

Bounding Evidence and Estimating Log-Likelihood in VAE

Łukasz Struski

Marcin Mazur

Paweł Batorski

Przemysław Spurek

Jacek Tabor

Jagiellonian University, Faculty of Mathematics and Computer Science, Kraków, Poland

Abstract

Many crucial problems in deep learning and statistical inference are caused by a variational gap, i.e., a difference between model evidence (log-likelihood) and evidence lower bound (ELBO). In particular, in a classical VAE setting that involves training via an ELBO cost function, it is difficult to provide a robust comparison of the effects of training between models, since we do not know a log-likelihood of data (but only its lower bound). In this paper, to deal with this problem, we introduce a general and effective upper bound, which allows us to efficiently approximate the evidence of data. We provide extensive theoretical and experimental studies of our approach, including its comparison to the other state-of-the-art upper bounds, as well as its application as a tool for the evaluation of models that were trained on various lower bounds.

1 INTRODUCTION

Many important models in deep learning (Bayer et al., 2021; Burda et al., 2015; Kingma and Welling, 2013), reinforcement learning (Duo, 2021; Todorov, 2008; Toussaint and Storkey, 2006) and statistical inference (Gao et al., 2017; Khan et al., 2020) suffer from the existence of a variational gap¹, which means a difference between the evidence and its lower bound (which follows from Jensen’s inequality), i.e.:

$$\text{variational gap} = f(\mathbb{E}X) - \mathbb{E}f(X), \quad (1)$$

where X is a random variable and f is a concave function. A simple visualization of this effect can be delivered us-

¹In fact, in such a general context it is rather known as Jensen’s gap, but here and henceforth we call it consequently the variational gap.

ing the concave function $f(x) = -x^2$ to transform Gaussian random variable $X \sim \mathcal{N}(0, 1)$. Indeed, in this case $f(\mathbb{E}X) = 0 > -1 = \mathbb{E}f(X)$ (note that $-f(X) \sim \chi^2(1)$).

Table 1: Estimated size of various variational gap bounds for the evidence of data (lower is better), calculated for VAE, IWAE-5, and IWAE-10 models, previously trained on MNIST, SVHN, and CelebA datasets. All computations were averaged over 3 collections of 2^{16} latent samples, and over the test dataset.

DATASET MODEL		VARIATIONAL GAP BOUND (VG-B)							
		IS (OUR)	CUBO _{1.5}	CUBO ₂	EUBO	TVO ₂	TVO ₅	TVO ₁₀	TVO ₅₀
MNIST	VAE	0.04	1.59	2.83	6.68	3.34	1.34	0.67	0.13
	IWAE-5	0.004	1.32	2.24	12.38	6.19	2.48	1.24	0.25
	IWAE-10	0.01	1.54	2.55	18.03	9.01	3.61	1.80	0.36
SVHN	VAE	0.44	3.16	4.84	21.08	10.54	4.22	2.11	0.42
	IWAE-5	0.31	2.74	4.30	28.82	14.41	5.76	2.88	0.58
	IWAE-10	0.28	2.83	4.40	32.00	16.00	6.40	3.20	0.64
CelebA	VAE	1.12	3.25	4.92	45.45	22.73	9.09	4.55	0.91
	IWAE-5	1.23	3.40	5.14	85.41	42.70	17.08	8.54	1.71
	IWAE-10	0.72	3.22	4.86	75.46	37.73	15.09	7.55	1.51

In machine learning literature, where typically $f = \log$, various approximations of the true evidence $f(\mathbb{E}X)$ were proposed (see Section 2 and references therein), but are often difficult to efficiently use in a deep neural network architecture. One of the reasons for this is that we train such models on mini-batches, and therefore the standard assumption is that the cost function factorizes as the sum over the input data set. In the other words, deep networks are designed to maximize (over network parameters) $\mathbb{E}f(X)$, rather than $f(\mathbb{E}X)$, which admits respective sample mean (unbiased) estimator.

However, there often naturally appear situations where maximization of $f(\mathbb{E}X)$ is an actual goal. Probably the most important such case is the variational autoencoder (VAE) (Kingma and Welling, 2013; Rezende et al., 2014), which is one of the most popular autoencoder-based generative models. Precisely, VAE uses an encoder network $q(z|x)$, which reduces the dimension of data and pro-

duces their latent codes (forced to follow approximately a given latent prior distribution $p(z)$), and a decoder network $p(x|z)$ that transforms the latent codes back to the data space. Both networks are jointly trained to maximize a variational lower bound for the log-likelihood (the evidence) of data:

$$\text{evidence} = \log \mathbb{E}_{z \sim q(\cdot|x)} \frac{p(x|z)p(z)}{q(z|x)}, \quad (2)$$

which is known as evidence lower bound or (briefly) ELBO:

$$\text{ELBO} = \mathbb{E}_{z \sim q(\cdot|x)} \log \frac{p(x|z)p(z)}{q(z|x)}. \quad (3)$$

The use of ELBO, instead of a direct value of log-likelihood, seems to be a fundamental problem in VAE. In practice, such optimization can lead to learning sub-optimal parameters (Burda et al., 2015), when we mean that our final goal is an approximation of data distribution. Hence, estimating, bounding, and reducing a difference between the evidence and ELBO, i.e., the variational gap, became important issues investigated so far by many authors from the machine learning community (see Section 2 for a respective overview of the literature). Moreover, such problems are (in a general context) strictly related to those concerning (reversed) Jensen’s inequality, which is a subject of studies in various fields of pure and applied mathematics, including statistical inference (Brnetic et al., 2015; Horváth, 2021; Saeed et al., 2022; Dragomir, 2013; Jebara and Pentland, 2001; Nielsen, 2010), reinforcement learning (Dayan and Hinton, 1997; Williams et al., 2017), or even biological studies (Ruel and Ayres, 1999).

Our work provides a comprehensive theoretical approach regarding the variational gap. Most of all, we construct novel upper bounds for $f(\mathbb{E}X)$ (and hence also for the size of variational gap), which are given as expected values of some random variables that depend on X , and then, inspired by (Burda et al., 2015), combine them with the technique of importance sampling, to derive tight bounds for the exact value of $f(\mathbb{E}X)$. Additionally, as an application in the field of deep learning, we use these general results for precisely estimating the log-likelihood of data for generative models, which are designed to learn only some lower bounds. Consequently, we obtain a method that allows comparing the effectiveness of the training process, which we examine using a few different experimental settings involving VAE-like architectures, i.e., the classical (Gaussian) VAE and two variants of the importance-weighted (Gaussian) autoencoder (IWAE) (Burda et al., 2015), all trained on MNIST, SVHN, and CelebA datasets.

Our contribution can be summarized as follows:

- we introduce novel upper bounds for the variational gap, based on the importance sampling technique, which allows us to calculate a tight approximation of $f(\mathbb{E}X)$ for any concave function f , and provide their formal (mathematical) justification,

- we apply these results for $f = \log$, to provide precise estimates for the evidence (log-likelihood) of data, which we treat as a practical method for validating the effects of training in generative models involving lower bound optimization, and remark both benefits and limitations of our approach,
- we perform experiments that confirm our (theoretical) claims and prove the superiority of the proposed bound estimations for the variational gap (see Table 1), in comparison to the other state-of-the-art techniques (we recall them briefly in Section 2, see also Appendix B).

2 RELATED WORK

One of the most popular generative, autoencoder-based models is variational autoencoder (VAE) (Kingma and Welling, 2013; Rezende et al., 2014), which aims to maximize the log-likelihood of data. However, since this likelihood is intractable, the main idea which stays behind estimating it is to calculate and optimize evidence lower bound (ELBO) instead, which results in appearing the variational gap.

One of the problems with the variational gap is its behavior since it can be tiny or tremendous, depending on the model distribution. The importance of taking care of gaps and their possible offending effects were mentioned in (Bayer et al., 2021). There are several techniques to deal with the variational gap, such as its direct estimation (Abramovich and Persson, 2016) or finding an upper or lower bound, to know how much we can lose. For example, Nowozin (2018) and Maddison et al. (2017) create lower bounds using big-O notation. Bounds for the variational gap were also derived in (Khan et al., 2020) and later used for deriving new inequalities (e.g., bounding Csiszár divergence or converse of the Hölder inequality), as well as in (Grosse et al., 2015), where bidirectional Monte Carlo simulations were involved.

