DVAE#: Discrete Variational Autoencoders with Relaxed Boltzmann Priors

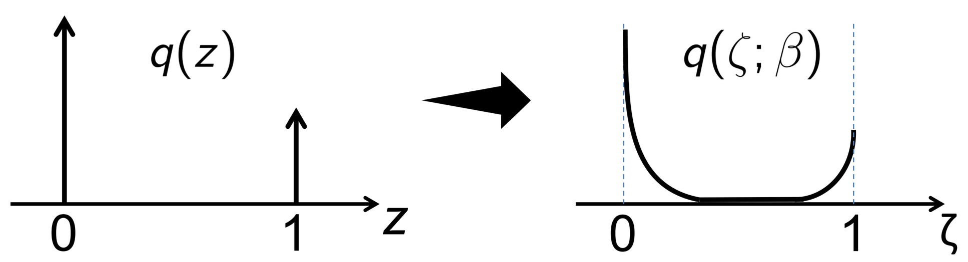
Arash Vahdat*, Evgeny Andriyash*, and William G. Macready

MOTIVATION

- Binary latent variable models have applications in topic modeling, semi-supervised learning, reinforcement learning
- Boltzmann machine as powerful distributions on binary vars.
- Variational autoencoders with Boltzmann priors:

$$\log p(x) \geq \mathbb{E}_{q(z|x)}[\log p(x|z)] - \mathrm{KL}(q(z|x)||p(z))$$

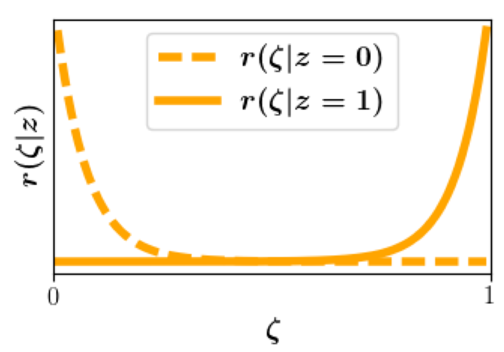
- Reparameterization doesn't work for binary variables
- Continuous relaxation of binary random:

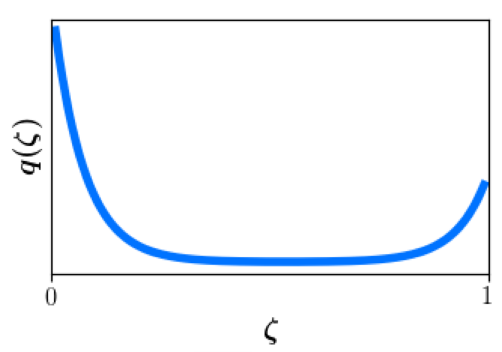


- We propose continuous relaxations for Boltzmann priors

OVERLAPPING TRANSFORMATIONS

- Use a mixture of two overlapping transformations:





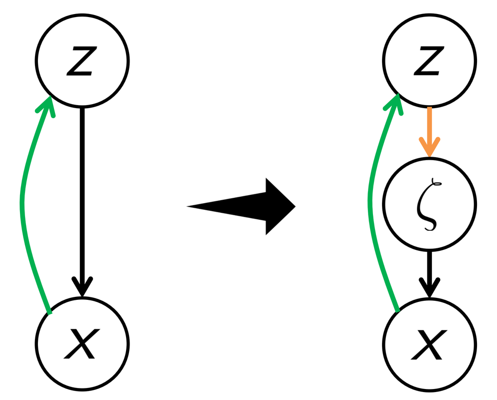
$r(\zeta|z=0) \propto e^{-\beta\zeta}$ $\qquad$ $q(\zeta) = \sum_{z\in\{0,1\}} q(z) r(\zeta|z)$

$r(\zeta|z=1) \propto e^{\beta(\zeta-1)}$ $\qquad$ $q(\zeta) \to q(z)$ as $\beta \to \infty$

- Inverse CDF of $q(\zeta)$ has a closed form for mixture of exponential and mixture of logistic distributions

UNDIRECTED PRIOR – ELBO (DVAE++)

$$p(z) = \frac{1}{Z} e^{a^T z + z^T W z} \qquad q(z|x) = \prod_i q_i(z_i|x)$$

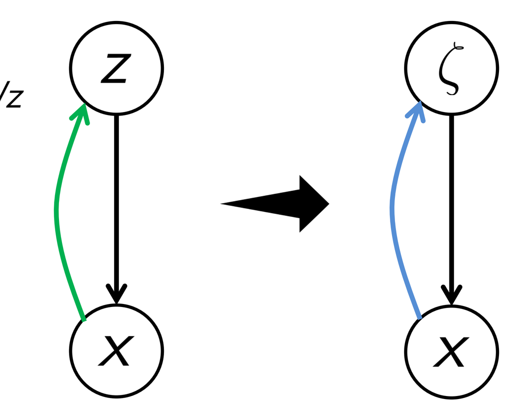


$$r(\zeta|z) = \prod_i r(\zeta_i|z_i) \qquad q(z,\zeta) = q(z) r(\zeta|z)$$

$$\mathcal{L}(x) = \mathbb{E}_{q(\zeta|x)}[\log p(x|\zeta)] - \mathrm{KL}(q(z,\zeta|x)||p(z,\zeta))$$
$$= \mathbb{E}_{q(\zeta|x)}[\log p(x|\zeta)] + \mathrm{H}(q(z|x)) - \mathbb{E}_{q(\zeta|x)}[\mathrm{H}(q(z|x,\zeta)||p(z)]$$

