UNIVERSITÄT BONN
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN

TRA MATTER
TRANSDISCIPLINARY RESEARCH AREA

Generative AI as the new frontier of sampling algorithms in Lattice QFT

Kim A. Nicoli

University of Bonn, TRA Matter, HISKP (Helmholtz-Instituts für Strahlen- und Kernphysik)

Talk based on PRE 101, 023304 (2020), PRL 126, 032001 (2021), PRD 108, 114501 (2023)

In collaboration with: P. Kessel, S. Nakajima, C. Anders, P. Stornati, L. Funcke, T. Hartung, K. Jansen

Friday, 1st March 2024 - AQTIVATE Workshop (Berlin)

KI write a poem which explains what is the standard model of particle physics

KI write a poem which explains what is the standard model of particle physics

The universe is filled with particles, so small and so precise,
Each with a role to play, each with a unique size.
To understand them all, we need a model that can describe,
The way they interact, the way they survive.

The Standard Model is that model, a framework to behold,
It tells us what's inside, what's hot, what's cold.
It's built upon particles, with quarks and leptons too,
And bosons that carry force, like glue.

Six types of quarks, and six types of leptons too,
All with different masses, all with different hues.
The bosons that they use, are force carriers all,
The photon, the W and Z, the gluon, and the Higgs, standing tall.

The photon is the light, that illuminates our world,
The W and Z, the weak force, that makes some particles swirled.
The gluon, the strong force, that binds quarks together tight,
And the Higgs, that gives mass, without which particles take flight.

KI write a poem which explains what is the standard model of particle physics

The universe is filled with particles, so small and so precise,
Each with a role to play, each with a unique size.
To understand them all, we need a model that can describe,
The way they interact, the way they survive.

The Standard Model is that model, a framework to behold,
It tells us what's inside, what's hot, what's cold.
It's built upon particles, with quarks and leptons too,
And bosons that carry force, like glue.

Six types of quarks, and six types of leptons too,
All with different masses, all with different hues.
The bosons that they use, are force carriers all,
The photon, the W and Z, the gluon, and the Higgs, standing tall.

The photon is the light, that illuminates our world,
The W and Z, the weak force, that makes some particles swirled.
The gluon, the strong force, that binds quarks together tight,
And the Higgs, that gives mass, without which particles take flight.

The Standard Model explains, how particles combine,
To form all we see, to make the universe shine.
It's not complete, with mysteries to unfold,
But it's a mighty tool, in the hands of the bold.

So when you think of particles, and all they can achieve,
Remember the Standard Model, and what it can conceive.
For in its tiny world, lies the secrets of the universe,
And with it, we can unlock, a future that's diverse.

KI write a poem which explains what is the standard model of particle physics

The universe is filled with particles, so small and so precise,
Each with a role to play, each with a unique size.
To understand them all, we need a model that can describe,
The way they interact, the way they survive.

The Standard Model is that model, a framework to behold,
It tells us what's inside, what's hot, what's cold.
It's built upon particles, with quarks and leptons too,
And bosons that carry force, like glue.

Six types of quarks, and six types of leptons too,
All with different masses, all with different hues.
The bosons that they use, are force carriers all,
The photon, the W and Z, the gluon, and the Higgs, standing tall.

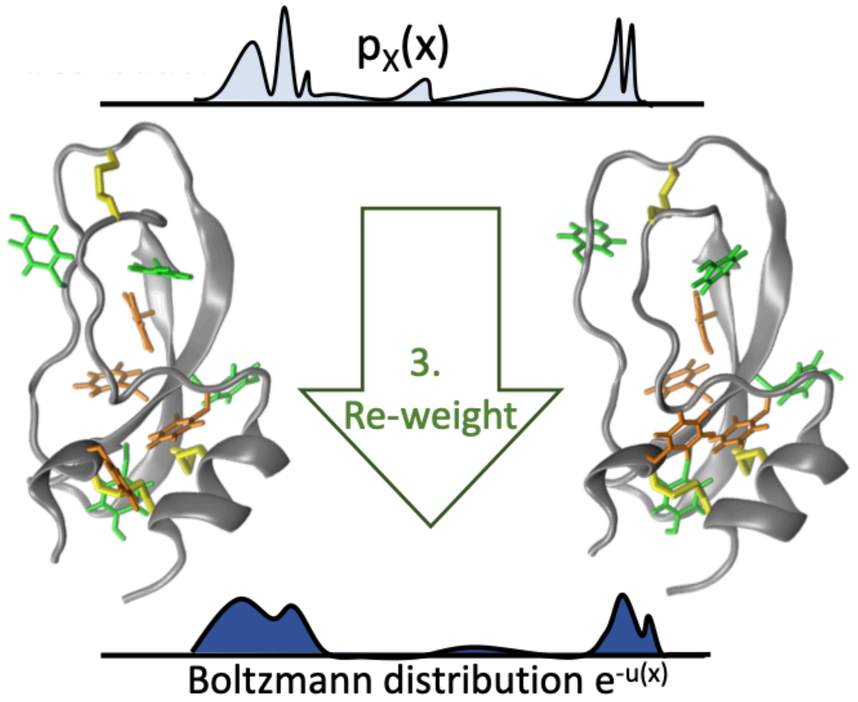
The photon is the light, that illuminates our world,
The W and Z, the weak force, that makes some particles swirled.
The gluon, the strong force, that binds quarks together tight,
And the Higgs, that gives mass, without which particles take flight.

The Standard Model explains, how particles combine,
To form all we see, to make the universe shine.
It's not complete, with mysteries to unfold,
But it's a mighty tool, in the hands of the bold.

So when you think of particles, and all they can achieve,
Remember the Standard Model, and what it can conceive.
For in its tiny world, lies the secrets of the universe,
And with it, we can unlock, a future that's diverse.

Generative AI for Physics : Past, Present and Future

F. Noé, et al., Science, eaaw1147 (2019)



(Image credits: F. Noé, et al., Science, eaaw1147 (2019))

Generative AI for Physics : Past, Present and Future

Scalar Field Theories (ϕ^4)

M. S. Albergo, et al., Phys. Rev. D 100, 034515 (2019)

K. A. Nicoli, et al., Phys. Rev. Lett. 126, 032001 (2021)

P. deHaan, et al., arXiv: 2110.02673 @ ML4Pys workshop (2021)

L. Vaitl et al., arXiv: 2206.09016 @ ICML (2022)

A. Matthews et al., arXiv:2201.13117 @ ICML (2022)

M. Caselle, et al., J. High Energ. Phys. 2022, 15 (2022)

M. Gerdes, et al., SciPost Phys. 15, 238 (2023)

A. Singha, et al., Phys. Rev. D 107, 014512 (2023)

Lattice Gauge Theories ($\mathrm{U}(1), \mathrm{SU}(N)$)

G. Kanwar, et al., Phys. Rev. Lett. 125, 121601 (2020)

S. Bacchio, et al., Phys. Rev. D 107, L051504 (2023)

R. Abbott et al., Phys. Rev. D 106, 074506 (2022)

M. S. Albergo, et al., Phys. Rev. D 106, 014514 (2022)

R. Abbott, et al., arXiv:2305.02402 (2023)

J. Finkenrath, arXiv: 2201.02216 (2022)

Sampling Multimodal Densities in QFT

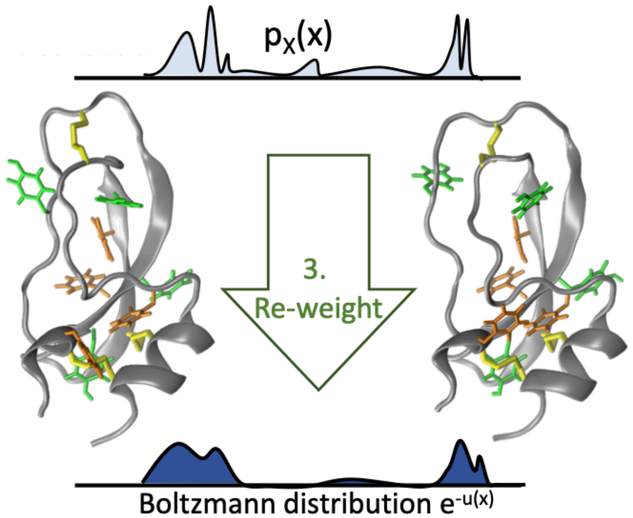
D.C. Hackett et al., arXiv:2107.00734 (2021)

K. A. Nicoli, et al., arXiv: 2111.11303 @ LATTICE21 (2021)

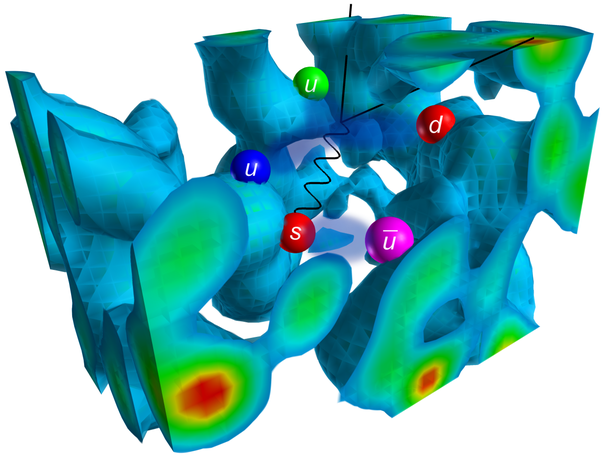
K. A. Nicoli, et al., Phys. Rev. D 108, 114501 (2023)

B. Maté et al., TMLR 2835-8856 (2023)

V. Kanaujia et al., arXiv:2401.15948 (2024)



(Image credits: F. Noé, et al., Science, eaaw1147 (2019))



(Image credits: Lattice QCD © Derek Leinweber/CSSM/University of Adelaide)

Scaling to Larger Lattices

L. Del Debbio, et al., Phys. Rev. D 104, 094507 (2021)

R. Abbott et al., Eur. Phys. J. A 59, 257 (2023)

A. Faraz et al., arXiv:2308.08615 (2022)

B. Maté et al., arXiv: 2401.00828 (2024)

R. Abbott et al., arXiv: 2401.10874v1 (2024)

J. Finkenrath, arXiv: 2402.12176 (2024)

Autoregressive Models in Stat. Mech.

