditional entropy estimates from EEVI can serve as a useful decision-making tool in this expert system.

Runtime. Fig. 5 compares runtime vs. accuracy profiles of EEVI to the nonparametric estimator of Kraskov et al. (2004) for estimating the joint entropy $H(\{o_i\}_{i=1}^k)$ of k-dimensional random variables in the HEPAR network (k = 10, 20, 40). For these queries, upper and lower bounds from EEVI converge between 1–5 seconds for each k. In contrast, the nonparametric estimator converges slower, as it is "model-free" and estimates log probabilities from simulated data without leveraging model structure. The plots also highlight a key feature of EEVI: the width of the interval quantifies the accuracy of the estimate at any given level of computation and can squeeze the true value, whereas the nonparametric estimator provides point estimates (typically lower bounds) whose accuracy at a given level of computation is unknown.

Variance Control. Following (11), bounds on the conditional entropy $H(d \mid t_j, \{o_i\})$ in Fig. 4 are a difference of two marginal entropy bounds (see also Fig. 1):

$$H^{\mathrm{lb}}(d \,|\, t_j, \{o_i\}) = H^{\mathrm{lb}}(d, t_j, \{o_i\}) - H^{\mathrm{ub}}(t_j, \{o_i\}), \quad (30)$$
$$H^{\mathrm{ub}}(d \,|\, t_j, \{o_i\}) = H^{\mathrm{ub}}(d, t_j, \{o_i\}) - H^{\mathrm{lb}}(t_j, \{o_i\}). \quad (31)$$

Fig. 6a shows 18 realizations of interval estimates of conditional entropy using shared samples (non-i.i.d. Estimator) and independent samples (i.i.d. Estimator) to bound the marginal entropies. The non-i.i.d. Estimator has lower variance as compared to the i.i.d. Estimator and ensures that the realized lower bound is smaller than the realized upper bound. Fig. 6b explains this behavior in terms of the correlation of the random weights (24) and (25) used to estimate the four marginal entropies in (30) and (31) (recall that $\mathrm{Var}[A - B] = \mathrm{Var}[A] + \mathrm{Var}[B] - 2\mathrm{Cov}[A, B]$ for any pair of real random variables A and B). We recommend that practitioners empirically assess the variance and width of the interval estimators as in Fig. 6b, as well as inspect scatter plots of the weights when adding/subtracting bounds from EEVI to bounds on derived information measures.



(a) Dynamic model of carbohydrate metabolism (Andreassen et al., 1991)

(b) Meal and Insulin Scenario 1

(c) Meal and Insulin Scenario 2

Figure 7: Inferring optimal pairs of times to measure blood glucose level (red) that maximize information about a patient's latent insulin sensitivity (blue). Each heatmap in (b)–(c) shows estimates from EEVI of the conditional mutual information of insulin sensitivity with a pair of blood glucose measurements for all pairs of times, under a certain scenario of the patient's insulin release (yellow) and meal (green) schedule. Each scenario has a different optimal pair of times to measure blood glucose: 12pm/1am and 9am/3pm, respectively.

4.2 Dynamic Insulin Model for Diabetes

We applied EEVI to solve a data acquisition task in a differential equation model of carbohydrate metabolism (Andreassen et al., 1991) used by physicians for insulin adjustment in diabetic patients. While most medications have two or three standard dosing options, insulin dose is highly individualized to each patient. Finding the correct dose often requires an iterative adjustment process that accounts for the patient's insulin response as well as their insulin sensitivities and changing clinical conditions. A variety of commercial insulin management software such as Glytec and EndoTool have been developed to aid clinicians with this nuanced and costly process.

In Fig. 7a the "insulin sensitivity" node (blue) is a global latent parameter that dictates how effectively the patient converts released insulin (from e.g., oral medications or injections) into biologically usable insulin. The "meal" (green) and "insulin release" (yellow) nodes are intervention variables based on the patient's victual and medication intake over a 25-hour period. At time t, the blood glucose level is a noisy function of insulin sensitivity and the following variables at $t-1$: blood glucose level, meal, insulin release, and 14 intermediate biological latent variables (white nodes); see Andreassen et al. (1991) for full details.

Suppose a diabetic patient is undergoing the insulin adjustment process and the physician is interested in the following problem: *Given the patient's meal and insulin release schedule, at which pairs of times should blood glucose level be measured to maximize information about insulin sensitivity?* We formalize the problem as ranking pairs of times (t_1, t_2) by decreasing CMI values with insulin sensitivity (for $0 \le t_1 < t_2 \le 24$), given the meal and insulin release schedule:

$$I(\text{ins sens}, (\text{BG}_{t_1}, \text{BG}_{t_2}) \mid \{(\text{meal}_t, \text{ins rel}_t)\}_{t=0}^{24}),$$

where BG_t indicates blood glucose measured at time t and "ins sense" the latent insulin sensitivity. To handle the temporal model, we compute entropy bounds using the SMC proposals in Example 3.2 and Algs. 3 and 4, with a number of particles P that squeezes the interval width to 10^{-2} nats. Figs. 7b and 7c show estimates of CMI for all pairs of time points under two different scenarios of meal and insulin release. Assuming that the model of Andreassen et al. (1991) is accurate, the optimal times from these heatmaps may represent valuable insights for the physician, as insulin sensitivity relates to blood glucose level through complex dynamics of carbohydrate metabolism that are challenging and costly to assess heuristically. Interval estimators of entropy enable quantitative analysis of information values of time points by probabilistic inference in the model for any meal and insulin schedule.

5 RELATED WORK

Several works have studied variational bounds of information measures in a known distribution (Barber and Agakov, 2004; Alemi and Fischer, 2018; Foster et al., 2019; Poole et al., 2019). Unlike EEVI, these previous estimators do not not apply to arbitrary subsets of random variables in a generative model and do not return interval estimates. Thus, the estimators in this paper can compute two-sided bounds in settings not handled previously. For example, upper bounding $I(X:Y)$ in Poole et al. (2019, Fig. 1) and Foster et al. (2019, Eq. (9)) assumes $p(y \mid x)$ is tractable. In contrast, Algs. 1 and 2 in this paper can be used to compute interval estimates of $I(X:Y)$ even when $p(y \mid x)$ and $p(x \mid y)$ are intractable; see applications in Sec. 4.

