

# Gradient estimators for normalising flows

## Training normalizing flows without action derivative

Piotr Białas

Institute of Applied Computer Science  
Faculty of Physics, Astronomy and Applied Computer Science  
Jagiellonian University, Kraków, Poland

Numerical Challenges in Lattice QCD  
Meinerzhagen, 16 August 2022

with P. Korcyl and T. Stebel arXiv:2202.01314

# Outline

- Neural Markov Chain Monte-Carlo
- Gradient estimators
  - Autoregressive networks
  - Normalizing flows
- Results ( $\phi^4$ )

# Independent Metropolised Sampler

$$p(\phi) = Z^{-1} e^{-\beta S(\phi)}, \quad P(\phi) = Z \cdot p(\phi)$$

Jun S. Liu. "Metropolized independent sampling with comparisons to rejection sampling and importance sampling". *Statistics and Computing* 6.2 (1996), pp. 113–119.

# Independent Metropolised Sampler

$$p(\phi) = Z^{-1} e^{-\beta S(\phi)}, \quad P(\phi) = Z \cdot p(\phi)$$

$$\phi_{trial} \sim q(\cdot)$$

Jun S. Liu. "Metropolized independent sampling with comparisons to rejection sampling and importance sampling". Statistics and Computing 6.2 (1996), pp. 113–119.

# Independent Metropolised Sampler

$$p(\phi) = Z^{-1} e^{-\beta S(\phi)}, \quad P(\phi) = Z \cdot p(\phi)$$

$$\phi_{trial} \sim q(\cdot)$$

$$p_a(\phi_{trial} | \phi_i) = \min \left\{ 1, \frac{p(\phi_{trial})}{q(\phi_{trial})} \frac{q(\phi_i)}{p(\phi_i)} \right\}$$

Jun S. Liu. "Metropolized independent sampling with comparisons to rejection sampling and importance sampling". Statistics and Computing 6.2 (1996), pp. 113–119.

## Learning $q(\phi)$

$$q(\phi) = q(\phi|\theta)$$

# Learning $q(\phi)$

$$q(\phi) = q(\phi|\theta)$$

$$\operatorname{argmin}_{\theta} D_{KL}(q(\cdot|\theta)||p)$$

## Kullback-Leibler divergence

$$D_{KL}(q_{\theta} \| p) = \int d\phi q(\phi|\theta) \log \frac{q(\phi|\theta)}{p(\phi)}$$

## Kullback-Leibler divergence

$$D_{KL}(q_{\theta} \| p) = \int d\phi q(\phi|\theta) \log \frac{q(\phi|\theta)}{p(\phi)}$$

$$D_{KL}(q_{\theta} \| p) \geq 0, \quad D_{KL}(q_{\theta} \| p) = 0 \iff p = q$$

## Kullback-Leibler divergence

$$D_{KL}(q_{\theta} \| p) = \int d\phi q(\phi | \theta) \log \frac{q(\phi | \theta)}{p(\phi)}$$

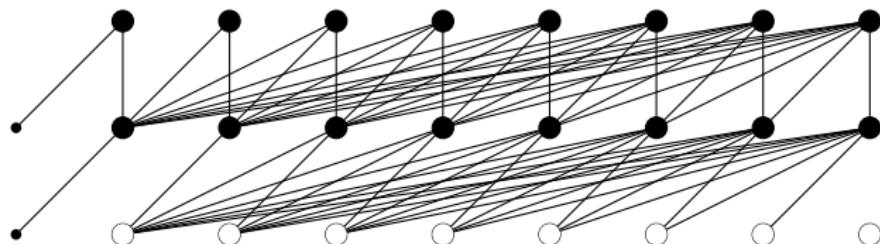
$$D_{KL}(q_{\theta} \| p) \geq 0, \quad D_{KL}(q_{\theta} \| p) = 0 \iff p = q$$

$$D_{KL}(q_{\theta} \| p) \neq D_{KL}(p \| q_{\theta})$$

# Autoregressive networks

$$q(s_1, s_2, \dots, s_N | \theta)$$

$$= q(s_1 | \theta) \cdot q(s_2 | s_1, \theta) \cdot q(s_3 | s_1, s_2, \theta) \cdots q(s_N | s_1, \dots, s_{N-1}, \theta)$$



Dian Wu, Lei Wang, and Pan Zhang. "Solving Statistical Mechanics Using Variational Autoregressive Networks". Phys. Rev. Lett. 122 (2019), p. 080602.

