1 Bayesian experimental design for model selection: variational and classification approaches

1.1 Introduction

Bayesian experimental design for model selection is a important and well-studied problem (Cavagnaro et al., 2010; Vanlier et al., 2014; Hainy et al., 2018). In this essay, we tackle two questions that are relevant to this problem. First, how do recently proposed variational methods for experimental design (Foster et al., 2019, 2020) translate into the model selection context? Second, how do these methods intersect with recently proposed classification-driven approaches to experimental design for model selection (Hainy et al., 2018)?

We begin by elucidating the key features of the model selection problem—it turns out that we can characterise the set-up as a semi-implicit model with a discrete latent variable of interest. The posterior or Barber–Agakov approach of Foster et al. (2019) involves training an amortised inference network from data simulated from the model. We find that, for model selection, this network is exactly a (neural) classifier that predicts the true model that synthesised an observation from that synthetic experimental observation. The marginal + likelihood method of Foster et al. (2019) also translates into the model selection case. This method involves variational density estimation of experimental outcomes for each possible model. In other words, it involves approximating the model evidence of the data for each possible model. Finally, we examine how the stochastic gradient design approach of Foster et al. (2020) applies here. This approach can build off the back of the Barber–Agakov bound, so it also utilises a classifier. The key difference here is that we differentiate the classifier output with respect to its input to learn the design at the same time as the classifier network parameters. This bears some similarities with adversarial approaches to neural network robustness (Carlini et al., 2019). Finally, we compare and contrast the variational approach with other classification driven approaches in the literature.

1.2 Characterising the problem

We denote experimental designs by ξ and experimental observations as y. Suppose there are K competing models $\{m_1, \dots, m_k\}$ and we have a prior distribution $p(m)$ on which model we think is likely to be correct. Given the choice of model, there are other model parameters $\psi \sim p(\psi|m)$. Conditional on the model, and on its parameters, we have a likelihood for the experiment $p(y|m, \psi, \xi)$ which we assume is known in closed form.

One important feature of the model selection problem is that we do *not* have a likelihood that directly relates the design ξ, observation y and the latent variable of interest m. Instead, we have to account for the auxiliary latent variable ψ. Indeed, we actually have $p(y|m, \xi) = \int_\Psi p(y|m, \psi, \xi)p(\psi|m)d\psi$. This case, where we have a closed form likelihood but for a larger set of variable, is referred to as a *semi-implicit* model.

In this essay, we focus on experimental design with the expected information gain (EIG) criterion, also called mutual information utility, that aims to reduce Shannon entropy in our beliefs about m. The EIG-optimal design is specifically,

$$\xi^* = \arg\max_\xi \mathbb{E}_{p(m)p(\psi|m)p(y|m,\psi,\xi)} \left[\log \frac{p(m|y, \xi)}{p(m)} \right]. \tag{1}$$

Finding ξ^* amounts to estimating the EIG objective function and optimising over the space of possible designs.

If we have already observed some data $\mathcal{D} = \{(\xi_1, y_1), \dots, (\xi_T, y_T)\}$, then we fit model-specific posteriors for the auxiliary variable ψ for each model $p(\psi|m, \mathcal{D})$, and we compute the posterior over models $p(m|\mathcal{D}) \propto p(m)p(\mathcal{D}|m)$. Thus, we update our priors $p(m)$ and $p(\psi|m)$ on the basis of past data.

1.3 The variational approach

1.3.1 Posterior lower bound

Foster et al. (2019) considered variational estimation of the EIG. Their general strategy was to optimise variational upper or lower bounds on the EIG. Their simplest bound was the posterior lower bound (also called the Barber–Agakov bound after Barber and Agakov (2003)). With the variables we have in this model, the bound would be expressed as

$$\mathbb{E}_{p(m)p(\psi|m)p(y|m,\psi,\xi)}\left[\log\frac{p(m|y,\xi)}{p(m)}\right] \geq \mathbb{E}_{p(m)p(\psi|m)p(y|m,\psi,\xi)}\left[\log\frac{q_\phi(m|y)}{p(m)}\right]. \tag{2}$$

The new term $q_\phi(m|y,\xi)$ was generically referred to as the *amortised approximate posterior* with variational parameters ϕ. It is an approximate posterior distribution on the latent variable m of interest. The amortisation here refers to the fact that we learn a function from y to a distribution over m (for different ξ, we would train separate functions). For the model selection approach, then, q_ϕ is a function from y to a distribution over the discrete model indicator m. First, since m is discrete, the choice of variational family is moot, because every distribution over m can be finitely represented. Second, q_ϕ has a very simple interpretation. It is a classifier that attempts to predict, on the basis of input y, which model of $m_1, \ldots, m_k$ generated that data, specifically trying to estimate the posterior probability $p(m|y,\xi)$ over the k different possibilities for m. Importantly though, rather than just attempting to predict the correct model that was responsible for generating the data y, it is essential that we have a *probabilistic* classifier that assigns probabilities to each possible model. For this probabilistic classifier, the issue of calibration becomes central, as we hope that our classifier probabilities will approach $p(m|y,\xi)$ during training.

We have established that q_ϕ is simply a probabilistic classifier for the model selection case. How should this classifier be trained? In general, Foster et al. (2019) proposed training q_ϕ by stochastic gradient methods (Robbins and Monro, 1951; Kingma and Ba, 2014) to maximise the lower bound with respect to ϕ

$$\phi^* = \arg\max_\phi \mathbb{E}_{p(m)p(\psi|m)p(y|m,\psi,\xi)}\left[\log\frac{q_\phi(m|y)}{p(m)}\right] \tag{3}$$

In model selection, training ϕ simply means training the parameters of the classifier. Maximising the posterior lower bound is equivalent to simply maximising the expected log likelihood under q, i.e.

$$\phi^* = \arg\max_\phi \mathbb{E}_{p(m)p(\psi|m)p(y|m,\psi,\xi)}\left[\log q_\phi(m|y)\right]. \tag{4}$$

This is true because $p(m)$ has no dependence on ϕ. So, we see that training q_ϕ to maximise the variational posterior lower bound amounts to maximum likelihood training of a neural classifier when we are in the setting of model selection. (Care may be needed to ensure the classifier produces good *probabilistic uncertainty*, as well as getting good predictions, as these probabilities are central to our method.)

In fact, we have an enhanced setting in which we can draw an infinite amount of training data by simulating from $p(m)p(\psi|m)p(y|m,\psi,\xi)$. To do this, we sample a random model m from its prior, then a random set of parameters $\psi \sim p(\psi|m)$ for the chosen model, and then simulate an experimental outcome under design ξ. Importantly, we do not need to draw a fixed training or test set, and we never need to show the classifier the same examples twice, we instead draw new batches on the fly. One particularly important consequence of this is that the spectre of *over-fitting* is much reduced in our case, as there is no fixed training set to overfit to.

We now see another important point—the negative log-likelihood loss of the classifier is essentially an estimate of the EIG, up to a constant. Suppose we have completed training and reached parameters $\hat{\phi}$. Then the EIG estimate is

$$\mathrm{EIG}(\xi) \approx \mathbb{E}_{p(m)p(\psi|m)p(y|m,\psi,\xi)}\left[\log\frac{q_{\hat{\phi}}(m|y)}{p(m)}\right] = \mathbb{E}_{p(m)p(\psi|m)p(y|m,\psi,\xi)}\left[\log q_{\hat{\phi}}(m|y)\right] + H[p(m)] \tag{5}$$

and we can estimate the expectation with new, independent batches simulated from the model.

In summary, the posterior lower bound method for model selection amounts to training a classifier on (infinite) simulated data to predict m from y. The optimal design ξ^* will be approximated by the classifier which has the best (lowest) validation loss, which is a good approximation of having the highest EIG.

1.3.2 Marginal + likelihood estimator

The posterior lower bound is not the only way to estimate the EIG proposed by Foster et al. (2019). Both the marginal and the VNMC methods require an explicit likelihood, so they are not suitable for the semi-implicit model selection scenario. The marginal + likelihood estimator is

$$\mathrm{EIG}(\xi) \approx \mathbb{E}_{p(m)p(\psi|m)p(y|m,\psi,\xi)}\left[\log \frac{q_\ell(y|m,\xi)}{q_p(y|\xi)}\right]. \tag{6}$$

This estimator translates, with some simplification, into the model selection setting. The 'approximate likelihood' $q_\ell(y|m,\xi)$ in the model selection setting is an approximation of the model evidence $q_\ell(y|m,\xi) \approx p(y|m,\xi)$. For model selection when m is discrete, we do not need to separately estimate q_p and q_ℓ, we can instead sum over m to obtain

$$q_p(y|\xi) = \sum_m p(m)q_\ell(y|m,\xi). \tag{7}$$

As shown in Appendix A.4 of Foster et al. (2019), the estimator actually becomes a lower bound

$$\mathrm{EIG}(\xi) \geq \mathbb{E}_{p(m)p(\psi|m)p(y|m,\psi,\xi)}\left[\log \frac{q_\ell(y|m,\xi)}{\sum_{m'} p(m')q_\ell(y|m',\xi)}\right] \tag{8}$$

on the EIG in this case, which is not generally the case for the marginal + likelihood method. (In fact, this lower bound is itself a special case of the likelihood-free ACE lower bound introduced in Foster et al. (2020). Indeed, if we take the prior as the variational posterior and let $L \to \infty$ in the LF-ACE bound, we recover this lower bound.)

This lower bound also has a nice interpretation in the model selection scenario. The best design will be the one where the lower bound is largest, which happens, loosely speaking, when $q_\ell(y|m,\xi)$ is much larger than $\sum_m p(m)q_\ell(y|m,\xi)$. That means the approximate model evidence for the observation y under the correct model m is much larger than its evidence under other models. Thus, using the experiment with design ξ and observing y will allow us to easily discriminate between models.

To explicitly use this method, we need to choose trainable density estimators for $q_\ell(y|m,\xi;\phi)$ with parameters ϕ. The simplest method would be to have a distinct set of variational parameters for each value of m and ξ. Whilst it is possible to use a Gaussian density model, we could use more sophisticated methods such as normalising flows (Rezende and Mohamed, 2015). The training approach is similar to that for the posterior method. We use infinite simulated data, and maximise the variational lower bound using stochastic gradient optimisers.

The last two sections highlight a general feature of the variational methods of Foster et al. (2019)—we can either make variational approximations to densities over m or over y. Both lead to valid bounds.

1.4 Stochastic gradient optimisation of the design

So far, we have focused on variational estimation of the EIG. As shown in Foster et al. (2020), it is only a short jump from variational estimation of the EIG to stochastic gradient optimisation of the design using a variational lower bound on EIG. The benefit here, of course, is that we do not have conduct a grid search, co-ordinate exchange or similar algorithm over the design space. What we require instead is a continuous design space and the ability to differentiate observations with respect to designs.

Whilst Foster et al. (2020) focused on explicit likelihood models, both the posterior (Barber–Agakov) lower bound and the LF-ACE bound are applicable to the semi-implicit model selection setting. There is just one thing to check, which is that we can compute a derivative $\partial y/\partial\xi$. In the semi-implicit case, this is often fine.

For example, if $p(y|m,\psi,\xi)$ takes the form $y = g(m,\psi,\xi,\epsilon)$ for a differentiable g and an independent noise random variable ϵ.

Assuming this is the case, we can train ξ by stochastic gradient using either the posterior bound or the simplified LF-ACE bound that was derived in equation (8). We focus on the posterior lower bound for simplicity. Recall that, for the posterior bound, we are training a classifier to predict m from y. We have

$$\mathrm{EIG}(\xi) \geq \mathbb{E}_{p(m)p(\psi|m)p(y|m,\psi,\xi)}\left[\log q_\phi(m|y)\right] + H[p(m)] \tag{9}$$

where q_ϕ is the classifier. One thing that we skimmed over slightly in the previous section was that ϕ implicitly depends on ξ via the training data, and different ξ will have different classifiers with different optimal values of the classifier parameters ϕ.

In Foster et al. (2020), rather than training separate classifiers with different designs ξ, we update ξ and ϕ togther in one stochastic gradient optimisation over the combined set of variables (ξ,ϕ). To explicitly write down the ξ gradient here, let's assume that we do have $y = g(m,\psi,\xi,\epsilon)$, so we can write

$$\mathcal{L}(\xi,\phi) = \mathbb{E}_{p(m)p(\psi|m)p(\epsilon)}\left[\log q_\phi(m|g(m,\psi,\xi,\epsilon))\right] + H[p(m)]. \tag{10}$$

In this form, the ξ gradient can be simply calculated as

$$\frac{\partial \mathcal{L}}{\partial \xi} = \mathbb{E}_{p(m)p(\psi|m)p(\epsilon)}\left[\left.\frac{\partial \log q_\phi}{\partial y}\right|_{m,g(m,\psi,\xi,\epsilon)} \left.\frac{\partial g}{\partial \xi}\right|_{m,\psi,\xi,\epsilon}\right]. \tag{11}$$

The beauty of modern auto-diff frameworks, of course, means that we do not even need to calculate this explicitly ourselves.

