Forget-me-not! Contrastive Critics for Mitigating Posterior Collapse

Sachit Menon[1] **David Blei**[1] **Carl Vondrick**[1]

[1]Computer Science Dept., Columbia University, New York, New York, USA

Appendices

A MUTUAL INFORMATION

Recall that the mutual information across the variational joint is defined

$$\begin{aligned} I_q(x;z) &= D_{\mathrm{KL}}(q_\phi(x,z)||p_{\mathcal{D}}(x)q(z)) \\ &= \mathbb{E}_{q_\theta(x,z)}\left[\log \frac{q_\phi(x,z)}{p_{\mathcal{D}}(x)q_\phi(z)}\right] \end{aligned} \tag{1}$$

If the observations and the latents are independent, the mutual information is zero; in our case, we want to encourage the model to preserve their dependence, so we want it to be higher. Increasing $I_q(x;z)$ works against posterior collapse by preventing the posterior from always matching the prior (since this would not preserve any information). We will show that the approach provides a new way to increase both of these MI measures that could be combined with existing approaches.

B BINARY CLASSIFIER AND MI

We use a categorical likelihood for multi-way classification to implement the critic. Another option that provides some intuition would be a binary classifier, that simply takes a pair and decides if they correspond or not in isolation. (See Appendix C for the details of why the multisample case is preferred.) The connection between the binary classifier and MI follows directly from application of the density ratio trick [Sugiyama et al., 2012], which tells us a binary probabilistic classifier between two distributions estimates the density ratio between them. In our case, then, the optimal classifier would correspond to $\frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{z})p(\mathbf{x})}$. We highlight that in a different context, this same binary-classification density-ratio trick is what is used to power GANs: the discriminator estimates a density ratio between real and fake samples. In GANs, we do not want to be able to distinguish these distributions so we train the critic adversarially; here, we *want* the critic to succeed. GANs also provide us some basis that we do not need to train the critic to optimality at every step, which would be too expensive - joint training of the critic and the model can yield the desired results [Goodfellow et al., 2014]. See Related Work for more discussion of GAN-related techniques.

Thus, applying the density trick to our distributions at hand would provide us the integrand (of the MI expectation), and we could compute the expectation via Monte Carlo using all of the samples in the batch to get an estimate of the mutual information. The general technique of using a density ratio to estimate mutual information is introduced in Suzuki et al. [2008], elaborated on in Sugiyama et al. [2012] and draws its roots to 2-sample testing via classifiers; we encourage the interested reader to refer to these for the history of the method.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

C MULTISAMPLE DENSITY RATIO

One practical reason we would be interested in using the information from all the samples is that the 'one-sample' estimate of every density ratio term in the Monte Carlo expectation for MI described in B will have very high variance; using information from multiple samples for each term and getting a 'multi-sample' estimate would be more stable per Poole et al. [2019]. When we use the multiclass objective pushing down the objective pushes up the MI implicitly, see Appendix D. The tightness of this bound increases with the number of samples, so this is another reason we opt for the multisample approach.

D MUTUAL INFORMATION BOUND

This follows from analogy to CPC [Oord et al., 2019] (Appendix). This is an immediate application of the InfoNCE bound introduced there, which we follow here (along with [Grewal, 2019]); this is further elaborated on theoretically in Poole et al. [2019].

Consider a classifier that, for a latent sample z_i, tries to pick which observation x from a set $X = \{x_1, \ldots, x_i, \ldots, x_K\}$ it corresponds to. (We can phrase the problem as the reverse as well - it doesn't matter since the density ratio and mutual information are symmetric.) We'll also follow the notation for the model critic (Equation 5 left) for simplicity, but the inference critic (Equation 5 right) follows the same steps. (Note we also drop subscripts on densities for clarity.)

Consider Equation 6. We know [Sugiyama et al., 2012], reshown by [Oord et al., 2019], [Song and Ermon, 2020]) that the classifier will estimate the density ratio up to a constant. That is,

$$f(x, z) \propto \frac{p(x, z)}{p(x)p(z)} = \frac{p(x|z)}{p(x)} \tag{2}$$

(where the second equality is a simple application of Bayes' rule.)

