Author Atrem Sobolev (http://artem.sobolev.name/pages/about.html) is at Bayesian Methods Research Group (https://bayesgroup.ru/), Moscow

Thoughts on Mutual Information: More Estimators

August 10, 2019

In this post I'd like to show how Self-Normalized Importance Sampling (IWHVI and IWAE) and Annealed Importance Sampling can be used to give (sometimes sandwich) bounds on the MI in many different cases.

Mutual Information (MI) is an important concept from the Information Theory that captures the idea of information one random variable X carries about the r.v. Z and is usually denoted $I(X, Z)$, however in order to emphasize the underlying joint distribution I'll be using a non-standard notation $\mathrm{MI}[p(x, z)]$. Formally the MI has the following definition:

$$\mathrm{MI}[p(x,z)] := \mathop{\mathbb{E}}_{p(x,z)} \log \frac{p(x,z)}{p(x)p(z)} = \mathop{\mathbb{E}}_{p(x,z)} \log \frac{p(x \mid z)}{p(x)}$$

Having such a nice information-measuring interpretation, MI is a natural objective and/or metric in many problems in Machine Learning. One particular application is evaluatiuon of encoder-decoder-like architectures as MI quantifies amount of information contained in the code. In particular, in Variational Autoencoders a good decoder should have high $\mathrm{MI}[p(x, z)]$, meaning the code is very useful for generations. A good encoder though... has to keep balance between providing enough information to the decoder, while not deviating too much from the prior by, for example, encoding redundant / unnecessary information. [1]

MI Estimation

In general estimating the MI is intractable as it requires knowing the log marginal density $\log p(x)$ or the intractable log posterior $\log p(z|x)$. However, there are many efficient variational bounds which can be employed to give tractable lower or upper bounds on the MI. Many existing bounds are reviewed in the On Variational Bounds of Mutual Information paper.

It is important to take into account what we know about the joint distribution of x and z. We'll consider several nested "layers" of decreasing complexity:

1. **Blackbox** case: Distributions $p(x, z)$ we can only sample from, but don't know any densities. This way we can form Monte Carlo estimates after some learning and surprisingly we can give some lower bounds already in this case.
2. **Known conditional** case: Distributions $p(x, z)$ we can sample from and know one conditional

distribution, say, $p(x|z)$.

3. **Known marginal** case: Distributions $p(x, z)$ we can sample from and know one marginal distribution, say, $p(z)$.
4. **Known joint** case: Distributions $p(x, z)$ we can sample from and know both a marginal and a conditional distributions, say, $p(z)$ and $p(x|z)$.
5. **Known everything** case: Distributions $p(x, z)$ which we can sample from and know all conditionals and marginals. This is a trivial case and doesn't require any bounds. The MI can be estimated using Monte Carlo directly, so we'll omit it from the discussion.

Bounds based on Self-Normalized Importance Sampling

Self-Normalized Importance Sampling (SNIS) has been shown (see my previous posts) to give both lower and upper bounds on the marginal log-likelihood:

$$\text{IWAE:} \qquad \log p(x) \geq \mathop{\mathbb{E}}_{q(z_{1:K}|x)} \log \frac{1}{K} \sum_{k=1}^{K} \frac{p(x, z_k)}{q(z_k|x)}$$

$$\text{IWHVI:} \qquad \log p(x) \leq \mathop{\mathbb{E}}_{p(z_0|x)} \mathop{\mathbb{E}}_{q(z_{1:K}|x)} \log \frac{1}{K+1} \sum_{k=0}^{K} \frac{p(x, z_k)}{q(z_k|x)}$$

We use such bounds to give sandwich bound on the intractable entropy term in the MI. Another useful insight is that

$$\omega(z_{0:K}|x) := \frac{\hat{\rho}(x, z_0)\tau(z_{1:K}|x)}{\frac{1}{K+1}\sum_{k=0}^{K} \frac{\hat{\rho}(x, z_k)}{\tau(z_k|x)}}$$

is a distribution (a valid pdf, to be precise) for almost any (the only condition is the same as in the standard importance sampling) unnormalized distribution $\hat{\rho}(x, z)$ (a joint distribution over z and x) and a normalized distribution $\tau(z|x)$ (a distribution over z possibly conditioned on x). The fact that $\omega(z_{0:K}|x)$ is a valid distribution allows us to consider the following KL divergence:

$$0 \leq \mathrm{KL}(p(x, z_0)\tau(z_{1:K}|x) \;||\; p(x)\omega(z_{0:K}|x)) = \mathop{\mathbb{E}}_{p(x, z_0)} \mathop{\mathbb{E}}_{\tau(z_{1:K}|x)} \log \frac{p(x, z_0)\tau(z_{1:K}|x)}{p(x)\omega(z_{0:K}|x)}$$