Several works have developed nonparametric entropy estimators given i.i.d. data from an unknown distribution (Kozachenko and Leonenko, 1987; Paninski, 2003; Kraskov et al., 2004; Pérez-Cruz, 2009; Belghazi et al., 2018; Goldfeld et al., 2020). This work assumes that the distribution is known. With our assumptions on the target model p, nonparametric estimators can often be used in model-based settings by applying them to i.i.d. data simulated from the model. A drawback of using nonparametric estimators in this way, however, is that they ignore the known model structure: Fig. 5 suggests that EEVI scales better, because the model structure is used to build a suitable proposal. A second advantage of EEVI is that the interval width indicates the quality of the estimate at a given level of computational effort, whereas nonparametric methods do not deliver two-sided bounds. The flip side is that building accurate proposals for EEVI needs more expertise as compared to nonparametric methods.

Rainforth et al. (2018) give a thorough treatment of consistency and convergence properties of a very general class of nested Monte Carlo estimators. The expressions in (8) and (9) used for EEVI are instances of nested Monte Carlo, where the inner expectation is obtained via pseudo-marginal methods (Andrieu and Roberts, 2009) and the non-linear mapping is log.

Grosse et al. (2015) introduced the idea of using annealed importance sampling or sequential harmonic mean to "sandwich" marginal log probabilities. Grosse et al. (2016) and Wu et al. (2017) applied these estimators to diagnose MCMC inference algorithms and analyze deep generative models, respectively. Our work develops log probability bounds for the new problem of squeezing entropy values and forming interval estimators, as well as composing the estimates to bound many information-theoretic quantities. Cusumano-Towner and Mansinghka (2017) use a similar family of auxiliary-variable importance samplers from Sec. 3.1 to upper bound symmetric KL divergences between a pair of distributions—in that setting, the normalizing constants in (17) are irrelevant as they cancel out so the weights (24) and (25) are only needed up to normalizing constants. In contrast, the normalizing constants in EEVI cannot be ignored as they are the essential quantities needed to bound entropies.

We implemented the EEVI (Algs. 1 and 2) as metaprograms in the Gen probabilistic programming system (Cusumano-Towner et al., 2019), available in the online code supplement. These implementations make it easy to apply EEVI to a broad set of generative models that are specified as probabilistic programs in Gen, provided that the target random variables correspond to random choices at addresses that exist in each execution of the program. Moreover, EEVI is more widely applicable than previous probabilistic programming-based estimators for information-theoretic quantities, such as Saad and Mansinghka (2017, Alg. 2a), Gehr et al. (2020, Fig. 11), and Narayanan and Shan (2020, Sec. 8.2)—these estimators assume that the probabilistic programming system can exactly compute any marginal or conditional density, which is rarely possible except in languages that restrict modeling expressiveness to enable exact inference (Saad and Mansinghka, 2016; Gehr et al., 2016; Narayanan et al., 2016; Saad et al., 2021).

6 CONCLUSION

We have introduced estimators of entropy via inference (EEVI), a new solution to the fundamental problem of estimating the entropy of arbitrary variables in a generative model by leveraging probabilistic inference. EEVI computes interval estimates of entropy that can be composed to accurately squeeze several other information-theoretic quantities. The experiments show that EEVI delivers information-theoretically optimal solutions to challenging problems from hepatology and endocrinology. Avenues for future work include using EEVI for optimal experiment design and information analysis in widely used medical expert systems (Zhou and Sordo, 2021) by developing the application in tandem with clinical domain experts. To increase the level of automation and accuracy of EEVI, it may be worthwhile to leverage recent probabilistic programming methods that automatically learn amortized proposal distributions (Paige and Wood, 2016; Ritchie et al., 2016; Le et al., 2017). Another direction is using EEVI to extend the class of models and queries that can be handled by existing probabilistic programming-based frameworks for tasks such as searching structured databases (Saad et al., 2017) and optimal experiment design (Ouyang et al., 2018) that need accurate estimates of entropy and information.

References

Alexander A. Alemi and Ian Fischer. GILBO: One metric to measure them all. In *NeurIPS 2018: Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018.

Steen Andreassen, Roman Hovorka, Jonathan Benn, Kristian G. Olesen, and Ewart R. Carson. A model-based approach to insulin adjustment. In *AIME 91: Proceedings of the Third Conference on Artificial Intelligence in Medicine*, volume 44 of *Lecture Notes in Informatics*, pages 239–248, Berlin, Heidelberg, 1991. Springer. doi:10.1007/978-3-642-48650-0_19.

Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009. doi:10.1214/07-AOS574.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010. doi:10.1111/j.1467-9868.2009.00736.x.

David Barber and Felix Agakov. The IM algorithm: A variational approach to information maximization. In *NIPS 2003: Advances in Neural Information Processing Systems 16*, pages 201–208. MIT Press, 2004.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML 2018: Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR, 2018.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *ICLR 2016: International Conference on Learning Representations*, 2016. doi:10.48550/arXiv.1509.00519.

Marco F. Cusumano-Towner and Vikash K. Mansinghka. AIDE: An algorithm for measuring the accuracy of probabilistic inference algorithms. In *NIPS 2017: Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.

Marco F. Cusumano-Towner, Feras A. Saad, Alexander K. Lew, and Vikash K. Mansinghka. Gen: A general-purpose probabilistic programming system with programmable inference. In *PLDI 2019: Proceedings of the 40th ACM SIGPLAN Conference on Programming Design and Implementation*, pages 221–236, New York, NY, USA, 2019. ACM. doi:10.1145/3314221.3314642.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006. doi:10.1111/j.1467-9868.2006.00553.x.