For model selection, equation (11) has a natural interpretation. We want to increase the lower bound $\mathcal{L}$ by moving to regions in which the classifier can confidently predict the correct model label m. This corresponds to moving y into regions in which $\log q_\phi(m|y)$ is larger *for the model that actually generated* y. In other words, we want the input to the classifier y to be pushed to regions where the classifier already finds it easy to classify correctly. That is, regions where deciding which model is correct is easier. We then exploit the differentiable relationship between ξ and y, and use this signal to 'improve' the input to the classifier by adjusting the design ξ to that such datasets y are more likely to be synthesised.

At the same time, we are constantly making gradient updates on the classifier parameters ϕ. This means that, as the distribution of (m,y) changes, the classifier can adjust accordingly.

If this sounds dubious, it is worth taking a step back. We are quite simply optimising the lower bound $\mathcal{L}(\xi,\phi)$ jointly with respect to ξ and ϕ, in the hopes that this global maximum may closely correspond to the EIG maximiser ξ^*. We actually have a guarantee that the value of $\mathcal{L}$ at our final trained variables $\hat{\xi},\hat{\phi}$ is a lower bound on $\mathrm{EIG}(\hat{\xi})$, i.e. the true value of $\hat{\xi}$ cannot be worse than the value we estimate for it.

Whilst the method is approximate, because we cannot quantify the discrepancy between $\mathcal{L}$ and the true EIG, it is highly scalable to very large design spaces. Other bounds presented in Foster et al. (2020) have the added benefit that they become equal to the EIG in a limit, providing some assurances that the global maximum of $\mathcal{L}$ is a good design. Foster et al. (2020) also introduced the evaluation method of establishing *lower and upper* bounds on chosen designs. This numerically bounds the discrepancy between the training objective $\mathcal{L}$ and the true EIG objective. Sadly, the upper bounds are only valid for explicit likelihood models; they don't work in the semi-implicit model selection case.

Finally, all of the above discussion carries over if we were to use the lower bound of equation (8) instead of the posterior bound.

1.5 Comparing with other classification approaches

We have established that the variational posterior approach of Foster et al. (2019) instructs us to learn a classifier to predict m from y and use the log probabilities $q_\phi(m|y)$ to estimate EIG. Other authors have considered supervised classification as a means to perform Bayesian experimental design for model selection.

Here, we focus on Hainy et al. (2018), which is "the first approach using supervised learning methods for optimal Bayesian design." This method trains a classifier that predicts m using y, with separate classifiers for different ξ. They focus on training decision trees and random forest classifiers (Breiman, 2001). Since random forests are not generally trained by stochastic gradient methods, this means that they fall back on simulating fixed training and test datasets of samples $(m_j, y_j)_{j=1}^J$ from $p(m)p(\psi|m)p(y|m,\psi,\xi)$. The training dataset is used to train the classifier model, whilst the test dataset gives unbiased estimates of the posterior loss. There is a danger that the classifier may overfit to the training set in this case. Compare this with the training of stochastic gradient classifiers in our previous sections—here we can draw fresh training batches on the fly, and avoid overfitting to a training set.

Decision trees and random forests do provide estimates of the class probabilities $q(m|y)$, but they are relatively noisy. For this reason, Hainy et al. (2018) focus on the 0–1 loss to evaluate designs. In the language of classification, therefore, they choose the design which gives the best *test accuracy*. Again, this is different to the variational approach which fits a neural classifier that automatically provides smooth probability estimates $q_\phi(m|y)$. The latter case was applied to estimate the information gain, which we showed is equivalent to choosing the design which gives the best *test loss*, assuming a negative log-likelihood loss function.

The trade-offs between these methods are clear when we consider optimising over a large design space. For the variational method, we have to train a number of neural networks to convergence. For the classification approach of Hainy et al. (2018), we train a number of random forest classifiers—this may be significantly more computationally efficient. Hainy et al. (2018) propose embedding their 0–1 loss estimation within a co-ordinate exchange algorithm (Meyer and Nachtsheim, 1995) to optimise over designs. The variational method, on the other hand, can naturally be embedded in a unified stochastic gradient optimisation to find the optimal design through stochastic gradient optimisation. The former may be more effective when the design space is not continuous, the latter can work well in a high-dimensional design space that is difficult to search using discrete methods.

2 Bayesian active learning by disagreement and Bayesian experimental design

2.1 Introduction

The purpose of this essay is to highlight the connection between the Bayesian Active Learning by Disagreement (BALD) score as estimated by Gal et al. (2017) and the Prior Contrastive Estimation (PCE) bound of Foster et al. (2020). There is a deep connection between Bayesian experimental design and Bayesian active learning. A significant touchpoint is the use of the mutual information score (Lindley, 1956)

$$I(\xi) = \mathbb{E}_{p(\theta)p(y|\theta,\xi)}\left[H[p(\theta)] - H[p(\theta|y,\xi)]\right]. \tag{12}$$

to acquire new information in a Bayesian model with parameters θ where, y is the as yet unobserved outcome, and ξ is the design to be chosen.

2.2 Bayesian Active Learning by Disagreement

One of the computational challenges inherent in estimating equation (12) directly is that it involves repeated estimation of posterior distributions $p(\theta|y,\xi)$ for different simulated observations y. To remove this particular bottleneck, Houlsby et al. (2011) introduced a rewriting of the mutual information score using Bayes rule

$$I(\xi) = H[p(y|\xi)] - \mathbb{E}_{p(\theta)}\left[H[p(y|\theta,\xi)]\right]. \tag{13}$$

Whilst this is exactly equal to the original mutual information score, the new way of expressing I removes the requirement to estimate posterior distributions over θ. They termed equation (13) the Bayesian Active Learning by Disagreement (BALD) score.

Unfortunately, the story does not end with the BALD score because it still typically involves some intractable computations that must be estimated. For example, Houlsby et al. (2011) focused on approximations for Gaussian Process models (Williams and Rasmussen, 2006).

The more recent work by Gal et al. (2017) estimated the BALD score in the context of Bayesian deep learning classifiers. In such a model, θ represents the parameters of a classification model, and $p(y|\theta,\xi)$ is a probability distribution over classes $y \in \{c_1, \dots, c_k\}$. Computing $p(y|\theta,\xi)$ involves a forward pass through the classifier with input ξ and parameters θ, the network generally ends in a softmax activation to produce a normalised distribution. To sample different values of θ, Gal et al. (2017) employed Monte Carlo Dropout (Gal and Ghahramani, 2016). Given independent samples $\theta_1, \dots, \theta_M$ from $p(\theta)$, they proposed the following Deep BALD (DBALD) estimator of $I(\xi)$

$$I(\xi) \approx \hat{I}_{\text{DBALD}}(\xi) = H\left[\frac{1}{M}\sum_{i=1}^{M} p(y|\theta_i,\xi)\right] - \frac{1}{M}\sum_{i=1}^{M} H[p(y|\theta_i,\xi)] \tag{14}$$

where $H[P(y)] = -\sum_c P(y=c)\log P(y=c)$.

Notation For comparison with the original paper, we used θ in place of $\boldsymbol{\omega}$, ξ in place of $\mathbf{x}$, M in place of T and $p(\theta)$ is used in place of $q_\theta^*(\boldsymbol{\omega})$.

2.3 Prior Contrastive Estimation

In the context of stochastic gradient optimisation of Bayesian experimental designs, Foster et al. (2020) also considered the mutual information score $I(\xi)$ and the rearrangement equation (13). They proved the following Prior Contrastive Estimation (PCE) lower bound on $I(\xi)$

$$I(\xi) \geq \mathbb{E}_{p(\theta_0)p(y|\theta_0,\xi)p(\theta_1)\dots p(\theta_L)}\left[\log \frac{p(y|\theta_0,\xi)}{\frac{1}{L+1}\sum_{\ell=0}^{L} p(y|\theta_\ell,\xi)}\right] \tag{15}$$

and used this bound to optimise ξ by stochastic gradient. One approach to estimate this bound using finite samples is the estimator

$$\hat{I}_{\text{PCE-naive}}(\xi) = \frac{1}{M}\sum_{m=1}^{M}\log\frac{p(y_m|\theta_{m0},\xi)}{\frac{1}{L+1}\sum_{\ell=0}^{L}p(y_m|\theta_{m\ell},\xi)}. \tag{16}$$

where $y_m, \theta_{m0} \sim p(y,\theta|\xi)$ and $\theta_{m\ell} \sim p(\theta)$ for $\ell \geq 1$. However, we can also re-use samples more efficiently to give the estimator

$$\hat{I}_{\text{PCE}}(\xi) = \frac{1}{M}\sum_{m=1}^{M}\log\frac{p(y_m|\theta_m,\xi)}{\frac{1}{M}\sum_{\ell=1}^{M}p(y_m|\theta_\ell,\xi)}. \tag{17}$$

where $y_m, \theta_m \sim p(y,\theta|\xi)$. (To check the expectation of this version matches the PCE bound with $L = M-1$, we simply move the $\mathbb{E}$ sign inside of the summation.) Finally, Foster et al. (2020) discussed a speed-up that is possible when y is a discrete random variable taking values in $\{c_1,\dots,c_k\}$. In this case, we can integrate out y by summing over it, rather than by drawing random samples of y. This method, called Rao-Blackwellisation, results in the estimator

$$\hat{I}_{\text{PCE-RB}}(\xi) = \frac{1}{M}\sum_{m=1}^{M}\sum_{c}p(y=c|\theta_m,\xi)\log\frac{p(y=c|\theta_m,\xi)}{\frac{1}{M}\sum_{\ell=1}^{M}p(y=c|\theta_\ell,\xi)}. \tag{18}$$

2.4 PCE and DBALD equivalence

We have looked at two parallel ways of approximating $I(\xi)$. The interesting result is that *the Rao-Blackwellised PCE estimator and the DBALD estimator are the same.* We can see this by direct calculation

$$\begin{aligned}
\hat{I}_{\text{PCE-RB}}(\xi) &= \frac{1}{M}\sum_{m=1}^{M}\sum_{c}p(y=c|\theta_m,\xi)\log\frac{p(y=c|\theta_m,\xi)}{\frac{1}{M}\sum_{\ell=1}^{M}p(y=c|\theta_\ell,\xi)} && (19)\\
&= \frac{1}{M}\sum_{m=1}^{M}\sum_{c}p(y=c|\theta_m,\xi)\log p(y=c|\theta_m,\xi) \\
&\quad - \frac{1}{M}\sum_{m=1}^{M}\sum_{c}p(y=c|\theta_m,\xi)\log\left(\frac{1}{M}\sum_{\ell=1}^{M}p(y=c|\theta_\ell,\xi)\right) && (20)\\
&= -\frac{1}{M}\sum_{m=1}^{M}H[p(y|\theta_m,\xi)] - \frac{1}{M}\sum_{m=1}^{M}\sum_{c}p(y=c|\theta_m,\xi)\log\left(\frac{1}{M}\sum_{\ell=1}^{M}p(y=c|\theta_\ell,\xi)\right) && (21)\\
&= -\frac{1}{M}\sum_{m=1}^{M}H[p(y|\theta_m,\xi)] - \sum_{c}\left(\frac{1}{M}\sum_{m=1}^{M}p(y=c|\theta_m,\xi)\right)\log\left(\frac{1}{M}\sum_{\ell=1}^{M}p(y=c|\theta_\ell,\xi)\right) && (22)\\
&= -\frac{1}{M}\sum_{m=1}^{M}H[p(y|\theta_m,\xi)] + H\left[\frac{1}{M}\sum_{m=1}^{M}p(y|\theta_m,\xi)\right] && (23)\\
&= \hat{I}_{\text{DBALD}}(\xi). && (24)
\end{aligned}$$

A major consequence of this result is that *the expectation of the DBALD score is a lower bound on the true mutual information score.* We also note that this estimator has been used by Vincent and Rainforth (2017) in the context of Bayesian experimental design, although they did not show that it was a stochastic lower bound.

2.5 New diagnostic for the DBALD score

One advantage of making this connection is that we can bring certain diagnostics that were applied by Foster et al. (2020) over to the active learning setting. In particular, Foster et al. (2020) paired their PCE lower bound with a complementary *upper bound* on $I(\xi)$. This provides a very useful diagnostic tool to tune the number of samples M used to compute the DBALD score. If the lower bound and upper bound are very close, we know that the difference between the DBALD score and the true mutual information must also be small. On the other hand, if the upper and lower bounds are far apart, then the DBALD score might not yet be close to the true mutual information.