Which gives the following lower bound on the MI:

$$\begin{aligned}\mathop{\mathbb{E}}_{p(x,z_0)}\mathop{\mathbb{E}}_{\tau(z_{1:K}|x)}\log\frac{p(x|z_0)}{p(x)} &\geq \mathop{\mathbb{E}}_{p(x,z_0)}\mathop{\mathbb{E}}_{\tau(z_{1:K}|x)}\log\frac{\omega(z_{0:K}|x)}{p(z_0)\tau(z_{1:K}|x)} \\ &= \mathop{\mathbb{E}}_{p(x,z_0)}\mathop{\mathbb{E}}_{\tau(z_{1:K}|x)}\log\frac{\hat{\rho}(x,z_0)}{\frac{1}{K+1}\sum_{k=0}^{K}\frac{\hat{\rho}(x,z_k)}{\tau(z_k|x)}} - \mathop{\mathbb{E}}_{p(z)}\log p(z)\end{aligned}$$

Equivalently, by reparametrizing the bound in terms of $\hat{\varrho}(x|z) = \hat{\rho}(x,z)/p(z)$ we have

$$\mathop{\mathbb{E}}_{p(x,z_0)}\mathop{\mathbb{E}}_{\tau(z_{1:K}|x)}\log\frac{p(x|z_0)}{p(x)} \geq \mathop{\mathbb{E}}_{p(x,z_0)}\mathop{\mathbb{E}}_{\tau(z_{1:K}|x)}\log\frac{\hat{\varrho}(x|z_0)}{\frac{1}{K+1}\sum_{k=0}^{K}\hat{\varrho}(x|z_k)\frac{p(z_k)}{\tau(z_k|x)}}$$

These lower bounds work for any ρ and ϱ, and the optimal choices are $p(x,z)$ and $p(x|z)$, correspondingly.

Known Joint

First, we'll consider the easiest case – when we know the joint distribution in the form of a prior + conditional. A prominent example of this class is VAEs decoders, which is defined by some prior in the latent space $p(z)$ and a decoder $p_\theta(x|z)$ that uses a neural network to generate a distribution over observations x for a particular z. Computing the MI of the decoder is arguably a natural way to measure the extent of posterior collapse[2], since MI can be expressed in the following form, which essentially measures the true posterior's deviation from the prior:

$$\mathrm{MI}[p(x,z)] := \mathop{\mathbb{E}}_{p(x,z)}\log\frac{p(z|x)}{p(z)} = \mathop{\mathbb{E}}_{p(x)} D_{KL}\left(p(z|x)\;||\;p(z)\right)$$

However the MI as introduced above requires knowing marginal $p(x)$, which is intractable. Luckily, the Multisample Variational Bounds allow us to give efficient variational sandwich bounds on MI (where $q_\phi(z|x)$ is an encoder with known density[3]):

$$\mathop{\mathbb{E}}_{\substack{p_\theta(x,z_0)\\ q_\phi(z_{1:K}|x)}}\log\frac{p_\theta(x|z_0)}{\frac{1}{K+1}\sum_{k=0}^{K}\frac{p_\theta(x,z_k)}{q_\phi(z_k|x)}} \leq \mathrm{MI}[p_\theta(x,z)] \leq \mathop{\mathbb{E}}_{\substack{p_\theta(x,z_0)\\ q_\phi(z_{1:K}|x)}}\log\frac{p_\theta(x|z_0)}{\frac{1}{K}\sum_{k=1}^{K}\frac{p_\theta(x,z_k)}{q_\phi(z_k|x)}}$$

Known Conditional

However, we might not have any of marginal densities in the closed form. For example, this is the case if we want to estimate the amount of information the encoder $q_\phi(z|x)$ puts into the code z. Since it only defines the conditional distribution, we pair it with some data-generating process $p(x)$ which defines the population we'd like to evaluate the encoder over.

Direct application of the aforementioned bounds leads to (where $\tau_\eta(x_k|z)$ is our variational inverse distribution)

$$\mathop{\mathbb{E}}_{\substack{p(x_0)\\q_\phi(z|x_0)\\\tau_\eta(x_{1:K}|z)}} \log \frac{q_\phi(z|x_0)}{\frac{1}{K+1}\sum_{k=0}^{K}\frac{q_\phi(z|x_k)p(x_k)}{\tau_\eta(x_k|z)}} \le \mathrm{MI}[q_\phi(z|x)p(x)] \le \mathop{\mathbb{E}}_{\substack{p(x_0)\\q_\phi(z|x_0)\\\tau_\eta(x_{1:K}|z)}} \log \frac{q_\phi(z|x_0)}{\frac{1}{K}\sum_{k=1}^{K}\frac{q_\phi(z|x_k)p(x_k)}{\tau_\eta(x_k|z)}}$$

However, we might not have an access to the density $p(x)$. In this case one can resort to SIVI-like bounds by setting $\tau_\eta(x|z) = p(x)$ and arrive to the following bounds:

$$\mathop{\mathbb{E}}_{\substack{p(x_{0:K})\\q_\phi(z|x_0)}} \log \frac{q_\phi(z|x_0)}{\frac{1}{K+1}\sum_{k=0}^{K} q_\phi(z|x_k)} \le \mathrm{MI}[q_\phi(z|x)p(x)] \le \mathop{\mathbb{E}}_{\substack{p(x_{0:K})\\q_\phi(z|x_0)}} \log \frac{q_\phi(z|x_0)}{\frac{1}{K}\sum_{k=1}^{K} q_\phi(z|x_k)}$$

These bounds are known as special case of InfoNCE bounds and are much worse due to uninformed proposal τ.