# Ancestral sampling

$$q(s_1, s_2, \dots, s_N | \theta)$$

$$= q(s_1 | \theta) \cdot q(s_2 | s_1, \theta) \cdot q(s_3 | s_1, s_2, \theta) \cdots q(s_N | s_1, \dots, s_{N-1}, \theta)$$

$$\begin{aligned} s_1 &\sim q(\cdot | \theta) \\ s_2 &\sim q(\cdot | s_1, \theta) \\ s_3 &\sim q(\cdot | s_1, s_2, \theta) \\ &\vdots \\ s_N &\sim q(\cdot | s_1, \dots, s_{N-1}, \theta) \end{aligned}$$

## $D_{KL}$ gradient

$$\phi = (s_1, \dots, s_N)$$

$$\begin{aligned} \frac{dD_{KL}(q|p)}{d\theta} &= \int d\phi \frac{\partial q(\phi|\theta)}{\partial \theta} (\log q(\phi|\theta) - \log p(\phi)) \\ &\quad + \int d\phi q(\phi|\theta) \frac{\partial}{\partial \theta} \log q(\phi|\theta) \end{aligned}$$

## $D_{KL}$ gradient

$$\begin{aligned}\frac{dD_{KL}(q||p)}{d\theta} &= \int d\phi \frac{\partial q(\phi|\theta)}{\partial \theta} (\log q(\phi|\theta) - \log p(\phi)) \\ &\quad + \int d\phi q(\phi|\theta) \frac{\partial}{\partial \theta} \log q(\phi|\theta)\end{aligned}$$

$$\int d\phi \frac{\partial q(\phi|\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \underbrace{\int d\phi q(\phi|\theta)}_1 = 0$$

Dian Wu, Lei Wang, and Pan Zhang. "Solving Statistical Mechanics Using Variational Autoregressive Networks". Phys. Rev. Lett. 122 (2019), p. 080602.

# Approximate gradient

$$\frac{dD_{KL}(q||p)}{d\theta} = \int d\phi q(\phi|\theta) \frac{\partial \log q(\phi|\theta)}{\partial \theta} (\log q(\phi|\theta) - \log p(\phi))$$

Dian Wu, Lei Wang, and Pan Zhang. "Solving Statistical Mechanics Using Variational Autoregressive Networks". Phys. Rev. Lett. 122 (2019), p. 080602.

# Approximate gradient

$$\frac{dD_{KL}(q|p)}{d\theta} = \int d\phi q(\phi|\theta) \frac{\partial \log q(\phi|\theta)}{\partial \theta} (\log q(\phi|\theta) - \log p(\phi))$$

$$\frac{dD_{KL}(q|p)}{d\theta} \approx \mathbf{g}_1[\{\phi\}] \equiv \frac{1}{N} \sum_{i=1}^N \frac{\partial \log q(\phi_i|\theta)}{\partial \theta} (\log q(\phi_i|\theta) - \log p(\phi_i))$$
$$\phi \sim q(\phi|\theta)$$

Dian Wu, Lei Wang, and Pan Zhang. "Solving Statistical Mechanics Using Variational Autoregressive Networks". Phys. Rev. Lett. 122 (2019), p. 080602.

# REINFORCE

$$\mathbf{g}_1[\{\phi\}] = \frac{1}{N} \sum_{i=1}^N \frac{\partial \log q(\phi_i|\theta)}{\partial \theta} \underbrace{(\log q(\phi_i|\theta) - \log p(\phi_i))}_{s_i}$$
$$\phi \sim q(\phi|\theta)$$

# REINFORCE

$$\mathbf{g}_1[\{\phi\}] = \frac{1}{N} \sum_{i=1}^N \frac{\partial \log q(\phi_i|\theta)}{\partial \theta} \underbrace{(\log q(\phi_i|\theta) - \log p(\phi_i))}_{s_i}$$
$$\phi \sim q(\phi|\theta)$$

$$\mathbf{g}_2[\{\phi\}] = \frac{1}{N} \sum_{i=1}^N \frac{\partial \log q(\phi_i|\theta)}{\partial \theta} (s_i - \bar{s})$$
$$\bar{s} = \frac{1}{N} \sum_{i=1}^N s_i$$

Dian Wu, Lei Wang, and Pan Zhang. "Solving Statistical Mechanics Using Variational Autoregressive Networks". Phys. Rev. Lett. 122 (2019), p. 080602.

