Mutual Information Constraints for Monte-Carlo Objectives to Prevent Posterior Collapse Especially in Language Modelling

Gábor Melis   MELISGL@GOOGLE.COM
DeepMind, London, UK; University College London, UK

András György   AGYORGY@GOOGLE.COM
DeepMind, London, UK

Phil Blunsom   PBLUNSOM@GOOGLE.COM
DeepMind, London, UK; Oxford University, UK

Editor: David Sontag

Abstract

Posterior collapse is a common failure mode of density models trained as variational autoencoders, wherein they model the data without relying on their latent variables, rendering these variables useless. We focus on two factors contributing to posterior collapse, that have been studied separately in the literature. First, the underspecification of the model, which in an extreme but common case allows posterior collapse to be the theoretical optimium. Second, the looseness of the variational lower bound and the related underestimation of the utility of the latents. We weave these two strands of research together, specifically the tighter bounds of multi-sample Monte-Carlo objectives and constraints on the mutual information between the observable and the latent variables. The main obstacle is that the usual method of estimating the mutual information as the average Kullback-Leibler divergence between the easily available variational posterior $q(z|x)$ and the prior does not work with Monte-Carlo objectives because their $q(z|x)$ is not a direct approximation to the model's true posterior $p(z|x)$. Hence, we construct estimators of the Kullback-Leibler divergence of the true posterior from the prior by recycling samples used in the objective, with which we train models of continuous and discrete latents at much improved rate-distortion and no posterior collapse. While alleviated, the tradeoff between modelling the data and using the latents still remains, and we urge for evaluating inference methods across a range of mutual information values.

Keywords: latent variables, variational inference, Monte-Carlo objective, mutual information

1. Introduction

The promise of latent variable models is in learning about the underlying generative process, discovering structure in the data, principled representation learning, improved generalization and controllable generation; all made possible by judicious choice of model structure, such as the prior, the likelihood, and any conditional independence assumptions. Variational autoencoders (VAEs, Kingma and Welling 2013; Rezende et al. 2014) provide a general framework for statistical inference in latent variable models of the form $p_\theta(x, z) = p_\theta(x|z)p(z)$, where x is the observable data, z is the vector of latent variables, and the objective is to learn the parameters θ so that the resulting marginal distribution $p_\theta(x)$ well approximates the empirical data distribution $p_D(x)$. The generality of VAEs comes at a price, as the variational posterior $q_\phi(z|x)$, used to approximate the true posterior $p_\theta(z|x)$,

usually underestimates the variance of the latter (Maddison et al. 2017),[1] which is often observed as the underuse of latent variables. In the extreme case, the underestimation leads to ignoring the latents entirely, which is known as posterior collapse (Zhao et al. 2019) and is the main focus of this work. The issue of posterior collapse is especially acute with auto-regressive decoders, which are capable of modelling the data without using the latents at all (Yang et al. 2017). Bowman et al. (2015) attributed this to a "difficult learning problem", and dozens of attempts to remedy it followed (Alemi et al. 2017; Dieng et al. 2017; van den Oord et al. 2017; Kim et al. 2018) to help VAEs fulfill their promise in representation learning.

This work aims to understand and remedy posterior collapse in VAEs with the long-term goal of facilitating research into latent variable models. While acknowledging that their ultimate evaluation is necessarily in terms of performance on down-stream tasks or as density models, we demonstrate that suboptimal inference can present a severe tradeoff between latent variable usage and data fit. This inefficiency of inference renders the posterior unfit for its purpose as a representation of the data. Therefore, instead of measuring the performance of the learned models on specific down-stream tasks, we evaluate this tradeoff in terms of their rate-distortion behaviour (Alemi et al. 2017), by measuring the rate as the mutual information between the observables x and the latents z, and distortion based on the negative log-likelihood assigned by the model to the data.

Several interacting factors (outlined in §2) play a role in posterior collapse, but the two most pertinent to this paper are the looseness of the lower bound and underspecification. *The looseness of the variational lower bound* (such as the ELBO objective) biases solutions found by VAEs away from the theoretical optimum. Hence, designing tighter lower bounds has been a mainstay of research on variational inference (Rezende and Mohamed 2015; Kingma et al. 2016; Tomczak and Welling 2018), and one approach is to take multiple samples from a variational posterior distribution $q_\phi(z|x)$ to form an approximation of the marginal likelihood $p_\theta(x)$.[2] In these so-called Monte-Carlo objectives (Mnih and Rezende 2016), such as IWAE (Burda et al. 2015), $q_\phi(z|x)$ does not represent the true posterior $p_\theta(z|x)$ explicitly, but it can be interpreted as a factor in a more elaborate, implicit approximate posterior (Cremer et al. 2017). In this context, it is thus more correct to refer to $q_\phi(z|x)$ as the proposal distribution.

The second factor we consider is *underspecification*. We use underspecification in the sense that models that are optimal in terms of marginal likelihood can differ greatly in their posteriors (Huszár 2017), thus the optimization aiming to fit the data by maximizing the likelihood is underspecified. This issue is most apparent in that, for VAEs with powerful function classes expressing the likelihoods, posterior collapse can be an optimal solution in terms of data fit because the usual evidence lower bound (ELBO) objective is neutral with respect to the mutual information between the latents and the data. However, as Huszár (2017) argues and as we show in §4, the marginal likelihood objective leaves the posterior underspecified, so underspecification is a shortcoming of the ELBO in only as much as it does not correct for it. Many proposed methods aim to address underspecification by constraining the mutual information between the observable and the latent variables (Higgins et al. 2017; Phuong et al. 2018; Zhao et al. 2019) relying on the availability of a good posterior approximation. As discussed next, with Monte-Carlo objectives we do not have access to a good

---

1. While we follow established terminology, strictly speaking q_ϕ does not "estimate" the variance of p_θ, and underestimation of the variance means that the distribution $q_\phi(z|x)$ has lower variance than that of $p_\theta(z|x)$.

2. Throughout we do not explicitly distinguish between densities and probability mass functions unless it is necessary; these are naturally dictated by the type (continuous/discrete) of the underlying variables.

posterior approximation, and this must be addressed if we hope constrain mutual information to reduce underspecification.

When the proposal $q_\phi(z|x)$ is close to $p_\theta(z|x)$, we can approximate the mutual information $I_p(X, Z) = \mathbb{E}_{p_\theta(x)} \text{KL}(p_\theta(z|x)\|p(z))$ (where KL denotes the Kullback-Leibler divergence) with $\mathbb{E}_{p_\theta(x)} \text{KL}(q_\phi(z|x)\|p(z))$, which features the proposal q_ϕ. Unfortunately, with Monte-Carlo objectives we cannot expect the proposal to approximate the posterior well (Mnih and Rezende 2016), and $\text{KL}(q_\phi(z|x)\|p(z))$ (called the *representational KL*) is a highly biased estimate of $\text{KL}(p_\theta(z|x)\|p(z))$ (the *true KL* from now on).

Our main contribution is a novel method to constrain the mutual information between the observable and the latent variables in the context of multi-sample Monte-Carlo objectives, bringing research on loose bounds and underspecification together. More specifically, we introduce an optimization objective which features two terms, one coming from the variational lower bound and another from the mutual information, where both terms are based on multiple samples taken from the proposal distribution q_ϕ. Compared to the single-sample case, we get the benefit of the tighter lower bounds Monte-Carlo objectives offer without having to give up control of the mutual information. At the same time, our multi-sample estimators for the mutual information are much more efficient than in the single-sample case and can better tolerate low-quality posterior approximations. Our mutual information term is computed from the recycled samples of the Monte-Carlo estimator of the marginal likelihood, hence the method has negligible computational overhead. Combined with best-of-breed gradient estimators, such as DReG (Tucker et al. 2018) and VIMCO (Mnih and Rezende 2016) for a multi-sample objective, we train models with continuous and discrete latents at much improved rate-distortion.

The rest of the paper is structured as follows.

- §2 provides an overview of the known causes of posterior collapse, which are all shortcomings of the inference method except for underspecification.
- In §4, we characterize underspecification as the lack of sufficient conditional independence assumptions, which may be partially offset by constraining mutual information.
- §5 proposes reusing samples from Monte-Carlo objectives to better estimate the mutual information, which is the main contribution of this paper.
- §6 and §7 show that the representational KL, which underlies many mutual information estimates, corresponds to the single-sample case of our estimators, and the single-sample objectives built on them are equivalent to the β-VAE objective of Higgins et al. (2017).
- §8 experimentally verifies the effectiveness of the proposed methods on synthetic and language modelling tasks, emphasizing evaluation in terms of the data fit vs latent usage tradeoff.

2. Variational Autoencoders and Posterior Collapse

This section introduces variational autoencoders and describes the known causes of posterior collapse. Contrary to what the name variational autoencoder may suggest, a VAE is not a model itself but an inference[3] mechanism for models of the form $p_\theta(x, z) = p_\theta(x|z)p(z)$ where $\{p_\theta(x|z)\}$ is a parametric family of conditional distributions (Kingma and Welling 2013). VAE training constructs an approximate maximum likelihood estimate of the model parameters θ with the aim of maximizing the probability over the empirical data distribution: $\arg\max_\theta \mathbb{E}_{x\sim p_D(x)}[\ln p_\theta(x)]$. Since $\ln p_\theta(x)$ has no analytic form in general, VAEs posit a variational family of distributions $\mathcal{Q} = \{q_\phi(z|x)\}$

---

3. We use *inference* in the statistical sense, commonly referred to as *training* or *learning* in the machine learning literature.

parameterized by ϕ to approximate the true posterior $p_\theta(z|x)$ and construct a lower bound on the marginal likelihood $p_\theta(x)$, also called the evidence:

$$\ln p_\theta(x) \geqslant \mathcal{L}_{\text{ELBO}}(x, \theta, \phi) = \ln p_\theta(x) - \text{KL}\left(q_\phi(z|x) \| p_\theta(z|x)\right) \tag{1}$$

$$= \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)}[\ln p_\theta(x|z)] - \text{KL}\left(q_\phi(z|x) \| p(z)\right). \tag{2}$$

As evident in the "posterior-contrastive" form of the ELBO (1), it is a lower bound on $\ln p_\theta(x)$ due to the non-negativity of the KL divergence (Kullback and Leibler 1951). In the "prior-contrastive" form (2), both the expectation and the KL divergence term can be estimated by taking a single sample from q, which forms the basis of its optimization. Alternatively, the KL may be computable analytically. Putting it all together, gradient-based optimization with VAEs is performed jointly over parameters θ of the model and ϕ of the approximate posterior as

$$\arg\max_{\theta, \phi} \mathop{\mathbb{E}}_{x \sim p_D(x)} \mathcal{L}_{\text{ELBO}}(x, \theta, \phi).$$

In practice, the expectation in (2) is approximated with a single sample from $q_\phi(z|x)$ and the expectation over $p_D(x)$ with a "minibatch", which goes by the name of doubly stochastic variational inference (Titsias and Lázaro-Gredilla 2014). For a broader context, we refer the reader to Zhang et al. (2018), who provide a comprehensive overview of developments in variational inference (Jordan et al. 1999). From this point on, wherever possible we drop the subscripts in p_θ and q_ϕ to declutter the notation.

The possibility of posterior collapse (also referred to as over-pruning, Yeung et al. 2017, or information preference, Zhao et al. 2019) is most evident in the prior-contrastive ELBO (2). If the likelihood $p(x|z)$ is able to model the distribution of x without the latents z, then the reconstruction term $\mathbb{E}_{q(z|x)} \ln p(x|z)$ is just $\ln p(x)$ independently of $q(z|x)$. Since q does not affect the reconstruction term, we can just set it to the prior $p(z)$, which is often in $\mathcal{Q}$, so that the KL penalty is zero. In this case, $q(z|x) = p(z|x)$ is also satisfied, but unfortunately the two posteriors have now "collapsed" to match the prior $p(z)$, which renders the latent variables useless.

The issues in what we observe as posterior collapse (or its milder form, the underuse of latents) span a number of causes. Figure 1 summarizes the main known contributors to posterior collapse and their interactions. At a glance, the immediate causes are high-variance gradient estimates, a loose lower bound, and underspecification.

- **Gradient noise**: Optimization can be adversely affected by high-variance gradient estimators (Roeder et al. 2017; Tucker et al. 2018) and minibatch noise in stochastic gradient descent (Titsias and Lázaro-Gredilla 2014). Unlike minibatch noise, sampling noise - the variance induced by the latents - can be eliminated by ignoring the latents. Coupled with SGD's preference for flat minima (Hochreiter and Schmidhuber 1995), this biases optimization towards posterior collapse.

- **Loose lower bound**: As its posterior-contrastive form (1) suggests, the ELBO is tight if $q(z|x)$ matches $p(z|x)$, and the worse the approximation, the looser the ELBO. However, perfect posterior approximation might be hard or impossible to achieve, depending on the following factors (Cremer et al. 2017).

  - The **approximation gap** is the distance of the true posterior from the variational family $\mathcal{Q}$ if $p(z|x) \notin \mathcal{Q}$ (Shu et al. 2018; Razavi et al. 2019).

  - The **amortization gap** is the gap caused by the encoder's inability to represent the optimal $q_{\phi^\star}(z|x)$ (which is $p_\theta(z|x)$) for all x with the same ϕ (Kim et al. 2018; Shu et al. 2018).
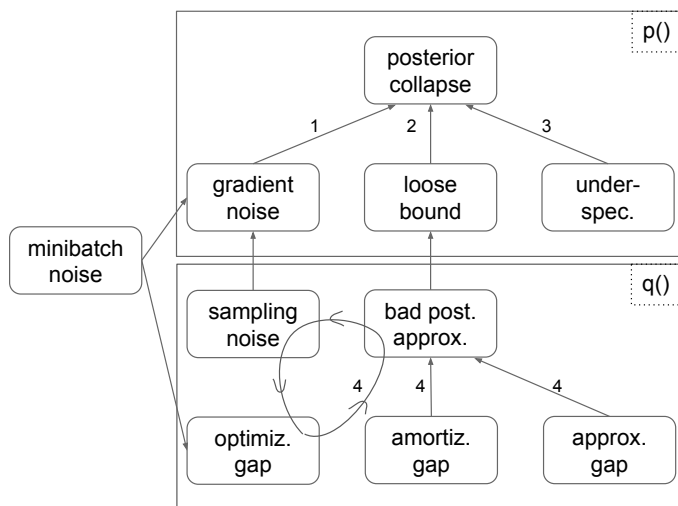
Figure 1: Causes of posterior collapse in VAEs. 1. High variance gradient estimates and SGD's preference for flat minima exerts pressure to reduce variance by ignoring the latents. 2. A loose lower bound underestimates the benefits of the latents. 3. Underspecification can allow posterior collapse to be the theoretical optimum. 4. Posterior collapse reduces or eliminates all of the gaps and the sampling noise.

- The **optimization gap** is caused the suboptimality of $q_\phi(z|x)$ found by the optimizer relative to $q_{\phi^\star}(z|x)$ (He et al. 2019).

Notably, there is a negative feedback loop: optimization difficulties cause bad posterior approximation, which increases the variance induced by sampling the latents, which makes optimization harder.