Known Prior

Sometimes we use complex implicit models as decoders and don't have closed-form densities for $p(x|z)$. The most popular instance of such models is Generative Adversarial Networks, which is similar to VAE with an exception of not having a well-defined decoder's density $p(x|z)$ (but the prior $p(z)$ is typically simple and known). In some sense, this density is degenerate: $p(x|z) = \delta(x - f(z))$ where f is generator's neural network. Unfortunately, we cannot use such density in the IWHVI bounds. Are we doomed then? Turns out, not quite so. Even in this case it's still possible to give an efficient multisample variational lower bound:

$$\mathrm{MI}[p(x,z)] \ge \mathop{\mathbb{E}}_{p(x,z_0)} \mathop{\mathbb{E}}_{q_\phi(z_{1:K}|x)} \log \frac{\hat\rho_\eta(x|z_0)}{\frac{1}{K+1}\sum_{k=0}^{K}\hat\rho_\eta(x|z_k)\frac{p(z_k)}{q_\phi(z_k|x)}}$$

Where $\hat\rho_\eta(x|z) = \exp(h_\eta(x|z))$ is any non-normalized energy-based model that essentially estimates the unknown density $p(x|z)$.

I am not aware of any good *upper* bounds and if they are possible. I would think the answer is negative due to hardness of upper-bounding the crossentropy in the black-box case.

Known Nothing

Staying in the realm of implicit models, let's assume we trained an implicit inference model (think of GANs with encoders) and would like estimate the MI much like in the case of VAE's encoder in the

"known conditional" section. Denoting our inference model $q(z|x)$ and the data-generating process $p(x)$ (both densities are unknown to us, but we can sample from them), we can adapt the previous section's lower bound by choosing the proposal $\tau_\eta(x|z) = p(x)$

$$\mathrm{MI}[q(z|x)p(x)] \geq \mathop{\mathbb{E}}_{p(x_{0:K})} \mathop{\mathbb{E}}_{q(z|x_0)} \log \frac{\hat{\rho}_\eta(z|x_0)}{\frac{1}{K+1}\sum_{k=0}^{K} \hat{\rho}_\eta(z|x_k)}$$

Where $\hat{\rho}_\eta(z|x) = \exp(h_\eta(z|x))$ is again a non-normalized energy-based model that estimates the unknown density $q(z|x)$.

While convenient in its wide applicability, this bound is known to be very loose in cases when the true MI is high. We'll discuss drawbacks and limitations in the next post on the topic.

Bounds based on Annealed Importance Sampling

SNIS is not the only way to obtain variational sandwich bounds on log marginal likelihood. Another widely known and powerful approach is Annealed Importance Sampling (AIS)

AIS uses two distributions, called forward and backward:

$$\begin{aligned}
q_{\rightarrow}(z_{1:T}|x) &= q(z_1|x)\mathcal{T}_2(z_2 \mid z_1, x)\cdots\mathcal{T}_T(z_T \mid z_{T-1}, x) \\
q_{\leftarrow}(z_{T:1}|x) &= p(z_T|x)\mathcal{T}_T(z_{T-1} \mid z_T, x)\cdots\mathcal{T}_2(z_1 \mid z_2, x) \\
q_{\leftarrow}(x, z_{T:1}) &= p(x, z_T)\mathcal{T}_T(z_{T-1} \mid z_T, x)\cdots\mathcal{T}_2(z_1 \mid z_2, x)
\end{aligned}$$

Where $\mathcal{T}_t$ is a transition operator that is designed to be invariant to $p_t(z|x) \propto q(z|x)^{1-\beta_t}p(x,z)^{\beta_t}$ and $\beta_{1:T+1}$ is a monotonically increasing sequence s.t. $\beta_1 = 0$ and $\beta_{T+1} = 1$. That is, in the *forward distribution* $q_{\rightarrow}(z_{1:T}|x)$ one starts with a sample z_1 from some proposal $q(z|x)$ and then transforms it into a sample from $p_2(z|x)$ using the $\mathcal{T}_t$ transition operator (typically a MCMC kernel). The sample z_2 is then analogously transformed into z_3 amd so on. The *backward distribution* $q_{\rightarrow}(z_{T:1}|x)$ is similar except it starts with the true posterior sample $z_T \sim p(z|x)$ and then sequentially transforms it into a sample from the proposal $z_1 \sim q(z|x)$.