UNDIRECTED PRIOR – IW BOUND

$$p(z) = \frac{1}{Z} e^{a^T z + z^T W z} \qquad q(z|x) = \prod_i q_i(z_i|x)$$



$$p(\zeta) = ? \qquad q(\zeta|x) = \prod_i q_i(\zeta_i|x)$$

$$\log p(x) \geq \mathcal{L}_K(x) = \mathbb{E}_{\zeta^{(k)}\sim q(\zeta|x)}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p(\zeta^{(k)})p(x|\zeta^{(k)})}{q(\zeta^{(k)}|x)}\right)\right] \geq \mathcal{L}_1(x)$$

OVERLAPPING RELAXATIONS

- Factorial overlapping transformation: $r(\zeta|z) = \prod_i r(\zeta_i|z_i)$
- Overlapping relaxation of Boltzmann machines:

$$p(\zeta) = \sum_z p(z,\zeta) = \sum_z p(z) r(\zeta|z)$$
$$\log p(\zeta) = \log\left(\sum_z p(z) r(\zeta|z)\right) = \log\left(\sum_z e^{-E_\theta(z) + b(\zeta)^T z + C(\zeta)}\right) - \log Z_\theta$$

- A mean-field distribution approximates the value and gradient of the first term

GAUSSIAN INTEGRAL RELAXATION

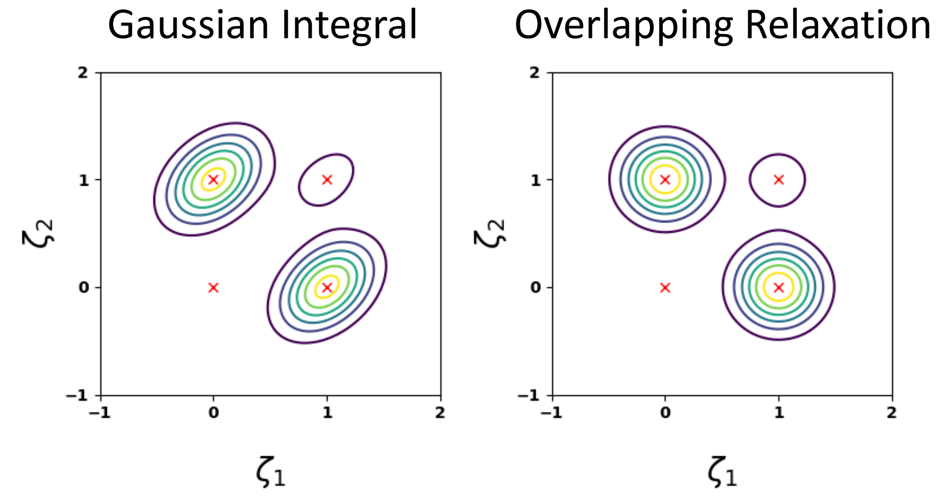
- The Gaussian integral trick: $r(\zeta|z) = \mathcal{N}(\zeta|z, (W+\beta I)^{-1})$
- The pairwise terms on z are removed in the joint:

$$p(z,\zeta) \propto e^{-\frac{1}{2}\zeta^T(W+\beta I)\zeta + z^T(W+\beta I)\zeta + (a - \frac{1}{2}\beta\mathbf{1})^T z}$$

- z is marginalized out easily
- Example:

$$a = \begin{bmatrix} 4.5 \\ 4.5 \end{bmatrix} \qquad W = \begin{bmatrix} 0 & -6 \\ -6 & 0 \end{bmatrix}$$

