Cycle-Consistent Adversarial Learning as Approximate Bayesian Inference

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

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

Summary

We cast the problem of learning inter-domain correspondences as approximate Bayesian inference in a latent variable model (LVM).

- We introduce **implicit latent variable models (ILVMs)**, where the prior over latent variables can be specified flexibly as an **implicit distribution**.
- We develop a new variational inference (VI) algorithm based on minimizing the **symmetric Kullback-Leibler (KL) divergence** between a variational and exact **joint distribution**.
- We demonstrate that the cycle-consistent adversarial learning (CYCLEGAN) models [1, 2] can be derived as a special case within our proposed VI framework.

Implicit Latent Variable Models

- Latent variable models (LVMs) are an indispensable tool for uncovering the hidden representations of observed data.
- Observation $\mathbf{x}$ is assumed governed by its underlying hidden variable $\mathbf{z}$. Joint distribution usually written as

$$p_{\theta}(\mathbf{x},\mathbf{z}) = p_{\theta}(\mathbf{x}\,|\,\mathbf{z})p^*(\mathbf{z}) \tag{1}$$

- **Implicit Prior.** Prior over latent variables is specified as an **implicit distribution** $p^*(\mathbf{z})$, given only by a finite collection $\mathbf{Z}^* = \{\mathbf{z}_m^*\}_{m=1}^M$ of its samples,

$$\mathbf{z}_m^* \sim p^*(\mathbf{z}). \tag{2}$$

Offers the utmost degree of flexibility in treatment of prior information.

- **Prescribed Likelihood.** Likelihood specified through mapping $\mathcal{F}_{\theta}$ which takes as input random noise ξ and latent variable $\mathbf{z}$,

$$\mathbf{x} \sim p_{\theta}(\mathbf{x}\,|\,\mathbf{z}) \\ \Leftrightarrow \mathbf{x} = \mathcal{F}_{\theta}(\xi;\mathbf{z}), \quad \xi \sim p(\xi) \tag{3}$$

(But restricted to **prescribed** likelihoods)

- **Example: Unpaired Image-to-Image Translation** Prior $p^*(\mathbf{z})$ is specified by images from one domain, while empirical distribution $q^*(\mathbf{x})$ is specified by images from another.

Symmetric Joint-Matching VI

- **Prescribed Variational Distribution.** Also specified through a mapping $\mathcal{G}_{\phi}$, with input noise ϵ and observed variable $\mathbf{x}$,

$$\mathbf{z} \sim q_{\phi}(\mathbf{z}\,|\,\mathbf{x}) \\ \Leftrightarrow \mathbf{z} = \mathcal{G}_{\phi}(\epsilon;\mathbf{x}), \quad \epsilon \sim p(\epsilon). \tag{4}$$

- Directly approximate the exact joint with **variational joint.**

$$q_{\phi}(\mathbf{x},\mathbf{z}) = q_{\phi}(\mathbf{z}\,|\,\mathbf{x})q^*(\mathbf{x}). \tag{5}$$

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

$$\mathrm{KL}_{\mathrm{SYMM}}\left[p_{\theta}(\mathbf{x},\mathbf{z}) \,\|\, q_{\phi}(\mathbf{x},\mathbf{z})\right]. \tag{6}$$

where $\mathrm{KL}_{\mathrm{SYMM}}[p \,\|\, q] := \mathrm{KL}[p \,\|\, q] + \mathrm{KL}[q \,\|\, p]$.

- Avoids under-/over-dispersed approximations.

Reverse KL Variational Objective

- **Reverse** KL divergence between **joints**,

$$\begin{aligned} &\mathrm{KL}\left[q_{\phi}(\mathbf{x},\mathbf{z}) \,\|\, p_{\theta}(\mathbf{x},\mathbf{z})\right] \\ &= \mathbb{E}_{q_{\phi}(\mathbf{x},\mathbf{z})}\left[\log q_{\phi}(\mathbf{x},\mathbf{z}) - \log p_{\theta}(\mathbf{x},\mathbf{z})\right] \quad (7)\\ &= \mathcal{L}_{\mathrm{NELBO}}(\theta,\phi) - \underbrace{\mathbb{H}[q^*(\mathbf{x})]}_{\text{constant}}. \quad (8)\end{aligned}$$

- Equivalent to maximizing **evidence lower bound (ELBO)**,

$$\mathcal{L}_{\mathrm{NELBO}}(\theta,\phi) = \underbrace{\mathbb{E}_{q^*(\mathbf{x})q_{\phi}(\mathbf{z}|\mathbf{x})}[-\log p_{\theta}(\mathbf{x}\,|\,\mathbf{z})]}_{\mathcal{L}_{\mathrm{NELL}}(\theta,\phi)} + \mathbb{E}_{q^*(\mathbf{x})}\mathrm{KL}\left[q_{\phi}(\mathbf{z}\,|\,\mathbf{x}) \,\|\, p^*(\mathbf{z})\right]. \tag{9}$$

- But KL term intractable as density $p^*(\mathbf{z})$ unavailable!

