Statistical estimation of mutual information

Adam Foster

September 2021

1 Introduction

Mutual information is a central statistical quantity that measures the relationship between two random variables. In machine learning, it has found use in blind source separation (Hyvärinen, 1999), representation learning (van den Oord et al., 2018), the information bottleneck (Tishby et al., 2000) and feature selection (Kwak and Choi, 2002). It is also a key quantity in Bayesian experimental design (Lindley, 1956). The mutual information between jointly distributed random variables $\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$ is defined as

$$I(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right]. \tag{1}$$

In this document, we focus on the estimation of mutual information in the *explicit likelihood* setting in which one of the conditional densities, say $p(\mathbf{y}|\mathbf{x})$ is known in closed form. In this case, asymptotically consistent estimators exist for the mutual information, and we are concerned in studying their convergence rates. In the *implicit likelihood* setting, the standard approach is to introduce a positive, unnormalised function $\kappa(\mathbf{x}, \mathbf{y})$ that is an estimate of the joint $p(\mathbf{x}, \mathbf{y})$. However, estimators that use κ as a surrogate for the true unknown density can only be guaranteed to produce lower bounds on the mutual information in the limit of infinite samples of $\mathbf{x}, \mathbf{y}$. The convergence rates, though, behave similarly.

2 Nested Monte Carlo and leave-one-out estimators

The Nested Monte Carlo (NMC) estimator (Ryan, 2003), also called the double loop estimator, for mutual information estimation with an explicit likelihood is defined as

$$A_{n,m} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{m}\sum_{j=1}^{m} p(\mathbf{y}_i|\mathbf{x}_{ij})} \tag{2}$$

where $\mathbf{x}_i, \mathbf{y}_i \overset{\text{i.i.d.}}{\sim} p(\mathbf{x}, \mathbf{y})$ and $\mathbf{x}_{ij} \overset{\text{i.i.d.}}{\sim} p(\mathbf{x})$ are independent. It is also possible to include some correlation in the $\mathbf{x}$ samples, for example we can repeatedly use $(\mathbf{x}_{1j})_{j=1}^{m}$

$$A'_{n,m} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{m}\sum_{j=1}^{m} p(\mathbf{y}_i|\mathbf{x}_{1j})}, \tag{3}$$

and we can use the original n samples, giving the leave-one-out (LOO) estimator (Poole et al., 2019)

$$\tilde{A}_n = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{n-1}\sum_{j\neq i} p(\mathbf{y}_i|\mathbf{x}_j)}. \tag{4}$$

Note that $\mathbb{E}[A_{n,m}] = \mathbb{E}[A'_{n,m}]$ and $\mathbb{E}[\tilde{A}_n] = \mathbb{E}[A_{n,n-1}]$, so the correlations only change the variance. Furthermore, estimators $A_{n,m}$ and $A'_{n,m}$ both cost $\mathcal{O}(mn)$ evaluations of the likelihood and $\tilde{A}_n$ costs $\mathcal{O}(n^2)$

evaluations of the likelihood. So, whilst $A'_{n,m}$ and $\tilde{A}_n$ appear more efficient in their use of samples, their theoretical computational complexity is not different to $A_{n,m}$.

Here, we focus on analysing the estimator $A_{n,m}$. Our results reaffirm previous analysis by Rainforth et al. (2018); Zheng et al. (2018); Beck et al. (2018). We focus on a rigorous approach to using Taylor's Theorem for the logarithm. Our techniques can then be used to analyse other estimators.

Theorem 1 (Expectation of $A_{n,m}$)**.** *Suppose there exist Hölder conjugate indices $p, q > 0$ with $1/p+1/q = 1$ such that*

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right] < \infty \text{ and } \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{q}\right] < \infty. \tag{5}$$

Then we have

$$\mathbb{E}[A_{n,m}] = I(\mathbf{x},\mathbf{y}) + \frac{1}{m}\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{2p(\mathbf{y})^2}\right] + \mathcal{O}\left(m^{-3/2}\right). \tag{6}$$

Proof. By linearity, $\mathbb{E}[A_{n,m}] = \mathbb{E}[A_{1,m}]$. To compute this expectation, we define

$$U_j = \frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)}. \tag{7}$$

with $E[U_j] = \mathbb{E}[\mathbb{E}[U_j|\mathbf{y}_1]] = 1$. Then,

$$A_{1,m} = \log\frac{p(\mathbf{y}_1|\mathbf{x}_1)}{p(\mathbf{y}_1)} - \log\left(\frac{1}{m}\sum_{j=1}^{m}U_j\right), \tag{8}$$

giving

$$\mathbb{E}[A_{n,m}] = I(\mathbf{x},\mathbf{y}) - \mathbb{E}\left[\log\left(\frac{1}{m}\sum_{j=1}^{m}U_j\right)\right]. \tag{9}$$

The standard approach to analysing the second term is to apply Taylor's Theorem to the logarithm function. However, a naive application does not work for several reasons: a) the Taylor series for the logarithm about 1 is convergent only on $(0, 2)$ rather than $(0,\infty)$, b) the derivatives of the logarithm are not bounded at 0, so the classical Delta Method (Lemma 9) does not apply. To get around these problems, we define the partial Taylor series

$$L_k(x) = \sum_{j=1}^{k}\frac{(-1)^{j+1}}{j}(x-1)^j, \tag{10}$$

in Lemma 10, we prove that $|\log x - L_k(x)| \le |x-1|^{k+1}\max(1, -\log x)$ on $(0,\infty)$. Taking $k = 2$, we have

$$\mathbb{E}\left[\log\left(\frac{1}{m}\sum_{j=1}^{m}U_j\right)\right] = -\frac{1}{2}\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}(U_j-1)\right)^2\right] + \mathbb{E}[\varepsilon] \tag{11}$$

and

$$|\mathbb{E}[\varepsilon]| \le \mathbb{E}[|\varepsilon|] \le \mathbb{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m}(U_j-1)\right|^3\max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m}U_j\right)\right)\right] \tag{12}$$

applying Hölder's Inequality

$$\le \mathbb{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m}(U_j-1)\right|^{3p}\right]^{1/p}\mathbb{E}\left[\max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m}U_j\right)\right)^q\right]^{1/q}. \tag{13}$$