Then one defines the importance weight

$$
\begin{aligned}
w(z_{1:T} \mid x) &= \frac{q_{\leftarrow}(z_{T:1}|x)}{q_{\rightarrow}(z_{1:T}|x)} = \frac{\hat{p}_2(z_1 \mid x)}{q(z_1|x)} \frac{\hat{p}_3(z_2 \mid x)}{\hat{p}_2(z_2 \mid x)} \cdots \frac{p(x, z_T)}{\hat{p}_T(z_T \mid x)} \\
&= \frac{\left(\frac{p(x,z_1)}{q(z_1|x)}\right)^{\beta_2} q(z_1|x)}{q(z_1|x)} \frac{\left(\frac{p(x,z_2)}{q(z_2|x)}\right)^{\beta_3} q(z_2|x)}{\left(\frac{p(x,z_2)}{q(z_2|x)}\right)^{\beta_2} q(z_2|x)} \cdots \frac{\left(\frac{p(x,z_T)}{q(z_T|x)}\right)^{\beta_{T+1}} q(z_T|x)}{\left(\frac{p(x,z_T)}{q(z_T|x)}\right)^{\beta_T} q(z_T|x)} \\
&= \left(\frac{p(x,z_1)}{q(z_1|x)}\right)^{\beta_2-\beta_1} \left(\frac{p(x,z_2)}{q(z_2|x)}\right)^{\beta_3-\beta_2} \cdots \left(\frac{p(x,z_T)}{q(z_T|x)}\right)^{\beta_{T+1}-\beta_T}
\end{aligned}
$$

Where the second identity is due to $\mathcal{T}_t$ satisfying the detailed balance equation. Then one can show that

$$
\begin{aligned}
\mathop{\mathbb{E}}_{q_{\rightarrow}(z_{1:T}|x)} w(z_{1:T} \mid x) = p(x) \quad &\Rightarrow \quad \mathop{\mathbb{E}}_{q_{\rightarrow}(z_{1:T}|x)} \log w(z_{1:T} \mid x) \leq \log p(x), \\
\mathop{\mathbb{E}}_{q_{\leftarrow}(z_{T:1}|x)} \frac{1}{w(z_{1:T} \mid x)} = \frac{1}{p(x)} \quad &\Rightarrow \quad \mathop{\mathbb{E}}_{q_{\leftarrow}(z_{T:1}|x)} \log w(z_{1:T} \mid x) \geq \log p(x)
\end{aligned}
$$

Which gives us another set of sandwich bounds, which we can use to sandwich bound the MI:

$$
\boxed{\mathop{\mathbb{E}}_{q_{\leftarrow}(x, z_{T:1}|x)} \log \frac{p_\theta(x|z_T)}{w(z_{1:T} \mid x)} \leq \mathrm{MI}[p_\theta(x, z)] \leq \mathop{\mathbb{E}}_{p_\theta(x,z_0)} \mathop{\mathbb{E}}_{q_{\rightarrow}(z_{1:T}|x)} \log \frac{p_\theta(x|z_0)}{w(z_{1:T} \mid x)}}
$$

One can also come up with a “decoder-free” version of the bound in a similar fashion to what we’ve done above. First, introduce the following distributions:

$$
\begin{aligned}
\gamma_{\rightarrow}(z_{1:T}|x) &= q(z_1|x)\mathcal{K}_2(z_2 \mid z_1, x) \cdots \mathcal{K}_T(z_T \mid z_{T-1}, x) \\
\gamma_{\leftarrow}(x, z_{T:1}) &= p(x, z_T)\mathcal{K}_T(z_{T-1} \mid z_T, x) \cdots \mathcal{K}_2(z_1 \mid z_2, x)
\end{aligned}
$$

Where $\mathcal{K}_t$ is now tailored to $\kappa_t(z|x) \propto q(z|x)^{1-\beta_t} \left(\hat{\varrho}(x|z)p(z)\right)^{\beta_t}$ for certain unnormalized $\hat{\varrho}(x|z)$

Now consider

$$
\begin{aligned}
0 \le \mathrm{KL}\left(\gamma_{\leftarrow}(x, z_{T:1}) \;||\; p(x)\gamma_{\rightarrow}(z_{1:T}|x)\right) &= \mathbb{E}_{\gamma_{\leftarrow}(x,z_{T:1})} \log \frac{\gamma_{\leftarrow}(x, z_{T:1})}{p(x)\gamma_{\rightarrow}(z_{1:T}|x)} \\
&= \mathbb{E}_{\gamma_{\leftarrow}(x,z_{T:1})} \log \left[\frac{p(x, z_T)}{p(x)q(z_1|x)} \frac{\mathcal{K}_2(z_1 \mid z_2, x)}{\mathcal{K}_2(z_2 \mid z_1, x)} \cdots \frac{\mathcal{K}_T(z_{T-1} \mid z_T, x)}{\mathcal{K}_T(z_T \mid z_{T-1}, x)} \right] \\
&= \mathbb{E}_{\gamma_{\leftarrow}(x,z_{T:1})} \log \left[\frac{p(x, z_T)}{p(x)q(z_1|x)} \frac{\kappa_2(z_1|x)}{\kappa_2(z_2|x)} \cdots \frac{\kappa_T(z_{T-1}|x)}{\kappa_T(z_T|x)} \right] \\
&= \mathbb{E}_{\gamma_{\leftarrow}(x,z_{T:1})} \log \left[\frac{p(x,z_T)}{p(x)q(z_1|x)} \frac{q(z_1|x)\left(\frac{\hat{\varrho}(x|z_1)p(z_1)}{q(z_1|x)}\right)^{\beta_2}}{q(z_2|x)\left(\frac{\hat{\varrho}(x|z_2)p(z_2)}{q(z_2|x)}\right)^{\beta_2}} \cdots \frac{q(z_{T-1}|x)\left(\frac{\hat{\varrho}(x|z_{T-1})p(z_{T-1})}{q(z_{T-1}|x)}\right)^{\beta_T}}{q(z_T|x)\left(\frac{\hat{\varrho}(x|z_T)p(z_T)}{q(z_T|x)}\right)^{\beta_T}} \right] \\
&= \mathbb{E}_{\gamma_{\leftarrow}(x,z_{T:1})} \log \left[\frac{p(x, z_T)}{p(x)} \frac{1}{p(z_T)\hat{\varrho}(x|z_T)} \prod_{t=1}^{T} \left(\frac{\hat{\varrho}(x|z_t)p(z_t)}{q(z_t|x)} \right)^{\beta_{t+1}-\beta_t} \right]
\end{aligned}
$$

