UNIVERSITEIT VAN AMSTERDAM

Self Normalizing Flows

T. Anderson Keller, Jorn Peters, Priyank Jaini,
Emiel Hoogeboom, Patrick Forré, Max Welling

https://arxiv.org/abs/2011.07248

International Conference on Machine Learning
July 18th 2021

UvA - BOSCH
DELTA LAB

Self Normalizing Flows

T. Anderson Keller

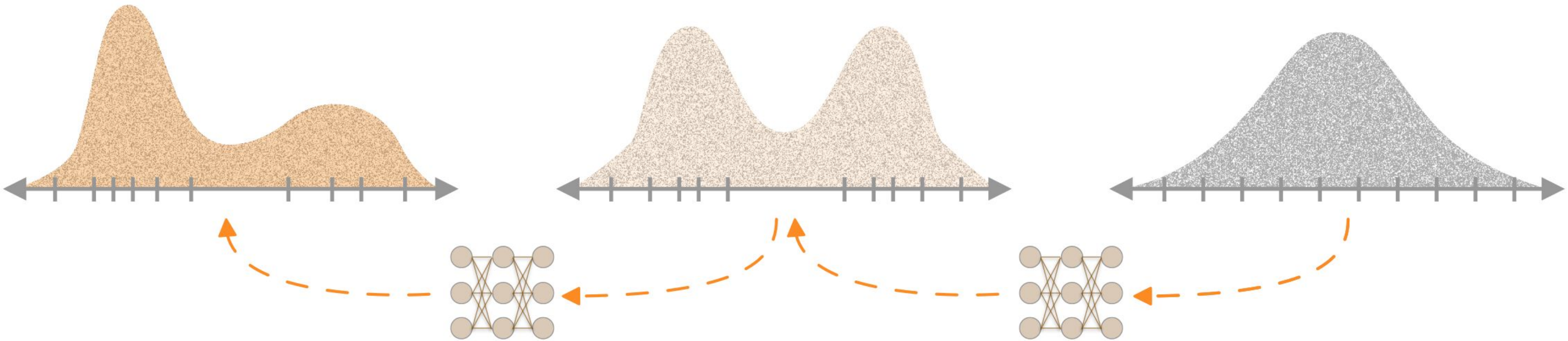
Jorn Peters

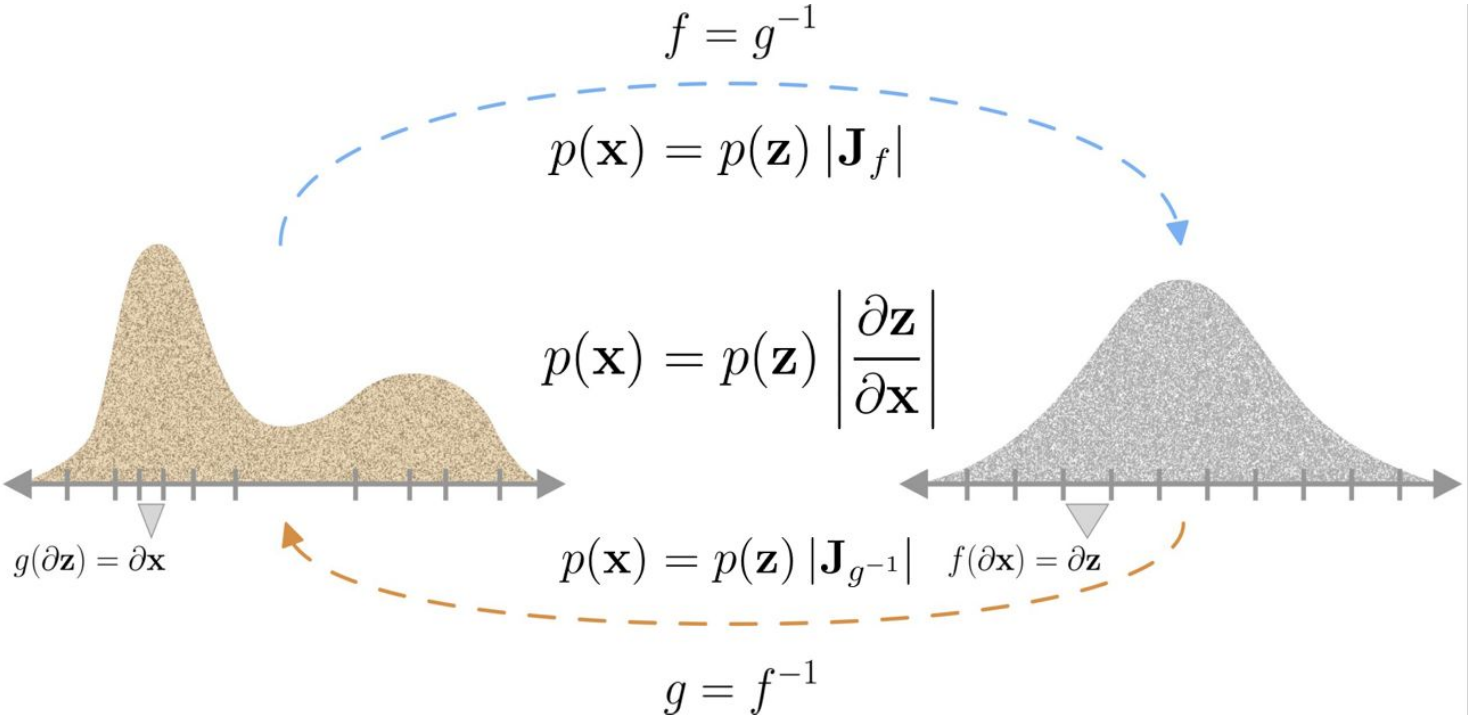
Priyank Jaini

