Path Gradients for Continuous Normalizing Flows

Lorenz Vaitl, Kim Nicoli, Shinichi Nakajima, Pan Kessel




Main idea:

- Variational Inference:

$$KL(q_\theta|p)$$

Main idea:

- Variational Inference:

$$KL(q_\theta|p)$$

- Continuous Normalizing Flow q_θ

Main idea:

- Variational Inference:

$$KL(q_\theta|p)$$

- Continuous Normalizing Flow q_θ
- Minimized by SGD:

$$\theta \leftarrow \theta - \lambda \nabla_\theta KL(q_\theta|p)$$

Main idea:

- Variational Inference:

$$KL(q_\theta|p)$$

- Continuous Normalizing Flow q_θ
- Minimized by SGD:

$$\theta \leftarrow \theta - \lambda \underbrace{\nabla_\theta KL(q_\theta|p)}_{\approx \mathcal{G}_{\text{total}}}$$

- Gradient typically estimated by total gradient estimator

Main idea:

- Variational Inference:

$$KL(q_\theta | p)$$

- Continuous Normalizing Flow q_θ
- Minimized by SGD:

$$\theta \leftarrow \theta - \lambda \underbrace{\nabla_\theta KL(q_\theta | p)}_{\approx \cancel{\mathcal{G}_{\text{total}}}\ \mathcal{G}_{\text{path}}}$$

- Gradient typically estimated by total gradient estimator
- We propose using the path gradient estimator

Main idea:

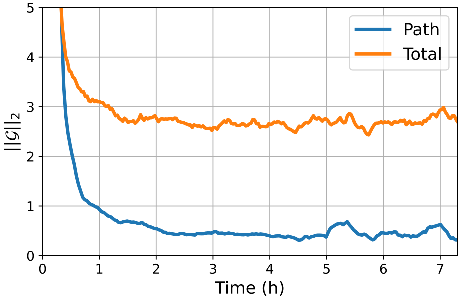
- Variational Inference:

$$KL(q_\theta|p)$$

- Continuous Normalizing Flow q_θ
- Minimized by SGD:

$$\theta \leftarrow \theta - \lambda \underbrace{\nabla_\theta KL(q_\theta|p)}_{\approx \cancel{\mathcal{G}_{\text{total}}}\ \mathcal{G}_{\text{path}}}$$

- Gradient typically estimated by total gradient estimator
- We propose using the path gradient estimator



-ELBO	Path (ours)	Total
MNIST	**82.09** $\pm .04$	$82.82 \pm .01$
Omniglot	**96.61** $\pm .17$	$98.33 \pm .09$
Caltech Silhouettes	**101.93** $\pm .63$	$104.03 \pm .43$
Frey Faces	**4.35** $\pm .00$	$4.39 \pm .01$

Variational Inference

- Target distribution:

$$p(x) \qquad\qquad \text{with tractable score } \frac{\partial \log p(x)}{\partial x}$$

Variational Inference

- Target distribution:

$$p(x) \qquad \text{with tractable score } \frac{\partial \log p(x)}{\partial x}$$

- Normalizing Flow:

$$x = g_\theta(z) \quad z \sim q_Z \qquad \log q_\theta(x) = \log q_Z(g_\theta^{-1}(x)) + \log \left| \det \frac{\partial g_\theta^{-1}(x)}{\partial x} \right|$$

Variational Inference

- Target distribution:

$$p(x) \qquad \text{with tractable score } \frac{\partial \log p(x)}{\partial x}$$

- Normalizing Flow:

$$x = g_\theta(z) \quad z \sim q_Z \qquad \log q_\theta(x) = \log q_Z(g_\theta^{-1}(x)) + \log \left| \det \frac{\partial g_\theta^{-1}(x)}{\partial x} \right|$$

- Reverse Kullback-Leiber Divergence:

$$KL(q_\theta|p) = \mathbb{E}_{x\sim q_\theta(x)}\left[\ln \frac{q_\theta(x)}{p(x)}\right] = \mathbb{E}_{z\sim q_z}\left[\ln \frac{q_\theta(g_\theta(z))}{p(g_\theta(z))}\right]$$

Path Gradients

- Reverse KL

$$KL(q_\theta|p) = \mathbb{E}_{x\sim q_\theta(x)}\left[\ln\frac{q_\theta(x)}{p(x)}\right] = \mathbb{E}_{z\sim q_z}\left[\ln\frac{q_\theta(g_\theta(z))}{p(g_\theta(z))}\right]$$