$\mathbf{g}_1$  &  $\mathbf{g}_2$  variance

$$E[\mathbf{g}_2[\{\phi\}]] = \frac{N-1}{N} E[\mathbf{g}_1[\{\phi\}]].$$

$\mathbf{g}_1$  &  $\mathbf{g}_2$  variance

$$E[\mathbf{g}_2[\{\phi\}]] = \frac{N-1}{N} E[\mathbf{g}_1[\{\phi\}]].$$

$$\text{var}[\mathbf{g}_1[\{\phi\}]]_{q(\phi|\theta)=p(\phi)} = \frac{1}{N} (\log Z)^2 \text{var} \left[ \frac{\partial \log q(\phi|\theta)}{\partial \theta} \right]_{q(\phi|\theta)=p(\phi)}$$

$\mathbf{g}_1$  &  $\mathbf{g}_2$  variance

$$E[\mathbf{g}_2[\{\phi\}]] = \frac{N-1}{N} E[\mathbf{g}_1[\{\phi\}]].$$

$$\text{var}[\mathbf{g}_1[\{\phi\}]]_{q(\phi|\theta)=p(\phi)} = \frac{1}{N} (\log Z)^2 \text{var} \left[ \frac{\partial \log q(\phi|\theta)}{\partial \theta} \right]_{q(\phi|\theta)=p(\phi)}$$

$$\text{var}[\mathbf{g}_2[\{\phi\}]]_{q(\phi|\theta)=p(\phi)} = 0$$

# Normalising flows

$$\mathbb{R}^D \ni \mathbf{z} \longrightarrow (\mathbf{q}_{pr}(\mathbf{z}), \varphi(\mathbf{z}|\boldsymbol{\theta})) \in (\mathbb{R}, \mathbb{R}^D)$$

↑  
bijection

$$\phi = \varphi(\mathbf{z}|\boldsymbol{\theta}), \quad q(\phi|\boldsymbol{\theta}) \equiv \mathbf{q}_{pr}(\mathbf{z}) J(\mathbf{z}|\boldsymbol{\theta})^{-1}$$

$$J(\mathbf{z}|\boldsymbol{\theta}) = \det \left( \frac{\partial \varphi(\mathbf{z}|\boldsymbol{\theta})}{\partial \mathbf{z}} \right)$$

Ivan Kobyzev, Simon Prince, and Marcus Brubaker. "Normalizing Flows: An Introduction and Review of Current Methods". IEEE Transactions on Pattern Analysis and Machine Intelligence(2020), pp. 1.

M. S. Albergo, G. Kanwar, and P. E. Shanahan. "Flow-based generative models for Markov chain Monte Carlo in lattice field theory". Phys. Rev. D100 (2019), p. 034515.

Michael S. Albergo et al. "Introduction to Normalizing Flows for Lattice Field Theory" (2021) arXiv:2101.08176

## Normalising flows - K-L divergence

$$D_{KL}(q|p) = \int d\mathbf{z} q_{pr}(\mathbf{z}) (\log q_z(\mathbf{z}|\theta) - \log p(\varphi(\mathbf{z}|\theta)))$$

## Normalising flows - K-L divergence

$$D_{KL}(q|p) = \int d\mathbf{z} q_{pr}(\mathbf{z}) (\log q_z(\mathbf{z}|\theta) - \log p(\varphi(\mathbf{z}|\theta)))$$

$$\frac{d D_{KL}(q|p)}{d\theta} = \int d\mathbf{z} q_{pr}(\mathbf{z}) \frac{d}{d\theta} (\log q(\mathbf{z}|\theta) - \log p(\varphi(\mathbf{z}|\theta)))$$