Emiel Hoogeboom

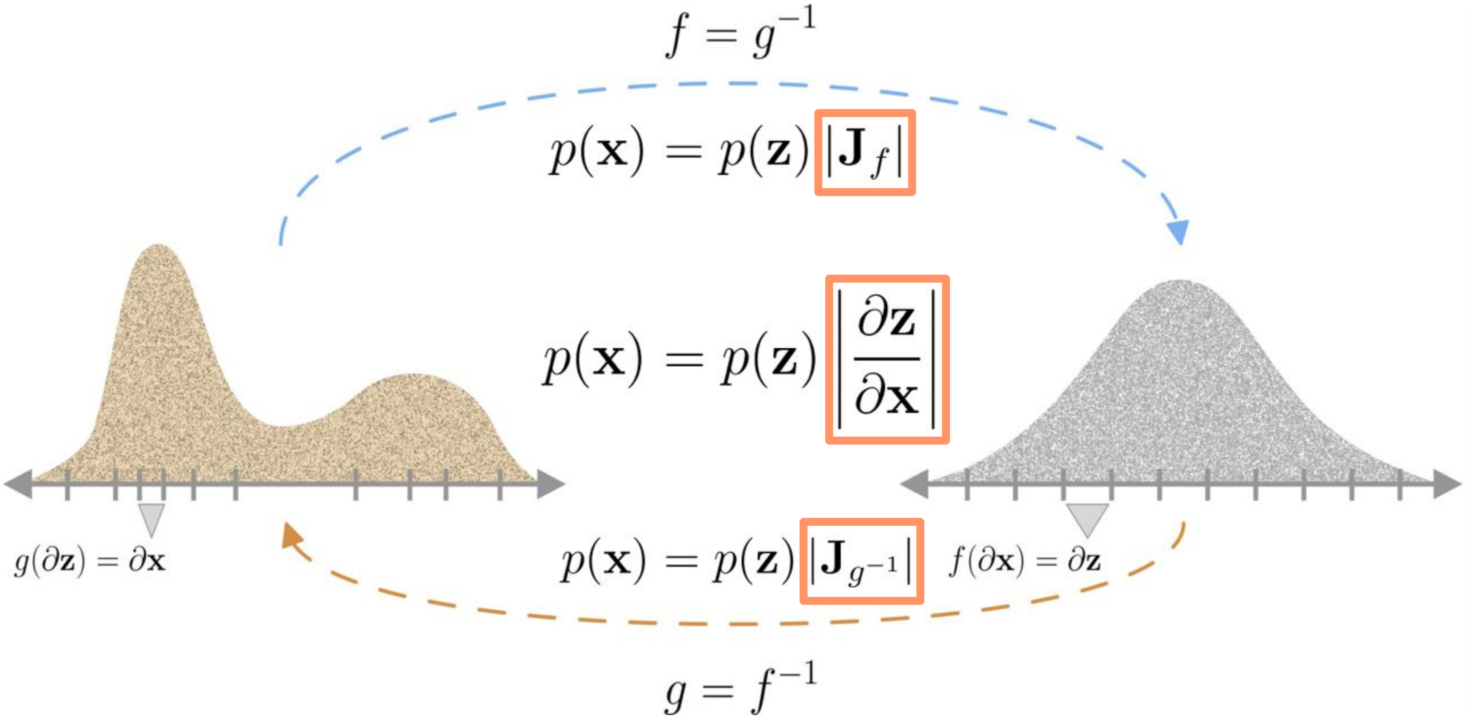
Patrick Forré

Max Welling





$f = g^{-1}$
$p(\mathbf{x}) = p(\mathbf{z})\,|\mathbf{J}_f|$
$p(\mathbf{x}) = p(\mathbf{z}) \left|\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right|$
$g(\partial \mathbf{z}) = \partial \mathbf{x}$
$p(\mathbf{x}) = p(\mathbf{z})\,|\mathbf{J}_{g^{-1}}|$
$f(\partial \mathbf{x}) = \partial \mathbf{z}$
$g = f^{-1}$



