[CoLike]
Complete Likelihood Objective
for Latent Variable Models

Mikhail Arkhipov
Mary Vikhreva

Generative Modelling Supervised Setting

Given a model

$$p_\theta(x, z) = p(z) p_\theta(x|z)$$

and a sample of data

$$\{(x_1, z_1), ..., (x_N, z_N)\}$$

Parameters θ can be obtained by maximization of complete log-likelihood of the data

$$\theta^* = \arg\max_\theta \mathscr{L}_c(\theta) = \arg\max_\theta \sum_i \log p_\theta(x_i, z_i)$$

Generative Modelling **Latent Variable** Setting

Given a model

$$p_\theta(x, z) = p(z) p_\theta(x|z)$$

and an incomplete sample of data

$$\{x_1, ..., x_N\}$$

Parameters θ can be obtained by maximization of **marginal** log-likelihood of the data

$$\theta^* = \arg\max_\theta \sum_i \log p_\theta(x_i) = \arg\max_\theta \sum_i \log \int_z p_\theta(x_i, z) dz$$

Additional Assumptions to Latent Variable Setting

Given a model

$$p_\theta(x, z) = p(z)p_\theta(x|z)$$

and an incomplete sample of data

$$\{x_1, ..., x_N\}$$

We add

a sample $\{z_1, ..., z_N\}$ from prior $p(z)$

and say that

$\{z_1, ..., z_N\}$ are pairs for $\{x_1, ..., x_N\}$ with unknown order

Generative Modelling with **Complete Likelihood** Objective

Given a model

$$p_\theta(x,z) = p(z)p_\theta(x|z)$$

and an incomplete sample of data

$$\{x_1,...,x_N\}$$

We add

a sample $\{z_1,...,z_N\}$ from prior $p(z)$

and say that

$\{z_1,...,z_N\}$ are pairs for $\{x_1,...,x_N\}$ with unknown order

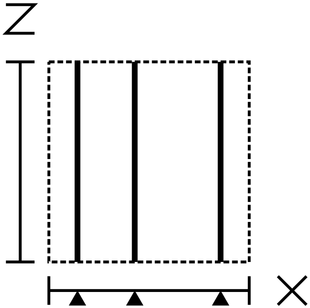
Under these assumptions Complete Likelihood is known up to a permutation

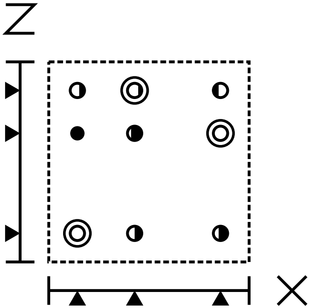
$$\mathcal{L}_c(\theta,\pi) = \sum_i \log p_\theta(x_i, z_{\pi(i)})$$

This allows us to perform Maximum Likelihood estimation over both permutation and parameters

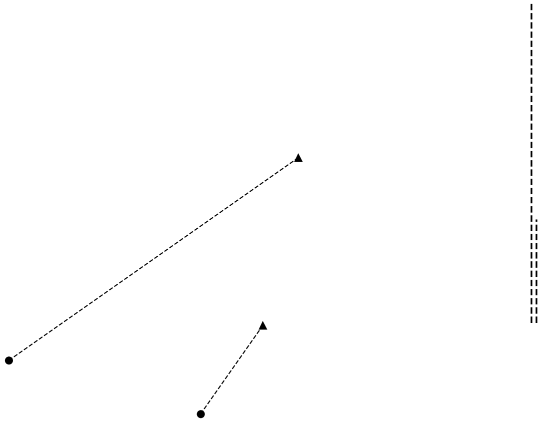
$$\theta^*,\pi^* = \arg\max_{\theta,\pi} \mathcal{L}_c(\theta,\pi)$$