• **Underspecification**: As argued by Alemi et al. (2017), the ELBO is neutral with respect to the mutual information, and even perfect optimization can land anywhere on the rate-distortion curve (the expected likelihood of the data as a function of the mutual information of the data and the latents). More generally, the optimization task for generative models of the form $p(x, z) = p(x|z)p(z)$ is often underspecified (Huszár 2017), which we explore in §4.

Unfortunately, posterior collapse reduces or eliminates all of the gaps and also the sampling noise, which makes the bound tight. For VAEs to use their latent variables reliably and optimally, all of the above issues must be addressed. The direction we take in this paper delegates the problem of dealing with the looseness of the bound to a Monte-Carlo estimator of the marginal likelihood, such as IWAE, and that of reducing gradient noise to a corresponding gradient estimator, such as DReG or VIMCO. Importantly, in addition to providing tighter bounds, Monte-Carlo estimators employ multiple samples from $q(z|x)$, which benefits our efforts to tackle the issue of underspecification by designing better estimators of the mutual information.

3. Related Work

With several interacting issues to disentangle, the body of work on posterior collapse has grown large and disparate. Many works aim to make the prior or the posterior more flexible to tighten the

lower bound (Rezende and Mohamed 2015; Kingma et al. 2016; Tomczak and Welling 2018). At the same time, Shu et al. (2018) argue that while it is tempting to think that the variational familiy $\mathcal{Q}$ and $q_\phi(z|x)$ should be as large and as flexible possible, a more restrictive choice can prove useful in performing posterior regularization. Kim et al. (2018) propose reducing the amortization gap by only initializing the value of the latent variables according to the encoder, then optimizing their values separately for each example as in non-amortized variational inference (Jordan et al. 1999). He et al. (2019) focus on improving the approximation quality of $q_\phi(z|x)$ by taking several optimization steps on ϕ for each update of θ in p_θ. Yang et al. (2017) replace an auto-regressive decoder with dilated convolutions to restrict the available context and to thus enforce the use of the latents. Van den Oord et al. (2017) introduce categorically distributed latents with a fixed KL cost and replace the prior with the aggregate posterior, which eliminates any pressure to ignore the latents.

Controlling the KL term in (2) by various means also received lots of attention. Downweighting or annealing it during training is common (Bowman et al. 2015; Higgins et al. 2017; Sønderby et al. 2016; Maaløe et al. 2019; Huang et al. 2018). Kingma et al. (2016) "flat spot" the KL term so that small KL values are not penalized. Razavi et al. (2019) achieve a similar effect by choosing the variational family and the approximate posterior such that there is always an approximation gap.

Several works (Alemi et al. 2017; Zhao et al. 2019; McCarthy et al. 2019; Rezaabad and Vishwanath 2020; Serdega and Kim 2020) recognize the issue of underspecification and propose to constrain the mutual information. All work with single sample VAEs and the representational KL. In contrast, we propose combining multi-sample Monte-Carlo objectives with estimates based on the true KL. As we will see, this results in our mutual information estimates being less tied to the quality of $q(z|x)$, while containing the single-sample, representational KL scenario as a degenerate case.

4. Conditional Independence Assumptions and Posterior Collapse

In this section, we explore *why* our models are underspecified to better understand whether constraining the mutual information is a good solution. More specifically, we ask under what conditional independence assumptions (CIA[4]) can posterior collapse be an optimal solution for *any* model to abstract away from finite capacities and optimization difficulties. The next proposition shows that the answer is in fact trivial: the independence assumptions can be satisfied with a model where the latent and the observable variables are independent and the model matches the data distribution perfectly if and only if the independence assumptions are compatible with the marginal distribution of the data.

In particular, we consider Bayesian networks, given as $p(x,z) = p(z) \prod_{i=1}^n p(x_i|\mathrm{pa}(i), \mathbb{1}_{[1,k]}(i)z)$, where $\mathrm{pa}(i)$ denotes the set of parent nodes of x_i, i.e. those on which it directly depends, and the first k x_i directly depend also on z.

Proposition 1. *Let $\{x_i\}$ ($i \in [1,n]$) be a partitioning of x, and $\mathrm{pa}(i) \subseteq \{x_i\}$. For any prior $p(z)$, there exists a distribution[5] $p(x,z)$ satisfying, for all x and z,*

(i) $p(x,z) = p(x)p(z)$ (posterior collapse);
(ii) $p(x) = p_D(x)$ (perfectly modelling the data);
(iii) $p(x,z) = p(z) \prod_{i=1}^n p(x_i|\mathrm{pa}(i), \mathbb{1}_{[1,k]}(i)z)$ for some $k \leqslant n$ (CIA)
if and only if $p_D(x) = \prod_{i=1}^n p_D(x_i|\mathrm{pa}(i))$ (the data distribution is compatible with the CIA of (iii)).

---

4. Due to CIA being a collective noun in other contexts, we let its singular form stand also for its plural.
5. Here we consider the set of all distributions and do not restrict attention to the parametrized family $\{p_\theta\}$.

Proof If $p_D(x) = \prod_{i=1}^n p_D(x_i|\text{pa}(i))$, then it is easy to check that $p(x,z) = p_D(x)p(z)$ satisfies the conditions (i)–(iii) with $k = 0$. To prove the other direction, assume (i)–(iii) hold. Then (i) and (iii) imply that $p(x) = \prod_{i=1}^n p(x_i|\text{pa}(i), \mathbb{1}_{[1,k]}(i)z)$ for all z. Also, from (i) it follows that $(\text{pa}(i), x_i) \perp\!\!\!\perp z$, and hence for all z, $p(x_i|\text{pa}(i), z) = p(x_i, \text{pa}(i)|z)/p(\text{pa}(i)|z) = p(x_i|\text{pa}(i))$, giving $p(x) = \prod_{i=1}^n p(x_i|\text{pa}(i))$. Finally, from (ii) we have that the two joints, $p_D(x)$ and $p(x)$ are the same, therefore their marginals and conditionals are the same too, which implies $p_D(x) = \prod_{i=1}^n p_D(x_i|\text{pa}(i))$. $\blacksquare$

The proposition says that just by specifying CIA for otherwise dependent parts of the data, any model that suffers posterior collapse (in the sense that x and z are independent) will be suboptimal in terms of the model evidence $p(x)$. This gives a degree of assurance that given such a structure, a well-optimized model with high enough capacity will not suffer posterior collapse. Latent variable image models, that model pixels or patches conditionally independently given the latents fall into this category, which explains while posterior collapse is not so prevalent in that case.

Conversely, in theory and in the absence of CIA, there is a trivial latent variable model $p(x,z) = p_D(x)p(z)$ which is optimal in terms of the marginal likelihood $p(x)$ but does not use the latents. The prototypical example for this case is auto-regressive likelihoods, such as RNN language models.

To summarize, *CIA must be made* to guarantee that latents are used given powerful enough models and inference methods. On the other hand, lacking the necessary CIA, we can still bias solutions by changing the objective. One such change to compensate for the lack of model structure is adding a constraint on mutual information.

5. Mutual Information Augmented Objectives

Mutual information is often used as a measure of latent variable usage in trained models or as part of the training objective to control latent usage and reduce underspecification. First, as a measure of latent usage, it is a diagnostic of the inference method. In this role, it is but a proxy for the generalization ability of the model or for the performance on down-stream tasks. Second, as a constraint during training, it can be seen as compensating for the lack of structure in the model. However, its role is not essential in either of these: evaluating representations without the down-stream tasks is fraught with peril, and equipping the model with structure sounds a rather more appealing direction to pursue. Still, finding a good model structure is easier said than done, and in practice mutual information is useful both as a diagnostic for inference and as a tool for model specification. Thus, we augment the marginal likelihood objective with a mutual information term and maximize

$$\mathop{\mathbb{E}}_{p_D(x)} \ln p(x) + \lambda I_p(X, Z), \tag{3}$$

where $p_D(x)$ is the data distribution and $\lambda \in \mathbb{R}, \lambda \geqslant 0$. Throughout, we assume that $I_p(X, Z)$ is bounded, which is satisfied by any model for which the optimization of $p(x)$ is well-posed and hence its $p(x|z)$ is bounded for all z. Also note that for any given $I_p(X, Z)$, the model achieves its global maximum when $p(x) = p_D(x)$, and we can reasonably expect models that suffer from posterior collapse to effectively balance data fit and mutual information.

Motivated by the identity $I_p(X, Z) = \mathbb{E}_{p(x)} \text{KL}(p(z|x)\|p(z))$, the mutual information term can be estimated by the average KL divergence. While this true KL is hard to compute in general due to

the intractable posterior, the availability the variational posterior offers the compelling alternative of estimating $I_p(X, Z)$ by

$$I_{p,q}(X, Z) := \mathop{\mathbb{E}}_{p_D(x)} \text{KL}\big(q(z|x)\|p(z)\big).$$

If we plan to use the representations obtained from the variational posterior $q(z|x)$ on some task, then $I_{p,q}$ is a natural quantity to track (Zhao et al. 2019; Rezaabad and Vishwanath 2020). However, in this work, our primary concern lies not with artifacts of variational inference but with the model $p(x, z)$ and its ability to capture information in the latents. Moreover, employing $I_{p,q}$ as a proxy objective is problematic because it overestimates I_p as q tends to underestimate the variance of the true posterior.[6] Even worse, with $I_{p,q}$ the quality of $q(z|x)$ would influence our conclusions about latent variable usage. Monte-Carlo objectives, whose $q(z|x)$ in itself is no longer a direct approximation to $p(z|x)$ (Mnih and Rezende 2016), exacerbate the problem with latent usage estimation. Experimentally, we found that, for models trained with Monte-Carlo objectives, $I_{p,q}$ can wildly under- or overestimate I_p. Thus, to form a better estimate of I_p, we replace $q(z|x)$ with $p(z|x)$ in $I_{p,q}$:

$$I_{p_D}^p(X, Z) := \mathop{\mathbb{E}}_{p_D(x)} \text{KL}\big(p(z|x)\|p(z)\big). \tag{4}$$

This "cross" mutual information $I_{p_D}^p$ is the average true KL over the data distribution p_D. In the typical doubly stochastic optimization setting (Titsias and Lázaro-Gredilla 2014), averaging over $p_D(x)$ instead of $p(x)$ allows us to estimate the mutual information based on the current minibatch without sampling from the model. The price for efficiency is a possible generalization problem: the average KL may be different over $p_D(x)$ and $p(x)$. Generalization may eventually become a pressing issue, but as our experiments will demonstrate (§8), we first have to deal with underfitting. In addition to being expedient, as we will show in §6 and §7, this choice makes the single-sample case of our estimators correspond to the representational KL and the β-VAE (Higgins et al. 2017).

Definition 2 (Mutual information augmented objective). *The mutual information augmented objective, which is a combination of the usual marginal likelihood objective and $I_{p_D}^p$, is defined as*

$$\mathcal{O}(\lambda) = \mathop{\mathbb{E}}_{p_D(x)} \ln p(x) + \lambda I_{p_D}^p(X, Z). \tag{5}$$

Furthermore, the pointwise version of the objective is defined as

$$\mathcal{O}(\lambda, x) = \ln p(x) + \lambda \text{KL}\big(p(z|x)\|p(z)\big), \tag{6}$$

which in turn satisfies $\mathcal{O}(\lambda) = \mathbb{E}_{p_D(x)} \mathcal{O}(\lambda, x)$.

In the following, we propose estimators of $\mathcal{O}(\lambda, x)$ and the true KL within it to estimate $\mathcal{O}(\lambda)$ and the mutual information in a manner suitable for the doubly stochastic optimization setting.

---

6. For VAEs, this follows from the properties of the KL divergence, while for Monte-Carlo objectives in general, it follows from how the looseness of the lower bound relates to the variance of the estimator (Maddison et al. 2017).

5.1 The KL Objective

To find the maximum of $\mathcal{O}(\lambda)$ in θ, both of its terms must be estimated well. We delegate the task of estimating the marginal log-likelihood $\ln p(x)$ to a "base" Monte-Carlo estimator of the form $\hat{S}^K(x, z_{1:K}) = \ln\left(\frac{1}{K}\sum_{i=1}^{K} f(x, z_i)\right)$, where $z_{1:K} = (z_1, \ldots, z_K)$ are independent samples from the proposal distribution $q(z|x)$, and f is some function of the observable and latent variables (Mnih and Rezende 2016). Ideally, $\hat{S}^K$ is chosen to have low bias and low variance, allowing optimization to strike a better balance with mutual information. For our first contribution, the KL objective, we rewrite the true KL in a form more amenable to importance sampling:

$$
\begin{aligned}
\mathrm{KL}\left(p(z|x)\|p(z)\right) &= \underset{p(z|x)}{\mathbb{E}} \ln\frac{p(z|x)}{p(z)} = \underset{p(z|x)}{\mathbb{E}}\left[\ln p(x|z)\right] - \ln p(x) \\
&= \underset{q(z|x)}{\mathbb{E}}\left[\frac{p(z|x)}{q(z|x)}\ln p(x|z)\right] - \ln p(x) \\
&= \frac{1}{p(x)}\underset{q(z|x)}{\mathbb{E}}\left[\frac{p(x,z)}{q(z|x)}\ln p(x|z)\right] - \ln p(x).
\end{aligned}
$$

Plugging this into the definition of $\mathcal{O}(\lambda, x)$ in (6) and grouping the $\ln p(x)$ terms, which can be estimated with the base Monte-Carlo estimator $\hat{S}^K$, we get

$$
\mathcal{O}(\lambda, x) = (1-\lambda)\ln p(x) + \lambda\frac{1}{p(x)}\underset{q(z|x)}{\mathbb{E}}\frac{p(x,z)}{q(z|x)}\ln p(x|z).
$$

This leaves the task of estimating the second term

$$
\frac{1}{p(x)}\underset{q(z|x)}{\mathbb{E}}\frac{p(x,z)}{q(z|x)}\ln p(x|z), \tag{7}
$$

in which $p(x)$ can be estimated with a simple K-sample importance sampling estimator

$$
\hat{p}^K(x, z_{1:K}) := \frac{1}{K}\sum_{i=1}^{K}\frac{p(x, z_i)}{q(z_i|x)}. \tag{8}
$$

However, combining $\hat{p}^K$ with an importance sampling estimate of the expectation in (7) using different samples begot very high variance in preliminary experiments. Instead, we approximate (7) with the self-normalized importance sampling estimator

$$
\hat{U}^K(x, z_{1:K}) := \hat{p}^K(x, z_{1:K})^{-1}\frac{1}{K}\sum_{i=1}^{K}\frac{p(x, z_i)}{q(z_i|x)}\ln p(x|z_i), \tag{9}
$$

which uses the same samples for estimating $1/p(x)$ and the expectation. This leads to our first estimator for $\mathcal{O}(\lambda, x)$.