$f = g^{-1}$
$p(\mathbf{x}) = p(\mathbf{z}) \left|\mathbf{J}_f\right|$
$p(\mathbf{x}) = p(\mathbf{z}) \left|\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right|$
$g(\partial \mathbf{z}) = \partial \mathbf{x}$
$p(\mathbf{x}) = p(\mathbf{z}) \left|\mathbf{J}_{g^{-1}}\right|$
$f(\partial \mathbf{x}) = \partial \mathbf{z}$
$g = f^{-1}$

Prior Work

- NICE (Non-linear independent components estimation) (Dinh et al., 2015)
- Real non-volume preserving flow (Real NVP) (Dinh et al., 2017)
- Inverse autoregressive flow (IAF) (Kingma et al., 2016)
- Masked autoregressive flow (MAF) (Papamakarios et al., 2017)
- Glow (Kingma and Dhariwal, 2018)
- Neural Autoregressive Flow (NAF) (Huang et al., 2018)
- block-NAF (B-NAF) (De Cao et al., 2019)
- Flow++ (Ho et al., 2019)
- Sums-of-squares Polynomial transformer (Jaini et al., 2019)

Prior Work

- Neural Spline Flows (Durkan et al., 2019)
- Residual Flows (Chen et al., 2019)
- Invertible Residual Networks (Jens Behrmann et al., 2018)
- Sylvester Flows (van den Berd et al., 2018)
- Radial Flows (Tabak and Turner, 2013)
- Planar Flows (Rezende and Mohamed, 2015)
- Emerging Convolutions (Hoogeboom et al., 2019)
- Integer Discrete Flows (Hoogeboom et al., 2019)
- The Convolution Exponential (Hoogeboom et al., 2020)

$$\frac{\partial \log |\mathbf{J}_f|}{\partial \mathbf{J}_f}$$

$$\frac{\partial \log |\mathbf{J}_f|}{\partial \mathbf{J}_f} = \mathbf{J}_f^{-T}$$

$$\frac{\partial \log |\mathbf{J}_f|}{\partial \mathbf{J}_f} = \mathbf{J}_f^{-T} = \mathbf{J}_{f^{-1}}^T$$

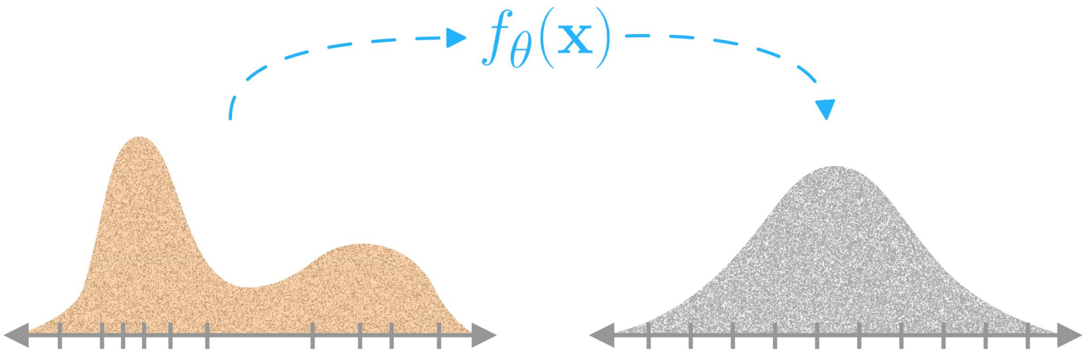
$$\frac{\partial \log |\mathbf{J}_f|}{\partial \mathbf{J}_f} = \mathbf{J}_f^{-T} = \mathbf{J}_{f^{-1}}^T$$

$$\text{if } g \approx f^{-1}$$

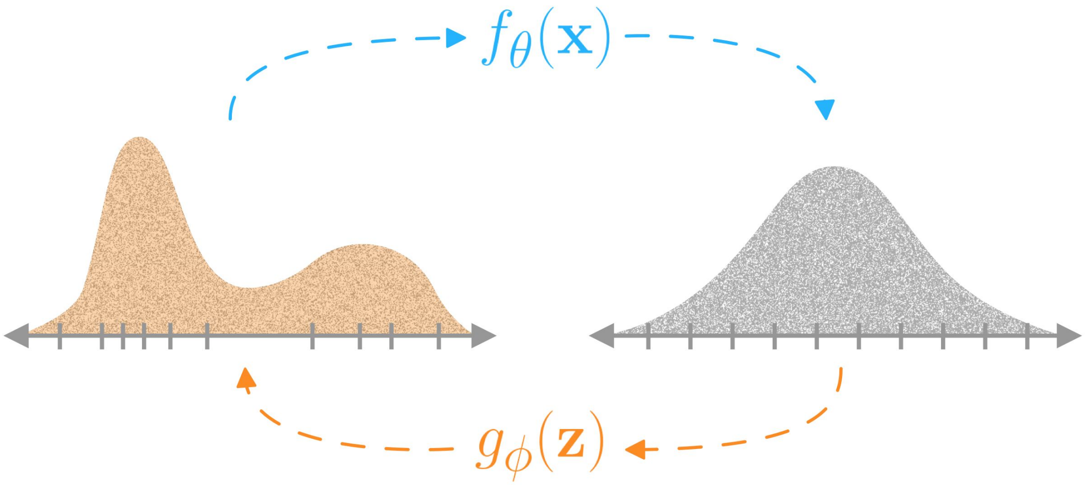
$$\frac{\partial \log |\mathbf{J}_f|}{\partial \mathbf{J}_f} = \mathbf{J}_f^{-T} = \mathbf{J}_{f^{-1}}^T \approx \mathbf{J}_g^T$$