We'll split the sum in the denominator of Equation 6 into 1) the term for the observation that corresponds to the latent at hand and 2) all the others. (In contrastive learning terminology, these are the positive and negatives respectively.)

$$\begin{aligned} \mathcal{L} &= \mathbb{E}\left[\log \frac{f\left(x^{+}, z^{+}\right)}{\sum_{x \in S} f\left(x, z^{+}\right)}\right] \\ &= \mathbb{E}\left[\log \frac{f\left(x^{+}, z^{+}\right)}{f\left(x^{+}, z^{+}\right) + \sum_{x_j \in X \setminus x_i} f\left(x_j, z^{+}\right)}\right] \end{aligned} \tag{3}$$

Since the classifier aims to estimate the density ratio (up to a constant), from Equation 2

$$\begin{aligned} &\approx \mathbb{E} \log \left[\frac{\frac{p\left(x^{+}|z^{+}\right)}{p(x^{+})} C}{\frac{p(x^{+}|z^{+})}{p(x^{+})} C + \sum_{x_j \in X \setminus x_i} \frac{p(x_j|z^{+})}{p(x_j)} C}\right] \\ &= \mathbb{E} \log \left[\frac{\frac{p\left(x^{+}|z^{+}\right)}{p(x^{+})}}{\frac{p(x^{+}|z^{+})}{p(x^{+})} + \sum_{x_j \in X \setminus x_i} \frac{p(x_j|z^{+})}{p(x_j)}}\right] \end{aligned} \tag{4}$$

We notice the 'positive' term appears in the numerator and denominator. Doing some algebraic manipulation,

$$\begin{aligned} &= \mathbb{E}\left(-\log\left[\frac{\frac{p\left(x^{+}|z^{+}\right)}{p(x^{+})} + \sum_{x_j \in X \setminus x_i} \frac{p\left(x_j|z^{+}\right)}{p(x_j)}}{\frac{p(x^{+}|z^{+})}{p(x^{+})}}\right]\right) \\ &= \mathbb{E}\left(-\log\left[1 + \frac{\sum_{x_j \in X \setminus x_i} \frac{p\left(x_j|z^{+}\right)}{p(x_j)}}{\frac{p(x^{+}|z^{+})}{p(x^{+})}}\right]\right) \\ &= \mathbb{E}\left(-\log\left[1 + \frac{p\left(x^{+}\right)}{p\left(x^{+} \mid z^{+}\right)} \sum_{x_j \in X \setminus x_i} \frac{p\left(x_j \mid z^{+}\right)}{p\left(x_j\right)}\right]\right) \end{aligned} \tag{5}$$

Now examine the sum over all terms but the 'positive' one. This can be considered a (scaled) expectation of the density ratio over the 'negative' terms - which should be 1, as for independent x, z the joint is the product of marginals. (Technically, since we are computing on samples, this is a Monte Carlo estimate of this expectation, but as noted by Oord et al. [2019] it is nearly exact even with relatively low K; Poole et al. [2019] shows a proof of the InfoNCE bound that does not use this approximation.)

$$\begin{aligned}
&\approx \mathbb{E}\left(-\log\left[1+\frac{p\left(x^{+}\right)}{p\left(x^{+} \mid z^{+}\right)}(K-1)\mathbb{E}_{x\sim X_{neg}}\left[\frac{p\left(x \mid z^{+}\right)}{p\left(x\right)}\right]\right]\right)\\
&=\mathbb{E}\left(-\log\left[1+\frac{p\left(x^{+}\right)}{p\left(x^{+} \mid z^{+}\right)}(K-1)\right]\right)\\
&=\mathbb{E}\log\left[\frac{1}{1+\frac{p(x^{+})}{p(x^{+}|z^{+})}(K-1)}\right]\\
&=\mathbb{E}\log\left[\frac{1}{K\frac{p(x^{+})}{p(x^{+}|z^{+})}-\frac{p(x^{+})}{p(x^{+}|z^{+})}+1}\right]\\
&=\mathbb{E}\log\left[\frac{1}{K\frac{p(x^{+})}{p(x^{+}|z^{+})}+\left(1-\frac{p(x^{+})}{p(x^{+}|z^{+})}\right)}\right]
\end{aligned} \tag{6}$$