Definition 3 (KL objective). *Let $\hat{S}^K$ be any K-sample Monte-Carlo estimator of $\ln p(x)$. Then the augmented objective $\mathcal{O}(\lambda, x)$ can be estimated by the KL objective*

$$
\mathcal{O}_{\mathrm{KL}}(\hat{S}, K, \lambda, x) = \underset{z_{1:K}\sim q(z|x)}{\mathbb{E}}\hat{O}_{\mathrm{KL}}(\hat{S}, K, \lambda, x, z_{1:K}), \tag{10}
$$

where

$$
\hat{O}_{\mathrm{KL}}(\hat{S}, K, \lambda, x, z_{1:K}) = (1-\lambda)\hat{S}^K(x, z_{1:K}) + \lambda\hat{U}^K(x, z_{1:K}). \tag{11}
$$

Note that $\hat{\mathcal{O}}_{\mathrm{KL}}(\hat{S}^K, K, \lambda, x, z_{1:K})$ uses the same samples $z_{1:K} \sim q(z|x)$ to estimate both terms of the augmented objective (6). Grouping the terms differently, we can separate out the estimate of the KL:

$$\hat{\mathcal{O}}_{\mathrm{KL}}(\hat{S}, K, \lambda, x, z_{1:K}) = \underbrace{\hat{S}^K(x, z_{1:K})}_{\approx \ln p(x)} + \lambda \underbrace{\left(\hat{U}^K(x, z_{1:K}) - \hat{S}^K(x, z_{1:K})\right)}_{\approx \mathrm{KL}(p(z|x)\|p(z))}. \qquad (12)$$

We assume throughout that the estimators $\hat{S}^K(x, z_{1:K})$ and $\hat{U}^K(x, z_{1:K})$ have finite variance. This implies the following properties of the estimators $\hat{\mathcal{O}}_{\mathrm{KL}}(\hat{S}^K, K, \lambda, x)$ of $\mathcal{O}(\lambda, x)$ and $\hat{U}^K(x, z_{1:K}) - \hat{S}^K(x, z_{1:K})$ of $\mathrm{KL}(p(z|x)\|p(z))$:

Proposition 4 (properties of the KL objective)**.**
 (i) *If $\hat{S}^K$ converges in probability or almost surely to $\ln p(x)$ as $K \to \infty$, then so do $\hat{\mathcal{O}}_{\mathrm{KL}}$ and $\hat{U}^K - \hat{S}^K$ to $\mathcal{O}(\lambda, x)$ and $\mathrm{KL}(p(z|x)\|p(z))$, respectively.*
 (ii) *If $\mathbb{E}_{z_{1:K} \sim q(z|x)} \hat{S}^K(x, z_{1:K}) \leqslant \ln p(x)$, then $\mathbb{E}_{z_{1:K} \sim q(z|x)}[\hat{U}^K(x, z_{1:K}) - \hat{S}^K(x, z_{1:K})] \geqslant \mathrm{KL}(p(z|x)\|p(z))$, that is, the estimator is biased upward.*
 (iii) *The bias of the self-normalized importance sampling estimator $\hat{U}^K$ is bounded if $p(x, z)$ is bounded as shown in Proposition 7 of Metelli et al. (2020).*
 (iv) *The variance of $\hat{U}^K$ decays with K, but unlike in non-normalized importance sampling, for any given K, there is no proposal distribution with which the variance is zero unless $\hat{U}^K$ is constant for all $z_{1:K}$ with probability 1.*

Proof These follow from the properties of self-normalized importance sampling (Owen 2013) except where noted. ∎

Note that although the objective is biased upwards, its bias is decreased with more samples and is bounded if $p(x, z)$ is bounded. It can be assumed that $p(x, z)$ is bounded, else the optimization of $\ln p(x)$ is ill-posed even without the mutual information term. Consequently we may be able to rely on λ to counteract the bias thus bounded. Importantly, the computation of $\hat{U}^K$ imposes minimal overhead as it needs to evaluate only $p(x|z_i)$, $p(z_i)$ and $q(z_i|x)$: the same quantities and same z_i as needed for computing $\hat{S}^K$. Referring back to (3), we argue that this KL objective allows for effective interpolation between fitting the data and capturing information in the latent variables.

5.2 The Rényi Objective

In this section, we introduce a second estimator of the augmented objective $\mathcal{O}(\lambda, x)$ (6), based on the Rényi divergence, to address a potential issue with the KL objective's estimate (11). This issue lies in the fact that the KL objective linearly combines two estimators ($\hat{S}^K$ and $\hat{U}^K$) of different quantities. How their biases and variances relate deserves some consideration. As we have seen, with $\hat{S}^K$ that underestimate $\ln p(x)$ (e.g. IWAE), $\hat{U}^K - \hat{S}^K$ overestimates the true KL. Luckily, both biases can be reduced with more samples.

However, taking more samples may not help if the two estimators have very different variances, in which case optimizing the objective may be difficult. This issue could be addressed by designing a $\hat{U}^K$ for every $\hat{S}^K$, but this would limit the applicability of our method in practice. Instead, we apply $\hat{S}$ not only to estimate $\ln p(x)$ but *a second time* too to estimate the Rényi divergence, itself a biased estimate of the true KL. As we will see later, this works surprisingly well in practice despite the presence of the bias.

The Rényi divergence between two distributions $f(x)$ and $g(x)$ is defined as $D_\alpha(f\|g) = \frac{1}{\alpha-1}\ln\mathbb{E}_{g(x)} f(x)^\alpha g(x)^{-\alpha}$, where α is a positive real number. For $\alpha < 1$, $D_\alpha(f\|g) \leqslant \mathrm{KL}(f\|g)$, while for $\alpha > 1$, $D_\alpha(f\|g) \geqslant \mathrm{KL}(f\|g)$. Since $\lim_{\alpha\to 1} D_\alpha(f\|g) = \mathrm{KL}(f\|g)$, $D_\alpha(f\|g)$ can approximate $\mathrm{KL}(f\|g)$ arbitrarily closely when α is sufficiently close to 1. This latter property motivates the use of $D_\alpha(p(z|x)\|p(z))$ as an approximation to the true KL. To construct an estimator, we first rewrite the Rényi divergence as:

$$
\begin{aligned}
(\alpha - 1)D_\alpha\big(p(z|x)\|p(z)\big) &= \ln \mathbb{E}_{p(z)} \frac{p(z|x)^\alpha}{p(z)^\alpha} \\
&= \ln \mathbb{E}_{p(z)} \left[\frac{p(z|x)^\alpha p(x)^\alpha}{p(z)^\alpha}\right] - \alpha \ln p(x) \\
&= \ln \mathbb{E}_{p(z)} \big[p(x|z)^\alpha\big] - \alpha \ln p(x) \\
&= \ln p^\alpha(x) - \alpha \ln p(x),
\end{aligned}
\tag{13}
$$

where $p^\alpha(x) := \mathbb{E}_{p(z)} p(x|z)^\alpha$. We note in passing that $p^\alpha(x)$ has an intuitive interpretation, particularly when $\alpha \in \mathbb{N}$. Consider the task of modelling the distribution of discrete data over some discrete set $\mathcal{X}$ duplicated α times, that is $p_D^\alpha(x, \dots, x) := p_D(x)$. That is, p_D^α is a probability distribution over $\mathcal{X}^\alpha$, whereas p_D is over $\mathcal{X}$, and $p_D^\alpha(x_1, \dots, x_\alpha) = 0$ unless all x_i are identical. On this "new" task, $\alpha \ln p(x) = \ln p(x)^\alpha$ acts as the uninformed baseline, in which a separate set of latents is used for each branch $p(x_i|z)$, thus the cost of information in the latents must be paid α times. This means that, based on its alternative form $\ln p^\alpha(x) - \alpha \ln p(x)$ in (13), we can interpret the Rényi divergence as a measure of how much better the α-duplicated model $p^\alpha(x)$ does at modelling the duplicated data compared to the worst-case solution $p(x)^\alpha$, which does not use the latents to amortize the cost of encoding the data multiple times.

We now derive a biased approximation to $\mathcal{O}(\lambda, x)$:

$$
\begin{aligned}
\mathcal{O}(\lambda, x) &= \ln p(x) + \lambda \, \mathrm{KL}\big(p(z|x)\|p(z)\big) \\
&\approx \ln p(x) + \lambda D_\alpha\big(p(z|x)\|p(z)\big) \\
&= \ln p(x) + \frac{\lambda}{\alpha - 1}\big(\ln p^\alpha(x) - \alpha \ln p(x)\big) \\
&= \frac{\lambda}{\alpha - 1}\ln p^\alpha(x) - \left(\frac{\lambda\alpha}{\alpha - 1} - 1\right)\ln p(x).
\end{aligned}
$$

Definition 5 (Rényi objective). *Let $\lambda, \alpha > 0$, and let $\hat{S}^K(x, z_{1:K})$ and $\hat{S}_\alpha^K(x, z_{1:K})$ be K-sample Monte-Carlo estimators for $\ln p(x)$ and $\ln p^\alpha(x)$, respectively. Then the augmented objective $\mathcal{O}(\lambda, x)$ can be estimated by the Rényi objective*

$$
\mathcal{O}_R(\hat{S}, K, \lambda, \alpha, x) = \mathbb{E}_{z_{1:K}\sim q(z|x)} \hat{\mathcal{O}}_R(\hat{S}, K, \lambda, \alpha, x, z_{1:K}),
\tag{14}
$$

where

$$
\hat{\mathcal{O}}_R(\hat{S}, K, \lambda, \alpha, x, z_{1:K}) = \frac{\lambda}{\alpha - 1}\hat{S}_\alpha^K(x, z_{1:K}) - \left(\frac{\lambda\alpha}{\alpha - 1} - 1\right)\hat{S}^K(x, z_{1:K}).
\tag{15}
$$

Separating out the estimate of the Rényi divergence yields the alternative form

$$\hat{\mathcal{O}}_R(\hat{S}, K, \lambda, \alpha, x, z_{1:K}) = \underbrace{\hat{S}^K(x, z_{1:K})}_{\approx \ln p(x)} + \lambda \underbrace{\left(\frac{1}{\alpha - 1} \Big(\hat{S}^K_\alpha(x, z_{1:K}) - \alpha \hat{S}^K(x, z_{1:K}) \Big) \right)}_{\approx D_\alpha(p(z|x) \| p(z))}. \quad (16)$$

Our goal was to address the mismatched biases and variances of the KL objective's $\hat{S}^K$ and $\hat{U}^K$. Having eliminated $\hat{U}^K$, we are left with only $\hat{S}^K$ and $\hat{S}^K_\alpha$, estimating two closely related quantities, $\ln p(x)$ and $\ln p^\alpha(x)$. Note that $\hat{S}^K_\alpha$ can be obtained with a slight modification of $\hat{S}^K$, as explained in the next section. In §8, we validate experimentally that the benefits this scheme affords outweigh the obvious drawback of additional bias in the Rényi objective.

5.2.1 ESTIMATING $\mathbf{p}^\alpha(\mathbf{x})$ WITH IWAE

Since $\ln p(x) = \ln \big(\mathbb{E}_{p(z)} p(x|z) \big)$ and $\ln p^\alpha(x) = \ln \big(\mathbb{E}_{p(z)} p(x|z)^\alpha \big)$, if $\hat{S}^K(x, z_{1:K})$ is computed explicitly in terms of $p(x|z)$, then $\hat{S}^K_\alpha(x, z_{1:K})$ can be derived simply by replacing $p(x|z)$ with $p(x|z)^\alpha$ in $\hat{S}^K(x, z_{1:K})$. All base estimators considered in this paper have this property and we elucidate the derivation through the example of the IWAE.

The ELBO of variational autoencoders can be rather loose, and the variance of the resulting approximate posterior is usually smaller than that of the true posterior. To tackle these issues, Burda et al. (2015) proposed the importance weighted autoencoder (IWAE). Whereas the ELBO is single-sample, the IWAE bound is based on $K \in \mathbb{N}$ samples:

$$\ln p(x) = \ln \mathop{\mathbb{E}}_{p(z)} p(x|z) = \ln \mathop{\mathbb{E}}_{\substack{z_1, \dots, z_K \\ \sim q(z|x)}} \frac{1}{K} \sum_{i=1}^{K} \frac{p(x|z_i)p(z_i)}{q(z_i|x)}$$

$$\geqslant \mathop{\mathbb{E}}_{\substack{z_1, \dots, z_K \\ \sim q(z|x)}} \ln \frac{1}{K} \sum_{i=1}^{K} \frac{p(x|z_i)p(z_i)}{q(z_i|x)}.$$

In importance sampling terms, $p(x|z)$ is the integrand, the function whose expectation we want to compute with respect to $p(z)$. While $p(x|z)$ here is a probability mass function, importance sampling does not require this. In fact, we can estimate the expectation of $p(x|z)^\alpha$ analogously:

$$\ln p^\alpha(x) \geqslant \mathop{\mathbb{E}}_{\substack{z_1, \dots, z_K \\ \sim q(z|x)}} \ln \frac{1}{K} \sum_{i=1}^{K} \frac{p(x|z_i)^\alpha p(z_i)}{q(z_i|x)}. \quad (17)$$

To estimate the expectation in the above lower bound, we can recombine quantities already computed for the base estimator with a negligible computational overhead.

Note that there is an optimal proposal distribution $q_{\text{opt}}(z|x) = p(x|z)^\alpha p(z)/p^\alpha(x)$ that leads to exactly computing $\ln p^\alpha(x)$ (i.e. estimating it with zero bias and variance). On the other hand, the other term in the Rényi objective is best estimated using, in general, a different proposal distribution. One solution would be to apply separate proposal distributions, another is to choose λ and α such that in $\ln p(x) + \lambda D_\alpha$ the $\ln p(x)$ term is cancelled out, which we explore briefly in §5.3 below.

5.2.2 Other Estimators

In the interest of space, we omit re-derivations of further estimators of $\ln p^\alpha(x)$ and their gradient estimators. As in the IWAE case, all we need to show is that the estimators do not depend on properties of $p(x|z)$ that $p^\alpha(x|z)$ doesn't have. This is true for the estimators used in our experiments: REINFORCE (Williams 1987; Mnih and Gregor 2014), VIMCO (Mnih and Rezende 2016), STL (Roeder et al. 2017) and DReG (Tucker et al. 2018).

5.3 The Power Objective

Notice that, in the Rényi estimator (15), if λ is set to $(\alpha - 1)/\alpha$, then the coefficient of the $\ln p(x)$ term becomes zero, yielding

$$\hat{\mathcal{O}}_P(\hat{S}, K, \alpha, x) = \alpha^{-1} \hat{S}_\alpha^K(x, z_{1:K}), \tag{18}$$

which we call the *power objective*. Note that for optimization the constant α^{-1} can be dropped from the objective and that if $\alpha = 1$, the power objective is equal to the log-likelihood $\ln p(x)$. Next we show that if $\alpha > 1$, maximizing $p^\alpha(x)$ optimizes a lower bound on $p(x)$, and this lower bound is tight when the latents fully determine the observables.

Proposition 6. *Assume that X is concentrated on the countable set $\mathcal{X}$, $\alpha > 1$, and let $x \in \mathcal{X}$ be arbitrary. Then $p^\alpha(x) \leqslant p(x)$ with equality for all $x \in \mathcal{X}$ if and only if $H_p(X|Z) = 0$. Furthermore, $(p(x))^\alpha \leqslant p^\alpha(x)$ and $\ln p^\alpha(x) - \alpha \ln p(x) = (\alpha - 1)D_\alpha(p(z|x)\|p(z))$.*

Proof Since X is countable, $p(x|z)$ is discrete, hence $p(x|z) \leqslant 1$ for all x and z. Therefore, since $\alpha > 1$, $p(x|z)^\alpha \leqslant p(x|z)$ with equality if and only if $p(x|z) \in \{0, 1\}$. Thus, $p^\alpha(x) = \mathbb{E}_{p(z)} p(x|z)^\alpha \leqslant \mathbb{E}_{p(z)} p(x|z) = p(x)$, with equality if and only if $p(x|z) \in \{0, 1\}$ for all x, for almost all z. The latter holds if and only if X is a deterministic function of Z with probability 1, which is equivalent to $H_p(X|Z) = 0$. The second statement of is a direct consequence of (13) and $\alpha - 1 > 0$. ∎

So the power objective with $\alpha > 1$ is an upper bound on the KL (and I_{PD}^p when averaged over x), but this upper bound nevertheless becomes tight when the latents determine the observable variables (i.e. $H_p(X|Z) = 0$).