To our best knowledge, approaches the most related to the results of the present paper were introduced in (Ding et al., 2017), where the authors proposed the χ upper bound (CUBO) for the evidence of the data, which was expressed in terms of the χ -divergence, and in (Ji and Shen, 2019), where the evidence upper bound (EUBO), involving the Kulback-Leibler divergence, was defined and explored. Following Masrani et al. (2019), it should be noticed that both EUBO and its generalization, i.e., an upper bound variant of the thermodynamic variational objective (TVO), can be derived as the right Riemann sum approximation of the log-likelihood of data expressed via thermodynamic integration method. Thus, we henceforth concentrate on the mentioned bounds when validating our approach, and (for readers’ convenience) in Appendix B we provide basic information concerning them. Nevertheless, we would like to emphasize that, unlike the others, our approach possesses a

more general theoretical background that does not exclude it from other potential applications².

On the other hand, we can find a broad usage of variational inference not only in the context of generative models (to which we limit our considerations). For example, Toussaint and Storkey (2006) use it to present the expectation-maximization algorithm (EM) for computing optimal policies by solving Markov decision processes. Furthermore, Botvinick and Toussaint (2012) say that people use probabilistic inference when they plan, Levine (2018) uses variational inference to derive a new view of reinforcement learning, where decision-making is an inference problem represented in a type of graphical model, and Duo (2021) proposes a policy optimization algorithm in the context of variational inference.

3 THEORETICAL STUDY

In this section, we present the main theoretical results of the paper. In the first subsection, in Theorem 1 we derive a general condition which, under the assumption of concavity, enables us to find an upper bound for the evidence. In the second subsection, inspired by (Burda et al., 2015), we describe the importance sampling technique that allows us to decrease the size of the gap by replacing the random variable with the mean of its independent copies. The third subsection contains the crucial results of the paper (Theorems 3 and 4), in which we provide a large class of upper bounds for the evidence. By respectively, choosing a suitable parameter C in Theorem 4, this allows us to obtain for the map $f = \log$ much tighter estimations than those given directly by Corollary 3 (see Theorem 5 and the experimental results supplied in Section 4). In the last subsection, we try to answer the question of where the evidence lies in the interval given by the lower bound and the upper bound. In particular, we show in Theorem 7 that for an important class of log-normal distributions, it is located in the middle of that interval.

3.1 Variational Gap

Let X be a random variable. By the classical Jensen inequality, for every concave function f we have $f(\mathbb{E}X) \geq \mathbb{E}f(X)$. The aim of this subsection is to obtain (under the above general assumptions) an upper bound for $f(\mathbb{E}X)$, which requires computing only the expected value of some random variable that depends on X . (Note that such an additional supposition ensures the additivity of the bound when applied for solving optimization problems by machine learning algorithms.) This easily follows from the following theorem, which can also be found in (Dragomir, 2013) in a more general form. Nevertheless, for completeness, we include the proof in Appendix A.

²We treat this as motivation for our future work.

Theorem 1. *Let f be a smooth concave function. Then*

$$f(\mathbb{E}X) \leq \mathbb{E}(f(X) + (Y - X)f'(X)), \quad (4)$$

where X and Y are two independent random variables with the same distribution.

Theorem 1 and Jensen’s inequality imply that the value of $f(\mathbb{E}X)$ is enclosed in the interval

$$[\mathbb{E}f(X), \mathbb{E}f(X) + \mathbb{E}((Y - X)f'(X))]. \quad (5)$$

Thus, the size of the variational gap is bounded from above by the length of the interval (5), i.e., by the value of $\mathbb{E}[(Y - X)f'(X)]$.

Proceeding to the most important case of $f = \log$, we directly obtain the following corollary.

Corollary 1. *We have*

$$\mathbb{E} \log X \leq \log \mathbb{E}X \leq \mathbb{E} \log X + \mathbb{E} \frac{Y}{X} - 1, \quad (6)$$

where X and Y are independent random variables with the same distribution.

Remark 1. Consider, for example³, the random variable $X \sim \text{Gamma}(a, \theta)$ and the concave function $f = \log$. Then $1/X \sim \text{Inv-Gamma}(a, 1/\theta)$. Therefore, assuming $a > 1$, we can calculate that

$$\mathbb{E} \frac{Y}{X} - 1 = \mathbb{E}Y \mathbb{E} \frac{1}{X} - 1 = \frac{a\theta}{\theta(a-1)} - 1 = \frac{1}{a-1}, \quad (7)$$

which means that in this case, we cannot bound the variational gap effectively when a approaches 1. Taking into account the properties of the Gamma distribution, this phenomenon means that when we sample from X , we obtain arbitrarily small (positive) values more and more likely. In practice, such a situation may appear in a VAE setting, especially when we are dealing with outliers, which is a direct motivation for the improvement introduced in the next subsection.

3.2 Reducing Variational Gap

In this subsection, addressing the problem outlined in Remark 1, we describe how to apply the importance sampling technique to obtain a tighter (additive) approximation of $f(\mathbb{E}X)$. It comes down to using (instead of X) the random variable

$$\bar{X}_k = \frac{1}{k}(X_1 + \dots + X_k), \quad (8)$$

representing the mean of a k -sample from X (consisting of k independent copies of X). Clearly, $\mathbb{E}\bar{X}_k = \mathbb{E}X$, which implies that $f(\mathbb{E}\bar{X}_k) = f(\mathbb{E}X)$. The following theorem is (in fact) a part of Theorem 1 from (Burda et al., 2015), restated in a general setting. However, for completeness, we include novel and independent proof.

³Here and henceforth, $\text{Gamma}(a, \theta)$ and $\text{Inv-Gamma}(a, \theta)$ denote Gamma and Inverse-Gamma distributions with the shape parameter $a > 0$ and the scale parameter $\theta > 0$, respectively.

Theorem 2. Let X be a random variable and f be a continuous concave function. Then for every $k > 0$ we have

$$\mathbb{E}f(\bar{X}_k) \leq \mathbb{E}f(\bar{X}_{k+1}). \quad (9)$$

Moreover, $f(\bar{X}_k)$ converges to $f(\mathbb{E}X)$ almost surely, and

$$\lim_{k \rightarrow \infty} \mathbb{E}f(\bar{X}_k) = f(\mathbb{E}X), \quad (10)$$

provided that the support of X is contained in some compact interval lying in the domain of f .

Proof. Let (Ω, μ) be a probabilistic space and $X: \Omega \rightarrow \mathbb{R}$ be a random variable. By the concavity of f we directly conclude that

$$\begin{aligned} f\left(\frac{1}{k+1} \sum_{i=1}^{k+1} x_i\right) &= f\left(\frac{1}{k+1} \sum_{i=1}^{k+1} \frac{1}{k} \sum_{j=1, j \neq i}^{k+1} x_j\right) \\ &\geq \frac{1}{k+1} \sum_{i=1}^{k+1} f\left(\frac{1}{k} \sum_{j=1, j \neq i}^{k+1} x_j\right). \end{aligned} \quad (11)$$

Then by monotonicity and linearity of the expected value, we obtain

$$\mathbb{E}f(\bar{X}_{k+1}) \geq \frac{1}{k+1} \sum_{i=1}^{k+1} \mathbb{E}f\left(\frac{1}{k} \sum_{j=1, j \neq i}^{k+1} X_j\right) = \mathbb{E}f(\bar{X}_k), \quad (12)$$

where X_1, \dots, X_{k+1} are independent copies of X . This gives the first assertion of the theorem.

Now consider the random variable \bar{X}_k . From the strong law of large numbers, it follows that \bar{X}_k converges to $\mathbb{E}X$ almost surely. Since f is continuous, this is also the case for $f(\bar{X}_k)$ and $f(\mathbb{E}X)$. Hence, we conclude that $\mathbb{E}f(\bar{X}_k) \rightarrow f(\mathbb{E}X)$ as $k \rightarrow \infty$, whenever the support of X is contained in some compact interval lying in the domain of f , which completes the proof. \square

Assuming bounded support for X in Theorem 2 we followed (Burda et al., 2015), where the respective theory behind the use of the importance sampling technique is based on this assumption and illustrates the underlying case in a simplified setting. Although, in some cases, the theorem is tending to “survive” without this restriction (see, e.g, Remark 2), the proof of a more general version would, however, cause some technical difficulties, resulting in reducing the clarity of the presentation.