$$\begin{aligned}
&\leq \mathbb{E}\log\left[\frac{1}{K\frac{p(x^{+})}{p(x^{+}|z^{+})}}\right]\\
&=\mathbb{E}\log\left[\frac{1}{K}\frac{p\left(x^{+} \mid z^{+}\right)}{p\left(x^{+}\right)}\right]\\
&=\mathbb{E}\left(\log\left[\frac{p\left(x^{+} \mid z^{+}\right)}{p\left(x^{+}\right)}\right]\right)-\log K\\
&=I(x;z)-\log K
\end{aligned} \tag{7}$$

where the last line is clearly less than $I(x;z)$. Thus

$$I(x;z)\geq\mathcal{L}+\log K \tag{8}$$

As the optimal loss (where the classifier is exactly a constant proportion of the density ratio) is bounded above by the MI, any suboptimal loss (which will be lower, since we are maximizing) will be bounded by the same. The key here is the $\log K$ term, which upper bounds the estimate of the MI (as $\mathcal{L}\leq 0$): intuitively, optimizing the loss pushes up the MI with a stick that is $\log K$ long. If the MI would otherwise fall to 0, the regularization aims to increase it by up to $\log K$ - past this, the bound is loose (our stick cannot reach), so it is advantageous to use higher K. (This is why the binary case as a lower bound is not practical.) Empirically, we increase $I_q(x;z)$ by almost exactly $\log K$ with an inference-side critic, showing this technique works as well as the theory might tell us it can (Table 1, Table 2).

E INFERENCE VS MODEL CRITICS

The model ('decoder'-side) critic corresponds to increasing $I_p(x;z)$ to prevent the likelihood from forgetting the latent, as previously discussed. For the inference ('encoder'-side) critic, the same analysis holds - instead of distinguishing the model joint from the product of the prior and the model approximate data distribution (marginal), it distinguishes the *variational* joint $q_\phi(x,z)$ from the product of the empirical data distribution $p_\mathcal{D}(x)$ and the aggregate posterior $q(z)$ (whose samples are obtained by ancestral sampling, analogously to the samples from the model approximate data distribution). Interestingly, adversarial variational Bayes [Mescheder et al., 2017] trains a similar critic adversarially, using this optimization to replace the ELBO. It also learns notably bad representations, so this is consistent, especially given that the ELBO terms they replace include the mutual information penalty (recall Equation 4), but this could be interesting to consider in more depth.

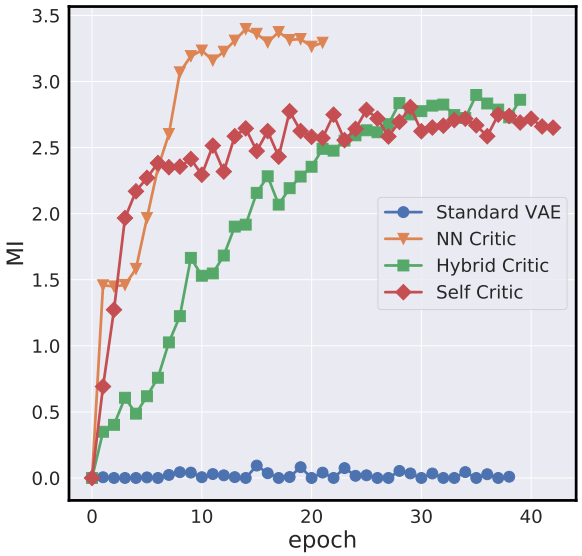

Figure 1: Comparison of mutual information across the variational family (I_q) for various critics vs baseline; the different endpoints are due to the termination condition for the experimental protocol depending on when a certain number of plateaus are reached.

