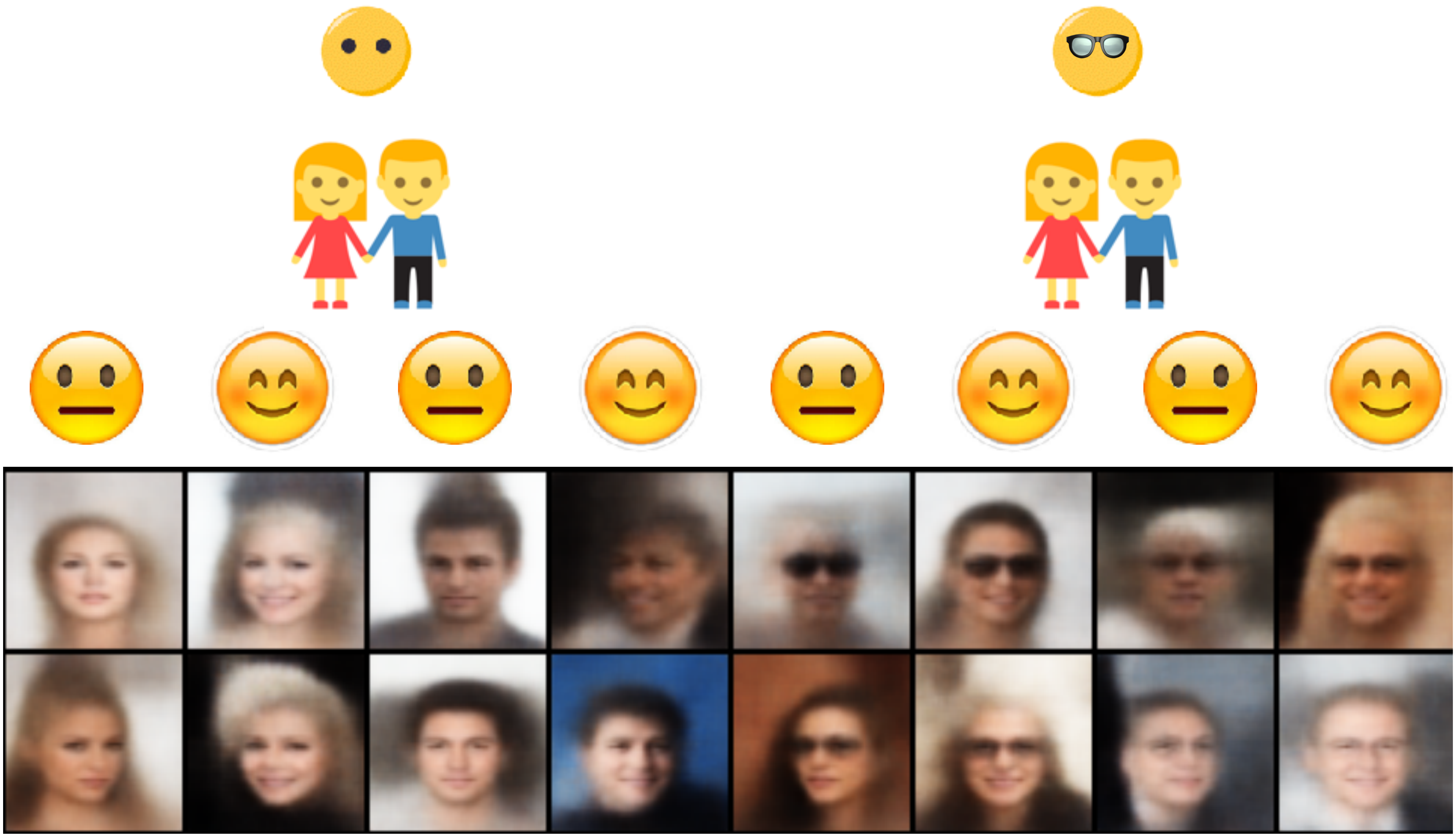
Direct Optimization through Argmax

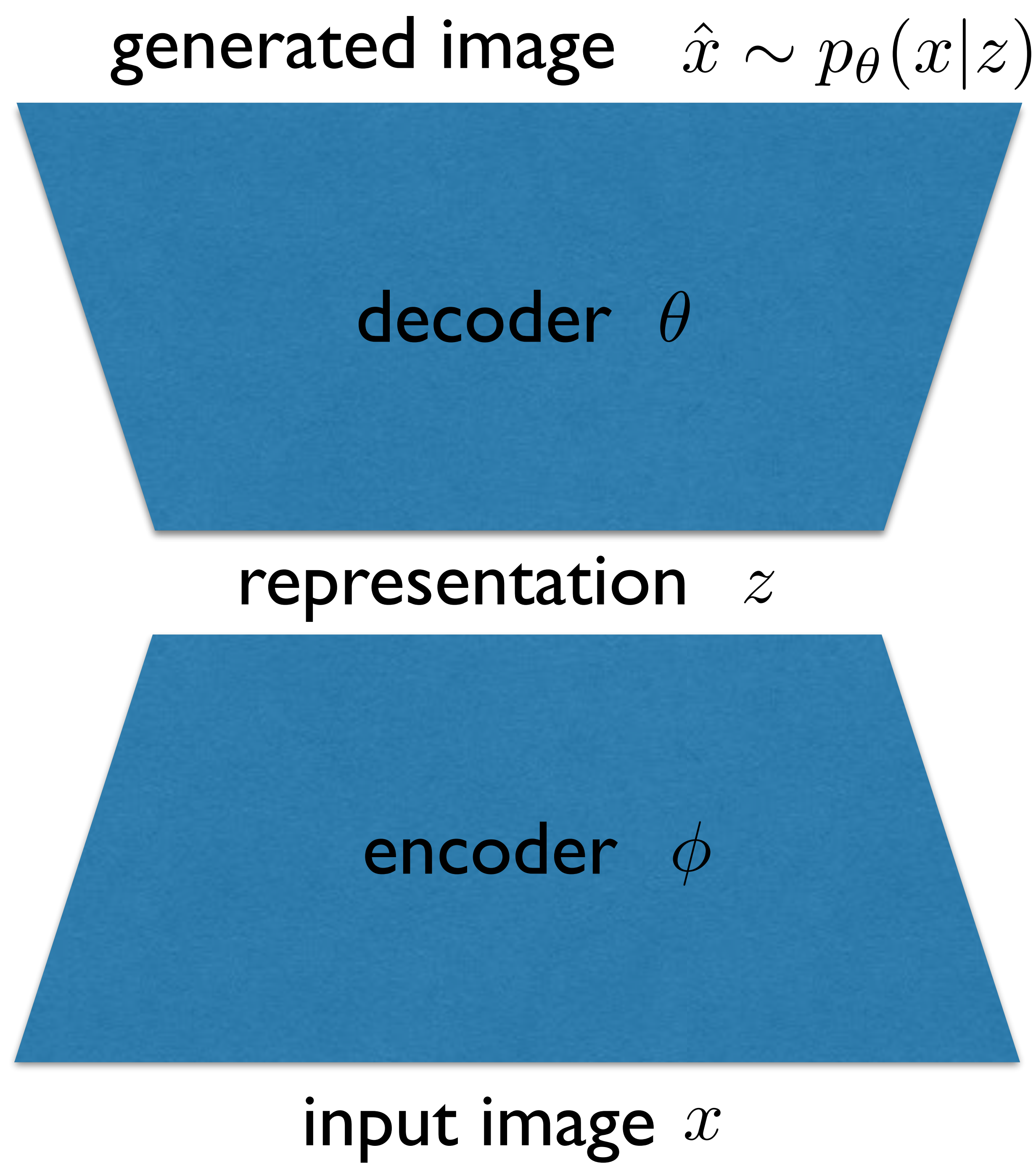
Guy Lorberbom, Andreea Gane, Tommi Jaakkola, Tamir Hazan

Generative learning



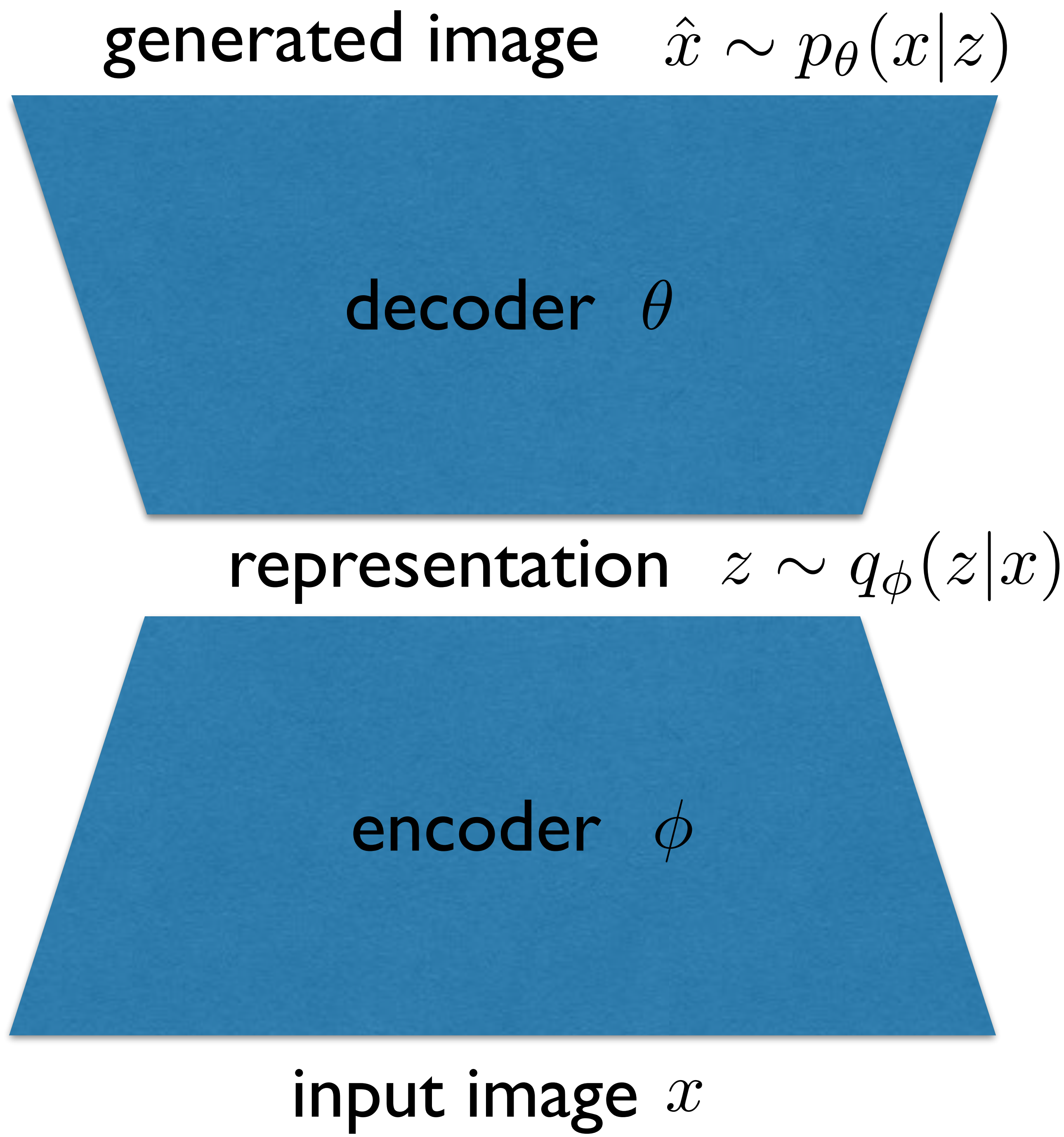
Variational Auto-Encoders

- Kingma and Welling, 2014; Reznde et al. 2014.