Arnaud Doucet and Adam M. Johansen. Tutorial on particle filtering and smoothing: Fifteen years later. In Dan Crisan and Boris Rozovskii, editors, *The Oxford Handbook of Nonlinear Filtering*, chapter 24, pages 656–704. Oxford University Press, Oxford, 2011.

Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. Variational Bayesian optimal experimental design. In *NeurIPS 2019: Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

Timon Gehr, Sasa Misailovic, and Martin Vechev. PSI: Exact symbolic inference for probabilistic programs. In *CAV 2016: Proceedings of the 28th International Conference on Computer Aided Verification*, volume 9779 of *Lecture Notes in Computer Science*, pages 62–83, Cham, 2016. Springer. doi:10.1007/978-3-319-41528-4_4.

Timon Gehr, Samuel Steffan, and Martin Vechev. λPSI: Exact inference for higher-order probabilistic programs. In *PLDI 2020: Proceedings of the 41st ACM SIGPLAN Conference on Programming Design and Implementation*, pages 883–897, New York, NY, USA, 2020. ACM. doi:10.1145/3385412.3386006.

Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, 2020. doi:10.1109/TIT.2020.2975480.

Roger B. Grosse, Zoubin Ghahramani, and Ryan P. Adams. Sandwiching the marginal likelihood using bidirectional Monte Carlo. *arXiv*, arXiv:1511.02543, 2015. doi:10.48550/arXiv.1511.02543.

Roger B. Grosse, Siddharth Ancha, and Daniel M. Roy. Measuring the reliability of MCMC inference with bidirectional Monte Carlo. In *NIPS 2016: Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016.

Te Sun Han. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, 36(2):133–156, 1978. doi:10.1016/S0019-9958(78)90275-9.

Geoffrey E. Hinton, Peter Dayan, and Radford M. Frey, Brendan J. Neal. The “wake-sleep” algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995. doi:10.1126/science.7761831.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR 2014: International Conference on Learning Representations*, 2014. doi:10.48550/arXiv.1312.6114.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019. doi:10.1561/2200000056.

L. F. Kozachenko and N. N. Leonenko. A statistical estimate for the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004. doi:10.1103/PhysRevE.69.066138.

Tuan Anh Le, Atilim Gunes Baydin, and Frank Wood. Inference compilation and universal probabilistic programming. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1338–1348. PMLR, 2017.

P. J. F. Lucas, R. W. Segaar, and A. R. Janssens. HEPAR: an expert system for the diagnosis of disorders of the liver and biliary tract. *Liver*, 9(5):266–275, October 1989. doi:10.1111/j.1600-0676.1989.tb00410.x.

Chris J. Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Whye Teh. Filtering variational objectives. In *NIPS 2017: Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.

Praveen Narayanan and Chung-chieh Shan. Symbolic disintegration with a variety of base measures. *ACM Transactions on Programming Language Systems*, 42(2), 2020. doi:10.1145/3374208.

Praveen Narayanan, Jacques Carette, Wren Romano, Chung-chieh Shan, and Robert Zinkov. Probabilistic inference by program transformation in Hakaru (system description). In *FLOPS 2016: Proceedings of the 13th International Symposium on Functional and Logic Programming*, volume 9613 of *Lecture Notes in Computer Science*, pages 62–79, Cham, 2016. Springer. doi:10.1007/978-3-319-29604-3_5.

Radford Neal. The harmonic mean of the likelihood: Worst Monte Carlo method ever. https://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/, 2008.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001. doi:10.1023/A:1008923215028.

Agnieszka Oniśko. *Probabilistic Causal Models in Medicine: Application to Diagnosis of Liver Disorders*. PhD thesis, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Science, 2003.

Long Ouyang, Michael H. Tessler, Daniel Ly, and Noah D. Goodman. webppl-oed: A practical optimal experiment design system. In *COGSCI 2018: Proceedings of the 40th Annual Cognitive Science Society Meeting*, pages 2192–2197, Austin, TX, USA, 2018. Cognitive Science Society.

Brooks Paige and Frank Wood. Inference networks for sequential Monte Carlo in graphical models. In *ICML 2016: Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 3040–3049, New York, New York, USA, 2016. PMLR.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003. doi:10.1162/089976603321780272.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, 1988. doi:10.1016/B978-0-08-051489-5.50001-1.

Fernando Pérez-Cruz. Estimation of information theoretic measures for continuous random variables. In *NIPS 2009: Advances in Neural Information Processing Systems 21*. Curran Associates, Inc., 2009.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *ICML 2019: Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180. PMLR, 2019.

Tom Rainforth, Rob Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting Monte Carlo estimators. In *ICML 2018: Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4267–4276. PMLR, 2018.

Daniel Ritchie, Paul Horsfall, and Noah D. Goodman. Deep amortized inference for probabilistic programs. *arXiv*, arXiv:1610.05735, 2016. doi:10.48550/arXiv.1610.05735.

Feras Saad and Vikash Mansinghka. A probabilistic programming approach to probabilistic data analysis. In *NIPS 2016: Advances in Neural Information Processing Systems 29*, pages 2011–2019. Curran Associates, Inc., 2016.

Feras Saad and Vikash Mansinghka. Detecting dependencies in sparse, multivariate databases using probabilistic programming and non-parametric

Bayes. In *AISTATS 2017: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 632–641. PMLR, 2017.

Feras Saad, Leonardo Casarsa, and Vikash Mansinghka. Probabilistic search for structured data via probabilistic programming and nonparametric Bayes. *arXiv*, arXiv:1704.01087, 2017. doi:10.48550/arXiv.1704.01087.

Feras A. Saad, Martin C. Rinard, and Vikash K. Mansinghka. SPPL: Probabilistic programming with fast exact symbolic inference. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Design and Implementation*, pages 804–819, New York, NY, USA, 2021. ACM. doi:10.1145/3453483.3454078.

Tim Salimans, Diederik Kingma, and Max Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In *ICML 2015: Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1218–1226. PMLR, 2015.

Mark J. Schervish. *Theory of Statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1995.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.

Hu Kuo Ting. On the amount of information. *Theory of Probability & Its Applications*, 7(4):439–447, 1962. doi:10.1137/1107041.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960. doi:10.1147/rd.41.0066.