One disadvantage of the model critic is that it requires sampling from the model, which can be expensive for strong model networks like autoregressive ones - which are where it would have the most effect. The inference critic does not have this restriction.

F EXPERIMENTAL PROTOCOL

Protocol reproduced from He et al. [2019].

Text experiments: LSTM parameters are initialized from $\mathcal{U}(-0.01, 0.01)$, with $\mathcal{U}(-0.1, 0.1)$ for embedding parameters. The final hidden representation produced by the inference network is used to predict the latent variable with a linear transformation. The SGD optimizer is used with an initial learning rate of 1.0, decayed by a factor of 2 upon a validation loss plateau for at least 2 epochs. Training ends once the learning rate has been thus decayed 5 times. No text preprocessing is performed. Dropout of 0.5 is used on the model network for the input embeddings and the pre-linear transformation output in vocabulary space.

Image experiments: train/val/test splits are used identically to He et al. [2019] and Kim et al. [2018]. The Adam optimizer is used with an initial learning rate of 0.001, decayed by a factor of 2 upon a validation loss plateau for at least 2 epochs. Training ends once the learning rate has been thus decayed 5 times. Images are dynamically binarized – that is, the input pixel values are considered parameters to Bernoulli random variables. Validation and test are performed with fixed binarization. The model network uses binary likelihood. The ResNet and PixelCNN are as described in He et al. [2019].

G MUTUAL INFORMATION COMPARISON – ALL TRAINING

H DISCUSSION OF VAE-MINE

Intuitively, our inference critics solve a classification task with a simple cross-entropy loss. This can be optimized with vanilla backprop. VAE-MINE adds a different term, based on MINE, to train an energy function that does not solve the same task; to optimize it, they resort to Taylor approximations and convex duality. This only implicitly results in contrasting the two distributions, while we directly train our inf. critic to do so. Yet, (per Poole et al. [2019] Sec 2.2,) their way of estimating MINE *is not even a correct bound on the MI*. Even if we ignore this, they lose the critical aspect of speed; for every size-n batch, their bound uses n^2 forward passes ([Poole et al., 2019] Sec 3) vs our $2n$ (vs base VAE's n). This scales poorly. (There is no code available for VAE-MINE for empirical comparison, but there is a decisive gap between their quadratic and our linear runtime.) Finally, ([Poole et al., 2019] App. A), the bound we use is lower variance than (the correct) MINE. Our method is theoretically appealing, correct, and fast.

REFERENCES

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. URL http://arxiv.org/abs/1406.2661. arXiv: 1406.2661.

Karan Grewal. Recent trends and mutual information-based objectives in unsupervised learning, 2019. URL http://karangrewal.ca/blog/mutual-information-objectives/.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging Inference Networks and Posterior Collapse in Variational Autoencoders. *arXiv:1901.05534 [cs, stat]*, January 2019. URL http://arxiv.org/abs/1901.05534. arXiv: 1901.05534.

Yoon Kim, Sam Wiseman, Andrew C. Miller, David Sontag, and Alexander M. Rush. Semi-Amortized Variational Autoencoders. *arXiv:1802.02550 [cs, stat]*, July 2018. URL http://arxiv.org/abs/1802.02550. arXiv: 1802.02550.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. January 2017. URL https://arxiv.org/abs/1701.04722v4.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv: 1807.03748.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On Variational Bounds of Mutual Information. *arXiv:1905.06922 [cs, stat]*, May 2019. URL http://arxiv.org/abs/1905.06922. arXiv: 1905.06922.

Jiaming Song and Stefano Ermon. Multi-label Contrastive Predictive Coding. In *NeurIPS*, July 2020. URL https://arxiv.org/abs/2007.09852v2.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, October 2012. ISSN 0020-3157, 1572-9052. doi: 10.1007/s10463-011-0343-8. URL http://link.springer.com/10.1007/s10463-011-0343-8.

Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating Mutual Information by Maximum Likelihood Density Ratio Estimation. In *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, pages 5–20. PMLR, September 2008. URL http://proceedings.mlr.press/v4/suzuki08a.html. ISSN: 1938-7228.