We tackle each term separately. Since the U_j are i.i.d conditional on $\mathbf{y}_1$, we can apply Corollary 8 that uses the Marcinkiewicz–Zygmund Inequality, and the Tower Law to conclude that there is a finite constant D_{3p} such that

$$\mathbb{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m}(U_j-1)\right|^{3p}\right]^{1/p} \leq D_{3p}^{1/p} m^{-3/2}\mathbb{E}\left[|U_1-1|^{3p}\right]^{1/p} \tag{14}$$

and

$$\mathbb{E}\left[|U_1-1|^{3p}\right] \leq 1+\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right] < \infty \text{ by assumption.} \tag{15}$$

So this term is $\mathcal{O}(m^{-3/2})$. For the latter term, we use the fact that $x \mapsto \max(1, -\log x)$ is a convex function. Thus

$$\begin{aligned}
\mathbb{E}\left[\max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m}U_j\right)\right)^{q}\right]^{1/q} &\leq \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}\max\left(1, -\log\left(U_j\right)\right)^{q}\right]^{1/q} && (16)\\
&= \mathbb{E}\left[\max\left(1, -\log\left(U_1\right)\right)^{q}\right]^{1/q} && (17)\\
&\leq \left(1+\mathbb{E}[|\log U_1|^{q}]\right)^{1/q} && (18)\\
&= \left(1+\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{q}\right]\right)^{1/q} && (19)\\
&< \infty \text{ by assumption.} && (20)
\end{aligned}$$

Overall, we have $\mathbb{E}[\varepsilon] = \mathcal{O}(m^{-3/2})$. Finally,

$$\frac{1}{2}\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}(U_j-1)\right)^{2}\right] = \frac{1}{2m}\mathbb{E}_{p(\mathbf{y})}\left[\mathbb{E}_{p(\mathbf{x})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}-1\right)^{2}\right]\right] = \frac{1}{m}\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{2p(\mathbf{y})^2}\right]. \tag{21}$$

This completes the proof. □

A simple application of Jensen's Inequality further shows that $\mathbb{E}[A_{n,m}] \geq I(\mathbf{x},\mathbf{y})$ for every value of n and m. Put another way, the NMC estimator is always a stochastic upper bound on the mutual information with bias of order $1/m$. Zheng et al. (2018) showed that the coefficient of the $1/m$ term is

$$\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{2p(\mathbf{y})^2}\right] = \frac{1}{2}\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}-1\right)^{2}\right] \tag{22}$$

which is the χ^2-divergence from $p(\mathbf{x},\mathbf{y})$ to $p(\mathbf{x})p(\mathbf{y})$.

Theorem 2 (Variance of $A_{n,m}$)**.** *Assume that there exist Hölder conjugate indices $p, q > 0$ such that*

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right] < \infty \text{ and } \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2q}\right] < \infty. \tag{23}$$

Then,

$$\begin{aligned}
\mathrm{Var}[A_{n,m}] = &\frac{1}{n}\mathrm{Var}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right]\\
&+\frac{1}{nm}\left(\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{p(\mathbf{y})^2}\right]+\mathrm{Cov}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, \frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right]\right) \qquad (24)\\
&+\mathcal{O}\left(n^{-1}m^{-3/2}\right).
\end{aligned}$$

Proof. We have

$$\mathrm{Var}[A_{n,m}] = \frac{1}{n}\mathrm{Var}[A_{1,m}]. \tag{25}$$

For the variance of $A_{1,m}$, we use the Tower Law for the Variance

$$\mathrm{Var}[A_{1,m}] = \mathbb{E}[\mathrm{Var}[A_{1,m}|\mathbf{x}_1, \mathbf{y}_1]] + \mathrm{Var}[\mathbb{E}[A_{1,m}|\mathbf{x}_1, \mathbf{y}_1]]. \tag{26}$$

For the conditional variance, we follow the proof of Theorem 1 to see that

$$\mathbb{E}\left[\mathrm{Var}[A_{1,m}|\mathbf{x}_1, \mathbf{y}_1]\right] = \mathbb{E}\left[\mathrm{Var}\left[\log\left(\frac{1}{m}\sum_{j=1}^m U_j\right)\middle|\mathbf{y}_1\right]\right] \text{ where } U_j = \frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)} \tag{27}$$

We will the form of the variance $\mathrm{Var}[A] = \mathbb{E}[A^2] - \mathbb{E}[A]^2$. We now study the function $x \mapsto \log(x)^2$. Taylor's Theorem suggests that $\log x = (x-1)^2 + ...$, but as before, we aim for a more rigorous approach. We have

$$|\log(x)^2 - (x-1)^2| = |(\log x - x + 1)(\log x + x - 1)| \leq |\log x - x + 1||\log x + x - 1|. \tag{28}$$

Using Lemma 10, we can show $|\log x - x + 1| \leq |x-1|^2 \max(1, -\log x)$. It is also elementary to check that $|\log x + x - 1| \leq 3|x-1|\max(1, -\log x)$. Hence

$$|\log(x)^2 - (x-1)^2| \leq 3|x-1|^3 \max(1, -\log x)^2. \tag{29}$$

We can now return to computing the conditional expectation of equation 27. We have

$$\mathbb{E}\left[\mathbb{E}\left[\log\left(\frac{1}{m}\sum_{j=1}^m U_j\right)^2\middle|\mathbf{y}_1\right]\right] = \mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^m (U_j - 1)\right)^2\right] + \mathbb{E}[\eta] \tag{30}$$

where our recent result guarantees that

$$|\mathbb{E}[\eta]| \leq \mathbb{E}[|\eta|] \leq 3\mathbb{E}\left[\left|\frac{1}{m}\sum_{j=1}^m (U_j - 1)\right|^3 \max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^m U_j\right)\right)^2\right]. \tag{31}$$

Without reproducing all the details, the approach of Theorem 1 shows us that this error term is $\mathcal{O}(m^{-3/2})$ provided that

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right] < \infty \text{ and } \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2q}\right] < \infty \tag{32}$$

where p, q are Hölder conjugate indices. Theorem 1 also shows that

$$\mathbb{E}\left[\mathbb{E}\left[\log\left(\frac{1}{m}\sum_{j=1}^m U_j\right)\middle|\mathbf{y}_1\right]^2\right] = \mathcal{O}(m^{-2}). \tag{33}$$

Putting these pieces together, we have

$$\mathbb{E}\left[\mathrm{Var}[A_{1,m}|\mathbf{x}_1, \mathbf{y}_1]\right] = \frac{1}{m}\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{p(\mathbf{y})^2}\right] + \mathcal{O}(m^{-3/2}). \tag{34}$$