Path Gradients

- Reverse KL total gradient

$$\frac{d}{d\theta}KL(q_\theta|p) = \frac{d}{d\theta}\mathbb{E}_{x\sim q_\theta(x)}\left[\ln\frac{q_\theta(x)}{p(x)}\right] = \frac{d}{d\theta}\mathbb{E}_{z\sim q_z}\left[\ln\frac{q_\theta(g_\theta(z))}{p(g_\theta(z))}\right]$$

Path Gradients

- Reverse KL total gradient

$$\frac{d}{d\theta} KL(q_\theta|p) = \frac{d}{d\theta}\mathbb{E}_{x\sim q_\theta(x)}\left[\ln\frac{q_\theta(x)}{p(x)}\right] = \mathbb{E}_{z\sim q_z}\left[\frac{d}{d\theta}\ln\frac{q_\theta(g_\theta(z))}{p(g_\theta(z))}\right]$$

Path Gradients

- Reverse KL total gradient

$$\frac{d}{d\theta} KL(q_\theta|p) = \frac{d}{d\theta}\mathbb{E}_{x\sim q_\theta(x)}\left[\ln \frac{q_\theta(x)}{p(x)}\right] = \mathbb{E}_{z\sim q_z}\left[\frac{d}{d\theta}\ln\frac{q_\theta(g_\theta(z))}{p(g_\theta(z))}\right]$$

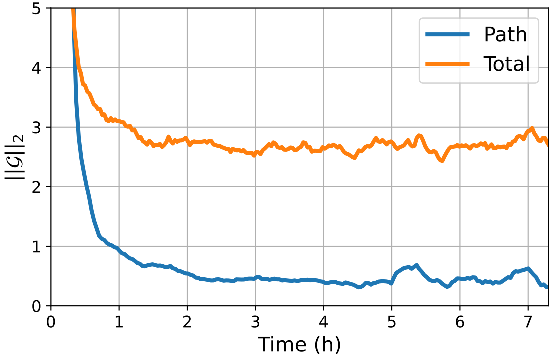
- Path Gradient

$$\frac{d}{d\theta} KL(q_\theta|p) = \mathbb{E}_{z\sim q_z}\left[\frac{\partial}{\partial g_\theta(z)}\left(\ln\frac{q_\theta(g_\theta(z))}{p(g_\theta(z))}\right)\frac{\partial g_\theta(z)}{\partial\theta} + \cancel{\frac{\partial \ln q_\theta(x)}{\partial\theta}\Big|_{x=g_\theta(z)}}\right]$$

[Roeder et al., 2017]

- Path Gradient

$$\frac{d}{d\theta} KL(q_\theta|p) = \mathbb{E}_{z\sim q_z}\left[\frac{\partial}{\partial g_\theta(z)}\left(\ln\frac{q_\theta(g_\theta(z))}{p(g_\theta(z))}\right)\frac{\partial g_\theta(z)}{\partial \theta} + \cancel{\left.\frac{\partial \ln q_\theta(x)}{\partial \theta}\right|_{x=g_\theta(z)}}\right]$$



- Path Gradient

$$\frac{d}{d\theta} KL(q_\theta | p) = \mathbb{E}_{z\sim q_z}\left[\frac{\partial}{\partial g_\theta(z)}\left(\ln\frac{q_\theta(g_\theta(z))}{p(g_\theta(z))}\right)\frac{\partial g_\theta(z)}{\partial\theta} + \cancel{\frac{\partial \ln q_\theta(x)}{\partial\theta}\bigg|_{x=g_\theta(z)}}\right]$$

- Path Gradient

$$\frac{d}{d\theta} KL(q_\theta | p) = \mathbb{E}_{z \sim q_z} \left[\frac{\partial}{\partial g_\theta(z)} \left(\ln \frac{q_\theta(g_\theta(z))}{p(g_\theta(z))} \right) \frac{\partial g_\theta(z)}{\partial \theta} + \cancel{\frac{\partial \ln q_\theta(x)}{\partial \theta} \Big|_{x = g_\theta(z)}} \right]$$

$$= \mathbb{E}_{z \sim q_z} \left[\left(\frac{\partial}{\partial g_\theta(z)} \ln q_\theta(g_\theta(z)) - \frac{\partial}{\partial g_\theta(z)} \ln p(g_\theta(z)) \right) \frac{\partial g_\theta(z)}{\partial \theta} \right]$$

- Path Gradient

$$\frac{d}{d\theta} KL(q_\theta|p) = \mathbb{E}_{z\sim q_z}\left[\frac{\partial}{\partial g_\theta(z)}\left(\ln\frac{q_\theta(g_\theta(z))}{p(g_\theta(z))}\right)\frac{\partial g_\theta(z)}{\partial\theta} + \cancel{\frac{\partial \ln q_\theta(x)}{\partial\theta}\Big|_{x=g_\theta(z)}}\right]$$

$$= \mathbb{E}_{z\sim q_z}\left[\left({\color{red}\frac{\partial}{\partial g_\theta(z)}\ln q_\theta(g_\theta(z))} - \frac{\partial}{\partial g_\theta(z)}\ln p(g_\theta(z))\right)\frac{\partial g_\theta(z)}{\partial\theta}\right]$$

