COMPLETE LIKELIHOOD OBJECTIVE FOR LATENT VARIABLE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we propose an alternative to the Marginal Likelihood (MaL) objective for training latent variable models, Complete Latent Likelihood (CoLLike). We analyze the objectives from the perspective of matching joint distributions. We show that MaL corresponds to a particular KL divergence between some target *joint* distribution and the model joint. Furthermore, the properties of the target joint explain such major malfunctions of MaL as uninformative latents (posterior collapse) and high deviation of the aggregated posterior from the prior. In CoLLike approach, we use a sample from the prior to construct a family of target joint distributions, which properties prevent these drawbacks. We utilize the complete likelihood both to choose the target from this family and to learn the model. We confirm our analysis by experiments with low-dimensional latents, which also indicate that it is possible to achieve high accuracy unsupervised classification using CoLLike objective.

1 INTRODUCTION

In the latent variable setting, the model defines a joint distribution over both observed variables x and latent variables z, while the training data contains only observed variables. The problem can be treated as an unknown $z|x$ target conditional distribution. There are at least two possible solutions to this problem: try to come up with a meaningful target $z|x$ distribution and train the model similar to a supervised setting, or give up and focus on matching only marginals in the x domain. The latter is the choice of the MaL objective. In this work, we follow the former approach. However, instead of picking up a single target conditional we construct an entire family of possible distributions and use the model likelihood to decide which conditional to use as a target.

To construct a family of possible conditionals we use a sample from prior of the same size as the dataset in the observed domain. All possible assignments of observed samples to latent ones span a family of empirical joint distributions. This can be represented as permutations of the latent samples. Despite the size of the permutations set being tremendous and growing as factorial of the dataset size, the search of the permutation with the best likelihood can be done efficiently using combinatorial optimization. The resulting optimization procedure resembles expectation maximization algorithm (Dempster et al., 1977), where expectation is replaced with the combinatorial assignment problem. Furthermore, since the proposed algorithm uses gradient-free optimization for obtaining the target distribution, the objective can be seamlessly applied to both continuous and discrete latent variables, while the discrete latents case is challenging for approaches based on the MaL(Mnih & Gregor, 2014; Mnih & Rezende, 2016; Tucker et al., 2017).

We analyze the objectives from the perspective of matching *joint* distributions. We show that MaL corresponds to a specific choice of the target $z|x$ conditional, while our approach takes into consideration family of possible conditionals. The choice of the target conditional is responsible for two major failures that arise during training with the MaL objective: inability to learn informative latents also known as "posterior collapse" (Bowman et al., 2016; Razavi et al., 2019; He et al., 2019) and divergence between the prior and the aggregated posterior (Hoffman & Johnson, 2016; Makhzani et al., 2015; Zhao et al., 2019; Kim & Mnih, 2018). These characteristics are vital for latent variable models because posterior collapse prevents learning meaningful representation and samples from the regions of high deviation of the latent marginals are subjected to severe quality degradation (Rosca et al., 2018). The form of the target joint also motivates the success of the complete likelihood

in these challenges. Namely, the target distribution for CoLLike has high mutual information and matches prior.

We verify our analysis with experiments. In this work, we focus on low-dimensional latent variables to perform direct comparison with exact MaL. Models trained with CoLLike stably maintain high mutual information and low divergence with the prior. In turn, MaL inevitably leads to either posterior collapse or highly divergent aggregated posterior. Previously, for simple linear models it has been shown that posterior collapse takes place during optimization of the exact likelihood Lucas et al. (2019). Our experiments demonstrate that it can as well happen with expressive models trained with exact likelihood. Along with informativeness and latent distribution matching, CoLLike indicates no degradation of likelihood compared to MaL. Furthermore, we show that CoLLike objective alone can achieve high accuracy in unsupervised classification. However, the resulting variance of the accuracy is high. We propose a latent ensembling algorithm to tackle this issue, which not only stabilizes but also significantly increases accuracy.

We show that CoLLike unifies a range of existing approaches that lack probabilistic justification. Constrained K-means (Bennett et al., 2000), Permutation Invariant Training (Yu et al., 2017; Luo & Mesgarani, 2019), and Noise as Target Bojanowski & Joulin (2017) are among these approaches. This allows to extend them to different factorizations of the joint and perform analysis from the probabilistic perspective. Furthermore, CoLLike bridges likelihood and optimal transport (OT) frameworks. From this perspective, the negative likelihood plays the role of both mapping from latent to visible domain and distance function.

In summary, we propose a new objective that compared to MaL allows to effectively train models with such desirable properties as informative latents and matching in the latent domain while maintaining similar likelihood levels. We propose a joint perspective on the objectives that explains MaL and CoLLike properties observed in experiments. We show that the proposed objective unifies and explains a range of existing approaches from probabilistic perspective. Besides, we show that it is possible to achieve high quality unsupervised classification using CoLLike and the proposed latent ensembling.

2 Complete Likelihood Objective

In the supervised setting, we are given a model $p_\theta(x, y)$ with parameters θ and a dataset represented by a collection of samples $\{(x_1, y_1), ..., (x_N, y_N)\}$. To mimic the data distribution with the model, a reasonable objective would be complete likelihood:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log p_\theta(x_i, y_i) \tag{1}$$

The motivation behind complete likelihood comes from the equivalence of maximizing (1) and minimizing the Kullback–Leibler divergence $KL(p_\delta(x, y)||p_\theta(x, y))$ which measures the discrepancy between the target empirical data distribution $p_\delta(x, y)$[1] and the model distribution $p_\theta(x, y)$ (Murphy, 2022, 4.2.2).

In the regular latent variable setting, we are given a dataset $\{x_1, ..., x_N\}$ and the model $p_\theta(x, z) = p_\theta(x|z)p(z)$. Missing labels z can be treated as missing $p_\delta(z|x)$ part of the target joint. If we cannot come up with a reasonable $z|x$ target we can at least match the marginals in the observed domain with $KL(p_\delta(x)||p_\theta(x))$ in hope that the model will learn an informative relation between x and z. This is equivalent to the maximization of MaL:

$$\mathcal{L}_{MaL}(\theta) = \sum_{i=1}^{N} \log p_\theta(x_i) = \sum_{i=1}^{N} \log \int p_\theta(x_i, z) dz \tag{2}$$

In general, we cannot compute $p_\theta(x)$ exactly due to the integration operation. This fact leads to an abundance of approximation techniques, which in the majority are aimed at getting better estimates of $p_\theta(x)$ (Hoffman et al., 2013; Kingma & Welling, 2014; Mnih & Gregor, 2014; Salimans et al., 2015; Burda et al., 2016) or $\nabla_\theta p_\theta(x)$ (Ruiz et al., 2021).