D. Wu et al., Phys. Rev. Lett. 122 (8), 080602 (2019)

K. A. Nicoli, et al., Phys. Rev. E 101 (2), 023304 (2019)

P. Bialas et al., Computer Physics Communications 281 (2022)

P. Bialas et al., Phys. Rev. E 107 (1), 015303 (2023)

Interesting Readings

F. Noé, et al., Science, eaaw1147 (2019)

S. Chen, et al., Phys. Rev. D 107, 056001 (2022)

B. Maté et al., arXiv: 2210.13772 (2022)

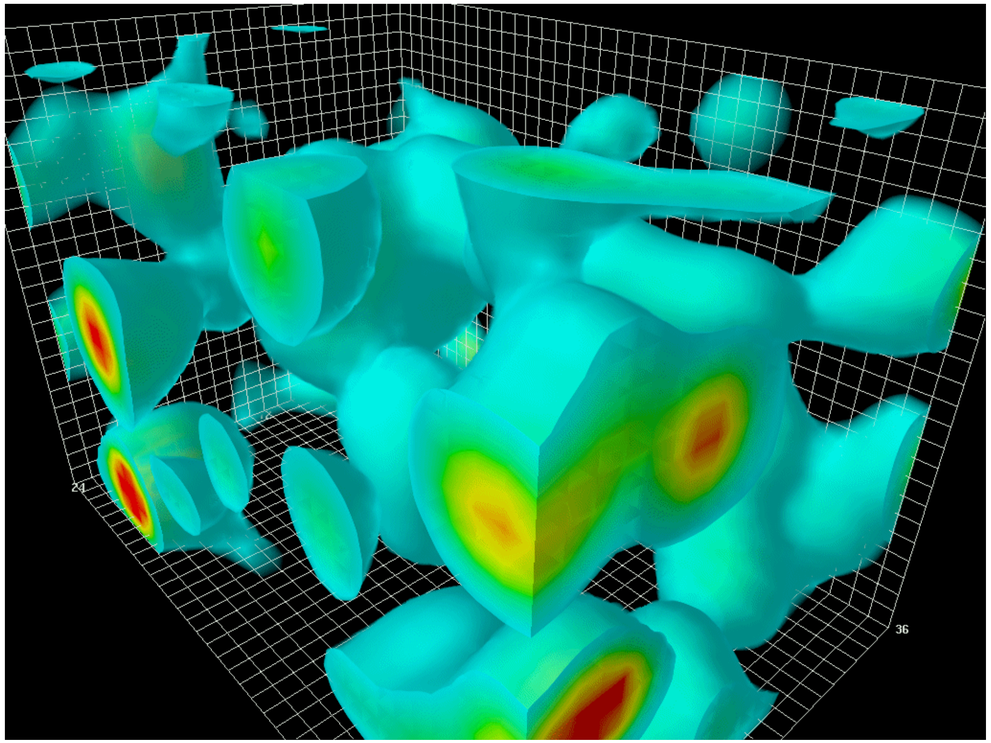
L. Vaitl et al., MLST 3 (4), 045006 (2022)

Caselle, M., et al., J. High Energ. Phys. 2024, 48 (2024)

Cranmer K. et al., Nature Reviews Physics 5 (9), 526-535

And much more...

Quantum Field Theory

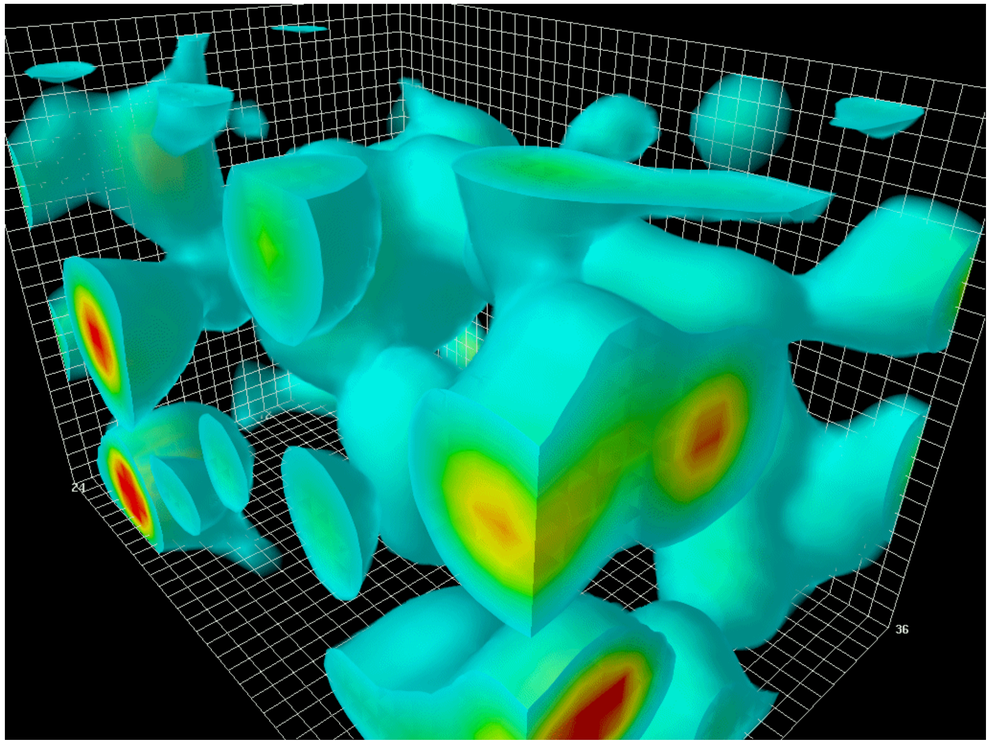


(credits: Derek Leinweber)

" […] [A quantum field] […] a ***bubbling soup*** *of* ***particles*** *and anti-particles, constantly created and destroyed "*

D. Tong

Quantum Field Theory

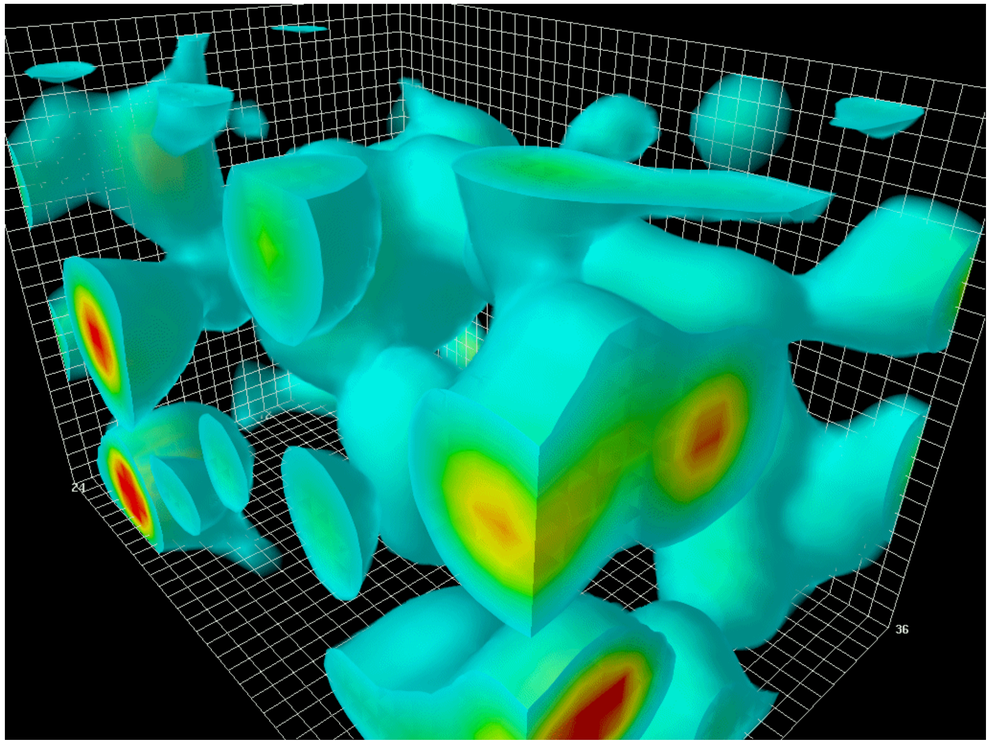


(credits: Derek Leinweber)