$$\text{if } g \approx f^{-1}$$

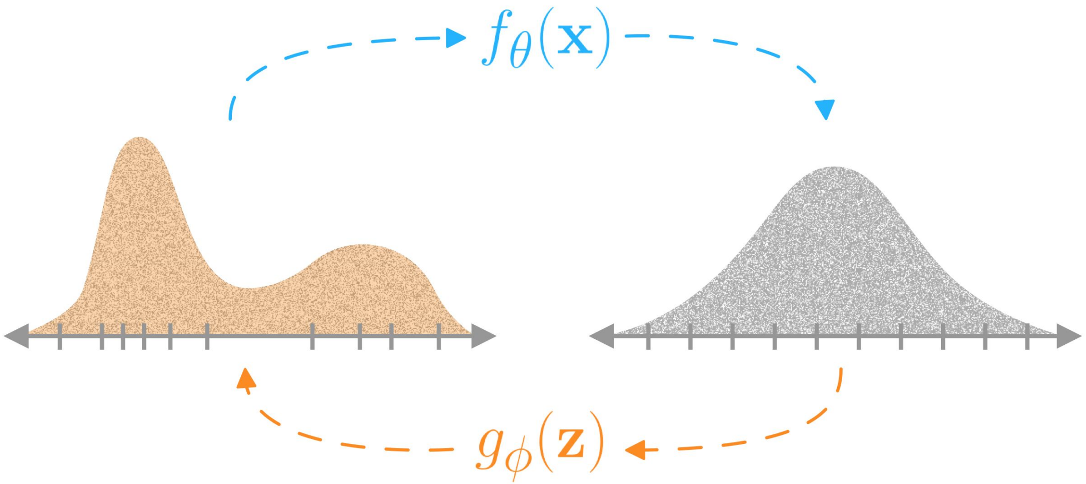
Self-Normalizing Flows



Self-Normalizing Flows

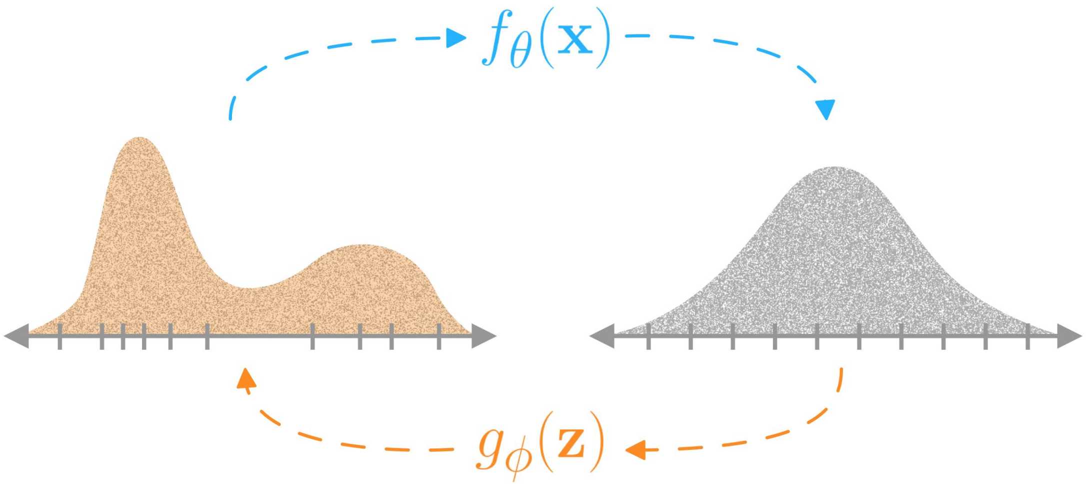


Self-Normalizing Flows



$$\mathcal{L}(\mathbf{x}) = ||g_\phi(f_\theta(\mathbf{x})) - \mathbf{x}||_2^2$$

Self-Normalizing Flows



$$\mathcal{L}(\mathbf{x}) = ||g_\phi(f_\theta(\mathbf{x})) - \mathbf{x}||_2^2$$

$$\log p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2} \log p^f_{\mathbf{X}}(\mathbf{x}) + \frac{1}{2} \log p^g_{\mathbf{X}}(\mathbf{x})$$