David Wingate, Andreas Stuhlmueller, and Noah Goodman. Lightweight implementations of probabilistic programming languages via transformational compilation. In *AISTATS 2011: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 770–778. PMLR, 2011.

Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the quantitative analysis of decoder-based generative models. In *ICLR 2017: International Conference on Learning Representations*, 2017. doi:10.48550/arXiv.1611.04273.

Li Zhou and Margarita Sordo. Expert systems in medicine. In Lei Xing, Maryellen L. Giger, and James K. Min, editors, *Artificial Intelligence in Medicine*, chapter 5, pages 75–100. Academic Press, 2021. doi:10.1016/B978-0-12-821259-2.00005-3.

Supplementary Material: Estimators of Entropy and Information via Inference in Probabilistic Models

A Special cases of interval estimators of entropy for various proposals

Algs. 1 and 2 in the main text present a general version of Monte Carlo lower and upper bounds for the entropy $H(Y) = -\mathbb{E}[\log p(Y)]$ using the identities (8)–(10), where the functions w and w' are as defined in (24) and (25). The general version of EEVI uses a proposal distribution $q(v, x; y)$ that is defined on an extended space $\mathcal{V} \times \mathcal{X}$ and parameterized by $\mathcal{Y}$, and the auxiliary proposals $r(v; x, y)$ are defined on $\mathcal{V}$ and parameterized by $\mathcal{X} \times \mathcal{Y}$. This appendix rephrases the general version of EEVI in terms of bounds on the negative entropy $-H(Y) = \mathbb{E}[\log p(Y)]$ and presents three special cases described in Sec. 3.1 of the main text.

1. Algs. 5 and 6 show Monte Carlo bounds on $\mathbb{E}[\log p(Y)]$ using a proposal $q(v, x; y)$ and auxiliary proposal $r(v; x, y)$ that have arbitrary auxiliary variables and are defined on the extended state-spaces $\mathcal{V} \times \mathcal{X}$ and $\mathcal{V}$, respectively. The proposals used for the lower and upper bounds need not be the same. Fig. 8 shows the estimation gap that results from using these estimators of $\mathbb{E}[\log p(Y)]$ to estimate $H(Y) = -\mathbb{E}[\log p(Y)]$.

2. Algs. 7 and 8 show Monte Carlo bounds for the case that the proposal q has no auxiliary variables, so that $\mathcal{V} = \{\omega\}$ is a singleton and $r(v; x, y)$ is a dirac measure on ω, as discussed in Sec. 3. Proposals of this form may arise when q is hand-constructed or learned via amortized variational inference.

3. Algs. 9 and 10 show Monte Carlo bounds for the case that the proposal is sampling-importance resampling (SIR) with P particles, using a base proposal $q_0(x; y)$ that has no auxiliary variables. The resulting proposal $q(v, x; y)$ and auxiliary proposal $r(v; x, y)$ are as in Example 3.1 from the main text.

4. Algs. 11 and 12 show Monte Carlo bounds for the case that the proposal is SIR with P particles, using a base proposal $q_0(v, x; y)$ that has auxiliary variables with a corresponding base auxiliary proposal $r_0(v; x, y)$. If the base proposal and auxiliary proposal follow the sequential Monte Carlo (SMC) scheme from Algs. 3 and 4 that are described in Example 3.2 of the main text, then the overall SIR scheme corresponds to P independent runs of SMC and conditional SMC each using P' particles, and a single resampling step that selects one of the P runs. In this case, the overall proposal $q(v, x; y)$ and auxiliary proposal $r(v; x, y)$ on the extended state space are

$$q(v_{1:P}, x; y) = \frac{1}{P} \sum_{k=1}^{P} q_0(v_k, x; y) \prod_{\substack{t=1 \\ t \neq k}}^{P} r_0(v_t; x, y),$$

$$r(v_{1:P}; x, y) = \prod_{t=1}^{P} r_0(v_t; x, y).$$

The extended weight (24) from the main text, which appears in Alg. 11 line 9, is then

$$\begin{aligned} &w(v_{1:P}, x; y) \\ &= \frac{p(x, y) \prod_{k=1}^{P} r_0(v_k; x, y)}{\frac{1}{P} \sum_{k=1}^{P} q_0(v_k, x; y) \prod_{\substack{t=1 \\ t \neq k}}^{P} r_0(v_t; x, y)} \\ &= \frac{p(x, y)}{\frac{1}{P} \sum_{k=1}^{P} \frac{q_0(v_k, x; y)}{r(v_k; x, y)}}. \end{aligned} \tag{32}$$

The extended weight (25) from the main text, which appears in Alg. 12 line 8, is the reciprocal of (32). The extended proposal $q(v_{1:P}, x; y)$ generates samples $(V_{1:P}, X)$ as follows:

- sample $(V_0, X) \sim q_0(v, x; y)$;
- sample selection index $k \sim \text{Uniform}(1 \ldots P)$;
- set $V_k \leftarrow V_0$;
- sample $V_j \sim r_0(v; X, y)$ from the base auxiliary proposal for $j = 1, \ldots, k-1, k+1, \ldots, P$.

As for the extended auxiliary proposal $r(v_{1:P}; x, y)$, it generates P i.i.d. samples from the base auxiliary proposal $r_0(v; x, y)$.