Applying (5) to the random variable \bar{X}_n , we conclude that

$$f(\mathbb{E}X) \in [\mathbb{E}f(\bar{X}_k), \mathbb{E}f(\bar{X}_k) + \mathbb{E}((\bar{Y}_k - \bar{X}_k)f'(\bar{X}_k))], \quad (13)$$

where X and Y are independent random variables with the same distribution, and X_1, \dots, X_k and Y_1, \dots, Y_k are independent copies of X and Y , respectively. By Theorem 2, the left end of the above interval converges to $f(\mathbb{E}X)$. Hence, a natural question arises, whether the same happens for the right end. In the following corollary, we show that this is the case.

Corollary 2. Let X be a random variable and f be a concave function. Assume also that $(m - x)f'(x)$ is convex in the support of X for arbitrary constant m that belongs to the support of X . Then the width of the interval given in (13), i.e., $\mathbb{E}((\bar{Y}_k - \bar{X}_k)f'(\bar{X}_k))$, is a decreasing sequence with k . Moreover, if the support of X is contained in the closed bounded interval lying in the domain of f' , then the limit is 0.

Proof. It is enough to apply Theorem 2 for the concave function: $(x - \mathbb{E}X)f'(x)$. \square

Now let us get back to the case of $f = \log$, where we directly obtain the following corollary.

Corollary 3. We have

$$\mathbb{E} \log \bar{X}_k \leq \log \mathbb{E}X \leq \mathbb{E} \log \bar{X}_k + \mathbb{E} \frac{\bar{Y}_k}{\bar{X}_k} - 1. \quad (14)$$

Moreover, $\mathbb{E} \frac{\bar{Y}_k}{\bar{X}_k}$ is a decreasing sequence with k , which converges to 1, provided that the support of X lies in some compact interval contained in $(0, \infty)$.

Remark 2. As we have already outlined, in some cases we can use Theorem 2 (and hence Corollaries 2 and 3) to make the size of a variational gap arbitrarily small, even if we cannot respectively bound values of X almost surely. Indeed, if we return to the example from Remark 1, we easily see that $\bar{X}_k \sim \text{Gamma}(ka, \theta/k)$ and $1/(\bar{X}_k) \sim \text{Inv-Gamma}(ka, k/\theta)$. Therefore, assuming $ka > 1$, we can calculate that

$$\mathbb{E} \frac{\bar{Y}_k}{\bar{X}_k} - 1 = \frac{a\theta k}{\theta(ka-1)} - 1 = \frac{1}{ka-1} \xrightarrow{k \rightarrow \infty} 0. \quad (15)$$

However, note that even though this means that we can find k large enough to decrease a variational gap sufficiently, we also see that when the value of a approaches 0, we may be forced to wait for such an effect quite a long while increasing the value of k . This is a direct motivation for the improvement introduced in the next subsection.

3.3 Improved Bounds for Variational Gap

In this subsection, we provide another technique, which is crucial in the estimation of the size of the variational gap and addresses the problem described in Remark 2. Although we start with a general proposition, the idea of which lies in generalizing the bounds obtained by the concavity, we eventually fix our attention on the case $f = \log$.

Proposition 1. Assume that f , g , and h are arbitrary functions such that

$$f(a) \leq g(x) + ah(x) \text{ for every } a \text{ and } x. \quad (16)$$

Then

$$f(\mathbb{E}X) \leq \mathbb{E}(g(X) + Yh(X)), \quad (17)$$

where X and Y are two independent random variables with the same distribution.

Proof. The proof follows the same lines as the proof of Theorem 1 (see Appendix A). We consider the probabilistic space (Ω, μ) and independent random variables $X, Y: \Omega \rightarrow \mathbb{R}$ with the same distribution. We use the notation

$$m = \mathbb{E}X = \int_{\Omega} X d\mu. \quad (18)$$

Clearly $\mathbb{E}X = \mathbb{E}Y$. Observe that, by the assumptions, for every $\omega \in \Omega$ we have

$$f(m) \leq g(X(\omega)) + mh(X(\omega)). \quad (19)$$

Integrating the above formula over all $\omega \in \Omega$, we get

$$\begin{aligned} f(\mathbb{E}X) &= \int_{\Omega} f(m) d\mu \leq \int_{\Omega} g(X(\omega)) + mh(X(\omega)) d\mu \\ &= \mathbb{E}g(X) + \mathbb{E}Y\mathbb{E}h(X) = \mathbb{E}(g(X) + Yh(X)), \end{aligned} \quad (20)$$

which ends the proof. \square

Now we focus our attention on the case when $f = \log$. We prove that given an arbitrary function g , we can easily compute the optimal h .

Lemma 1. *Let $g: (0, \infty) \rightarrow \mathbb{R}$ be an arbitrary function. Then*

$$\log a \leq g(x) + a \exp(-g(x) - 1) \text{ for all } a, x > 0. \quad (21)$$

Moreover, for any function $h: (0, \infty) \rightarrow \mathbb{R}$ satisfying

$$\log a \leq g(x) + ah(x), \quad (22)$$

we have

$$h(x) \geq \exp(-g(x) - 1). \quad (23)$$

Proof. Consider an arbitrary function h . Let $x > 0$ be fixed. We are going to find an equivalent condition for h so that

$$\log a \leq g(x) + ah(x) \text{ for all } a > 0. \quad (24)$$

Note that verifying (24) is equivalent to checking whether

$$h(x) \geq \frac{\log a - g(x)}{a} \text{ for all } a > 0, \quad (25)$$

or, equivalently,

$$h(x) \geq \sup_{a>0} \frac{\log a - g(x)}{a}. \quad (26)$$

One can easily check that if $w: (0, \infty) \rightarrow \mathbb{R}$ is a function defined as

$$w(a) = \frac{\log a - g(x)}{a}, \quad (27)$$

then $w'(a) = \frac{1 - (\log a - g(x))}{a^2}$. Consequently, w reaches the maximal value at $a_x = \exp(1 + g(x))$. Thus, the equivalent condition for h to satisfy (24) is

$$h(x) \geq w(a_x) = \exp(-1 - g(x)), \quad (28)$$

which proves all assertions. \square

As a direct consequence of Proposition 1 and Lemma 1 we obtain the following theorem.

Theorem 3. *Let g be an arbitrary function. Then*

$$\log \mathbb{E}X \leq \mathbb{E}(g(X) + Y \exp(-g(X) - 1)), \quad (29)$$

where X and Y are two positive independent random variables with the same distribution.

Now consider a one-parameter family of functions

$$g_C(x) = \log x - 1 + C \quad (C \in \mathbb{R}). \quad (30)$$

Then by applying any function g_C as g in Proposition 1, we obtain the following theorem.

Theorem 4. *Let C be arbitrarily chosen. Then*

$$\mathbb{E} \log X \leq \log \mathbb{E}X \leq \mathbb{E} \log X - 1 + C + \exp(-C) \mathbb{E} \frac{Y}{X}, \quad (31)$$

where X and Y are two independent positive random variables with the same distribution.

Observe that increasing C decreases the last component of the right-hand side formula in (31). Hence, the optimal value of C (i.e., minimizing the upper bound for $\log \mathbb{E}X$) can be obtained as the one that minimizes the function $W(C) = C + \exp(-C) \mathbb{E} \frac{Y}{X}$. This easily leads us to the following corollary.

Corollary 4. *Under the assumptions of Theorem 4, the optimal value of C is $C = \log \mathbb{E} \frac{Y}{X}$, which gives the following estimates:*

$$\mathbb{E} \log X \leq \log \mathbb{E}X \leq \mathbb{E} \log X + \log \mathbb{E} \frac{Y}{X}. \quad (32)$$

Note that although the upper bound for $\log \mathbb{E}X$ provided by (32) does not have an additive form, it tells us about the most optimal estimation we can obtain. Moreover, in this case, we can also apply the importance sampling technique resulting in reducing the variational gap, which is an immediate consequence of Corollaries 3 and 4.