Variational Auto-Encoders

$$\log \frac{1}{p_\theta(x)} \leq \mathbb{E}_{z \sim q_\phi} \log \frac{1}{p_\theta(x|z)} + KL(q_\phi(z|x) || p_\theta(z))$$



Variational Auto-Encoders

$$\log \frac{1}{p_\theta(x)} \leq \mathbb{E}_{z \sim q_\phi} \log \frac{1}{p_\theta(x|z)} + KL(q_\phi(z|x) || p_\theta(z))$$

Variational Auto-Encoders

$$\log \frac{1}{p_\theta(x)} \leq \mathbb{E}_{z \sim q_\phi} \log \frac{1}{p_\theta(x|z)} + KL(q_\phi(z|x)||p_\theta(z))$$

$$p_\theta(x|z) = e^{\theta(x,z)}$$

$$q_\phi(z|x) = e^{\phi(x,z)}$$

Variational Auto-Encoders

$$\log \frac{1}{p_\theta(x)} \leq \mathbb{E}_{z\sim q_\phi} \log \frac{1}{p_\theta(x|z)} + KL(q_\phi(z|x)||p_\theta(z))$$

$$p_\theta(x|z) = e^{\theta(x,z)}$$

$$q_\phi(z|x) = e^{\phi(x,z)}$$

- discrete latent space $\quad \mathbb{E}_{z\sim q_\phi} \log p_\theta(x|z) = \sum_z e^{\phi(x,z)}\theta(x,z)$

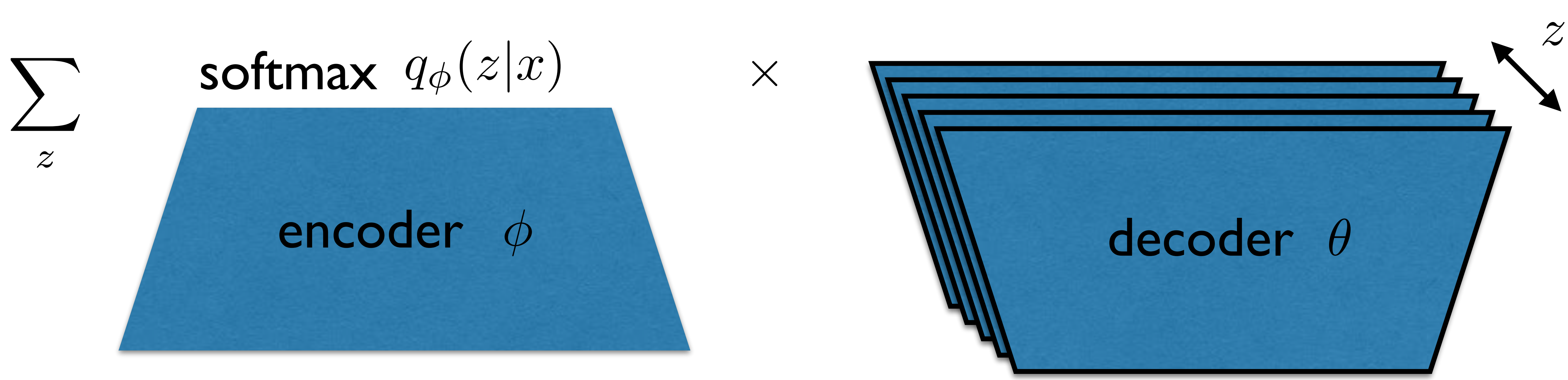
Variational Auto-Encoders

$$\log \frac{1}{p_\theta(x)} \leq \mathbb{E}_{z \sim q_\phi} \log \frac{1}{p_\theta(x|z)} + KL(q_\phi(z|x)||p_\theta(z))$$

$$p_\theta(x|z) = e^{\theta(x,z)}$$

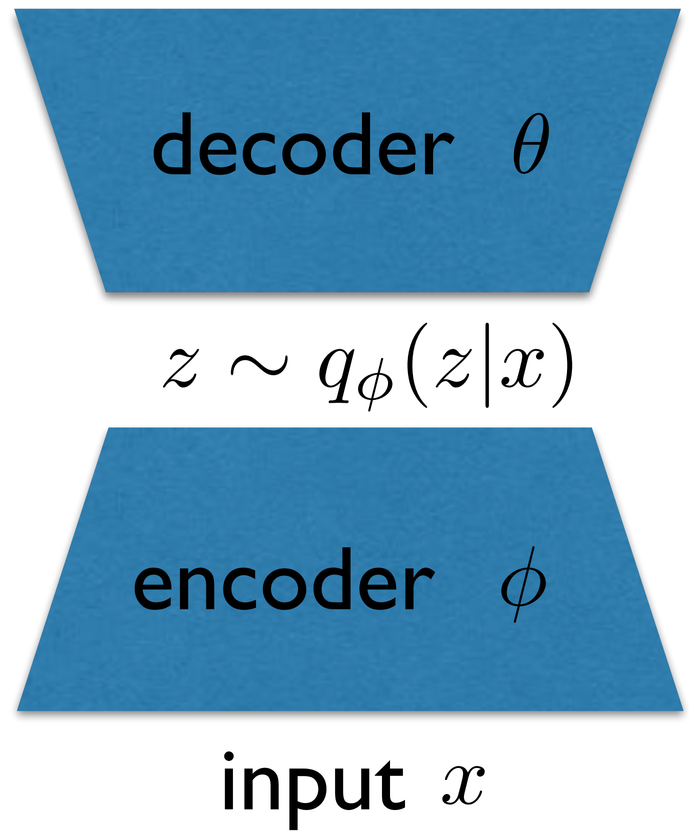
$$q_\phi(z|x) = e^{\phi(x,z)}$$

- discrete latent space $\mathbb{E}_{z \sim q_\phi} \log p_\theta(x|z) = \sum_z e^{\phi(x,z)} \theta(x,z)$

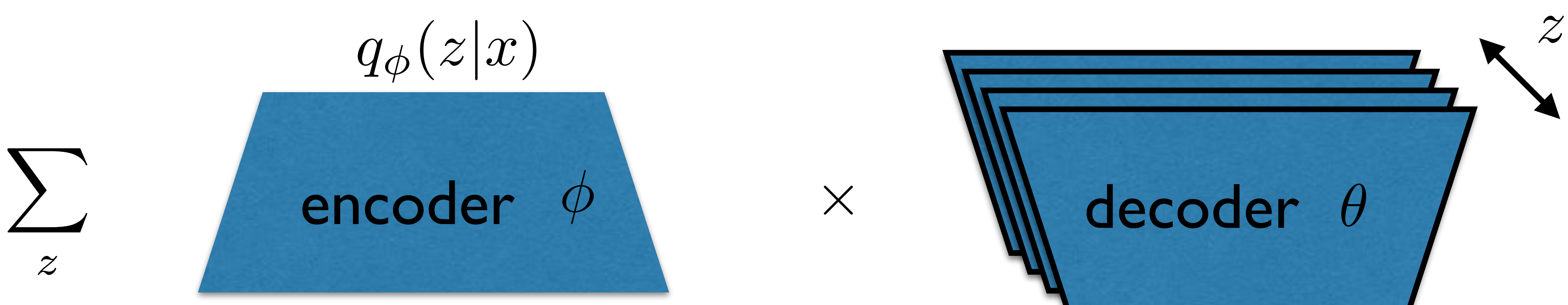


Variational Auto-Encoders

- continuous setting (with reparameterization)



- Discrete setting (without reparameterization)



Reperameterizing Discrete VAEs

$$q_\phi(z|x) = e^{\phi(x,z)}$$

- Theorem: (Fisher 1928, Gumbel 1953, McFadden 1973)

  Let $\gamma(z)$ be i.i.d. with Gumbel distribution with zero mean

$$G(t) \stackrel{def}{=} \mathbb{P}[\gamma(z) \leq t] = e^{-e^{-t+c}}$$

Reperameterizing Discrete VAEs

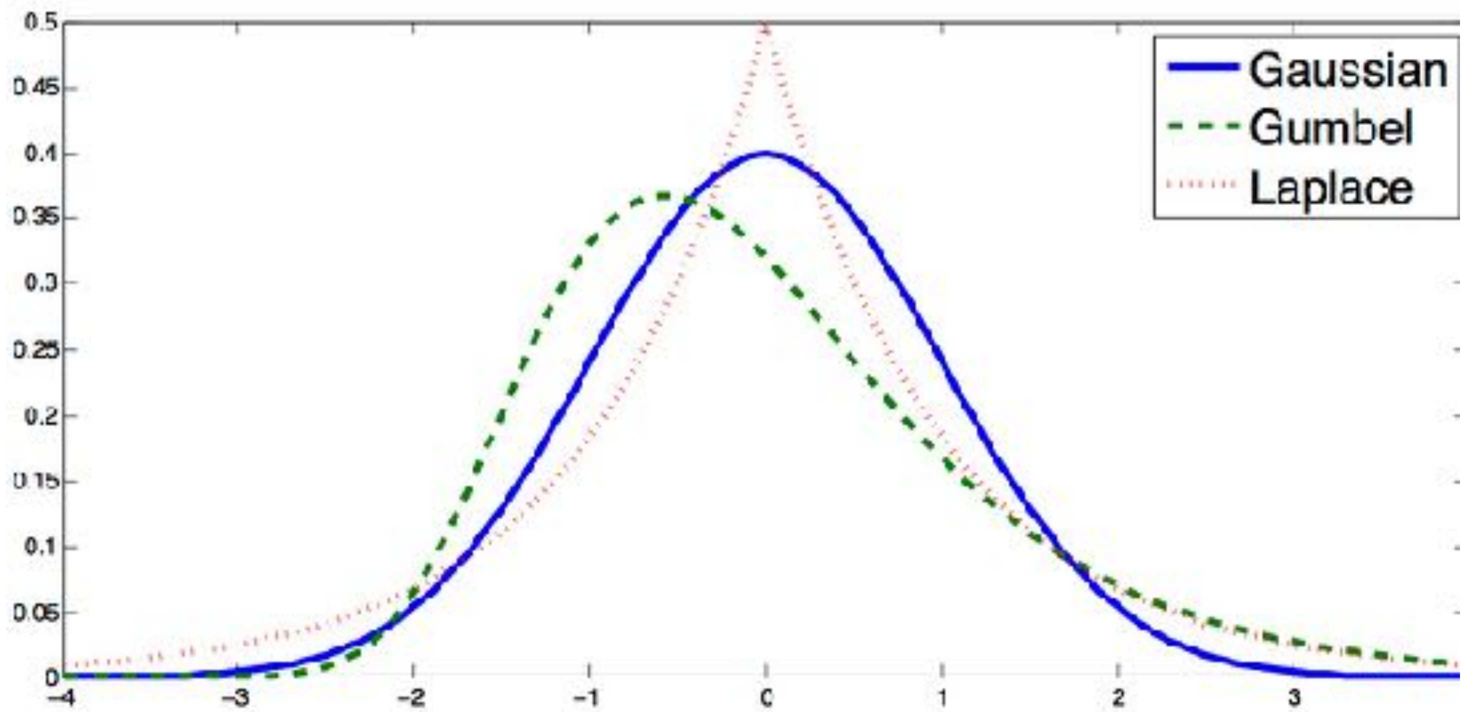
$$q_\phi(z|x) = e^{\phi(x,z)}$$

- Theorem: (Fisher 1928, Gumbel 1953, McFadden 1973)

Let $\gamma(z)$ be i.i.d. with Gumbel distribution with zero mean

$$G(t) \stackrel{def}{=} \mathbb{P}[\gamma(z) \leq t] = e^{-e^{-t+c}}$$

$$g(t) = G'(t)$$



Reparameterizing Discrete VAEs

$$q_\phi(z|x) = e^{\phi(x,z)}$$

- Theorem: (Fisher 1928, Gumbel 1953, McFadden 1973)