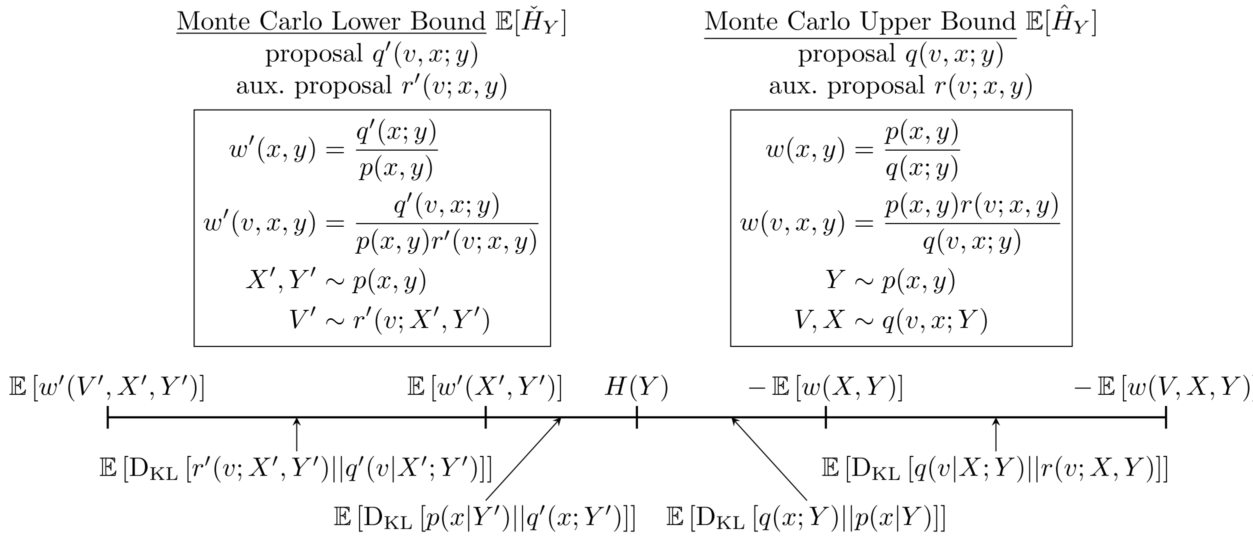


Figure 8: Characterization of the estimation gaps of the upper and lower bounds $\hat{H}_Y$ and $\check{H}_Y$ in Algs. 1 and 2.

B Deferred Proofs

This appendix establishes (16)–(20) from Sec. 3 in the main text using elementary arguments. The basic setup is as follows: let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a measurable space and suppose that $G(\mathrm{d}x)$ is a probability measure that admits a density $g(x)$ with respect to some σ-finite measure $\nu(\mathrm{d}x)$. Therefore,

$$G(A) = \int_{\mathcal{X}} \mathbb{I}[x \in A] g(x)\nu(\mathrm{d}x) \quad (A \in \mathcal{B}(\mathcal{X})), \tag{33}$$

and for any bounded measurable function $\phi : \mathcal{X} \to \mathbb{R}$,

$$\int_{\mathcal{X}} \phi(x)G(\mathrm{d}x) = \int_{\mathcal{X}} \phi(x)g(x)\nu(\mathrm{d}x). \tag{34}$$

Assume now that $g(x) = \tilde{g}(x)/Z_g$ is known only up to a normalizing constant

Proposition B.1. *The normalizing constant Z_g of g satisfies $Z_g = \int_{\mathcal{X}} \tilde{g}(x)\nu(\mathrm{d}x)$.*

Proof.

$$1 = G(\mathcal{X}) = \int_{\mathcal{X}} g(x)\nu(\mathrm{d}x) \tag{35}$$

$$= \int_{\mathcal{X}} [\tilde{g}(x)/Z_g]\nu(\mathrm{d}x) \tag{36}$$

$$= \int_{\mathcal{X}} [\tilde{g}(x)/Z_g]\nu(\mathrm{d}x). \tag{37}$$

□

Proposition B.2. *Let $\mathcal{G} \coloneqq \{x \mid g(x) = 0\}$ be the set of values for which g is zero and let $G' \coloneqq \mathcal{X} \setminus \mathcal{G}$ be its complement. Then $G(\mathcal{G}) = 0$.*

Proof. As g is a measurable function, the set $\mathcal{G}$ is also measurable and has G-probability

$$G(\mathcal{G}) = \int_{\mathcal{X}} \mathbb{I}[x \in \mathcal{G}]g(x)\nu(\mathrm{d}x) = \int_{\mathcal{G}} g(x)\nu(\mathrm{d}x) \tag{38}$$

$$= \int_{\mathcal{G}} 0\nu(\mathrm{d}x) \tag{39}$$

$$= 0. \tag{40}$$

□

Corollary B.3. *For measurable function $\phi : \mathcal{X} \to \mathbb{R}$, we have*

$$\int_{\mathcal{X}} \phi(x)G(\mathrm{d}x) = \int_{\mathcal{G}'} \phi(x)G(\mathrm{d}x). \tag{41}$$

Let H be a probability measure that is absolutely continuous with respect to ν with density $h(x) = \tilde{h}(x)/Z_h$. Prop. B.1 implies that $Z_h = \int_{\mathcal{X}} \tilde{h}(x)\nu\,\mathrm{d}x$. Further, let

$$\tilde{w}(x) \coloneqq \begin{cases} \dfrac{\tilde{h}(x)}{\tilde{g}(x)} & \text{if } x \in \mathcal{G}', \\ 1 & \text{if } x \in \mathcal{G}, \end{cases} \tag{42}$$

$$w(x) \coloneqq \frac{Z_h}{Z_g}\tilde{w}(x) = \begin{cases} \dfrac{h(x)}{g(x)} & \text{if } x \in \mathcal{G}', \\ 1 & \text{if } x \in \mathcal{G}. \end{cases} \tag{43}$$

Algorithm 5 Lower bound on $\mathbb{E}[\log p(Y)]$

Require: Target distribution $p(x, y)$
Proposal distribution $q(v, x; y)$
Auxiliary proposal distribution $r(v; x, y)$
Number of samples n, m

for $i = 1 \dots n$ **do**
  $(\tilde{X}, Y) \sim p(x, y)$
  for $j = 1 \dots m$ **do**
    $(V, X) \sim q(v, x; Y)$
    $t'_j \leftarrow \log \dfrac{p(X, Y) r(V; X, Y)}{q(V, X; y)}$
  $t_i \leftarrow \frac{1}{m} \sum_{j=1}^{m} t'_j$
return $\frac{1}{n} \sum_{i=1}^{n} t_i$