[1] We find the Greek letter δ especially suitable for data distribution because it is consonant with "data" and reflects the delta-function-like form of the empirical distribution.

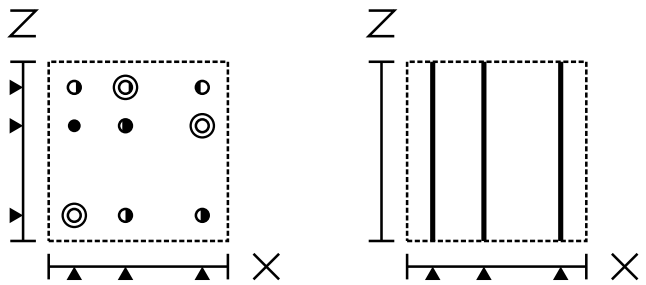


Figure 1: Illustration of the CoLLike (left) and MaL (right) objectives. Double circles and bold lines indicate areas of the joint to be maximized and circle filling represents likelihood.

Despite the family of all possible target $p(z|x)$ distributions being tremendous we do not need to consider it entirely. Firstly, the marginal of the target conditional in the latent domain should match the prior $p(z)$. Secondly, as in any real-world dataset, it should confidently assign a label z to each x in the dataset. It is not hard to get a rich family of distributions with such properties. We can obtain a collection $\{z_1, ..., z_N\}$ by sampling from the prior and pair this collection with the dataset $\{x_1, ..., x_N\}$. Sampling from the prior ensures the first requirement, while the assignment of a single z to each x assures the second. We express each pairing as some permutation π, which produces a complete collection $\{(x_1, z_{\pi(1)}), ..., (x_N, y_{\pi(N)})\}$ and empirical joint $p_{\delta\pi}(x, z) = p_\delta(x)p_\pi(z|x)$. Given a family of distributions we need to decide which member of the family is our target. We propose to pick the one with the highest complete likelihood relying on the model inductive biases. For this target we optimize once again the complete likelihood of the $(x_i, z_{\pi^*(i)})$ pairs with the optimal permutation π^*. These considerations lead us to the CoLLike objective:

$$\mathcal{L}_{CL}(\theta, \pi) = \sum_{i=1}^{N} \log p_\theta \left(x_i, z_{\pi(i)}\right) \tag{3}$$

which we maximize both with respect to θ and π. An alternative view on the objective can be the following: we sample z values from prior and assume that they are ground truth targets for the training dataset with unknown pairing. Figure 1 depicts the main difference between the objectives: CoLLike maximizes specific points of the joint distribution, while MaL is aimed at maximization of whole lines along the joint.

3 Objective Analysis

We start our analysis by proving that MaL corresponds to matching of a specific joint distribution and the model joint:

$$KL(p_\delta(x)p_\theta(z|x)||p_\theta(x,z)) = \mathbb{E}_{x,z\sim p_\delta(x)p_\theta(z|x)}\left[\log \frac{p_\delta(x)p_\theta(z|x)}{p_\theta(x)p_\theta(z|x)}\right] = \mathbb{E}_{x\sim p_\delta(x)}\left[\log \frac{p_\delta(x)}{p_\theta(x)}\right]$$
$$= \mathbb{E}_{x\sim p_\delta(x)}\left[\log p_\delta(x)\right] - \mathbb{E}_{x\sim p_\delta(x)}\left[\log p_\theta(x)\right] = C - \frac{1}{N}\sum_i \log p_\theta(x_i) = C - \frac{1}{N}\mathcal{L}_{MaL}(\theta)$$

where C is a constant. The joint KL form of the MaL brings new perspectives on the objective. It might be tempting to think about MaL as a workaround for unknown latents that allows you not to specify the target $z|x$ conditional. However, the joint form reveals that the target conditional is actually specified and equals $p_\theta(z|x)$ if we ask what distribution we want to mimic. This implies that we are aimed at keeping the model posterior unchanged. In addition, the form also highlights the intimate connection between MaL and posterior.

CoLLike and a common variational (Jordan et al., 1999) approximation of MaL, Evidence Lower Bound (ELBO), can also be expressed as KL divergences between joint distributions (see Table 1). We refer to Appendix A for derivation of the equivalence. Note the elegant similarity between objectives which becomes obvious in the joint KL form. All divergences share the model $p_\theta(x, z)$ as the second argument, which implies that the first argument is the target *joint* distribution. For all objectives the target joint contains the data distribution $p_\delta(x)$ as a marginal in x domain, thus the

Table 1: Considered objectives and their joint KL forms.

	Original Objective	**Joint KL form**
CoLLike	$\sum_{i=1}^{N} \log p_\theta\left(x_i, z_{\pi(i)}\right)$	$KL(p_\delta(x)p_\pi(z\|x)\|\|p_\theta(x,z))$
MaL	$\sum_{i=1}^{N} \log p_\theta(x_i)$	$KL(p_\delta(x)p_\theta(z\|x)\|\|p_\theta(x,z))$
ELBO[2]	$\sum_{i=1}^{N} \mathbb{E}_{z\sim q_\phi(z\|x_i)}\left[\log \frac{p_\theta(x_i,z)}{q_\phi(z\|x_i)}\right]$	$KL(p_\delta(x)q_\phi(z\|x)\|\|p_\theta(x,z))$

only difference is in the target $z|x$ conditional. Therefore, all considered objectives belong to the same family of the form:

$$\begin{aligned}\mathcal{L}(\theta) &= KL(p_\delta(x)p(z|x)||p_\theta(x,z)) = KL(p_\delta(x)p(z|x)||p_\theta(x)p_\theta(z|x)) \\ &= KL(p_\delta(x)||p_\theta(x)) + \mathbb{E}_{p_\delta(x)}\left[KL(p(z|x)||p_\theta(z|x))\right]\end{aligned} \tag{4}$$

Since the second term in (4) is non-negative, all objectives in the family are lower bounds on the likelihood up to an additive constant. Note that the $z|x$ target conditional is used to minimize the overall divergence. This affects the second term of (4) to make the lower bound tighter.

Despite the common traits, the objectives are different. We will highlight a few differences and go deeper in the following sections. Firstly, the target conditional for CoLLike $p_\pi(z|x)$ is empirical, while its counterparts $p_\theta(z|x)$ and $q_\phi(z|x)$ aren't. Secondly, in MaL approach, we construct a particular joint distribution $p_\delta(x)p_\theta(z|x)$ and use it as a target joint, while, in CoLLike, we construct an entire family of joint distributions with desired properties. Thirdly, the target posterior is readily available in CoLLike and ELBO cases, while for MaL it could be intractable. Lastly, the CoLLike objective allows learning models with reverse factorization $p_\theta(x)p_\theta(z|x)$, while MaL and ELBO do not.