One upper bound upper by Foster et al. (2020) was the Nested Monte Carlo (NMC) (Vincent and Rainforth, 2017) estimator. For the discrete y case with Rao-Blackwellisation, the estimator is

$$\hat{I}_{\text{NMC-RB}}(\xi) = -\frac{1}{M}\sum_{m=1}^{M}\sum_{c} p(y=c|\theta_m,\xi)\log\left(\frac{1}{M-1}\sum_{\ell\neq m} p(y=c|\theta_\ell,\xi)\right) - \frac{1}{M}\sum_{m=1}^{M} H[p(y|\theta_m,\xi)] \tag{25}$$

$$= \frac{1}{M}\sum_{m=1}^{M} H\left[p(y|\theta_m,\xi), \frac{1}{M-1}\sum_{\ell\neq m} p(y|\theta_\ell,\xi)\right] - \frac{1}{M}\sum_{m=1}^{M} H[p(y|\theta_m,\xi)] \tag{26}$$

where $H[p,q]$ is the cross-entropy. The expectation of this mutual information estimator is always an upper bound on $I(\xi)$. So both the DBALD score and the NMC-RB estimator converge to $I(\xi)$ as $M\to\infty$, but from opposite directions. We suggest NMC-RB as a diagnostic for the parameter M.

2.6 BALD estimators for regression

The connection to PCE may also be helpful when considering regression models. The standard parametrisation of a Bayesian neural network for regression is for the output of the network with parameters θ and input ξ to be the predictive mean μ and standard deviations σ of a Gaussian $y|\theta,\xi \sim N(\mu(\theta,\xi),\sigma(\theta,\xi)^2)$. (It is normal for y, μ and σ to be vector-valued and for the Gaussian to have a diagonal covariance matrix.)

For the DBALD estimator for a regression model, the entropy of a Gaussian is known in closed form, so $H[p(y|\theta_i,\xi)] = \frac{1}{2}\log\left(2\pi e\sigma(\theta_i,\xi)^2\right)$. However, the entropy of a mixture of Gaussians $H\left[\frac{1}{M}\sum_{i=1}^{M} p(y|\theta_i,\xi)\right]$ cannot be computed analytically. Instead, we could estimate this mixture of Gaussians entropy using Monte Carlo by sampling $i\in\{1,\ldots,M\}$ uniformly, sampling y from $p(y|\theta_i,\xi)$ and calculating the log-density at y.

Despite the fact that we are using an analytic entropy for one term, and a Monte Carlo estimate for the other, it's easy to see that this new estimator is a *partially Rao-Blackwellised* PCE estimator. (This can be proved starting from equation (17).) That means all the existing facts, such as the estimator being a stochastic lower bound on $I(\xi)$, carry over naturally to the regression case.

3 Deep Adaptive Design and Bayesian reinforcement learning

3.1 Introduction

The purpose of this essay is to discuss the connections between the recently proposed Deep Adaptive Design (DAD) (Foster et al., 2021) method and the field of Bayesian reinforcement learning (Ghavamzadeh et al., 2016). That such a connection exists is hinted at by a high-level appraisal of the DAD method—it solves a sequential decision making problem to optimise a certain objective function, decision optimality is dependent on a *state* which is the experimental data already gathered, and the automated decision maker is a design *policy* network. We begin by showing how the sequential Bayesian experimental design problem solved by DAD can be viewed as a Bayes Adaptive Markov Decision Process (BAMDP) (Ross et al., 2007; Guez et al., 2012), making this connection formally precise. We also isolate some of the key differences between the problem DAD is solving and a conventional Bayesian RL problem, noting that the reward in DAD is intractable. Much of the effort of DAD is in establishing a differentiable surrogate for the true objective. The differentiability of the surrogate reward is also a key feature of the DAD problem, which facilitates the direct policy optimisation approach taken to train the policy that is rarely applicable in standard RL problems. We also highlight other features of the DAD method, such as its avoidance of explicitly estimating any posterior distributions, i.e. the avoidance of explicit belief state estimation.

Having studied DAD in some detail, we consider possible extensions of the method that make use of the RL connection. First, there are rather natural extensions of DAD to more general objective functions that incorporate design costs, terminal decisions and other functionals of the posterior distribution. Second, more standard approaches to (Bayesian) RL, such as Q-learning (Watkins and Dayan, 1992; Dearden et al., 1998) can be applicable to the sequential Bayesian experimental design problem. They may be particularly useful for long- or infinite-horizon problems.

3.2 Background on Bayesian Reinforcement Learning

3.2.1 Markov Decision Processes

The Markov Decision Process (MDP) (Bellman, 1957; Duff, 2002) is a highly successful mathematical framework for sequential decision problems in a known environment. Formally, a MDP consists of a state space S, an action space A, a transition model $\mathcal{P}$, a reward distribution R, a discount factor $0 \leq \gamma \leq 1$ and a time horizon T which may be infinite. An agent operates in the MDP by moving between different states in discrete time. For example, if the agent is in state s_t at time t and chooses to play action a_t, then the next state s_{t+1} will be sampled randomly according to the transition model $s_{t+1} \sim \mathcal{P}(s|s_t, a_t)$. Since the distribution over the next state depends only on s_t and a_t, the transitions are Markovian. Finally, by making the transition $s_t \xrightarrow{a_t} s_{t+1}$, the agent receives a random reward $r_t \sim R(r|s_t, a_t, s_{t+1}) \in \mathbb{R}$. The agent's objective is to maximise the discounted sum of rewards $\sum_{t=0}^{T} \gamma^t r_t$. Given the Markovian nature of the problem, it is sufficient to choose actions according to some *policy* π, where $a_t = \pi(s_t)$. The optimality condition for a policy is

$$\pi^* = \arg\max_{\pi} \mathcal{J}(\pi), \tag{27}$$

where

$$\mathcal{J}(\pi) = \mathbb{E}_{s_0 \sim p(s_0) \prod_{t=0}^{T} a_t = \pi(s_t), s_{t+1} \sim \mathcal{P}(s|s_t, a_t), r_t \sim R(r|s_t, a_t, s_{t+1})} \left[\sum_{t=0}^{T} \gamma^t r_t \right]. \tag{28}$$

In a classical MDP, we assume that $\mathcal{P}$ and R are known during the planning phase, when the agent devises their policy π. Of particular utility in planning a policy is the value function, defined as

$$V^\pi(s) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s, \pi(s)), r \sim R(r|s, \pi(s), s')} \left[r + \gamma V^\pi(s') \right] \tag{29}$$

and the Q-function

$$Q^\pi(s, a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s, a), r \sim R(r|s, a, s')} \left[r + \gamma V^\pi(s') \right]. \tag{30}$$

These equations are valid when $T = \infty$, for finite time horizon we also have to take account of time t in state evaluations.

3.2.2 Bayes Adaptive Markov Decision Processes

The BAMDP (Duff, 2002; Ross et al., 2007; Guez et al., 2012; Ghavamzadeh et al., 2016) is one approach to generalising the MDP to deal with unknown transition models. In the BAMDP, the agent retains an explicit posterior distribution over the transition model called a belief state. This allows a formally elegant approach to behaviour under uncertainty which can trade off exploration (learning the transition model) and exploitation (executing actions that receive a high reward).

To set this up formally using the notation of Guez et al. (2012), we begin by considering an outer probabilistic model over the transition probabilities with prior $P(\mathcal{P})$. Given a history of states, actions and rewards $h_t = s_0 a_0 \dots r_{t-1} a_{t-1} s_t$, we can compute a posterior distribution on $\mathcal{P}$ by

$$P(\mathcal{P}|h_t) \propto P(\mathcal{P})P(h_t|\mathcal{P}) = P(\mathcal{P}) \prod_{\tau=0}^{t} \mathcal{P}(s_{\tau+1}|s_\tau, a_\tau). \tag{31}$$

To bring this back into the MDP formulation, we consider an augmented state space S^+ which consists of entire histories, and which encapsulates both the current state and our beliefs about the transition model. Transitions in the augmented state space S^+ are given by integrating over the current beliefs on $\mathcal{P}$

$$\mathcal{P}^+(h_{t+1}|h_t, a_t) = \int P(\mathcal{P}|h_t)\mathcal{P}(s_{t+1}|s_t, a_t)\, d\mathcal{P}. \tag{32}$$

It is also possible for BAMDPs to incorporate unknown reward distributions (see e.g. Zintgraf et al. (2019)), where an outer model over reward distributions is updated on the basis of h_t in the same manner as for the transition probabilities. Specifically, if we have a prior $P(R)$ over reward distributions, then the reward function for playing action a_t in augmented state h_t is

$$R^+(r|h_t, a_t, h_{t+1}) = \int P(R|h_{t+1})R(r|s_t, a_t, s_{t+1})\, dR. \tag{33}$$

Combining these gives a new MDP with state space S^+ of histories, unchanged action space A, augmented transition model $\mathcal{P}^+$, augmented reward distribution R^+, discount factor γ and time horizon T. Optimal action in this new MDP gives the optimal trade-off between exploration and exploitation.

3.3 The Bayesian RL formulation of DAD

In DAD (Foster et al., 2021), we choose a sequence of designs $\xi_1, \dots, \xi_T$ with a view to maximising the expected information gained about a latent parameter of interest θ. To place DAD in a Bayesian RL setting, we begin by associating the design ξ_t chosen before observing an outcome with the action a_{t-1}. The difference in time labels is necessary because ξ_t is chosen before y_t is observed. Since the observation distribution $p(y|\xi, \theta)$ depends on the unknown θ, we are not in a MDP, but rather a BAMDP. As in the previous section, it seems sensible to consider the state space for DAD as the space of histories $h_t = \xi_1 y_1 \dots \xi_t y_t$. Uncertainty over the transition model in DAD is captured by uncertainty in θ. Specifically, we have the following transition distribution for history states

$$p(h_{t+1}|h_t, \xi_{t+1}) = \int p(\theta|h_t)p(y_{t+1}|\xi_{t+1}, \theta)\, d\theta \tag{34}$$

which is the analogue of equation (32), but now expressed in the notation of experimental design. Unlike the standard reinforcement learning setting, there are no external rewards in DAD. Instead, rewards are defined in terms of information gathered about θ. Specifically, we can take the reward distribution on augmented states $R^+(r|h_t, a_t, h_{t+1})$ to be a deterministic function of h_{t+1} that represents the information gained about θ by moving from h_t to h_{t+1}. This is given by the reduction in entropy

$$R^+(h_t, a_t, h_{t+1}) = H[p(\theta|h_t)] - H[p(\theta|h_{t+1})]. \tag{35}$$

To complete the BAMDP specification, we take $\gamma = 1$ and we use a time horizon of T. This gives the objective function for policies

$$\mathcal{J}(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \mathbb{E}_{p(\theta)p(h_T|\theta,\pi)}\left[\sum_{t=1}^{T} H[p(\theta|h_{t-1})] - H[p(\theta|h_t)]\right]. \tag{36}$$

To connect this with the objective that is used in DAD, we apply Theorem 1 of Foster et al. (2021), which tells us that

$$\mathcal{J}(\pi) = \mathbb{E}_{p(\theta)p(h_T|\theta,\pi)}\left[\sum_{t=1}^{T} H[p(\theta|h_{t-1})] - H[p(\theta|h_t)]\right] \overset{\text{Theorem 1}}{=} \mathcal{I}_T(\pi) \tag{37}$$

where

$$\mathcal{I}_T(\pi) = \mathbb{E}_{p(\theta)p(h_T|\theta,\pi)}\left[\log \frac{p(h_T|\theta,\pi)}{\mathbb{E}_{p(\theta')}[p(h_T|\theta',\pi)]}\right]. \tag{38}$$

In summary, we can cast the problem that DAD solves as a BAMDP. We identify designs with actions, experimental histories with augmented states, we use the probabilistic model to give a natural transition distribution on these states, we introduce non-random rewards that are one-step information gains, we set $\gamma = 1$ and generally assume a finite number of experiment iterations T.