Marginal Likelihood vs. Complete Likelihood

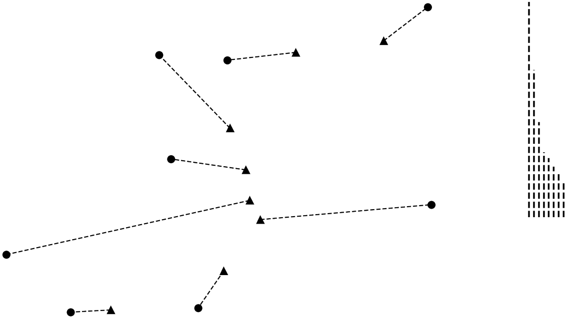




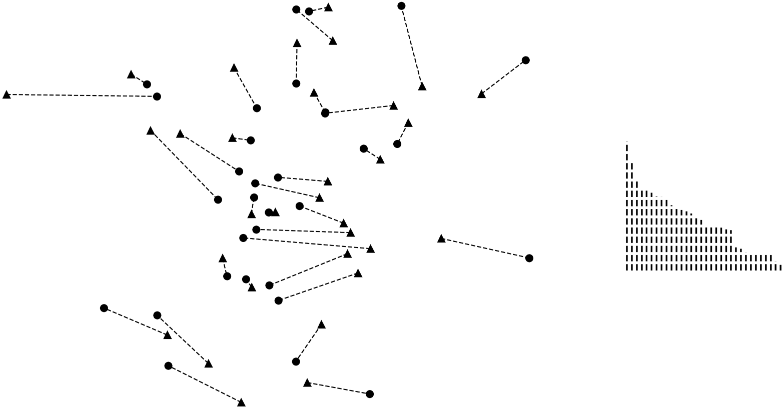
How strong is the assumption?



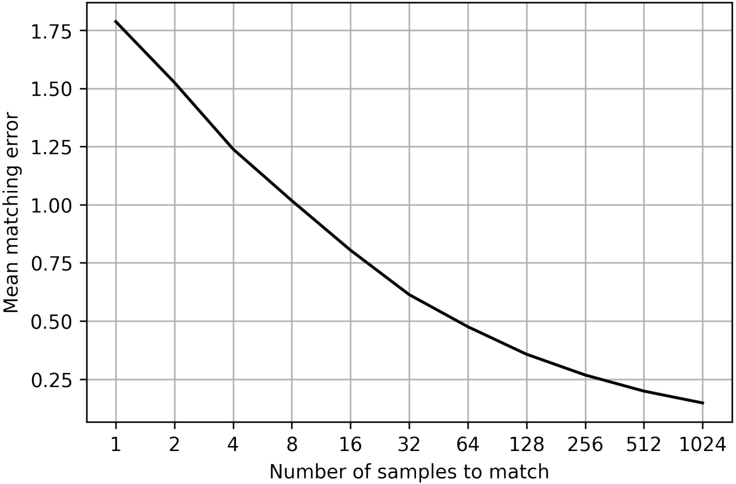
How strong is the assumption?



How strong is the assumption?



How strong is the assumption?



Links to Supervised Settings

Threre is a known supervised problem Broken Sample[1] or Linear Regression With Shuffled Labels[2]:

$$y_i = \mathbf{w}^T \mathbf{x}_{\pi(i)} + \varepsilon_i$$

This problem is common e.g. in tracking[3] and signal processing[4].

[1] [DeGroot et al. 1976] The Matching Problem for Multivariate Normal Data
[2] [Hsu et. al 2017] Linear Regression Without Correspondence
[3] [Bewley et al. 2016] Simple Online and Realtime Tracking
[4] [Haghighatshoar and Caire 2017] Signal Recovery from Unlabeled Samples

Link to Unsupervised Learning by Predicting Noise[5]

Setting:

data sample $\{x_1, ..., x_N\}$

sample from spherical prior $\{y_1, ..., y_N\}$

model f_θ that maps x to y domain

Objective:

$$\max_{\theta} \max_{\pi} \sum_{i} \mathbf{y_i}^T f_\theta(\mathbf{x}_{\pi(i)})$$