Algorithm 6 Upper bound on $\mathbb{E}[\log p(Y)]$

Require: Target distribution $p(x, y)$
Proposal distribution $q(v, x; y)$
Auxiliary proposal distribution $r(v; x, y)$
Number of samples n, m

for $i = 1 \dots n$ **do**
  $(X_1, Y) \sim p(x, y)$
  $(X_{2:m}) \sim \mathrm{MCMC}_{X_1}$ targeting $p(x \mid Y)$
  for $j = 1 \dots m$ **do**
    $V \sim r(v; X_j, Y)$
    $t'_j \leftarrow \log \dfrac{q(V, X_j; Y)}{p(X_j, Y) r(V; X_j, Y)}$
  $t_i \leftarrow -\frac{1}{m} \sum_{j=1}^{m} t'_j$
return $\frac{1}{n} \sum_{i=1}^{n} t_i$

Algorithm 7 Lower bound on $\mathbb{E}[\log p(Y)]$ using a proposal distribution without auxiliary variables

Require: Target distribution $p(x, y)$
Proposal distribution $q(x; y)$
Number of samples n, m

for $i = 1 \dots n$ **do**
  $Y \sim p(x, y)$
  for $j = 1 \dots m$ **do**
    $X \sim q(x; Y)$
    $t'_j \leftarrow \log \dfrac{p(X, Y)}{q(X; y)}$
  $t_i \leftarrow \frac{1}{m} \sum_{j=1}^{m} t'_j$
return $\frac{1}{n} \sum_{i=1}^{n} t_i$

Algorithm 8 Upper bound on $\mathbb{E}[\log p(Y)]$ using a proposal distribution without auxiliary variables

Require: Target distribution $p(x, y)$
Proposal distribution $q(x; y)$
Number of samples n, m

for $i = 1 \dots n$ **do**
  $(X_1, Y) \sim p(x, y)$
  $(X_{2:m}) \sim \mathrm{MCMC}_{X_1}$ targeting $p(x \mid Y)$
  for $j = 1 \dots m$ **do**
    $t'_j \leftarrow \log \dfrac{q(X_j; Y)}{p(X_j, Y)}$
  $t_i \leftarrow -\frac{1}{m} \sum_{j=1}^{m} t'_j$
return $\frac{1}{n} \sum_{i=1}^{n} t_i$

Algorithm 9 Lower bound on $\mathbb{E}[\log p(Y)]$ using SIR and a base proposal without auxiliary variables

Require: Target distribution $p(x, y)$
Base proposal distribution $q_0(x; y)$
Number of samples n, m, P

for $i = 1 \dots n$ **do**
  $(\tilde{X}, Y) \sim p(x, y)$
  for $j = 1 \dots m$ **do**
    for $k = 1 \dots P$ **do**
      $X \sim q_0(x; Y)$
      $\xi_k \leftarrow \dfrac{p(X, Y)}{q_0(X; Y)}$
    $t'_j \leftarrow \log \left[\frac{1}{P} \sum_{k=1}^{P} \xi_k \right]$
  $t_i \leftarrow \frac{1}{m} \sum_{j=1}^{m} t'_j$
return $\frac{1}{n} \sum_{i=1}^{n} t_i$

Algorithm 10 Upper bound on $\mathbb{E}[\log p(Y)]$ using SIR and a base proposal without auxiliary variables

Require: Target distribution $p(x, y)$
Base proposal distribution $q_0(x; y)$
Number of samples n, m, P

for $i = 1 \dots n$ **do**
  $(X_1, Y) \sim p(x, y)$
  $(X_{2:m}) \sim \mathrm{MCMC}_{X_1}$ targeting $p(x \mid Y)$
  for $j = 1 \dots m$ **do**
    $X'_1 \leftarrow X_j$
    for $k = 2 \dots P$ **do**
      $X'_k \sim q_0(x; Y)$
    for $k = 1 \dots P$ **do**
      $\xi_k \leftarrow \dfrac{p(X'_k, Y)}{q_0(X'_k; y)}$
    $t'_j \leftarrow \log \left[1 / \frac{1}{P} \sum_{k=1}^{P} \xi_k \right]$
  $t_i \leftarrow -\frac{1}{m} \sum_{j=1}^{m} t'_j$
return $\frac{1}{n} \sum_{i=1}^{n} t_i$

Algorithm 11 Lower bound on $\mathbb{E}[\log p(Y)]$ using SIR and a base proposal with auxiliary variables

Require: Target distribution $p(x, y)$
Base proposal distribution $q_0(v, x; y)$
Base auxiliary proposal dist $r_0(v; x, y)$
Number of samples n, m, P

for $i = 1 \dots n$ **do**
  $(\tilde{X}, Y) \sim p(x, y)$
  for $j = 1 \dots m$ **do**
    $(V_1, X) \sim q_0(v, x; Y)$
    for $k = 2 \dots P$ **do**
      $V_k \sim r_0(v; X, Y)$
    for $k = 1 \dots P$ **do**
      $\xi_k \leftarrow \dfrac{q(V_k, X; Y)}{r(V_k; X, Y)}$
    $t'_j \leftarrow \log \dfrac{p(X, Y)}{\frac{1}{P} \sum_{k=1}^{P} \xi_k}$
  $t_i \leftarrow \frac{1}{m} \sum_{j=1}^{m} t'_j$
return $\frac{1}{n} \sum_{i=1}^{n} t_i$

Algorithm 12 Upper bound on $\mathbb{E}[\log p(Y)]$ using SIR and a base proposal with auxiliary variables

Require: Target distribution $p(x, y)$
Base proposal distribution $q_0(v, x; y)$
Base auxiliary proposal dist $r_0(v; x, y)$
Number of samples n, m, P

for $i = 1 \dots n$ **do**
  $(X_1, Y) \sim p(x, y)$
  $(X_{2:m}) \sim \mathrm{MCMC}_{X_1}$ targeting $p(x \mid Y)$
  for $j = 1 \dots m$ **do**
    for $k = 1 \dots P$ **do**
      $V_k \sim r(v; X_j, Y)$
      $\xi_k \leftarrow \dfrac{q(V_k, X_j; Y)}{r(V_k; X_j, Y)}$
    $t'_j \leftarrow \log \dfrac{\frac{1}{P} \sum_{k=1}^{P} \xi_k}{p(X_j, Y)}$
  $t_i \leftarrow -\frac{1}{m} \sum_{j=1}^{m} t'_j$
return $\frac{1}{N} \sum_{i=1}^{N} t_i$