Turning to the variance of the conditional expectation, recall from Theorem 1 that

$$\mathbb{E}\left[A_{1,m}|\mathbf{x}_1, \mathbf{y}_1\right] = \log\frac{p(\mathbf{y}_1|\mathbf{x}_1)}{p(\mathbf{y}_1)} + \frac{1}{m}\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}_1|\mathbf{x})]}{2p(\mathbf{y}_1)^2} + \mathcal{O}\left(m^{-3/2}\right). \tag{35}$$

Taking the variance gives

$$\mathrm{Var}\left[\mathbb{E}\left[A_{1,m}|\mathbf{x}_1,\mathbf{y}_1\right]\right] = \mathrm{Var}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] + \frac{1}{m}\mathrm{Cov}\left(\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, \frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right) + \mathcal{O}(m^{-3/2}). \tag{36}$$

Thus,

$$\begin{aligned}\mathrm{Var}[A_{1,m}] = {} & \mathrm{Var}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] \\ & + \frac{1}{m}\left(\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{p(\mathbf{y})^2}\right] + \mathrm{Cov}_{p(\mathbf{x},\mathbf{y})}\left(\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, \frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right)\right) + \mathcal{O}(m^{-3/2})\end{aligned} \tag{37}$$

and the full result follows. □

Combining the last two theorems establishes that

$$\mathbb{E}\left[|A_{n,m} - I(\mathbf{x},\mathbf{y})|^2\right] = \mathcal{O}\left(\frac{1}{n} + \frac{1}{m^2}\right). \tag{38}$$

The computational cost of $A_{n,m}$ is $\mathcal{O}(mn)$. Thus it is optimal to set $m \propto \sqrt{n}$. Then the estimator converges to $I(\mathbf{x},\mathbf{y})$ at a rate $T^{-1/3}$ in root mean square, where T is the total computational budget.

Finally, in the case that $p(\mathbf{y}|\mathbf{x})$ is not known, we can repeat this analysis using a positive function $\kappa(\mathbf{x},\mathbf{y})$ in its place. In this case,

$$A^{(\kappa)}_{n,m} \to \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{\kappa(\mathbf{x},\mathbf{y})}{\kappa(\mathbf{y})}\right] \text{ as } m,n\to\infty \tag{39}$$

where $\kappa(\mathbf{y}) = \mathbb{E}_{p(\mathbf{x})}[\kappa(\mathbf{x},\mathbf{y})]$. The same convergence rates apply.

3 Prior Contrastive Estimation and InfoNCE

We now consider the Prior Contrastive Estimation (PCE) estimator (Foster et al., 2020)

$$B_{n,m} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{m+1}\left(p(\mathbf{y}_i|\mathbf{x}_i) + \sum_{j=1}^{m}p(\mathbf{y}_i|\mathbf{x}_{ij})\right)}. \tag{40}$$

where $\mathbf{x}_i,\mathbf{y}_i \overset{\text{i.i.d.}}{\sim} p(\mathbf{x},\mathbf{y})$ and $\mathbf{x}_{ij} \overset{\text{i.i.d.}}{\sim} p(\mathbf{x})$ are independent. We can also re-use samples to make the variant

$$\tilde{B}_n = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{n}\sum_{j=1}^{n}p(\mathbf{y}_i|\mathbf{x}_j)}. \tag{41}$$

It is more common to utilise this estimator in the case that $p(\mathbf{y}|\mathbf{x})$ is not known, leading to the InfoNCE estimator (van den Oord et al., 2018)

$$\tilde{B}^{(\kappa)}_n = \frac{1}{n}\sum_{i=1}^{n}\log\frac{\kappa(\mathbf{x}_i,\mathbf{y}_i)}{\frac{1}{n}\sum_{j=1}^{n}\kappa(\mathbf{x}_j,\mathbf{y}_i)} \tag{42}$$

for some positive function κ. Here, we focus on analysing the estimator $B_{n,m}$.

Before computing the asymptotic expansion of $B_{n,m}$, we present a basic result on its expectation.

Proposition 3 (Bounding the expectation of $B_{n,m}$)**.** *Assume*

$$\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] < \infty. \tag{43}$$

Then,

$$0 \le I(\mathbf{x},\mathbf{y}) - \mathbb{E}[B_{n,m}] \le \frac{1}{m+1}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} - 1\right]. \tag{44}$$

This shows $B_{m,n}$ is negatively biased with bias of order $1/m$.

Proof. See Theorems 1 and 3 of Foster et al. (2020). □

Theorem 4 (Expectation of $B_{n,m}$)**.** *Suppose there exist Hölder conjugate indices $p, q > 0$ with $1/p+1/q = 1$ such that*

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right] < \infty \text{ and } \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{q}\right] < \infty. \tag{45}$$

Then we have

$$\begin{aligned}\mathbb{E}[B_{n,m}] =& I(\mathbf{x},\mathbf{y}) \\ & - \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} - 1\right] + \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{2p(\mathbf{y})^2}\right] \\ & + \mathcal{O}\left(m^{-3/2}\right).\end{aligned} \tag{46}$$

Proof. By linearity, $\mathbb{E}[B_{n,m}] = \mathbb{E}[B_{1,m}]$. To compute this we define U_j as in Theorem 1, and we define $U_0 = p(\mathbf{x}_1|\mathbf{y}_1)/p(\mathbf{y}_1)$. We have

$$\mathbb{E}[B_{1,m}] = I(\mathbf{x},\mathbf{y}) - \mathbb{E}\left[\log\left(\frac{1}{m+1}\sum_{j=0}^{m}U_j\right)\right]. \tag{47}$$

To reduce this to a more manageable form, we have

$$\mathbb{E}\left[\log\left(\frac{1}{m+1}\sum_{j=0}^{m}U_j\right)\right] = \log\left(\frac{m}{m+1}\right) + \mathbb{E}\left[\log\left(1+\frac{U_0}{m}+\frac{1}{m}\sum_{j=1}^{m}(U_j-1)\right)\right] \tag{48}$$

$$= \log\left(\frac{m}{m+1}\right) + \mathbb{E}\left[\log\left(1+\frac{U_0}{m}\right)\right] + \mathbb{E}\left[\log\left(1+\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)\right]. \tag{49}$$