3.1 Mutual Information of the Target Distribution

Mutual Information (MI) is the key property of the joint distribution in a latent variable setting. It characterizes how dependent the observed and latent variables are. We would like to know what MI value our model is targeted at for each objective. Since our objective can be expressed as KL divergence between model and target joint distributions (Table 1), we can investigate MI values for each target joint. We define MI between x and z under $p(x,z)$ distribution as:

$$MI(p(x,z)) = \mathbb{E}_{x,z\sim p(x,z)}\left[\log \frac{p(x,z)}{p(x)p(z)}\right] \tag{5}$$

For MaL, the MI of the target $p_\delta(x)p_\theta(z|x)$ is determined by the model's current posterior $p_\theta(z|x)$. Most models have no class preferences at initialization, which results in low MI of $p_\delta(x)p_\theta(z|x)$. Moreover, we are aimed at keeping it unchanged, since we are using the current posterior as our target posterior. So, low MI at initialization might induce learning non-meaningful factorized joint throughout the training procedure. Since for ELBO the approximate posterior aligns to the true model posterior this argument is applicable to ELBO too. Furthermore, uninformative posterior is a common problem when learning a latent variable model (Bowman et al., 2016; Alemi et al., 2018; Lucas et al., 2019; Razavi et al., 2019; He et al., 2019) known as "posterior collapse".

CoLLike target is an empirical joint distribution. It represents a deterministic mapping and has constantly high MI by construction, as shown in Appendix B. Therefore, we are aimed at mimicking a high MI distribution with our model distribution. Furthermore, CoLLike can be interpreted as some realization of InfoMax principle Huszár (2017), where prior limits the entropy and deterministic mapping maximize MI.

3.2 Matching in the Latent Domain

The joint form of the objectives from Table 1 is convenient for obtaining a perspective on distribution matching in the latent space. After treating $p_\delta(x)p_\theta(z|x)$ as a joint $p_{\delta\theta}(x,z)$ and rewriting the

[2] $q_\phi(z|x)$ is an approximate posterior distribution parametrized by ϕ.

original MaL objective as:

$$KL(p_\delta(x)||p_\theta(x)) = KL(p_{\delta\theta}(x,z)||p_\theta(x,z)) = \mathbb{E}_{x,z\sim p_{\delta\theta}(x|z)p_{\delta\theta}(z)}\left[\log \frac{p_{\delta\theta}(x|z)p_{\delta\theta}(z)}{p_\theta(x|z)p_\theta(z)}\right]$$
$$= \mathbb{E}_{z\sim p_{\delta\theta}(z)}\left[KL\left(p_{\delta\theta}(x|z)||p_\theta(x|z)\right)\right] + KL(p_{\delta\theta}(z)||p_\theta(z)) \tag{6}$$

we see that matching in x space requires matching in z space. Namely, $KL(p_{\delta\theta}(z)||p_\theta(z)) = 0$, where $p_{\delta\theta}(z)$ is called an aggregated posterior. It signifies that even though MaL is constructed such that z given x conditional part of the KL between joints is zero, we end up in a situation where none of the model marginals match target marginals. Moreover, the learning signal from the first term of (6) might be significantly larger compared to the second term signal if the dimensions of x and z differ a lot. This might lead to a sacrifice of the second divergence in favor of the first one.

Matching in a latent domain is considered as a known challenge of latent variable modelling (Hoffman & Johnson, 2016). Mismatch with prior results in unnatural samples from areas with high deviation of aggregated posterior from the prior (Rosca et al., 2018). A number of works is focused on this problem. They either utilize additional losses that penalize discrepancy between marginals (Makhzani et al., 2015; Zhao et al., 2019; Kim & Mnih, 2018) or introduce a learnable prior (Bauer & Mnih, 2019; Tomczak & Welling, 2018).

In turn, CoLLike addresses this problem by constructing a conditional, which marginal matches prior in the latent domain. Obviously, the target marginal in x domain for CoLLike is always $p_\delta(x)$. In turn, the target aggregate posterior is always a sample from the prior since $p_{\delta\pi}(z) = \int_x p_\delta(x)p_\pi(z|x)dx = p_\epsilon(z)$ for all π values, where $p_\epsilon(z)$ is the distribution of the sample produced by sampling from the prior. While it is intuitively obvious that the empirical distribution converges to the underlying distribution, one can show that KL between the empirical sample and the prior converges in probability to 0 (Cover & Thomas, 2006, Theorem 11.2.1).

3.3 Gradient Estimation with Respect to Sampling

The main challenge in optimization of ELBO is to estimate the gradient with respect to the approximate posterior parameters ϕ. Approximate posterior $q_\phi(z|x)$ appears both in the probability ratio and in the expectation:

$$KL(p_\delta(x)q_\phi(z|x)||p_\theta(x,z)) = \mathbb{E}_{x,z\sim p_\delta(x)q_\phi(z|x)}\left[\log \frac{p_\delta(x)q_\phi(z|x)}{p_\theta(x,z)}\right] \tag{7}$$

Expectation is usually estimated by sampling. Differentiation with respect to the sampling procedure is hard. For a limited range of continuous distributions, the problem can be solved by reparametrization trick (Kingma & Welling, 2014; Salimans & Knowles, 2012). For other cases, including discrete z, a general purpose score-function estimator can be applied (Williams, 1992). Nevertheless, it requires significant efforts to obtain reliable gradients (Mnih & Gregor, 2014; Mnih & Rezende, 2016; Tucker et al., 2017).

For CoLLike, the target distribution $p_\delta(x)p_\pi(z|x)$ is parameterized by permutation π. This parameterization does not rely on an encoder, so the approach is encoderless. The permutation cannot be tuned by gradient-based techniques, however, since the set of all possible π values is countable and finite, optimization can be performed by an exhaustive search.

4 Algorithm

The complete likelihood objective 3 is function of the permutation π and the model parameters θ. We approach the objective by alternating between finding optimal pairing π^* and gradient-based optimization of θ. The procedure resembles EM algorithm, where the expectation is replaced with the assignment x samples to z samples. To tackle $N!$ search space of possible pairings we evaluate the likelihood of all (x, z) pairs and treat the task as a combinatorial linear sum assignment problem with $\mathcal{O}(N^3)$ complexity, as shown in Appendix D. The following python-like pseudocode describes the proposed algorithm:

```
x_collection = load_dataset()
n_samples = len(x_collection)
z_collection = model.p_z.sample(n_samples)
optim = optimizer()

for epoch in range(n_epochs):
  log_p_xz_mat = empty(n_samples, n_samples)
  for nx, x in enumerate(x_collection):
    for nz, z in enumerate(z_collection):
      log_p_xz_mat[nx, nz] = model.log_p_xz(x, z)

  opt_rows, opt_cols = maximize_linear_assignment(log_p_xz_mat)
  loss = 0
  for opt_nx, opt_nz in zip(opt_rows, opt_cols):
    loss += -log_p_xz_mat[opt_nx, opt_nz]
  optim.apply_grads(compute_grads(loss, model.parameters))
```

The main challenges of the proposed algorithm are $\mathcal{O}(N^2)$ calls of the model and $\mathcal{O}(N^3)$ complexity of combinatorial optimization. While former might be more challenging due to an extensive amount of computations inside the model, the later exhibits faster growth. We purpose a range of techniques to reduce the complexity below.