Let $\gamma(z)$ be i.i.d. with Gumbel distribution with zero mean

$$G(t) \stackrel{def}{=} \mathbb{P}[\gamma(z) \le t] = e^{-e^{-t+c}}$$

then

$$e^{\phi(x,z)} = \mathbb{P}_{\gamma \sim g}[z^{\phi+\gamma} = z]$$

$$z^{\phi+\gamma} = \arg\max_{\hat{z}} \{\phi(x,\hat{z}) + \gamma(\hat{z})\}]$$

Reparameterizing Discrete VAEs

- Gumbel-Argmax reparameterization

$$z^{\phi+\gamma} = \arg\max_{\hat{z}}\{\phi(x, \hat{z}) + \gamma(\hat{z})\}]$$

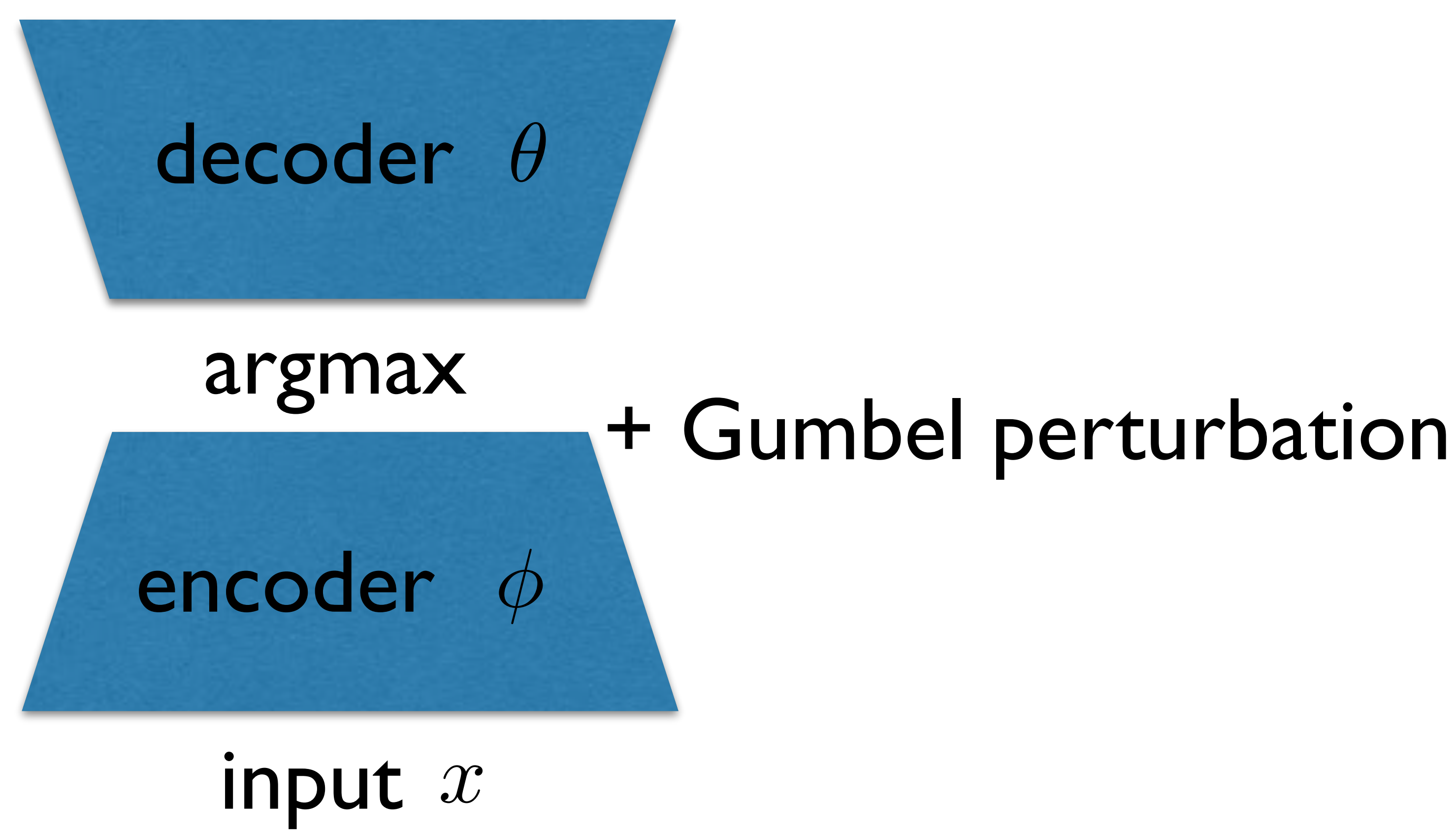
$$\mathbb{E}_{z\sim q_\phi} \log p_\theta(x|z) = \mathbb{E}_{\gamma\sim g}[\theta(x, z^{\phi+\gamma})]$$

Reparameterizing Discrete VAEs

- Gumbel-Argmax reparameterization

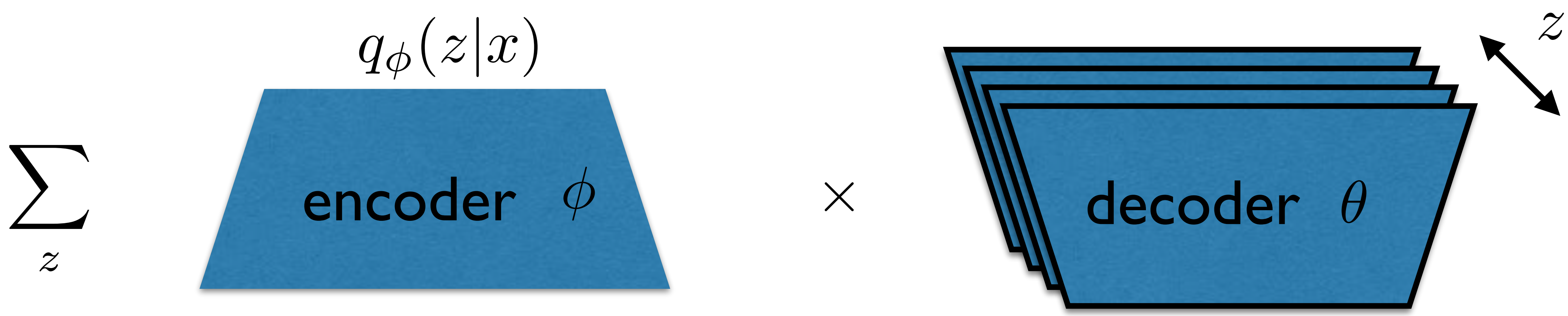
$$z^{\phi+\gamma} = \arg\max_{\hat{z}}\{\phi(x,\hat{z}) + \gamma(\hat{z})\}]$$

$$\mathbb{E}_{z\sim q_\phi} \log p_\theta(x|z) = \mathbb{E}_{\gamma\sim g}[\theta(x, z^{\phi+\gamma})]$$

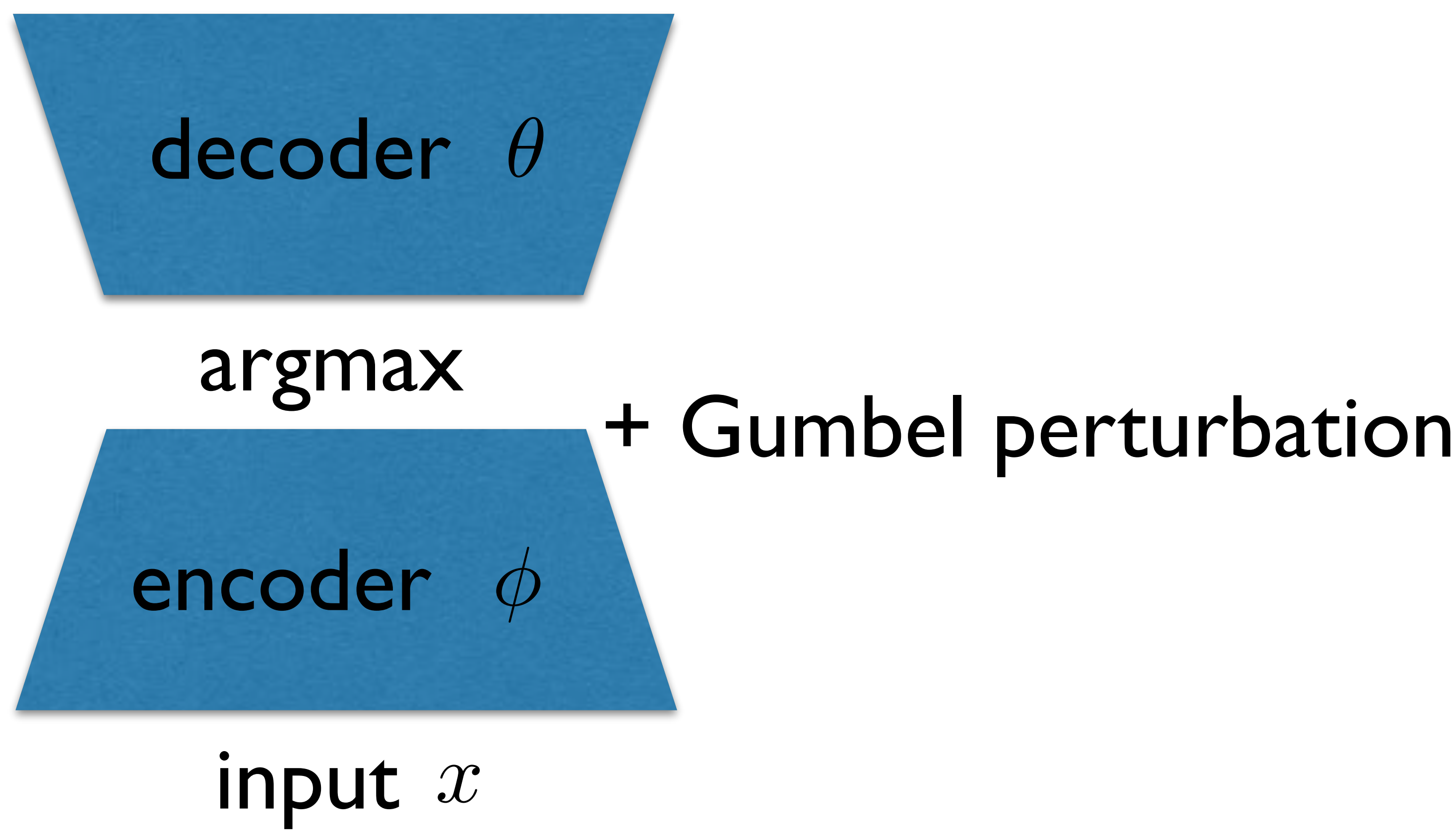


Reperameterizing Discrete VAEs

- discrete VAEs (without reperameterization)

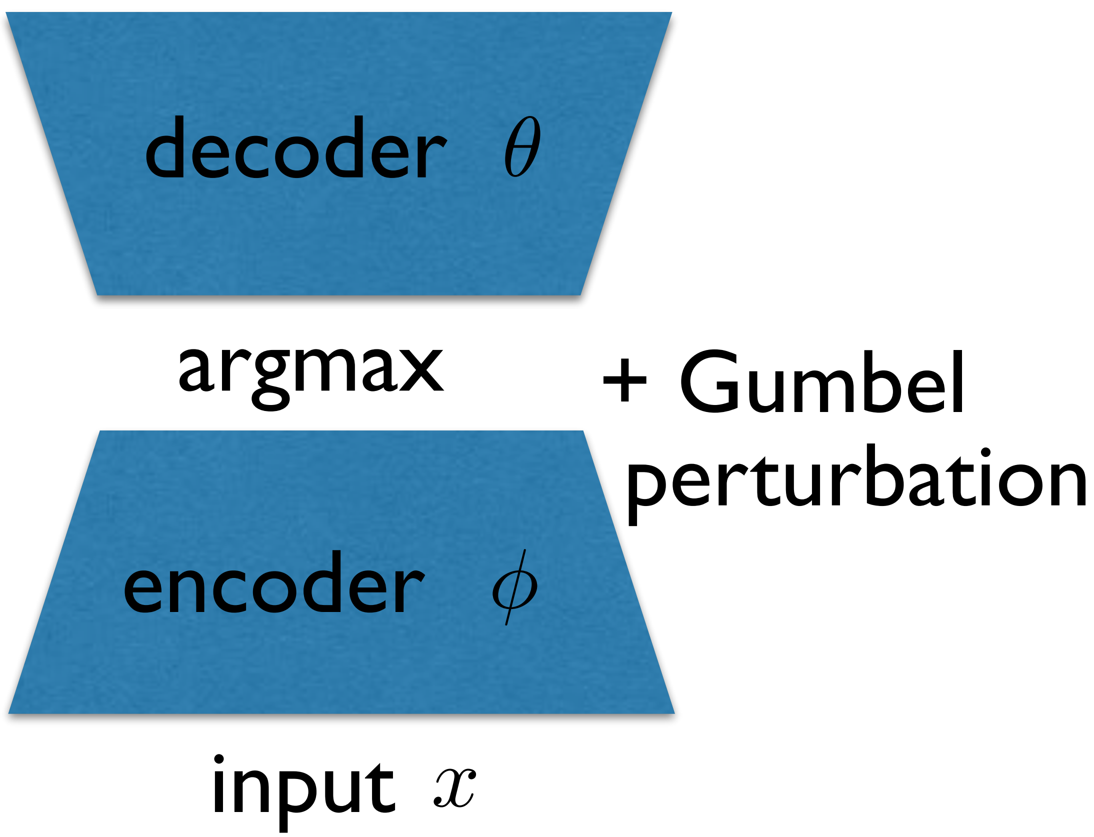


- discrete VAEs (with reperameterization)