Hence

$$
\mathrm{MI}[p(x,z)] \ge \mathbb{E}_{\gamma_{\leftarrow}(x,z_{T:1})} \log \frac{\hat{\varrho}(x|z_T)}{\prod_{t=1}^{T} \left(\frac{\hat{\varrho}(x|z_t)p(z_t)}{q(z_t|x)} \right)^{\beta_{t+1}-\beta_t}}
$$

Now we can again reparametrize this formula in terms of $\hat{\rho}(x,z) = \hat{\varrho}(x|z)p(z)$:

$$
\mathrm{MI}[p(x,z)] \ge \mathbb{E}_{\gamma_{\leftarrow}(x,z_{T:1})} \log \frac{\hat{\rho}(x, z_T)}{\prod_{t=1}^{T} \left(\frac{\hat{\rho}(x,z_t)}{q(z_t|x)} \right)^{\beta_{t+1}-\beta_t}} - \mathbb{E}_{p(z)} \log p(z)
$$

It's tempting to simply put $q(z|x) = p(z)$ to obtain a blackbox AIS-based analogue of the InfoNCE bound, however notice that $\gamma_{\leftarrow}(x, z_{T:1})$ relies on an MCMC kernel that gradually transforms a sample $z_T \sim p(z_T|x)$ into $z_1 \sim p(z_1)$ and thus needs to know this density. Thus I don't think one can use AIS in the blackbox mode.

Finally, note while the bound is valid for any $\hat{\rho}$ and $\hat{\varrho}$, it's not quite differentiable w.r.t. their parameters as any proper MCMC method would require an accept-reject step which is not differentiable. Thus if one seeks to learn $\hat{\rho}$ or $\hat{\varrho}$, an alternative objective should be used. Luckily, the SNIS-based lower bound with small K would work just fine.

Conclusion

I presented two different ways to give bounds on the MI in cases of variable complexity. The SNIS-based approach seem to be somewhat novel (the special case of InfoNCE has been already known), and is applicable in many different problems. The AIS-based one is based on well-known sandwich bounds

on the log marginal likelihoods, but I haven't seen it being applied to the problem of MI estimation. The reason might be that it's more restrictive that the SNIS-based estimations: AIS only works for continuous variables, requires complicated MCMC to function and does not seem to allow blackbox estimators. On the positive side of things, AIS-based estimator should perform much better in high-dimensional problems with large MI, especially if one uses gradient-based kernels like HMC.

Next, I'll share some of my thoughts on drawbacks of these (and some other) bounds, in particular in light of the notorious "Formal Limitations" paper.

---

1. See ELBO surgery: yet another way to carve up the variational evidence lower bound↩

2. Oftentimes $D_{KL}(q_\phi(z|x) \mid\mid p(z))$ is used to measure the extent of the so called posterior collapse. One can argue this is an indirect metric as it does not use *decoder at all* and relies on the encoder's ability to approximate the true posterior better than the prior. Moreover, high $D_{KL}(q_\phi(z|x) \mid\mid p(z))$ is likely to be an indicator of poor encoder $q(z|x)$, whereas the Mutual Information is monotonic in decoder's quality.↩

3. As I have argued in the IWHVI paper, it might be beneficial to fit two different encoders for lower and upper bounds, correspondingly.↩

Thoughts on Mutual Information: Formal Limitations

August 14, 2019

This posts continues the discussion started in the Thoughts on Mutual Information: More Estimators. This time we'll focus on drawbacks and limitations of these bounds.

Let's start with a elephant in the room: a year ago an interesting preprint has been uploaded to arxiv: Formal Limitations on the Measurement of Mutual Information in which authors essentially argue that if you don't know any densities (the hardest case, according to my hierarchy), then any **distribution-free high-confidence lower bound** on the MI would require $\exp(\mathrm{MI})$ number of samples and thus black-box MI lower bounds should be deemed impractical.

