Cycle-Consistent Adversarial Learning as Approximate Bayesian Inference

Louis C. Tiao[1] Edwin V. Bonilla[2] Fabio Ramos[1]

July 22, 2018

[1]University of Sydney, [2]University of New South Wales

Motivation: Unpaired Image-to-Image Translation

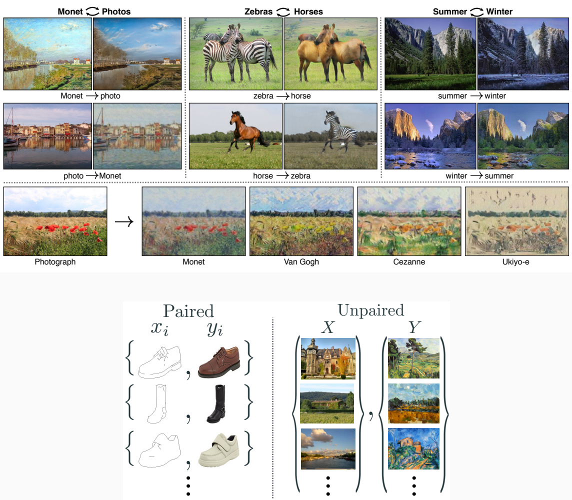


Figure 1: From Zhu et al. (2017)

Cycle-Consistent Adversarial Learning (CycleGAN)

- Introduced by Kim et al. (2017); Zhu et al. (2017)
- Forward and reverse mappings $\mathbf{m}_{\boldsymbol{\phi}} : \mathbf{x} \mapsto \mathbf{z}$ and $\boldsymbol{\mu}_{\boldsymbol{\theta}} : \mathbf{z} \mapsto \mathbf{x}$
- Discriminators $\mathrm{D}_{\boldsymbol{\alpha}}$ and $\mathrm{D}_{\boldsymbol{\beta}}$

Distribution matching (GAN objectives)

Yield realistic outputs in the other domain.

$$\ell_{\mathrm{GAN}}^{\mathrm{reverse}}(\boldsymbol{\alpha}; \boldsymbol{\phi}) = \mathbb{E}_{p^*(\mathbf{z})}[\log \mathrm{D}_{\boldsymbol{\alpha}}(\mathbf{z})] + \mathbb{E}_{q^*(\mathbf{x})}[\log(1 - \mathrm{D}_{\boldsymbol{\alpha}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x})))],$$
$$\ell_{\mathrm{GAN}}^{\mathrm{forward}}(\boldsymbol{\beta}; \boldsymbol{\theta}) = \mathbb{E}_{p^*(\mathbf{x})}[\log \mathrm{D}_{\boldsymbol{\beta}}(\mathbf{x})] + \mathbb{E}_{p^*(\mathbf{z})}[\log(1 - \mathrm{D}_{\boldsymbol{\beta}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z})))].$$

Cycle-consistency losses

Encourage tighter correspondences—must be able to reconstruct output from input and vice versa. May alleviate mode-collapse

$$\ell_{\mathrm{CONST}}^{\mathrm{reverse}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q^*(\mathbf{x})}[\|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}))\|_\rho^\rho],$$
$$\ell_{\mathrm{CONST}}^{\mathrm{forward}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{p^*(\mathbf{z})}[\|\mathbf{z} - \mathbf{m}_{\boldsymbol{\phi}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}))\|_\rho^\rho].$$

Contributions

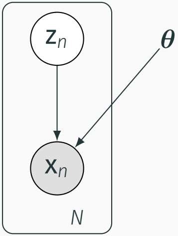
We cast the problem of learning inter-domain correspondences without paired data as **approximate Bayesian inference** in a **latent variable model (LVM)**.

1. We introduce **implicit latent variable models (ILVMs)**,
   - prior over latent variables specified flexibly as **implicit distribution**.
2. We develop a new variational inference (VI) algorithm based on
   - minimizing the **symmetric Kullback-Leibler (KL) divergence**
   - between a variational and exact **joint distribution**.
3. We demonstrate that CYCLEGAN (Kim et al., 2017; Zhu et al., 2017) can be instantiated as a **special case** of our framework.