Minibatch Assignment. The proposed algorithm is aimed at full batch optimization. Aside from squared complexity in dataset size N, there are two reasons to develop a minibatch version of it. Firstly, combinatorial optimization can be prohibitive for large N. Secondly, large batches can potentially harm generalization You et al. (2019; 2020). We apply a minibatch technique similar to Bojanowski & Joulin (2017). We optimize over permutation set and perform the gradient step only for minibatch. However, we store an array of z values and permute minibatches inside it according to the optimal permutation. The number of model forward passes is now quadratic in terms of minibatch size instead of dataset size. However, as gracefully shown by Huszár (2017), this kind of minibatch combinatorial optimization provides only locally optimal solutions.

Low-dimensional Discrete Latents. Discrete latent variables with a number of categories K lower than N can provide additional speed-ups. In the case, all possible values z for each x is $K < N$ which results in cost $\mathcal{O}(NK)$ instead of $\mathcal{O}(N^2)$. For instance, if there is a single binary random variable, the computational cost is proportional to $2N$.

Factorized Conditional. If $p_\theta(x|z)$ part of the model factorizes (each x dimension is predicted independently given z like in the original VAE Kingma & Welling (2014)) only one forward pass is required. Concretely, the distribution for all dimensions of x is obtained by passing z to $p_\theta(x|z)$. In contrast to, for instance, autoregressive $p_\theta(x|z)$, the distribution does not depend on previous dimensions of x. Having a distribution for all dimensions of x, it is regularly cheap to evaluate likelihood of a particular x.

5 Connections

Connections with existing techniques not only give alternative perspectives on CoLLike objective, but also provide probabilistic grounding to some existing algorithms. Many well-known objectives actually use CoLLike while being motivated as an ad-hoc empirical risk minimization. We show that these objectives not only seem reasonable but are also probabilistically motivated.

While traditional K-means algorithm (MacQueen, 1967; Lloyd, 1982) has a probabilistic interpretation (Murphy, 2022, 21.4.1.1), its constrained counterpart (Bennett et al., 2000) lacks probabilistic grounds. Constrained K-means is equivalent to CoLLike under factorized Gaussian $p_\theta(x|z)$ and uniform categorical $p(z)$, which has a number of states equal to the number of clusters. This connection allows extending the constrained K-means approach to different generative distributions and priors.

Permutation Invariant Training (PIT) (Yu et al., 2017; Luo & Mesgarani, 2019) used in source separation solutions can also be expressed as CoLLike objective. For instance, in cocktail part problem, we want to separate a mixture of K sources. During training, we have K isolated mixture components and a network that produces K estimates of the components based on a single mixture. We don't know which network output corresponds to which source and we pick a permutation that pro-

duces minimal total mismatch between outputs and sources. This procedure corresponds to training a latent variable model with CoLLike objective, where a categorical latent variable of dimension K determines the source identity. In this setting, we treat mixture components as samples in the dataset.

The closest predecessor of the CoLLike is Noise As Target (NAT) (Bojanowski & Joulin, 2017). This is an unsupervised approach to learn an image encoder. In this approach, the representations produced by a network are assigned to a fixed collection of vectors sampled from the uniform distribution on a sphere. After this, the network parameters are adjusted to make encodings closer to the assigned vectors. This approach is equivalent to CoLLike with reverse model factorization $p_\theta(x, z) = p_\theta(z|x)p(x)$ and factorized Gaussian $p_\theta(z|x)$. Another approaches that obtain clear probabilistic interpretation using CoLLike include: Sinkhorn Autoencoders (Patrini et al., 2019), simultaneous clustering and representation learning (Asano et al., 2020), and (Jeong & Song, 2019).

Bojanowski & Joulin (2017) noticed that NAT objective has Optimal Transport (OT) roots. OT framework can be used to measure discrepancy between distributions. Particularly, for a given non-negative cost function c the optimal transport distance between distributions p_δ and p_ϵ is defined as

$$OT(p_\delta, p_\epsilon) = \min_{\gamma \in \Gamma(p_\delta, p_\epsilon)} \mathbb{E}_{x,z\sim\gamma(x,z)} \left[c(x, z)\right]$$

where $\Gamma(p_\delta, p_\epsilon)$ is the set of all joint distributions on x and z with marginals $p_\delta(x)$ and $p_\epsilon(z)$ respectively. Furthermore, if we use a parametric model p_θ in place of p_ϵ we can fit it by minimizing the distance. Note that in this case we minimize the function that already has a $\min$ function inside.

When both p_δ and p_ϵ are empirical, the search space Γ becomes countable and finite. Now it contains only pairings between points in p_δ and points in p_ϵ. Given an arbitrary initial pairing, we can express all other pairings through permutation applied to either x or z. In this case, the cost becomes

$$OT(p_\delta, p_\epsilon) = \min_{\pi \in \Pi} \sum_i c\left(x_i, z_{\pi(i)}\right)$$

where Π is the set of all permutation functions. This expression is almost the CoLLike objective (3). Choosing the cost function c to be $-\log p_\theta(x, z)$ and switching to maximization make them equivalent[3]. Thus, CoLLike bridges maximum likelihood methods with OT. This connection allows bringing latest developments in OT to improve likelihood-based methods. Furthermore, in Appendix C, we provide an example of the equivalence between CoLLike and Wasserstein distance. In the case, the model's complete likelihood plays the roles of both a mapping from z to x domain and a distance metric.

6 EXPERIMENTS

In this work, we focus on low-dimensional discrete latents. This type of latent variables allows to perform direct comparison with the exact likelihood. Furthermore, we emphasize our focus on learning useful $z|x$ instead of simplifying the model with factorized $x|z$ conditional.

6.1 TRACTABLE LIKELIHOOD

Models with tractable likelihood are perfect for comparison of likelihood-based algorithms because they remove the problem of the likelihood estimation precision. For this type of models all quantities of interest can be computed exactly. Moreover, tractable likelihood allows comparing CoLLike directly with MaL instead of its approximations like ELBO.