Self-Normalizing Flows

$$\log p_{\mathbf{X}}^{f}(\mathbf{x}) = \log p_{\mathbf{Z}}\left(f_{\theta}(\mathbf{x})\right) + \log\left|\mathbf{J}_{f}\right|$$

$$\log p_{\mathbf{X}}^{g}(\mathbf{x}) = \log p_{\mathbf{Z}}\left(g_{\phi}^{-1}(\mathbf{x})\right) + \log\left|\mathbf{J}_{g^{-1}}\right|$$

Self-Normalizing Flows

$$\log p_{\mathbf{X}}^{f}(\mathbf{x}) = \log p_{\mathbf{Z}}\left(f_{\theta}(\mathbf{x})\right) + \log |\mathbf{J}_{f}|$$

$$\frac{\partial}{\partial\theta} \log p_{\mathbf{X}}^{f}(\mathbf{x}) = \frac{\partial}{\partial\theta} \log p_{\mathbf{Z}}\left(f_{\theta}(\mathbf{x})\right) + \frac{\partial(\text{vec } \mathbf{J}_{f})^{T}}{\partial\theta}(\text{vec } \mathbf{J}_{f}^{-T})$$

$$\log p_{\mathbf{X}}^{g}(\mathbf{x}) = \log p_{\mathbf{Z}}\left(g_{\phi}^{-1}(\mathbf{x})\right) + \log |\mathbf{J}_{g^{-1}}|$$

$$\frac{\partial}{\partial\phi} \log p_{\mathbf{X}}^{g}(\mathbf{x}) = \frac{\partial}{\partial\phi} \log p_{\mathbf{Z}}\left(g_{\phi}^{-1}(\mathbf{z})\right) + \frac{\partial(\text{vec } \mathbf{J}_{g^{-1}})^{T}}{\partial\phi}(\text{vec } \mathbf{J}_{g^{-1}}^{-T})$$

Self-Normalizing Flows

$$\log p_{\mathbf{X}}^{f}(\mathbf{x}) = \log p_{\mathbf{Z}}\left(f_{\theta}(\mathbf{x})\right) + \log |\mathbf{J}_f|$$

$$\frac{\partial}{\partial\theta} \log p_{\mathbf{X}}^{f}(\mathbf{x}) = \frac{\partial}{\partial\theta} \log p_{\mathbf{Z}}\left(f_{\theta}(\mathbf{x})\right) + \frac{\partial(\text{vec } \mathbf{J}_f)^T}{\partial\theta}(\text{vec } \mathbf{J}_f^{-T})$$

$$\approx \frac{\partial}{\partial\theta} \log p_{\mathbf{Z}}\left(f_{\theta}(\mathbf{x})\right) + \frac{\partial(\text{vec } \mathbf{J}_f)^T}{\partial\theta}(\text{vec } \mathbf{J}_g^{T})$$

$$\log p_{\mathbf{X}}^{g}(\mathbf{x}) = \log p_{\mathbf{Z}}\left(g_{\phi}^{-1}(\mathbf{x})\right) + \log |\mathbf{J}_{g^{-1}}|$$

$$\frac{\partial}{\partial\phi} \log p_{\mathbf{X}}^{g}(\mathbf{x}) = \frac{\partial}{\partial\phi} \log p_{\mathbf{Z}}\left(g_{\phi}^{-1}(\mathbf{z})\right) + \frac{\partial(\text{vec } \mathbf{J}_{g^{-1}})^T}{\partial\phi}(\text{vec } \mathbf{J}_{g^{-1}}^{-T})$$

$$\approx \frac{\partial}{\partial\phi} \log p_{\mathbf{Z}}\left(g_{\phi}^{-1}(\mathbf{z})\right) - \frac{\partial(\text{vec } \mathbf{J}_g)^T}{\partial\phi}(\text{vec } \mathbf{J}_f^{T})$$

Fully-Connected

$f(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x} = \boldsymbol{z}$

$g(\boldsymbol{z}) = \boldsymbol{R}\boldsymbol{z}$

$\boldsymbol{W}^{-1} \approx \boldsymbol{R}$

$$\frac{\partial}{\partial \mathbf{W}} \log p^f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{W}} \log p_{\mathbf{Z}}(\mathbf{W}\mathbf{x}) + \mathbf{W}^{-T}$$

$$\approx \delta_{\mathbf{z}}\mathbf{x}^T + \mathbf{R}^T$$

$$\frac{\partial}{\partial \mathbf{R}} \log p^g_{\mathbf{X}}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{R}} \log p_{\mathbf{Z}}(\mathbf{R}^{-1}\mathbf{x}) - \mathbf{R}^{-T}$$

$$\approx -\delta_{\mathbf{x}}\mathbf{z}^T - \mathbf{W}^T$$

$\boldsymbol{\delta}_{\boldsymbol{x}} = \frac{\partial \log p_{\boldsymbol{Z}}(\boldsymbol{z})}{\partial \boldsymbol{x}}$