Theorem 5. *We have*

$$\mathbb{E} \log \bar{X}_k \leq \log \mathbb{E}X \leq \mathbb{E} \log \bar{X}_k + \log \mathbb{E} \frac{\bar{Y}_k}{\bar{X}_k}. \quad (33)$$

Moreover, $\log \mathbb{E} \frac{\bar{Y}_k}{\bar{X}_k}$ is a decreasing sequence with k , which converges to 0, provided that the support of X lies in some compact interval contained in $(0, \infty)$.

Remark 3. Continuing the example involving the Gamma distributed random variable X (see Remarks 1 and 2), which goes beyond the assumptions of Theorem 5, when $ka > 1$ we have

$$\log \mathbb{E} \frac{\bar{Y}_k}{\bar{X}_k} = \log \frac{ka}{ka-1} \xrightarrow{k \rightarrow \infty} 0. \quad (34)$$

This means that in this case, the last conclusion of Theorem 5 “survives”, too.

3.4 Quality of Estimations

We have already proved (see (5)) that for any concave function f we have

$$f(\mathbb{E}X) \in [\mathbb{E}f(X), \mathbb{E}(f(X) + (Y - X)f'(X))], \quad (35)$$

where X and Y are independent random variables with the same distribution. In this section, we are going to show that the optimal choice for an approximation of $f(\mathbb{E}X)$ is the middle of the above interval.

Let us start with the following general result, which relates to a special case of the delta method (Bickel and Doksum, 2015, Section 5.3.1). For completeness, we include the proof in Appendix A.

Theorem 6. *Assume that f is a smooth function. Let X and Y be independent random variables with the same distribution, which attain only values ε -close to $\mathbb{E}X$, where $\varepsilon > 0$ is small. Then*

$$f(\mathbb{E}X) = Ef(X) + \frac{1}{2}\mathbb{E}((Y - X)f'(X)) + o(\varepsilon^2). \quad (36)$$

Even though the assumptions of Theorem 6 are somewhat unrealistic, it was formulated to improve readers' intuition. Note that the importance sampling technique leads to random variables being more and more concentrated around their means, which we utilize in our experiments using sufficiently large samples.

Now we proceed to the case when $f = \log$ and we are going to apply the bounds for \bar{X}_k . Clearly, by the central limit theorem, for large k the distribution of \bar{X}_k can be considered Gaussian. However, this approximation is not satisfactory from our point of view, since we are limited to the class of positive random variables (which are proper arguments for the log function). Based on the results of (Mouri, 2013), it is known that typically the distribution of the mean of independent positive random variables can be better approximated by the log-normal distribution \mathcal{LN} . In the other words, we can write $\bar{X}_k \approx \mathcal{LN}(m, \sigma)$. In the following theorem, we prove that for a log-normal random variable X , the value of $\log \mathbb{E}X$ lies exactly in the middle of the interval given by Corollary 4.

Theorem 7. *Let X and Y be independent random variables with the same log-normal distribution. Then*

$$\log \mathbb{E}X = \mathbb{E} \log X + \frac{1}{2} \log \mathbb{E} \frac{Y}{X}. \quad (37)$$

Proof. Let $X \sim \mathcal{LN}(m, \sigma)$, which means that $\log X \sim \mathcal{N}(m, \sigma^2)$. Then $\mathbb{E}X = \exp(m + \sigma^2/2)$ and, consequently,

$$\log \mathbb{E}X = m + \frac{\sigma^2}{2}, \quad \mathbb{E} \log X = m. \quad (38)$$

Moreover, $1/X \sim \mathcal{LN}(-m, \sigma)$ and hence

$$\log \mathbb{E} \frac{Y}{X} = \log \mathbb{E}Y + \log \mathbb{E} \frac{1}{X} = m + \frac{\sigma^2}{2} - m + \frac{\sigma^2}{2} = \sigma^2. \quad (39)$$

By applying (38) and (39) in (37), we obtain the conclusion. \square

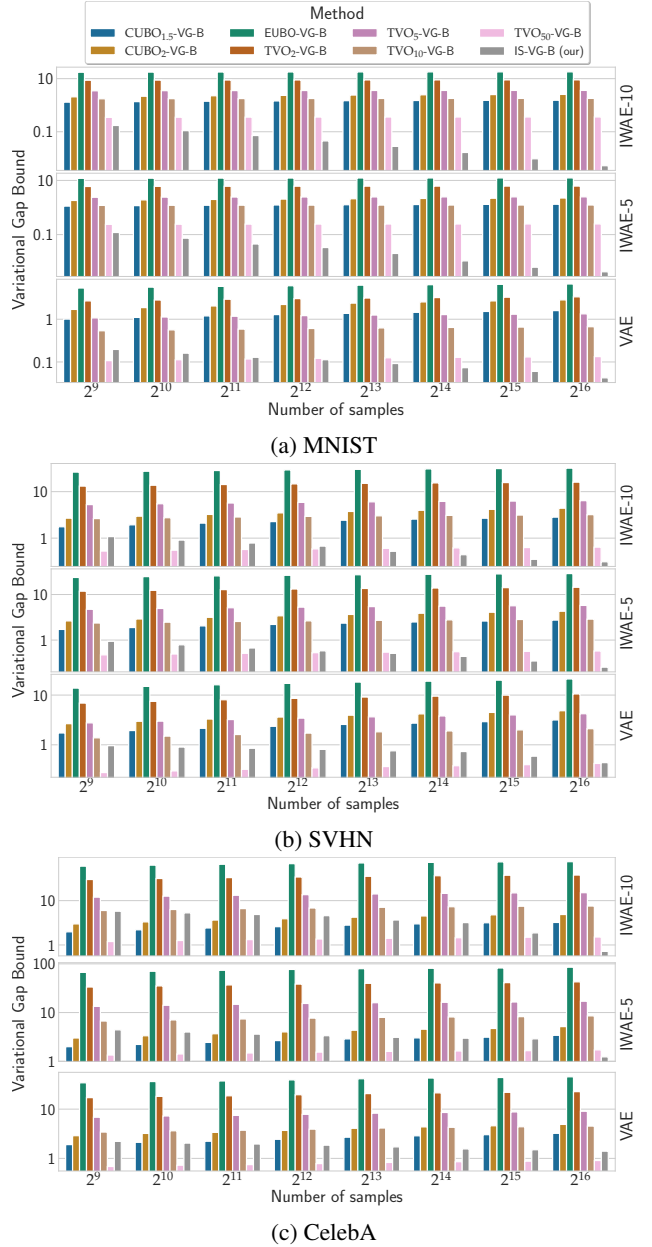


Figure 1: Estimated size of various variational gap bounds for the evidence of data (lower is better), calculated for VAE, IWAE-5, and IWAE-10 models, previously trained on MNIST, SVHN, and CelebA datasets, vs. the number of latent samples. All computations were averaged over 3 collections of samples, and over the test dataset.

4 EXPERIMENTS

In this section, we consider variational generative models (like VAE). Let us first establish some standard notation.

By $p(x)$ we denote the distribution induced by the model on the input space $\mathcal{X} = \mathbb{R}^N$. By $p(z)$ we denote the prior distribution on the latent space \mathcal{Z} , while $q(z|x)$ denotes the variational encoder and $p(x|z)$ denotes the variational decoder.

Now, given a point $x \in \mathbb{R}^N$, its model log-likelihood (evidence) is expressed as follows:

$$\log p(x) = \log \mathbb{E}_{z \sim q(\cdot|x)} \frac{p(x|z)p(z)}{q(z|x)}. \quad (40)$$

To simplify the notation, we put

$$R(x, z) = \frac{p(x|z)p(z)}{q(z|x)}. \quad (41)$$

In the classical VAE, we maximize

$$\text{ELBO} = \mathbb{E}_{z \sim q(\cdot|x)} \log R(x, z), \quad (42)$$

which is the lower bound for the log-likelihood.