Implicit Latent Variable Models

Join Distribution

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = \underbrace{p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}_{\text{likelihood}} \underbrace{p^*(\mathbf{z})}_{\text{prior}}$$



Prescribed Likelihood

Likelihood $p_{\boldsymbol{\theta}}(\mathbf{x}_n \mid \mathbf{z}_n)$ is **prescribed** (as usual)

Implicit Prior

Prior $p^*(\mathbf{z})$ over latent variables specified as **implicit** distribution

- Given only by a finite collection $\mathbf{Z}^* = \{\mathbf{z}^*_m\}_{m=1}^M$ of its samples,

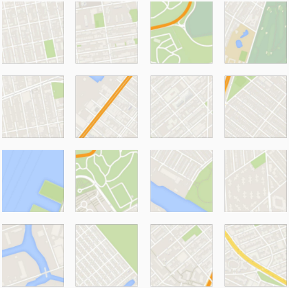
$$\mathbf{z}^*_m \sim p^*(\mathbf{z})$$

- **Offers utmost degree of flexibility in treatment of prior information.**

Implicit Latent Variable Models: Example

Unpaired Image-to-Image Translation

- **Prior distribution** $p^*(\mathbf{z})$ specified by images $\mathbf{Z}^* = \{\mathbf{z}_m^*\}_{m=1}^M$ from one domain.
- **Empirical data distribution** $q^*(\mathbf{x})$ specified by images $\mathbf{X}^* = \{\mathbf{x}_n\}_{n=1}^N$ from another domain.



(a) samples from $p^*(\mathbf{z})$

(b) a sample from $q^*(\mathbf{x})$

Inference in Implicit Latent Variable Models

Having specified the generative model, our aims are

- Optimize $\boldsymbol{\theta}$ by maximizing marginal likelihood $p_{\boldsymbol{\theta}}(\mathbf{x})$
- Infer hidden representations $\mathbf{z}$ by computing posterior $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$

Both require intractable $p_{\boldsymbol{\theta}}(\mathbf{x})$

- **must resort to approximate inference**

Classical Variational Inference

- Approximate exact posterior $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$ with variational posterior $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$
- Reduces **inference** problem to **optimization** problem

$$\min_{\boldsymbol{\phi}} \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})\right]$$

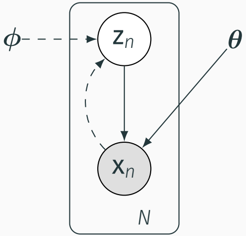
Symmetric Joint-Matching Variational Inference

Variational Joint

- Consider instead directly approximating the exact **joint** with variational **joint**

$$q_{\phi}(x, z) = q_{\phi}(z \mid x) q^{*}(x)$$

  - variational posterior $q_{\phi}(z \mid x)$ also **prescribed**



Symmetric Joint-Matching Variational Inference

Minimize **symmetric KL divergence** between **joints**

$$\mathrm{KL}_{\mathrm{SYMM}}\left[p_{\boldsymbol{\theta}}(\mathrm{x}, \mathrm{z}) \parallel q_{\boldsymbol{\phi}}(\mathrm{x}, \mathrm{z})\right]$$

where

$$\mathrm{KL}_{\mathrm{SYMM}}\left[p \parallel q\right] = \underbrace{\mathrm{KL}\left[p \parallel q\right]}_{\text{forward KL}} + \underbrace{\mathrm{KL}\left[q \parallel p\right]}_{\text{reverse KL}}$$

Why?

1. Because we can:
    - $\mathrm{KL}_{\mathrm{SYMM}}\left[p_{\boldsymbol{\theta}}(\mathrm{x}, \mathrm{z}) \parallel q_{\boldsymbol{\phi}}(\mathrm{x}, \mathrm{z})\right]$ tractable
    - $\mathrm{KL}_{\mathrm{SYMM}}\left[p_{\boldsymbol{\theta}}(\mathrm{z} \mid \mathrm{x}) \parallel q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x})\right]$ **intractable**
2. Helps avoid under/over-dispersed approximations (see paper for details)

Reverse KL Variational Objective

- Minimizing **reverse** KL divergence between **joints** equivalent to maximizing usual **evidence lower bound (ELBO)**,

$$\begin{aligned}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathrm{x},\mathrm{z}) \,\|\, p_{\boldsymbol{\theta}}(\mathrm{x},\mathrm{z})\right] &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathrm{x},\mathrm{z})}\left[\log q_{\boldsymbol{\phi}}(\mathrm{x},\mathrm{z}) - \log p_{\boldsymbol{\theta}}(\mathrm{x},\mathrm{z})\right] \\ &= \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathrm{x},\mathrm{z})}\left[\log q_{\boldsymbol{\phi}}(\mathrm{z}\,|\,\mathrm{x}) - \log p_{\boldsymbol{\theta}}(\mathrm{x},\mathrm{z})\right]}_{\mathcal{L}_{\text{NELBO}}(\boldsymbol{\theta},\boldsymbol{\phi})} - \underbrace{\mathbb{H}[q^*(\mathrm{x})]}_{\text{constant}}\end{aligned}$$