3.4 What makes the experimental design problem distinctive?

Having established a theoretical connection between sequential Bayesian experimental design and Bayesian RL, one might naturally ask whether there is any reason to develop specialist algorithms for experimental design when general purpose Bayesian RL algorithms are applicable. First, we focus on the reward structure of the Bayesian experimental design problem. The rewards $r_t = H[p(\theta|h_{t-1})] - H[p(\theta|h_t)]$ are generally intractable, requiring Bayesian inference on θ. Rather than attempting to estimate this reward, DAD proposes the sPCE lower bound on the total expected informationn gain under policy π, namely

$$\mathcal{I}_T(\pi) \geq \mathcal{L}_T(\pi, L) = \mathbb{E}_{p(\theta_0)p(h_T|\theta_0,\pi)p(\theta_{1:L})}\left[\log \frac{p(h_T|\theta_0,\pi)}{\frac{1}{L+1}\sum_{\ell=0}^{L} p(h_T|\theta_\ell,\pi)}\right]. \tag{39}$$

Interestingly, there is a way to interpret the sPCE objective within the RL framework. First, we use *root sampling* to sample θ_0 and h_T together. We also fix the contrasts $\theta_{1:L}$. Finally, we use the surrogate rewards

$$\tilde{r}_t = \log \frac{p(h_t|\theta_0,\pi)}{\frac{1}{L+1}\sum_{\ell=0}^{L} p(h_t|\theta_\ell,\pi)} - \log \frac{p(h_{t-1}|\theta_0,\pi)}{\frac{1}{L+1}\sum_{\ell=0}^{L} p(h_{t-1}|\theta_\ell,\pi)}. \tag{40}$$

Since these rewards depend on θ_0, we can treat them as randomised rewards if we are only conditioning on h_t.

One important feature of these rewards is that, whilst intractable, the surrogate $\mathcal{L}_T(\pi, L)$ is differentiable with respect to the designs $(\xi_t)_{t=1}^T$ and observations $(y_t)_{t=1}^T$. In the simplest form of DAD, we further assume a differentiable relationship between y_t and ξ_t that is encapsulated by a reparametrisable way to sample $p(y|\theta, \xi)$. Concretetly, for example, we might have $y|\theta, \xi = \mu(\theta, \xi) + \sigma(\theta, \xi)\varepsilon$ where $\varepsilon \sim N(0, 1)$ and μ and σ are differentiable functions. The result of these assumptions is that we can directly differentiate the surrogate objective $\mathcal{L}_T(\pi, L)$ with respect to the parameters ϕ of the policy network π_ϕ that generates the designs $(\xi_t)_{t=1}^T$ according to the formula $\xi_t = \pi_\phi(h_{t-1})$. DAD optimises the policy π_ϕ directly by gradient descent on $\mathcal{L}_T(\pi, L)$.

Thus, DAD can be characterised in RL language as a direct policy optimisation method. Whilst direct policy optimisation methods (Lorberbom et al., 2019; Howell et al., 2021) are used in RL, they are far from the norm, with methodologies such as Q-learning (Watkins and Dayan, 1992) and actor–critic (Konda and Tsitsiklis, 2000) being more dominant. This may be because RL does not typically assume that the reward function is differentiable—for example, rewards from a real environment rarely come with gradient information. It may also be because discrete action problems are more the focus.

DAD also contrasts with many approaches to *Bayesian* RL in that it avoids the estimation of the posteriors $p(\theta|h_t, \pi)$. In Bayesian RL, these posterior distributions are referred to as *belief states*. Many methods for tackling Bayesian RL problems utilise the estimation of belief states (Ghavamzadeh et al., 2016; Igl et al., 2018; Zintgraf et al., 2019). DAD instead relies on an approach that is closer to the method of root sampling (Guez et al., 2012). This is also one difference between DAD and the previous approach to non-greedy sequential Bayesian experimental design of Huan and Marzouk (2016).

3.5 New objective functions for DAD

Seeing DAD in the framework of Bayesian RL naturally invites the question of whether the general DAD methodology can be applied to objective functions (rewards) that are not information gains. The preceding discussion suggests that, using root sampling so a dependence on θ is possible, we could consider rewards of the form

$$r_t^{\text{general}} = R(\theta, h_t, \epsilon_t) \tag{41}$$

where R is a known differentiable function and ϵ_t is an independent noise random variable. Clearly, the information gain reward r_t fits this pattern, being a function of h_t only. Combining the differentiable reward function with the reparametrisation assumption would mean that the general reward

$$\mathcal{J}^{\text{general}}(\pi) = \mathbb{E}_{p(\theta)p(h_T)p(\epsilon_{1:T})}\left[\sum_{t=1}^{T} r_t^{\text{general}}\right] \tag{42}$$

can be optimised with respect to π by direct policy gradients. In the experimental design context, this opens the door to two relatively simple extensions of DAD. First, we can assign a (differentiable) cost to each design. Suppose we augment the original expected information gain objective with the negative sum of the costs of the designs. Using λ to trade off cost and information, we arrive at

$$\mathcal{J}^{\text{costed}}(\pi) = \mathcal{I}_T(\pi) - \lambda\mathbb{E}\left[\sum_{t=1}^{T} C(\xi_t)\right] \tag{43}$$

which we can tackle using an approach that is essentially the same as DAD. Second, we can consider different measures of the quality of the final posterior distribution. For instance, with a one-dimensional θ, we might be more interested in reducing posterior *variance* than posterior entropy. We could take the reward function

$$r_t^{\text{variance}} = \mathrm{Var}_{p(\theta|h_{t-1})}[\theta] - \mathrm{Var}_{p(\theta|h_t)}[\theta]. \tag{44}$$

Whilst there are certain reasons why the entropy approach is considered more theoretically well-justified (Lindley, 1956), using a different functional of the posterior distribution as a reward signal does fit relatively naturally into the DAD framework. The remaining piece of the puzzle would be whether that functional could be estimated efficiently as DAD estimates the information gain using sPCE. For the variance, we have

$$\mathbb{E}_{p(\theta)p(h_T|\theta,\pi)}\left[\sum_{t=1}^{T} r_t^{\text{variance}}\right] \geq \mathrm{Var}_{p(\theta)}[\theta] - \mathbb{E}_{p(\theta)p(h_T|\theta,\pi)}\left[(\theta - f_{\phi'}(h_T))^2\right] \tag{45}$$

where $f_{\phi'}$ is a learnable function. Note the similarity with the Barber–Agakov bound (Barber and Agakov, 2003; Foster et al., 2019, 2020).

3.6 RL algorithms for Bayesian experimental design

To conclude, making the formal connection between sequential Bayesian experimental design opens up the possibility of using the vast literature on Bayesian RL and control theory to improve our ability to plan sequential experiments. Whilst the direct policy optimisation approach of DAD works remarkably well, understanding the connection to RL should aid us when this training method begins to break down. The application of existing Bayesian RL algorithms to experimental design is an exciting area for new research that is well within reach.

A case of potential difficulty for DAD, where such insights may be useful, is in long-horizon experiments. In order to plan effectively for long experiments, DAD simulates thousands of possible experimental trajectories. However, the efficiency of this simulation is likely to drop as T increases. DAD is extremely data hungry—it resimulates completely new trajectories at each gradient step. This avoids any problems of the training data becoming out-of-date, but it increases the training cost.

It is also conceivable that, in some settings, it is impossible to plan for all future eventualities. The RL analogy would be a strongly stochastic environment in which a game is selected at random from a long list at the start of play. The agent, therefore, has to first discover which game it is playing, and then to play it successfully. If all planning is conducted up-front, then the RL agent has to learn how to play every single game well before starting on the real environment. The alternative is to introduce some real data and retrain the policy as we go. In the RL setting, that would mean discovering which game is being played before knowing how to play the games, which could be achieved with a much simpler policy. Once this discovery is made with good confidence, we can retrain to learn to play that specific game. In the experimental design setting, we are often in the 'unknown game' setting. This is because, until we have observed some data, it is almost impossible to know which later experiments will be optimal to run. The DAD approach is to simulate different possibilities and learn to 'play' well across the board. The retraining alternative would be a hybrid approach between the standard greedy method and DAD in which some real data is used to retrain the policy as we progress.

4 Statistical estimation of mutual information

4.1 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

$$I(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right]. \tag{46}$$

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.

4.2 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

$$A_{n,m} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{m}\sum_{j=1}^{m} p(\mathbf{y}_i|\mathbf{x}_{ij})} \tag{47}$$

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 $(\mathbf{x}_{1j})_{j=1}^{m}$

$$A'_{n,m} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{m}\sum_{j=1}^{m} p(\mathbf{y}_i|\mathbf{x}_{1j})}, \tag{48}$$

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

$$\tilde{A}_n = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{n-1}\sum_{j\neq i} p(\mathbf{y}_i|\mathbf{x}_{j})}. \tag{49}$$

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*

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right] < \infty \text{ and } \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{q}\right] < \infty. \tag{50}$$

Then we have

$$\mathbb{E}[A_{n,m}] = I(\mathbf{x}, \mathbf{y}) + \frac{1}{m}\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{2p(\mathbf{y})^2}\right] + \mathcal{O}\left(m^{-3/2}\right). \tag{51}$$

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

$$U_j = \frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)}. \tag{52}$$

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

$$A_{1,m} = \log \frac{p(\mathbf{y}_1|\mathbf{x}_1)}{p(\mathbf{y}_1)} - \log \left(\frac{1}{m} \sum_{j=1}^{m} U_j \right), \tag{53}$$

giving

$$\mathbb{E}[A_{n,m}] = I(\mathbf{x}, \mathbf{y}) - \mathbb{E}\left[\log \left(\frac{1}{m} \sum_{j=1}^{m} U_j \right) \right]. \tag{54}$$

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

$$L_k(x) = \sum_{j=1}^{k} \frac{(-1)^{j+1}}{j} (x-1)^j, \tag{55}$$

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

$$\mathbb{E}\left[\log \left(\frac{1}{m} \sum_{j=1}^{m} U_j \right) \right] = -\frac{1}{2} \mathbb{E}\left[\left(\frac{1}{m} \sum_{j=1}^{m} (U_j - 1) \right)^2 \right] + \mathbb{E}[\varepsilon] \tag{56}$$

and

$$|\mathbb{E}[\varepsilon]| \leq \mathbb{E}[|\varepsilon|] \leq \mathbb{E}\left[\left| \frac{1}{m} \sum_{j=1}^{m} (U_j - 1) \right|^3 \max\left(1, -\log\left(\frac{1}{m} \sum_{j=1}^{m} U_j \right)\right) \right] \tag{57}$$

applying Hölder's Inequality

$$\leq \mathbb{E}\left[\left| \frac{1}{m} \sum_{j=1}^{m} (U_j - 1) \right|^{3p} \right]^{1/p} \mathbb{E}\left[\max\left(1, -\log\left(\frac{1}{m} \sum_{j=1}^{m} U_j \right)\right)^q \right]^{1/q}. \tag{58}$$

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

$$\mathbb{E}\left[\left| \frac{1}{m} \sum_{j=1}^{m} (U_j - 1) \right|^{3p} \right]^{1/p} \leq D_{3p}^{1/p} m^{-3/2} \mathbb{E}\left[|U_1 - 1|^{3p} \right]^{1/p} \tag{59}$$

and

$$\mathbb{E}\left[|U_1 - 1|^{3p} \right] \leq 1 + \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} \right)^{3p} \right] < \infty \text{ by assumption.} \tag{60}$$

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

$$\begin{aligned}
\mathbb{E}\left[\max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m} U_j\right)\right)^q\right]^{1/q} &\leq \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}\max\left(1, -\log\left(U_j\right)\right)^q\right]^{1/q} && (61)\\
&= \mathbb{E}\left[\max\left(1, -\log\left(U_1\right)\right)^q\right]^{1/q} && (62)\\
&\leq \left(1 + \mathbb{E}[|\log U_1|^q]\right)^{1/q} && (63)\\
&= \left(1 + \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^q\right]\right)^{1/q} && (64)\\
&< \infty \text{ by assumption.} && (65)
\end{aligned}$$

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

$$\frac{1}{2}\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}(U_j - 1)\right)^2\right] = \frac{1}{2m}\mathbb{E}_{p(\mathbf{y})}\left[\mathbb{E}_{p(\mathbf{x})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} - 1\right)^2\right]\right] = \frac{1}{m}\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{2p(\mathbf{y})^2}\right]. \tag{66}$$

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

$$\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{2p(\mathbf{y})^2}\right] = \frac{1}{2}\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} - 1\right)^2\right] \tag{67}$$

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*

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right] < \infty \textit{ and } \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2q}\right] < \infty. \tag{68}$$

Then,

$$\begin{aligned}
\mathrm{Var}[A_{n,m}] = &\frac{1}{n}\mathrm{Var}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right]\\
&+ \frac{1}{nm}\left(\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{p(\mathbf{y})^2}\right] + \mathrm{Cov}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, \frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right]\right) && (69)\\
&+ \mathcal{O}\left(n^{-1}m^{-3/2}\right).
\end{aligned}$$

Proof. We have

$$\mathrm{Var}[A_{n,m}] = \frac{1}{n}\mathrm{Var}[A_{1,m}]. \tag{70}$$

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

$$\mathrm{Var}[A_{1,m}] = \mathbb{E}[\mathrm{Var}[A_{1,m}|\mathbf{x}_1, \mathbf{y}_1]] + \mathrm{Var}[\mathbb{E}[A_{1,m}|\mathbf{x}_1, \mathbf{y}_1]]. \tag{71}$$