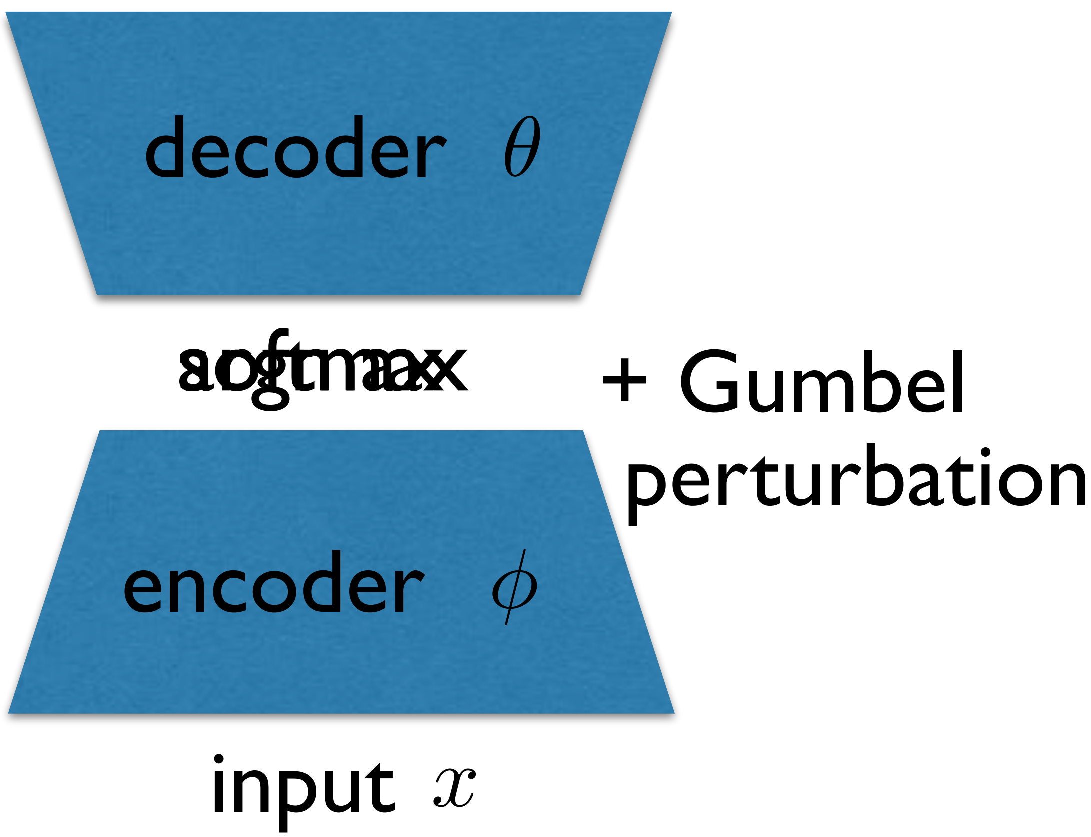
Reperameterizing Discrete VAEs

- Propagating gradients?

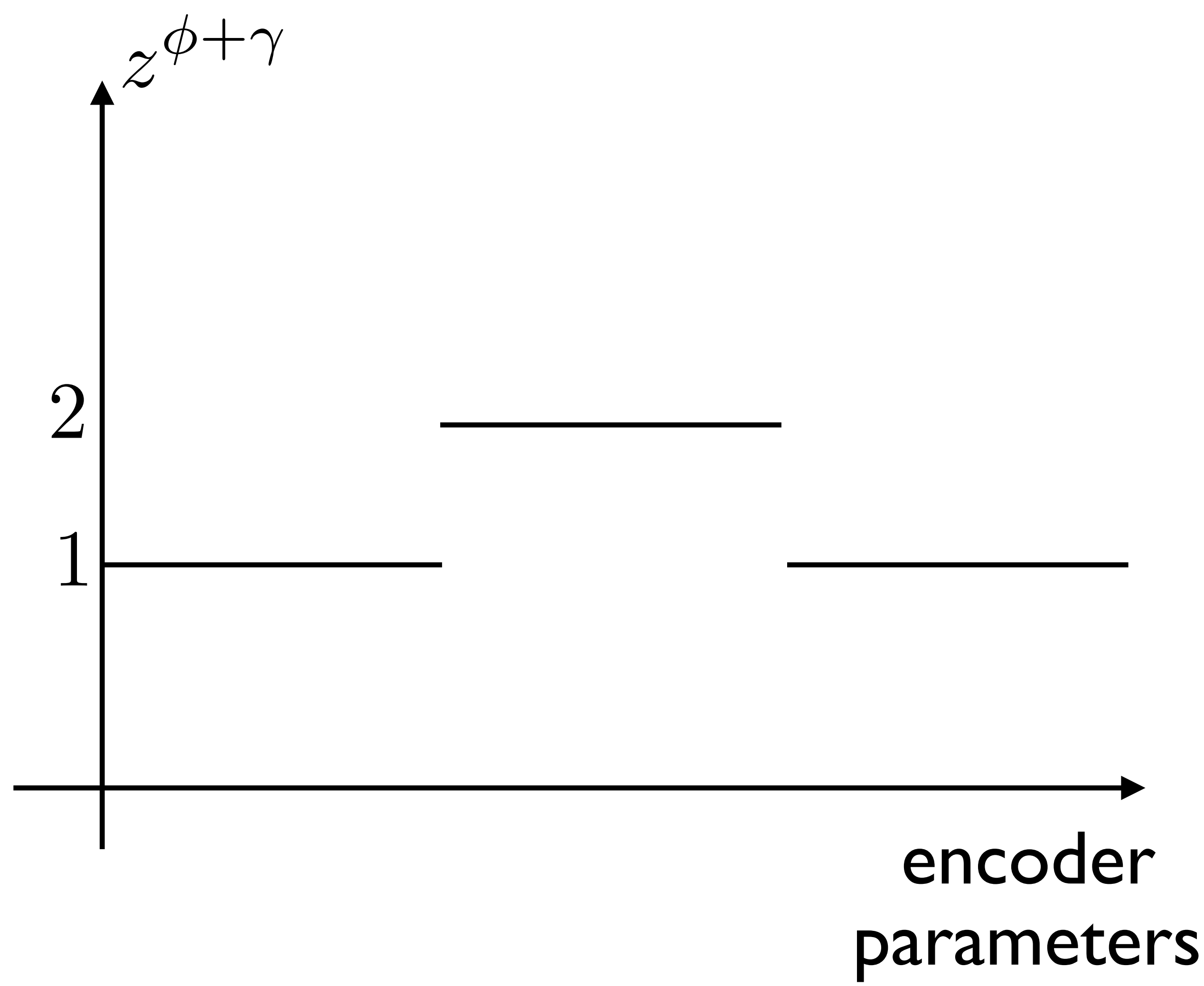


Reperameterizing Discrete VAEs

- Propagating gradients?