The idea behind the IWAE model (Burda et al., 2015) is to obtain an (asymptotically optimal) approximation of the evidence by maximizing

$$\text{IW-ELBO}_k = \mathbb{E}_{z_i \sim q(\cdot|x)} \log \frac{1}{k} \sum_{i=1}^k R(x, z_i), \quad (43)$$

which is a closer (than ELBO) lower bound for the log-likelihood.

To obtain upper bounds for the evidence, it is enough to apply Theorems 4 and 5 for $X = \frac{1}{k} \sum_{i=1}^k X_i$, where $X_i = R(x, z_i)$ and $Y_i = R(x, \tilde{z}_i)$, and all z_i and \tilde{z}_i are independently sampled from $q(\cdot|x)$. Then, we conclude that the size of the respective variational gap is bounded from above by the following value, which we will call *importance sampling variational gap bound (IS-VG-B)*:

$$\text{IS-VG-B} = C - 1 + \exp(-C) \mathbb{E}_{z_i, \tilde{z}_i \sim q(\cdot|x)} \frac{\sum_{i=1}^k R(x, \tilde{z}_i)}{\sum_{i=1}^k R(x, z_i)}, \quad (44)$$

where C may be chosen arbitrarily, with the optimum equal to

$$C^{\text{opt}} = \log \mathbb{E}_{z_i, \tilde{z}_i \sim q(\cdot|x)} \frac{\sum_{i=1}^k R(x, \tilde{z}_i)}{\sum_{i=1}^k R(x, z_i)}. \quad (45)$$

Starting with a case study for simple synthetic one-dimensional data generated from the Laplace distribution, in the following few paragraphs we provide and discuss the results of the experiments, in which we compare our approach to those presented in Dieng et al. (2017); Ji and Shen (2019); Masrani et al. (2019). Mainly, we apply all considered estimation techniques for the variational gap to selected Gaussian autoencoders (i.e., classical VAE and two different IWAE models, on MNIST, SVHN, and CelebA datasets), previously learned using their own objectives and VAE experimental setup.

Case Study on Synthetic Data Suppose that our data are drawn from a known distribution $p(x)$. Then obviously, the

true evidence of data⁴ is expressed as $\int p(x) \log p(x) dx$. Suppose that by training the VAE model, we construct the approximation of $p(x)$ in the class of distributions $p_\theta(x)$, where θ denotes the weights of the neural networks. Then the (model) evidence is given as $\int p(x) \log p_\theta(x) dx$. Consequently,

$$\text{true evidence} = \text{evidence} + D_{\text{KL}}(p(x) \| p_\theta(x)), \quad (46)$$

where D_{KL} denotes the Kullback-Leibler divergence. Obviously, if (which is a common case) $p(x)$ does not belong to the family of distributions $(p_\theta(x))_{\theta \in \Theta}$ (more precisely, $D_{\text{KL}}(p(x) \| p_\theta(x)) > 0$), then the evidence is less than the true-evidence. Since we used the VAE model, we also have ELBO which is less than the evidence. Finally, we obtain upper and lower bounds for the evidence of data, which become tighter with the increasing number of samples from the latent distribution $q_\theta(\cdot|x)$ (here x represents any data point drawn from $p(x)$).

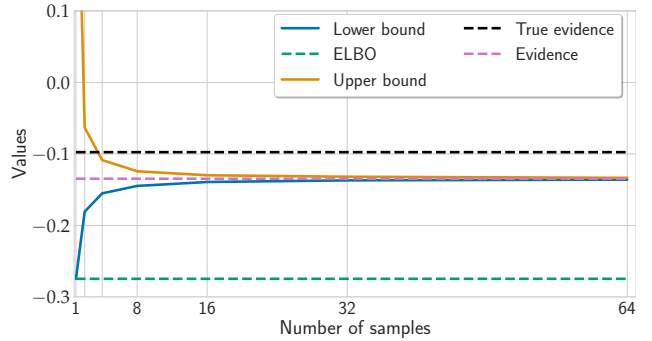


Figure 2: Behavior of lower and the upper bounds for the evidence vs. the number of samples from the latent.

We illustrate the above reasoning in the case of simple synthetic one-dimensional data generated from the Laplace distribution $\text{Laplace}(0, 0.2)$. We trained the VAE model (for architecture details, see Appendix C) with one-dimensional latent space. The experimental results (see Figure 2) are consistent with the above-mentioned theoretical discussion. The true evidence of the data coming from the Laplace distribution equals -0.097 . On the other hand, ELBO (calculated as a value of the cost function of VAE) is a strong lower bound for the (model) evidence which, in turn, is less than the true evidence. Moreover, the IW-ELBO lower bound and our upper bound converge to the evidence of data, thus providing its tight estimate in the interval $[-0.137, -0.132]$, with the ends calculated using 64 latent samples.

⁴In this paragraph, unlike before, the notion “evidence” refers to the expected value of the log-likelihood of data. In practice, this corresponds to taking the average over the whole dataset, which we use anyway in our experiments. Moreover, to avoid possible misunderstanding, we draw readers’ attention to the fact that in our paper, the evidence of data is based on the model distribution (see (2) and (40)). This paragraph is the only place where we also mention the true evidence.

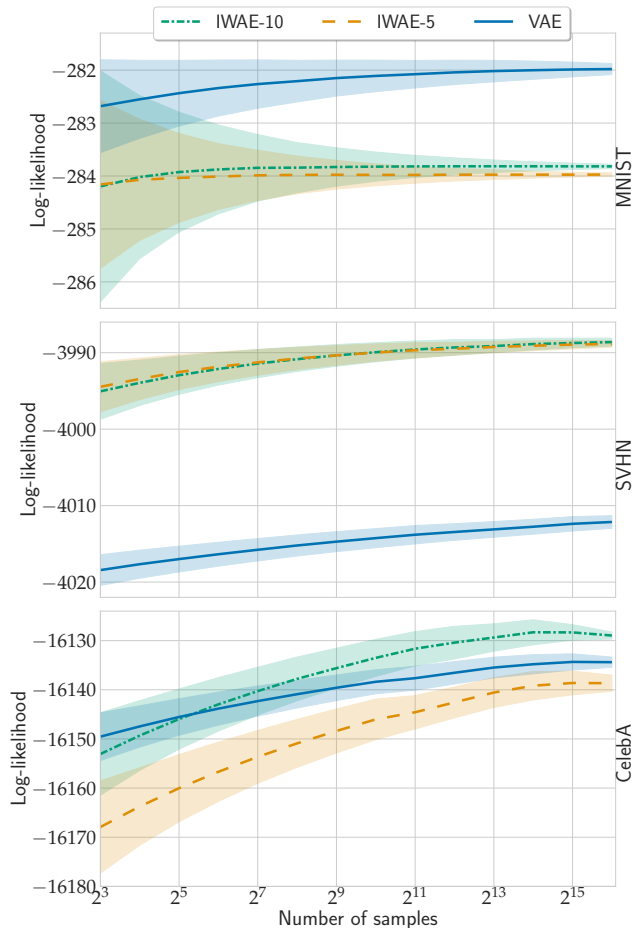


Figure 3: Behavior of lower and upper bounds for the evidence of data, calculated by our method for VAE, IWAE-5, and IWAE-10 models, previously trained on MNIST, SVHN, and CelebA datasets, vs. the number of latent samples. All computations were averaged over 3 collections of samples, and over the test dataset.

Experiments for VAE and IWAE Models In our further experiments, we validate our importance sampling variational gap bound (given by (44) and (45)) against the other state-of-the-art bounds, denoted by CUBO-VG-B, EUBO-VG-B, and TVO-VG-B, depending on the evidence upper bound used. All comparisons are made for VAE, IWAE-5, and IWAE-10 models, for which the respective bounds are calculated using (in total) the same number of latent samples z_i . Each model was previously trained on three classical datasets, i.e., MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), and CelebA (Liu et al., 2015), using its own objective and VAE experimental setup (see Appendix C for the details). The code for all experiments is available on the GitHub repository https://github.com/gmum/Bounding_Evidence_Estimating_LL.