- Continuous Normalizing Flow
  - Sampling

$$x \equiv z_T = g_\theta(z_0)$$
$$= z_0 + \int_0^T \mathrm{d}t\, f_\theta(z_t, t)$$

  - Density

$$\ln q_\theta(x) = \ln q_Z(z_0) - \int_0^T \mathrm{tr}\left(\frac{\partial f_\theta(z_t,t)}{\partial z_t}\right)\mathrm{d}t$$

- Path Gradient

$$\frac{d}{d\theta} KL(q_\theta | p) = \mathbb{E}_{z\sim q_z}\left[\frac{\partial}{\partial g_\theta(z)}\left(\ln \frac{q_\theta(g_\theta(z))}{p(g_\theta(z))}\right)\frac{\partial g_\theta(z)}{\partial \theta} + \cancel{\frac{\partial \ln q_\theta(x)}{\partial \theta}\bigg|_{x=g_\theta(z)}}\right]$$

$$= \mathbb{E}_{z\sim q_z}\left[\left(\underbrace{\frac{\partial}{\partial g_\theta(z)}\ln q_\theta(g_\theta(z))}_{\frac{\partial}{\partial z_T}\ln q_\theta(z_T)} - \frac{\partial}{\partial g_\theta(z)}\ln p(g_\theta(z))\right)\frac{\partial g_\theta(z)}{\partial \theta}\right]$$

- Continuous Normalizing Flow
  - Sampling

$$x \equiv z_T = g_\theta(z_0)$$
$$= z_0 + \int_0^T \mathrm{d}t\, f_\theta(z_t, t)$$

  - Density

$$\ln q_\theta(x) = \ln q_Z(z_0) - \int_0^T \mathrm{tr}\left(\frac{\partial f_\theta(z_t, t)}{\partial z_t}\right)\mathrm{d}t$$

Theorem

The derivative $\frac{\partial \ln q_\theta(z_T)}{\partial z_T}$ *can be obtained by solving the initial value problem*

$$\frac{d}{dt}\frac{\partial \ln q_\theta(z_t)}{\partial z_t} = -\frac{\partial \ln q_\theta(z_t)}{\partial z_t}^{\intercal}\frac{\partial f_\theta(z_t,t)}{\partial z_t} - \partial_{z_t} tr\left(\frac{\partial f_\theta(z_t,t)}{\partial z_t}\right), \tag{1}$$

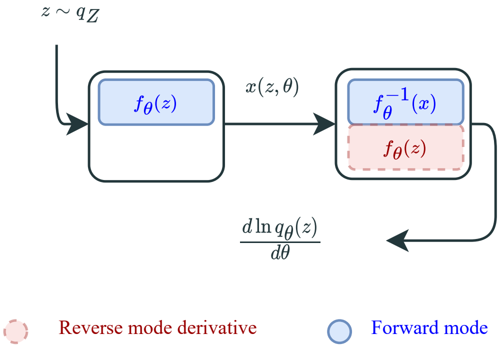
with initial condition

$$\frac{\partial \ln q_\theta(z_0)}{\partial z_0} = \frac{\partial \ln q_Z(z_0)}{\partial z_0}.$$

Path gradients for CNFs

CNF Total Gradient

(Chen et al. 2018)

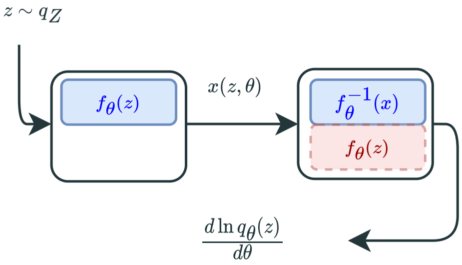


Lorenz Vaitl, Kim Nicoli, Shinichi Nakajima, Pan Kessel

Path gradients for CNFs

CNF Total Gradient

(Chen et al. 2018)

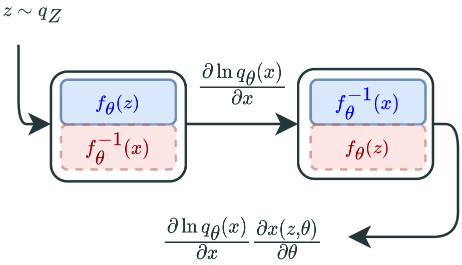


Reverse mode derivative

Forward mode

CNF Path Gradient

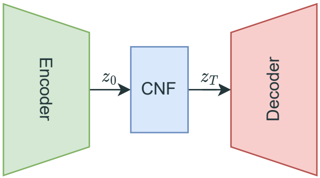
(ours)