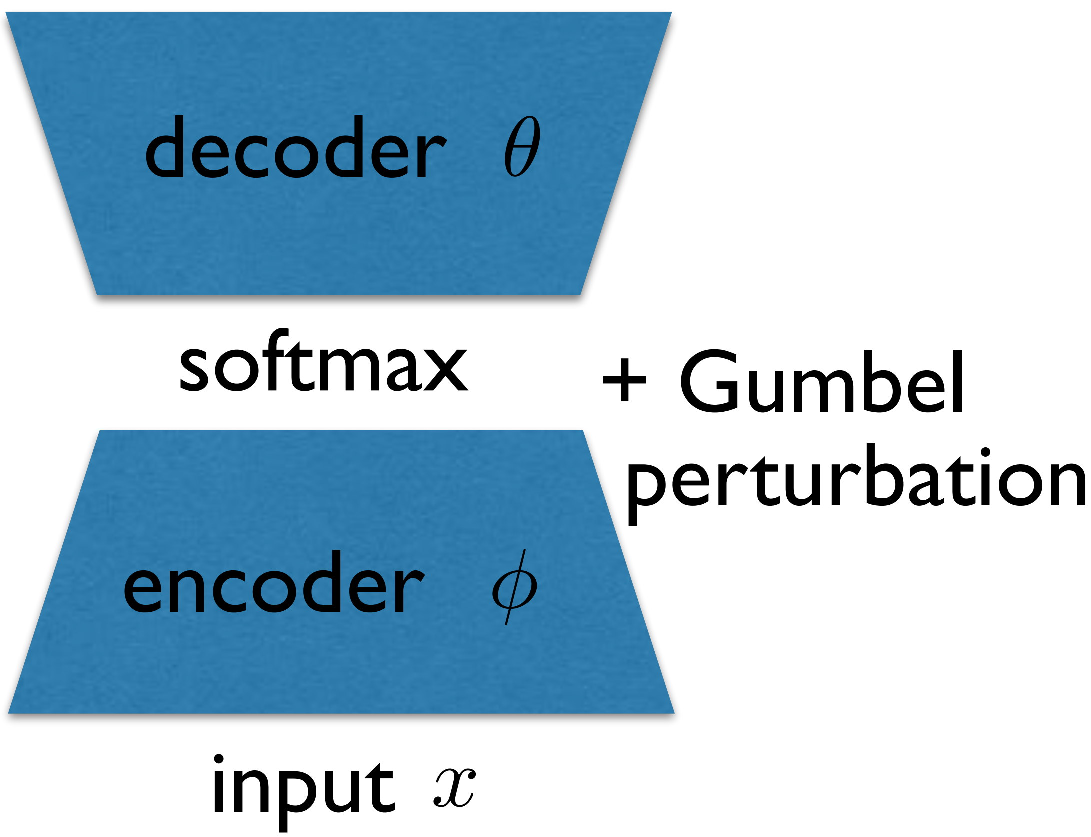


- The argmax derivative is not informative



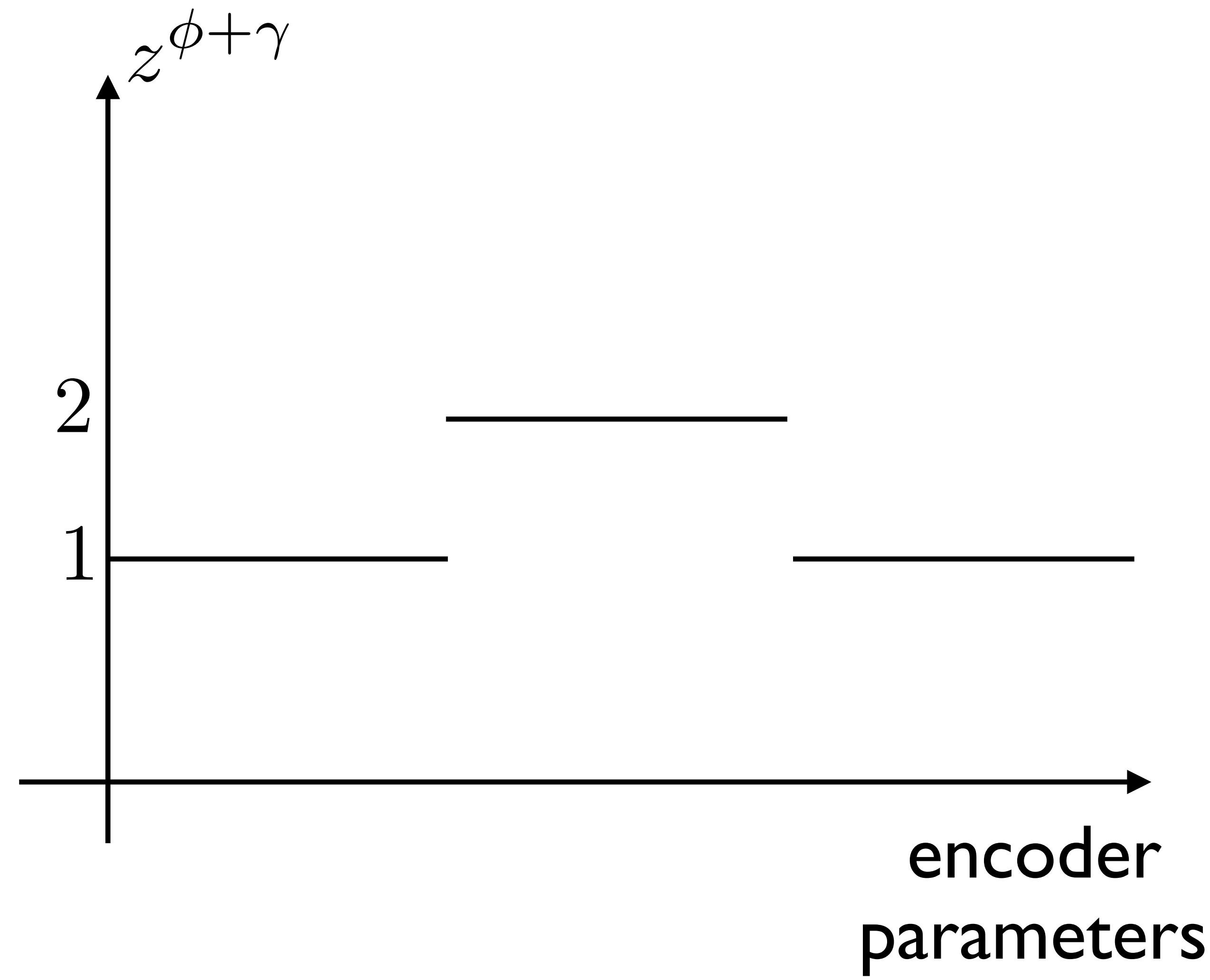
Reperameterizing Discrete VAEs

- Propagating gradients?



- Gumbel-Softmax
  Maddison et al., 2016
  Jang et al., 2016

- The argmax derivative is not informative



Propagating Gradients through Argmax

- Theorem

$$\nabla_w \mathbb{E}_\gamma[\theta(x, z^{\phi+\gamma})] = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \Big(\mathbb{E}_\gamma[\nabla_w \phi(x, z^{\epsilon\theta+\phi+\gamma}; w) - \nabla_w \phi(x, z^{\phi+\gamma}; w)] \Big)$$

Propagating Gradients through Argmax

- Theorem

$$\nabla_w \mathbb{E}_\gamma[\theta(x, z^{\phi+\gamma})] = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \Big(\mathbb{E}_\gamma [\nabla_w \phi(x, z^{\epsilon\theta+\phi+\gamma}; w) - \nabla_w \phi(x, z^{\phi+\gamma}; w)] \Big)$$

- Proof sketch

$G(w, \epsilon) = \mathbb{E}_\gamma[\max_{\hat{z}} \{\epsilon\theta(x, \hat{z}) + \phi(x, \hat{z}; w) + \gamma(\hat{z})\}]$ is smooth

$$\partial_w \partial_\epsilon G(w, 0) = \nabla_w \mathbb{E}_\gamma[\theta(x, z^{\phi+\gamma})]$$

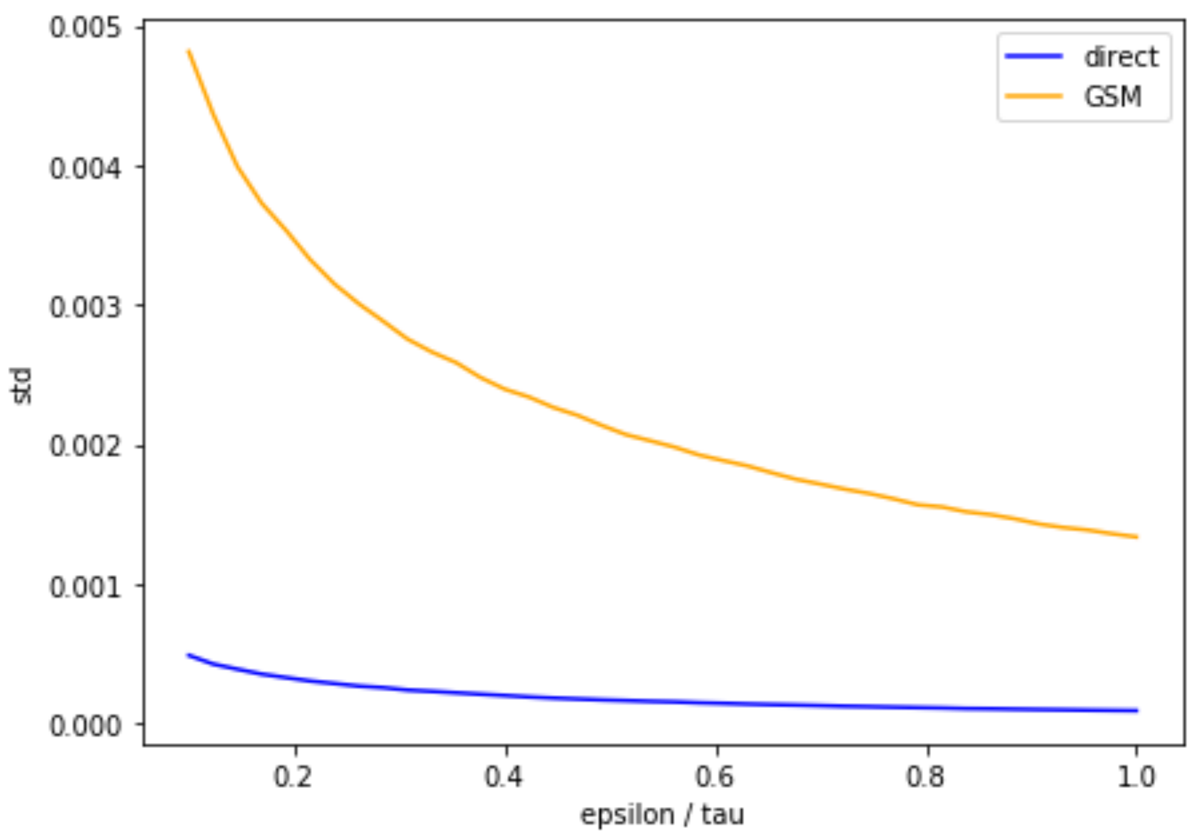
$$\partial_\epsilon \partial_w G(w, 0) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} (E_\gamma[\nabla_w \phi(x, z^{\epsilon\theta+\phi+\gamma}; w) - \nabla_w \phi(x, z^{\phi+\gamma}; w)])$$

$$\partial_w \partial_\epsilon G(w, \epsilon) = \partial_\epsilon \partial_w G(w, \epsilon)$$

Built-in control variates

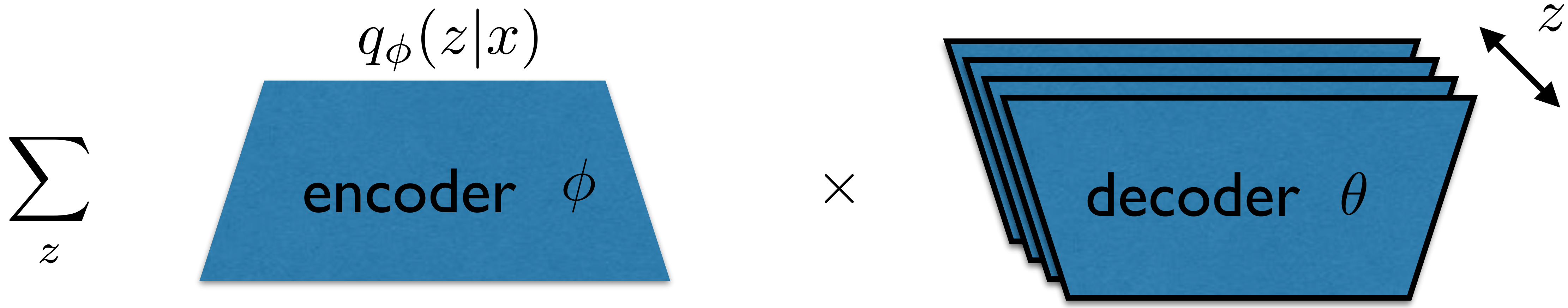
$$\nabla_w \mathbb{E}_\gamma[\theta(x, z^{\phi+\gamma})] = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \Big(\mathbb{E}_\gamma[\nabla_w \phi(x, z^{\epsilon\theta+\phi+\gamma}; w) - \nabla_w \phi(x, z^{\phi+\gamma}; w)] \Big)$$

- if $e^{\phi(x,z)} = \mathbb{P}_{\gamma \sim g}[z^{\phi+\gamma} = z]$ then $\mathbb{E}_\gamma[\nabla_w \phi(x, z^{\phi+\gamma}; w)] = 0$