Proposition B.4. *If $H \ll G$ then $w(x)$ is a density of H with respect to G.*

Proof. Since $H \ll G$ and $G(\mathcal{G}) = 0$ it also holds that $H(\mathcal{G}) = 0$. Thus Corr. B.3 also applies when taking expectations under H. For any $A \in \mathcal{B}(\mathcal{X})$

$$
\begin{aligned}
&\int_{\mathcal{X}} \mathbb{I}[x \in A] w(x) G(\mathrm{d}x) && (44)\\
&= \int_{\mathcal{G}'} \mathbb{I}[x \in A] w(x) G(\mathrm{d}x) && (45)\\
&= \int_{\mathcal{G}'} \mathbb{I}[x \in A] [h(x)/g(x)] G(\mathrm{d}x) && (46)\\
&= \int_{\mathcal{G}'} \mathbb{I}[x \in A] [h(x)/g(x)] g(x) \nu(\mathrm{d}x) && (47)\\
&= \int_{\mathcal{G}'} \mathbb{I}[x \in A] h(x) \nu(\mathrm{d}x) && (48)\\
&= H(A). && (49)
\end{aligned}
$$

□

Corollary B.5. *If $H \ll G$ and $X \sim G$, then $\mathbb{E}[w(X)] = 1$.*

Prop. B.6 establishes (16) from the main text.

Proposition B.6. *If $H \ll G$ and $X \sim G$, then $\mathbb{E}[\tilde{w}(X)] = Z_h/Z_g$.*

Proof. Eq. (42) implies that $\tilde{w}(x) = [Z_h/Z_g]w(x)$ for all $x \in \mathcal{X}$. Applying Corr. B.5:

$$
\begin{aligned}
\mathbb{E}[\tilde{w}(X)] &= \mathbb{E}[[Z_h/Z_g] w(X)] && (50)\\
&= [Z_h/Z_g]\, \mathbb{E}[w(X)] && (51)\\
&= Z_h/Z_g. && (52)
\end{aligned}
$$

□

Props. B.7 and B.8 establish (17) from the main text.

Proposition B.7. *If $G \ll H$ then $\tilde{w}(x)$ is G-almost surely positive.*

Proof. Suppose towards a contradiction that there exists a measurable set A such that $G(A) > 0$ and $\tilde{w}(x) = 0$ for $x \in A$. From (42), it must be that $h(x) = 0$ whenever $x \in A$. But then $H(A) = \int_A h(x) H(\mathrm{d}x) = 0$, a contradiction to $G \ll H$. □

Proposition B.8. *If $G \ll H$ and $X \sim G$, then $\mathbb{E}[\log \tilde{w}(X)] = \log(Z_h/Z_g) - \mathrm{D}_{\mathrm{KL}}[G||H]$.*

Proof. Prop. B.7 and the definition $w(x)$ in (42) together imply that $\mathrm{D}_{\mathrm{KL}}[G||H] = -\mathbb{E}[\log w(X)]$ for $X \sim G$. Since $G \ll H$, we have

$$
\begin{aligned}
\mathbb{E}[\log \tilde{w}(X)] &= \mathbb{E}\left[\log \frac{Z_h}{Z_g} w(X)\right] && (53)\\
&= \log(Z_h/Z_g) + \mathbb{E}[\log w(X)] && (54)\\
&= \log(Z_h/Z_g) - \mathrm{D}_{\mathrm{KL}}[G||H]. && (55)
\end{aligned}
$$

□

Proposition B.9. *If $G \ll H$ and $X \sim G$, then $\mathrm{Var}[\log \tilde{w}(X)] = \mathbb{E}\left[\log^2 w(X)\right] - (\mathrm{D}_{\mathrm{KL}}[G||H])^2$.*

Proof. Recall that for any real random variable B and constant c, $\mathrm{Var}[B] = \mathbb{E}\left[B^2\right] - (\mathbb{E}[B])^2$ and $\mathrm{Var}[B + c] = \mathrm{Var}[B]$. Applying this property to the random variable $\log \tilde{w}(X)$ gives

$$
\begin{aligned}
&\mathrm{Var}[\log \tilde{w}(X)] && (56)\\
&= \mathrm{Var}\left[\log \frac{Z_h}{Z_g} + \log w(X)\right] && (57)\\
&= \mathrm{Var}[\log w(X)] && (58)\\
&= \mathbb{E}\left[\log^2 w(X)\right] - (\mathbb{E}[\log w(X)])^2 && (59)\\
&= \mathbb{E}\left[\log^2 w(X)\right] - (-\mathrm{D}_{\mathrm{KL}}[G||H])^2 && (60)\\
&= \mathbb{E}\left[\log^2 w(X)\right] - (\mathrm{D}_{\mathrm{KL}}[G||H])^2. && (61)
\end{aligned}
$$

□

Prop. B.9 establishes (18). We next establish (19).

Proposition B.10. *If $H \ll G$ and $X \sim G$, then*

$$
\Pr[\log \tilde{w}(X) \ge t + \log(Z_h/Z_g)] \le e^{-t} \qquad (62)
$$

for any $t > 0$.

Proof. Prop. B.6 gives $\mathbb{E}[\tilde{w}(X)] = Z_h/Z_g$. Further, from Prop. B.7 $\tilde{w}$ is G-almost surely positive. Thus, $\Pr[\tilde{w}(X) \ge e^t(Z_h/Z_g)] \le e^{-t}$ for $t > 0$. Applying log to both sides as in (19) completes the proof. □

Eq. (20) in the main text is a special case of a result in Burda et al. (2016, Appx. B) to importance sampling estimators that satisfy Props. B.8 and B.10. For completeness, we give the proof in two stages using the notation in this paper below.