The obtained results are presented in Table 1 and Figures 1

and 3. Namely, from Table 1 we learn that the proposed method for bounding the evidence of data is superior in comparison to all competitors, excluding TVO₅₀-VG-B for VAE trained on SVHN and CelebA datasets, as far as we make calculations using the largest considered number of latent samples (i.e., 2^{16}). However, estimating bounds for such sets of data is much more demanding because of complicated encoded data distributions. We believe increasing the sample size up to 2^{17} or 2^{18} (although quite time-consuming) would further weaken the impact of outliers and allow our method to definitely win. Additionally, a closer inspection of Figure 1 shows that our approach is the only one guaranteeing to decrease a computed variational gap bound with the increasing number of samples, which agrees with the conclusion of Theorem 5. This is due to applying the importance sampling technique, which (to our best knowledge) is not the case for the other methods (note that for them, the biased estimation may result in bounds growing with the sample size).

More complete results of our approach, including the dependence of the obtained lower and upper bounds on the number of latent samples used in the computations, are presented in Figure 3. In practice, they allow us (for any given dataset) to compare the effects of training between all considered models. Note that even though the IWAE models are learned using more latent samples, they do not always deliver better results, i.e., greater values of the log-likelihood of data (see also Rainforth et al., 2018). For example, VAE trained on the MNIST dataset delivers the best (the greatest) log-likelihood estimates⁵.

To prevent possible readers' concerns, we would like to explain that the gaps calculated on SVHN and CelebA datasets for different numbers of latent samples do not overlap because we use estimators (and not strict values, provided in (44) and (45)) suffering from the presence of outliers. To confirm this, we removed outliers and reran all computations. The obtained results are presented in Figure 4. Note that we observe no such phenomenon anymore, which supports our assertion.

Dependence of Variational Gap Bound on C We conducted additional experiments to explore the dependence of the proposed variational gap bound on the choice of constant C in (44). In Figure 5 we present the results obtained for VAE, IWAE-5, and IWAE-10 models, previously trained on MNIST, SVHN, and CelebA datasets. Note that in each case, C^{opt} (see (45)) is a value for which IS-VG-B reaches a minimum.

⁵We would like to emphasize that, although our method is also able to evaluate the effects of even suboptimal training, we work on properly learned models. For example, the FID score for our Gaussian VAE equals 28.73 and is better than, e.g., 40.47 provided in (Knop et al., 2020), where a comparable experimental setup was used.

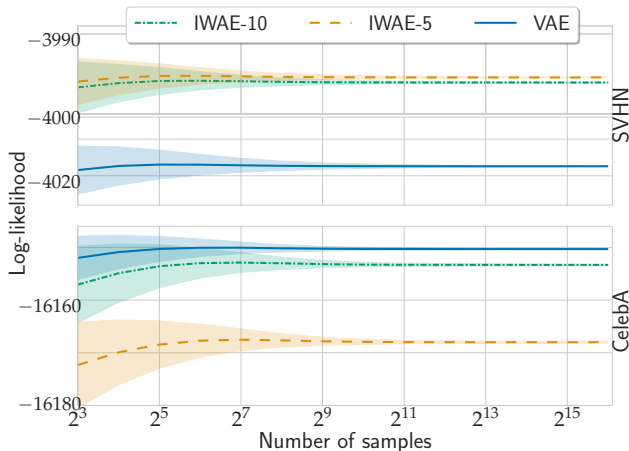


Figure 4: Behavior of lower and upper bounds for the evidence of data, calculated without 1% outliers by our method for VAE, IWAE-5, and IWAE-10 models, previously trained on SVHN and CelebA datasets, vs. the number of latent samples. All computations were averaged over 3 collections of samples, and over the test dataset.

More details on the experimental results can be found in Appendix C.

5 CONCLUSION

In this paper, we proposed novel upper bounds for the variational gap (see (4) for a basic version and (17) for an improved version), and the use of the importance sampling technique to tighten them, which allowed us to calculate a precise estimation of $f(\mathbb{E}X)$ for any concave function f . We focused particular attention on the most important case when $f = \log$ (see (6) for a basic version as well as (29) and (31) for improved versions), which led us to a practical method for validating the effects of training in generative models involving lower bound optimization. We conducted experiments that proved the superiority of our approach in comparison to the other state-of-the-art techniques.

Limitation In this contribution we did not consider possible applications of our approach outside VAEs, which might be a natural direction for future work. Moreover, our experience shows that the introduced novel bounds have rather limited utility in the training process. Finally, calculating bounds in our experiments is based on estimators rather than strict values, which makes it sensitive to the presence of outliers, and results in nonrigorous bounds (although we note the discussed convergence results). However, we would like to emphasize that our proposed variational gap bounds are strict bounds as far as we limit ourselves to the theoretical approach (where we deal with random variables and their means).

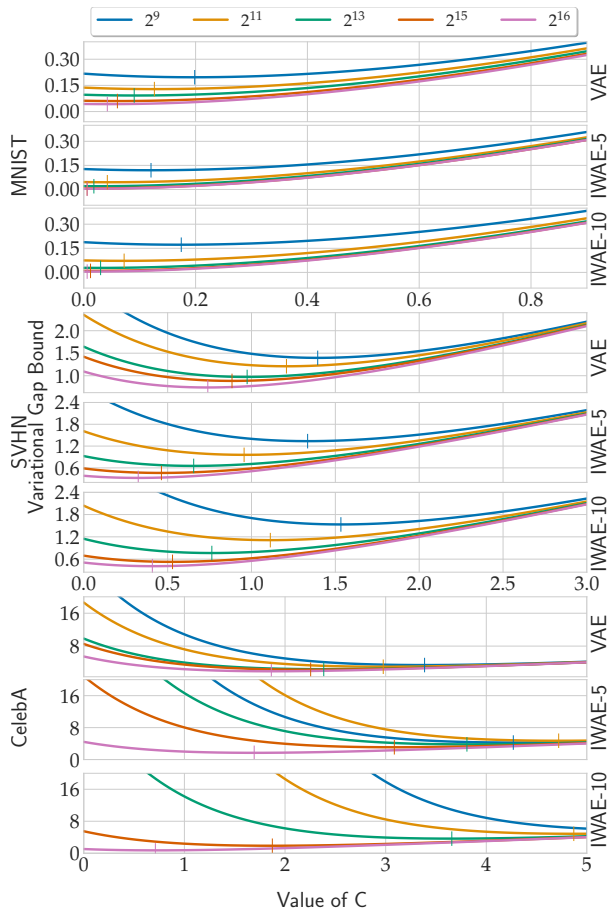


Figure 5: Estimated size of the proposed variational gap bound, calculated using (44) for VAE, IWAE-5, and IWAE-10 models, previously trained on MNIST, SVHN, and CelebA datasets, vs. constant C . In each case, the optimal value C^{opt} is marked with a short vertical line. All computations were averaged over the test dataset.

Societal Impact There are no foreseen ethical or societal consequences for the research presented herein.

Acknowledgements

This research was partially funded by the National Science Centre, Poland, grants no. 2020/39/D/ST6/01332 (work by Łukasz Struski), 2021/43/B/ST6/01456 (work by Przemysław Spurek), and 2021/41/B/ST6/01370 (work by Jacek Tabor). Some experiments were performed on servers purchased with funds from the flagship project entitled “Artificial Intelligence Computing Center Core Facility” from the DigiWorld Priority Research Area within the Excellence Initiative – Research University program at Jagiellonian University in Kraków. Marcin Mazur’s participation in the Conference was supported by the Faculty of Mathematics and Computer Science of the Jagiellonian University from the funds of the Excellence Initiative at the Jagiellonian University.