" […] [A quantum field] […] a ***bubbling soup*** *of* ***particles*** *and anti-particles, constantly created and destroyed "*

D. Tong

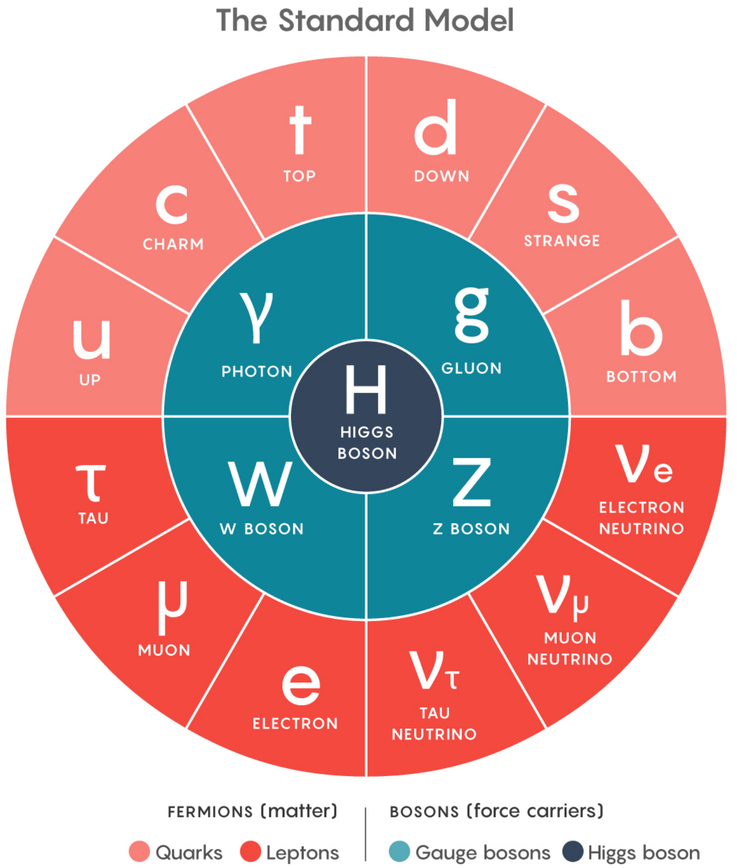
Quantum Field Theory



(credits: Derek Leinweber)

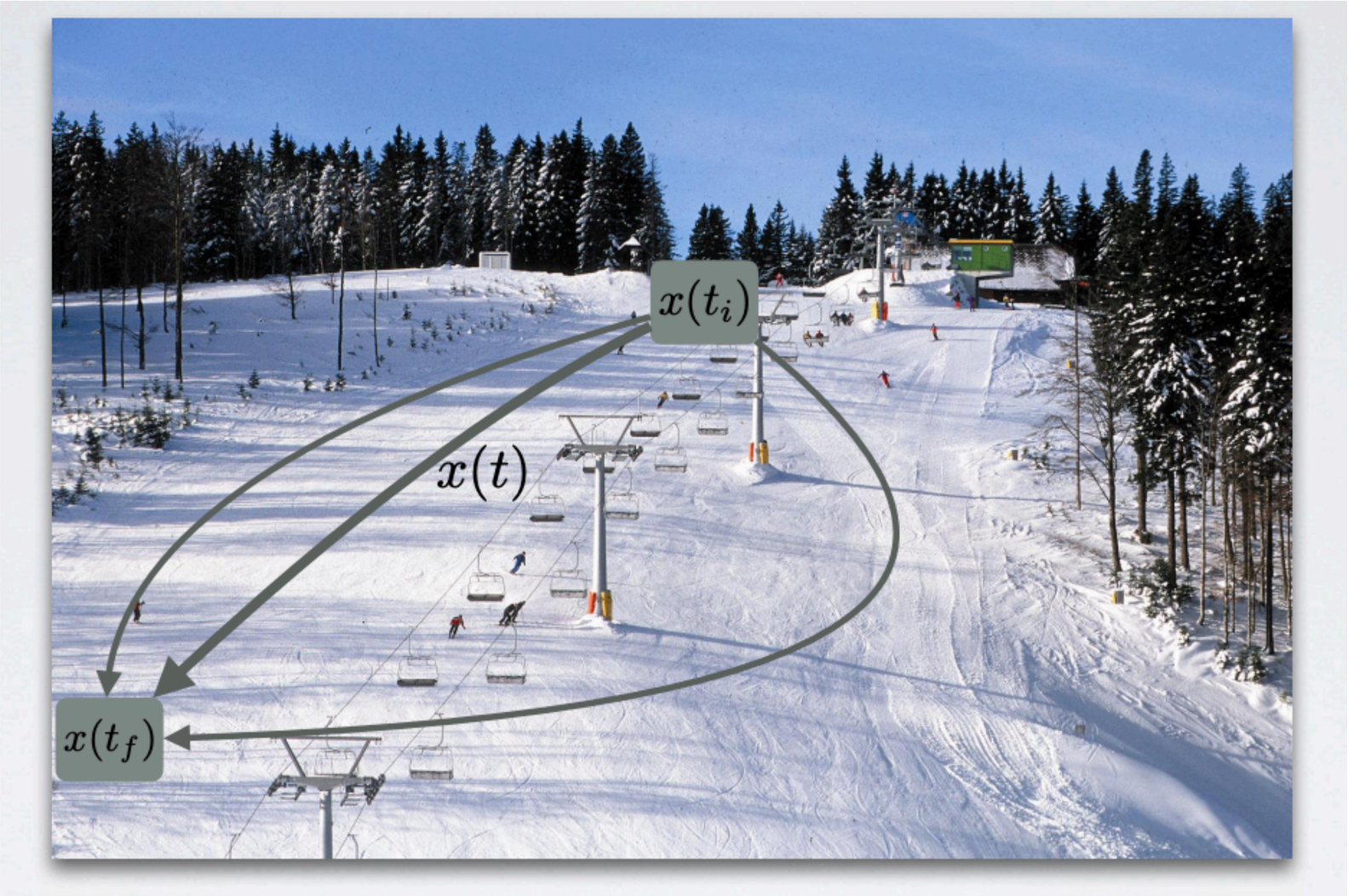
*" […] [A quantum field] […] a **bubbling soup** of **particles** and anti-particles, constantly created and destroyed "*

D. Tong



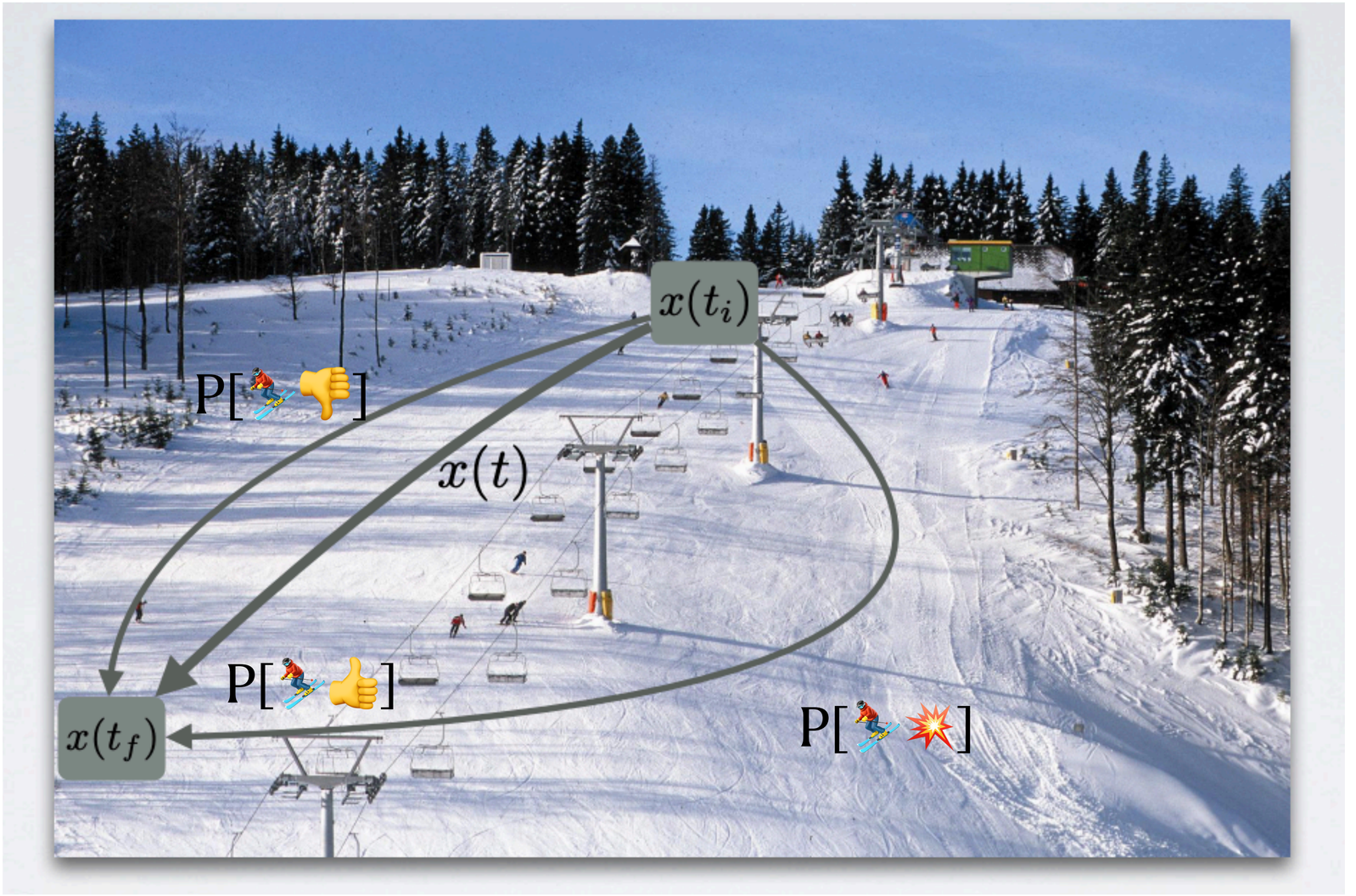
(Image credits: Quanta Magazine)