$\boldsymbol{\delta}_{\boldsymbol{z}} = \frac{\partial \log p_{\boldsymbol{Z}}(\boldsymbol{z})}{\partial \boldsymbol{z}}$

Fully-Connected

$f(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x} = \boldsymbol{z}$

$g(\boldsymbol{z}) = \boldsymbol{R}\boldsymbol{z}$

$\boldsymbol{W}^{-1} \approx \boldsymbol{R}$

$$\frac{\partial}{\partial \mathbf{W}} \log p^f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{W}} \log p_{\mathbf{Z}}(\mathbf{W}\mathbf{x}) + \mathbf{W}^{-T}$$
$$\approx \delta_{\mathbf{z}}\mathbf{x}^T + \mathbf{R}^T$$

$$\frac{\partial}{\partial \mathbf{R}} \log p^g_{\mathbf{X}}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{R}} \log p_{\mathbf{Z}}(\mathbf{R}^{-1}\mathbf{x}) - \mathbf{R}^{-T}$$
$$\approx -\delta_{\mathbf{x}}\mathbf{z}^T - \mathbf{W}^T$$

$\boldsymbol{\delta}_{\boldsymbol{x}} = \frac{\partial \log p_{\boldsymbol{Z}}(\boldsymbol{z})}{\partial \boldsymbol{x}}$

$\boldsymbol{\delta}_{\boldsymbol{z}} = \frac{\partial \log p_{\boldsymbol{Z}}(\boldsymbol{z})}{\partial \boldsymbol{z}}$

Convolutional

$f(\boldsymbol{x}) = \boldsymbol{w} \star \boldsymbol{x} = \boldsymbol{z}$

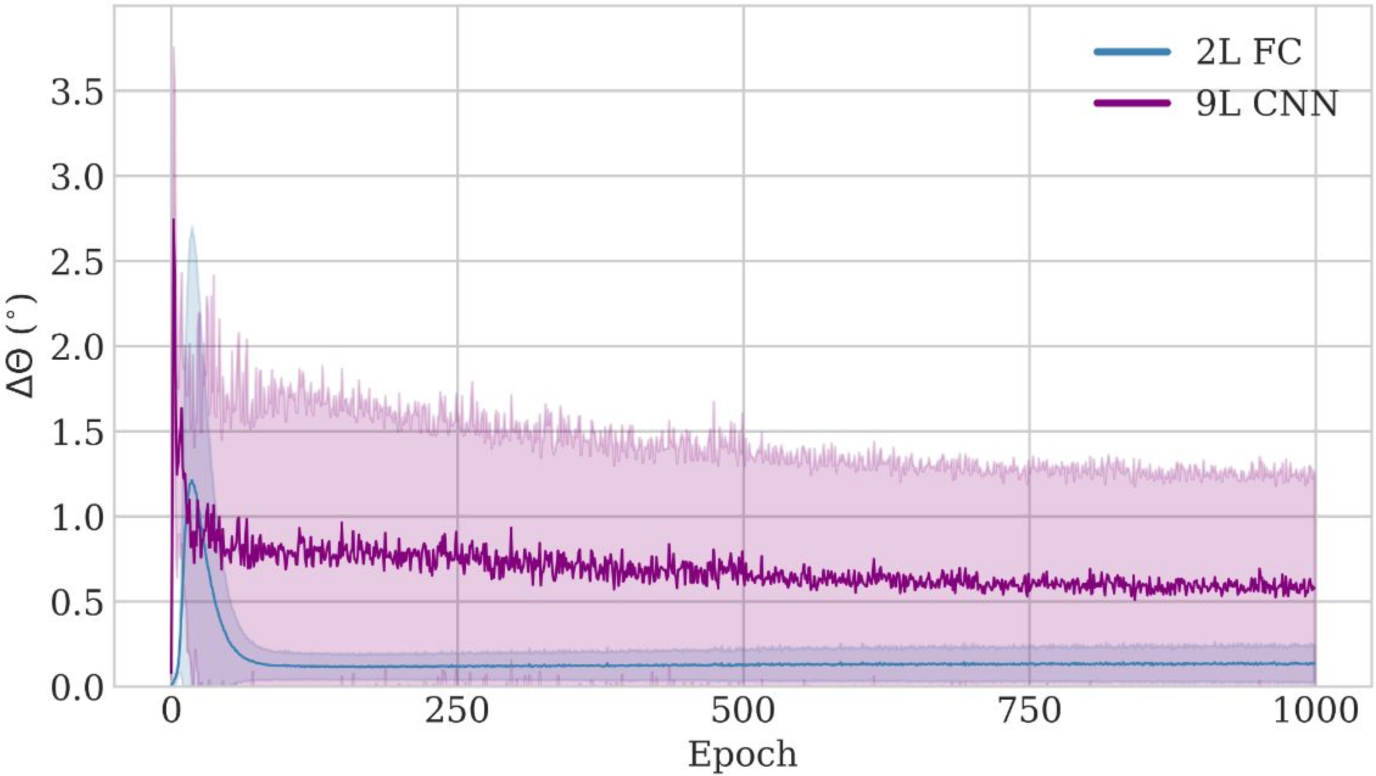
$g(\boldsymbol{z}) = \boldsymbol{r} \star \boldsymbol{z}$