Forward KL Variational Objective

- **Forward** KL divergence between **joints**,

$$\begin{aligned} &\mathrm{KL}\left[p_{\theta}(\mathbf{x},\mathbf{z}) \,\|\, q_{\phi}(\mathbf{x},\mathbf{z})\right] \\ &= \mathbb{E}_{p_{\theta}(\mathbf{x},\mathbf{z})}\left[\log p_{\theta}(\mathbf{x},\mathbf{z}) - \log q_{\phi}(\mathbf{x},\mathbf{z})\right] \quad (10)\\ &= \mathcal{L}_{\mathrm{NAPLBO}}(\theta,\phi) - \underbrace{\mathbb{H}[p^*(\mathbf{z})]}_{\text{constant}}. \quad (11)\end{aligned}$$

- Define new variational objective,

$$\mathcal{L}_{\mathrm{NAPLBO}}(\theta,\phi) = \underbrace{\mathbb{E}_{p^*(\mathbf{z})p_{\theta}(\mathbf{x}|\mathbf{z})}[-\log q_{\phi}(\mathbf{z}\,|\,\mathbf{x})]}_{\mathcal{L}_{\mathrm{NELP}}(\theta,\phi)} + \mathbb{E}_{p^*(\mathbf{z})}\mathrm{KL}\left[p_{\theta}(\mathbf{x}\,|\,\mathbf{z}) \,\|\, q^*(\mathbf{x})\right]. \tag{12}$$

- Tractable, unlike $\mathrm{KL}\left[p_{\theta}(\mathbf{z}\,|\,\mathbf{x}) \,\|\, q_{\phi}(\mathbf{z}\,|\,\mathbf{x})\right]$!
- But KL term intractable as density $q^*(\mathbf{x})$ unavailable!

Approximate Divergence Minimization

- Well-known generalized lower bound [3],

$$\mathbb{E}_{q^*(\mathbf{x})}\mathcal{D}_f\left[p^*(\mathbf{z}) \,\|\, q_{\phi}(\mathbf{z}\,|\,\mathbf{x})\right] \geq \max_{\alpha} \mathcal{L}_f^{\mathrm{latent}}(\alpha;\phi), \tag{13}$$

where

$$\begin{aligned}\mathcal{L}_f^{\mathrm{latent}}(\alpha;\phi) = &\,\mathbb{E}_{q^*(\mathbf{x})q_{\phi}(\mathbf{z}|\mathbf{x})}[f'(r_{\alpha}(\mathbf{z};\mathbf{x}))] \\ &- \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[f^*(f'(r_{\alpha}(\mathbf{z};\mathbf{x})))],\end{aligned} \tag{14}$$

and r_{α} is a neural net with parameters α, with equality at

$$r_{\alpha}^*(\mathbf{z};\mathbf{x}) = \frac{q_{\phi}(\mathbf{z}\,|\,\mathbf{x})}{p^*(\mathbf{z})} \tag{15}$$

- For $f_{\mathrm{KL}}(u) = u\log u$, we instantiate KL lower bound,

$$\underbrace{\mathbb{E}_{q^*(\mathbf{x})}\mathrm{KL}\left[q_{\phi}(\mathbf{z}\,|\,\mathbf{x}) \,\|\, p^*(\mathbf{z})\right]}_{\text{intractable}} \geq \underbrace{\max_{\alpha} \mathcal{L}_{\mathrm{KL}}^{\mathrm{latent}}(\alpha;\phi)}_{\text{tractable}} \tag{16}$$

where

$$\begin{aligned}\mathcal{L}_{\mathrm{KL}}^{\mathrm{latent}}(\alpha;\phi) = &\,\mathbb{E}_{q^*(\mathbf{x})q_{\phi}(\mathbf{z}|\mathbf{x})}[\log r_{\alpha}(\mathbf{z};\mathbf{x})] \\ &- \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[r_{\alpha}(\mathbf{z};\mathbf{x}) - 1].\end{aligned} \tag{17}$$

Related to **KL importance estimation procedure (KLIEP)** [4].