## Action derivative

$$\frac{dD_{KL}(q||p)}{d\theta} \approx \mathbf{g}_3[\{\phi\}] \equiv \frac{1}{N} \sum_{i=1}^N \left( \frac{d}{d\theta} \log q(\mathbf{z}_i|\theta) - \frac{d}{d\theta} \log p(\varphi(\mathbf{z}_i|\theta)) \right)$$

$$\mathbf{z}_i \sim q_{pr}(\cdot|\theta)$$

## Action derivative

$$\frac{dD_{KL}(q||p)}{d\theta} \approx \mathbf{g}_3[\{\phi\}] \equiv \frac{1}{N} \sum_{i=1}^N \left( \frac{d}{d\theta} \log q(\mathbf{z}_i|\theta) - \frac{d}{d\theta} \log p(\varphi(\mathbf{z}_i|\theta)) \right)$$

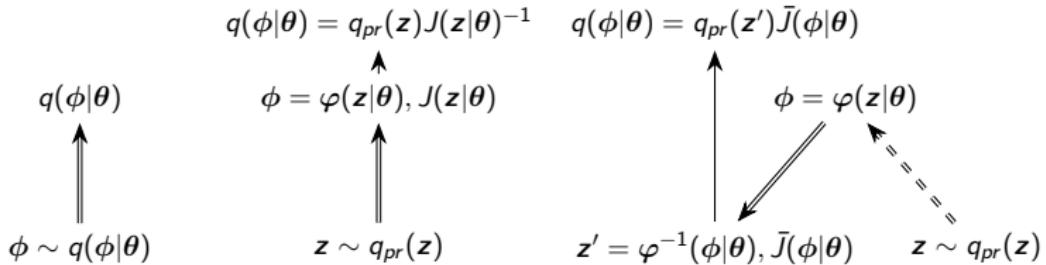
$$\mathbf{z}_i \sim q_{pr}(\cdot|\theta)$$

$$\frac{d}{d\theta} \log p(\varphi(\mathbf{z}_i|\theta)) = \left. \frac{d}{d\phi} \log p(\phi) \right|_{\phi=\varphi(\mathbf{z}_i|\theta)} \frac{d}{d\theta} \varphi(\mathbf{z}_i|\theta)$$

$$\frac{d}{d\phi} \log p(\phi) = -\frac{d}{d\phi} S(\phi)$$

$\mathbf{g}_3$  variance

$$\text{var} [\mathbf{g}_3[\{\phi\}]]_{q(\phi|\theta)=p(\phi)} \neq 0$$



No action derivative

No gradient calculations

$$z_i \sim q_{pr}$$

No action derivative

No gradient calculations

$$z_i \sim q_{pr}$$

$$\phi = q(z|\theta)$$

No action derivative

No gradient calculations

$$z_i \sim q_{pr}$$

$$\phi = q(z|\theta)$$

Gradient calculations

$$z'_i = \varphi^{-1}(\phi_i|\theta)$$

No action derivative

No gradient calculations

$$z_i \sim q_{pr}$$

$$\phi = q(z|\theta)$$

Gradient calculations

$$z'_i = \varphi^{-1}(\phi_i|\theta)$$

$$q(\phi|\theta) = q_{pr}(z')\bar{J}(\phi|\theta)$$

No action derivative

No gradient calculations

$$z_i \sim q_{pr}$$

$$\phi = q(z|\theta)$$

Gradient calculations

$$z'_i = \varphi^{-1}(\phi_i|\theta)$$

$$q(\phi|\theta) = q_{pr}(z') \bar{J}(\phi|\theta)$$

$$\bar{J}(\phi|\theta) \equiv \det \left( \frac{\partial \varphi^{-1}(\phi|\theta)}{\partial \phi} \right)$$