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

$$\mathbb{E}\left[\mathrm{Var}[A_{1,m}|\mathbf{x}_1, \mathbf{y}_1]\right] = \mathbb{E}\left[\mathrm{Var}\left[\log\left(\frac{1}{m}\sum_{j=1}^{m} U_j\right)\middle|\, \mathbf{y}_1\right]\right] \text{ where } U_j = \frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)} \tag{72}$$

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

$$|\log(x)^2 - (x-1)^2| = |(\log x - x + 1)(\log x + x - 1)| \leq |\log x - x + 1||\log x + x - 1|. \tag{73}$$

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

$$|\log(x)^2 - (x-1)^2| \leq 3|x-1|^3 \max(1, -\log x)^2. \tag{74}$$

We can now return to computing the conditional expectation of equation (72). We have

$$\mathbb{E}\left[\mathbb{E}\left[\log\left(\frac{1}{m}\sum_{j=1}^m U_j\right)^2 \middle| \mathbf{y}_1\right]\right] = \mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^m (U_j - 1)\right)^2\right] + \mathbb{E}[\eta] \tag{75}$$

where our recent result guarantees that

$$|\mathbb{E}[\eta]| \leq \mathbb{E}[|\eta|] \leq 3\mathbb{E}\left[\left|\frac{1}{m}\sum_{j=1}^m (U_j - 1)\right|^3 \max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^m U_j\right)\right)^2\right]. \tag{76}$$

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

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right] < \infty \text{ and } \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2q}\right] < \infty \tag{77}$$

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

$$\mathbb{E}\left[\mathbb{E}\left[\log\left(\frac{1}{m}\sum_{j=1}^m U_j\right) \middle| \mathbf{y}_1\right]^2\right] = \mathcal{O}(m^{-2}). \tag{78}$$

Putting these pieces together, we have

$$\mathbb{E}\left[\mathrm{Var}[A_{1,m}|\mathbf{x}_1, \mathbf{y}_1]\right] = \frac{1}{m}\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{p(\mathbf{y})^2}\right] + \mathcal{O}(m^{-3/2}). \tag{79}$$

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

$$\mathbb{E}\left[A_{1,m}|\mathbf{x}_1, \mathbf{y}_1\right] = \log \frac{p(\mathbf{y}_1|\mathbf{x}_1)}{p(\mathbf{y}_1)} + \frac{1}{m}\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}_1|\mathbf{x})]}{2p(\mathbf{y}_1)^2} + \mathcal{O}\left(m^{-3/2}\right). \tag{80}$$

Taking the variance gives

$$\mathrm{Var}\left[\mathbb{E}\left[A_{1,m}|\mathbf{x}_1, \mathbf{y}_1\right]\right] = \mathrm{Var}_{p(\mathbf{x},\mathbf{y})}\left[\log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] + \frac{1}{m}\mathrm{Cov}\left(\log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, \frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right) + \mathcal{O}(m^{-3/2}). \tag{81}$$

Thus,

$$\begin{aligned}\mathrm{Var}[A_{1,m}] = {} & \mathrm{Var}_{p(\mathbf{x},\mathbf{y})}\left[\log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] \\ & + \frac{1}{m}\left(\mathbb{E}_{p(\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x})}[p(\mathbf{y}|\mathbf{x})]}{p(\mathbf{y})^2}\right] + \mathrm{Cov}_{p(\mathbf{x},\mathbf{y})}\left(\log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, \frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right)\right) + \mathcal{O}(m^{-3/2})\end{aligned} \tag{82}$$

and the full result follows. □

Combining the last two theorems establishes that

$$\mathbb{E}\left[|A_{n,m} - I(\mathbf{x},\mathbf{y})|^2\right] = \mathcal{O}\left(\frac{1}{n} + \frac{1}{m^2}\right). \tag{83}$$

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,

$$A^{(\kappa)}_{n,m} \to \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{\kappa(\mathbf{x},\mathbf{y})}{\kappa(\mathbf{y})}\right] \text{ as } m,n\to\infty \tag{84}$$

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

4.3 Prior Contrastive Estimation and InfoNCE

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

$$B_{n,m} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{m+1}\left(p(\mathbf{y}_i|\mathbf{x}_i) + \sum_{j=1}^{m} p(\mathbf{y}_i|\mathbf{x}_{ij})\right)}. \tag{85}$$

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

$$\tilde{B}_n = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{n}\sum_{j=1}^{n} p(\mathbf{y}_i|\mathbf{x}_j)}. \tag{86}$$

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)

$$\tilde{B}^{(\kappa)}_n = \frac{1}{n}\sum_{i=1}^{n}\log\frac{\kappa(\mathbf{x}_i,\mathbf{y}_i)}{\frac{1}{n}\sum_{j=1}^{n} \kappa(\mathbf{x}_j,\mathbf{y}_i)} \tag{87}$$

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*

$$\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] < \infty. \tag{88}$$

Then,

$$0 \le I(\mathbf{x},\mathbf{y}) - \mathbb{E}[B_{n,m}] \le \frac{1}{m+1}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} - 1\right]. \tag{89}$$

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*

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right] < \infty \text{ and } \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{q}\right] < \infty. \tag{90}$$

Then we have

$$\begin{aligned}\mathbb{E}[B_{n,m}] =& I(\mathbf{x},\mathbf{y}) \\ &- \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} - 1\right] + \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{2p(\mathbf{y})^2}\right] \\ &+ \mathcal{O}\left(m^{-3/2}\right).\end{aligned} \tag{91}$$

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

$$\mathbb{E}[B_{1,m}] = I(\mathbf{x},\mathbf{y}) - \mathbb{E}\left[\log\left(\frac{1}{m+1}\sum_{j=0}^{m} U_j\right)\right]. \tag{92}$$

To reduce this to a more manageable form, we have

$$\mathbb{E}\left[\log\left(\frac{1}{m+1}\sum_{j=0}^{m} U_j\right)\right] = \log\left(\frac{m}{m+1}\right) + \mathbb{E}\left[\log\left(1 + \frac{U_0}{m} + \frac{1}{m}\sum_{j=1}^{m}(U_j - 1)\right)\right] \tag{93}$$

$$= \log\left(\frac{m}{m+1}\right) + \mathbb{E}\left[\log\left(1 + \frac{U_0}{m}\right)\right] + \mathbb{E}\left[\log\left(1 + \frac{1}{m}\sum_{j=1}^{m}\frac{U_j - 1}{1 + U_0/m}\right)\right]. \tag{94}$$

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

$$\mathbb{E}\left[\log\left(1 + \frac{1}{m}\sum_{j=1}^{m}\frac{U_j - 1}{1 + U_0/m}\right)\right] = -\frac{1}{2}\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j - 1}{1 + U_0/m}\right)^2\right] + \mathbb{E}[\zeta] \tag{95}$$

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

$$|\zeta| \leq \left|\frac{1}{m}\sum_{j=1}^{m}\frac{U_j - 1}{1 + U_0/m}\right|^3 \max\left(1, -\log\left(1 + \frac{1}{m}\sum_{j=1}^{m}\frac{U_j - 1}{1 + U_0/m}\right)\right). \tag{96}$$

Since $U_0 > 0$, we must have

$$\left|\frac{U_j - 1}{1 + U_0/m}\right| \leq |U_j - 1|, \tag{97}$$

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

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j - 1}{1 + U_0/m}\right)^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j - 1}{1 + U_0/m}\right)^2 \middle| \mathbf{x}_1, \mathbf{y}_1\right]\right] \tag{98}$$

$$= \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{1}{1 + U_0/m}\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right] \tag{99}$$

$$= \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{1}{1 + \frac{p(\mathbf{y}|\mathbf{x})}{mp(\mathbf{y})}}\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right], \tag{100}$$

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

$$\frac{1}{1 + \frac{p(\mathbf{y}|\mathbf{x})}{mp(\mathbf{y})}} = 1 + \mathcal{O}(m^{-1}) \tag{101}$$

and so we can drop the extract factor, giving

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j - 1}{1 + U_0/m}\right)^2\right] = \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right] + \mathcal{O}(m^{-2}) \tag{102}$$

We also need to expand

$$\log\left(\frac{m}{m+1}\right) + \mathbb{E}\left[\log\left(1+\frac{U_0}{m}\right)\right] = \mathbb{E}\left[\log\left(1+\frac{U_0-1}{m+1}\right)\right] \tag{103}$$

$$= \mathbb{E}\left[\frac{U_0-1}{m+1}\right] + \mathcal{O}(m^{-3/2}) \tag{104}$$

$$= \frac{1}{m+1}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} - 1\right] + \mathcal{O}(m^{-3/2}) \tag{105}$$

$$= \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} - 1\right] + \mathcal{O}(m^{-3/2}) \text{ as the difference is order } m^{-2}. \tag{106}$$

Combining these gives the result. □

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

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right] < \infty \text{ and } \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2q}\right] < \infty. \tag{107}$$

Then,

$$\begin{aligned}\mathrm{Var}[B_{n,m}] = &\frac{1}{n}\mathrm{Var}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] \\ &+ \frac{1}{nm}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{2p(\mathbf{y})^2}\right] \\ &+ \frac{1}{nm}\mathrm{Cov}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, -\frac{2p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} + \frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right] \\ &+ \mathcal{O}\left(n^{-1}m^{-3/2}\right).\end{aligned} \tag{108}$$

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

$$\mathrm{Var}[B_{n,m}] = \frac{1}{n}\mathrm{Var}[B_{1,m}]. \tag{109}$$

By Tower Law,

$$\mathrm{Var}[B_{1,m}] = \mathbb{E}[\mathrm{Var}[B_{1,m}|\mathbf{x}_1,\mathbf{y}_1]] + \mathrm{Var}[\mathbb{E}[B_{1,m}|\mathbf{x}_1,\mathbf{y}_1]]. \tag{110}$$

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

$$\begin{aligned}\mathbb{E}[\mathrm{Var}[B_{1,m}|\mathbf{x}_1,\mathbf{y}_1]] = &\mathbb{E}\left[\mathbb{E}\left[\log\left(1+\frac{1}{m}\sum_{j=1}^m\frac{U_j-1}{1+U_0/m}\right)^2\middle|\mathbf{x}_1,\mathbf{y}_1\right]\right] \\ &- \mathbb{E}\left[\mathbb{E}\left[\log\left(1+\frac{1}{m}\sum_{j=1}^m\frac{U_j-1}{1+U_0/m}\right)\middle|\mathbf{x}_1,\mathbf{y}_1\right]^2\right].\end{aligned} \tag{111}$$

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

$$|\log(x)^2 - (x-1)^2| \le |x-1|^3\max(1, -\log x)^2. \tag{112}$$

Thus,

$$\mathbb{E}\left[\mathbb{E}\left[\log\left(1+\frac{1}{m}\sum_{j=1}^m\frac{U_j-1}{1+U_0/m}\right)^2\middle|\mathbf{x}_1,\mathbf{y}_1\right]\right] = \mathbb{E}\left[\left(1+\frac{1}{m}\sum_{j=1}^m\frac{U_j-1}{1+U_0/m}\right)^2\right] + \mathbb{E}[\nu] \tag{113}$$

where

$$|\mathbb{E}[\nu]| \le \mathbb{E}[|\nu|] \le \mathbb{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right|^3 \max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)\right)^2\right] \tag{114}$$

$$\overset{\text{Hölder}}{\le} \mathbb{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right|^{3p}\right]^{1/p} \mathbb{E}\left[\max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)\right)^{2q}\right]^{1/q} \tag{115}$$

$$\le \mathbb{E}\left[\left|\frac{1}{m}\sum_{j=1}^{m}U_j-1\right|^{3p}\right]^{1/p} \mathbb{E}\left[\max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m}U_j-1\right)\right)^{2q}\right]^{1/q} \tag{116}$$

$$\overset{\text{Corollary 8}}{\le} D_{3p}^{1/p} m^{-3/2}\mathbb{E}\left[|U_1-1|^{3p}\right]^{1/p} \mathbb{E}\left[\max\left(1, -\log\left(\frac{1}{m}\sum_{j=1}^{m}U_j-1\right)\right)^{2q}\right]^{1/q} \tag{117}$$

$$\overset{\text{convexity}}{\le} D_{3p}^{1/p} m^{-3/2}\mathbb{E}\left[|U_1-1|^{3p}\right]^{1/p}\left(1+\mathbb{E}\left[|\log U_1|^{2q}\right]\right)^{1/q} \tag{118}$$

$$\le D_{3p}^{1/p} m^{-3/2}\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left(\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right)^{3p}\right]^{1/p}\left(1+\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2q}\right]\right)^{1/q}. \tag{119}$$

We also have, as previously

$$\mathbb{E}\left[\left(1+\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)^2\right] = \frac{1}{m}\mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{2p(\mathbf{y})^2}\right] + \mathcal{O}(m^{-2}). \tag{120}$$

On the other hand, Theorem 4 shows that

$$\mathbb{E}\left[\mathbb{E}\left[\log\left(1+\frac{1}{m}\sum_{j=1}^{m}\frac{U_j-1}{1+U_0/m}\right)\middle|\mathbf{x}_1,\mathbf{y}_1\right]^2\right] = \mathcal{O}(m^{-2}). \tag{121}$$

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

$$\mathbb{E}[B_{1,m}|\mathbf{x}_1,\mathbf{y}_1] = \log\frac{p(\mathbf{y}_1|\mathbf{x}_1)}{p(\mathbf{y}_1)} + \frac{1}{m}\left(1-\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}+\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{2p(\mathbf{y})^2}\right) + \mathcal{O}\left(m^{-3/2}\right). \tag{122}$$

Thus,

$$\begin{aligned}\mathrm{Var}[\mathbb{E}[B_{1,m}|\mathbf{x}_1,\mathbf{y}_1]] &= \mathrm{Var}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] \\ &\quad + \frac{1}{m}\mathrm{Cov}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, -\frac{2p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}+\frac{\mathrm{Var}_{p(\mathbf{x}')}[p(\mathbf{y}|\mathbf{x}')]}{p(\mathbf{y})^2}\right] \\ &\quad + \mathcal{O}\left(m^{-3/2}\right).\end{aligned} \tag{123}$$

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

$$-\frac{1}{nm}\mathrm{Cov}_{p(\mathbf{x},\mathbf{y})}\left[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}, \frac{2p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right]. \tag{124}$$

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.