[5] [Bojanowski and Joulin 2017] Unsupervised Learning by Predicting Noise

Links to Optimal Transport

The Optimal Transport task for matching distributions P_X and P_Y is stated as

$$W_c(P_X, P_Y) = \inf_{\Gamma(P_X, P_Y)} \mathbb{E}_{X,Y \sim \Gamma}[c(X, Y)]$$

where $\Gamma(P_X, P_Y)$ is a family of joint distributions with marginals P_X and P_Y.
This problem is equivalent to[6]

$$W_c(P_X, P_Y) = \inf_{\substack{Q(Z|X): \\ Q_Z = P_Z}} \mathbb{E}_{X \sim P_X} \mathbb{E}_{Z \sim Q(Z|X)}[c(X, G(Z))]$$

where Q_Z is the marginal of $Q(Z|X)$, while G and $Q(Z|X)$ are a deterministic functions

[6][Tolstikhin et al. 2017] Wasserstein auto-encoders

Links to Optimal Transport

The objective is hard to optimize due to $Q_Z = P_Z$ constraint

$$W_c(P_X, P_Y) = \inf_{\substack{Q(Z|X): \\ \mathbf{Q_Z = P_Z}}} \mathbb{E}_{X \sim P_X} \mathbb{E}_{Z \sim Q(Z|X)} [c(X, G(Z))]$$

Wasserstein AutoEncoder suggests relaxation of original objective

$$\min_{G} \min_{Q(Z|X)} \mathbb{E}_{X \sim P_X} \mathbb{E}_{Z \sim Q(Z|X)} [c(X, G(Z)] + \beta D(Q_Z, P_Z)$$

where D is arbitrary discrepancy measure

Links to Optimal Transport

Consider encoderless case

$$W_c = \sup_{\Gamma \in \Pi(P_X, P_Y)} \mathbb{E}_{X,Y \sim \Gamma}\left[c(X,Y)\right] = \sup_{\Gamma \in \Pi(P_X, P_Z)} \mathbb{E}_{X,Z \sim \Gamma}\left[c(X,G(Z))\right]$$

When both P_X and P_Z are empirical distributions[7]

$$W_c = \sup_{\pi} \sum_i c\left(x_i, y_{\pi(i)}\right)$$

and for quadratic cost function

$$W_c = \sum_i ||x_i - y_{\pi(i)}||_2^2 = \sum_i ||x_i - G(z_{\pi(i)})||_2^2$$

[7] [Patrini et al. 2018] Sinkhorn AutoEncoders

Links to Optimal Transport

The function

$$W_c = \sum_i ||x_i - y_{\pi(i)}||_2^2 = \sum_i ||x_i - G(z_{\pi(i)})||_2^2$$

is identical to CoLike objective for Gaussian $p_\theta(x|z)$ and uniform prior $p(z)$

$$\begin{aligned}
\mathcal{L}_c &= \sup_\pi \sum_i \log p_\theta\left(x_i \middle| z_{\pi(i)}\right) p\left(z_{\pi(i)}\right) \\
&= \log c_z + \sup_\pi \sum_i \left(-\frac{d}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left\|x_i - G\left(z_{\pi(i)}\right)\right\|_2^2\right) \\
&= C + \sum_i \left\|x_i - G\left(z_{\pi(i)}\right)\right\|_2^2
\end{aligned} \tag{1}$$

Algorithm

Direct evaluation of CoLike objective

$$\mathcal{L}_c(\theta,\pi) = \sum_i \log p_\theta(x_i, z_{\pi(i)})$$

requires evaluations of $p_\theta(x_i, z_k)$ for every pair x_i and z_k for $i \in \{1,\ldots,N\}$ and $k \in \{1,\ldots,N\}$.
This amounts to N^2 evaluations of $p_\theta(x|z)$.
However, for non-autoregressive models, the neural network can be evaluated only N times, since decoder $p_\theta(x|z)$ takes only z as an input, while autoregressive models require x as an input.