$f^{-1} \approx g$

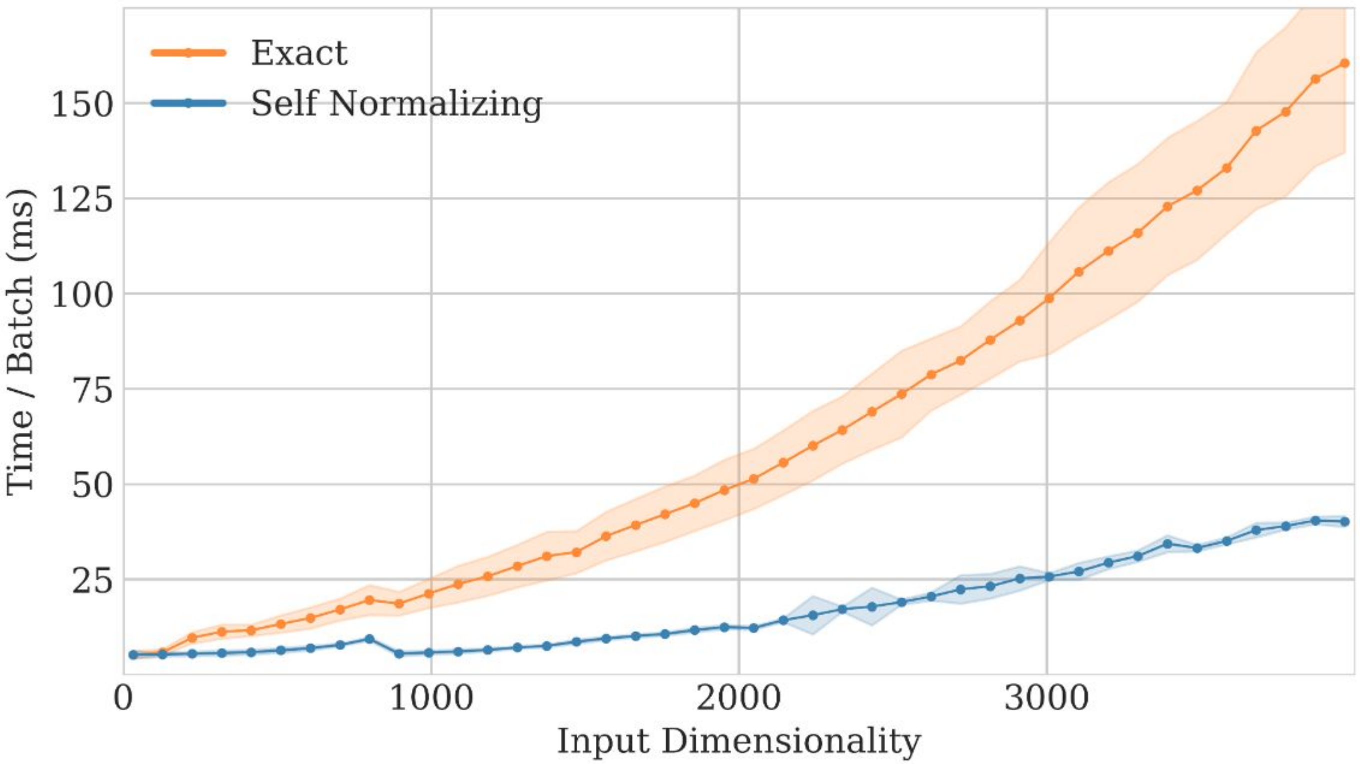
$$\frac{\partial}{\partial \boldsymbol{w}} \log p^f_{\boldsymbol{X}}(\boldsymbol{x}) = \boldsymbol{\delta}^f_{\boldsymbol{z}} \star \boldsymbol{x} + \frac{\partial (\text{vec}\, \mathcal{T}(\boldsymbol{w}))^T}{\partial \boldsymbol{w}} \left(\text{vec}\, \mathcal{T}(\boldsymbol{w})^{-T}\right)$$
$$\approx \boldsymbol{\delta}_{\boldsymbol{z}} \star \boldsymbol{x} + \text{flip}(\boldsymbol{r}) \odot \boldsymbol{m}$$

$$\frac{\partial}{\partial \boldsymbol{r}} \log p^g_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\partial (\text{vec}\, \mathcal{T}(\boldsymbol{r}))^T}{\partial \boldsymbol{r}} \left(\text{vec}\, \left[-\mathcal{T}(\boldsymbol{r})^{-T} \boldsymbol{\delta}^g_{\boldsymbol{z}} \boldsymbol{x}^T \mathcal{T}(\boldsymbol{r})^{-T}\right] - \text{vec}\, \mathcal{T}(\boldsymbol{r})^{-T}\right)$$
$$\approx -\boldsymbol{\delta}_{\boldsymbol{x}} \star \boldsymbol{z} - \text{flip}(\boldsymbol{w}) \odot \boldsymbol{m}$$