4.4 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

$$P_\ell = A_{1,M_\ell} \tag{125}$$

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

$$\begin{aligned} Z_\ell = &- \log\left(\frac{1}{M_\ell}\sum_{j=1}^{M_\ell} p(\mathbf{y}_1|\mathbf{x}_{1j})\right) \\ &+ \frac{1}{2}\left[\log\left(\frac{1}{M_{\ell-1}}\sum_{j=1}^{M_{\ell-1}} p(\mathbf{y}_1|\mathbf{x}_{1j})\right) + \log\left(\frac{1}{M_{\ell-1}}\sum_{j=1+M_{\ell-1}}^{M_\ell} p(\mathbf{y}_1|\mathbf{x}_{1j})\right)\right]. \end{aligned} \tag{126}$$

The key property of Z_ℓ is

$$\mathbb{E}[Z_\ell] = \mathbb{E}[P_\ell - P_{\ell-1}] \tag{127}$$

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*

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^p\right] < \infty \quad \textit{and} \quad \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^q\right] < \infty. \tag{128}$$

Then,

$$\mathbb{E}\left[|Z_\ell|\right] = O(2^{-a\ell}), \qquad \mathrm{Var}(Z_\ell) = O(2^{-r\ell}) \tag{129}$$

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

Proof. First, define

$$\beta_\ell^{(a)} = \frac{1}{M_\ell}\sum_{j=1}^{M_\ell}\frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)} \tag{130}$$

$$\beta_\ell^{(b)} = \frac{1}{M_\ell}\sum_{j=1+M_\ell}^{M_{\ell+1}}\frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)} \tag{131}$$

$$\text{so} \quad Z_\ell = -\log\beta_\ell^{(a)} + \frac{1}{2}\left(\log\beta_{\ell-1}^{(a)} + \log\beta_{\ell-1}^{(b)}\right) \tag{132}$$

$$\text{and} \quad \beta_\ell^{(a)} = \frac{1}{2}\left(\beta_{\ell-1}^{(a)} + \beta_{\ell-1}^{(b)}\right). \tag{133}$$

We then have

$$Z_\ell = -\log\beta_\ell^{(a)} + \frac{1}{2}\left(\log\beta_{\ell-1}^{(a)} + \log\beta_{\ell-1}^{(b)}\right) \tag{134}$$

$$= -\left(\log \beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right) + \frac{1}{2}\left(\log \beta_{\ell-1}^{(a)} - \beta_{\ell-1}^{(a)} + 1\right) + \frac{1}{2}\left(\log \beta_{\ell-1}^{(b)} - \beta_{\ell-1}^{(b)} + 1\right) \tag{135}$$

$$= 2\left[-\frac{1}{2}\left(\log \beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right) + \frac{1}{4}\left(\log \beta_{\ell-1}^{(a)} - \beta_{\ell-1}^{(a)} + 1\right) + \frac{1}{4}\left(\log \beta_{\ell-1}^{(b)} - \beta_{\ell-1}^{(b)} + 1\right)\right] \tag{136}$$

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

$$|Z_\ell|^2 \leq 2\left|\log \beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right|^2 + \left|\log \beta_{\ell-1}^{(a)} - \beta_{\ell-1}^{(a)} + 1\right|^2 + \left|\log \beta_{\ell-1}^{(b)} - \beta_{\ell-1}^{(b)} + 1\right|^2. \tag{137}$$

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

$$|\log x - x + 1| \leq |x-1|^r \max(-\log x, 1) \tag{138}$$

which gives

$$\left|\log \beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right|^2 \leq \left|\beta_\ell^{(a)} - 1\right|^{2r}\left(\max\left(-\log \beta_\ell^{(a)}, 1\right)\right)^2. \tag{139}$$

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

$$\mathbb{E}\left[\left|\log \beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right|^2\right] \leq \left\|\left|\beta_\ell^{(a)} - 1\right|^{2r}\right\|_{L^s}\left\|\left(\max\left(-\log \beta_\ell^{(a)}, 1\right)\right)^2\right\|_{L^t}. \tag{140}$$

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

$$\left\|\left|\beta_\ell^{(a)} - 1\right|^{2r}\right\|_{L^s} \leq D_{2rs}^{1/s}\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2sr}\right]^{1/s}(M_0 2^\ell)^{-r}, \tag{141}$$

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

$$\left\|\left(\max\left(-\log \beta_\ell^{(a)}, 1\right)\right)^2\right\|_{L^t} \leq \left\|\left(\frac{1}{M_\ell}\sum_{j=1}^{M_\ell}\max\left(-\log\frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)}, 1\right)\right)^2\right\|_{L^t} \tag{142}$$

$$= \left(\mathbb{E}\left[\left(\frac{1}{M_\ell}\sum_{j=1}^{M_\ell}\max\left(-\log\frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)}, 1\right)\right)^{2t}\right]\right)^{1/t} \tag{143}$$

$$\leq \left(\frac{1}{M_\ell}\mathbb{E}\left[\sum_{j=1}^{M_\ell}\max\left(-\log\frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)}, 1\right)^{2t}\right]\right)^{1/t} \tag{144}$$

$$\leq \left(\frac{1}{M_\ell}\mathbb{E}\left[\sum_{j=1}^{M_\ell}\left|\log\frac{p(\mathbf{y}_1|\mathbf{x}_{1j})}{p(\mathbf{y}_1)}\right|^{2t} + 1\right]\right)^{1/t} \tag{145}$$

$$= \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^{2t} + 1\right]^{1/t}. \tag{146}$$

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

$$\mathbb{E}\left[\left|\log \beta_\ell^{(a)} - \beta_\ell^{(a)} + 1\right|^2\right] \leq A_0 2^{-r\ell} \tag{147}$$

where

$$A_0 = D_{2rs}^{1/s}\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^p\right]^{1/s}\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}\left[\left|\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right|^q + 1\right]^{1/t} M_0^{-r}. \tag{148}$$

Since we can bound the other two terms of (137) 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

$$\begin{aligned} Z_{n,\ell} =&\frac{1}{n}\sum_{i=1}^{n}\left[-\log\left(\frac{1}{M_\ell}\sum_{j=1}^{M_\ell}p(\mathbf{y}_i|\mathbf{x}_{ij})\right)\right. \\ &\left.+\frac{1}{2}\left[\log\left(\frac{1}{M_{\ell-1}}\sum_{j=1}^{M_{\ell-1}}p(\mathbf{y}_i|\mathbf{x}_{ij})\right)+\log\left(\frac{1}{M_{\ell-1}}\sum_{j=1+M_{\ell-1}}^{M_\ell}p(\mathbf{y}_i|\mathbf{x}_{ij})\right)\right]\right]. \end{aligned} \tag{149}$$

Then

$$Z_L^{\mathrm{MLMC}}=\sum_{\ell=0}^{L}Z_{N_\ell,\ell} \tag{150}$$

with

$$\mathbb{E}\left[\left|Z_L^{\mathrm{MLMC}}-I(\mathbf{x},\mathbf{y})\right|^2\right]=\sum_{\ell=0}^{L}\frac{\mathrm{Var}[Z_\ell]}{N_\ell}+\left[\mathbb{E}[P_L]-I(\mathbf{x},\mathbf{y})\right]^2. \tag{151}$$

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).

4.5 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*

$$\mathbb{E}\left(\left|\sum_{i=1}^{m}(X_i-\mu)\right|^p\right)\le D_p\mathbb{E}\left(\left(\sum_{i=1}^{m}|X_i|^2\right)^{p/2}\right) \tag{152}$$

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*

$$\mathbb{E}\left(\left|\frac{1}{m}\sum_{i=1}^{m}(X_i-\mu)\right|^p\right)\le D_p m^{-p/2}\mathbb{E}\left[|X_1|^p\right] \tag{153}$$

Proof. Applying the Marcinkiewicz–Zygmund Inequality, we have

$$\mathbb{E}\left(\left|\frac{1}{m}\sum_{i=1}^{m}(X_i-\mu)\right|^p\right)\le D_p m^{-p/2}\mathbb{E}\left(\left(\frac{1}{m}\sum_{i=1}^{m}|X_i|^2\right)^{p/2}\right), \tag{154}$$

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

$$\le D_p m^{-p/2}\mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}|X_i|^p\right) \tag{155}$$

$$= D_p m^{-p/2} \mathbb{E}\left[|X_1|^p\right]. \tag{156}$$

□

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*

$$\mathbb{E}\left[f\left(\frac{1}{m}\sum_{i=1}^m X_i\right)\right] = \sum_{j=0}^k \frac{f^{(j)}(\mu)}{j!}\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^m (X_i-\mu)\right)^j\right] + \mathcal{O}\left(m^{-(k+1)/2}\right). \tag{157}$$

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

$$f(x) = \sum_{j=0}^k \frac{f^{(j)}(\mu)}{j!}(x-\mu)^k + \frac{f^{(k+1)}(\xi)}{(k+1)!}(x-\mu)^{k+1}. \tag{158}$$

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

$$\mathbb{E}\left[f\left(\frac{1}{m}\sum_{i=1}^m X_i\right)\right] = \sum_{j=0}^k \frac{f^{(j)}(\mu)}{j!}\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^m (X_i-\mu)\right)^j\right] + \mathbb{E}\left[\frac{f^{(k+1)}(\Xi)}{(k+1)!}\left(\frac{1}{m}\sum_{i=1}^m (X_i-\mu)\right)^{k+1}\right] \tag{159}$$

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

$$\mathbb{E}\left(\left|\sum_{i=1} X_i - \mu\right|^{k+1}\right) \le D_{k+1} m^{(k+1)/2}\mathbb{E}\left[|X_1|^{k+1}\right]. \tag{160}$$

Hence we conclude that

$$\left|\mathbb{E}\left[\frac{f^{(k+1)}(\Xi)}{(k+1)!}\left(\frac{1}{m}\sum_{i=1}^m (X_i-\mu)\right)^{k+1}\right]\right| \le \frac{M D_{k+1}\mathbb{E}\left[|X_1|^{k+1}\right] m^{-(k+1)/2}}{(k+1)!} = \mathcal{O}(m^{-(k+1)/2}). \tag{161}$$

□

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

Lemma 10. *Define*

$$L_k(x) = \sum_{j=1}^k \frac{(-1)^{j+1}}{j}(x-1)^j. \tag{162}$$

Then $|\log x - L_k(x)| \le |x-1|^{k+1}\max(1, -\log x)$ 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

$$\log x = L_k(x) + \frac{(-1)^{k+2}}{\xi^{k+1}}(x-\xi)^k(x-1) \tag{163}$$

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

$$\frac{\xi - x}{\xi} = 1 - x/\xi \text{ and } 0 \leq 1 - x/\xi \leq 1 - x \text{ since } x \leq \xi \leq 1. \tag{164}$$

Thus, the magnitude of the remainder term becomes

$$\left| \frac{(-1)^{k+2}}{\xi^{k+1}} (x - \xi)^k (x - 1) \right| = \left| \left(\frac{\xi - x}{\xi} \right)^k \frac{x - 1}{\xi} \right| \leq (1 - x)^k \left| \frac{x - 1}{\xi} \right| \leq \frac{(1 - x)^{k+1}}{x} \tag{165}$$