Path Integral Formulation



(Image credits: adapted from Tom Luu)

Path Integral Formulation

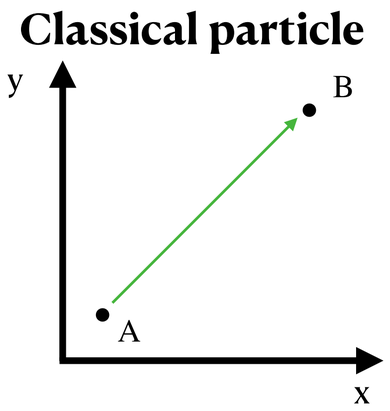


(Image credits: adapted from Tom Luu)

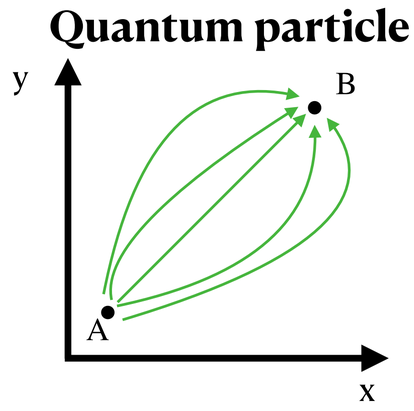
Path Integral Formulation

Path-integral is the basic tool for quantising fields and compute expectation values of physical observables $\mathcal{O}$:

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \int D[\phi]\, \mathcal{O}(\phi)\, \exp\{-S(\phi)\}$$



Least action ≡ Most probable



Path Integral Formulation

Path-integral is the basic tool for quantising fields and compute expectation values of physical observables $\mathcal{O}$:

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \int D[\phi]\, \mathcal{O}(\phi)\, \exp\{-S(\phi)\}$$

Path Integral Formulation

Path-integral is the basic tool for quantising fields and compute expectation values of physical observables $\mathcal{O}$:

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \int D[\phi]\, \mathcal{O}(\phi)\, \exp\{-S(\phi)\}$$

Path Integral Formulation

Path-integral is the basic tool for quantising fields and compute expectation values of physical observables $\mathcal{O}$:

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \int D[\phi] \, \mathcal{O}(\phi) \exp\{-S(\phi)\}$$

Path Integral Formulation

Path-integral is the basic tool for quantising fields and compute expectation values of physical observables $\mathcal{O}$:

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \int D[\phi]\, \mathcal{O}(\phi)\, \exp\{-S(\phi)\}$$

Path Integral Formulation

Path-integral is the basic tool for quantising fields and compute expectation values of physical observables $\mathcal{O}$:

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \int D[\phi]\, \mathcal{O}(\phi)\, \exp\{-S(\phi)\}$$

Path Integral Formulation

Path-integral is the basic tool for quantising fields and compute expectation values of physical observables $\mathcal{O}$:

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \int D[\phi]\, \mathcal{O}(\phi)\, \exp\{-S(\phi)\}$$

computed over a Boltzmann-like density:

$$p(\phi) = \frac{e^{-S(\phi)}}{Z} \longrightarrow$$

known in **closed form** up to a **numerically intractable** normalisation

$$Z = \int D[\phi] e^{-S(\phi)}$$

Lattice Quantum Field Theory

- The **path-integral** reduces from a **functional integral** to a **high-dimensional** ordinary integral

$$\langle \mathcal{O} \rangle = \int \prod_{x \in \Lambda} \mathrm{d}[\phi(x)] \, \mathcal{O}(\phi) \, p(\phi)$$

- The field configuration $\phi(x)$ is now a **random variable** of size Λ (lattice volume).
- With Markov-Chain Monte-Carlo algorithm we can **sample** and estimate observables.

$$\langle \mathcal{O} \rangle_p = \int D[\phi] \, \mathcal{O}(\phi) \, p(\phi) \approx \frac{1}{N} \sum_{i=1}^{N} \mathcal{O}(\phi_i)$$

where $\phi_i \sim p$ (Boltzmann-like density)

Markov Chain Monte-Carlo

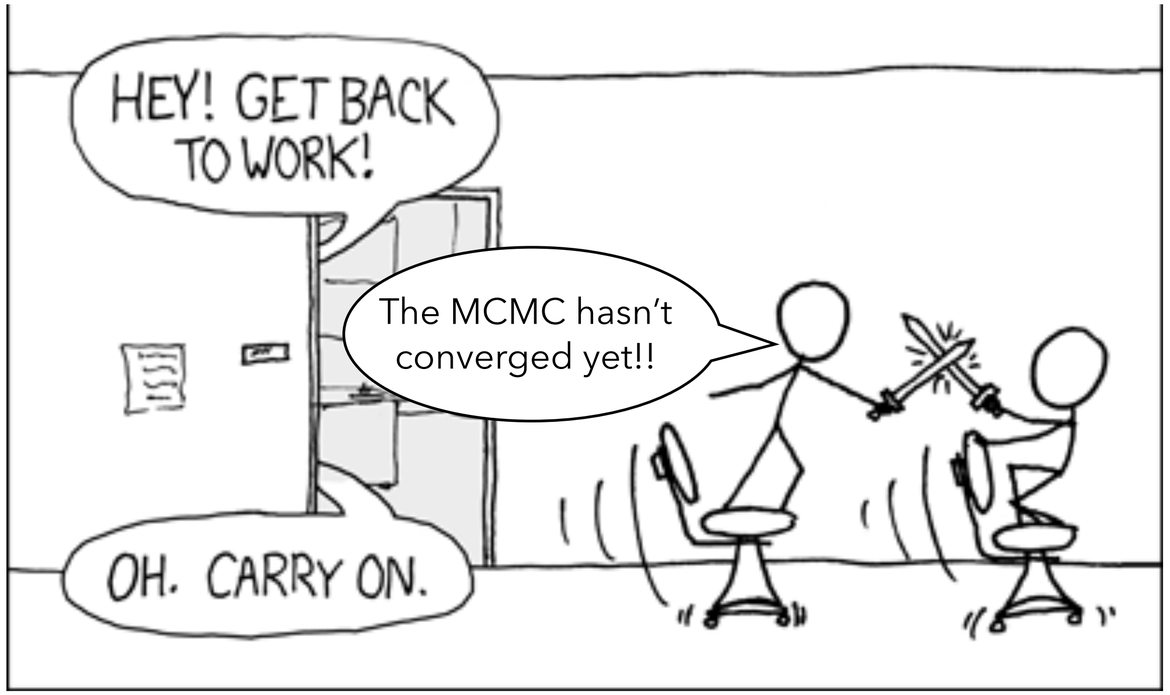
MCMC: sequentially proposeS the next sample and guarantees to eventually converge to a target density.

Markov Chain Monte-Carlo

MCMC: sequentially proposeS the next sample and guarantees to eventually converge to a target density.