We use MNIST (LeCun et al., 1998) and CIFAR (Krizhevsky, 2009) datasets for image modality and AG News (Zhang et al., 2015) for text domain. All these datasets are equipped with class labels. For images, we train a Glow-like normalizing flow conditioned on a discrete latent variable with 10 categories through all coupling layers. For text we use a Transformer Language Model conditioned on a discrete latent variable with 4 categories using additive embedding for all tokens. The size of

[3] A cautious reader might note that for continuous variables non-negativity of $\log p_\theta$ can be violated, however, all model densities used in practice are finite and the corresponding cost can be made positive just by an additive constant which does not change the optimization problem.

Table 2: Results for tractable categorical latents. MNIST, CIFAR – BPD; AG News – NLL.

Dataset	Objective	Accuracy ↑	NLL/BPD ↓	Agg. KL ↓	MI ↑
CIFAR	CoLLike	14.5	3.45	0.01	2.20
CIFAR	MaL	14.0	3.46	0.74	1.50
MNIST	CoLLike	14.1	1.27	0.01	1.95
MNIST	MaL	12.5	1.29	1.61	0.58
AG News	CoLLike	82.1	250.79	0.00	1.32
AG News	MaL	31.6	249.73	0.00	0.00

the discrete variables is equal to the number of classes in the underlying dataset. Small number of categories allows to compute exact marginal likelihood value and speed up CoLLike to $\mathcal{O}(NK)$.

Table 2 presents the results of training the latent variable models for CoLLike and MaL objectives averaged across 4 runs. Both objectives exhibit similar performance in terms of likelihood across datasets. However, other characteristics vary.

MI is high for CoLLike objective on every dataset. Furthermore, it attains approximately maximal value for AG News and CIFAR. MI for MaL objective ranges from zero to values significantly lower than those of CoLLike. Zero MI indicates posterior collapse cases, which are mainly observed in ELBO optimization and recently discovered by Lucas et al. (2019) for MaL applied to simple linear models. This experiment indicates important observation: posterior collapse can as well happen in deep latent variable models during optimization of exact MaL despite usually being corresponded to the structure of ELBO. Importantly, for MNIST dataset, half of the experiments exhibits posterior collapse.

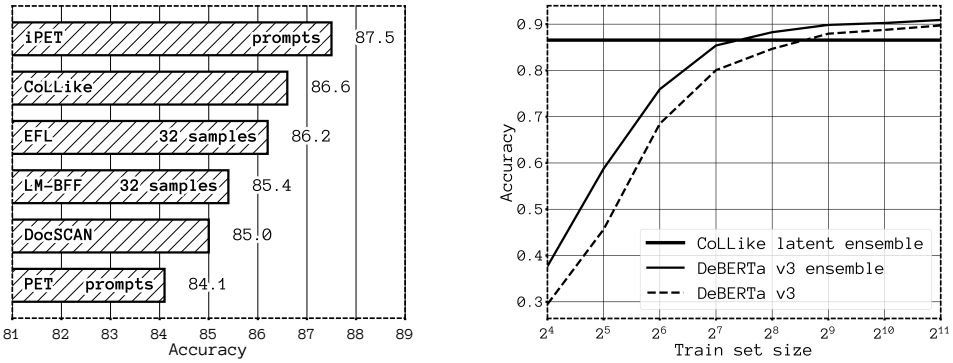
CoLLike exhibits near zero aggregated KL for all experiments. It implies that the model joint marginal in latent domain perfectly matches prior. For MaL, aggregated KL is zero only for AG News dataset which also has uninformative factorized joint. For other datasets, aggregated posterior significantly deviates from the prior. We also note that for MNIST dataset, MaL puts all probability mass to a single category in half of the runs.

To estimate unsupervised classification quality we perform the optimal assignment of latent categories to classes. For all cases except CoLLike objective on AG News dataset, the quality of the unsupervised classification is similar and is low. On AG News the unsupervised accuracy is exceptionally good. However, the variance of the proposed solution is relatively high. Standard deviation of the accuracy across 4 runs is 5.4 with the highest value of 87.1 and the lowest of 73.3. In the following section we show that it is possible to achieve significantly higher unsupervised accuracy and lower variance by latent variable ensembling.

Overall, CoLLike clearly outperforms MaL in the tractable likelihood setting. Moreover, it shows high unsupervised classification accuracy for text modality. For MaL, experiments depict a variety of possible failures from posterior collapse to degenerate aggregated posterior, which extends findings of (Lucas et al., 2019) to expressive models and exact likelihood. However, despite CoLLike producing informative latents in terms of MI, unsupervised classification might be challenging even in these cases. We believe that the key to high-performance unsupervised classification should be in the right inductive biases in conditioning and probabilistic model type.

6.2 Latent Ensembling

To reduce the high variance of CoLLike unsupervised classification and increase its accuracy we propose to perform ensembling of multiple models trained on the same data but using different seeds at initialization. Although there is no correspondence between labels for latent variable models, we can try to find the labels assignment based on the agreement between them. This approach is motivated by direct cluster ensembling (Boongoen & Iam-on, 2018). The agreement between two labels of different ensemble members is the number of intersecting samples with those labels. To align the latents we iteratively find the assignment with the highest intersection between labels. Finally, we find the assignment between aligned latents and ground truth labels.



(a) CoLLike ensemble vs. unsupervised and semi-supervised approaches.

(b) Supervised DeBERTa v3 base vs. CoLLike ensemble.

Figure 2: Comparison of ensembled CoLLike with supervised (a) and unsupervised/few-shot methods(b).

In our experiments, we use 8 models per ensemble and train 4 independent ensembles. The simplest ensembling method is averaging of the predictions. It increases the mean unsupervised accuracy from 82.1 to 84.5 and reduces the standard deviation from 5.4 to 1.7. We further significantly improve these results by utilizing the agreement score, which is also used for alignment of the labels. We pick top-k models with highest maximum coherence across other models in the ensemble. Averaging predictions of those top-k models further increases accuracy to $\mathbf{86.6}$ and lowers the standard deviation to $\mathbf{0.2}$.

We compare CoLLike results with the following unsupervised and supervised approaches: PET and iPET (Schick & Schütze, 2021), EFL (Wang et al., 2021), LM-BFF (Gao et al., 2021), DocSCAN (Stammbach & Ash, 2021). DocSCAN is purely unsupervised, while other approaches rely on engineering multiple textual descriptions of classes (prompts) or labeled data. All methods use heavy pre-trained Transformers (Vaswani et al., 2017) as an initialization, while in CoLLike we use small 2-layer Transformer with random initialization. Figure 2a presents the comparison of the methods. CoLLike clearly outperforms both unsupervised and most of the supervised methods. To determine how much training data we need without, possibly laborious, prompt engineering we use DeBERTa v3 (He et al., 2021). We vary training set sizes from 32 to 2048 and apply additional ensembling of 8 models with different initializations and train-validation splits. Figure 2b reveals that CoLLike can be a better alternative to labeling more than hundred samples, which, in turn, requires an extensive data analysis. Besides, note the high difference between the ensemble and the single model for small dataset sizes in a supervised setting, which is an interesting result by itself.