- Recall (negative) ELBO,

$$\mathcal{L}_{\text{NELBO}}(\boldsymbol{\theta},\boldsymbol{\phi}) = \underbrace{\mathbb{E}_{q^*(\mathrm{x})q_{\boldsymbol{\phi}}(\mathrm{z}\,|\,\mathrm{x})}\left[-\log p_{\boldsymbol{\theta}}(\mathrm{x}\,|\,\mathrm{z})\right]}_{\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta},\boldsymbol{\phi})} + \underbrace{\mathbb{E}_{q^*(\mathrm{x})}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathrm{z}\,|\,\mathrm{x}) \,\|\, p^*(\mathrm{z})\right]}_{\text{intractable}}$$

- **KL term is intractable as prior $p^*(\mathrm{z})$ is unavailable—can only sample!**

Forward KL Variational Objective

- Minimizing **forward** KL divergence between **joints**

$$\mathrm{KL}\left[p_{\boldsymbol\theta}(\mathsf{x},\mathsf{z}) \parallel q_{\boldsymbol\phi}(\mathsf{x},\mathsf{z})\right] = \mathbb{E}_{p_{\boldsymbol\theta}(\mathsf{x},\mathsf{z})}\left[\log p_{\boldsymbol\theta}(\mathsf{x},\mathsf{z}) - \log q_{\boldsymbol\phi}(\mathsf{x},\mathsf{z})\right]$$
$$= \underbrace{\mathbb{E}_{p_{\boldsymbol\theta}(\mathsf{x},\mathsf{z})}\left[\log p_{\boldsymbol\theta}(\mathsf{x}\,|\,\mathsf{z}) - \log q_{\boldsymbol\phi}(\mathsf{x},\mathsf{z})\right]}_{\mathcal{L}_{\mathrm{NAPLBO}}(\boldsymbol\theta,\boldsymbol\phi)} - \underbrace{\mathbb{H}[p^*(\mathsf{z})]}_{\text{constant}}$$

- New variational objective, **aggregate posterior lower bound (APLBO)**

$$\mathcal{L}_{\mathrm{NAPLBO}}(\boldsymbol\theta,\boldsymbol\phi) = \underbrace{\mathbb{E}_{p^*(\mathsf{z})p_{\boldsymbol\theta}(\mathsf{x}\,|\,\mathsf{z})}\left[-\log q_{\boldsymbol\phi}(\mathsf{z}\,|\,\mathsf{x})\right]}_{\mathcal{L}_{\mathrm{NELP}}(\boldsymbol\theta,\boldsymbol\phi)} + \underbrace{\mathbb{E}_{p^*(\mathsf{z})}\mathrm{KL}\left[p_{\boldsymbol\theta}(\mathsf{x}\,|\,\mathsf{z}) \parallel q^*(\mathsf{x})\right]}_{\text{intractable}}$$

- **KL term is intractable as empirical data distribution $q^*(\mathsf{x})$ is unavailable—can only sample!**

Density Ratio Estimation and f-divergence Approximation

General f-divergence lower bound (Nguyen et al., 2010)

For convex lower-semicontinuous function $f : \mathbb{R}_+ \to \mathbb{R}$,

$$\underbrace{\mathbb{E}_{q^*(\mathrm{x})} \mathcal{D}_f \left[p^*(\mathrm{z}) \parallel q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x}) \right]}_{\text{intractable}} \geq \max_{\boldsymbol{\alpha}} \underbrace{\mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha}; \boldsymbol{\phi})}_{\text{tractable}},$$

where

$$\mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha}; \boldsymbol{\phi}) = \mathbb{E}_{q^*(\mathrm{x}) q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x})} [f'(r_{\boldsymbol{\alpha}}(\mathrm{z}; \mathrm{x}))] - \mathbb{E}_{q^*(\mathrm{x}) p^*(\mathrm{z})} [f^\star(f'(r_{\boldsymbol{\alpha}}(\mathrm{z}; \mathrm{x})))]$$

- Turns **divergence estimation** into an **optimization problem**
- Estimate divergence using a l.b. that just requires samples!
- $r_{\boldsymbol{\alpha}}$ is a neural net with parameters $\boldsymbol{\alpha}$, with equality at

$$r^*_{\boldsymbol{\alpha}}(\mathrm{z}; \mathrm{x}) = \frac{q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x})}{p^*(\mathrm{z})}$$