Reverse mode derivative

Forward mode

Results VAE



[Grathwohl et al., 2019]

-ELBO	Path	Total
MNIST	**82.09** $\pm .04$	$82.82 \pm .01$
Omniglot	**96.61** $\pm .17$	$98.33 \pm .09$
Caltech Silhouettes	**101.93** $\pm .63$	$104.03 \pm .43$
Frey Faces	**4.35** $\pm .00$	$4.39 \pm .01$

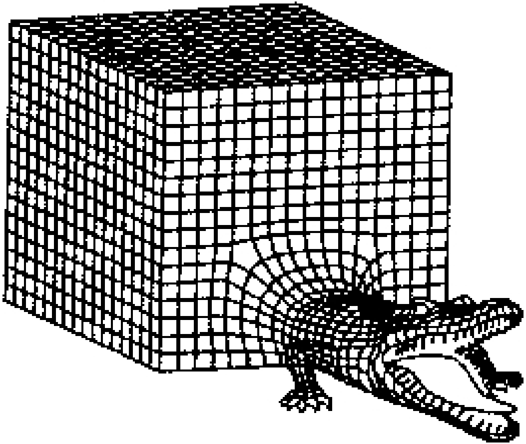
Lattice Field Theory:

- Target

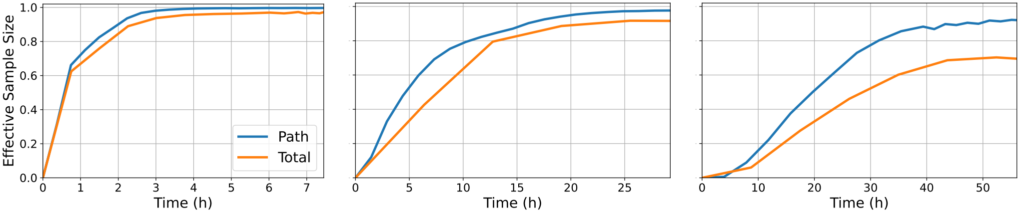
$$p(x) = \frac{1}{Z} e^{-S(x)}$$

  - Intractable
  - Known in closed form

- Can be approximated by CNF with inductive biases [de Haan et al., 2021]



Lattice size	Path	Total
12x12	**99.66**$\% \pm 0.07$	$98.01\% \pm 0.44$
20x20	**97.65**$\% \pm 0.14$	$91.56\% \pm 1.13$
32x32	**91.81**$\% \pm 1.32$	$69.53\% \pm 5.59$



Main idea:

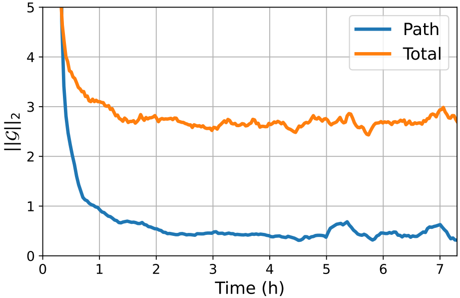
- Variational Inference:

$$KL(q_\theta | p)$$

- Continuous Normalizing Flow q_θ
- Minimized by SGD:

$$\theta \leftarrow \theta - \lambda \underbrace{\nabla_\theta KL(q_\theta | p)}_{\approx \mathcal{G}_{\text{total}} \; \mathcal{G}_{\text{path}}}$$

- Leads to better performance as demonstrated in our experiments for VAEs and Lattice Field Theory



-ELBO	Path (ours)	Total
MNIST	**82.09** $\pm .04$	$82.82 \pm .01$
Omniglot	**96.61** $\pm .17$	$98.33 \pm .09$
Caltech Silhouettes	**101.93** $\pm .63$	$104.03 \pm .43$
Frey Faces	**4.35** $\pm .00$	$4.39 \pm .01$

Thank you for your attention

References i

de Haan, P., Rainone, C., Cheng, M. C. N., and Bondesan, R. (2021).
Scaling up machine learning for quantum field theory with equivariant continuous flows.
CoRR, abs/2110.02673.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., and Duvenaud, D. (2019).
Scalable reversible generative models with free-form continuous dynamics.
In *International Conference on Learning Representations*.

References ii

Roeder, G., Wu, Y., and Duvenaud, D. (2017).
Sticking the landing: Simple, lower-variance gradient estimators for variational inference.
In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6925–6934.