Algorithm

Furthermore, finding optimal permutation requires solving combinatorial optimization problem.
The solution can be found with **Hungarian** algorithm with complexity $O(N^3)$.

Algorithm

Problem:

For large datasets, N^2 evaluations might be infeasible.

Solution:

Minibatch approximation for optimal permutation.

0. Sample z_i for each x_i
1. Sample minibatch of pairs (x_i, z_i)
2. Find optimal permutatin π^* for minibatch
3. Permute z_i in the training set according to π^*
4. Compute loss and perform gradient step for θ according to π^*
5. Go to step 1.

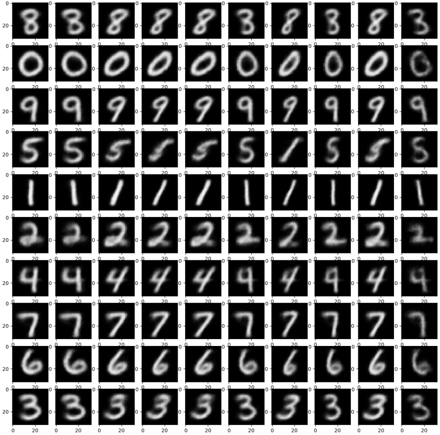
Discrete Latents out of the box

Setting:

- MNIST dataset
- Convolutional $p_\theta(x|z)$ from DCGAN
- 1 categorical latent with 10 classes
- 2 uniform continuous latents

Rows - distinct categorical latents.

Columns - random samples of continuos variables.



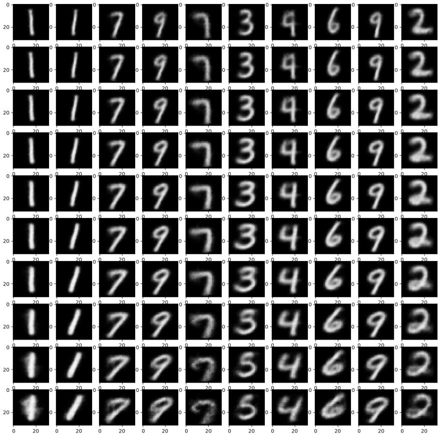
Discrete Latents out of the box

Setting:

- MNIST dataset
- Convolutional $p_\theta(x|z)$ from DCGAN
- 1 categorical latent with 10 classes
- 2 uniform continuous latents

Rows - traverse of one continuos latent.

Columns - samples of categorical and one continuous latents.



Likelihood on binarized MNIST

Setting:

- Statically binarized MNIST dataset
- 2 hidden layer MLP $p_\theta(x|z)$ with hidden dimensionality 500
- Gaussian latents
- RealNVP to approximate posterior after training
- Evidince is estimated using 1000 importnace weigthed samples from RealNVP posterior

Mehod	$dim(z)$	$p(x)$	Dataset size
VAE	2	-205.07	32
CoLike	2	-200.65	32
VAE	2	-171.43	256
CoLike	2	-170.04	256
VAE	2	-176.80	1024
CoLike	2	-151.18	1024
VAE	2	-152.61	50000
CoLike	2	-157.92	50000

Likelihood on binarized MNIST

Mehod	$dim(z)$	$p(x)$	Dataset size	Active Units
VAE	2	-152.61	50000	2
CoLike	2	-157.92	50000	2
VAE	8	-100.94	50000	8
CoLike	8	-108.36	50000	8
VAE	32	-93.80	50000	18
CoLike	32	-110.09	50000	32

A latent unit (a single dimension of z) is active when the variance of its expectation with respect to x is larger than 0.01[8]: $A_u > 0.01$ where $A_u = Cov_x(\mathbb{E}_{u \sim q_\phi(u|x)}[u])$

[8] [Burda et al. 2015] Importance weighted autoencoders

Language Modelling with Latent Variables

Setting:

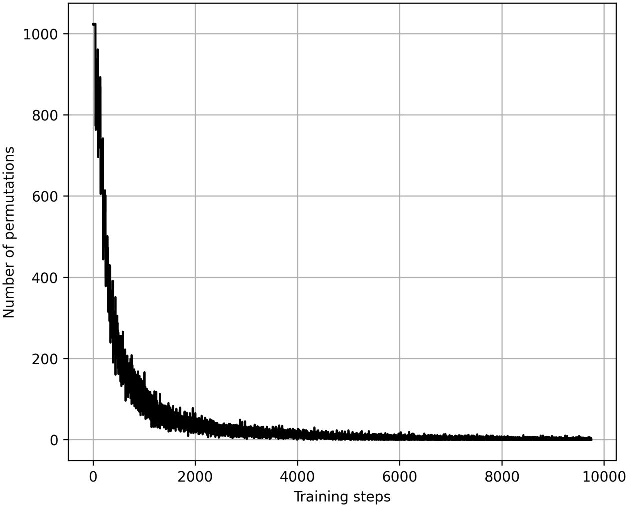
- SNLI dataset
- Single layer autoregressive unidirectional LSTM $p_\theta(x|z)$
- z is concatenated to each input
- LSTM hidden size - 512
- Gaussian latents
- $dim(z) = 32$

Mehod	$dim(z)$	PPL	Active Units
VAE	32	21.67	1
VAE FB[9]	32	22.00	32
CoLike	8	25.81	32

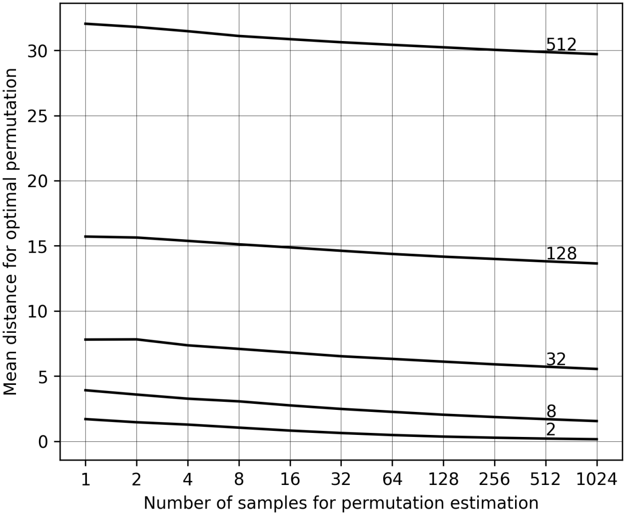
[9] VAE FB - VAE with free bits objective with KL constraint to be grater that 7.0

[Kingma et al. 2017] Improved Variational Inferencewith Inverse Autoregressive Flow

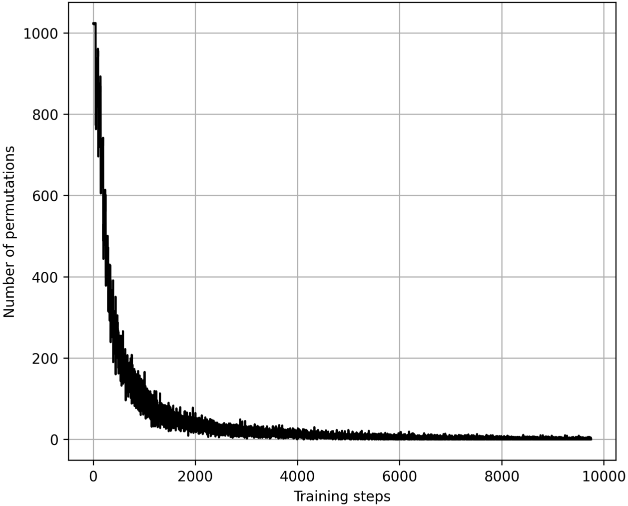
Challenges. Permutation saturation



Challenges. Dimensionality



Challenges. Permutation saturation



Conclusion

- New probabilistic objective for training latent variables models
- Approximate solution with partial permutation is proposed
- Promising results for discrete latentents and low-dimensional latents
- Robustness to posterior collapse for autoregressive models

SPASIBO