7 Discussion and Future Work

In this work, we propose to switch from the MaL paradigm of matching only marginals in the observed domain to CoLLike paradigm of finding an exact target joint by selection from a family of joints with desirable properties. Furthermore, we show that matching of marginals utilized by MaL corresponds to a specific choice of target joint, which motivates such failures as posterior collapse and divergence between target and model marginals in the latent domain. We experimentally show the ability of CoLLike to learn useful representations. Connection of CoLLike with OT allows to borrow techniques from the latter. For instance, Sinkhorn Relaxation (Cuturi, 2013) can be used to speed up the assignment problem. Investigation of alternatives to complete likelihood for target selection is of special interest. The right inductive biases for inducing useful properties using CoLLike are still to be discovered, at least until we want to get the desired without specifying what we want. We believe that the further extension of CoLLike to high-dimensional latents would be exciting and challenging. Other lines of research can be devoted to the application of other divergences to the constructed family of joint target candidates and extension of CoLLike to learnable priors.

8 REPRODUCIBILITY

To promote reproducibility we open-source our code *the link is hidden for double-blind review, check the supplementary materials*. Furthermore, we describe details of flow architectures in Appendix E.1 and Transformers in Appendix E.2. Along with models, we describe details of the training procedures and data pre-processing. We also devote special attention to setting all necessary seeds, including CUDA, and to removing stochasticity from the BPE tokenizer.

REFERENCES

Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 159–168. PMLR, 2018. URL http://proceedings.mlr.press/v80/alemi18a.html.

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Hyx-jyBFPr.

Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 66–75. PMLR, 2019. URL http://proceedings.mlr.press/v89/bauer19a.html.

K.P. Bennett, P.S. Bradley, and A. Demiriz. Constrained k-means clustering. Technical Report MSR-TR-2000-65, May 2000. URL https://www.microsoft.com/en-us/research/publication/constrained-k-means-clustering/.

Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 517–526. PMLR, 2017. URL http://proceedings.mlr.press/v70/bojanowski17a.html.

Tossapon Boongoen and Natthakan Iam-on. Cluster ensembles: A survey of approaches with recent extensions and applications. *Comput. Sci. Rev.*, 28:1–25, 2018. doi: 10.1016/j.cosrev.2018.01.003. URL https://doi.org/10.1016/j.cosrev.2018.01.003.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In Yoav Goldberg and Stefan Riezler (eds.), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pp. 10–21. ACL, 2016. doi: 10.18653/v1/k16-1002. URL https://doi.org/10.18653/v1/k16-1002.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1509.00519.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9. URL http://www.elementsofinformationtheory.com/.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL http://www.jstor.org/stable/2984875.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 3816–3830. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.295. URL https://doi.org/10.18653/v1/2021.acl-long.295.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rylDfnCqF7.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John W. Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347, 2013. doi: 10.5555/2567709.2502622. URL https://dl.acm.org/doi/10.5555/2567709.2502622.

Ferenc Huszár. Unsupervised learning by predicting noise: an information maximization view. https://www.inference.vc/unsupervised-learning-by-predicting-noise-an-information-maximization-view-2/, 2017. Accessed: 2022-09-27.

Yeonwoo Jeong and Hyun Oh Song. Learning discrete and continuous factors of data via alternating disentanglement. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3091–3099. PMLR, 2019. URL http://proceedings.mlr.press/v97/jeong19d.html.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, 1999. doi: 10.1023/A:1007665907178. URL https://doi.org/10.1023/A:1007665907178.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2654–2663. PMLR, 2018. URL http://proceedings.mlr.press/v80/kim18b.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, 2019. doi: 10.1561/2200000056. URL https://doi.org/10.1561/2200000056.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–136, 1982. doi: 10.1109/TIT.1982.1056489. URL https://doi.org/10.1109/TIT.1982.1056489.

James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Don't blame the elbo! A linear VAE perspective on posterior collapse. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 9403–9413, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/7e3315fe390974fcf25e44a9445bd821-Abstract.html.

Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 27(8):1256–1266, 2019. doi: 10.1109/TASLP.2019.2915167. URL https://doi.org/10.1109/TASLP.2019.2915167.

J MacQueen. Classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *CoRR*, abs/1511.05644, 2015. URL http://arxiv.org/abs/1511.05644.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 1791–1799. JMLR.org, 2014. URL http://proceedings.mlr.press/v32/mnih14.html.

Andriy Mnih and Danilo Jimenez Rezende. Variational inference for monte carlo objectives. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 2188–2196. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/mnihb16.html.

Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL probml.ai.

Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, 1982. ISBN 0-13-152462-3.

Giorgio Patrini, Rianne van den Berg, Patrick Forré, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. In Amir Globerson and Ricardo Silva (eds.), *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pp. 733–743. AUAI Press, 2019. URL http://proceedings.mlr.press/v115/patrini20a.html.

Ali Razavi, Aäron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-vaes. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BJe0Gn0cY7.

Mihaela Rosca, Balaji Lakshminarayanan, and Shakir Mohamed. Distribution matching in variational inference. *CoRR*, abs/1802.06847, 2018. URL http://arxiv.org/abs/1802.06847.

Francisco J. R. Ruiz, Michalis K. Titsias, A. Taylan Cemgil, and Arnaud Doucet. Unbiased gradient estimation for variational auto-encoders using coupled markov chains. In Cassio P. de Campos, Marloes H. Maathuis, and Erik Quaeghebeur (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, volume 161 of *Proceedings of Machine Learning Research*, pp. 707–717. AUAI Press, 2021. URL https://proceedings.mlr.press/v161/ruiz21a.html.

Tim Salimans and David A. Knowles. Fixed-form variational posterior approximation through stochastic linear regression. *CoRR*, abs/1206.6679, 2012. URL http://arxiv.org/abs/1206.6679.

Tim Salimans, Diederik P. Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1218–1226. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/salimans15.html.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 255–269. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.20. URL https://doi.org/10.18653/v1/2021.eacl-main.20.

Dominik Stammbach and Elliott Ash. Docscan: Unsupervised text classification via learning from neighbors. *CoRR*, abs/2105.04024, 2021. URL https://arxiv.org/abs/2105.04024.

Jakub M. Tomczak and Max Welling. VAE with a vampprior. In Amos J. Storkey and Fernando Pérez-Cruz (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1214–1223. PMLR, 2018. URL http://proceedings.mlr.press/v84/tomczak18a.html.