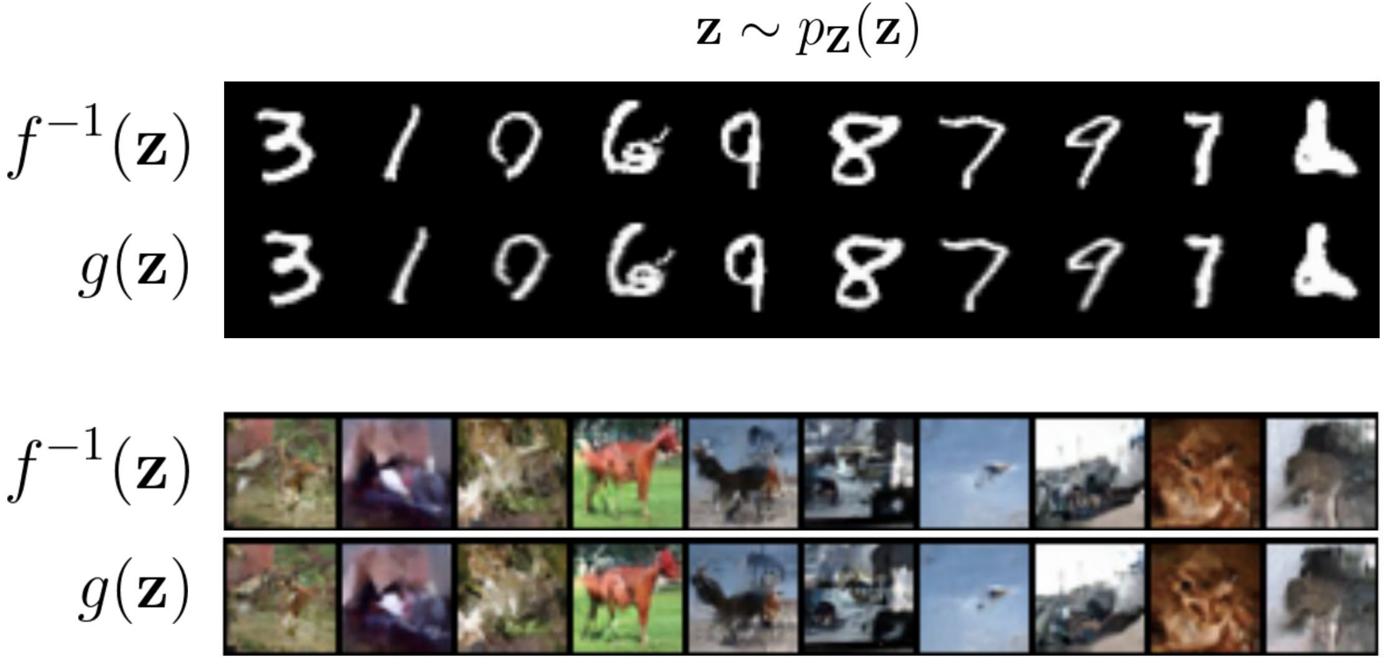
Experiments



Experiments



Experiments



Experiments

Model	$-\log p_{\boldsymbol{X}}(\boldsymbol{x})$
Relative Grad. FC 2-Layer [13]	1096.5 ± 0.5
Exact Gradient FC 2-Layer	947.6 ± 0.2
SNF FC 2-Layer (ours)	947.1 ± 0.2
Emerging Conv. 9-Layer [16]	645.7 ± 3.6
SNF Conv. 9-Layer (ours)	638.6 ± 0.9
Conv. Exponential 9-Layer [15]	638.1 ± 1.0
Exact Gradient Conv. 9-Layer	637.4 ± 0.2
Glow-like 32-Layer [20]	575.7 ± 0.8
SNF Glow 32-Layer (ours)	575.4 ± 1.4

Experiments

Model	CIFAR-10	ImageNet32
Glow	3.36 ± 0.002	4.12 ± 0.002
SNF Glow	3.37 ± 0.004	4.14 ± 0.007

Thank you!

Paper: https://arxiv.org/abs/2011.07248

Code: https://github.com/akandykeller/SelfNormalizingFlows

Blog: http://keller.org/research/2020-10-21-self-normalizing-flows/

Contact: T.Anderson.Keller@Gmail.com