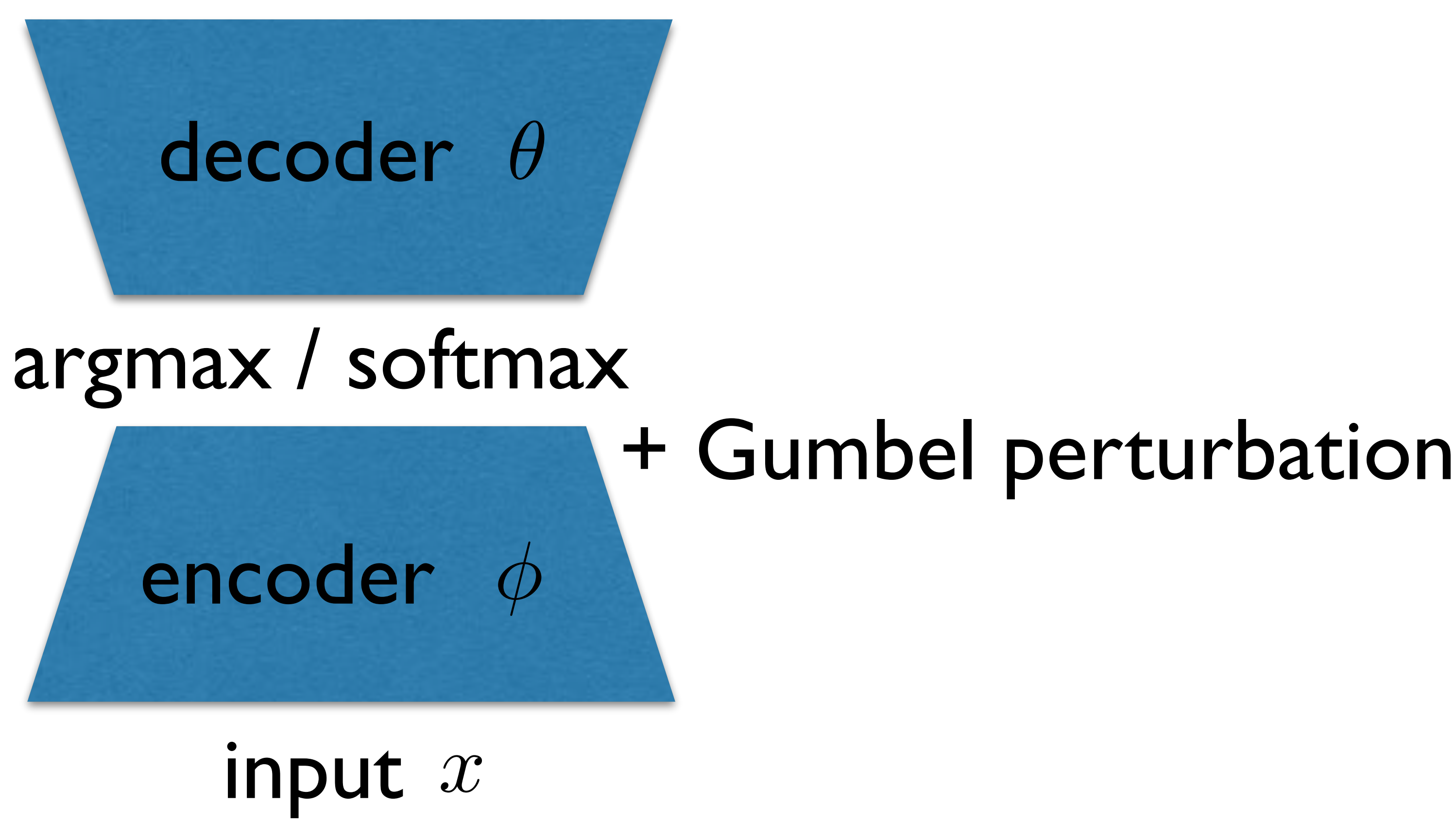


Comparing to

- Unbiased gradient



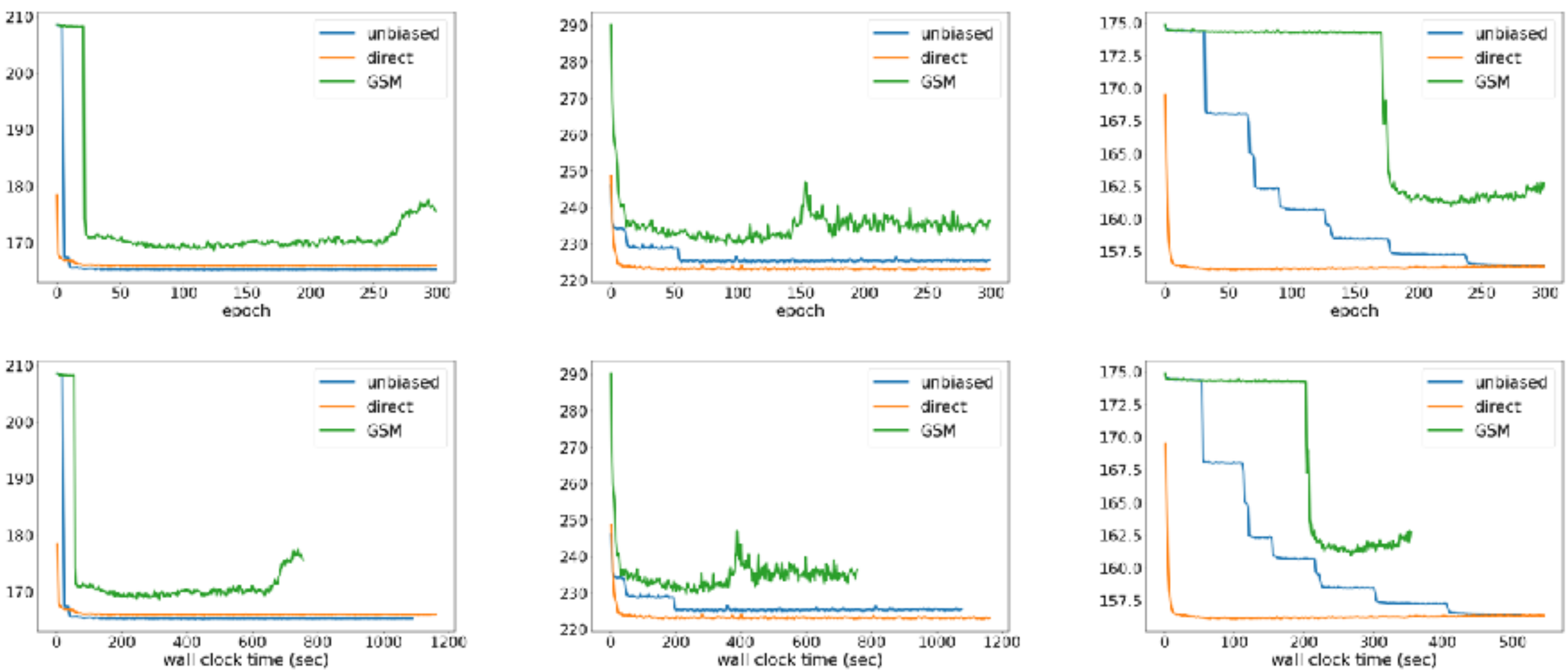
- Gumbel-Argmax / Gumbel-Softmax



Evaluating discrete VAEs

10 discrete latent assignments

MNIST | Fashion MNIST | Omniglot

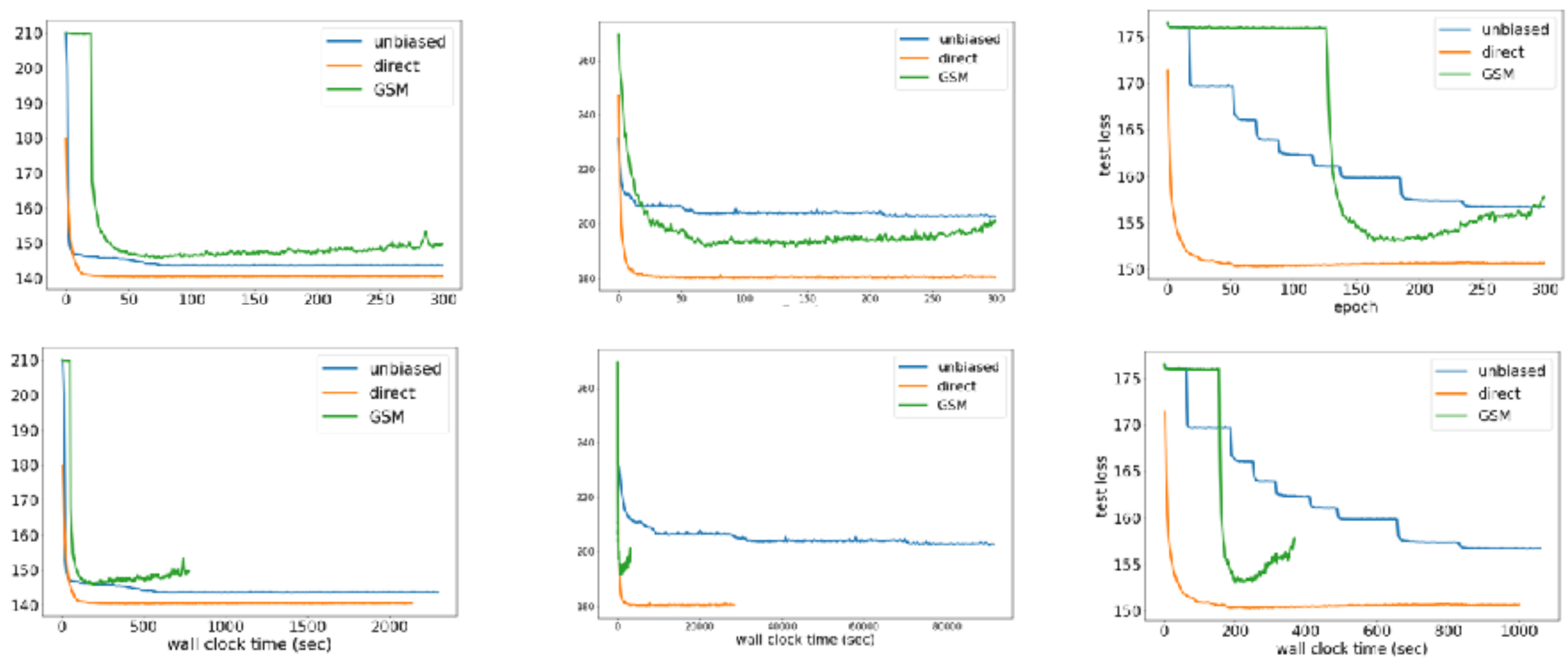


- Surprisingly the unbiased gradient descent is slower to converge
- Surprisingly Gumbel-Argmax wall-clock time is comparable