Proposition B.11. *Let B be a real random variable with finite expectation and $\mu := \mathbb{E}[B]$ denote its expectation. Then $\mathbb{E}[|B - \mu|] = 2\,\mathbb{E}[\max(0, B - \mu)]$.*

Proof. By additivity of expectation

$$
\begin{aligned}
\mathbb{E}[|B - \mu|] = &\ \mathbb{E}[|B - \mu| \cdot \mathbf{1}_{B>\mu}] && (63)\\
&+ \mathbb{E}[|B - \mu| \cdot \mathbf{1}_{B<\mu}]\\
&+ \mathbb{E}[|B - \mu| \cdot \mathbf{1}_{B=\mu}].
\end{aligned}
$$

As the third term in the right hand side of (63) is zero, it suffices to establish that the first two terms on the right hand side are equal:

$$\mu = \mathbb{E}\left[B\mathbf{1}_{B>\mu}\right] + \mathbb{E}\left[B\mathbf{1}_{B<\mu}\right] + \mathbb{E}\left[B\mathbf{1}_{B=\mu}\right] \tag{64}$$

$$\mu = \mathbb{E}\left[B\mathbf{1}_{B>\mu}\right] + \mathbb{E}\left[B\mathbf{1}_{B<\mu}\right] + \mu\,\mathbb{E}\left[(1-(\mathbf{1}_{B>\mu}+\mathbf{1}_{B<\mu}))\right] \tag{65}$$

$$\mu = \mathbb{E}\left[B\mathbf{1}_{B>\mu}\right] + \mathbb{E}\left[B\mathbf{1}_{B<\mu}\right] + \mu - \mu\,\mathbb{E}\left[\mathbf{1}_{B>\mu}\right] - \mu\,\mathbb{E}\left[\mathbf{1}_{B<\mu}\right] \tag{66}$$

$$0 = \mathbb{E}\left[(B-\mu)\mathbf{1}_{B>\mu}\right] - \mathbb{E}\left[(\mu-B)\mathbf{1}_{B<\mu}\right]. \tag{67}$$

Using the fact that $-X\mathbf{1}_{X<0} = |X|\mathbf{1}_{X<0}$ and $X\mathbf{1}_{X>0} = |X|\mathbf{1}_{X>0}$ a.s., for any random variable X, we have

$$\mathbb{E}\left[(\mu-B)\mathbf{1}_{B<\mu}\right] = \mathbb{E}\left[(B-\mu)\mathbf{1}_{B>\mu}\right] \tag{68}$$

$$\mathbb{E}\left[|\mu-B|\cdot\mathbf{1}_{B<\mu}\right] = \mathbb{E}\left[|B-\mu|\cdot\mathbf{1}_{B>\mu}\right]. \tag{69}$$

Combining (63), (68) and (69) gives.

$$\mathbb{E}\left[|B-\mu|\right] = 2\,\mathbb{E}\left[|B-\mu|\cdot\mathbf{1}_{B>\mu}\right] \tag{70}$$

$$= 2\,\mathbb{E}\left[(B-\mu)\cdot\mathbf{1}_{B>\mu}\right] \tag{71}$$

$$= 2\,\mathbb{E}\left[\max(0,B-\mu)\right]. \tag{72}$$

□

Proposition B.12. *If H and G are mutually absolutely continuous, i.e., $G \ll H$, $H \ll G$, and $X \sim G$, then*

$$\mathbb{E}\left[\left|\log\tilde{w}(X) - \mathbb{E}\left[\log\tilde{w}(X)\right]\right|\right] \leq 2 + 2\,\mathrm{D}_{\mathrm{KL}}\left[G||H\right]. \tag{73}$$

Proof. For any real random variable B, let $(B)_+ \coloneqq \max(0,B)$ denote the positive part. Then

$$\mathbb{E}\left[\left|\log\tilde{w}(X) - \mathbb{E}\left[\log\tilde{w}(X)\right]\right|\right] \tag{74}$$

$$= 2\,\mathbb{E}\left[(\log\tilde{w}(X) - \mathbb{E}\left[\log\tilde{w}(X)\right])_+\right] \tag{75}$$

$$= 2\mathbb{E}\Big[\Big(\log\tilde{w}(X) - \log(Z_h/Z_g) + \log(Z_h/Z_g) - \mathbb{E}\left[\log\tilde{w}(X)\right]\Big)_+\Big] \tag{76}$$

$$\leq 2\mathbb{E}\big[(\log\tilde{w}(X) - \log(Z_h/Z_g))_+ + (\log(Z_h/Z_g) - \mathbb{E}\left[\log\tilde{w}(X)\right])_+\big] \tag{77}$$

$$= 2\,\mathbb{E}\left[(\log\tilde{w}(X) - \log(Z_h/Z_g))_+\right] + 2\left(\log(Z_h/Z_g) - \mathbb{E}\left[\log\tilde{w}(X)\right]\right)_+ \tag{78}$$

$$= 2\,\mathbb{E}\left[(\log\tilde{w}(X) - \log(Z_h/Z_g))_+\right] + 2\,\mathrm{D}_{\mathrm{KL}}\left[G||H\right] \tag{79}$$

$$= 2\int_0^\infty \Pr[\log\tilde{w}(X) - \log(Z_h/Z_g) > t]\,\mathrm{d}t + 2\,\mathrm{D}_{\mathrm{KL}}\left[G||H\right] \tag{80}$$

$$= 2\int_0^\infty \Pr[\log\tilde{w}(X) > \log(Z_h/Z_g) + t]\,\mathrm{d}t + 2\,\mathrm{D}_{\mathrm{KL}}\left[G||H\right] \tag{81}$$

$$\leq 2\int_0^\infty \exp(-t)\,\mathrm{d}t + 2\,\mathrm{D}_{\mathrm{KL}}\left[G||H\right] \tag{82}$$

$$= 2 + 2\,\mathrm{D}_{\mathrm{KL}}\left[G||H\right]]. \tag{83}$$

□