In summary, the power objective has three important differences from the Rényi objective, of which it is a special case. First, with $\alpha > 1$, it is guaranteed to be a lower bound on $\ln p(x)$. Second, since there is only a single quantity, $\ln p^\alpha(x)$, being estimated, the question of using separate $q(z|x)$ proposal distributions to estimate the different terms does not arise. Third, λ and α are tied, so it may be harder to find a good balance between ease of optimization and low bias.

6. Connection to the Representational KL

Proposition 7 (Single-sample KL estimate). *In the single-sample case with $\hat{S}^1(x, z_1) = \ln \frac{p(x,z_1)}{q(z_1|x)}$, the KL objective's estimate of the true KL in (12) is the representational KL in expectation:*

$$\mathbb{E}_{z_1 \sim q(z|x)} \left[\hat{U}^1(x, z_1) - \hat{S}^1(x, z_1) \right] = \text{KL}\left(q(z|x) \| p(z) \right).$$

Proof With $\hat{p}^1(x, z_1) = p(x, z_1)/q(z_1|x)$ from (8), we have that

$$\hat{U}^1(x, z_1) - \hat{S}^1(x, z_1) = \hat{p}^1(x, z_1)^{-1} \frac{p(x, z_1)}{q(z_1|x)} \ln p(x|z_1) - \ln \frac{p(x, z_1)}{q(z_1|x)}$$

$$= \ln p(x|z_1) - \ln \frac{p(x, z_1)}{q(z_1|x)} = \ln \frac{q(z_1|x)}{p(z_1)}$$

∎

So not only are the expectations the same, but our single-sample estimate is the same as the trivial single-sample estimate of the representational KL. We now prove a similar result for the Rényi objective.

Proposition 8 (Single-sample Rényi estimate). *In the single-sample case with $\hat{S}^1(x, z_1) = \ln \frac{p(x,z_1)}{q(z_1|x)}$, the Rényi objective's estimate of $D_\alpha(p(z|x)\|p(z))$ in (16) is the representational KL in expectation:*

$$\mathbb{E}_{z_1 \sim q(z|x)} \left[\frac{1}{\alpha - 1} \Big(\hat{S}_\alpha^1(x, z_1) - \alpha \hat{S}^1(x, z_1) \Big) \right] = \text{KL} \big(q(z|x)\|p(z) \big).$$

Proof

$$\frac{1}{\alpha - 1} \Big(\hat{S}_\alpha^1(x, z_1) - \alpha \hat{S}^1(x, z_1) \Big) = \frac{1}{\alpha - 1} \left(\ln \frac{p(x|z_1)^\alpha p(z_1)}{q(z_1|x)} - \alpha \ln \frac{p(x, z_1)}{q(z_1|x)} \right)$$

$$= \frac{1}{\alpha - 1} \left(\ln \frac{p(z_1)}{q(z_1|x)} - \alpha \ln \frac{p(z_1)}{q(z_1|x)} \right)$$

$$= \ln \frac{q(z_1|x)}{p(z_1)}$$

∎

In the single-sample case, where $q(z|x)$ approximates the true posterior much better compared to the multi-sample Monte-Carlo estimators, the representational KL approximates the true KL with a relatively small bias. This bias is, however, asymptotically eliminated by our KL estimator as the number of samples K grows (and reduced for our Rényi estimator), which significantly improves the usage of the latent variables. This is demonstrated in our experiments (e.g. Figures 4 and 7) where we compare methods using KL estimates based on different number of samples from $q(z|x)$.

7. Connection to the β-VAE

We have seen that in the single-sample case our estimators are equal to the representational KL in expectation. Still in the single-sample case, we show that the β-VAE (Higgins et al. 2017) objective defined as

$$\mathcal{O}_{\beta\text{-VAE}}(\beta, x) = \mathbb{E}_{z_1 \sim q(z|x)} \big[\ln p(x|z_1) \big] - \beta \, \text{KL} \big(q(z|x)\|p(z) \big) \tag{19}$$

is equal to the KL, Rényi and power objectives with suitably chosen parameters.

Proposition 9 (β-VAE equivalence). *In the single-sample case with $\hat{S}^1(x, z_1) = \ln \frac{p(x,z_1)}{q(z_1|x)}$, the β-VAE and our single-sample objectives are equal with a suitable choice of λ or α:*

$$
\begin{aligned}
\mathcal{O}_{\beta\text{-VAE}}(\beta, x) &= \mathcal{O}_{\text{KL}}(\hat{S}, 1, 1 - \beta, x) && (\lambda = 1 - \beta) \\
&= \mathcal{O}_R(\hat{S}, 1, 1 - \beta, \alpha, x) && (\lambda = 1 - \beta) \\
&= \mathcal{O}_P(\hat{S}, 1, \beta^{-1}, x) && (\alpha = \beta^{-1}).
\end{aligned}
$$

Proof We prove the KL objective case from (12) and Proposition 7:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{z_1 \sim q(z|x)} \hat{\mathcal{O}}_{\text{KL}}(\hat{S}, 1, \lambda, x, z_1) &= \mathop{\mathbb{E}}_{z_1 \sim q(z|x)} \left[\hat{S}^1(x, z_1) \right] + \lambda \mathop{\mathbb{E}}_{z_1 \sim q(z|x)} \left[\hat{U}^1(x, z_1) - \hat{S}^1(x, z_1) \right] \\
&= \mathop{\mathbb{E}}_{z_1 \sim q(z|x)} \left[\ln \frac{p(x, z_1)}{q(z_1|x)} \right] + \lambda \, \text{KL}\left(q(z|x) \| p(z) \right) \\
&= \mathop{\mathbb{E}}_{z_1 \sim q(z|x)} \left[\ln p(x|z_1) \right] + (\lambda - 1) \, \text{KL}\left(q(z|x) \| p(z) \right).
\end{aligned}
$$

The proof for the Rényi case follows a similar route, starting from (16) and using Proposition 8. Finally, for the power objective, $\lambda = (\alpha - 1)/\alpha$, so $\alpha = \beta^{-1}$ recovers $\lambda = 1 - \beta$. ∎

8. Experiments

The goal of our experiments is to demonstrate the difficulty of inference with VAEs and Monte-Carlo objectives and to evaluate the proposed methods. Our results indicate the presence of a severe tradeoff between data fit and latent variable usage. We emphasize that, for progress to be made, the choice of evaluation method must acknowledge the existence of this tradeoff. In this work, evaluation is performed in terms of Pareto frontiers of data likelihood vs latent usage curves; reporting a single, best data likelihood would always pick the point with zero latent usage. Results with our proposed estimators, either with continuous latents and DReG or discrete latents and the VIMCO base estimator, markedly improve on their baselines, which do not have multiple samples or cannot use them as efficiently. The improvement is especially significant with discrete latents.

Instead of trying to improve the predictive performance directly, first we demonstrate the difficulty of inference on a simple, synthetic data set and a model to which posterior collapse comes easy. These experiments focus exclusively on data fit to highlight the tradeoff against latent variable usage. After the experiments on synthetic data, we move on to language modelling with recurrent networks, a very hostile application for VAEs (Bowman et al. 2015).

8.1 Experiments with Synthetic Data

Every data point is a single symbol drawn from a discrete uniform distribution over a vocabulary of 10000 symbols. Thus, the optimal solution is to assign probability 1/10000 to each symbol, which can be trivially satisfied by a simple model that ignores the latents. Note that $\lambda > 0$ makes such solutions suboptimal. In fact, for all priors such that $H(Z) \geqslant H(X)$, the optimal solution must capture all information in the latents (i.e. $I(X, Z) = H(X)$). Our goal with these experiments is to demonstrate the difficulty of fitting the data while using the latents.

8.1.1 MODEL ARCHITECTURE

The encoder implementing q assigns an embedding (Mikolov et al. 2013) to each word in the vocabulary, feeds the embedding to a two-layer, densely connected neural network with $\tanh$ non-linearities. For continuous latents, denoting the output of the last encoder layer for a given x with o, $q(z|x)$ follows an isotropic normal distribution $\mathcal{N}(f(o), \mathrm{diag}(\exp(g(o))))$, where f and g are affine transformations parameterized as densely connected neural network layers. For categorical latents, $q(z|x)$ is proportional to $\exp(o_z)$ (i.e. it is $Categorical(\mathrm{softmax}(o))$). The decoder is a neural network similar to the encoder but with a final softmax layer. In the decoder, values of continuous latents are fed directly as input to the first layer, but values of categorical latents are first embedded. To compensate for this discrepancy, decoders with categorical latents have only one hidden layer. All embeddings and hidden layers are of size 128.

8.1.2 EVALUATION METHODOLOGY

Since there is no single number to summarize the tradeoff between latent usage and the quality of the model, we plot the model's latent usage and the expected negative log-likelihood (NLL). The NLL of optimal solutions is the entropy of the data distribution, $\ln 10000 \approx 9.210$. The models' NLLs are estimated using IWAE with 100 samples. We quantify latent variable usage as the mutual information $I_p(X, Z)$, estimated as the average $\hat{U}^K(x, z_{1:K}) - \ln \hat{p}^K(x, z_{1:K})$ (9) with $K = 100$ samples. We validated empirically that the variance of these estimates is small (< 0.01) over all feasible latent usage values. Zero latent usage corresponds to posterior collapse.

These plots (like Figure 3b in Alemi et al. 2017) carry the same information as rate-distortion curves, which can be recovered by subtracting the rate I_p from the NLL. Only the measurements on the Pareto frontier are shown, and we tune hyperparemeters, such as the learning rate and λ, of the augmented objective (5) to push the Pareto frontier towards more efficient latent usage. See Appendix C for details.

8.1.3 OVERVIEW OF RESULTS

Results for continuous and discrete latents are presented in separate sections, following a similar progression:

- We first show that the choice of base estimator matters: the variance issues of IWAE and REINFORCE limit the potential of the proposed method.
- Next, with DReG and VIMCO we present clear improvements over the single-sample baselines.
- Finally, as an ablation study, we verify that using multiple samples in both the marginal likelihood and the mutual information terms of the objective is necessary for best performance.

We also note that without augmenting the objective with a mutual information term, our model tends to fit the data perfectly with negligible latent usage. These degenerate curves consisting of a single point are omitted from the plots.

8.1.4 CONTINUOUS LATENTS

With 40 continuous latents and an isotropic standard normal prior, the base estimator is either the standard ELBO, IWAE or IWAE-DReG. In the ELBO, the KL term (2) is computed analytically, while in IWAE 1, the single sample from the latents is used to estimate it. An improvement to IWAE 1 is the STL estimator from Roeder et al. (2017), which removes a zero-expectation term
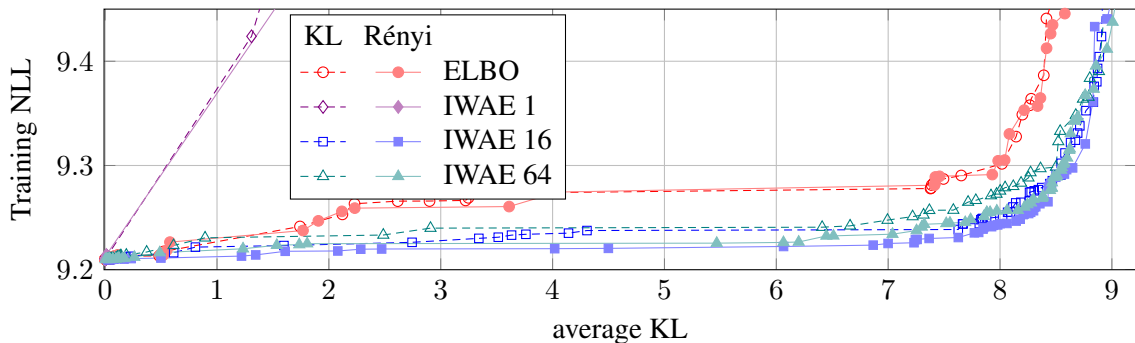
Figure 2: KL and Rényi objectives (empty and full markers) on continuous synthetic data with base estimators ELBO and IWAE N, where N is the number of samples used for estimating both $p(x)$ and $\mathrm{KL}(p(z|x)\|p(z))$. Here and in the following plots, without augmenting the objective with the mutual information term, models tend to fit the data perfectly with negligible latent usage. These degenerate curves consisting of a single point are omitted from the plots. Also, with base estimators ELBO and IWAE 1, our estimators are equivalent to a β-VAE with analytical and estimated KL (see §7), respectively.
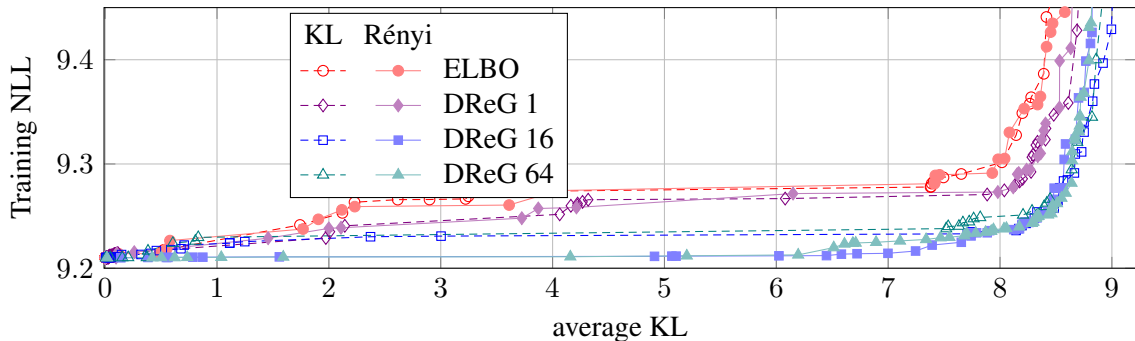
Figure 3: KL and Rényi objectives on continuous synthetic data with base estimators ELBO, DReG.

from the objective and whose gradients have zero variance when the variational approximation is exact. IWAE-DReG, whose objective is also identical to that of IWAE, is a generalization of STL to multiple samples, fixing the signal-to-noise problem of gradient estimates of IWAE (Rainforth et al. 2018), wherein the magnitude of the gradient decreases faster with more samples than the variance of the gradient estimates. Although DReG can also be applied to Reweighted Wake-Sleep (Bornschein and Bengio 2014) and Jackknife Variational Inference (Nowozin 2018), we only use IWAE-DReG in our experiments and let DReG stand for IWAE-DReG without ambiguity.

Our results in Figures 2 and 3 are in agreement with these previous works. In the context of this work, our findings can be summarized as follows.