Evaluating discrete VAEs

50 discrete latent assignments

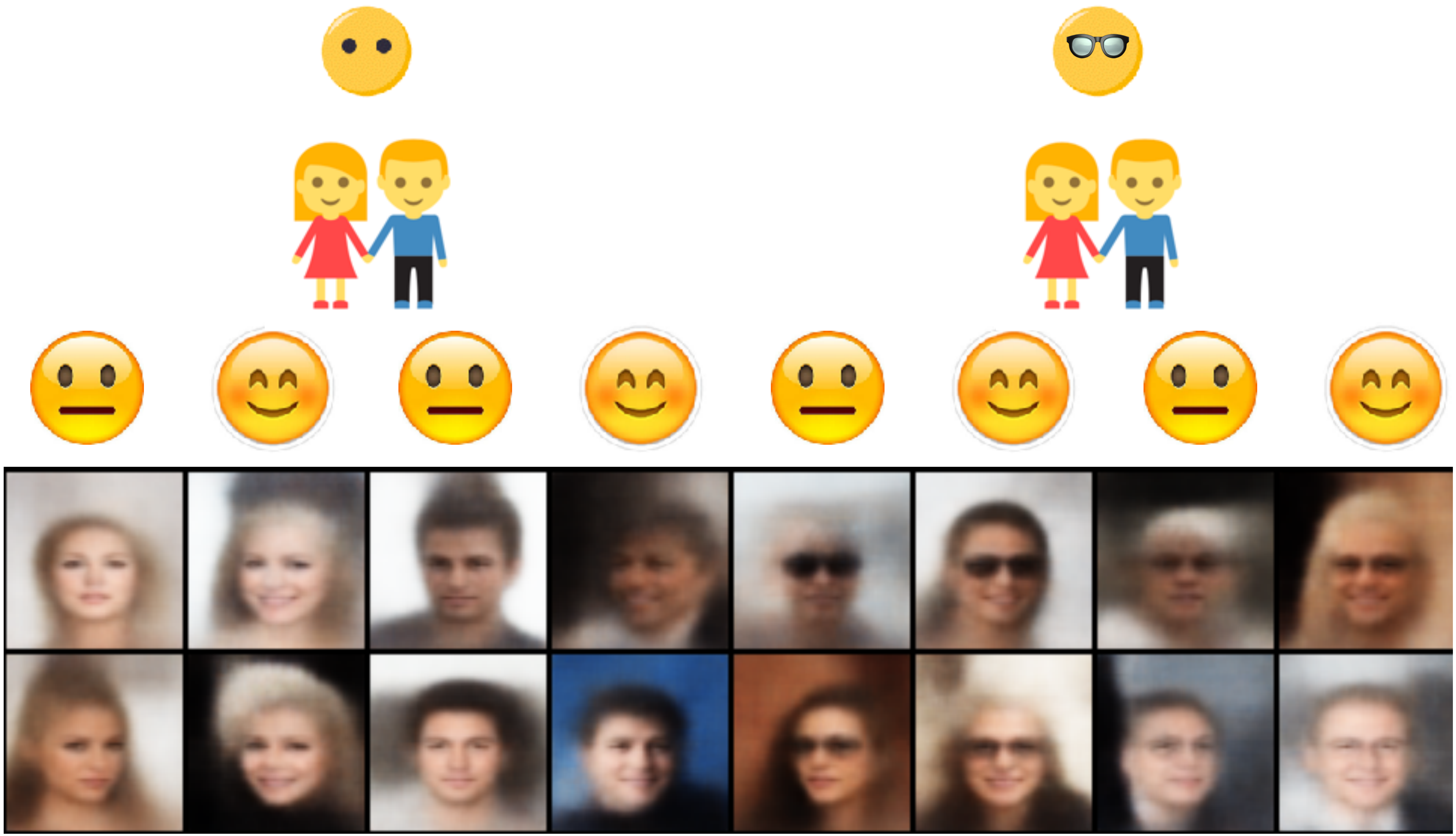
MNIST | Fashion MNIST | Omniglot



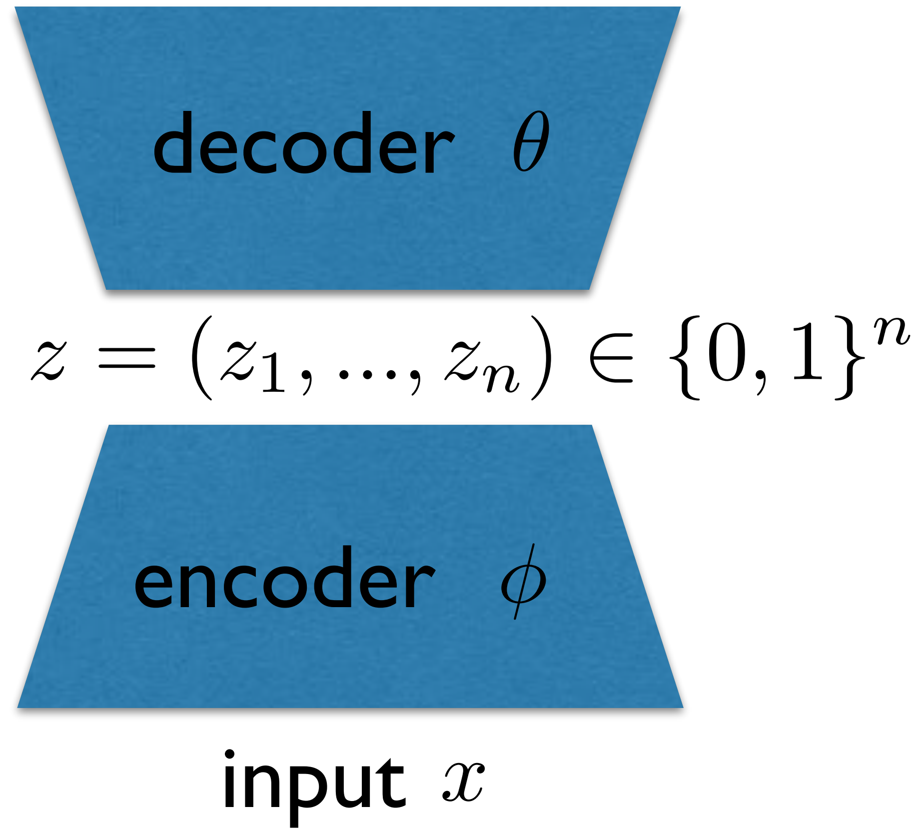
- The unbiased gradient descent is <u>much</u> slower to converge
- Surprisingly Gumbel-Argmax wall-clock time is comparable

High dimensional VAEs

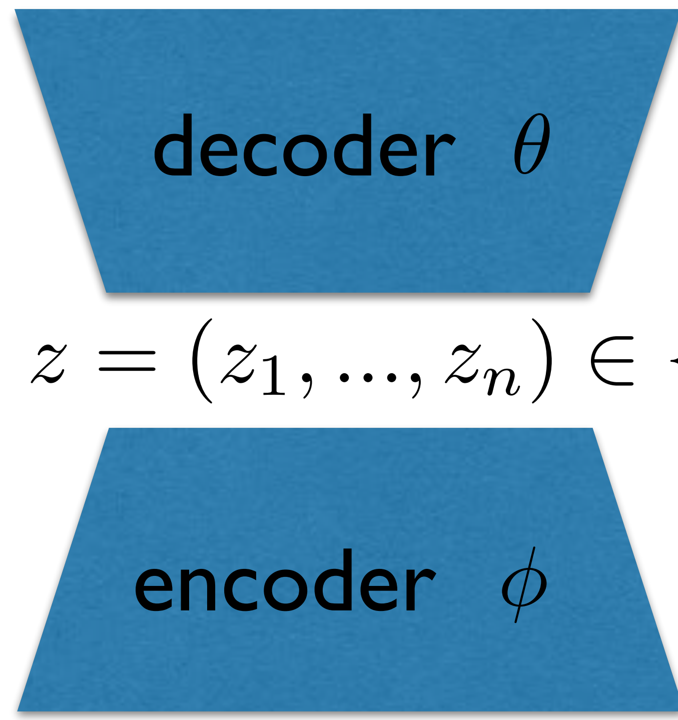
$$z_1, \ldots, z_n \in \{0, 1\}^n$$



High dimensional VAEs



High dimensional VAEs

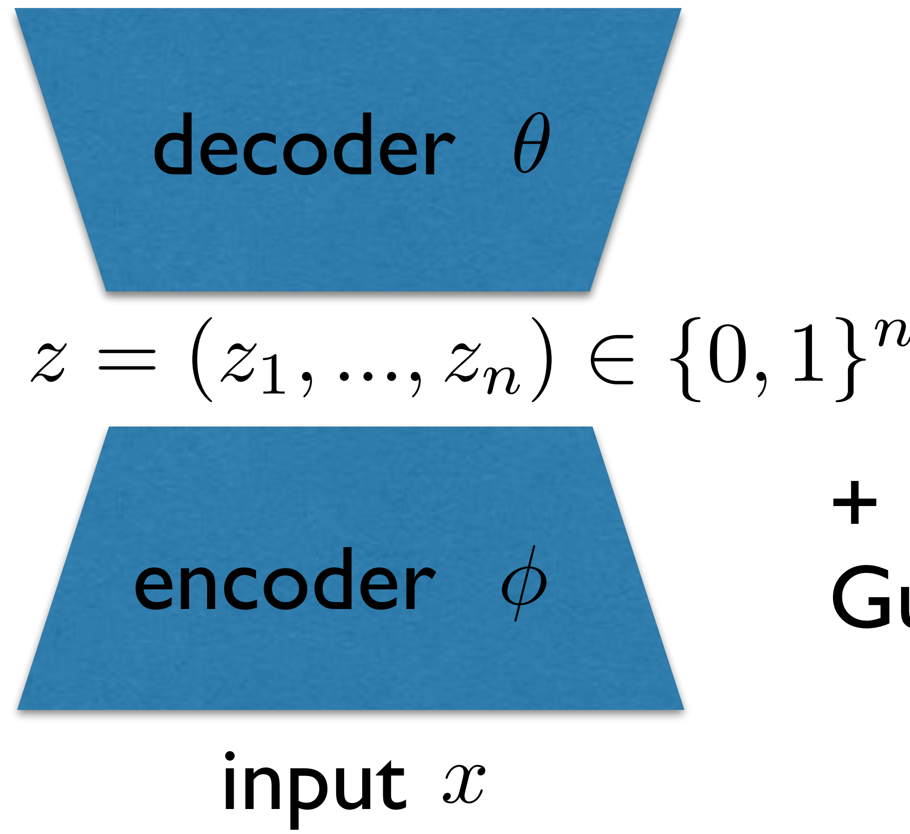


$$z = (z_1, \ldots, z_n) \in \{0, 1\}^n$$

$$\phi(x, z) = \sum_i \phi_i(x, z_i)$$

input x

High dimensional VAEs

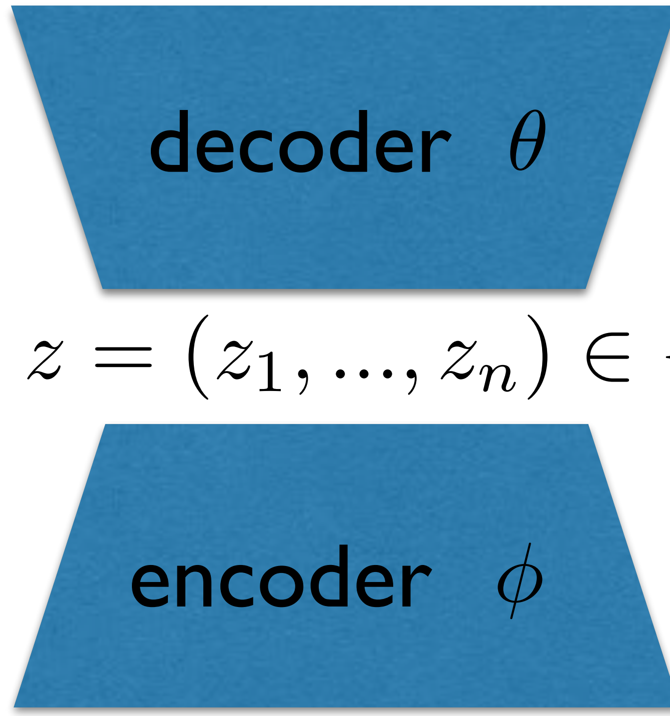


$$\phi(x, z) = \sum_i \phi_i(x, z_i)$$

\+ low dimensional Gumbel perturbation

$$z_i^{\phi+\gamma} = \arg\max_{z_i} \phi_i(x, z_i) + \gamma_i(z_i)$$

High dimensional VAEs



$$z = (z_1, \ldots, z_n) \in \{0,1\}^n$$

input x

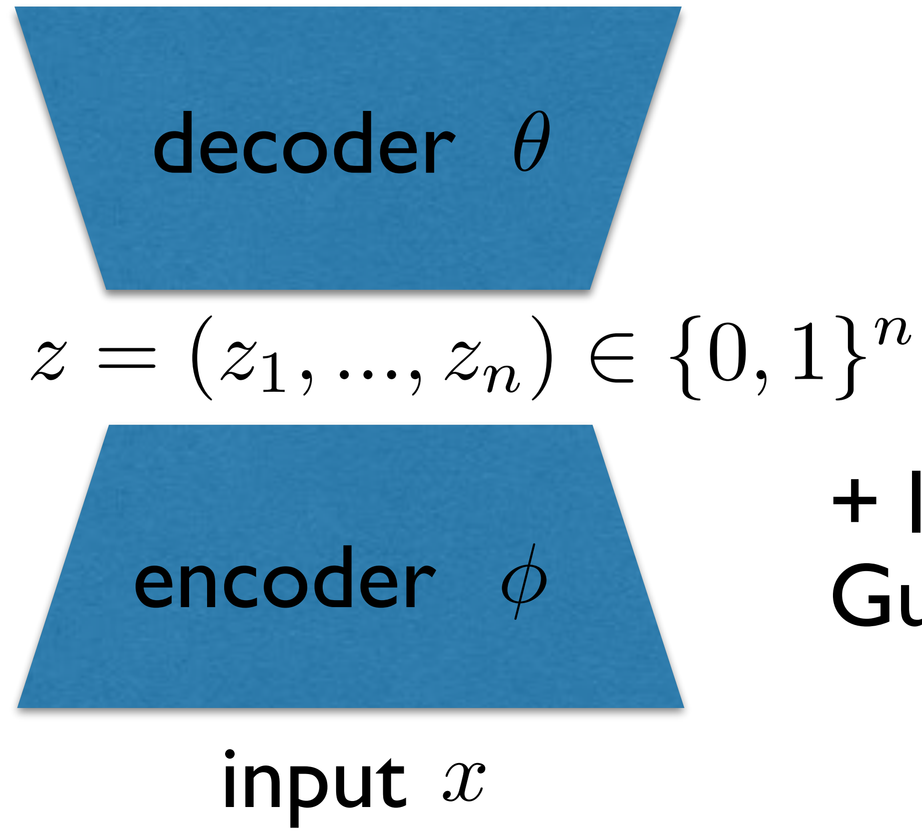
$$\phi(x, z) = \sum_i \phi_i(x, z_i)$$

+ low dimensional Gumbel perturbation

$$z_i^{\phi+\gamma} = \arg\max_{z_i} \phi_i(x, z_i) + \gamma_i(z_i)$$

$$z^{\epsilon\theta+\phi+\gamma} = \arg\max_{z_1,\ldots,z_n} \theta(z_1, \ldots, z_n) + \sum_{i=1}^n \phi_i(x, z_i) + \sum_{i=1}^n \gamma_i(z_i)$$