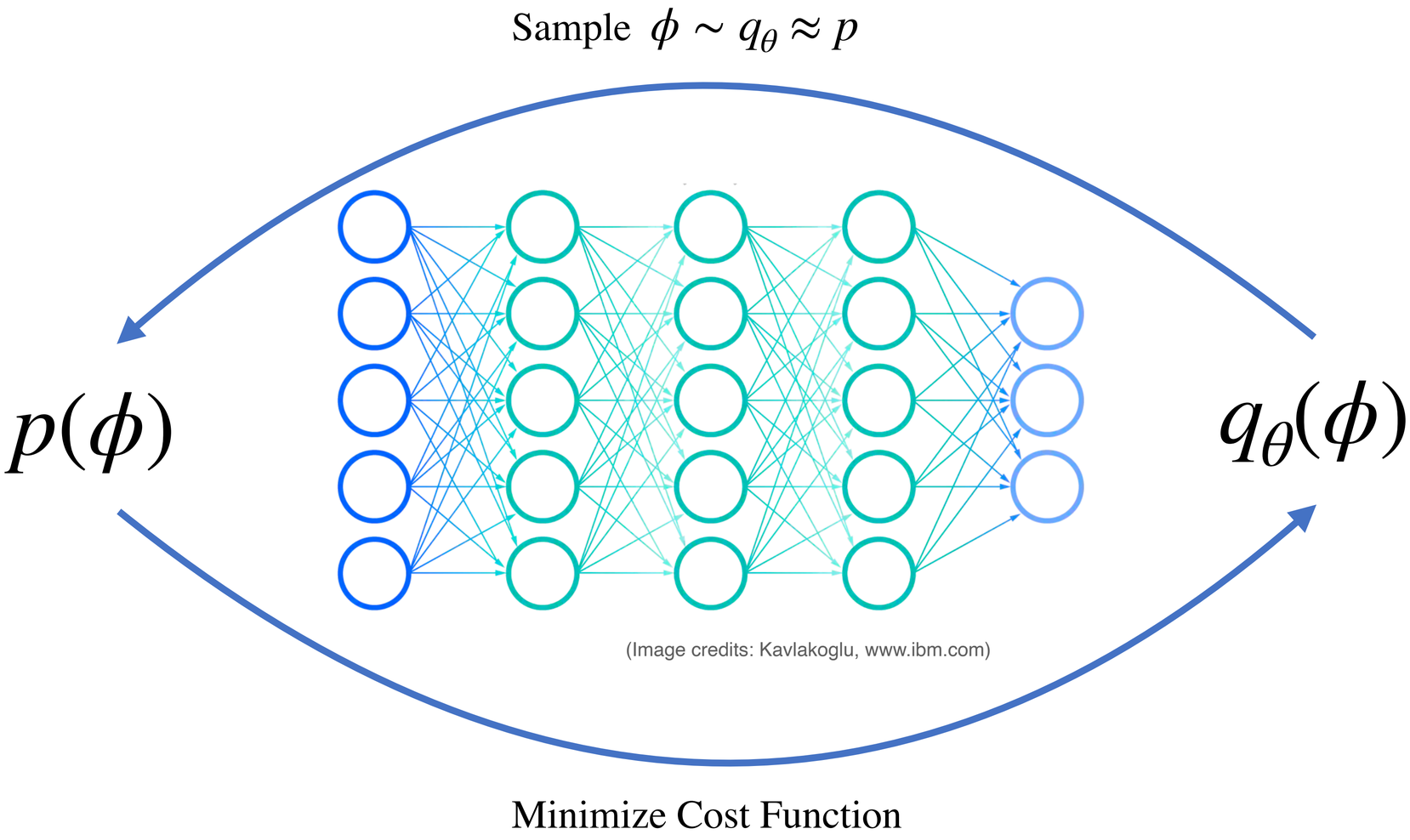
However MCMC algorithms come at a **cost**:

- 👎 **Sequential** $\Longrightarrow$ MCMC chains can't be parallelized.
- 👎 **Critical slowing down** $\Longrightarrow$ Struggles around phase transitions.
- 👎 Long range autocorrelations $\Longrightarrow$ **large statistical errors.**
- 👎 The partition function Z **is unknown**.
- 👎 No direct estimation of **thermodynamic observables.**



(Adapted from: xkcd/303)

Density Estimation with Deep Generative Models



(Image credits: Kavlakoglu, www.ibm.com)

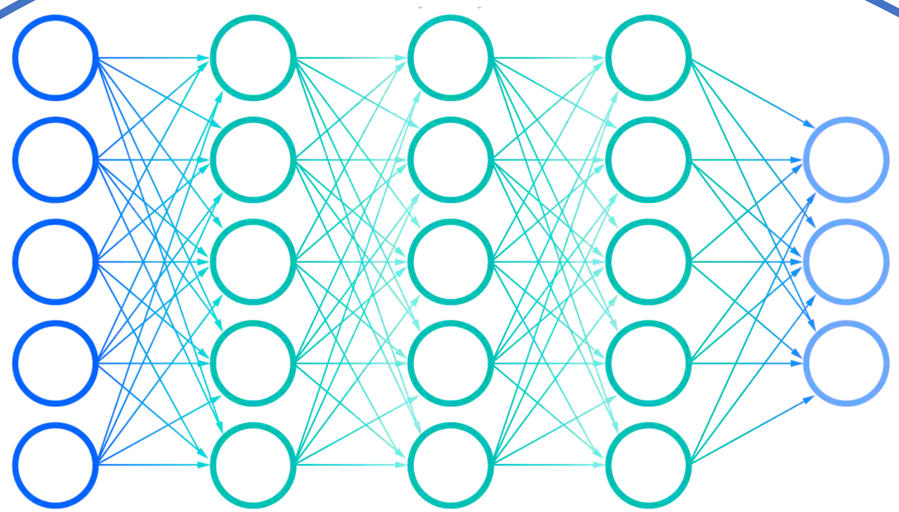
Density Estimation with Deep Generative Models

Sample $\phi \sim q_\theta \approx p$

Autoregressive Models
(Pixel CNN)
Van der Oord et al. (2016)

Discrete RV
e.g., Ising

$p(\phi)$



$q_\theta(\phi)$

(Image credits: Kavlakoglu, www.ibm.com)

Minimize Cost Function

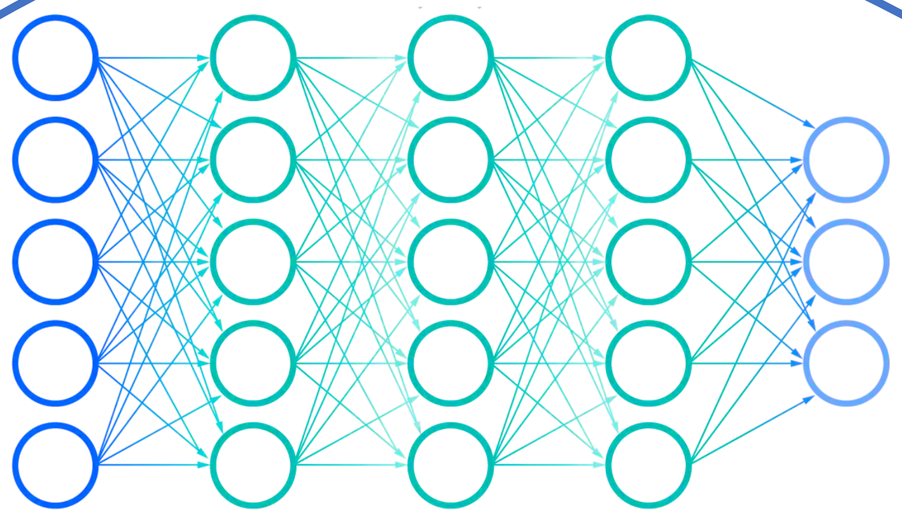
Density Estimation with Deep Generative Models

Sample $\phi \sim q_\theta \approx p$

Autoregressive Models
(Pixel CNN)
Van der Oord et al. (2016)

Discrete RV
e.g., Ising

$p(\phi)$



(Image credits: Kavlakoglu, www.ibm.com)

$q_\theta(\phi)$

Continuous RV
e.g., QFT

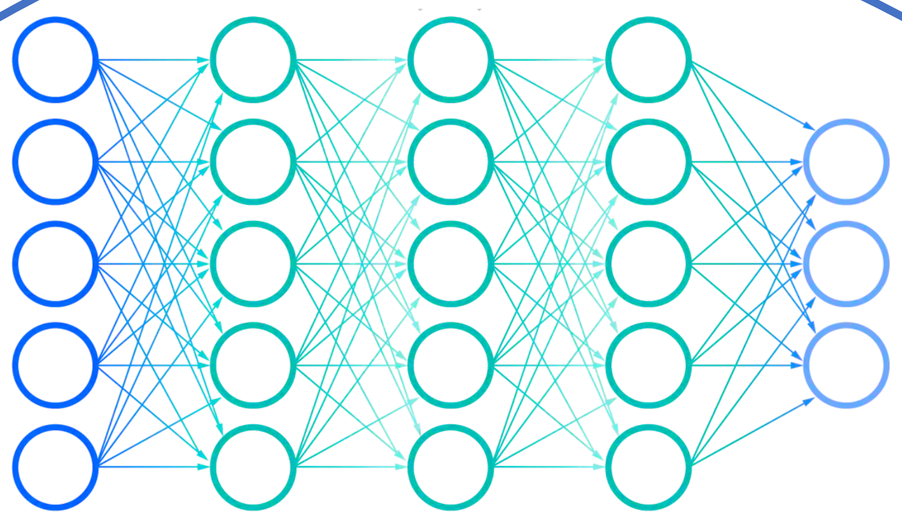
Normalizing Flows
Rezende and Mohamed (2015)

Minimize Cost Function

Density Estimation with Deep Generative Models

Sample $\phi \sim q_\theta \approx p$

$p(\phi)$



(Image credits: Kavlakoglu, www.ibm.com)

Minimize Cost Function

Autoregressive Models
(Pixel CNN)
Van der Oord et al. (2016)

Discrete RV
e.g., Ising

$q_\theta(\phi)$

Exact Likelihood
Density Estimation

Continuous RV
e.g., QFT

Normalizing Flows
Rezende and Mohamed (2015)