z_1	z_2	$p(z)$
0	0	0.00
0	1	0.45
1	0	0.45
1	1	0.10

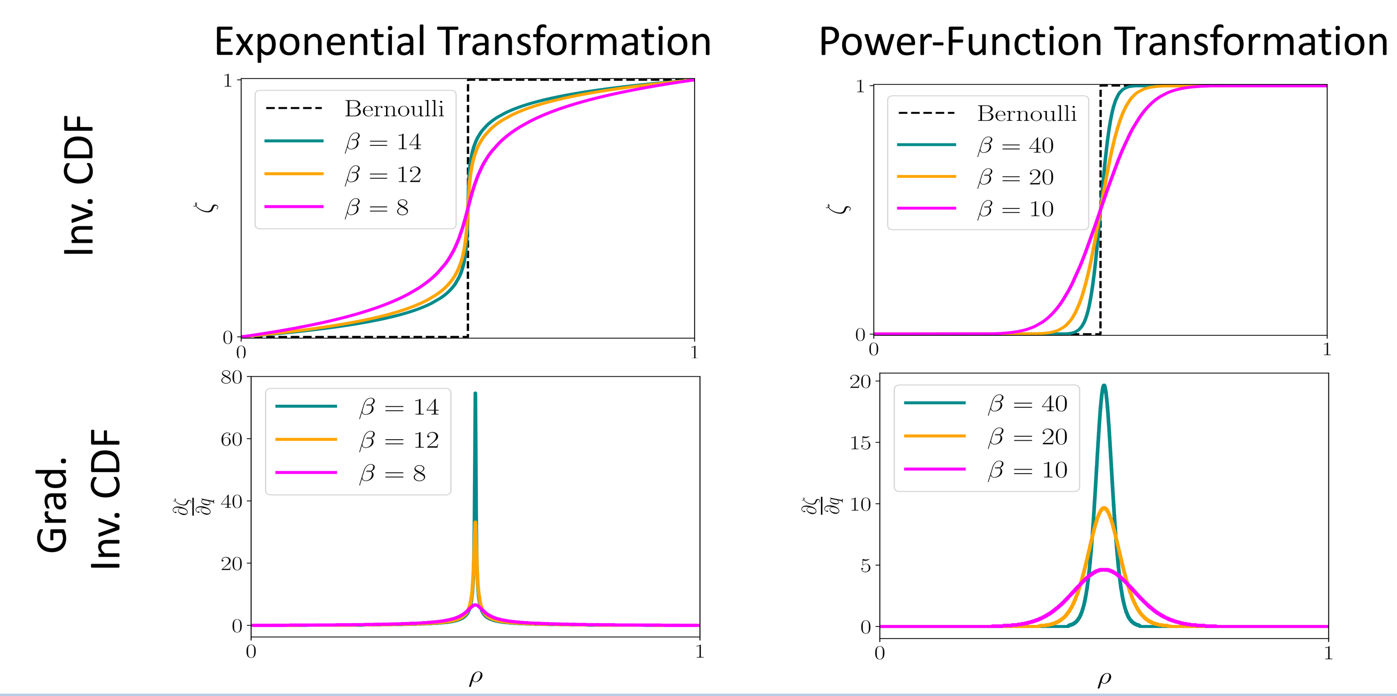


GENERALIZED OVERLAPPING TRANSFORMATIONS

- Reparameterized sampling from $q(\zeta|x)$ required computing inverse CDF in DVAE++

$$q(\zeta|x) = (1-q) r(\zeta|z=0) + q r(\zeta|z=1)$$
$$CDF_{q(\zeta|x)}(\zeta) = (1-q) R(\zeta|z=0) + q R(\zeta|z=1) = \rho$$

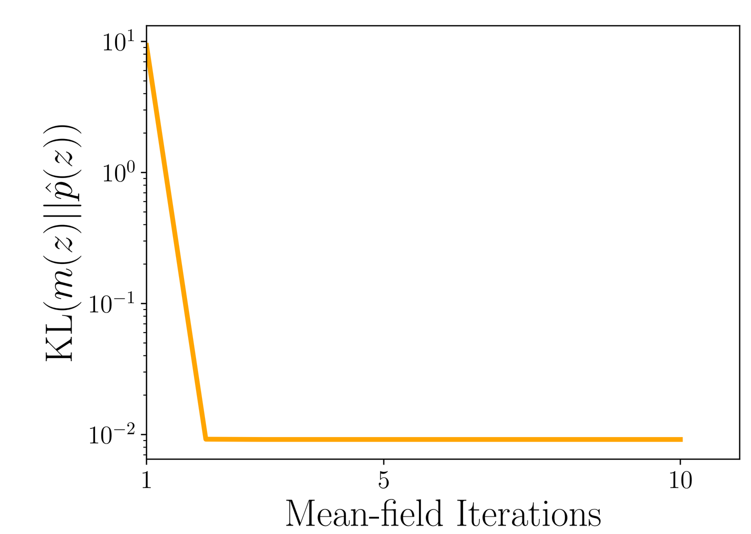
- Use implicit differentiation for the gradient of ζ w.r.t q
- Extend $r(\zeta|x)$ to Normal and power-function dist.



EXPERIMENTS

	K	Variational Bound: DVAE Spike-Exp	Variational Bound: DVAE++ Exp	Variational Bound: DVAE++ Power	Importance Weighted Bound: DVAE# Guass. Int.	DVAE# Gauss.	DVAE# Exp	DVAE# Unifor-Exp	DVAE# Power
MNIST	1	83.97	84.15	83.62	84.30	84.35	83.96	83.54	**83.37**
	5	83.74	84.85	83.57	83.68	83.61	83.70	83.33	**82.99**
	25	84.19	85.49	83.58	83.39	83.26	83.76	83.30	**82.85**
OMNIGLOT	1	103.10	101.34	100.42	102.07	102.84	100.38	**99.84**	**99.75**
	5	100.88	100.55	99.51	100.85	101.43	99.93	99.57	**99.24**
	25	100.55	100.31	99.49	100.20	100.45	100.10	99.59	**98.93**

Negative log-likelihood, lower is better



Code available at:
github.com/QuadrantAI/dvae

QuPA Sampling Library:
try.quadrant.ai/qupa

http://quadrant.ai arash@quadrant.ai