Here, the third term involves a sum of conditionally i.i.d. random variables with mean zero. We now expand this third term with Taylor's Theorem

$$\mathbb{E}\left[\log\left(1+\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)\right] = -\frac{1}{2}\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)^2\right] + \mathbb{E}[\zeta] \tag{50}$$

We focus on controlling the ζ term. By Lemma 10 with $k = 2$ we have

$$|\zeta| \le \left|\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right|^3 \max\left(1, -\log\left(1+\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)\right). \tag{51}$$

Since $U_0 > 0$, we must have

$$\left|\frac{U_j-1}{1+U_0/m}\right| \le |U_j - 1|, \tag{52}$$

thus we can bound $\mathbb{E}[|\zeta|]$ by the exact error term that was considered in Theorem 1. This shows that $\mathbb{E}[|\zeta|] = \mathcal{O}(m^{-3/2})$. To calculate the expectation, we have

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)^2\middle|\mathbf{x}_1,\mathbf{y}_1\right]\right] \tag{53}$$

$$= \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{1}{1+U_0/m}\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right] \tag{54}$$

$$= \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{1}{1+\frac{p(\mathbf{y}|\mathbf{x})}{mp(\mathbf{y})}}\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right], \tag{55}$$

this form offers easy comparison with Theorem 1. However, we have

$$\frac{1}{1+\frac{p(\mathbf{y}|\mathbf{x})}{mp(\mathbf{y})}} = 1 + \mathcal{O}(m^{-1}) \tag{56}$$

and so we can drop the extract factor, giving

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j - 1}{1+U_0/m}\right)^2\right] = \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right] + \mathcal{O}(m^{-2}) \tag{57}$$

We also need to expand

$$\begin{aligned}
\log\left(\frac{m}{m+1}\right) + \mathbb{E}\left[\log\left(1+\frac{U_0}{m}\right)\right] &= \mathbb{E}\left[\log\left(1+\frac{U_0-1}{m+1}\right)\right] && (58)\\
&= \mathbb{E}\left[\frac{U_0-1}{m+1}\right] + \mathcal{O}(m^{-3/2}) && (59)\\
&= \frac{1}{m+1}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} - 1\right] + \mathcal{O}(m^{-3/2}) && (60)\\
&= \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} - 1\right] + \mathcal{O}(m^{-3/2}) \text{ as the difference is order } m^{-2}. && (61)
\end{aligned}$$

Combining these gives the result. □

Theorem 5 (Variance of $B_{m,n}$). *Assume that there exist Hölder conjugate indices $p, q > 0$ such that*

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right] < \infty \textit{ and } \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2q}\right] < \infty. \tag{62}$$

Then,

$$\begin{aligned}
\mathrm{Var}[B_{n,m}] = &\frac{1}{n}\mathrm{Var}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right]\\
&+ \frac{1}{nm}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{2p(\mathbf{y})^2}\right]\\
&+ \frac{1}{nm}\mathrm{Cov}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, -\frac{2p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} + \frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right]\\
&+ \mathcal{O}\left(n^{-1}m^{-3/2}\right).
\end{aligned} \tag{63}$$

Proof. We proceed using the same general strategy as Theorem 2. We have

$$\mathrm{Var}[B_{n,m}] = \frac{1}{n}\mathrm{Var}[B_{1,m}]. \tag{64}$$

By Tower Law,

$$\mathrm{Var}[B_{1,m}] = \mathbb{E}[\mathrm{Var}[B_{1,m}|\mathbf{x}_1, \mathbf{y}_1]] + \mathrm{Var}[\mathbb{E}[B_{1,m}|\mathbf{x}_1, \mathbf{y}_1]]. \tag{65}$$

For the conditional variance, using the notation of Theorem 4 we have

$$\begin{aligned}\mathbb{E}[\mathrm{Var}[B_{1,m}|\mathbf{x}_1,\mathbf{y}_1]] =& \mathbb{E}\left[\mathbb{E}\left[\log\left(1+\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)^2\middle|\mathbf{x}_1,\mathbf{y}_1\right]\right] \\ &- \mathbb{E}\left[\mathbb{E}\left[\log\left(1+\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)\middle|\mathbf{x}_1,\mathbf{y}_1\right]^2\right].\end{aligned} \tag{66}$$

For the first term of this variance, we use the analysis of $x \mapsto \log(x)^2$ that was done in Theorem 2 showing

$$|\log(x)^2 - (x-1)^2| \le |x-1|^3 \max(1, -\log x)^2. \tag{67}$$

Thus,

$$\mathbb{E}\left[\mathbb{E}\left[\log\left(1+\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)^2\middle|\mathbf{x}_1,\mathbf{y}_1\right]\right] = \mathbb{E}\left[\left(1+\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)^2\right] + \mathbb{E}[\nu] \tag{68}$$

where

$$\begin{aligned}|\mathbb{E}[\nu]| \le \mathbb{E}[|\nu|] &\le \mathbb{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right|^3 \max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)\right)^2\right] && (69)\\ &\overset{\text{Hölder}}{\le} \mathbb{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right|^{3p}\right]^{1/p} \mathbb{E}\left[\max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)\right)^{2q}\right]^{1/q} && (70)\\ &\le \mathbb{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m}U_j-1\right|^{3p}\right]^{1/p} \mathbb{E}\left[\max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m}U_j-1\right)\right)^{2q}\right]^{1/q} && (71)\\ &\overset{\text{Corollary 8}}{\le} D_{3p}^{1/p} m^{-3/2}\mathbb{E}\left[|U_1-1|^{3p}\right]^{1/p} \mathbb{E}\left[\max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m}U_j-1\right)\right)^{2q}\right]^{1/q} && (72)\\ &\overset{\text{convexity}}{\le} D_{3p}^{1/p} m^{-3/2}\mathbb{E}\left[|U_1-1|^{3p}\right]^{1/p} \left(1+\mathbb{E}\left[|\log U_1|^{2q}\right]\right)^{1/q} && (73)\\ &\le D_{3p}^{1/p} m^{-3/2}\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right]^{1/p} \left(1+\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2q}\right]\right)^{1/q}. && (74)\end{aligned}$$

We also have, as previously

$$\mathbb{E}\left[\left(1+\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)^2\right] = \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{2p(\mathbf{y})^2}\right] + \mathcal{O}(m^{-2}). \tag{75}$$

On the other hand, Theorem 4 shows that

$$\mathbb{E}\left[\mathbb{E}\left[\log\left(1+\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)\middle|\mathbf{x}_1,\mathbf{y}_1\right]^2\right] = \mathcal{O}(m^{-2}). \tag{76}$$