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

$$\begin{aligned} \log x - L_k(x) &= \sum_{j=k+1}^{\infty} \frac{(-1)^{j+1}}{j} (x - 1)^j & (166) \\ &= (x - 1)^{k+1} (-1)^k \left(\frac{1}{k+1} - \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{k+1+j} (x - 1)^j \right) & (167) \end{aligned}$$

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

$$= -|x - 1|^{k+1} \left(\frac{1}{k+1} + \sum_{j=1}^{\infty} \frac{1}{k+1+j} |x - 1|^j \right). \tag{168}$$

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

$$\frac{1}{k+1} + \sum_{j=1}^{\infty} \frac{1}{k+1+j} |x - 1|^j \leq \frac{1}{k+1} + \sum_{j=1}^{\infty} \frac{1}{k+1+j} |e - 1|^j \tag{169}$$

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

$$\begin{aligned} \frac{1}{k+1} + \sum_{j=1}^{\infty} \frac{1}{k+1+j} |x - 1|^j &\leq \frac{1}{k+1} + \frac{|x - 1|}{k+2} + \sum_{j=2}^{\infty} \frac{|x - 1|^j}{k+1+j} & (170) \\ &\leq |x - 1| + \sum_{j=2}^{\infty} \frac{|x - 1|^j}{j} = -\log x, & (171) \end{aligned}$$

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

$$\frac{1}{k+1} + \sum_{j=1}^{\infty} \frac{1}{k+1+j} |x - 1|^j \leq \max(-\log x, \log e) = \max(-\log x, 1). \tag{172}$$

$\square$

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,*

$$\mathbb{E}\left[\log\left(\frac{1}{m}\sum_{i=1}^{m} U_i\right)\right] = \sum_{j=2}^{k} \frac{(-1)^{j+1}}{j} \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}(U_i - 1)\right)^j\right] + E_k \tag{173}$$

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

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

$$\left|\log\left(\frac{1}{m}\sum_{i=1}^{m}U_i\right) - L_k\left(\frac{1}{m}\sum_{i=1}^{m}U_i\right)\right| \leq \left|\frac{1}{m}\sum_{i=1}^{m}U_i\right|^{k+1} \max\left(-\log\left(\frac{1}{m}\sum_{i=1}^{m}U_i\right), 1\right). \tag{174}$$

We see that

$$\mathbb{E}\left[L_k\left(\frac{1}{m}\sum_{i=1}^{m}U_i\right)\right] = \sum_{j=1}^{k}\frac{(-1)^{j+1}}{j}\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}(U_i-1)\right)^j\right] \tag{175}$$

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

The error term E_k is bounded in L_1 by

$$\mathbb{E}[|E_k|] \leq \mathbb{E}\left[\left|\frac{1}{m}\sum_{i=1}^{m}U_i\right|^{k+1} \max\left(-\log\left(\frac{1}{m}\sum_{i=1}^{m}U_i\right), 1\right)\right] \tag{176}$$

apply Hölder's Inequality to give

$$\leq \mathbb{E}\left[\left|\frac{1}{m}\sum_{i=1}^{m}U_i\right|^{p(k+1)}\right]^{1/p} \mathbb{E}\left[\max\left(-\log\left(\frac{1}{m}\sum_{i=1}^{m}U_i\right), 1\right)^q\right]^{1/q}. \tag{177}$$

For the first term, Corollary 8 shows that

$$\mathbb{E}\left[\left|\frac{1}{m}\sum_{i=1}^{m}U_i\right|^{p(k+1)}\right]^{1/p} \leq D_{(k+1)p}^{1/p} m^{-(k+1)/2}\mathbb{E}\left[U_1^{(k+1)p}\right]^{1/p} \tag{178}$$

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

$$\max\left(-\log\left(\frac{1}{m}\sum_{i=1}^{m}U_i\right), 1\right)^q \leq \frac{1}{m}\sum_{i=1}^{m}\max\left(-\log\left(U_i\right), 1\right)^q \tag{179}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m}\left(|\log U_i| + 1\right)^q, \tag{180}$$

hence

$$\mathbb{E}\left[\max\left(-\log\left(\frac{1}{m}\sum_{i=1}^{m}U_i\right), 1\right)^q\right]^{1/q} \leq \left(\mathbb{E}\left[|\log U_1|^q\right] + 1\right)^{1/q}. \tag{181}$$

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

$$\mathbb{E}[|E_k|] \leq m^{-(k+1)/2}\, D_{(k+1)p}^{1/p}\mathbb{E}\left[U_1^{(k+1)p}\right]^{1/p} \left(\mathbb{E}\left[|\log U_1|^q\right] + 1\right)^{1/q}, \tag{182}$$

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 $p = 1, q = \infty$.

5 The generalized Donsker-Varadhan representation

5.1 Introduction

In this essay, we present a generalization of the classical Donsker-Varadhan representation of the KL-divergence (Donsker and Varadhan, 1975). Our purpose is twofold. Firstly, the new representation of the KL-divergence sheds further light on mutual information and may motivate the development of new statistical estimators of information. Secondly, our new representation is a powerful tool that connects a number of *existing* mutual information estimators under one umbrella. An important feature of the generalized Donsker-Varadhan representation is that it includes self-normalized bounds such as InfoNCE (van den Oord et al., 2018) as a special case, something which is not true of the classical Donsker-Varadhan representation.

5.2 Information-theoretic quantities

Throughout machine learning, we have cause to consider the entropy of probability measure p

$$\mathrm{H}\left(p\right) = \mathbb{E}_{p(\mathbf{x})}[-\log p(\mathbf{x})], \tag{183}$$

the KL divergence between two probability measures $p \ll q$

$$\mathrm{KL}\left(p \,\|\, q\right) = \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\right] \tag{184}$$

and the mutual information between jointly distributed random variables $\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$

$$I(\mathbf{x}, \mathbf{y}) = \mathrm{KL}\left(p(\mathbf{x}, \mathbf{y}) \,\|\, p(\mathbf{x})p(\mathbf{y})\right). \tag{185}$$

These are foundational quantities in information theory (Shannon, 1948), Bayesian experimental design (Lindley, 1956) and deep learning (Linsker, 1988). A key result in information theory is the following.

Theorem 12 (Gibbs' Inequality). *For any probability measures $p \ll q$, $\mathrm{KL}\left(p \,\|\, q\right) \geq 0$.*

5.3 The Donsker-Varadhan representation

An important lower bound on the KL divergence is the Donsker-Varadhan (DV) representation.

Theorem 13 (Donsker and Varadhan (1975)). *Let $p \ll q$ be probability measures on $\mathcal{X}$, then*

$$\mathrm{KL}\left(p \,\|\, q\right) = \sup_{T:\mathcal{X}\to\mathbb{R} \text{ measurable}} \mathbb{E}_{p(\mathbf{x})}[T(\mathbf{x})] - \log\left(\mathbb{E}_{q(\mathbf{x})}\left[\exp(T(\mathbf{x}))\right]\right) \tag{186}$$

One important bound that can be obtained as a consequence of the Donsker-Varadhan representation is the following.

Corollary 14 (Barber and Agakov (2003)). *Let $q(\mathbf{y}|\mathbf{x})$ be a conditional distribution. Then*

$$I(\mathbf{x}, \mathbf{y}) \geq \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\log \frac{q(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right] \tag{187}$$

Proof. Since mutual information is defined as a KL divergence, the DV representation is applicable. Let $T(\mathbf{x}, \mathbf{y}) = \log q(\mathbf{x}, \mathbf{y})/p(\mathbf{y})$ in Theorem 13. We have

$$\mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}[q(\mathbf{y}|\mathbf{x})/p(\mathbf{y})] = 1 \tag{188}$$

so the bound is *self-normalized*. The result follows. $\square$

The Barber-Agakov bound can be written as

$$I(\mathbf{x},\mathbf{y}) \geq \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\log q(\mathbf{y}|\mathbf{x})\right] + \mathrm{H}\left(p(\mathbf{y})\right) \tag{189}$$

which can be helpful in cases in which the $\mathrm{H}\left(p(\mathbf{y})\right)$ term is unknown but also unneeded for e.g. gradient estimation. Another bound, that appears in Nguyen et al. (2010); Nowozin et al. (2016); Belghazi et al. (2018) has a connection to the theory of f-divergences. Applying the inequality $\log x \leq e^{-1}x$ to Theorem 13 gives the **NWJ bound**

$$I(\mathbf{x},\mathbf{y}) \geq \mathbb{E}_{p(\mathbf{x})}[T(\mathbf{x})] - e^{-1}\mathbb{E}_{q(\mathbf{x})}\left[\exp(T(\mathbf{x}))\right]. \tag{190}$$

An advantage of this looser bound is that it can be directly estimated by samples.

5.4 A generalization of the Donsker-Varadhan representation

To generalize Theorem 13, suppose we extend the sample space to $\mathcal{X} \times \mathcal{S}$, where $\mathcal{S}$ represents 'side-information'. Suppose we have a conditional distribution $p(\mathbf{s}|\mathbf{x})$. Then we can extend the Donsker-Varadhan representation as follows.

Theorem 15 (Generalized Donsker-Varadhan representation)**.** *Under the assumptions of Theorem 13, let $p(\mathbf{s}|\mathbf{x})$ be a valid conditional distribution for each $\mathbf{x} \in \mathcal{X}$. Then,*

$$\mathrm{KL}\left(p \,\|\, q\right) = \sup_{U:\mathcal{X}\times\mathcal{S}\to\mathbb{R}\text{ measurable}} \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[U(\mathbf{x},\mathbf{s})] - \log\left(\mathbb{E}_{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}\left[\exp(U(\mathbf{x},\mathbf{s}))\right]\right) \tag{191}$$

Proof. Since any function $T : \mathcal{X} \to \mathbb{R}$ can be extended to a new function on $\mathcal{X} \times \mathcal{S}$ by ignoring the side information, Theorem 13 immediately tells us that

$$\mathrm{KL}\left(p \,\|\, q\right) \leq \sup_{U:\mathcal{X}\times\mathcal{S}\to\mathbb{R}\text{ measurable}} \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[U(\mathbf{x},\mathbf{s})] - \log\left(\mathbb{E}_{q(\mathbf{x})p(\mathbf{s},\mathbf{x})}\left[\exp(U(\mathbf{x},\mathbf{s}))\right]\right). \tag{192}$$

To prove the $\geq$ inequality, we consider some measurable $U : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$. We have

$$\mathrm{KL}\left(p \,\|\, q\right) = \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\right] \tag{193}$$

$$= \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}\left[\log \frac{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}\right] \tag{194}$$

define $V(\mathbf{x},\mathbf{s}) = \exp(U(\mathbf{x},\mathbf{s}))/\mathbb{E}_{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\exp(U(\mathbf{x},\mathbf{s}))]$

$$= \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}\left[\log \frac{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})V(\mathbf{x},\mathbf{s})}\right] + \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\log V(\mathbf{x},\mathbf{s})] \tag{195}$$

now note that by definition of V, $\int_{\mathcal{X}\times\mathcal{S}} q(\mathbf{x})p(\mathbf{x}|\mathbf{s})V(\mathbf{x},\mathbf{s}) = 1$, so $q(\mathbf{x})p(\mathbf{x}|\mathbf{s})V(\mathbf{x},\mathbf{s})$ is a probability measure

$$= \mathrm{KL}\left(p(\mathbf{x})p(\mathbf{s}|\mathbf{x}) \,\|\, q(\mathbf{x})p(\mathbf{s}|\mathbf{x})V(\mathbf{x}|\mathbf{s})\right) + \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\log V(\mathbf{x},\mathbf{s})] \tag{196}$$

now by Gibbs' Inequality

$$\geq \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\log V(\mathbf{x},\mathbf{s})] \tag{197}$$

$$= \mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[U(\mathbf{x},\mathbf{s})] - \log\left(\mathbb{E}_{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}\left[\exp(U(\mathbf{x},\mathbf{s}))\right]\right). \tag{198}$$

This completes the proof. $\square$

5.5 Self-normalized bounds

One particular use of Theorem 15 is for cases in which $\mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\exp(U(\mathbf{x},\mathbf{s}))] = 1$. For such a self-normalized bound, the task of estimating the potentially high-dimensional term $\mathbb{E}_{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\exp(U(\mathbf{x},\mathbf{s}))]$ is removed, and the bound reduces to $\mathbb{E}_{p(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[U(\mathbf{x},\mathbf{s})]$ for which unbiased estimators can be constructed directly from samples.