- Similar lower bound for $\mathbb{E}_{p^*(\mathbf{z})}\mathcal{D}_f\left[q^*(\mathbf{x}) \,\|\, p_{\theta}(\mathbf{x}\,|\,\mathbf{z})\right]$.

CycleGAN as a Special Case

- Mappings $\mathbf{m}_{\phi}: \mathbf{x} \mapsto \mathbf{z}$ and $\mu_{\theta}: \mathbf{z} \mapsto \mathbf{x}$, discriminators $\mathbf{D}_{\alpha}$, $\mathbf{D}_{\beta}$.
- **Cycle-consistency losses**

$$\ell_{\mathrm{CONST}}^{\mathrm{reverse}}(\theta,\phi) = \mathbb{E}_{q^*(\mathbf{x})}[\|\mathbf{x} - \mu_{\theta}(\mathbf{m}_{\phi}(\mathbf{x}))\|_{\rho}^{\rho}], \tag{18}$$

$$\ell_{\mathrm{CONST}}^{\mathrm{forward}}(\theta,\phi) = \mathbb{E}_{p^*(\mathbf{z})}[\|\mathbf{z} - \mathbf{m}_{\phi}(\mu_{\theta}(\mathbf{z}))\|_{\rho}^{\rho}]. \tag{19}$$

- **Distribution matching** (GAN) objectives

$$\begin{aligned}\ell_{\mathrm{GAN}}^{\mathrm{reverse}}(\alpha;\phi) = &\,\mathbb{E}_{p^*(\mathbf{z})}[\log \mathbf{D}_{\alpha}(\mathbf{z})] \\ &+ \mathbb{E}_{q^*(\mathbf{x})}[\log(1 - \mathbf{D}_{\alpha}(\mathbf{m}_{\phi}(\mathbf{x})))],\end{aligned} \tag{20}$$

$$\begin{aligned}\ell_{\mathrm{GAN}}^{\mathrm{forward}}(\beta;\theta) = &\,\mathbb{E}_{p^*(\mathbf{x})}[\log \mathbf{D}_{\beta}(\mathbf{x})] \\ &+ \mathbb{E}_{p^*(\mathbf{z})}[\log(1 - \mathbf{D}_{\beta}(\mu_{\theta}(\mathbf{z})))].\end{aligned} \tag{21}$$

Cycle-consistency as Conditional Probability Maximization

For Gaussian likelihood and variational posterior

$$p_{\theta}(\mathbf{x}\,|\,\mathbf{z}) = \mathcal{N}(\mathbf{x}\,|\,\mu_{\theta}(\mathbf{z}),\tau^2\mathbf{I}), \qquad q_{\phi}(\mathbf{z}\,|\,\mathbf{x}) = \mathcal{N}(\mathbf{z}\,|\,\mathbf{m}_{\phi}(\mathbf{x}),t^2\mathbf{I}),$$
$$\Leftrightarrow \mathcal{F}_{\theta}(\xi;\mathbf{z}) = \mu_{\theta}(\mathbf{z}) + \tau\xi, \qquad \Leftrightarrow \mathcal{G}_{\phi}(\epsilon;\mathbf{x}) = \mathbf{m}_{\phi}(\mathbf{x}) + t\epsilon$$

- **$\ell_{\mathrm{CONST}}^{\mathrm{reverse}}(\theta,\phi)$ can be recovered from $\mathcal{L}_{\mathrm{NELL}}(\theta,\phi)$ as posterior $q_{\phi}(\mathbf{z}\,|\,\mathbf{x})$ becomes *degenerate*,**

$$\mathcal{L}_{\mathrm{NELL}}(\theta,\phi) \to \gamma_1\ell_{\mathrm{CONST}}^{\mathrm{reverse}}(\theta,\phi) + \delta_1 \quad \text{as } t \to 0$$

for constants γ_1 and δ_1.

Similarly, **$\ell_{\mathrm{CONST}}^{\mathrm{forward}}(\theta,\phi)$ can be recovered from $\mathcal{L}_{\mathrm{NELP}}(\theta,\phi)$ as likelihood $p_{\theta}(\mathbf{x}\,|\,\mathbf{z})$ becomes *degenerate*,**

$$\mathcal{L}_{\mathrm{NELP}}(\theta,\phi) \to \gamma_2\ell_{\mathrm{CONST}}^{\mathrm{forward}}(\theta,\phi) + \delta_2 \quad \text{as } \tau \to 0$$

for constants γ_2 and δ_2.