Formal Limitations

The paper mounts a massive attack on distribution-free lower bounds on the Mutual Information. Not only McAllester and Stratos show that existing bounds are inferior, but also kill any blackbox lower bounds on the KL divergence. The core result that warrants impossibility of cheap and good lower bounds on the Mutual Information is the Theorem 2, which states (in a slightly reformulated notation)

> ***Theorem**: Let B be any distribution-free high-confidence lower bound on $\mathbb{H}[p(x)]$ computed from a sample $x_{1:N} \sim p(x)$ More specifically, let $B(x_{1:N}, \delta)$ be any real-valued function of a sample and a confidence parameter δ such that for any $p(x)$, with probability at least $(1-\delta)$ over a draw of $x_{1:N}$ from $p(x)$, we have*
>
> $$\mathbb{H}[p(x)] \geq B(x_{1:N}, \delta).$$
>
> *For any such bound, and for $N \geq 50$ and $k \geq 2$, with probability at least $1 - \delta - 1.01/k$ over the draw of $x_{1:N}$ we have*
>
> $$B(x_{1:N}, \delta) \leq \log(2kN^2)$$

Indeed, since (in discrete case) $I(X, X) = H(X)$ [1], any good[2] lower bound on the MI would give a good lower bound on the entropy, and the theorem above says there are no such bounds (only those that are either exponentially expensive to compute or are not high-confidence or are not black-box). Authors then argue that one can have good estimators if they forgo the lower bound guarantee and settle with an estimate that is neither a lower or an upper bound. However, this is undesirable in many cases, especially when we'd like to compare two numbers.

Unfortunately, I found the paper hard to digest and as far as I know, it's still not published, so probably we should be cautious about the presented result. Nevertheless, I'll show below that several often used bounds do indeed seem to have this limitation.

The Nguyen-Wainwright-Jordan Bound

The process we've followed so far to derive a lower bound on the MI has been somewhat cumbersome: we first decomposed the MI into some expectation and then used fancy bounds on some of the terms. An alternative and easier approach is to recall that the MI is a certain KL divergence take any off-the-shelf lower bounds on the KL divergence.

One such lower bound can be obtained using the Fenchel conjugate functions (Nguyen et al., alternatively see the f-GANs paper):

$$KL(p(x) \,||\, q(x)) \geq \mathbb{E}_{p(x)} f(x) - \mathbb{E}_{q(x)} \exp(f(x)) + 1$$

Where $f(x)$ (a critic) is any function that takes x as input and outputs a scalar. The optimal choice can be shown to be $f^*(x) = \ln \frac{p(x)}{q(x)}$. And all is nice, except the lurking menace of the exp term. Consider a Monte Carlo estimate in the case of optimal critic ($x_{1:N} \sim p(x), y_{1:M} \sim q(y)$):

$$\frac{1}{N}\sum_{n=1}^{N} \ln \frac{p(x_n)}{q(x_n)} - \frac{1}{M}\sum_{m=1}^{M} \frac{p(y_m)}{q(y_m)} + 1$$

The first term is exactly the Monte Carlo estimate of the KL divergence, while the second (the balancing term as it's counterweights the first one) in expectation gives 1. However, ratio of densities might take on extremely large values and in general has enormous variance. Indeed, the variance of the balancing term is

$$\begin{aligned}
\mathbb{V}_{q(y_{1:M})}\left[\frac{1}{M}\sum_{m=1}^{M} \frac{p(y_m)}{q(y_m)}\right] &= \frac{1}{M}\mathbb{V}_{q(y)}\left[\frac{p(y)}{q(y)}\right] = \frac{1}{M}\mathop{\mathbb{E}}_{q(y)}\left[\left(\frac{p(y)}{q(y)}\right)^2 - 1\right] \\
&= \frac{1}{M}\mathop{\mathbb{E}}_{p(y)}\left[\frac{p(y)}{q(y)} - 1\right] = \frac{\mathbb{E}_{p(y)}\left[\exp\log\frac{p(y)}{q(y)}\right] - 1}{M} \\
&\geq \frac{\exp\mathbb{E}_{p(y)}\left[\log\frac{p(y)}{q(y)}\right] - 1}{M} = \frac{\exp(\mathrm{KL}(p(y) \,||\, q(y))) - 1}{M}
\end{aligned}$$

So, one can see that indeed, the NWJ bound can't give us high-confidence few-samples lower bound on any KL, not only the MI. This is because the second term would bias the bound by contributing large zero-mean noise. The only way to drive the magnitude of this noise down is to take more samples, and

as the analysis above shows, number of samples should be exponential in the KL (The statement could be made more precise by appealing to the Chebyshev's inequality).