- Compared to the theoretical optimum, a horizontal line at $\ln 10000 \approx 9.210$, all Pareto curves slope increasingly upwards with more latent usage.
- Higher-variance estimators have steeper curves than lower variance estimators, which is most apparent in the contrast between two of our baselines, IWAE 1 and DReG 1.
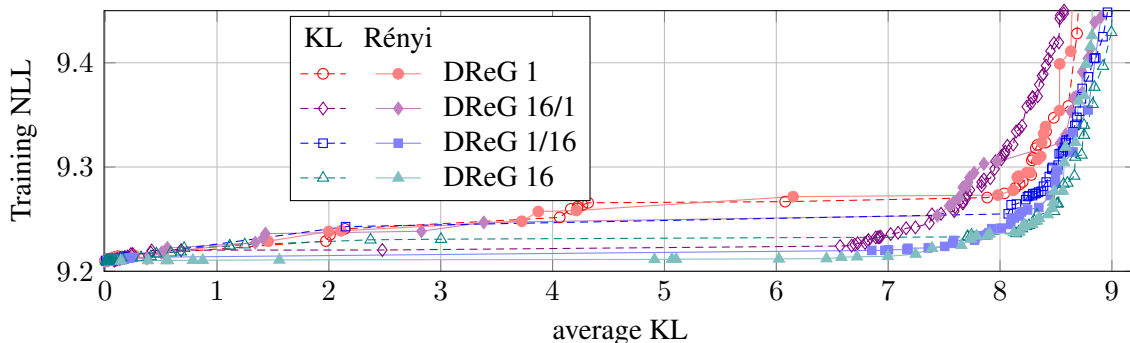
Figure 4: KL and Rényi objectives on continuous synthetic data with base estimator DReG. DReG 1 uses a single sample to estimate both $p(x)$ and $\mathrm{KL}(p(z|x)\|p(z))$. DReG 16/1 uses 16 samples to estimate $p(x)$ and 1 sample for the KL. DReG 1/16 uses 1 sample to estimate $p(x)$ and 16 samples for the KL. Finally, DReG 16 uses the same 16 samples for both terms.

- The single sample estimators ELBO, IWAE 1, DReG 1 - which are equivalent to a β-VAE (§7) - are less efficient in their latent usage than our proposed multi-sample estimators.
- The Rényi objective performs slightly better than the KL objective with the same base estimator.

In more detail, Figure 2 shows that both of our estimators perform much worse with IWAE 1 as the base estimator than with the ELBO. As the number of samples grows, this is reversed, although with a high number of samples we do see the predicted degradation (Rainforth et al. 2018).

Next, Figure 3 confirms DReG's advantage over the IWAE, performing more efficiently than the ELBO even with a single sample. Our multi-sample objectives both improve on the stronger baseline DReG 1 represents. The Rényi objective outperforms the KL objective by a small but consistent margin before all estimators start degenerating quickly nearing the maximal average KL possible.

In Figure 4, we take a closer look at the best performing DReG 16 base estimator to determine whether the improvements are due to a better estimate of the marginal likelihood $\ln p(x)$, or the mutual information term $I_{PD}^p(X, Z)$ in (5). DReG 16/1 and DReG 1/16 (which use multiple samples to estimate only the marginal likelihood or the mutual information, respectively) are less efficient than DReG 16, sometimes being outperformed even by DReG 1. Out of the two, DReG 1/16 performs better, indicating that the variance of the base estimator DReG 1 is low and the problem lies with the mutual information estimate. These results support not only our observations in §5 on the unsuitability of $I_{p,q}$ as an estimate of I_p in the multi-sample case (DReG 16/1), when $q(z|x)$ is not a good approximation of $p(z|x)$, but also demonstrate that $I_{p,q}$ can be improved by taking multiple samples even when $\ln p(x)$ is estimated with a single sample and $q(z|x)$ better approximates the true posterior $p(z|x)$.

8.1.5 DISCRETE LATENTS

Next we performed experiments with 8 categorical latent variables, each with a uniform prior over 10 categories. Using the high-variance REINFORCE base estimator (Figure 5), we could only get a small improvement over the single-sample case with 16 samples and a similar degradation with 64.

However, with the much lower variance VIMCO estimator, we achieved almost perfect results (see Figure 6) with 16 samples. For our purposes, it suffices to think of VIMCO as IWAE applied to
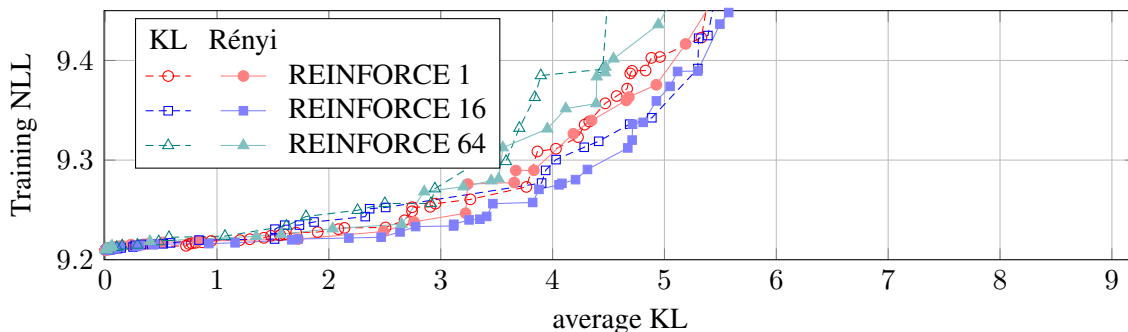
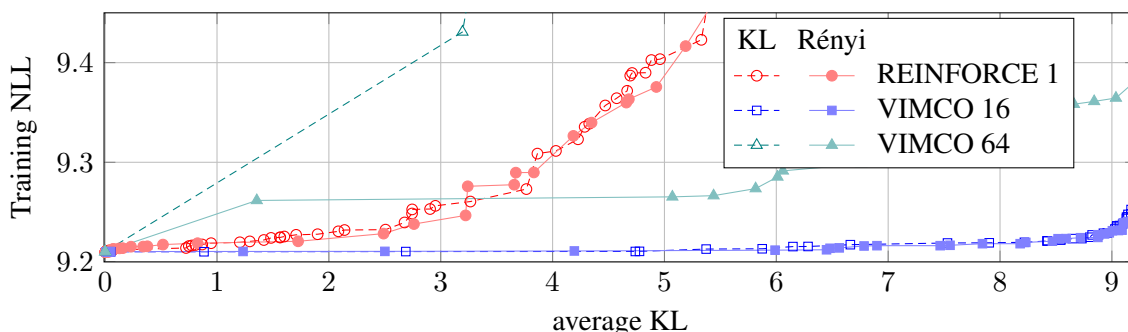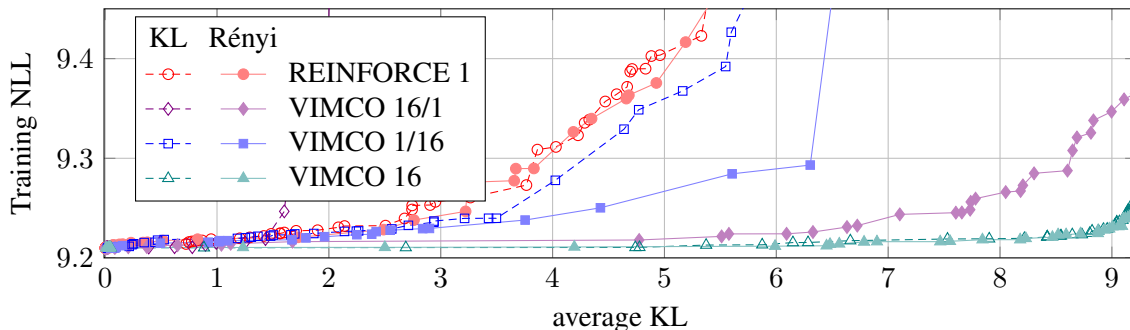Figure 5: KL and Rényi objectives on discrete synthetic data with base estimator REINFORCE.

Figure 6: KL and Rényi objectives on discrete synthetic data with base estimator VIMCO. The single-sample VIMCO is equivalent to REINFORCE.

discrete latents with lower variance gradients. It may be surprising that these results are even better than with continuous latents, especially at near-maximal latent usage. To intuit why, consider the minimum variance posterior achievable for a given level of average KL. For discrete latents, a hard posterior (i.e. of zero variance) is possible depending on the number of latents and categories. For continuous latents, the posterior can never be hard: the mutual information determines a lower bound on the average variance of the posterior.

Increasing the number of samples further to 64 made results much worse, indicating a potential issue with VIMCO, which may be similar to the signal-to-noise issue that DReG addresses, but this is beyond the scope of this paper. Note that there is no single-sample VIMCO since its baseline for the contribution of a sample is computed as the average over the rest of the samples, which is undefined. Assuming a zero baseline, we recover REINFORCE 1, which we use for comparison wherever VIMCO 1 would be needed.

Similarly to Figure 4 in the continuous case, in Figure 7, we try to tease apart the contributions of using multiple samples for estimating the marginal likelihood and mutual information terms of (5). The results are much more pronounced here. Both VIMCO 16/1 and VIMCO 1/16 improve significantly on REINFORCE 1 but only with the Rényi objective. Still, both fall way short of VIMCO 16, which employs multiple samples for both terms. Again, these results support not only our observations in §5 on the unsuitability of $I_{p,q}$ as an estimate of I_p in the multi-sample case

Figure 7: KL and Rényi objectives on discrete synthetic data with base estimator VIMCO. REINFORCE 1 (standing in for VIMCO 1) uses a single sample to estimate both $p(x)$ and $\text{KL}(p(z|x)\|p(z))$. VIMCO 16/1 uses 16 samples to estimate $p(x)$ and 1 samples for the KL. VIMCO 1/16 uses 1 sample to estimate $p(x)$ and 16 samples for the KL. Finally, VIMCO 16 uses the same 16 samples for both terms.

(DReG 16/1), when $q(z|x)$ is not a good approximation of $p(z|x)$, but also demonstrate that $I_{p,q}$ can be improved by taking multiple samples even when $\ln p(x)$ is estimated with a single sample and $q(z|x)$ better approximates the true posterior $p(z|x)$.

In Appendix A, we also present results for the VQ-VAE (van den Oord et al. 2017) *without* augmenting its objective with a mutual information term. Since the KL cost in VQ-VAEs is determined by the latent space and is fixed during training, we tune the number of latent variables and the number of categories to control the mutual information. Results of the VQ-VAE are better than REINFORCE but much worse than VIMCO 16 with either the KL or the Rényi objective.

8.2 Language Modelling Experiments

One of the most challenging applications of variational autoencoders is language modelling with per-sentence latents, as found by Bowman et al. (2015). They recognize the generality of the underspecification issue and attribute the increased difficulty to "the sensitivity of the LSTM to subtle variations in its hidden state as introduced by the sampling process". In this section, we first show that the data fit vs latent usage tradeoff is even more pronounced in the language modelling case than on the synthetic task, then confirm that the proposed estimators improve validation set results in terms of the Pareto frontiers. Once again, the improvement is strongest with discrete latents.

8.2.1 DATA SET

We do sentence-level language modelling on the Penn Treebank (PTB) corpus by Marcus et al. (1993) with preprocessing from Mikolov et al. (2010). Our goal here is to compare inference methods, not to establish a new state of the art, so to reduce the computational burden brought about by hyperparemeter tuning, we truncated sentences to 35 tokens in both the training and validation sets, with a reduction of 3% in the total number of tokens. This truncation is non-standard.
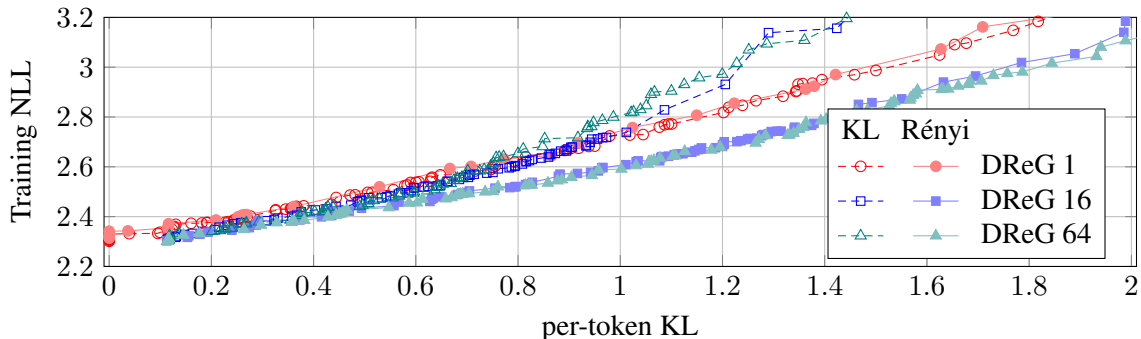
Figure 8: KL and Rényi objectives (empty and full markers, respectively) on Penn Treebank with base estimators DReG N, where N is the number of samples used for estimating both $p(x)$ and $\mathrm{KL}(p(z|x)\|p(z))$.

8.2.2 Model Architecture

The model architecture is like in the synthetic case except the encoder embeds the input tokens and feeds them to a one-layer, bidirectional LSTM (Hochreiter and Schmidhuber 1997) and the output o (from which the parameters of the variational posterior $q(z|x)$ are computed) is the average of the last states of LSTMs corresponding to the two directions. There is a fixed number of latents for all sentences, regardless of their length. The decoder is a unidirectional LSTM whose input is the embedding of the previous token plus the values of the latent variables. Unless stated otherwise, the embedding and hidden sizes are all 128. Similarly to the synthetic case, we use either 40 real-valued latents with an isotropic standard normal prior or 8 categorical latent variables, each with a uniform prior over 10 categories.

8.2.3 Evaluation Methodology

Following previous works, both the reported NLL and the average KL values are averages over tokens in the data set. Since there is only a single set of latents per sentence, this means that the average sentence-level KL is about 21 (the average number of tokens per sentence) times larger. For expediency, only the base estimators that performed best on the synthetic data set are considered, leaving us with DReG for continuous, and VIMCO for discrete latents. Training and validation NLLs are estimated using IWAE with 100 samples. Test NLLs in Table 1 are estimated using IWAE with 500 samples.

8.2.4 Training Fit

In terms of training fit, the language modelling results are similar to those of the synthetic case. In Figure 8, there is improvement in the continuous case with multiple samples, initially with both the KL and the Rényi objectives but only with the Rényi at high latent usage.