Taxonomy of Generative Models

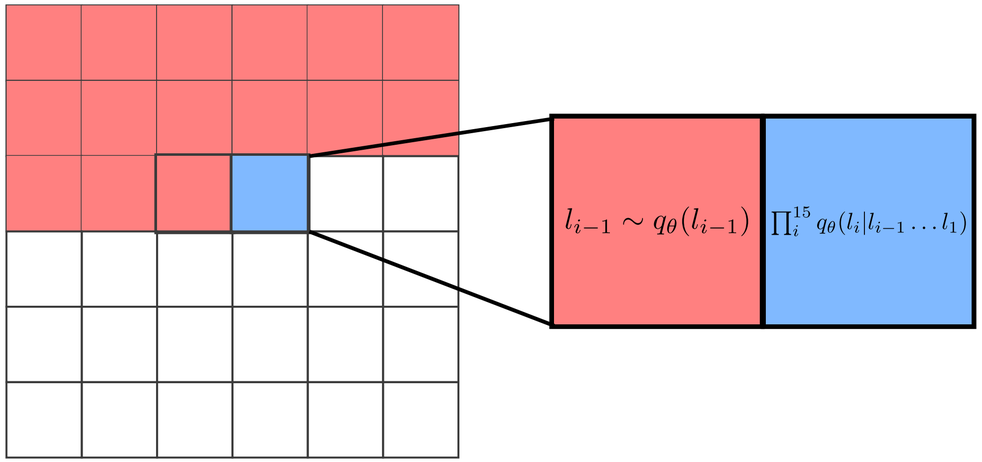
Table 2.1. **Exact likelihood estimation for different types of generative models**: overview on different types of deep generative models (DGMs) with few examples of their application in the physics domain. Generative adversarial networks (generative adversarial network (GAN)) usually do not have access to sampling probabilities while variational auto-encoders (VAEs) only allows for approximate ones. The cases of normalizing flows (NFs) and autoregressive neural networks (ARNNs) are instead substantially different as they allow access to normalized tractable densities.

DGM	sampling probability	Normalization
GAN	none	X
VAE	approximate	✓
ARNN	exact	✓
NF	exact	✓
DDPM	approximate*	✓

* can be exact under some circumstances, e.g., Sec 4.3 from arXiv:2209.00796.

Autoregressive Models

- Probability density of sampling a configuration L **(ancestral sampling)**



Nicoli, PhD diss., Technische Universität Berlin, 2023.

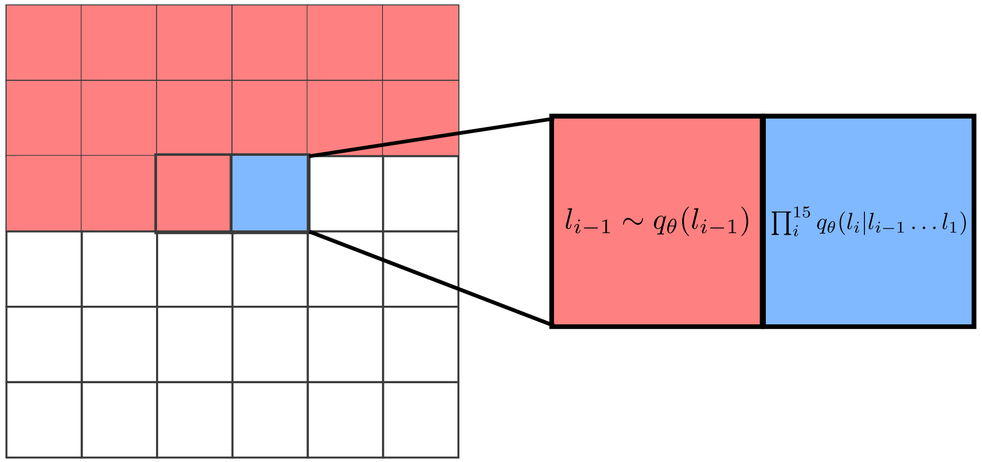
$$q_\theta(L) = \prod_{i=1}^{N^2} q_\theta(l_i \,|\, l_{i-1}\dots l_1)$$

Images (PixelCNN)

A. Van der Oord et al., arXiv:1601.06759 (2016)

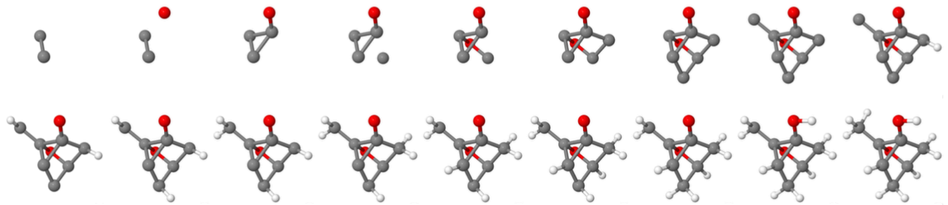
Autoregressive Models

- Probability density of sampling a configuration L **(ancestral sampling)**



Nicoli, PhD diss., Technische Universität Berlin, 2023.

Molecular Graphs



N. Gebauer et al., arXiv:1810.11347 (2018)

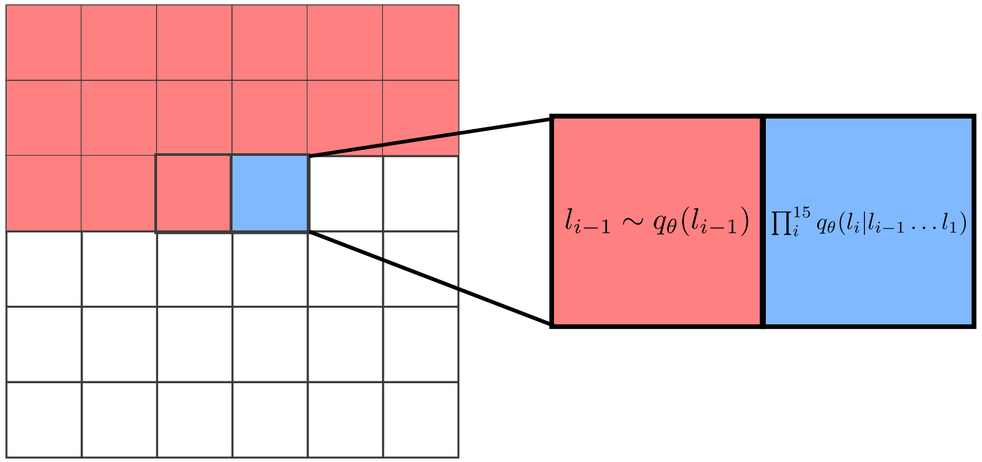
$$q_\theta(L) = \prod_{i=1}^{N^2} q_\theta(l_i \,|\, l_{i-1}\dots l_1)$$

Images (PixelCNN)

A. Van der Oord et al., arXiv:1601.06759 (2016)

Autoregressive Models

- Probability density of sampling a configuration L **(ancestral sampling)**



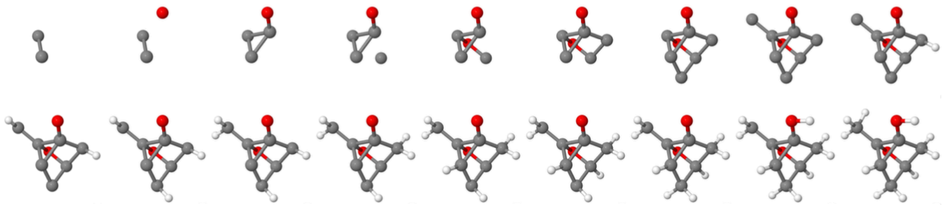
Nicoli, PhD diss., Technische Universität Berlin, 2023.

$$q_\theta(L) = \prod_{i=1}^{N^2} q_\theta(l_i \,|\, l_{i-1} \dots l_1)$$

Images (PixelCNN)

A. Van der Oord et al., arXiv:1601.06759 (2016)

Molecular Graphs



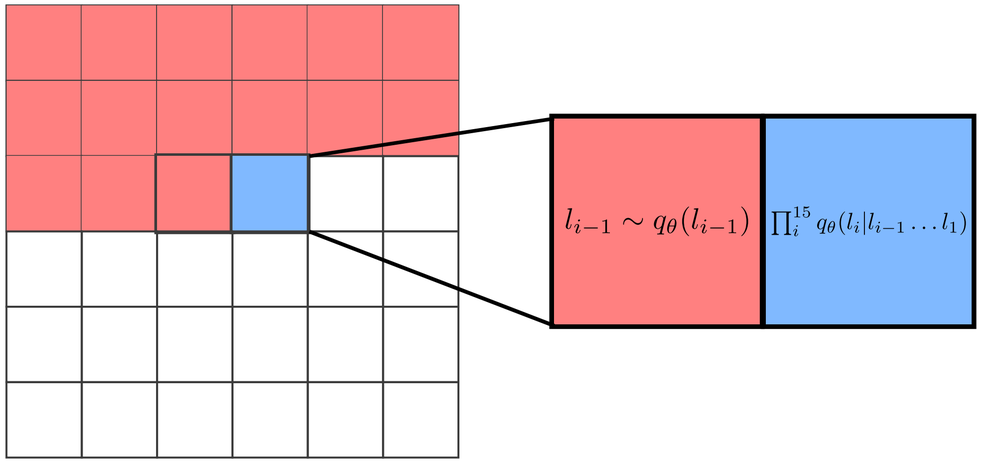
N. Gebauer et al., arXiv:1810.11347 (2018)