KL divergence lower bound

Example: KL divergence lower bound

For $f(u) = u \log u$, we instantiate the KL lower bound

$$\underbrace{\mathbb{E}_{q^*(\mathrm{x})} \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x}) \parallel p^*(\mathrm{z})\right]}_{\text{intractable}} \geq \max_{\boldsymbol{\alpha}} \underbrace{\mathcal{L}_{\mathrm{KL}}^{\text{latent}}(\boldsymbol{\alpha}; \boldsymbol{\phi})}_{\text{tractable}}$$

where

$$\mathcal{L}_{\mathrm{KL}}^{\text{latent}}(\boldsymbol{\alpha}; \boldsymbol{\phi}) = \mathbb{E}_{q^*(\mathrm{x}) q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x})}[\log r_{\boldsymbol{\alpha}}(\mathrm{z}; \mathrm{x})] - \mathbb{E}_{q^*(\mathrm{x}) p^*(\mathrm{z})}[r_{\boldsymbol{\alpha}}(\mathrm{z}; \mathrm{x}) - 1]$$

Yields estimate of the ELBO where all terms are tractable,

$$\begin{aligned} \mathcal{L}_{\mathrm{NELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}) &= \underbrace{\mathcal{L}_{\mathrm{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi})}_{\text{tractable}} + \underbrace{\mathbb{E}_{q^*(\mathrm{x})} \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x}) \parallel p^*(\mathrm{z})\right]}_{\text{intractable}} \\ &\geq \max_{\boldsymbol{\alpha}} \underbrace{\mathcal{L}_{\mathrm{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi})}_{\text{tractable}} + \underbrace{\mathcal{L}_{\mathrm{KL}}^{\text{latent}}(\boldsymbol{\alpha}; \boldsymbol{\phi})}_{\text{tractable}} \end{aligned}$$

CycleGAN as a Special Case

Cycle-consistency as Conditional Probability Maximization

For Gaussian likelihood and variational posterior

$$p_{\boldsymbol{\theta}}(\mathsf{x} \mid \mathsf{z}) = \mathcal{N}(\mathsf{x} \mid \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathsf{z}), \tau^2 \mathbf{I}), \qquad q_{\boldsymbol{\phi}}(\mathsf{z} \mid \mathsf{x}) = \mathcal{N}(\mathsf{z} \mid \mathsf{m}_{\boldsymbol{\phi}}(\mathsf{x}), t^2 \mathbf{I})$$

Can instantiate $\ell^{\text{reverse}}_{\text{CONST}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ from $\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi})$

as posterior $q_{\boldsymbol{\phi}}(\mathsf{z} \mid \mathsf{x})$ *degenerates* (as $t \to 0$)

Can instantiate $\ell^{\text{forward}}_{\text{CONST}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ from $\mathcal{L}_{\text{NELP}}(\boldsymbol{\theta}, \boldsymbol{\phi})$

as likelihood $p_{\boldsymbol{\theta}}(\mathsf{x} \mid \mathsf{z})$ *degenerates* (as $\tau \to 0$)

Cycle-consistency corresponds to maximizing conditional probabilities:

- **ELL.** forces $q_{\boldsymbol{\phi}}(\mathsf{z} \mid \mathsf{x})$ to place mass on hidden representations that *recover* the data
- **ELP.** forces $p_{\boldsymbol{\theta}}(\mathsf{x} \mid \mathsf{z})$ to generate observations that *recover* the prior

Distribution Matching as Regularization

For appropriate setting of f, and simplifying the mappings and discriminators,

- Can instantiate $\ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha}; \boldsymbol{\phi})$ from $\mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha}; \boldsymbol{\phi})$
- Can instantiate $\ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta}; \boldsymbol{\theta})$ from $\mathcal{L}_f^{\text{observed}}(\boldsymbol{\beta}; \boldsymbol{\theta})$

Approximately minimizes intractable divergences:

- $\mathcal{D}_f\left[p^*(\mathsf{z}) \parallel q_{\boldsymbol{\phi}}(\mathsf{z} \mid \mathsf{x})\right]$ — forces $q_{\boldsymbol{\phi}}(\mathsf{z} \mid \mathsf{x})$ to match prior $p^*(\mathsf{z})$
- $\mathcal{D}_f\left[q^*(\mathsf{x}) \parallel p_{\boldsymbol{\theta}}(\mathsf{x} \mid \mathsf{z})\right]$ — forces $p_{\boldsymbol{\theta}}(\mathsf{x} \mid \mathsf{z})$ to match data $q^*(\mathsf{x})$