## Toy model

$$\phi(z|\theta) = -\frac{1}{\theta} \log(1-z), \quad z \in [0, 1)$$

## Toy model

$$\phi(z|\theta) = -\frac{1}{\theta} \log(1-z), \quad z \in [0, 1)$$

$$q(\phi|\theta) = \theta e^{-\theta\phi}, \quad z \sim [0, 1)$$

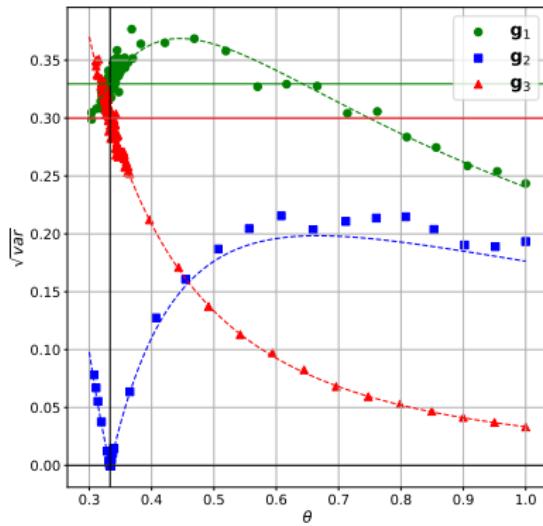
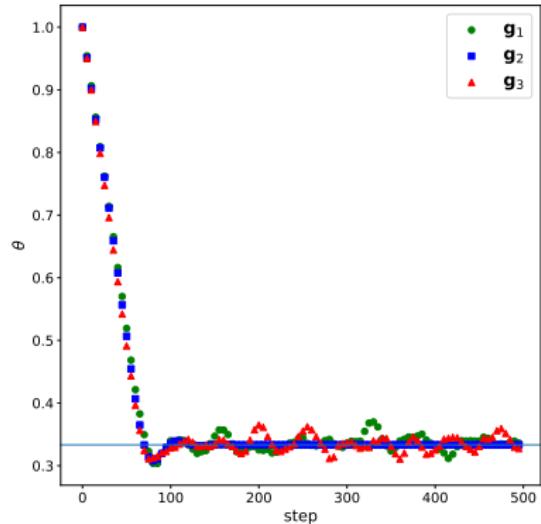
## Toy model

$$\phi(z|\theta) = -\frac{1}{\theta} \log(1-z), \quad z \in [0, 1)$$

$$q(\phi|\theta) = \theta e^{-\theta\phi}, \quad z \sim [0, 1)$$

$$P(\phi) = e^{-\lambda\phi} \quad \lambda = \frac{1}{3}$$

# Toy model



# $\phi^4$ theory ( $16 \times 16$ )

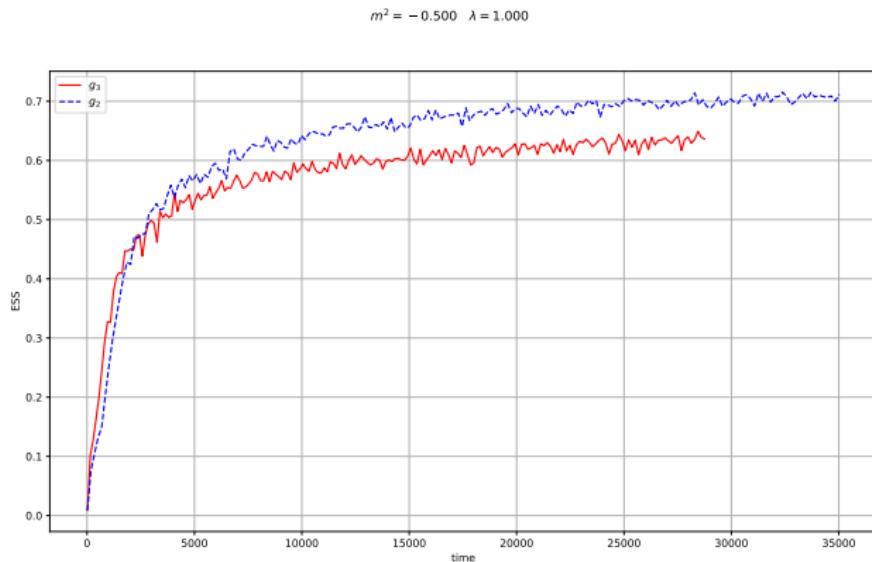
$$\begin{aligned} S(\phi|m^2, \lambda) = & \\ & \sum_{i,j=0}^{L-1} \phi_{i,j} (2\phi_{i,j} - \phi_{i-1,j} - \phi_{i+1,j} + 2\phi_{i,j} - \phi_{i,j-1} - \phi_{i,j+1}) \\ & + \sum_{i,j=0}^{L-1} (m^2 \phi_{i,j} + \lambda \phi_{i,j}^4), \end{aligned}$$