Language Models (GPT)

T. Brown et al., arXiv: 2005.14165 (2020)

Autoregressive Models

- Probability density of sampling a configuration L **(ancestral sampling)**



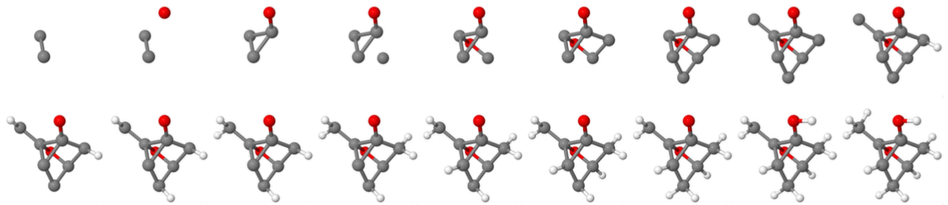
Nicoli, PhD diss., Technische Universität Berlin, 2023.

$$q_\theta(L) = \prod_{i=1}^{N^2} q_\theta(l_i \,|\, l_{i-1}\dots l_1)$$

Images (PixelCNN)

A. Van der Oord et al., arXiv:1601.06759 (2016)

Molecular Graphs

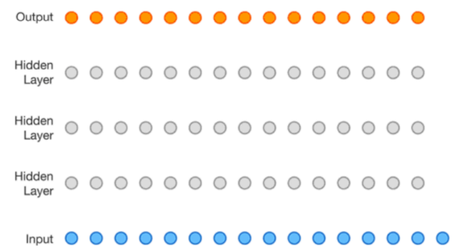


N. Gebauer et al., arXiv:1810.11347 (2018)

Language Models (GPT)

T. Brown et al., arXiv: 2005.14165 (2020)

Speech (WaveNet)



A. Van der Oord et al., arXiv:1609.03499 (2016)

Sampling with Normalizing Flows

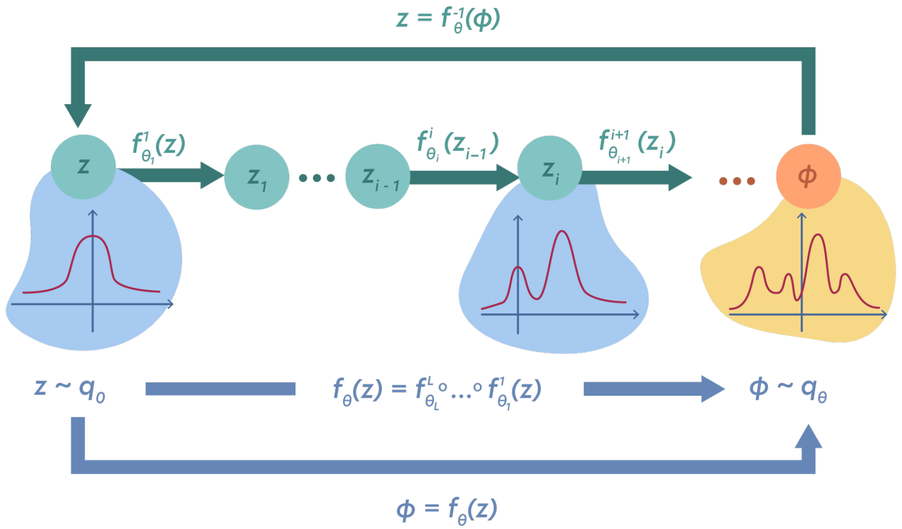
We use a parametric function f_θ (a **diffeomorphism**) to transform Gaussian samples $z \sim q_0$ into physical configurations $\phi \sim q_\theta$

$$f_\theta : z \in \mathcal{Z} \sim q_z \rightarrow x = f_\theta(z) \in \mathcal{X} \sim q_\theta \, .$$

The parametric function needs to fulfill certain criteria:

- Bijective transformation $\phi = f_\theta(z)$
- Invertible and differentiable
- Tractable Jacobian

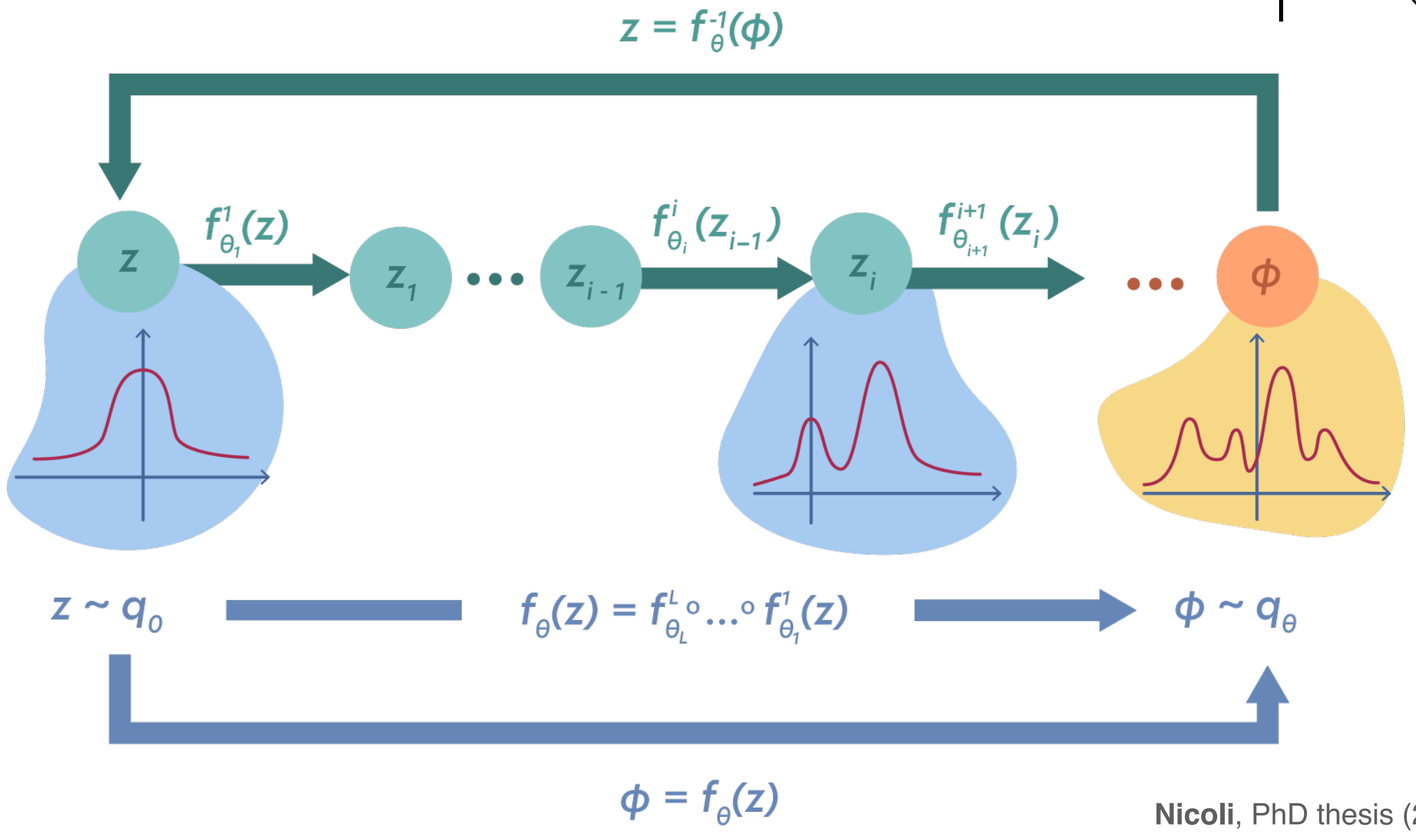
N.B. The function f_θ is parametrized by a Neural Network.



Nicoli, PhD thesis (2023)

Sampling with Normalizing Flows

The likelihood of q_θ can be computed exactly: $$q_\theta(\phi) = q_0(f_\theta^{-1}(\phi)) \left| \det \left(\frac{\partial f_\theta}{\partial z} \right) \right|^{-1}$$



Nicoli, PhD thesis (2023)

How do we train a generative model?

Often the variational density q_θ is trained by minimizing the **Reverse-KL** divergence:

$$KL(q_\theta||p) = \int D[\phi] q_\theta(\phi) \ln \frac{q_\theta(\phi)}{p(\phi)} \equiv \mathbb{E}_{q_\theta}\left[\ln \frac{q_\theta(\phi)}{p(\phi)}\right]$$

How do we train a generative model?

Often the variational density q_θ is trained by minimizing the **Reverse-KL** divergence:

$$KL(q_\theta||p) = \int D[\phi] q_\theta(\phi) \ln \frac{q_\theta(\phi)}{p(\phi)} \equiv \mathbb{E}_{q_\theta}\left[\ln \frac{q_\theta(\phi)}{p(\phi)}\right]$$

since we know the target $p(\phi)$ is a Boltzmann distribution $p(\phi) = Z^{-1}\exp\{-S(\phi)\}$

$$KL(q_\theta||p) = \mathbb{E}_{q_\theta}\left[\ln \frac{q_\theta(\phi)}{p(\phi)}\right] = \mathbb{E}_{q_\theta}\left[\ln q_\theta(\phi) + S(\phi) + \ln Z\right]$$

How do we train a generative model?