- **Cycle-consistency corresponds to maximizing likelihood $p_{\theta}(\mathbf{x}\,|\,\mathbf{z})$ and approximate posterior $q_{\phi}(\mathbf{z}\,|\,\mathbf{x})$.**

Distribution Matching as Regularization

- For $f_{\mathrm{GAN}}(u) = u\log u - (u+1)\log(u+1)$, we instantiate

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}}^{\mathrm{reverse}}(\alpha;\phi) := &\,\mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[\log \mathcal{D}_{\alpha}(\mathbf{z};\mathbf{x})] \\ &+ \mathbb{E}_{q^*(\mathbf{x})q_{\phi}(\mathbf{z}|\mathbf{x})}[\log(1 - \mathcal{D}_{\alpha}(\mathbf{z};\mathbf{x}))],\end{aligned} \tag{22}$$

where discriminator $\mathcal{D}_{\alpha}(\mathbf{z};\mathbf{x}) := 1 - \sigma(\log r_{\alpha}(\mathbf{z};\mathbf{x}))$.

- By fixing discriminator to ignore auxiliary input $\mathbf{x}$,

$$\mathcal{D}_{\alpha}(\mathbf{z};\mathbf{x}) = \mathbf{D}_{\alpha}(\mathbf{z}), \tag{23}$$

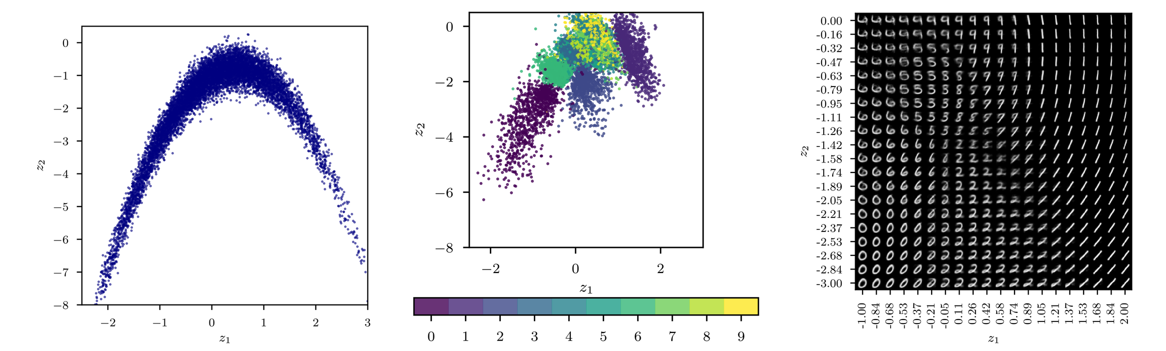
and fixing mapping to ignore stochastic input ϵ,

$$\mathcal{G}_{\phi}(\epsilon;\mathbf{x}) = \mathbf{m}_{\phi}(\mathbf{x}), \tag{24}$$

$\mathcal{L}_{\mathrm{GAN}}^{\mathrm{reverse}}(\alpha;\phi)$ reduces to $\ell_{\mathrm{GAN}}^{\mathrm{reverse}}(\alpha;\phi)$.

- Can be viewed as another way to estimate density ratio $r_{\alpha}^*(\mathbf{z};\mathbf{x})$ of eq. (15).
- **Regularizes approximate posterior $q_{\phi}(\mathbf{z}\,|\,\mathbf{x})$ by approximately minimizes intractable divergence $\mathcal{D}_f\left[p^*(\mathbf{z}) \,\|\, q_{\phi}(\mathbf{z}\,|\,\mathbf{x})\right]$ from prior $p^*(\mathbf{z})$.**
- Setting f_{KL} can help alleviate vanishing gradients, and results in usual prior-contrastive KL term of ELBO.
- Similar results for $\ell_{\mathrm{GAN}}^{\mathrm{forward}}(\beta;\theta)$.

Experiment: MNIST with Implicit Prior



(a) 10k samples from prior $\tilde{p}(\mathbf{z})$.

(b) Mean of $q_{\phi}(\mathbf{z}\,|\,\mathbf{x})$ for every $\mathbf{x}$ from held-out test set of size 10k, colored by digit class.

(c) Mean of $p_{\theta}(\mathbf{x}\,|\,\mathbf{z})$ for 20×20 values of $\mathbf{z}$ along a uniform grid.

Figure: Visualization of 2D latent space and the corresponding observed space manifold.

Table: Mean-squared errors of reconstructions.

METHOD	MSE z	MSE x
SJMVI (OURS)	**0.17**	**0.04**
VAE [5]	0.88	**0.04**
AVB [6]	0.29	**0.04**

References

[1] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[2] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 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, 2017.

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

[4] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, Dec 2008.

[5] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR) 2014*, Dec 2014.

[6] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 2391–2400, Jan 2017.