We can now turn to the variance of the conditional expectation. From Theorem 4, we know

$$\mathbb{E}[B_{1,m}|\mathbf{x}_1,\mathbf{y}_1] = \log\frac{p(\mathbf{y}_1|\mathbf{x}_1)}{p(\mathbf{y}_1)} + \frac{1}{m}\left(1 - \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} + \frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{2p(\mathbf{y})^2}\right) + \mathcal{O}\left(m^{-3/2}\right). \tag{77}$$

Thus,

$$\begin{aligned}\mathrm{Var}[\mathbb{E}[B_{1,m}|\mathbf{x}_1,\mathbf{y}_1]] = {}& \mathrm{Var}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] \\ &+ \frac{1}{m}\mathrm{Cov}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, -\frac{2p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} + \frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right] \\ &+ \mathcal{O}\left(m^{-3/2}\right).\end{aligned} \tag{78}$$

Putting the pieces together gives the final result. □

Finally, we note the key difference between the variance of the NMC and PCE estimators is the term

$$-\frac{1}{nm}\mathrm{Cov}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, \frac{2p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right]. \tag{79}$$

We would expect the covariance between a random variable and its logarithm to be positive, indicating that this term as a whole is negative. This, in turn, suggests that the PCE estimator has a lower variance than its NMC counterpart. However, focusing on the dominant terms, we still have the same overall NMC convergence rate of $T^{-1/3}$ in the total computational budget T.

4 Multi-level Monte Carlo

The following section covers material in Goda et al. (2020a), with Goda et al. (2020b) covering the extension to gradient estimators.

To begin, we define the random variables using the NMC estimator $A_{n,m}$ as our base

$$P_\ell = A_{1,M_\ell} \tag{80}$$

where M_ℓ is an increasing sequence of positive integers. From previous remarks, we know that $\mathbb{E}[P_\ell] \to I(\mathbf{x},\mathbf{y})$ as $\ell \to \infty$. We now take $M_\ell = M_0 2^\ell$. We define the random variables Z_ℓ as follows

$$\begin{aligned}Z_\ell = {}& -\log\left(\frac{1}{M_\ell}\sum_{j=1}^{M_\ell} p(\mathbf{y}_1|\mathbf{x}_{1j})\right) \\ &+ \frac{1}{2}\left[\log\left(\frac{1}{M_{\ell-1}}\sum_{j=1}^{M_{\ell-1}} p(\mathbf{y}_1|\mathbf{x}_{1j})\right) + \log\left(\frac{1}{M_{\ell-1}}\sum_{j=1+M_{\ell-1}}^{M_\ell} p(\mathbf{y}_1|\mathbf{x}_{1j})\right)\right].\end{aligned} \tag{81}$$

The key property of Z_ℓ is

$$\mathbb{E}[Z_\ell] = \mathbb{E}[P_\ell - P_{\ell-1}] \tag{82}$$

and the cost of computing Z_ℓ is bounded by $c2^\ell$. The main technical challenge is to bound the expectation and variance of Z_ℓ. We have the following theorem.

Theorem 6 (Goda et al. (2020a)). *Suppose there exist constants $p, q > 2$ such that $(p-2)(q-2) \geq 4$ such that*

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^p\right] < \infty \quad \textit{and} \quad \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^q\right] < \infty. \tag{83}$$

Then,

$$\mathbb{E}\left[|Z_\ell|\right] = O(2^{-a\ell}), \qquad \mathrm{Var}(Z_\ell) = O(2^{-r\ell}) \tag{84}$$

where $a = \min\left(\frac{p(q-1)}{2q}, 1\right)$, $r = \min\left(\frac{p(q-2)}{2q}, 2\right)$.

Proof. First, define

$$\beta_\ell^{(a)} = \frac{1}{M_\ell}\sum_{j=1}^{M_\ell} \frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)} \tag{85}$$

$$\beta_\ell^{(b)} = \frac{1}{M_\ell}\sum_{j=1+M_\ell}^{M_{\ell+1}} \frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)} \tag{86}$$

$$\text{so}\quad Z_\ell = -\log\beta_\ell^{(a)} + \frac{1}{2}\left(\log\beta_{\ell-1}^{(a)} + \log\beta_{\ell-1}^{(b)}\right) \tag{87}$$

$$\text{and}\quad \beta_\ell^{(a)} = \frac{1}{2}\left(\beta_{\ell-1}^{(a)} + \beta_{\ell-1}^{(b)}\right). \tag{88}$$

We then have

$$\begin{aligned} Z_\ell &= -\log\beta_\ell^{(a)} + \frac{1}{2}\left(\log\beta_{\ell-1}^{(a)} + \log\beta_{\ell-1}^{(b)}\right) && (89)\\ &= -\left(\log\beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right) + \frac{1}{2}\left(\log\beta_{\ell-1}^{(a)} - \beta_{\ell-1}^{(a)} + 1\right) + \frac{1}{2}\left(\log\beta_{\ell-1}^{(b)} - \beta_{\ell-1}^{(b)} + 1\right) && (90)\\ &= 2\left[-\frac{1}{2}\left(\log\beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right) + \frac{1}{4}\left(\log\beta_{\ell-1}^{(a)} - \beta_{\ell-1}^{(a)} + 1\right) + \frac{1}{4}\left(\log\beta_{\ell-1}^{(b)} - \beta_{\ell-1}^{(b)} + 1\right)\right] && (91) \end{aligned}$$

By convexity of $x \mapsto |x|^2$, we have

$$|Z_\ell|^2 \le 2\left|\log\beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right|^2 + \left|\log\beta_{\ell-1}^{(a)} - \beta_{\ell-1}^{(a)} + 1\right|^2 + \left|\log\beta_{\ell-1}^{(b)} - \beta_{\ell-1}^{(b)} + 1\right|^2. \tag{92}$$

We use the following elementary inequality that holds for $1 \le r \le 2$

$$|\log x - x + 1| \le |x-1|^r \max(-\log x, 1) \tag{93}$$

which gives

$$\left|\log\beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right|^2 \le \left|\beta_\ell^{(a)} - 1\right|^{2r}\left(\max\left(-\log\beta_\ell^{(a)}, 1\right)\right)^2. \tag{94}$$