Figure 9 shows the discrete latents case. Once again, with either the KL or the Rényi objective, the results improve markedly, outperforming the continuous models. Note that the KL range is limited by the entropy of the latent space at around $\ln(10^8)/21 \approx 0.87$.
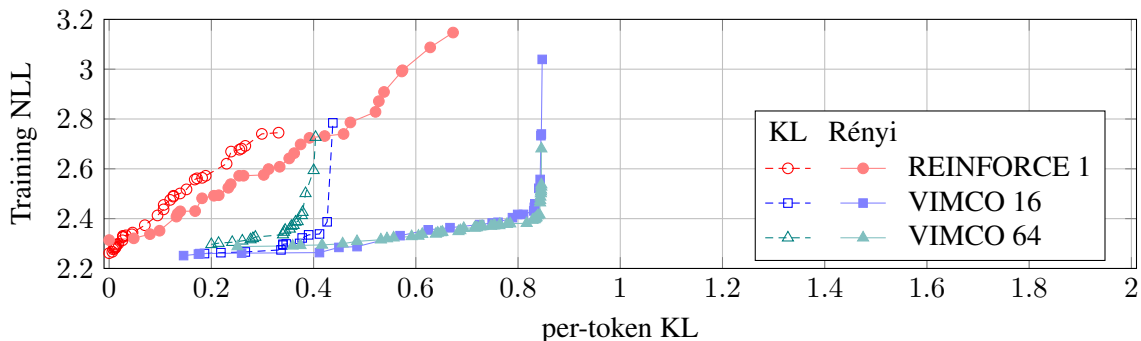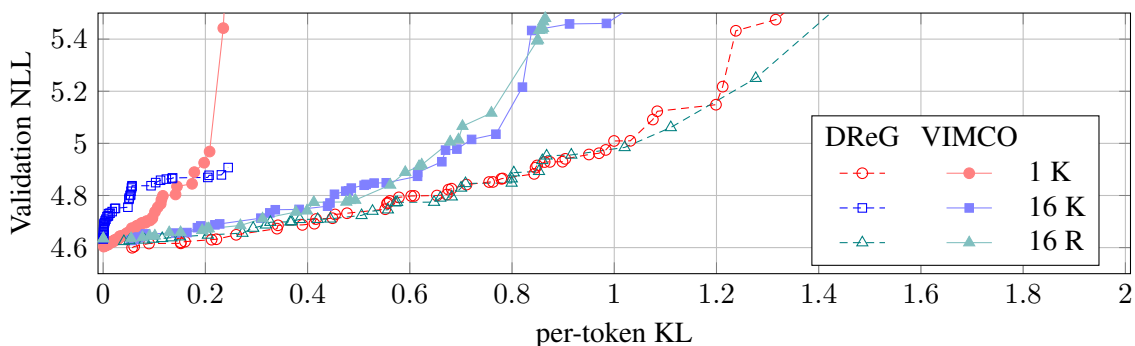
Figure 9: KL and Rényi objectives on PTB with base estimator VIMCO.

Figure 10: Validation NLL with naive dropout using DReG and VIMCO on PTB. The row with *16 K* refers to 16-sample DReG or VIMCO with the KL objective, while *R* stands for the Rényi objective.

Furthermore, as Appendix B.2 shows, improvements in training fit require multiple samples in both terms, even more so than in the experiments on the synthetic data earlier.

8.2.5 VALIDATION AND TEST RESULTS

Our initial results on the validation set were not very positive. We tuned hyperparameters on the validation set, where following Melis et al. (2017), we introduced three additional hyperparameters in the decoder: the rate of dropout applied to the input embedding, the recurrent state, and the output embedding. As Figure 10 shows, with this strong form of regularization, the multi-sample results brought no improvement with continuous latents. With discrete latents, the results improved but not nearly to the extent observed on the training set, and discrete latents performed considerably worse than their continuous counterparts.

We suspected that the stochasticity from dropout may accentuate the problems of variational inference. To verify, we repeated the validation experiment with L2 regularization instead of dropout (see Figure 11). In this setting, the validation results are more consistent with training fit. As before, multiple samples significantly improve efficiency. Unlike the training fit though, validation NLL with discrete latents degrades faster than with continuous ones. These results suggest that the additional stochasticity of dropout indeed poses a challenge. On the other hand, with only L2 regularization,
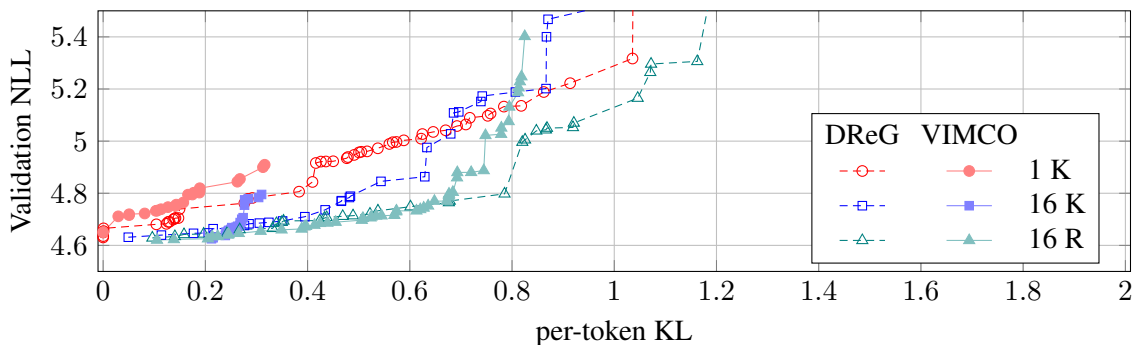
Figure 11: Validation NLL with L2 regularization using DReG and VIMCO on PTB.
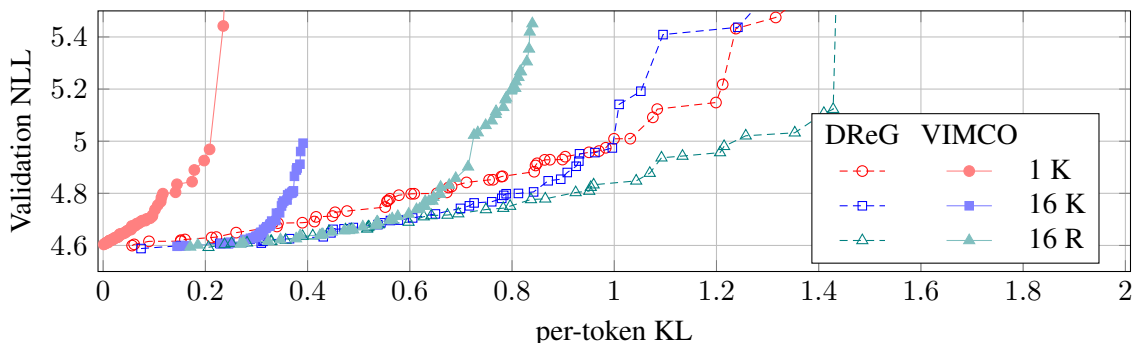
Figure 12: Validation NLL with DReG and VIMCO on PTB, using the same dropout mask for all latent samples belonging to the same sentence in a minibatch.

the best NLL is higher than with dropout. This translates to a few perplexity points, which may seem small on the graphs presented here, but in language modelling, kingdoms have been won or lost on such differences (Merity et al. 2017).

To have the best of both worlds, the best NLL of dropout and the Pareto curves of L2 regularization, we went back to dropout, but this time we tried using the same dropout mask for all latent samples belonging to the same sentence in a minibatch. As Figure 12 shows, this change was successful, and we observed that our proposed estimators improve latent usage for both continuous and discrete latents, and discrete and continuous latents are on par up to an average KL of about 0.5. This constitutes a significant advance in modelling with discrete latents.

Contrary to results of Pelsmaeker and Aziz (2019), we did not find that introducing latent variables significantly improves outright perplexity. As Table 1 shows, the best test perplexity improves only slightly with more samples for this combination of a small model and strong regularization.

We also performed experiments with the power objective (Appendix B.3) to better understand the tradeoff it represents. We found that its trivial implementation cost comes at the price of decreased efficiency relative to the general Rényi objective, of which it is a special case.

In closing, we would like to emphasize the importance of the results in Figure 12. It is not only that small and large improvements have been made, but that the evaluation throughout focussed on the apparent tradeoff between data fit and latent usage. Every point on the Pareto curves is the result

	Base	DReG			VIMCO		
	#Samples	1	16		1	16	
		K / R	K	R	K / R	K	R
hs=128	Perplexity	92.5	91.7	91.6	94.6	92.1	92.5
	NLL	4.527	4.518	4.517	4.549	4.523	4.527
	Per-token KL	0.047	0.131	0.126	0.000	0.164	0.120
hs=256	Perplexity	83.9	82.8	82.2	84.7	83.4	83.1
	NLL	4.429	4.416	4.409	4.439	4.424	4.420
	Per-token KL	0.000	0.140	0.138	0.001	0.167	0.094

Table 1: Best test results for DReG and VIMCO in the continuous and discrete cases, estimated with IWAE and 500 samples. These optima are at low (or, in the single-sample case, mostly negligible) latent usage.
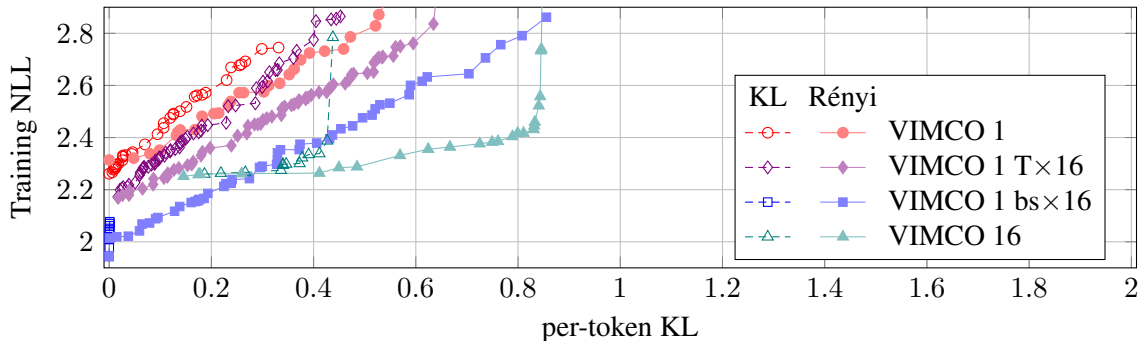
Figure 13: Training NLL on PTB with KL and Rényi objectives and base estimator VIMCO. VIMCO 1 T×16 is trained 16 times longer. VIMCO 1 bs×16 has a 16 times larger batch size. While their best NLL at very low latent usage is lower than that of VIMCO 16, they lose this advantage at higher latent usage.

of tuning several hyperparemeters: the learning rate, λ, α for the Rényi objective, and three different dropout rates. These curves capture and communicate what most published experiments do not and what single numbers (e.g. in Table 1) cannot: a reliable comparison of a latent variable model to a strongly regularized baseline over a wide range of latent usage.

8.2.6 SINGLE- VS MULTI-SAMPLE AT THE SAME COMPUTATIONAL BUDGET

The argument for using more samples in Monte-Carlo objectives is that their lower bound is tighter, resulting in a better data fit vs latent usage tradeoff. Having equipped multi-sample Monte-Carlo objectives with mutual information constraints, we have indeed demonstrated improvements in this tradeoff. Here, we present additional experiments to answer whether performing more computation with a single-sample estimator can compensate the advantages of multiple samples. To this end, we
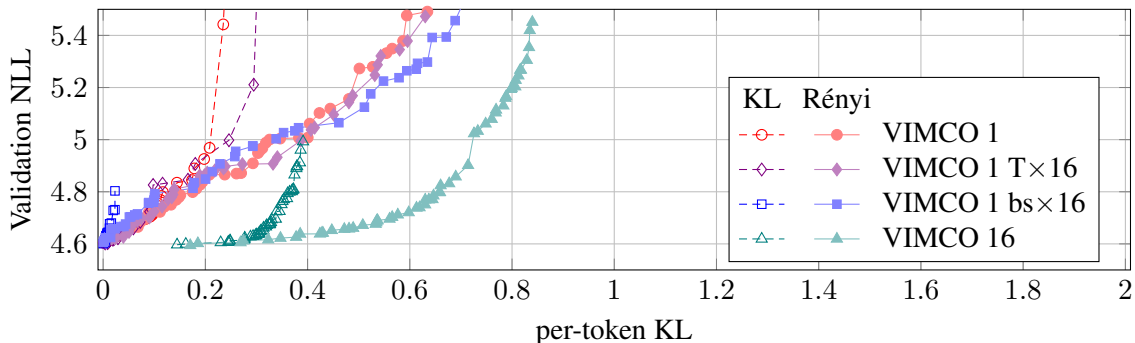
Figure 14: Validation NLL on PTB with KL and Rényi objectives and base estimator VIMCO. VIMCO 1 T×16 is trained 16 times longer. VIMCO 1 bs×16 has a 16 times larger batch size. VIMCO 16 performs better than either.

compare 16-sample to single-sample estimators, where the latter are trained for 16 times more optimization steps (*T×16*) or with a batch size that is 16 times larger (*bs×16*) to make the computational cost more equal. In fact, these perform an equal number of gradient calculations but more forward computation as they evaluate $q(z|x)$ 16 times more than the 16-sample Monte-Carlo estimator.

In Figure 13, we observe that with more computation, the best training fit of VIMCO 1 improves beyond that of VIMCO 16 , but the tradeoff remains severe, as evidenced by that the shapes of the Pareto curves hardly change from the single-sample baseline.

With regards to validation results, we expected the strong regularization provided by tuning three different dropout rates to compensate for possible regularization effects of fewer optimization steps and smaller batch sizes. Consequently, all single-sample estimators shall exhibit similar validation results at zero latent usage. Figure 14 confirms this, and shows that the steeper training curves translate to steeper validation curves, indicating that increasing the computation does not address the ineffeciency of the inference method, and the apparent improvement in training comes at the cost of more overfitting. In Appendix B.1, Figure 24 and Figure 25 tell a similar story for DReG, but the results are less conclusive there since the curves are much closer.

9. Conclusions

We identified the underspecification of the model and the looseness of the lower bound as two important issues that cause posterior collapse and put forward a natural combination of Monte-Carlo objectives and mutual information constraints to address both at the same time. The proposed estimators of the mutual information reuse samples from the Monte-Carlo objective of the marginal likelihood to estimate the KL of the true posterior from the prior. We showed that the representational KL, often used in mutual information constraints, corresponds to the single-sample version of our estimators. Taking more samples both tightens the lower bound and reduces the variance of the estimate of the true KL. Our experimental results support these theoretical predictions and underline the need to use multiple samples for both terms of the objective.

Recognizing that the problem of underfitting becomes more acute with increased latent usage, we emphasized evaluating estimators on the training set, where regularization does not cloud the
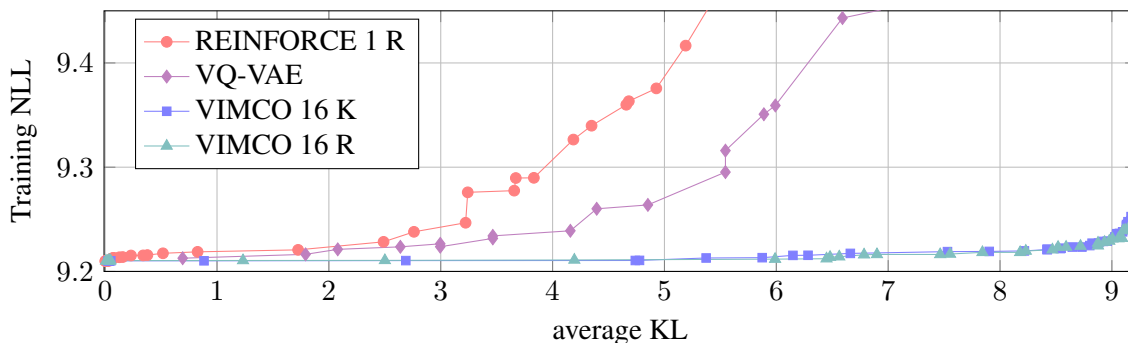
Figure 15: VQ-VAE on synthetic data with tuned latent sizes compared to REINFORCE 1 and VIMCO 16 with the KL and Rényi objectives.

picture, instead of going outright for improvements in held-out performance. In addition, evaluation was performed in terms of the Pareto frontier of negative log-likelihood vs latent usage curves since reporting a single number cannot capture the tradeoff between the two quantities.