Summary

$$\mathcal{L}_{\text{NELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \geq \max_{\boldsymbol{\alpha}} \underbrace{\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi})}_{\ell_{\text{CONST}}^{\text{reverse}}(\boldsymbol{\theta}, \boldsymbol{\phi})} + \underbrace{\mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\alpha}; \boldsymbol{\phi})}_{\ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha}; \boldsymbol{\phi})}$$

$$\mathcal{L}_{\text{NAPLBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \geq \max_{\boldsymbol{\beta}} \underbrace{\mathcal{L}_{\text{NELP}}(\boldsymbol{\theta}, \boldsymbol{\phi})}_{\ell_{\text{CONST}}^{\text{forward}}(\boldsymbol{\theta}, \boldsymbol{\phi})} + \underbrace{\mathcal{L}_{\text{KL}}^{\text{observed}}(\boldsymbol{\beta}; \boldsymbol{\theta})}_{\ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta}; \boldsymbol{\theta})}$$

Conclusion

- Formulated **implicit latent variable models**, which introduces **implicit prior** over latent variables
  - Offers utmost degree of flexibility in incorporating prior knowledge
- Developed new paradigm for variational inference
  - directly approximates exact **joint distribution**
  - minimizes the **symmetric KL divergence**
- Provided theoretical treatment of the links between **CycleGAN methods** and **Variational Bayes**

Poster Session

To find out more, come visit us at our poster!

Poster #14, Session 4 (17:10-18:00 Saturday, 14 July)

Questions?

References

Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. (2017). Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 1857–1865.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *IEEE Trans. Information Theory*, 56(11):5847–5861.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*.

Symmetric Joint-Matching KL Minimization i

- KL divergence is asymmetric $\text{KL}\left[p \parallel q\right] \neq \text{KL}\left[q \parallel p\right]$
- $\text{KL}\left[q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x}) \parallel p_{\boldsymbol{\theta}}(\mathrm{z} \mid \mathrm{x})\right]$ (reverse) underestimates support
- $\text{KL}\left[p_{\boldsymbol{\theta}}(\mathrm{z} \mid \mathrm{x}) \parallel q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x})\right]$ (forward) overestimates support
- Consider symmetric KL: $\text{KL}_{\text{SYMM}}\left[p \parallel q\right] = \text{KL}\left[p \parallel q\right] + \text{KL}\left[q \parallel p\right]$

- Forward KL involves expectation under intractable posterior $p_{\boldsymbol{\theta}}(\mathrm{z} \mid \mathrm{x})$—what we're trying to approximate in the first place

$$\text{KL}\left[p_{\boldsymbol{\theta}}(\mathrm{z} \mid \mathrm{x}) \parallel q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x})\right] = \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathrm{z} \mid \mathrm{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathrm{z} \mid \mathrm{x})}{q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x})}\right]$$

Symmetric Joint-Matching KL Minimization ii

- Can show

$$\arg\min_{\boldsymbol{\phi}} \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x}) \parallel p_{\boldsymbol{\theta}}(\mathrm{z} \mid \mathrm{x})\right] = \arg\min_{\boldsymbol{\phi}} \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathrm{x}, \mathrm{z}) \parallel p_{\boldsymbol{\theta}}(\mathrm{x}, \mathrm{z})\right]$$

$$\arg\min_{\boldsymbol{\phi}} \mathrm{KL}\left[p_{\boldsymbol{\theta}}(\mathrm{z} \mid \mathrm{x}) \parallel q_{\boldsymbol{\phi}}(\mathrm{z} \mid \mathrm{x})\right] = \arg\min_{\boldsymbol{\phi}} \mathrm{KL}\left[p_{\boldsymbol{\theta}}(\mathrm{x}, \mathrm{z}) \parallel q_{\boldsymbol{\phi}}(\mathrm{x}, \mathrm{z})\right]$$

- Already showed

$$\arg\max_{\boldsymbol{\phi}} \mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \arg\min_{\boldsymbol{\phi}} \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathrm{x}, \mathrm{z}) \parallel p_{\boldsymbol{\theta}}(\mathrm{x}, \mathrm{z})\right]$$

- Can we find something similar for $\mathrm{KL}\left[p_{\boldsymbol{\theta}}(\mathrm{x}, \mathrm{z}) \parallel q_{\boldsymbol{\phi}}(\mathrm{x}, \mathrm{z})\right]$?