The Donsker-Varadhan Estimator

Donsker and Varadhan have proposed essentially a tighter bound on the KL divergence, of the form

$$KL(p(x) \mid\mid q(x)) \geq \mathop{\mathbb{E}}_{p(x)} f(x) - \log \mathop{\mathbb{E}}_{q(x)} \exp(f(x)) + 1$$

With the same $f^*(x) = \ln \frac{p(x)}{q(x)}$ being an optimal critic. There are two key differences to the previous bound: the first is that it uses a logarithm in front of the balancing term, preventing it from contributing huge variance (but this variance still has to go somewhere, and we'll see where it goes), and the second (and the most important) is that this bound is no loger amendable to (unbiased) Monte Carlo estimation due to the logarithm outside of the expectation. In practice people just take an empirical average under the expectation thus obtaining a biased estimate (which in general is neither a lower nor an upper bound):

$$\frac{1}{N}\sum_{n=1}^{N} \ln \frac{p(x_n)}{q(x_n)} - \log \frac{1}{M}\sum_{m=1}^{M} \frac{p(y_m)}{q(y_m)}$$

It can be shown that the balancing term now has huge bias and is always negative. It's also easy to see that the biased converges to 0 as we take more samples M, so one might hope that with moderately many samples we'd have some tolerable bias. Well, this doesn't seem to be the case.

Take a closer look at the bias of the balancing term

$$\mathop{\mathbb{E}}_{q(y_{1:M})} \log \frac{1}{M}\sum_{m=1}^{M} \frac{p(y_m)}{q(y_m)}$$

It can be seen as an asymptotically unbiased estimate (a lower bound for all finite M) of the log-normalizing constant of $p(y)$ (which is 1 since it's already normalized) and is well-studied. In particular, Domke and Sheldon have shown (Theorem 3) that, essentially, the bias of the balancing term converges to 0 with the following rate:

$$O\left(M^{-1}\mathbb{V}_{q(y)}\left[\frac{p(y)}{q(y)}\right]\right)$$

Which 1) shows us where the variance has gone; 2) hints that in order to eliminate the bias we'd again need to take exponential number of samples.I don't know what happens to the actual variance of the

balancing term, but it can only make things worse.

The Contrastive-Predictive-Coding Bound

Let's leave the realm of lower bounds on KL now. Previously I have already presented the InfoNCE bound:

$$\mathrm{MI}[p(x,z)] \geq \mathop{\mathbb{E}}_{p(z_{0:K})} \mathop{\mathbb{E}}_{p(x|z_0)} \log \frac{p(x|z_0)}{\frac{1}{K+1}\sum_{k=0}^{K} p(x|z_k)}$$

Importantly, this bound does not have access to any marginals of the $p(x,z)$ joint. It's easy to show that this lower bound is upper bounded by $\log(K+1)$, which confirms the thesis:

$$\begin{aligned}\mathop{\mathbb{E}}_{p(z_{0:K})} \mathop{\mathbb{E}}_{p(x|z_0)} \log \frac{p(x|z_0)}{\frac{1}{K+1}\sum_{k=0}^{K} p(x|z_k)} &= \log(K+1) + \mathop{\mathbb{E}}_{p(z_{0:K})} \mathop{\mathbb{E}}_{p(x|z_0)} \log \frac{p(x|z_0)}{\sum_{k=0}^{K} p(x|z_k)} \\ &\leq \log(K+1)\end{aligned}$$

Which is due to the log's argument being between 0 and 1. So this $\log(K+1)$ upper bound on the lower bound means that if the true MI is much larger than this value, the bound will be very loose.

Given all these negative results, one might ask themselves if knowing the marginal $p(z)$ would do much better. Consider the "known prior" case:

$$\mathrm{MI}[p(x,z)] \geq \mathop{\mathbb{E}}_{p(x,z_0)} \mathop{\mathbb{E}}_{q_\phi(z_{1:K}|x)} \log \frac{\hat{\varrho}_\eta(x|z_0)}{\frac{1}{K+1}\sum_{k=0}^{K} \hat{\varrho}_\eta(x|z_k)\frac{p(z_k)}{q_\phi(z_k|x)}}$$

Then we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{p(x,z_0)\\q_\phi(z_{1:K}|x)}} \log \frac{\hat{\varrho}_\eta(x|z_0)}{\frac{1}{K+1}\sum_{k=0}^K \hat{\varrho}_\eta(x|z_k)\frac{p(z_k)}{q_\phi(z_k|x)}} \\
&\qquad = \log(K+1) + \mathop{\mathbb{E}}_{\substack{p(x,z_0)\\q_\phi(z_{1:K}|x)}} \left[\log \frac{\hat{\varrho}_\eta(x|z_0)\frac{p(z_0)}{q_\phi(z_0|x)}}{\sum_{k=0}^K \hat{\varrho}_\eta(x|z_k)\frac{p(z_k)}{q_\phi(z_k|x)}} - \log \frac{p(z_0)}{q_\phi(z_0|x)}\right] \\
&\qquad \le \log(K+1) + \mathop{\mathbb{E}}_{\substack{p(x,z_0)\\q_\phi(z_{1:K}|x)}} \log \frac{q_\phi(z_0|x)p(x|z_0)}{p(z_0)p(x|z_0)} \\
&\qquad = \log(K+1) + \mathop{\mathbb{E}}_{p(x,z_0)} \log \frac{q_\phi(z|x)p(x|z)}{p(z|x)p(x)} \\
&\qquad = \log(K+1) + \mathop{\mathbb{E}}_{p(x,z)} \log \frac{p(x|z)}{p(x)} - \mathop{\mathbb{E}}_{p(x,z)} \log \frac{p(z|x)}{q_\phi(z|x)} \\
&\qquad = \log(K+1) + \mathrm{MI}[p(x,z)] - \mathrm{KL}(p(x,z) \mid\mid q_\phi(z|x)p(x))
\end{aligned}
$$

Which shows that by choosing $q_\phi(z|x) = p(z)$ we essentially threw the baby out with the bathwater. Yes, K still needs to be exponential, but this time not in the original MI, but rather in $\mathrm{KL}(p(x,z) \mid\mid q_\phi(z|x)p(x))$, which can be made much smaller with a good choice of the variational distribution $q_\phi(z|x)$.