We now take the expectation and apply Hölder's Inequality with $1/s + 1/t = 1$, giving

$$\mathbb{E}\left[\left|\log\beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right|^2\right] \le \left\|\left|\beta_\ell^{(a)} - 1\right|^{2r}\right\|_{L^s}\left\|\left(\max\left(-\log\beta_\ell^{(a)}, 1\right)\right)^2\right\|_{L^t}. \tag{95}$$

For the first term, we apply Corollary 8 to conclude that

$$\left\|\left|\beta_\ell^{(a)} - 1\right|^{2r}\right\|_{L^s} \le D_{2rs}^{1/s}\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2sr}\right]^{1/s}(M_0 2^\ell)^{-r}, \tag{96}$$

for the second term, we use the fact that the functions $x \mapsto \max(-\log x, 1)$ and $x \mapsto x^{2t}$ are convex to give

$$\begin{aligned} \left\|\left(\max\left(-\log\beta_\ell^{(a)}, 1\right)\right)^2\right\|_{L^t} &\le \left\|\left(\frac{1}{M_\ell}\sum_{j=1}^{M_\ell}\max\left(-\log\frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)}, 1\right)\right)^2\right\|_{L^t} && (97)\\ &= \left(\mathbb{E}\left[\left(\frac{1}{M_\ell}\sum_{j=1}^{M_\ell}\max\left(-\log\frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)}, 1\right)\right)^{2t}\right]\right)^{1/t} && (98)\\ &\le \left(\frac{1}{M_\ell}\mathbb{E}\left[\sum_{j=1}^{M_\ell}\max\left(-\log\frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)}, 1\right)^{2t}\right]\right)^{1/t} && (99) \end{aligned}$$

$$\leq \left(\frac{1}{M_\ell}\mathbb{E}\left[\sum_{j=1}^{M_\ell}\left|\log\frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)}\right|^{2t}+1\right]\right)^{1/t} \tag{100}$$

$$= \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2t}+1\right]^{1/t}. \tag{101}$$

We now choose $s = q/(q-2), t = q/2$ and $r = \min(p(q-2)/2q, 2)$. This gives

$$\mathbb{E}\left[\left|\log\beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right|^2\right] \leq A_0 2^{-r\ell} \tag{102}$$

where

$$A_0 = D_{2rs}^{1/s}\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^p\right]^{1/s}\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^q+1\right]^{1/t}M_0^{-r}. \tag{103}$$

Since we can bound the other two terms of (92) in a similar way, we obtain a bound on $\mathrm{Var}(Z_\ell)$ that is of order $2^{-r\ell}$. A similar proof gives the bound for $\mathbb{E}[|Z_\ell|]$. □

An important result of this theorem is that we can obtain a MLMC estimator of $I(\mathbf{x}, \mathbf{y})$ with total cost T that converges at a rate $\mathcal{O}(T^{-1/2})$ in root mean square. This is achieved using standard MLMC technology (Giles, 2008). We define, analogously to the NMC case

$$\begin{aligned} Z_{n,\ell} =& \frac{1}{n}\sum_{i=1}^{n}\left[-\log\left(\frac{1}{M_\ell}\sum_{j=1}^{M_\ell}p(\mathbf{y}_i|\mathbf{x}_{ij})\right)\right. \\ &\left.+\frac{1}{2}\left[\log\left(\frac{1}{M_{\ell-1}}\sum_{j=1}^{M_{\ell-1}}p(\mathbf{y}_i|\mathbf{x}_{ij})\right)+\log\left(\frac{1}{M_{\ell-1}}\sum_{j=1+M_{\ell-1}}^{M_\ell}p(\mathbf{y}_i|\mathbf{x}_{ij})\right)\right]\right]. \end{aligned} \tag{104}$$

Then

$$Z_L^{\mathrm{MLMC}} = \sum_{\ell=0}^{L} Z_{N_\ell,\ell} \tag{105}$$

with

$$\mathbb{E}\left[\left|Z_L^{\mathrm{MLMC}} - I(\mathbf{x},\mathbf{y})\right|^2\right] = \sum_{\ell=0}^{L}\frac{\mathrm{Var}[Z_\ell]}{N_\ell} + \left[\mathbb{E}[P_L] - I(\mathbf{x},\mathbf{y})\right]^2. \tag{106}$$

The cost of the estimator Z_L^{MLMC} is $\mathcal{O}(N_L M_L)$. What Theorem 6 shows is that the bias and variance of the Z_ℓ decay fast enough to offset the growth in cost. For full details, see Goda et al. (2020a).

5 A rigorous delta method for the natural logarithm

This self-contained section includes some of the mathematical machinery that is relied upon by the rest of this work. As previously mentioned, most analyses of mutual information estimators (Zheng et al., 2018; Beck et al., 2018; Rainforth et al., 2018) utilise the delta method for moments. Unfortunately, the standard delta method that we derive here in Lemma 9 is not valid for the natural logarithm function, because none of its derivatives are bounded on $(0, \infty)$. In this section, we derive a rigorous delta method for the logarithm. Whilst this is *not* sufficient for all the Theorems in the preceding sections, it highlights and essentialises the key technical pieces required.

We begin with the Marcinkiewicz–Zygmund Inequality, which is used to derive the standard delta method.

Lemma 7 (Marcinkiewicz and Zygmund (1937))**.** *Let $X_1, \ldots, X_m$ be independent random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{E}\left[|X_i|^p\right] < \infty$. Then there exists a constant D_p such that*

$$\mathbb{E}\left(\left|\sum_{i=1}^{m}(X_i - \mu)\right|^p\right) \leq D_p \mathbb{E}\left(\left(\sum_{i=1}^{m}|X_i|^2\right)^{p/2}\right) \tag{107}$$

Corollary 8. *Let $X_1, \ldots, X_m$ be i.i.d. random variables with $\mathbb{E}[X_1] = \mu$ and $\mathbb{E}\left[|X_1|^p\right] < \infty$. Then there exists a constant D_p such that*

$$\mathbb{E}\left(\left|\frac{1}{m}\sum_{i=1}^{m}(X_i - \mu)\right|^p\right) \leq D_p m^{-p/2} \mathbb{E}\left[|X_1|^p\right] \tag{108}$$

Proof. Applying the Marcinkiewicz–Zygmund Inequality, we have

$$\mathbb{E}\left(\left|\frac{1}{m}\sum_{i=1}^{m}(X_i - \mu)\right|^p\right) \leq D_p m^{-p/2} \mathbb{E}\left(\left(\frac{1}{m}\sum_{i=1}^{m}|X_i|^2\right)^{p/2}\right), \tag{109}$$

by the convexity of $x \mapsto x^{p/2}$ on $(0, \infty)$, we have

$$\leq D_p m^{-p/2} \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}|X_i|^p\right) \tag{110}$$

$$= D_p m^{-p/2} \mathbb{E}\left[|X_1|^p\right]. \tag{111}$$

□

Notice that Corollary 8 essentially gives the asymptotic moments that would be expected from the Central Limit Theorem, although they cannot be derived from the standard Central Limit Theorem which gives convergence *in distribution* only.