Theorem 16 (Self-normalized KL bound)**.** *Let $k : \mathcal{X} \to \mathbb{R}$ be any measurable function. Then we have the following bound on the KL divergence*

$$\mathrm{KL}\left(p \,\|\, q\right) \leq \mathbb{E}_{p(\mathbf{x}_1)q(\mathbf{x}_2)\ldots q(\mathbf{x}_m)}\left[\log \frac{\exp(k(\mathbf{x}_1))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))}\right]. \tag{199}$$

Proof. We apply Theorem 15 with $\mathbf{x} = \mathbf{x}_1$, $\mathcal{S} = \mathcal{X}^{m-1}$, $\mathbf{s} = (\mathbf{x}_2, ..., \mathbf{x}_m)$ and $p(\mathbf{s}|\mathbf{x}) = q(\mathbf{x}_2) \cdot ... \cdot q(\mathbf{x}_m)$ is independent of $\mathbf{x}_1$. We have

$$U(\mathbf{x},\mathbf{s}) = \log \frac{\exp(k(\mathbf{x}))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))} \tag{200}$$

To apply the theorem, we consider

$$\mathbb{E}_{q(\mathbf{x})p(\mathbf{s}|\mathbf{x})}[\exp(U(\mathbf{x},\mathbf{s}))] = \mathbb{E}_{q(\mathbf{x}_1)\ldots q(\mathbf{x}_m)}\left[\frac{\exp(k(\mathbf{x}))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))}\right]. \tag{201}$$

Since the $\mathbf{x}_1, ..., \mathbf{x}_m$ are all equal in distribution, we can replace the index of the sample used in the numerator by any $j \in \{1, ..., m\}$

$$= \mathbb{E}_{q(\mathbf{x}_1)\ldots q(\mathbf{x}_m)}\left[\frac{\exp(k(\mathbf{x}_j))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))}\right] \tag{202}$$

we can take the mean over all possible values of j

$$= \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{q(\mathbf{x}_1)\ldots q(\mathbf{x}_m)}\left[\frac{\exp(k(\mathbf{x}_j))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))}\right] \tag{203}$$

now by linearity of the expectation we have

$$= \mathbb{E}_{q(\mathbf{x}_1)\ldots q(\mathbf{x}_m)}\left[\frac{\frac{1}{m}\sum_{j=1}^{m}\exp(k(\mathbf{x}_j))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))}\right] \tag{204}$$

$$= 1. \tag{205}$$

Thus the bound is self-normalized and the result follows. □

We note that this bound cannot typically recover the KL divergence, because

$$\log \frac{\exp(k(\mathbf{x}_1))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i))} \leq \log \frac{\exp(k(\mathbf{x}))}{\frac{1}{m}\exp(k(\mathbf{x}))} = \log m. \tag{206}$$

We can apply a related idea to mutual information. The following theorem provides a self-normalized bound on $I(\mathbf{x},\mathbf{y})$ that is closely related to the popular InfoNCE (van den Oord et al., 2018) bound.

Theorem 17 (Self-normalized information bound)**.** *Let $k : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be any measurable function. Then we have the following bound on the mutual information*

$$I(\mathbf{x},\mathbf{y}) \leq \mathbb{E}_{p(\mathbf{x}_1,\mathbf{y}_1)p(\mathbf{x}_2)\ldots p(\mathbf{x}_m)}\left[\log \frac{\exp(k(\mathbf{x}_1,\mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^{m}\exp(k(\mathbf{x}_i,\mathbf{y}_1))}\right]. \tag{207}$$

Proof. Since $I(\mathbf{x}, \mathbf{y}) = \mathrm{KL}\left(p(\mathbf{x}, \mathbf{y}) \,\|\, p(\mathbf{x})p(\mathbf{y})\right)$, we can apply Theorem 15. We set $\mathcal{S} = \mathcal{X}^{m-1}$ and $\mathbf{s} = (\mathbf{x}_2, ..., \mathbf{x}_m)$. We have

$$U((\mathbf{x}_1, \mathbf{y}_1), \mathbf{s}) = \log \frac{\exp(k(\mathbf{x}_1, \mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \exp(k(\mathbf{x}_i, \mathbf{y}_1))}. \tag{208}$$

To show that this bound is self-normalized, we consider

$$\mathbb{E}_{p(\mathbf{x}_1)p(\mathbf{y}_1)p(\mathbf{s})}[\exp(U((\mathbf{x}_1, \mathbf{y}_1), \mathbf{s}))] = \mathbb{E}_{p(\mathbf{x}_1)...p(\mathbf{x}_m)p(\mathbf{y}_1)}\left[\frac{\exp(k(\mathbf{x}_1, \mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \exp(k(\mathbf{x}_i, \mathbf{y}_1))}\right], \tag{209}$$

for any $\ell \in \{1, ..., m\}$, we have

$$= \mathbb{E}_{p(\mathbf{x}_1)...p(\mathbf{x}_m)p(\mathbf{y}_1)}\left[\frac{\exp(k(\mathbf{x}_\ell, \mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \exp(k(\mathbf{x}_i, \mathbf{y}_1))}\right] \tag{210}$$

since the $\mathbf{x}_i$ are all equal in distribution. Then,

$$\begin{aligned}
&= \frac{1}{m}\sum_{\ell=1}^m \mathbb{E}_{p(\mathbf{x}_1)...p(\mathbf{x}_m)p(\mathbf{y}_1)}\left[\frac{\exp(k(\mathbf{x}_\ell, \mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \exp(k(\mathbf{x}_i, \mathbf{y}_1))}\right] && (211)\\
&= \mathbb{E}_{p(\mathbf{x}_1)...p(\mathbf{x}_m)p(\mathbf{y}_1)}\left[\frac{\frac{1}{m}\sum_{\ell=1}^m \exp(k(\mathbf{x}_\ell, \mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \exp(k(\mathbf{x}_i, \mathbf{y}_1))}\right] && (212)\\
&= 1. && (213)
\end{aligned}$$

This completes the proof. ☐

Finally, it is possible to change the distribution that is used to generate $\mathbf{s}$ as long as we compensate with importance weighting. The following theorem gives a bound that is closely connected to the likelihood-free Adaptive Contrastive Estimation bound of Foster et al. (2020) eq. (14).

Theorem 18 (Importance weighted self-normalized information bound)**.** *Let $k : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be any measurable function. Consider a conditional distribution $q(\mathbf{x}'|\mathbf{y})$ on $\mathcal{X}$. Then we have the following bound on the mutual information*

$$I(\mathbf{x}, \mathbf{y}) \leq \mathbb{E}_{p(\mathbf{x}_1, \mathbf{y}_1)q(\mathbf{x}_2|\mathbf{y}_1)...q(\mathbf{x}_m|\mathbf{y}_1)}\left[\log \frac{\exp(k(\mathbf{x}_1, \mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \frac{\exp(k(\mathbf{x}_i, \mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}}\right]. \tag{214}$$

Proof. Following the same strategy as the previous two proofs, we consider

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{x}_1)p(\mathbf{y}_1)p(\mathbf{s})}[\exp(U((\mathbf{x}_1, \mathbf{y}_1), \mathbf{s}))] &= \mathbb{E}_{p(\mathbf{x}_1)p(\mathbf{y}_1)q(\mathbf{x}_{2:m}|\mathbf{y}_1)}\left[\frac{\exp(k(\mathbf{x}_1, \mathbf{y}_1))}{\frac{1}{m}\sum_{i=1}^m \frac{\exp(k(\mathbf{x}_i, \mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}}\right] && (215)\\
&= \mathbb{E}_{p(\mathbf{y}_1)q(\mathbf{x}_{1:m}|\mathbf{y}_1)}\left[\frac{\frac{\exp(k(\mathbf{x}_1, \mathbf{y}_1))p(\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{y}_1)}}{\frac{1}{m}\sum_{i=1}^m \frac{\exp(k(\mathbf{x}_i, \mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}}\right] && (216)
\end{aligned}$$

for any $\ell \in \{1, ..., m\}$, we have

$$= \mathbb{E}_{p(\mathbf{y}_1)q(\mathbf{x}_{1:m}|\mathbf{y}_1)}\left[\frac{\frac{\exp(k(\mathbf{x}_\ell, \mathbf{y}_1))p(\mathbf{x}_\ell)}{q(\mathbf{x}_\ell|\mathbf{y}_1)}}{\frac{1}{m}\sum_{i=1}^m \frac{\exp(k(\mathbf{x}_i, \mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}}\right] \tag{217}$$

since the $\mathbf{x}_i$ are all now equal in distribution. Then,

$$= \frac{1}{m}\sum_{\ell=1}^m \mathbb{E}_{p(\mathbf{y}_1)q(\mathbf{x}_{1:m}|\mathbf{y}_1)}\left[\frac{\frac{\exp(k(\mathbf{x}_\ell, \mathbf{y}_1))p(\mathbf{x}_\ell)}{q(\mathbf{x}_\ell|\mathbf{y}_1)}}{\frac{1}{m}\sum_{i=1}^m \frac{\exp(k(\mathbf{x}_i, \mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}}\right] \tag{218}$$

$$= \mathbb{E}_{p(\mathbf{y}_1)q(\mathbf{x}_{1:m}|\mathbf{y}_1)} \left[\frac{\frac{1}{m} \sum_{\ell=1}^{m} \frac{\exp(k(\mathbf{x}_\ell, \mathbf{y}_1))p(\mathbf{x}_\ell)}{q(\mathbf{x}_\ell|\mathbf{y}_1)}}{\frac{1}{m} \sum_{i=1} \frac{\exp(k(\mathbf{x}_i, \mathbf{y}_1))p(\mathbf{x}_i)}{q(\mathbf{x}_i|\mathbf{y}_1)}} \right] \tag{219}$$

$$= 1. \tag{220}$$

This completes the proof. □

A limitation of this bound is that we need to know the density $p(\mathbf{x})$.

References

David Barber and Felix Agakov. The IM algorithm: a variational approach to information maximization. *Advances in Neural Information Processing Systems*, 16:201–208, 2003.

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.

Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R Devon Hjelm, and Aaron Courville. MINE: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, 6(5):679–684, 1957.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Daniel R Cavagnaro, Jay I Myung, Mark A Pitt, and Janne V Kujala. Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural computation*, 22 (4):887–905, 2010.

Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian q-learning. In *Aaai/iaai*, pages 761–768, 1998.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Michael O'Gordon Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002.

Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. Variational Bayesian Optimal Experimental Design. In *Advances in Neural Information Processing Systems 32*, pages 14036–14047. Curran Associates, Inc., 2019.

Adam Foster, Martin Jankowiak, Matthew O'Meara, Yee Whye Teh, and Tom Rainforth. A unified stochastic gradient approach to designing bayesian-optimal experiments. In *International Conference on Artificial Intelligence and Statistics*, pages 2959–2969. PMLR, 2020.

Adam Foster, Desi R Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. *arXiv preprint arXiv:2103.02438*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *arXiv preprint arXiv:1609.04436*, 2016.

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, and Wataru Kitade. Unbiased mlmc stochastic gradient-based optimization of bayesian experimental designs. *arXiv preprint arXiv:2005.08414*, 2020b.

Arthur Guez, David Silver, and Peter Dayan. Efficient bayes-adaptive reinforcement learning using sample-based search. In *Advances in neural information processing systems*, pages 1025–1033, 2012.

Markus Hainy, David J Price, Olivier Restif, and Christopher Drovandi. Optimal bayesian design for model discrimination via classification. *arXiv preprint arXiv:1809.05301*, 2018.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Taylor A Howell, Chunjiang Fu, and Zachary Manchester. Direct policy optimization using deterministic sampling and collocation. *IEEE Robotics and Automation Letters*, 6(3):5324–5331, 2021.

Xun Huan and Youssef M Marzouk. Sequential bayesian optimal experimental design via approximate dynamic programming. *arXiv preprint arXiv:1604.08320*, 2016.

Aapo Hyvärinen. Survey on independent component analysis. 1999.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. In *International Conference on Machine Learning*, pages 2117–2126. PMLR, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

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.

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

Guy Lorberbom, Chris J Maddison, Nicolas Heess, Tamir Hazan, and Daniel Tarlow. Direct policy gradients: Direct optimization of policies in discrete action spaces. *arXiv preprint arXiv:1906.06062*, 2019.

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

Ruth K Meyer and Christopher J Nachtsheim. The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, 37(1):60–69, 1995.

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

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016.

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.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayes-Adaptive POMDPs. In *NIPS*, pages 1225–1232, 2007.

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.

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

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.

Joep Vanlier, Christian A Tiemann, Peter AJ Hilbers, and Natal AW van Riel. Optimal experiment design for model selection in biochemical networks. *BMC systems biology*, 8(1):1–16, 2014.

Benjamin T Vincent and Tom Rainforth. The DARC toolbox: automated, flexible, and efficient delayed and risky choice experiments using bayesian adaptive design. *Retrieved from psyarxiv.com/yehjb*, 2017.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

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.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.