Also recall that we can reparametrize the bound in terms of $\hat{\rho}_\eta(x,z) = \hat{\varrho}_\eta(x|z)p(z)$

$$
\begin{aligned}
\mathrm{MI}[p(x,z)] &\ge \mathop{\mathbb{E}}_{p(x,z_0)} \mathop{\mathbb{E}}_{q_\phi(z_{1:K}|x)} \log \frac{\hat{\rho}_\eta(x,z_0)}{\frac{1}{K+1}\sum_{k=0}^K \frac{\hat{\rho}_\eta(x,z_k)}{q_\phi(z_k|x)}} - \mathop{\mathbb{E}}_{p(z)} \log p(z) \\
\mathrm{MI}[p(x,z)] &= \mathop{\mathbb{E}}_{p(x,z)} \log p(z|x) - \mathop{\mathbb{E}}_{p(z)} \log p(z)
\end{aligned}
$$

Hence

$$
\mathop{\mathbb{E}}_{p(x,z_0)} \log p(z_0|x) \ge \mathop{\mathbb{E}}_{p(x,z_0)} \mathop{\mathbb{E}}_{q_\phi(z_{1:K}|x)} \log \frac{\hat{\rho}_\eta(x,z_0)}{\frac{1}{K+1}\sum_{k=0}^K \frac{\hat{\rho}_\eta(x,z_k)}{q_\phi(z_k|x)}}
$$

Now we can choose the $p(x)$ marginal freely. Let $p(x) = \delta(x - \tilde{x})$. Then

$$
\mathop{\mathbb{E}}_{p(z_0|\tilde{x})} \log p(z_0|\tilde{x}) \ge \mathop{\mathbb{E}}_{p(z_0|\tilde{x})} \mathop{\mathbb{E}}_{q_\phi(z_{1:K}|\hat{x})} \log \frac{\hat{\rho}_\eta(\tilde{x},z_0)}{\frac{1}{K+1}\sum_{k=0}^K \frac{\hat{\rho}_\eta(\tilde{x},z_k)}{q_\phi(z_k|\tilde{x})}}
$$

So in a sense (and this is indeed how we derived the bound in the first place), this “known prior” lower

bound is based on a distribution-free *upper* bound on the entropy of $z|x$ and avoids lower bounding any entropies.

But why does it work in practice?

Despite the negative results above, there's a lot of empirical evidence of successful applications of all of the distribution-free bounds presented above. So what's going on? Quite possibly, this was the question folks from the Google Brain have asked themselves in their recent preprint On Mutual Information Maximization for Representation Learning. In this paper researchers investigated representations obtained by the Mutual Information maximization principle. For example, one finding is that tighter MI estimates surprisingly led to worse performance. Overall, the apparent conclusion of the paper is that MI estimation perspective does not seem to explain the observed behavior. Authors then suggest the metric learning perspective and reinterpret lower bounds on the MI as metric learning objectives.

Conclusion

All this evidence suggests that estimating the MI is an even harder problem than we used to think. In particular, blackbox MI estimation seems to be intractable in non-toy cases. Luckily, representation learning works nevertheless, probably due to a different phenomena.

However, for many problems it'd be really nice to have a way to quantify the dependence between x and z. A possible approach here is to consider different divergences between the joint $p(x, z)$ and the product of marginals $p(x)p(z)$. For example, one possible direction is to replace the KL divergence with some other f-divergence (see Lautum Information, for example), or, Wasserstein distance. And there's already some works in this direction: Wasserstein Dependency Measure for Representation Learning explores, unsurprisingly, the Wasserstein distance, or, Learning deep representations by mutual information estimation and maximization considers Jensen-Shannon divergence instead of the KL divergence. It'd curious to see some theorems / efficient bounds for these and other divergences.

Finally, one additional contribution of the Formal Limitations paper is the impossibility or good lower bounds on the KL divergence (supported by the reasoning above). This raises the question: given the whole family of f-divergences and their Fenchel conjugate-based blackbox lower bounds, do all of them exhibit such computationally unfavorable behavior? If no, which ones do?

Thanks to Ben Poole, Evgenii Egorov and Arseny Kuznetsov for valuable discussions.

1. Or, in my notation, $\mathrm{MI}[p(x)\delta(z-x)] = \mathbb{H}[p(x)]$.↩

2. By “good lower bound” I mean distribution-free high-confidence lower bound that uses some decent amount of samples, say, polynomial or even linear in the MI.↩