High dimensional VAEs



$$\phi(x,z) = \sum_i \phi_i(x, z_i)$$

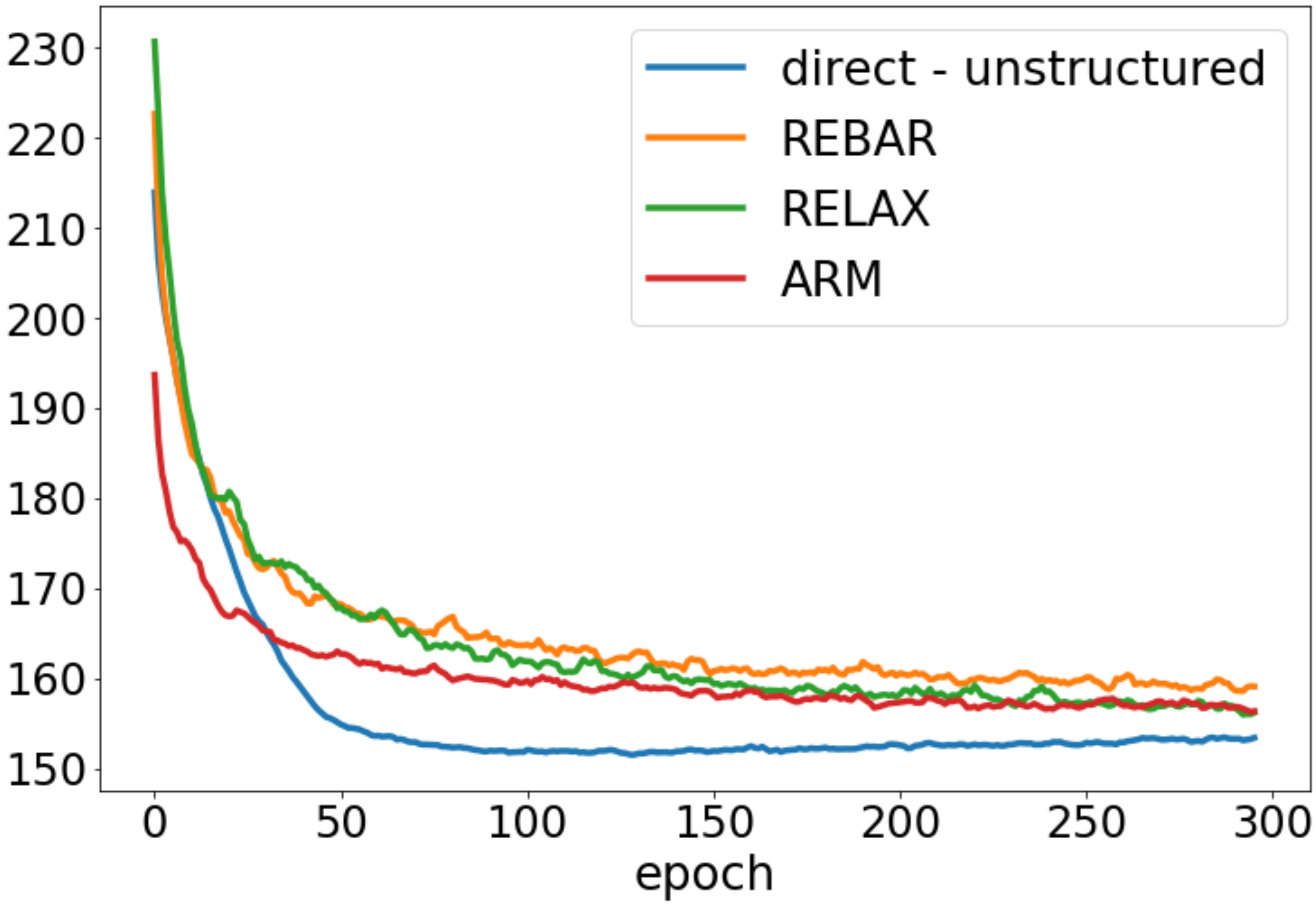
+ low dimensional Gumbel perturbation

$$z_i^{\phi+\gamma} = \arg\max_{z_i} \phi_i(x, z_i) + \gamma_i(z_i)$$

$$z^{\epsilon\theta+\phi+\gamma} = \arg\max_{z_1,\ldots,z_n} \theta(z_1, ..., z_n) + \sum_{i=1}^n \phi_i(x, z_i) + \sum_{i=1}^n \gamma_i(z_i)$$ 😱

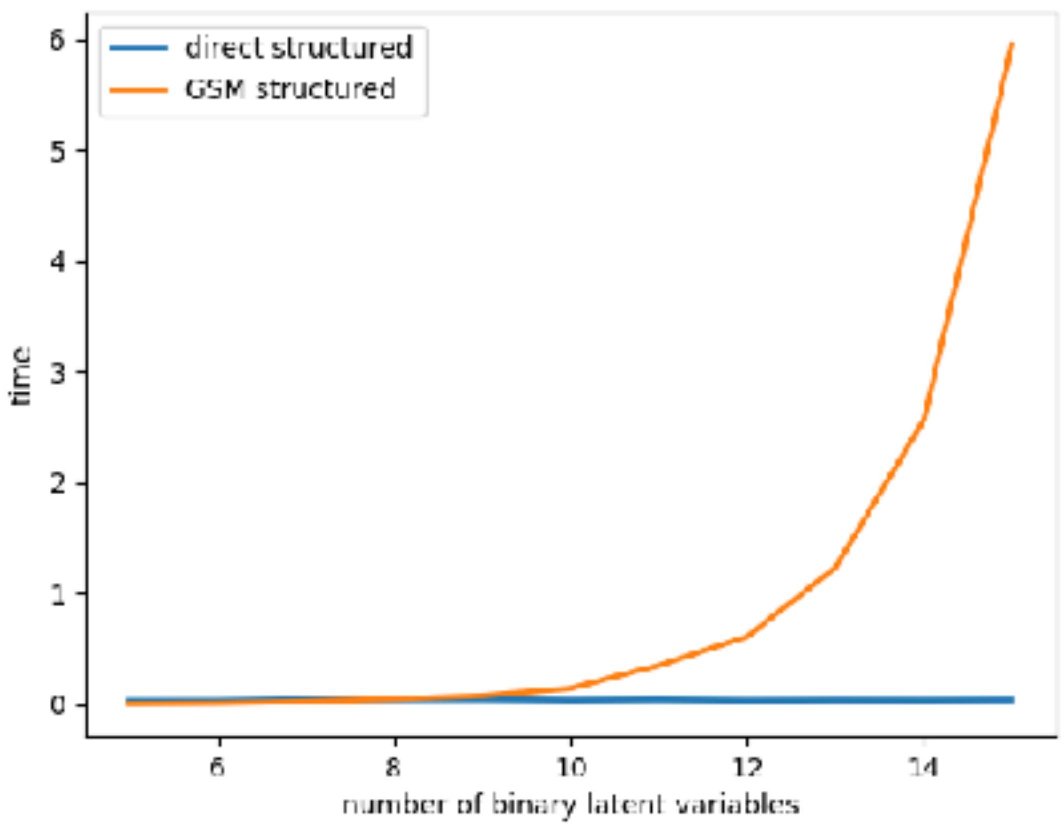
$$z_i^{\epsilon\theta+\phi+\gamma} \approx \arg\max_{z_i} \theta(z_1^{\phi+\gamma}, ..., z_i, ..., z_n^{\phi+\gamma}) + \phi_i(x, z_i) + \gamma_i(z_i)$$ 😅

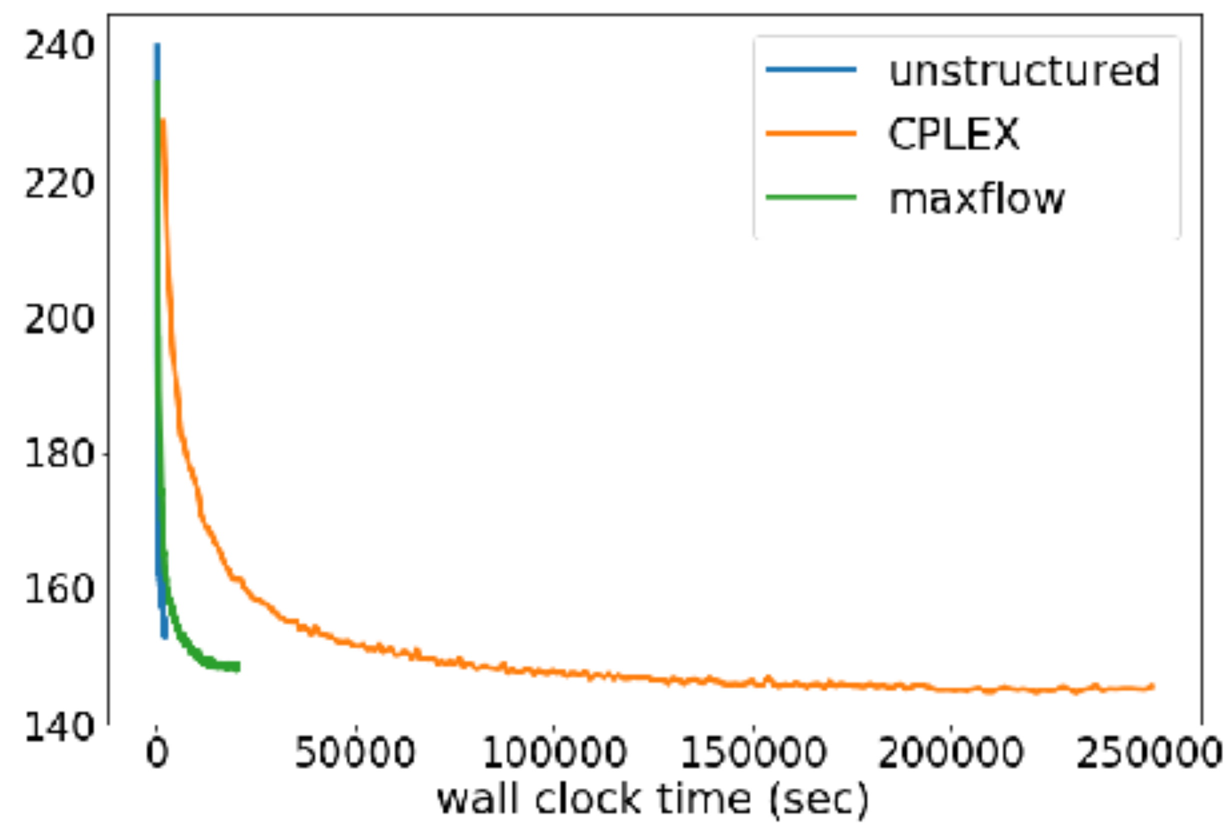
High dimensional VAEs



High dimensional VAEs

- Unstructured encoders $\phi(x, z) = \sum_i \phi_i(x, z_i)$
- Structured encoders $\phi(x, z) = \sum_i \phi_i(x, z_i) + \sum_{i,j} \phi(x, z_i, z_j)$





Semi-supervised VAEs

$$\sum_{x \in S} \mathbb{E}_\gamma[\theta(x, z^{\phi+\gamma})] + \sum_{(x,z) \in S_1} \mathbb{E}_\gamma[\ell(z, z^{\phi+\gamma})] + \sum_{x \in S} KL(q_\phi(z|x)||p_\theta(z))$$

unsupervised

semisupervised

	MNIST				Fashion-MNIST			
	accuracy		bound		accuracy		bound	
#labels	direct	GSM	direct	GSM	direct	GSM	direct	GSM
50	92.6%	84.7%	90.24	91.23	63.3%	61.2%	129.66	129.813
100	95.4%	88.4%	90.93	90.64	67.2%	64.2%	130.822	129.054
300	96.4%	91.7%	90.39	90.01	70.0%	69.3%	130.653	130.371
600	96.7%	92.3%	90.78	89.77	72.1%	71.6%	130.81	129.973
1200	96.8%	92.7%	90.45	90.37	73.7%	73.2%	130.921	130.063