References

- Abramovich, S. and Persson, L.-E. (2016). Some new estimates of the ‘Jensen gap’. *Journal of Inequalities and Applications*, 2016(1):1–9.
- Bayer, J., Soelch, M., Mirchev, A., Kayalibay, B., and van der Smagt, P. (2021). Mind the gap when conditioning amortised inference in sequential latent-variable models. *arXiv preprint arXiv:2101.07046*.
- Bickel, P. J. and Doksum, K. A. (2015). *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC.
- Botvinick, M. and Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, 16(10):485–488.
- Brnetic, I., Khan, K. A., and Pecaric, J. (2015). Refinement of Jensen’s inequality with applications to cyclic mixed symmetric means and cauchy means. *J. Math. Inequal*, 9(4):1309–1321.
- Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.
- Dayan, P. and Hinton, G. E. (1997). Using expectation-maximization for reinforcement learning. *Neural Computation*, 9(2):271–278.
- Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. (2017). Variational inference via χ upper bound minimization. *Advances in Neural Information Processing Systems*, 30.
- Dragomir, S. S. (2013). Some reverses of the Jensen inequality with applications. *Bulletin of the Australian Mathematical Society*, 87(2):177–194.
- Duo, X. (2021). Improving actor-critic reinforcement learning via Hamiltonian Monte Carlo method. In *Deep RL Workshop NeurIPS 2021*.
- Gao, X., Sitharam, M., and Roitberg, A. E. (2017). Bounds on the Jensen gap, and implications for mean-concentrated distributions. *arXiv preprint arXiv:1712.05267*.
- Grosse, R. B., Ghahramani, Z., and Adams, R. P. (2015). Sandwiching the marginal likelihood using bidirectional monte carlo. *arXiv preprint arXiv:1511.02543*.
- Horváth, L. (2021). Extensions of recent combinatorial refinements of discrete and integral Jensen inequalities. *Aequationes Mathematicae*, pages 1–21.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR.
- Jebara, T. and Pentland, A. (2001). On reversing Jensen’s inequality. *Advances in Neural Information Processing Systems*, pages 231–237.
- Ji, C. and Shen, H. (2019). Stochastic variational inference via upper bound. *arXiv preprint arXiv:1912.00650*.
- Khan, M. A., Khan, S., and Chu, Y. (2020). A new bound for the Jensen gap with applications in information theory. *IEEE Access*, 8:98001–98008.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.
- Knop, S., Spurek, P., Tabor, J., Podolak, I., Mazur, M., and Jastrzebski, S. (2020). Cramer-Wold auto-encoder. *Journal of Machine Learning Research*, 21(164):1–28.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*.
- Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, volume 29.
- Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738.
- Maddison, C. J., Lawson, D., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., and Teh, Y. W. (2017). Filtering variational objectives. *arXiv preprint arXiv:1705.09279*.
- Masrani, V., Le, T. A., and Wood, F. (2019). The thermodynamic variational objective. *Advances in Neural Information Processing Systems*, 32.
- Mouri, H. (2013). Log-normal distribution from a process that is not multiplicative but is additive. *Physical Review E*, 88(4):042124.
- Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Svhn: Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.
- Nielsen, F. (2010). A family of statistical symmetric divergences based on Jensen’s inequality. *arXiv preprint arXiv:1009.4004*.
- Nowozin, S. (2018). Debiasing evidence approximations: On importance-weighted autoencoders and jack-knife variational inference. In *International Conference on Learning Representations*.
- Rainforth, T., Kosiorek, A., Le, T. A., Maddison, C., Igl, M., Wood, F., and Teh, Y. W. (2018). Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pages 4277–4285. PMLR.
- Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st*

International Conference on Machine Learning, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286.

- Ruel, J. J. and Ayres, M. P. (1999). Jensen’s inequality predicts effects of environmental variation. *Trends in Ecology & Evolution*, 14(9):361–366.
- Saeed, T., Khan, M. A., and Ullah, H. (2022). Refinements of Jensen’s inequality and applications. *AIMS Mathematics*, 7(4):5328–5346.
- Todorov, E. (2008). General duality between optimal control and estimation. In *2008 47th IEEE Conference on Decision and Control*, pages 4286–4292. IEEE.
- Toussaint, M. and Storkey, A. (2006). Probabilistic inference for solving discrete and continuous state Markov decision processes. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 945–952.
- Williams, G., Wagener, N., Goldfain, B., Drews, P., Rehg, J. M., Boots, B., and Theodorou, E. A. (2017). Information theoretic mpc for model-based reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1714–1721. IEEE.

A MISSING PROOFS

In this section, we provide proofs omitted from the main article.

Theorem 1. Let f be a smooth concave function. Then

$$f(\mathbb{E}X) \leq \mathbb{E}(f(X) + (Y - X)f'(X)), \quad (47)$$

where X and Y are two independent random variables with the same distribution.

Proof. Let (Ω, μ) be a probabilistic space and $X, Y: \Omega \rightarrow \mathbb{R}$ be independent random variables with the same distribution. We use notation

$$m = \mathbb{E}X = \int_{\Omega} X d\mu. \quad (48)$$

Clearly $\mathbb{E}X = \mathbb{E}Y$. Observe that applying Taylor's expansion and concavity of f (which means that $f'' \leq 0$), for every $\omega \in \Omega$ we have

$$f(m) \leq f(X(\omega)) + (m - X(\omega))f'(X(\omega)). \quad (49)$$

Integrating the above formula over all $\omega \in \Omega$ and making use of the fact that X and Y are independent random variables with the same distribution, we obtain

$$\begin{aligned} f(\mathbb{E}X) &= \int_{\Omega} f(m) d\mu \leq \int_{\Omega} f(X(\omega)) + (m - X(\omega))f'(X(\omega)) d\mu \\ &= \mathbb{E}f(X) + \mathbb{E}Y\mathbb{E}f'(X) - \mathbb{E}(Xf'(X)) = \mathbb{E}(f(X) + (Y - X)f'(X)), \end{aligned} \quad (50)$$

which completes the proof. \square

Theorem 6. Assume that f is a smooth function. Let X and Y be independent random variables with the same distribution, which attain only values ε -close to $\mathbb{E}X$, where $\varepsilon > 0$ is small. Then

$$f(\mathbb{E}X) = \mathbb{E}f(X) + \frac{1}{2}\mathbb{E}((Y - X)f'(X)) + o(\varepsilon^2). \quad (51)$$

Proof. Let (Ω, μ) be a probabilistic space and $X, Y: \Omega \rightarrow \mathbb{R}$ be random variables satisfying the assumptions of the theorem. Put $m = \mathbb{E}X = \mathbb{E}Y$. By applying Taylor's expansion for f and f' , taking any $\omega \in \Omega$ we have

$$f(m) = f(X(\omega)) + (m - X(\omega))f'(X(\omega)) + \frac{1}{2}(m - X(\omega))^2 f''(X(\omega)) + o(\varepsilon^2) \quad (52)$$

and

$$f'(m) = f'(X(\omega)) + (m - X(\omega))f''(X(\omega)) + o(\varepsilon). \quad (53)$$

Hence,

$$f(m) = f(X(\omega)) + \frac{1}{2}(m - X(\omega))f'(m) + \frac{1}{2}(m - X(\omega))f'(X(\omega)) + o(\varepsilon^2). \quad (54)$$

Then integrating the above formula over all $\omega \in \Omega$ and making use of the fact that X and Y are independent random variables with the same distribution, we obtain

$$f(\mathbb{E}X) = \mathbb{E}f(X) + \frac{1}{2}(\mathbb{E}X\mathbb{E}f'(X) - \mathbb{E}(Xf'(X))) + o(\varepsilon^2) = \mathbb{E}(f(X) + \frac{1}{2}(Y - X)f'(X)) + o(\varepsilon^2), \quad (55)$$

which completes the proof. \square

B OTHER BOUNDS

In this section, we follow the notation established in Section 4 of the main paper.

χ **Upper Bound (CUBO)** The (general) χ upper bound for the evidence $\log p(x)$ was derived by Dieng et al. (2017), with the use of the χ divergence. It is given by the following formula:

$$\text{CUBO}_n = \frac{1}{n} \log \mathbb{E}_{z \sim q(\cdot|x)} \left(\frac{p(x, z)}{q(z|x)} \right)^n = \frac{1}{n} \log \mathbb{E}_{z \sim q(\cdot|x)} R(x, z)^n, \quad (56)$$

where $n \geq 1$. Let us note that CUBO is strictly related to the variational Rényi bound provided by Li and Turner (2016).