The results demonstrated increased efficiency in latent usage on both the synthetic and language modelling tasks. For discrete latent spaces in particular, the improvements have been dramatic: from a very weak baseline, data fit improved beyond that of models with continuous latents on both data sets. In terms of validation results on Penn Treebank, the best continuous and discrete models and estimators are closely matched up until a significant, per-token KL of 0.5 (about 10.5 as a per-sentence KL).

This work is towards opening the door to making the latents truly useful. We believe that substantial gains in genaralization and utility in down-stream tasks can be achieved by shaping the latent space. It also remains to be explored how to best encode the true posterior as the representation for use in down-stream models since the approximate posterior is no longer a suitable choice in the context of Monte-Carlo objectives.

In summary, our Mutual Information constrained Monte-Carlo Objectives (MICMCOs) help achieve a better tradeoff between modelling the data and using the latent variables to drive the generative process: a prerequisite for fulfilling the promise of generative modelling. This tradeoff is still quite severe though, and there is a lot of room for improvement.

Appendix A. Additional Experiments on Synthetic Data

Since the KL cost in VQ-VAEs is determined by the latent space and is fixed during training, we tune the number of latent variables and the number of categories to control the mutual information. As Figure 15 shows, the VQ-VAE is generally better than REINFORCE 1 but quite far from the optimum and much worse than VIMCO 16 with either the KL or the Rényi objective.
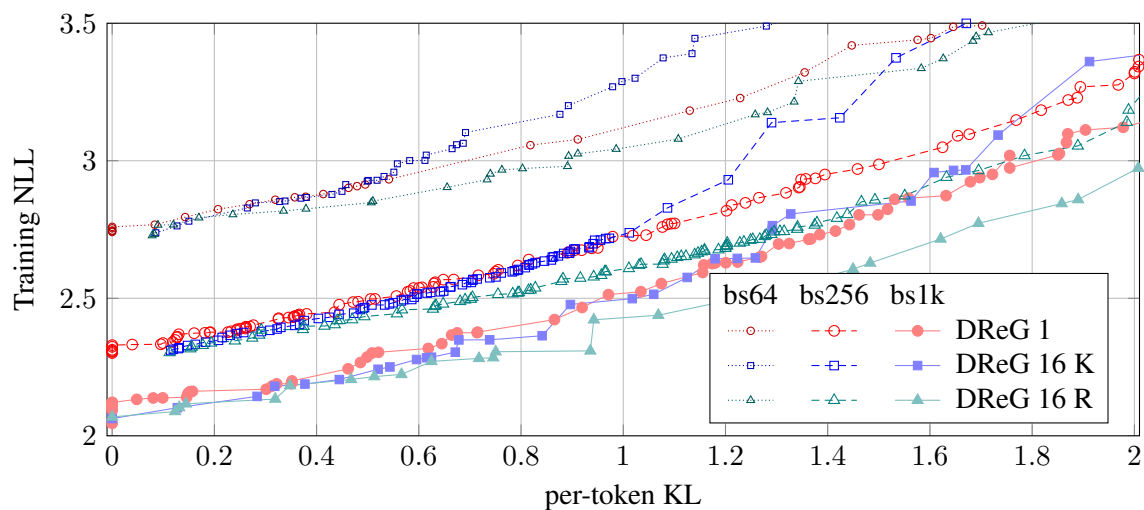
Figure 16: The effect of batch size with DReG and 128 hidden units on PTB.
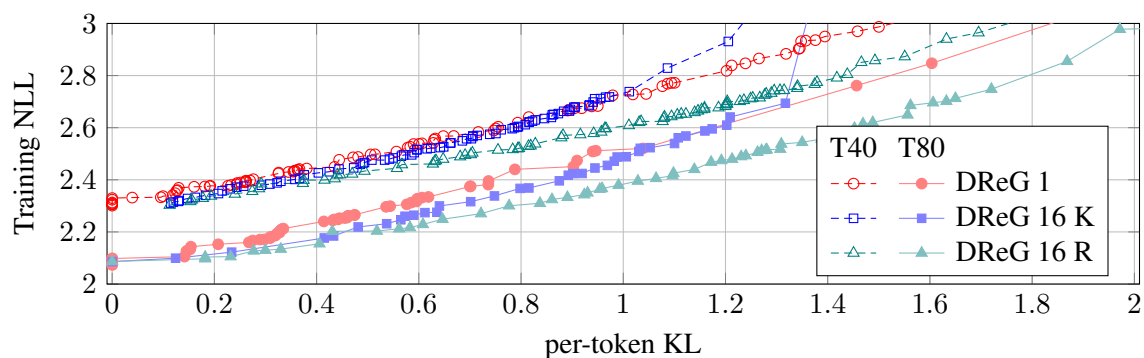

Figure 17: The effect of optimization length (40 or 80 thousand optimization steps) with DReG and 128 hidden units on PTB.
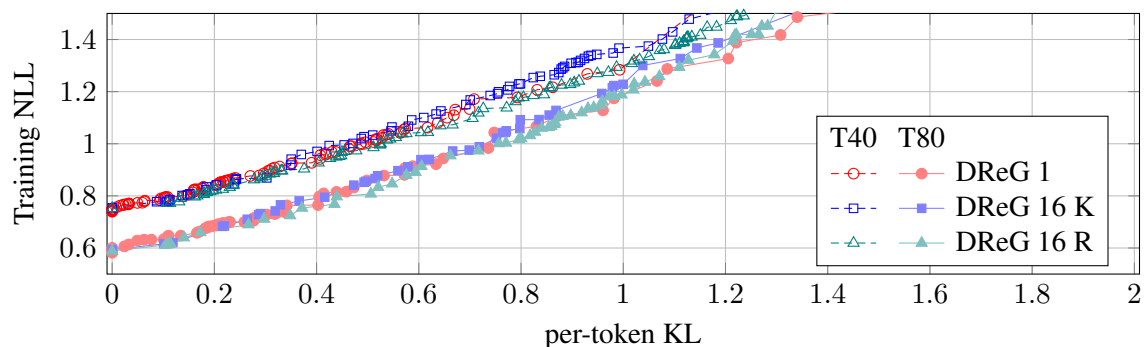

Figure 18: The effect of optimization length (40 or 80 thousand optimization steps) with DReG with 256 hidden units on PTB.
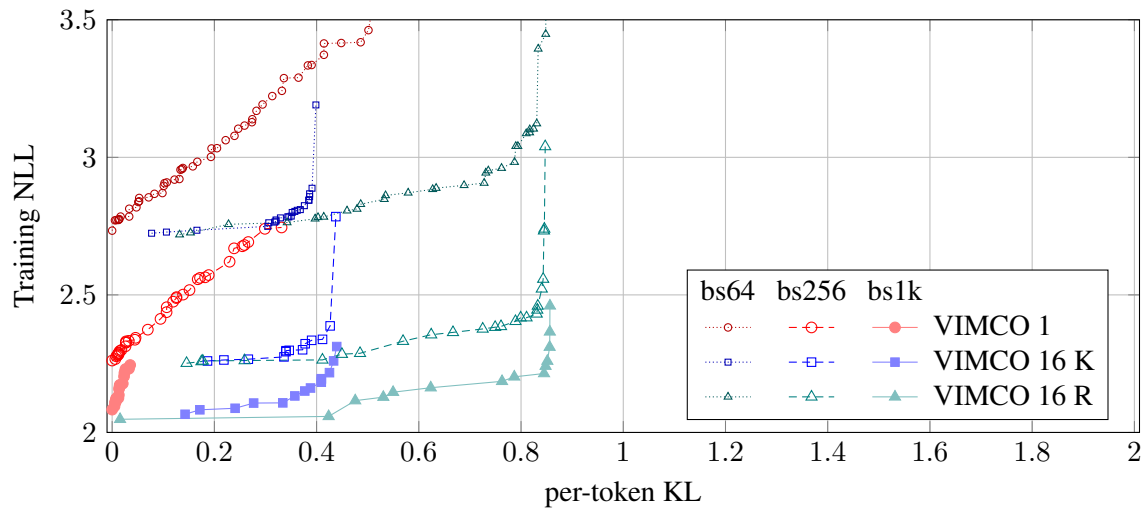
Figure 19: The effect of batch size with VIMCO and 128 hidden units on PTB.
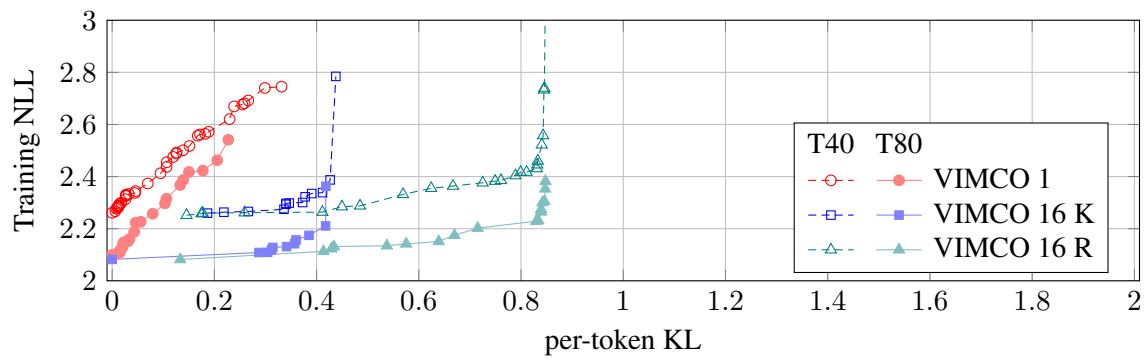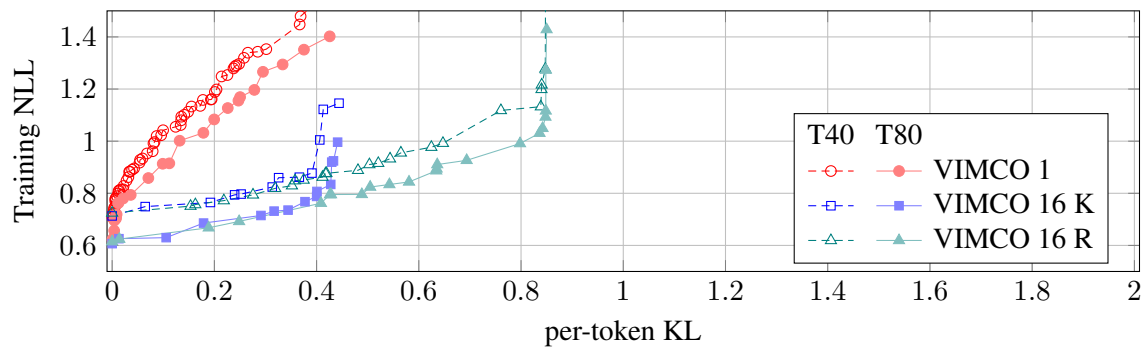

Figure 20: The effect of optimization length (40 or 80 thousand optimization steps) with VIMCO and 128 hidden units on PTB.

Figure 21: The effect of optimization length (40 or 80 thousand optimization steps) with VIMCO and 256 hidden units on PTB.
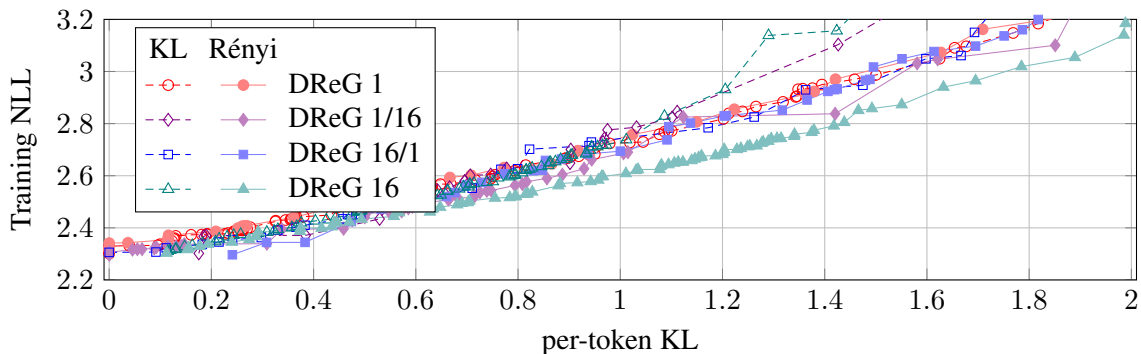
Figure 22: KL and Rényi objectives on PTB with base estimator DReG. DReG 1 uses a single sample to estimate both $p(x)$ and $\mathrm{KL}(p(z|x)\|p(z))$. DReG 16/1 uses 16 samples to estimate $p(x)$ and 1 samples for the KL. DReG 1/16 uses 1 sample to estimate $p(x)$ and 16 samples for the KL. Finally, DReG 16 uses the same 16 samples for both terms.
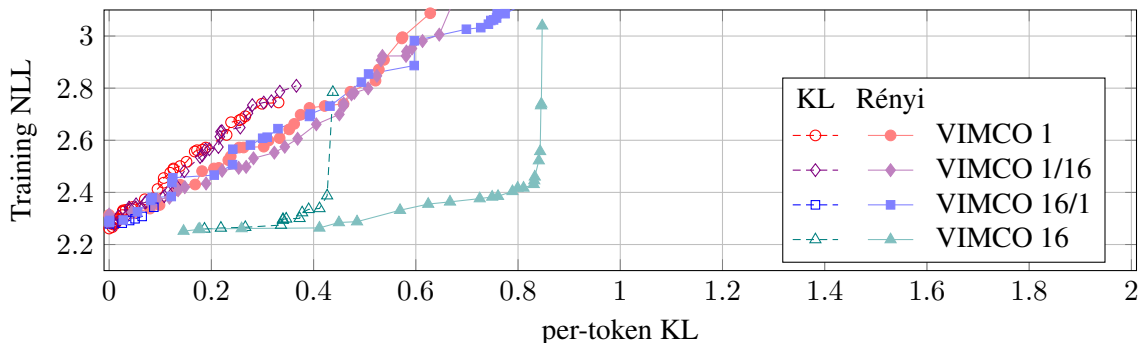

Figure 23: KL and Rényi objectives on PTB with base estimator VIMCO. VIMCO 1 uses a single sample to estimate both $p(x)$ and $\mathrm{KL}(p(z|x)\|p(z))$. VIMCO 16/1 uses 16 samples to estimate $p(x)$ and 1 samples for the KL. VIMCO 1/16 uses 1 sample to estimate $p(x)$ and 16 samples for the KL. Finally, VIMCO 16 uses the same 16 samples for both terms.