Lemma 9 (Delta method of order k)**.** *Let X_i be a sequence of i.i.d. random variables with mean μ and $\mathbb{E}\left[|X_1|^{k+1}\right] < \infty$, and let f be a smooth function with $\|f^{(k+1)}\|_\infty = M < \infty$. Then*

$$\mathbb{E}\left[f\left(\frac{1}{m}\sum_{i=1}^{m}X_i\right)\right] = \sum_{j=0}^{k}\frac{f^{(j)}(\mu)}{j!}\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}(X_i - \mu)\right)^j\right] + \mathcal{O}\left(m^{-(k+1)/2}\right). \tag{112}$$

Proof. By Taylor's Theorem with Lagrange's form of the remainder, we have for any x and for some ξ between x and μ

$$f(x) = \sum_{j=0}^{k}\frac{f^{(j)}(\mu)}{j!}(x - \mu)^k + \frac{f^{(k+1)}(\xi)}{(k+1)!}(x - \mu)^{k+1}. \tag{113}$$

Applying this to $\frac{1}{m}\sum_{i=1}^{m}X_i$ and taking the expectation gives

$$\mathbb{E}\left[f\left(\frac{1}{m}\sum_{i=1}^{m}X_i\right)\right] = \sum_{j=0}^{k}\frac{f^{(j)}(\mu)}{j!}\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}(X_i - \mu)\right)^j\right] + \mathbb{E}\left[\frac{f^{(k+1)}(\Xi)}{(k+1)!}\left(\frac{1}{m}\sum_{i=1}^{m}(X_i - \mu)\right)^{k+1}\right] \tag{114}$$

where Ξ is a random variable between μ and $\frac{1}{m}\sum_i X_i$. By assumption, we have $f^{(k+1)}(\Xi) \leq M$. By Corollary 8, we have

$$\mathbb{E}\left(\left|\sum_{i=1} X_i - \mu\right|^{k+1}\right) \leq D_{k+1} m^{(k+1)/2} \mathbb{E}\left[|X_1|^{k+1}\right]. \tag{115}$$

Hence we conclude that

$$\left|\mathbb{E}\left[\frac{f^{(k+1)}(\Xi)}{(k+1)!}\left(\frac{1}{m}\sum_{i=1}^{m}(X_i-\mu)\right)^{k+1}\right]\right| \leq \frac{MD_{k+1}\mathbb{E}\left[|X_1|^{k+1}\right]m^{-(k+1)/2}}{(k+1)!} = \mathcal{O}(m^{-(k+1)/2}). \tag{116}$$

□

We now turn to the logarithm function in particular, bounding the difference between the function and its series approximation.

Lemma 10. *Define*

$$L_k(x) = \sum_{j=1}^{k}\frac{(-1)^{j+1}}{j}(x-1)^j. \tag{117}$$

Then $|\log x - L_k(x)| \leq |x-1|^{k+1}\max(1, -\log x)$ *for* $0 < x < \infty$.

Proof. By Taylor's Theorem with Cauchy's form of the remainder, for any $0 < x < \infty$ there exists ξ that is between 1 and x such that

$$\log x = L_k(x) + \frac{(-1)^{k+2}}{\xi^{k+1}}(x-\xi)^k(x-1) \tag{118}$$

For $x > 1$, we must have $\xi^{k+1} > 1$, so $|\log x - L_k(x)| < |x-\xi|^k|x-1| < |x-1|^{k+1}$.

For $x \leq 1$, we have

$$\frac{\xi - x}{\xi} = 1 - x/\xi \text{ and } 0 \leq 1 - x/\xi \leq 1 - x \text{ since } x \leq \xi \leq 1. \tag{119}$$

Thus, the magnitude of the remainder term becomes

$$\left|\frac{(-1)^{k+2}}{\xi^{k+1}}(x-\xi)^k(x-1)\right| = \left|\left(\frac{\xi-x}{\xi}\right)^k\frac{x-1}{\xi}\right| \leq (1-x)^k\left|\frac{x-1}{\xi}\right| \leq \frac{(1-x)^{k+1}}{x} \tag{120}$$

which shows that the Taylor series for the logarithm is convergent on $(0,1]$. Therefore, we have

$$\log x - L_k(x) = \sum_{j=k+1}^{\infty}\frac{(-1)^{j+1}}{j}(x-1)^j \tag{121}$$

$$= (x-1)^{k+1}(-1)^k\left(\frac{1}{k+1} - \sum_{j=1}^{\infty}\frac{(-1)^{j+1}}{k+1+j}(x-1)^j\right) \tag{122}$$

noting that $x - 1 \leq 0$ we see that each term of the sum has the same sign, giving

$$= -|x-1|^{k+1}\left(\frac{1}{k+1} + \sum_{j=1}^{\infty}\frac{1}{k+1+j}|x-1|^j\right). \tag{123}$$

If $x \geq e^{-1}$, we have

$$\frac{1}{k+1} + \sum_{j=1}^{\infty}\frac{1}{k+1+j}|x-1|^j \leq \frac{1}{k+1} + \sum_{j=1}^{\infty}\frac{1}{k+1+j}|e-1|^j \tag{124}$$

by monotonicity. If $x \leq e^{-1}$, we have

$$\frac{1}{k+1} + \sum_{j=1}^{\infty}\frac{1}{k+1+j}|x-1|^j \leq \frac{1}{k+1} + \frac{|x-1|}{k+2} + \sum_{j=2}^{\infty}\frac{|x-1|^j}{k+1+j} \tag{125}$$

$$\leq |x-1| + \sum_{j=2}^{\infty} \frac{|x-1|^j}{j} = -\log x, \tag{126}$$

for any $k \geq 1$. Combining these, we have

$$\frac{1}{k+1} + \sum_{j=1}^{\infty} \frac{1}{k+1+j} |x-1|^j \leq \max(-\log x, \log e) = \max(-\log x, 1). \tag{127}$$

□

For the following Proposition, the logic is inspired by Goda et al. (2020a).