The authors of (Dieng et al., 2017) prove (see the sandwiching theorem therein) that CUBO_n is a non-decreasing function of $n \geq 1$ and

$$\text{IW-ELBO}_k \leq \log p(x) \leq \text{CUBO}_n \quad (57)$$

for any $k \geq 1$. Hence, for comparison with our approach, as the respective bound for the variational gap (CUBO-VG-B) we take a difference between CUBO_n and IW-ELBO_k , i.e.

$$\text{CUBO}_n\text{-VG-B} = \text{CUBO}_n - \text{IW-ELBO}_k. \quad (58)$$

Thermodynamic Variational Objective (TVO) Masrani et al. (2019) provide upper and lower bounds for the evidence $\log p(x)$, obtained by applying the thermodynamic integration technique. It is based on bounding (via left and right Riemann sum) a one-dimensional integral of the expected values of instantaneous ELBO, calculated under latent distributions π_β for $\beta \in [0, 1]$, lying on a geometric path between $q(z|x)$ and $p(x, z)$ (hence $\beta_0 = q(z|x)$ and $\beta_1 = p(z|x)$). This leads to a lower and upper version of the thermodynamic variational objective (TVO), given by the following formulas:

$$\text{TVO}_K^L = \frac{1}{K} \sum_{l=0}^{K-1} \mathbb{E}_{z \sim \pi_{\beta_l}} \log \frac{p(x, z)}{q(z|x)} = \frac{1}{K} \sum_{l=0}^{K-1} \mathbb{E}_{z \sim \pi_{\beta_l}} \log R(x, z), \quad (59)$$

and

$$\text{TVO}_K^U = \frac{1}{K} \sum_{l=1}^K \mathbb{E}_{z \sim \pi_{\beta_l}} \log \frac{p(x, z)}{q(z|x)} = \frac{1}{K} \sum_{l=1}^K \mathbb{E}_{z \sim \pi_{\beta_l}} \log R(x, z), \quad (60)$$

where $\beta_l = l/K$. Let us note that $\text{TVO}_1^L = \text{ELBO}$ and TVO_1^U coincides with the evidence upper bound (EUBO), which was introduced by Ji and Shen (2019).

It was proved in (Masrani et al., 2019) that for any $K \geq 1$ we have

$$\text{TVO}_K^L \leq \log p(x) \leq \text{TVO}_K^U. \quad (61)$$

Hence, for our purpose, as the respective bound for the variational gap (TVO-VG-B) we take a difference between TVO_K^U and TVO_K^L , i.e.:

$$\text{TVO}_K\text{-VG-B} = \text{TVO}_K^U - \text{TVO}_K^L. \quad (62)$$

In particular,

$$\text{EUBO-VG-B} = \text{TVO}_1\text{-VG-B} = \text{TVO}_1^U - \text{TVO}_1^L = \text{EUBO} - \text{ELBO}. \quad (63)$$

C EXPERIMENTAL DETAILS AND EXTENSIONS

Experimental Results for VAE and IWAE Models We validated our importance sampling variational gap bound (IS-VG-B) against the other state-of-the-art bounds, denoted by CUBO-VG-B, EUBO-VG-B, and TVO-VG-B, depending on the evidence upper bound used. All computations were made for VAE and IWAE models, previously trained with the use of ELBO, IW-ELBO₅, and IW-ELBO₁₀ objectives. For the IWAE models, we used the same neural architectures as in VAE.

Table 2, which is an extension of Table 1 from the main paper, presents the size of all considered variational gap bounds for the evidence of data, computed with the use of various numbers of latent samples and averaged over the test dataset.



Figure 6: Sampled images for the VAE, IWAE-5, and IWAE-10 models (from left to right) trained on MNIST, SVHN, and CelebA datasets (from top to bottom).

following formulas:

$$\mathbb{E}_{z \sim q(\cdot|x)} R(x, z)^n \approx \frac{1}{k} \sum_{i=1}^k R(x, z_i)^n \text{ for } n \geq 1, \quad (64)$$

$$\mathbb{E}_{z \sim \pi_{\beta_l}} \log R(x, z) \approx \sum_{i=1}^k \bar{w}_i^l \log R(x, z_i) \text{ for } l = 0, \dots, K, \quad (65)$$

$$\mathbb{E}_{z_i \sim q(\cdot|x)} \log \frac{1}{k} \sum_{i=1}^k R(x, z_i) \approx \frac{1}{m} \sum_{j=1}^m \log \frac{1}{k} \sum_{i=1}^k R(x, z_i^j), \quad (66)$$

$$\mathbb{E}_{z_i, \tilde{z}_i \sim q(\cdot|x)} \frac{\sum_{i=1}^k R(x, \tilde{z}_i)}{\sum_{i=1}^k R(x, z_i)} \approx \frac{1}{m} \sum_{j=1}^m \frac{\sum_{i=1}^k R(x, \tilde{z}_i^j)}{\sum_{i=1}^k R(x, z_i^j)}, \quad (67)$$

where z_i , z_i^j , and \tilde{z}_i^j were sampled from $q(\cdot|x)$, $w_i^l = R(x, z_i)^{\beta_l}$, and $\bar{w}_i^l = w_i^l / \sum_{j=1}^k w_j^l$. In each case, we forced the use of the same number of latent samples (z_i or z_i^j), and therefore computations of CUBO and TVO were averaged over m repetitions. In our experiments we set $m = 3$ and $k = 2^3, \dots, 2^{16}$.

Training and Architecture Details All experiments were performed on a single Tesla v100 GPU. We used a convolutional VAE architecture (see below for more details) with weights optimized by the Adam optimizer and a learning rate of 0.0001. The networks were trained for 100 epochs with a batch size of 64 for the CelebA dataset and 50 epochs for both MNIST and SVHN datasets. In all reported experiments, we used Euclidean latent spaces $\mathcal{Z} = \mathbb{R}^d$ for $d = 8, 32, 128$, depending on the used dataset (respectively: MNIST, SVHN, and CelebA). We took standard Gaussian priors $p(z) \sim \mathcal{N}(0, I_d)$. We used Gaussian encoders $q(z|x) \sim \mathcal{N}(\mu_x, \Sigma_x)$, with a mean μ_x and a diagonal covariance matrix Σ_x , and Gaussian decoders $p(x|z) \sim \mathcal{N}(q(z|x), \sigma^2 I)$ with $\sigma^2 = 0.3$.

We applied three different models, each for a different dataset. For the MNIST dataset, we used a network architecture that contained two parts: an encoder and a decoder. The encoder consisted mainly of a 2-layer fully-connection, and the decoder consisted of a 3-layer fully-connection. Between each layer, we used the ReLU activation function.

In the case of the SVHN dataset, we used deeper architectures for the encoder and decoder. Both networks consisted mainly of 4 layers. The encoder had only 2D convolutions, between which we used leaky ReLU with leakiness of 0.2.

In the decoder, we applied 2D transposed convolutions and ReLU as activation functions. For the last layer, we used the sigmoid activation function.

For the CelebA dataset, we applied network architectures that consisted mainly of replicated 5-layer blocks. In the encoder network, each block was built with a 2D convolution layer, batch normalization (Ioffe and Szegedy, 2015), and leaky ReLU with leakiness of 0.2. A single block in the decoder network contained an operation that applied a 2D nearest-neighbor upsampling to an input signal composed of several input channels. Then, similar to the block of the encoder network, there was a 2D convolution layer, batch normalization, and leaky ReLU with a leakiness of 0.2.

Details for Experiment with Laplace Distributed Data We generated 10^4 observations from Laplace distribution $\text{Laplace}(0, 0.2)$. Then we use them to calculate the (average) log-likelihood, ELBO, IW-ELBO lower bound, and our proposed upper bound (taking $C = 0$) for the previously learned VAE with one-dimensional latent space. To estimate lower and upper bounds, we sampled from the latent different numbers of times (from 1 to 64), to examine how the number of draws will affect bound positions. In the training procedure of the underlying VAE model, we selected batches of size 1000 and SGD optimizer with the learning rate 10^{-7} . The encoder consisted of one hidden layer with 4 neurons and a ReLU activation function in it, with linear activation in the output layer. We set an identical network architecture for the decoder.

Simple Qualitative Evaluation We tried to confirm the quantitative results presented in Figure 3 in the main paper by comparing visually samples randomly generated by various models. We expected to obtain images of the best quality for VAE trained on MNIST and for IWAE models trained on SVHN, as well as images of comparable quality for all considered models trained on CelebA. The respective samples are presented in Figure 6. Although some slight visual effects might be visible after a closer look, the differences are not impressive. Hence, in this case, simple qualitative evaluation is a rather inadequate method.