Appendix B. Additional Language Modelling Experiments

B.1 Robustness

We performed additional experiments to verify that our results about the relative merits of the estimators are robust to different choices of batch size, number of parameters and optimization length. With continuous latents, we focussed on DReG, the best performing base estimator and varied the batch size (Figure 16), the length of optimization (Figure 17), and the length of optimization again at two times the model size (Figure 18). For VIMCO, the best performing base estimator for the discrete latents, Figures 19 to 21 tell a similar story. In most cases, we observed that varying these nuisance factors only shifted the Pareto curves downwards or upwards, leaving their relative positions the same. The exception is Figure 18, DREG with 256 hidden units, where the differences between the three estimators are very small.
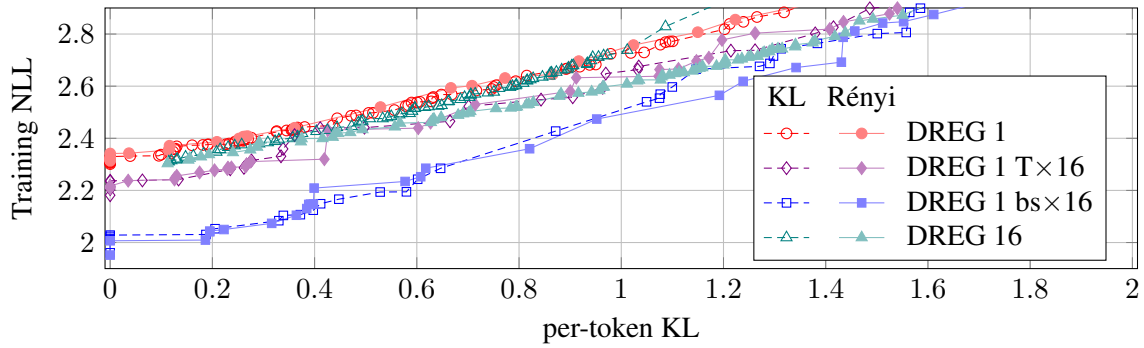
Figure 24: Training NLL on PTB with KL and Rényi objectives and base estimator DREG. DREG 1 T×16 is trained 16 times longer. DREG 1 bs×16 has a 16 times larger batch size. While their best NLL at very low latent usage is lower than that of DREG 16, they lose this advantage at higher latent usage.
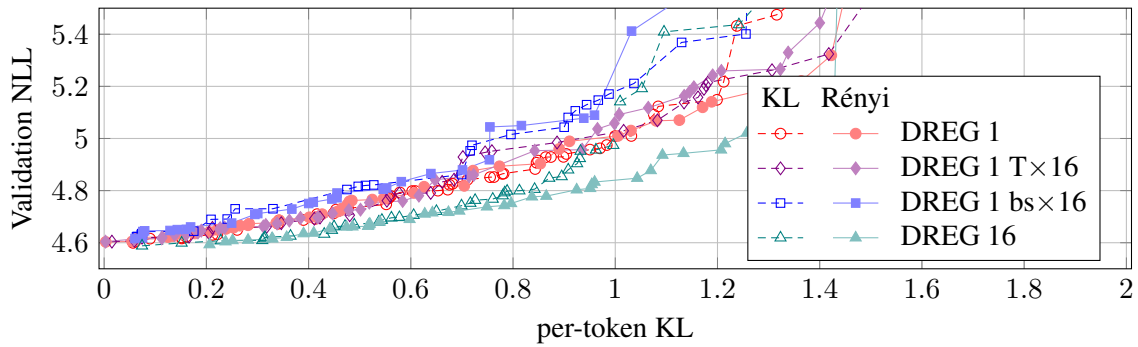

Figure 25: Validation NLL on PTB with KL and Rényi objectives and base estimator DREG. DREG 1 T×16 is trained 16 times longer. DREG 1 bs×16 has a 16 times larger batch size. DREG 16 generalizes better than either.
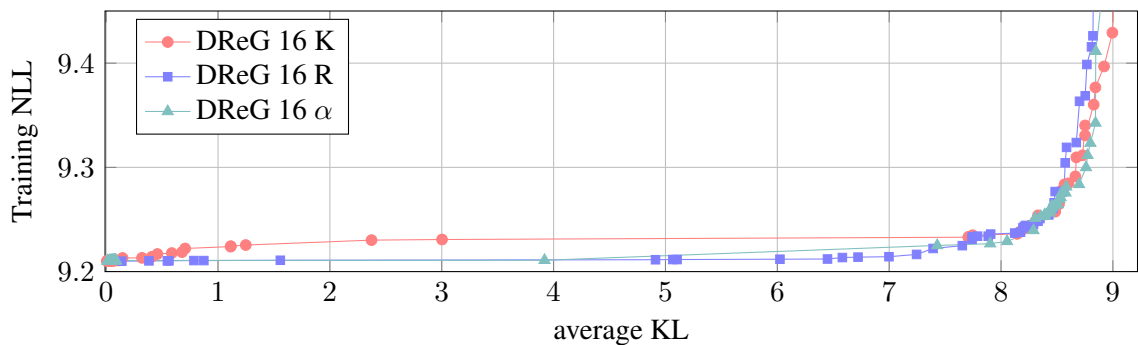

Figure 26: The power objective with DReG on the synthetic data set compared to the KL and Rényi objectives.
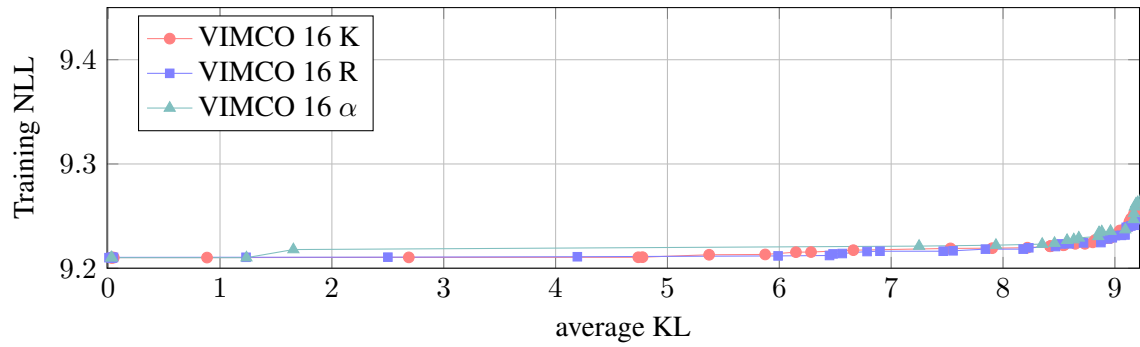
Figure 27: The power objective with VIMCO on the synthetic data set compared to the KL and Rényi objectives.
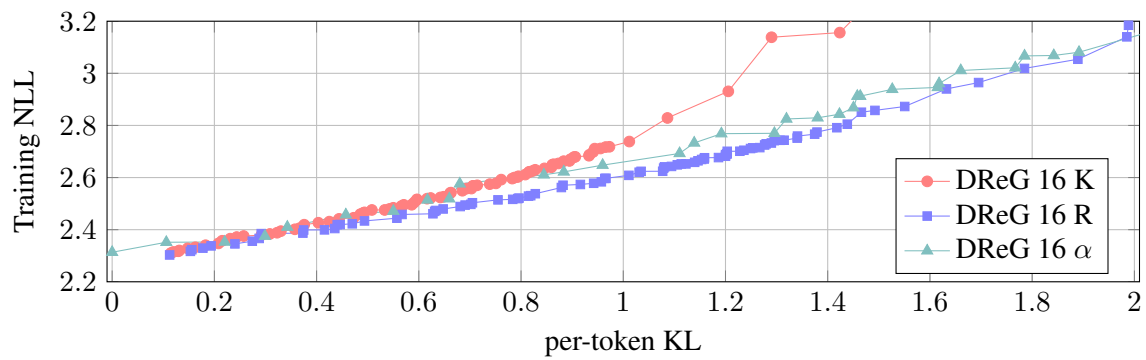

Figure 28: The power objective with DReG on PTB compared to the KL and Rényi objectives.
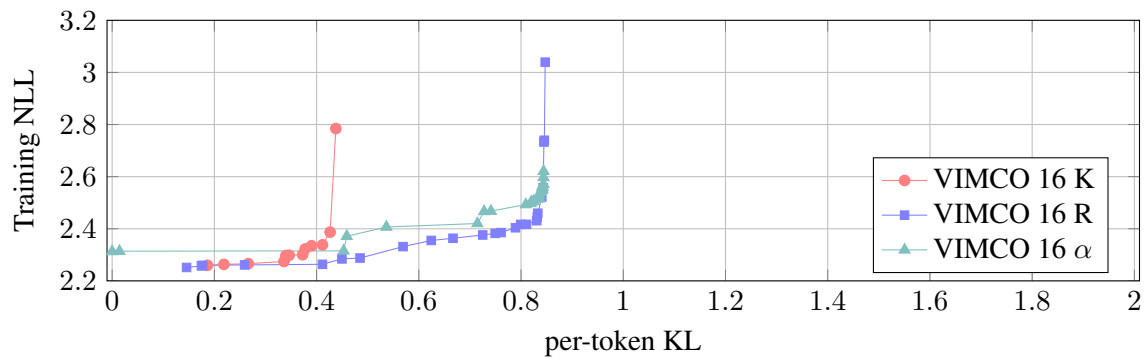

Figure 29: The power objective with VIMCO on PTB compared to the KL and Rényi objectives.

Finally, similar to those in §8.2.6, we present experiments with single-sample DReG at an increased computational budget. As training (Figure 24) and validation (Figure 25) results show, the overall picture may be the same as for VIMCO, (that is, steeper training curves translate to steeper validation curves), but it is harder to assess this since the curves are closer.

B.2 Asymmetric Samples

We now investigate whether the improvements are due to a better estimate of the marginal likelihood $\ln p(x)$, or the mutual information term $I_{p_D}^p(X, Z)$ in (5). Figures 22 and 23 show that improvements in training fit require multiple samples in both terms, even more so than in the experiments on the synthetic data earlier (Figures 4 and 7).

B.3 Experiments with the Power Objective

Recall that the power objective (18) is a special case of the Rényi objective (15) with the choice of $\lambda = (\alpha - 1)/\alpha$. With this λ, the objective simplifies to $\mathbb{E}_{p_D(x)} \ln p^\alpha(x)$. Since $\ln p^\alpha(x) = \ln(p(x|z)^\alpha p(z)) = \alpha \ln p(x|z) + \ln p(z)$, implementing the power objective is as easy as upweighting the log-likelihood term $\ln p(x|z)$. Here we investigate whether at the same time, by tying λ and α, we can maintain parity with the Rényi objective in terms efficiency of latent usage. As Figures 26 to 29 show, the power objective is often better than the KL objective but lags the Rényi objective as α determines both λ, the weight of the mutual information term and its bias with respect to the KL.

Appendix C. Optimization Settings

In all experiments, we use the Adam optimizer (Kingma and Ba 2014) with $\beta_1 = 0$, $\beta_2 = 0.999$, and $\epsilon = 1e - 8$. We tune hyperparameters using a black-box hyperparameter tuner based on batched Gaussian Process Bandits (Golovin et al. 2017). Hyperparameters and their ranges are listed in Table 2. The learning rate is the only hyperparameter tuned in all experiments. The rest only apply in specific circumstances. *Input and output dropout* are the dropout rates applied to the inputs and outputs of the LSTM, respectively, while *state dropout* is the dropout rate for the LSTM's recurrent state from the previous time step (Gal and Ghahramani 2016). For a description of the VQ-VAE parameters see van den Oord et al. (2017).

References

Alexander A Alemi, Ben Poole, Ian Fischer, Joshua V Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. *arXiv preprint arXiv:1711.00464*, 2017.

Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. 2014.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

hyperparameter	min	max	scale	
learning rate	0.0001	0.01	log	
λ	-0.05	0.9999		KL and Rényi only
α	0.86	16.0	log	power objective only
input dropout	0.0	0.9		validation only
state dropout	0.0	0.8		validation only
output dropout	0.0	0.95		validation only
L2 penalty coefficient	5e-6	1e-3	log	validation only
number of latents	1	8		VQ-VAE only
number of categories of latents	2	20		VQ-VAE only
VQ β	0.01	20.0	log	VQ-VAE only
VQ decay	0.9	0.99999	log	VQ-VAE only

Table 2: Hyperparameter tuning ranges.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741, 2017.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1019–1027, 2016.

Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D Sculley. Google Vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1487–1495. ACM, 2017.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *International Conference on Learning Representations*, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Machine Learning*, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in neural information processing systems*, pages 529–536, 1995.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Chin-Wei Huang, Shawn Tan, Alexandre Lacoste, and Aaron C Courville. Improving explorability in variational inference with annealed variational objectives. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Ferenc Huszár. Is maximum likelihood useful for representation learning? http://web.archive.org/web/20190704042553/https://www.inference.vc/maximum-likelihood-for-representation-learning-2/, 2017. Accessed: 2020-04-15.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Yoon Kim, Sam Wiseman, Andrew C Miller, David Sontag, and Alexander M Rush. Semi-amortized variational autoencoders. *arXiv preprint arXiv:1802.02550*, 2018.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751, 2016.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modeling. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, pages 6573–6583, 2017.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The Penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

Arya D McCarthy, Xian Li, Jiatao Gu, and Ning Dong. Improved variational neural machine translation by promoting mutual information. *arXiv preprint arXiv:1909.09237*, 2019.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*, 2017.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing LSTM language models. *arXiv preprint arXiv:1708.02182*, 2017.

Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. Importance sampling techniques for policy optimization. *J. Mach. Learn. Res.*, 21:141–1, 2020.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3, 2010.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. *arXiv preprint arXiv:1402.0030*, 2014.

Andriy Mnih and Danilo J Rezende. Variational inference for monte carlo objectives. *arXiv preprint arXiv:1602.06725*, 2016.

Sebastian Nowozin. Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference. 2018.

Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.

Tom Pelsmaeker and Wilker Aziz. Effective estimation of deep generative language models. *arXiv preprint arXiv:1904.08194*, 2019.

Mary Phuong, Max Welling, Nate Kushman, Ryota Tomioka, and Sebastian Nowozin. The mutual autoencoder: Controlling information in latent code representations, 2018. URL `https://openreview.net/forum?id=HkbmWqxCZ`.

Tom Rainforth, Adam R Kosiorek, Tuan Anh Le, Chris J Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. *arXiv preprint arXiv:1802.04537*, 2018.

Ali Razavi, Aaron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-VAEs. In *International Conference on Learning Representations*, 2019.

Ali Lotfi Rezaabad and Sriram Vishwanath. Learning representations by maximizing mutual information in variational autoencoders. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2729–2734. IEEE, 2020.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems 30*, 2017.

Andriy Serdega and Dae-Shik Kim. Vmi-vae: Variational mutual information maximization framework for vae with discrete and continuous priors. *arXiv preprint arXiv:2005.13953*, 2020.

Rui Shu, Hung H. Bui, Shengjia Zhao, Mykel J. Kochenderfer, and Stefano Ermon. Amortized inference regularization. In *NeurIPS*, pages 4398–4407, 2018.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29:3738–3746, 2016.

Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *International Conference on Machine Learning*, pages 1971–1979, 2014.

Jakub Tomczak and Max Welling. Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223. PMLR, 2018.

George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J Maddison. Doubly reparameterized gradient estimators for monte carlo objectives. *arXiv preprint arXiv:1810.04152*, 2018.

Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017.

R Williams. A class of gradient-estimation algorithms for reinforcement learning in neural networks. In *Proceedings of the International Conference on Neural Networks*, pages II–601, 1987.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *International conference on machine learning*, pages 3881–3890. PMLR, 2017.

Serena Yeung, Anitha Kannan, Yann Dauphin, and Li Fei-Fei. Tackling over-pruning in variational autoencoders. *arXiv preprint arXiv:1706.03643*, 2017.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5885–5892, 2019.