Proposition 11 (Rigorous delta method for the logarithm)**.** *Let $U_1, \ldots, U_m$ be a sequence of i.i.d. positive random variables with $\mathbb{E}[U_1] = 1$. Fix a natural number $k \geq 1$. Suppose that for Hölder conjugate indices $p, q > 0$ with $1/p + 1/q = 1$, we have $\mathbb{E}\left[U_1^{(k+1)p}\right] < \infty$ and $\mathbb{E}[|\log U_1|^q] < \infty$. Then,*

$$\mathbb{E}\left[\log\left(\frac{1}{m}\sum_{i=1}^m U_i\right)\right] = \sum_{j=2}^{k} \frac{(-1)^{j+1}}{j} \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^m (U_i - 1)\right)^j\right] + E_k \tag{128}$$

where $E_k = \mathcal{O}\left(m^{-(k+1)/2}\right)$.

Proof. Define L_k as in Lemma 10. By that Lemma, we have

$$\left|\log\left(\frac{1}{m}\sum_{i=1}^m U_i\right) - L_k\left(\frac{1}{m}\sum_{i=1}^m U_i\right)\right| \leq \left|\frac{1}{m}\sum_{i=1}^m U_i\right|^{k+1} \max\left(-\log\left(\frac{1}{m}\sum_{i=1}^m U_i\right), 1\right). \tag{129}$$

We see that

$$\mathbb{E}\left[L_k\left(\frac{1}{m}\sum_{i=1}^m U_i\right)\right] = \sum_{j=1}^{k} \frac{(-1)^{j+1}}{j} \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^m (U_i - 1)\right)^j\right] \tag{130}$$

and $\mathbb{E}[U_i - 1] = 0$.

The error term E_k is bounded in L_1 by

$$\mathbb{E}[|E_k|] \leq \mathbb{E}\left[\left|\frac{1}{m}\sum_{i=1}^m U_i\right|^{k+1} \max\left(-\log\left(\frac{1}{m}\sum_{i=1}^m U_i\right), 1\right)\right] \tag{131}$$

apply Hölder's Inequality to give

$$\leq \mathbb{E}\left[\left|\frac{1}{m}\sum_{i=1}^m U_i\right|^{p(k+1)}\right]^{1/p} \mathbb{E}\left[\max\left(-\log\left(\frac{1}{m}\sum_{i=1}^m U_i\right), 1\right)^q\right]^{1/q}. \tag{132}$$

For the first term, Corollary 8 shows that

$$\mathbb{E}\left[\left|\frac{1}{m}\sum_{i=1}^m U_i\right|^{p(k+1)}\right]^{1/p} \leq D_{(k+1)p}^{1/p} m^{-(k+1)/2} \mathbb{E}\left[U_1^{(k+1)p}\right]^{1/p} \tag{133}$$

for the second term we use the fact that $x \mapsto \max(-\log x, 1)$ is a convex function, so

$$\max\left(-\log\left(\frac{1}{m}\sum_{i=1}^m U_i\right), 1\right)^q \leq \frac{1}{m}\sum_{i=1}^m \max\left(-\log\left(U_i\right), 1\right)^q \tag{134}$$

$$\leq \frac{1}{m} \sum_{i=1}^{m} \left(|\log U_i| + 1 \right)^q , \tag{135}$$

hence

$$\mathbb{E}\left[\max\left(-\log\left(\frac{1}{m} \sum_{i=1}^{m} U_i \right), 1 \right)^q \right]^{1/q} \leq \left(\mathbb{E}\left[|\log U_1|^q \right] + 1 \right)^{1/q} . \tag{136}$$

By assumption, we have $\mathbb{E}\left[U_1^{(k+1)p} \right] < \infty$ and $\mathbb{E}\left[|\log U_1|^q \right] < \infty$. Putting the pieces together, we have

$$\mathbb{E}[|E_k|] \leq m^{-(k+1)/2} \, D_{(k+1)p}^{1/p} \mathbb{E}\left[U_1^{(k+1)p} \right]^{1/p} \left(\mathbb{E}\left[|\log U_1|^q \right] + 1 \right)^{1/q} , \tag{137}$$

so E_k is $\mathcal{O}\left(m^{-(k+1)/2} \right)$ as required. □

Notice that we recover the regular delta method with $\log U_i$ bounded if $p = 1, q = \infty$.

References

Joakim Beck, Ben Mansour Dia, Luis FR Espath, Quan Long, and Raul Tempone. Fast bayesian experimental design: Laplace-based importance sampling for the expected information gain. *Computer Methods in Applied Mechanics and Engineering*, 334:523–553, 2018.

Adam Foster, Martin Jankowiak, Matthew O'Meara, Yee Whye Teh, and Tom Rainforth. A unified stochastic gradient approach to designing bayesian-optimal experiments. volume 108 of *Proceedings of Machine Learning Research*, pages 2959–2969, Online, 26–28 Aug 2020. PMLR.

Michael B Giles. Multilevel monte carlo path simulation. *Operations research*, 56(3):607–617, 2008.

Takashi Goda, Tomohiko Hironaka, and Takeru Iwamoto. Multilevel Monte Carlo estimation of expected information gains. *Stochastic Analysis and Applications*, 38(4):581–600, 2020a.

Takashi Goda, Tomohiko Hironaka, Wataru Kitade, and Adam Foster. Unbiased MLMC stochastic gradient-based optimization of Bayesian experimental designs. *arXiv preprint arXiv:2005.08414*, 2020b.

Aapo Hyvärinen. Survey on independent component analysis. 1999.

Nojun Kwak and Chong-Ho Choi. Input feature selection by mutual information based on parzen window. *IEEE transactions on pattern analysis and machine intelligence*, 24(12):1667–1671, 2002.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

Józef Marcinkiewicz and Antoni Zygmund. Quelques théoremes sur les fonctions indépendantes. *Fund. Math*, 29:60–90, 1937.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180, 2019.

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

Kenneth J Ryan. Estimating expected information gains for experimental designs with application to the random fatigue-limit model. *Journal of Computational and Graphical Statistics*, 12(3):585–603, 2003.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

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

Sue Zheng, Jason Pacheco, and John Fisher. A robust approach to sequential information theoretic planning. In *International Conference on Machine Learning*, pages 5941–5949, 2018.