## Results – Effective Sample Size

$$w(\phi) = \frac{p(\phi)}{q(\phi|\theta)}$$

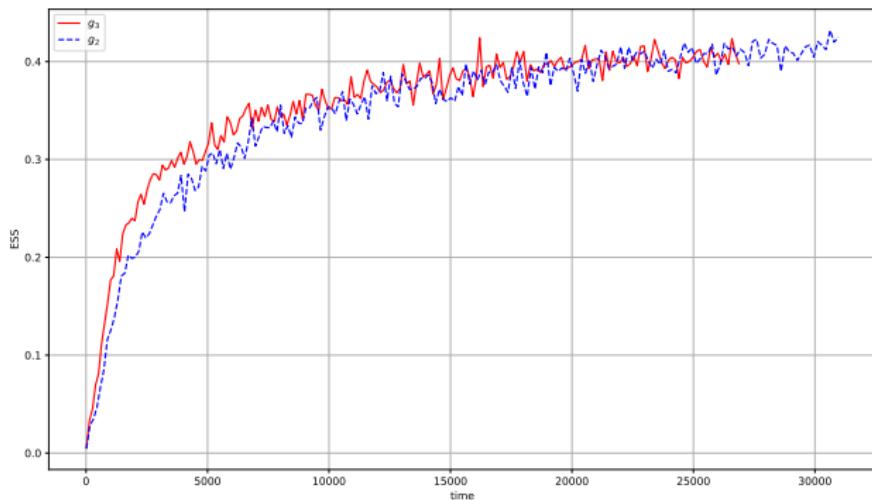
$$ESS = \frac{\langle w \rangle_q^2}{\langle w^2 \rangle_q}$$