George Tucker, Andriy Mnih, Chris J. Maddison, Dieterich Lawson, and Jascha Sohl-Dickstein. REBAR: low-variance, unbiased gradient estimates for discrete latent variable models. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 2627–2636, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/ebd6d2f5d60ff9afaeda1a81fc53e2d0-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. *CoRR*, abs/2104.14690, 2021. URL https://arxiv.org/abs/2104.14690.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Yang You, Jonathan Hseu, Chris Ying, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large-batch training for lstm and beyond. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362290. doi: 10.1145/3295500.3356137. URL https://doi.org/10.1145/3295500.3356137.

Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training BERT in 76 minutes. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Syx4wnEtvH.

Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pp. 241–245. IEEE, 2017. doi: 10.1109/ICASSP.2017.7952154. URL https://doi.org/10.1109/ICASSP.2017.7952154.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 649–657, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 5885–5892. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33015885. URL https://doi.org/10.1609/aaai.v33i01.33015885.

A Derivation of KL Forms of the Considered Objectives

Equivalence between CoLLike objective (3) and its KL divergence form from Table 1 can be derived as follows:

$$\begin{aligned} KL(p_\delta(x)p_\pi(z|x)||p_\theta(x,z)) &= \mathbb{E}_{x,z\sim p_\delta(x)p_\pi(z|x)}\left[\log\frac{p_\delta(x)p_\pi(z|x)}{p_\theta(x,z)}\right] \\ &= \mathbb{E}_{x,z\sim p_\delta(x)p_\pi(z|x)}\left[\log p_\delta(x)p_\pi(z|x) - \log p_\theta(x,z)\right] \qquad (8) \\ &= C - \mathbb{E}_{x,z\sim p_\delta(x)p_\pi(z|x)}\left[\log p_\theta(x,z)\right] \\ &= C - \frac{1}{N}\sum_{i=1}^{N}\log p_\theta(x_i, z_{\pi(i)}) \\ &= C - \frac{1}{N}\mathcal{L}_{CLL}(\theta,\pi) \qquad (9) \end{aligned}$$

where the first term in (8) is treated as constant with the assumption that all samples from $p_\delta(x)$ take distinct values, which is reasonable for such high-dimensional objects as images, texts, and sounds. Thus, the KL form of the objective is equivalent to the CoLLike objective up to a multiplicative factor and an additive term. For proof of the constancy see Appendix B.

The derivation of equivalence between (2) and its KL from Table 1 is as follows

$$
\begin{aligned}
KL(p_\delta(x)p_\theta(z|x)||p_\theta(x,z)) &= \mathbb{E}_{x,z\sim p_\delta(x)p_\theta(z|x)}\left[\log \frac{p_\delta(x)p_\theta(z|x)}{p_\theta(x)p_\theta(z|x)}\right] \\
&= \mathbb{E}_{x\sim p_\delta(x)}\left[\log \frac{p_\delta(x)}{p_\theta(x)}\right] \\
&= \mathbb{E}_{x\sim p_\delta(x)}\left[\log p_\delta(x)\right] - \mathbb{E}_{x\sim p_\delta(x)}\left[\log p_\theta(x)\right] \\
&= C - \frac{1}{N}\sum_i \log p_\theta(x_i) \\
&= C - \frac{1}{N}\mathcal{L}_{MaL}(\theta)
\end{aligned}
$$

The KL form of ELBO objective from Table 1 can be found in many works Zhao et al. (2019); Kingma & Welling (2019), however, we provide a derivation here to make the paper self-contained.

$$
\begin{aligned}
KL(p_\delta(x)q_\phi(z|x)||p_\theta(x,z)) &= \mathbb{E}_{x,z\sim p_\delta(x)q_\phi(z|x)}\left[\log \frac{p_\delta(x)q_\phi(z|x)}{p_\theta(x,z)}\right] \\
&= \mathbb{E}_{x\sim p_\delta(x)}\left[\log p_\delta(x)\right] + \mathbb{E}_{x,z\sim p_\delta(x)q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{p_\theta(x,z)}\right] \\
&= C - \mathbb{E}_{x\sim p_\delta(x)}\mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \\
&= C - \frac{1}{N}\sum_i \mathcal{L}_{ELBO}(x_i,\phi,\theta)
\end{aligned}
$$

B Entropy and Mutual Information of Empirical Joint

In this appendix we derive some useful properties of the empirical joint distributions produced by sampling from the prior. The joint distribution $p_\delta(x)p_\pi(z|x)$ depends on π. We focus on how π influences such distribution characteristics as entropy and mutual information.

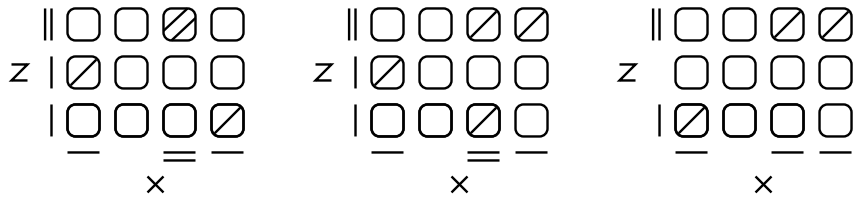


Figure 3: Example of three joint distributions with discrete x and z. Lines on the left and in the bottom depict the number of samples from empirical marginals with the corresponding value of the random variable. Squares reflect the joint probability value. Each line crossing the square corresponds to $1/N$ probability added to the corresponding (x, z) random variable pair.

Consider a joint distribution over discrete x and z. This kind of distribution can be visualized as a table, such as depicted in Figure 3. If there are multiple samples taking the same value both in x and z domain, the permutation can change the entropy of the joint. For instance, for the distribution on the left, the entropy $H(p_\delta(x)p_\pi(z|x)) = -(\frac{1}{4}\log\frac{1}{4} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{2}\log\frac{1}{2}) \approx 1.04$ nats. For the distribution in the center, the entropy $H(p_\delta(x)p_\pi(z|x)) = -(4\cdot\frac{1}{4}\log\frac{1}{4}) \approx 1.39$ nats. So, depending on π we might end up with more and less entropic distributions.

However, when we restrict any empirical marginal to take only distinct values, like in the right part of Figure 3, the situation changes. Namely, each distinct pair (x, z) can be chosen at most once, because to choose it twice we need a duplicate sample in both domains. This can be verified using the right part of Figure 3. Just try to construct a joint with some square having grater than one

line assuming that each value x marginal has only one line. Moreover, for x, this is a reasonable assumption since usually the domain of x is high-dimensional. For the case, the joint will contain N non-zero points each with probability $1/N$. Thus, the entropy of the empirical distribution is equal $-\sum_{n=1}^{N} \frac{1}{N} \log \frac{1}{N} = \log N$.