# Results



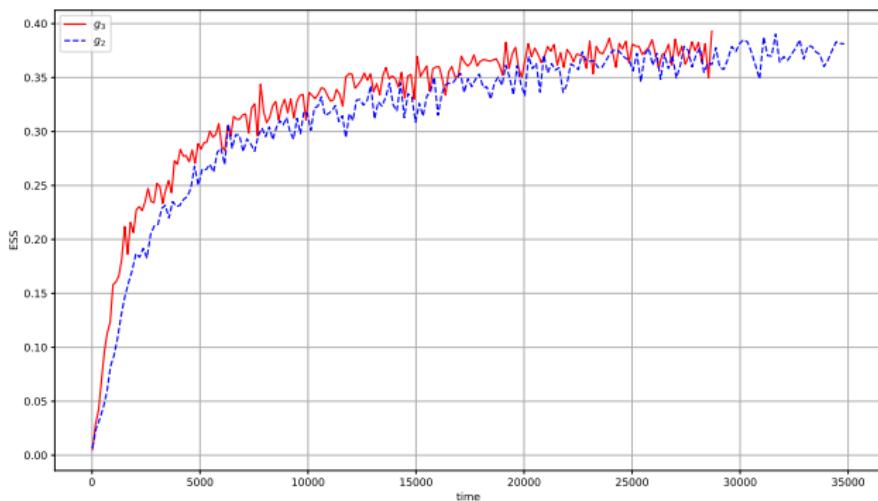
# Results

$$m^2 = -0.950 \quad \lambda = 1.000$$

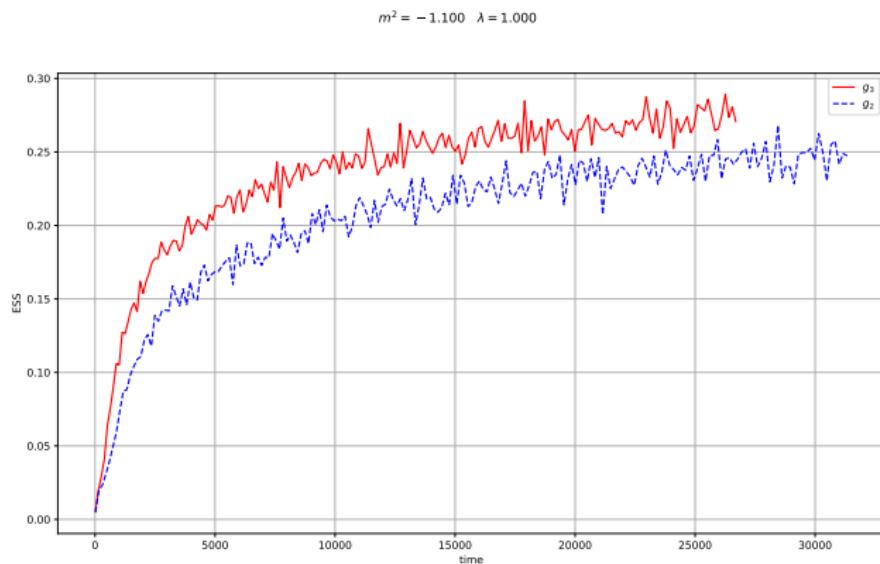


# Results

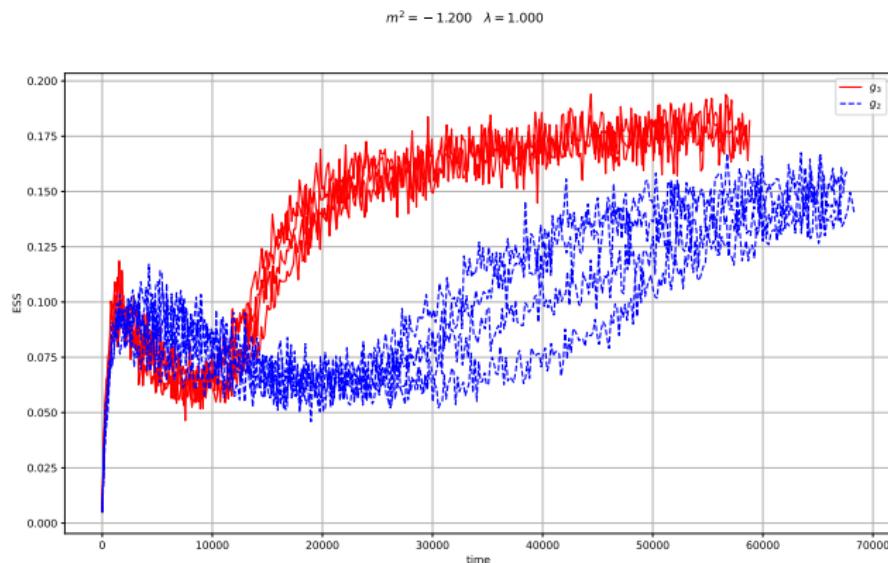
$$m^2 = -1.000 \quad \lambda = 1.000$$



# Results

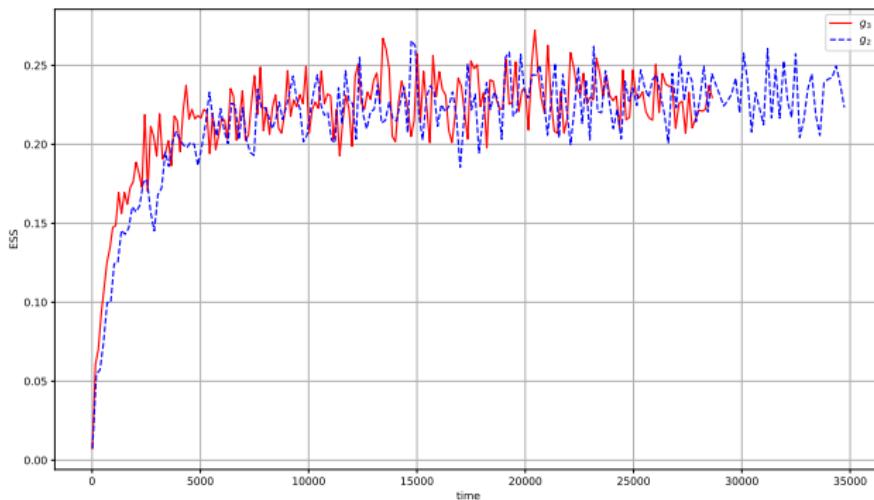


# Results



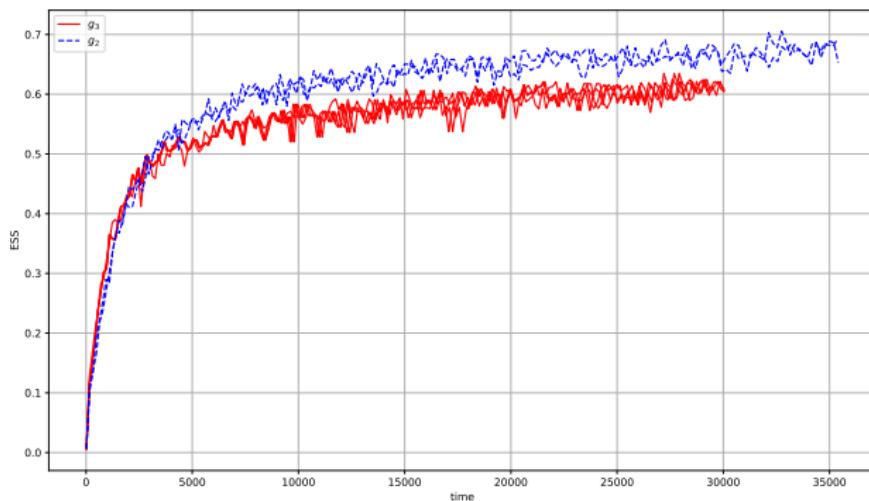
# Results

$$m^2 = -1.250 \quad \lambda = 1.000$$



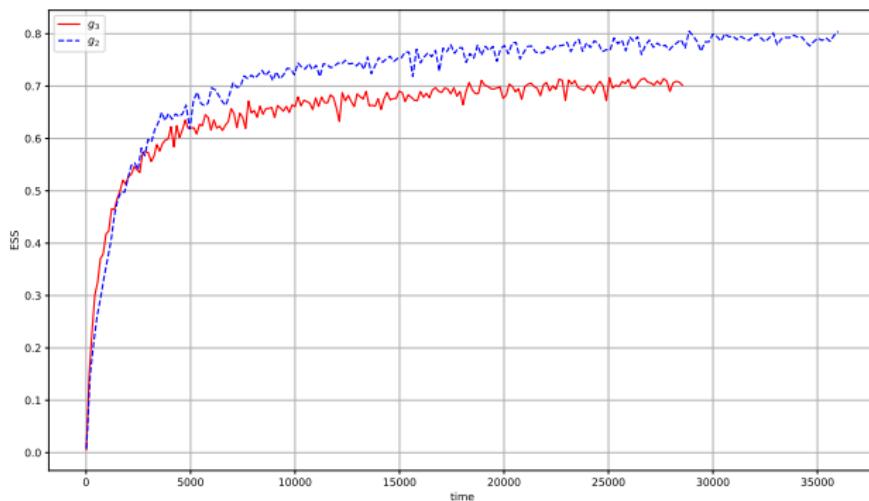
# Results

$$m^2 = -1.400 \quad \lambda = 1.000$$



# Results

$$m^2 = -1.500 \quad \lambda = 1.000$$



# Conclusions

- There is more than one way of estimating gradients for normalizing flows.

# Conclusions

- There is more than one way of estimating gradients for normalizing flows.
- Normalising flows can be trained without taking the action derivative.

# Conclusions

- There is more than one way of estimating gradients for normalizing flows.
- Normalising flows can be trained without taking the action derivative.
- At slight cost in performance ( $\phi^4$ ).

# Conclusions

- There is more than one way of estimating gradients for normalizing flows.
- Normalising flows can be trained without taking the action derivative.
- At slight cost in performance ( $\phi^4$ ).
- Different convergence (training) properties.