The observation above allows to easily derive mutual information of the empirical joint. Mutual information is defined as

$$MI(p(x,z)) = \mathbb{E}_{x,z\sim p(x,z)} \log \frac{p(x,z)}{p(x)p(z)} = \mathbb{E}_{x,z\sim p(x,z)} \log \frac{p(z|x)}{p(z)}$$

Under the assumption that $p_\delta(x)$ contains only distinct elements the conditional $p_{\delta\pi}(z|x) \equiv 1$ for all values x from the $p_\delta(x)$. So, choosing x uniquely determines the value of z, as can be seen from the right part of Figure 3. Then the mutual information is given by

$$MI(p_\delta(x)p_\pi(z|x)) = \mathbb{E}_{x,z\sim p_\delta(x)p_\pi(z|x)} \log \frac{p_{\delta\pi}(z|x)}{p_{\delta\pi}(z)} = \mathbb{E}_{z\sim p_{\delta\pi}(z)} \log \frac{1}{p_{\delta\pi}(z)} = H(p_{\delta\pi}(z)) \tag{10}$$

So, the mutual information is always equal to the entropy of the empirical prior. It is possible to show that the value is the maximum possible one. This becomes obvious from the entropic factorization of the mutual information

$$MI(p(x,z)) = H(p(z)) - \mathbb{E}_{x\sim p(x)} \left[H(p(z|x))\right] \tag{11}$$

Since the entropy is non-negative, the mutual information can be decreased only through the second term of (11), which equals 0 because z value is completely determined by x.

When we try to extend the observations above to continuous cases we face the following challenge: empirical distribution has infinite values at the sample points. This drives the differential entropy as well as mutual information to infinity. However, adding noise to the empirical distribution solves this problem. Adding uniform noise with the interval smaller than the precision of the floating point makes the entropy finite and constant with respect to π. One can show that the resulting mutual information of the empirical joint also equals $\log N$.

C WASSERSTEIN DISTANCE AND COLLIKE

Optimal Transport cost becomes Wasserstein distance when c is a metric. A very illustrative example from this family is equality of Wasserstein-2 (c is the Euclidean distance) and CoLLike for some setups. Specifically, the following objective can be produced both by Wasserstein distance and CoLLike

$$\mathcal{L}_W(\theta) = \min_{\pi\in\Pi} \sum_i \left(x_i - f_\theta(z_{\pi(i)})\right)^2$$

To get this objective from OT perspective we define the model distribution to be produced by passing a fixed sample from prior through a deterministic decoder $f_\theta(z)$. The result is an empiric distribution in x domain. Wasserstein distance between two empiric distributions is determined by optimal pairing between points from data distribution $p_\delta(x)$ and model distribution spanned by empiric latents. The same objective is produced by factorized Gaussian $p_\theta(x|z)$ and uniform prior $p(z)$.

This connection demonstrates that the model $p_\theta(x|z)$ defines both mapping from z to x domain and "topology" of the x space (how we measure distance between objects). However, approaches based on the Wasserstein distance are limited to continuous variables, while CoLLike is applicable both for discrete and continuous domains. Moreover, CoLLike provides a probabilistic basis for the choice of the cost function.

D OPTIMAL PAIRING BY COMBINATORIAL OPTIMIZATION

Having at hand log-likelihood values for all possible x_i z_j pairs, we are ready to find the optimal permutation. A naive way to do so is to evaluate the sum 3 for every possible permutation π. Despite we need only to sum different pre-computed values, the search space for π is tremendous $N!$. However, we can cast this problem to a combinatorial optimization one. Following Papadimitriou & Steiglitz (1982), the assignment problem is stated as follows

$$\begin{aligned} \text{minimize} \quad & \sum_{i,j} c_{i,j} a_{i,j} \\ \text{subjected to} \quad & \sum_{i} a_{i,j} = 1 \quad j = 0, ..., N \\ & \sum_{j} a_{i,j} = 1 \quad i = 0, ..., N \\ & a_{i,j} \in \{0, 1\} \end{aligned}$$

where $c_{i,j}$ is the cost of picking the element i, j and $a_{i,j}$ is the indicator variable. The constrains of this problem define a set of permutation matrices. By choosing the cost to be negative log-likelihood and replacing indicator variable with permutation we end up with CoLLike objective. This combinatorial optimization problem can be solved efficiently with Hungarian algorithm Kuhn (1955) with complexity of $\mathcal{O}(N^3)$.

E MODELS DESCRIPTION

E.1 NORMALIZING FLOWS

We use Glow-like normalizing flow for all image experiments. We choose the learning rate by starting from $1e^{-2}$ and gradually decrease it until there is no instabilities during training. No extensive learning rate search was done. Below we provide details on the model parameters.

CIFAR model:

- architecture: Glow
- number of flows per scale: 7
- flow coupling: affine
- coupling net: ResNet, 3 blocks, hidden size 96
- permutation flow: invertible 1×1 convolution with lower-upper factorization
- normalization flow: ActNorm
- number of scales: 3
- scale factor: 2 height, 2 width

MNIST model (same as CIFAR except the number of scales):

- architecture: Glow
- number of flows per scale: 7
- flow coupling: affine
- coupling net: ResNet, 3 blocks, hidden size 96
- permutation flow: invertible 1×1 convolution with lower-upper factorization
- normalization flow: ActNorm
- number of scales: 2
- scale factor: 2 height, 2 width

During training we use marginal likelihood to validate both MaL and CoLLike. We did not search over possible optimizers, and use Adam (Kingma & Ba, 2015) with default parameters. The validation split is chosen to be 0.05 because no significant variations of the likelihood were observed during training. To summarize, we use the following parameters both for CoLLike and MaL:

- epochs: 256
- learning rate: $5e^{-5}$ - MNIST; $2e^{-4}$ - CIFAR
- batch size: 64
- validation part of the training set: 0.05
- validation criterion: marginal likelihood
- optimizer: Adam, $\beta = (0.9, 0.999); \epsilon = 1e^{-8}$

As data pre-processing step, we used only dequantization with uniform noise, with the range equal to the quantization step.

E.2 Transformer

We used simple two-layer transformer across our experiments. The model description:

- number of layers: 2
- hidden size: 128
- feedforward dimension: 128
- embedding dimension: 128
- number of attention heads: 4
- number of embeddings: 4000

The training details are similar to CIFAR configuration:

- epochs: 256
- learning rate: $2e^{-4}$ - CIFAR
- batch size: 64
- validation part of the training set: 0.05
- validation criterion: marginal likelihood
- optimizer: Adam, $\beta = (0.9, 0.999); \epsilon = 1e^{-8}$

In pre-processing step, we truncate the sequences longer than 192 tokens. Truncation affects less than 0.3% of the samples. Nevertheless, the tokenizer is trained on the full-length sequences. Data pre-processing can be summarized as follows:

- maximum length truncation: 192
- BPE tokenization
- vocabulary size: 4000