Often the variational density q_θ is trained by minimizing the **Reverse-KL** divergence:

$$KL(q_\theta||p) = \int D[\phi] q_\theta(\phi) \ln \frac{q_\theta(\phi)}{p(\phi)} \equiv \mathbb{E}_{q_\theta}\left[\ln \frac{q_\theta(\phi)}{p(\phi)}\right]$$

since we know the target $p(\phi)$ is a Boltzmann distribution $p(\phi) = Z^{-1}\exp\{-S(\phi)\}$

$$\nabla_\theta KL(q_\theta||p) = \mathbb{E}_{q_\theta}\left[\nabla_\theta \ln q_\theta(\phi) + \nabla_\theta S(\phi) + \cancel{\ln Z}\right]$$

How do we train a generative model?

Often the variational density q_θ is trained by minimizing the **Reverse-KL** divergence:

$$KL(q_\theta||p) = \int D[\phi] q_\theta(\phi) \ln \frac{q_\theta(\phi)}{p(\phi)} \equiv \mathbb{E}_{q_\theta}\left[\ln \frac{q_\theta(\phi)}{p(\phi)}\right]$$

since we know the target $p(\phi)$ is a Boltzmann distribution $p(\phi) = Z^{-1}\exp\{-S(\phi)\}$

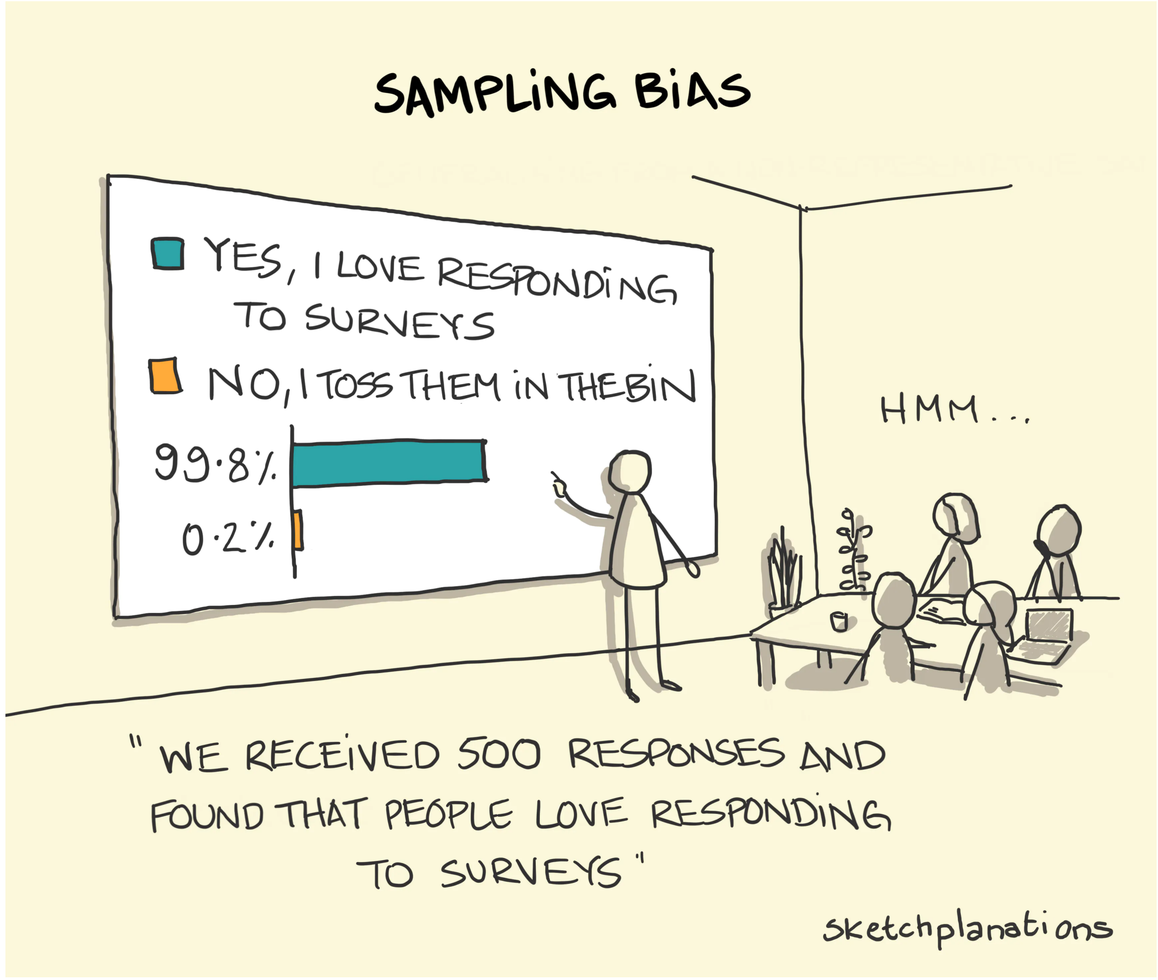
$$\nabla_\theta KL(q_\theta||p) = \mathbb{E}_{q_\theta}\left[\nabla_\theta \ln q_\theta(\phi) + \nabla_\theta S(\phi) + \cancel{\ln Z}\right]$$

Training can be performed **by self-sampling** from the model we are training**!**

Shall we Trust These Samplers?

(Image credits: Gymclimber)

YES! But…



(Image credits: sketchplanations)

Neural Importance Sampling (NIS)

$$p(\phi) \approx q_\theta \sim \phi_i \qquad \text{where} \qquad p(\phi) = \frac{\exp\{-S(\phi)\}}{Z}$$

$$Z = \int D[\phi]\ \exp\{-S(\phi)\} = \int D[\phi]\, q_\theta(\phi)\, \tilde{w}(\phi) \quad \text{where} \quad \tilde{w}(\phi) = \frac{\exp\{-S(\phi)\}}{q_\theta(\phi)}$$

Neural Importance Sampling (NIS)

$$p(\phi) \approx q_\theta \sim \phi_i \qquad \text{where} \qquad p(\phi) = \frac{\exp\{-S(\phi)\}}{Z}$$

$$Z = \int D[\phi]\ \exp\{-S(\phi)\} = \int D[\phi]\, q_\theta(\phi)\, \tilde{w}(\phi) \quad \text{where} \quad \tilde{w}(\phi) = \frac{\exp\{-S(\phi)\}}{q_\theta(\phi)}$$

Neural Importance Sampling (NIS)

$$p(\phi) \approx q_\theta \sim \phi_i \qquad \text{where} \qquad p(\phi) = \frac{\exp\{-S(\phi)\}}{Z}$$

$$Z = \int D[\phi]\ \exp\{-S(\phi)\} = \int D[\phi]\, q_\theta(\phi)\, \tilde{w}(\phi) \quad \text{where} \quad \tilde{w}(\phi) = \frac{\exp\{-S(\phi)\}}{q_\theta(\phi)}$$

$$Z \overset{\text{MC}}{\approx} \hat{Z} = \frac{1}{N} \sum_{i=1}^{N} \tilde{w}(\phi_i) \qquad \phi_i \sim q_\theta$$

Asymptotically Unbiased Estimation of Physical Observables

$$\langle \mathcal{O} \rangle_p = \langle w\mathcal{O} \rangle_{q_\theta} \overset{\text{MC}}{\approx} \frac{1}{N} \sum_{i=1}^{N} w(\phi_i)\mathcal{O}(\phi_i) \quad \phi_i \sim q_\theta$$

Asymptotically Unbiased Estimation of Physical Observables

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

$$\langle \mathcal{O} \rangle_p = \langle w\mathcal{O} \rangle_{q_\theta} \overset{\text{MC}}{\approx} \frac{1}{N} \sum_{i=1}^{N} w(\phi_i)\mathcal{O}(\phi_i) \quad \phi_i \sim q_\theta$$

Asymptotically Unbiased Estimation of Physical Observables

$$\langle \mathcal{O} \rangle_p = \langle w\mathcal{O} \rangle_{q_\theta} \overset{\text{MC}}{\approx} \frac{1}{N} \sum_{i=1}^{N} w(\phi_i)\mathcal{O}(\phi_i) \quad \phi_i \sim q_\theta$$

Direct Estimation of Thermodynamic Quantities

Nicoli, Anders, Funcke, et al., Phys. Rev. Lett. (2021)

$$\hat{F} = -T \ln \hat{Z} \quad ⚠$$

$$\langle \mathcal{O} \rangle_p = \langle w\mathcal{O} \rangle_{q_\theta} \overset{\text{MC}}{\approx} \frac{1}{N} \sum_{i=1}^{N} w(\phi_i)\mathcal{O}(\phi_i) \qquad \phi_i \sim q_\theta$$

$$\hat{p}, \hat{H}$$

The 2D Ising Model

Sample $\phi \sim q_\theta \approx p$

Autoregressive Models
(Pixel CNN)
Van der Oord et al. (2016)

Discrete RV
e.g., Ising

Deep neural network

(Image credits: Kavlakoglu, www.ibm.com)

$p(\phi)$

$q_\theta(\phi)$

Exact Likelihood
Density Estimation

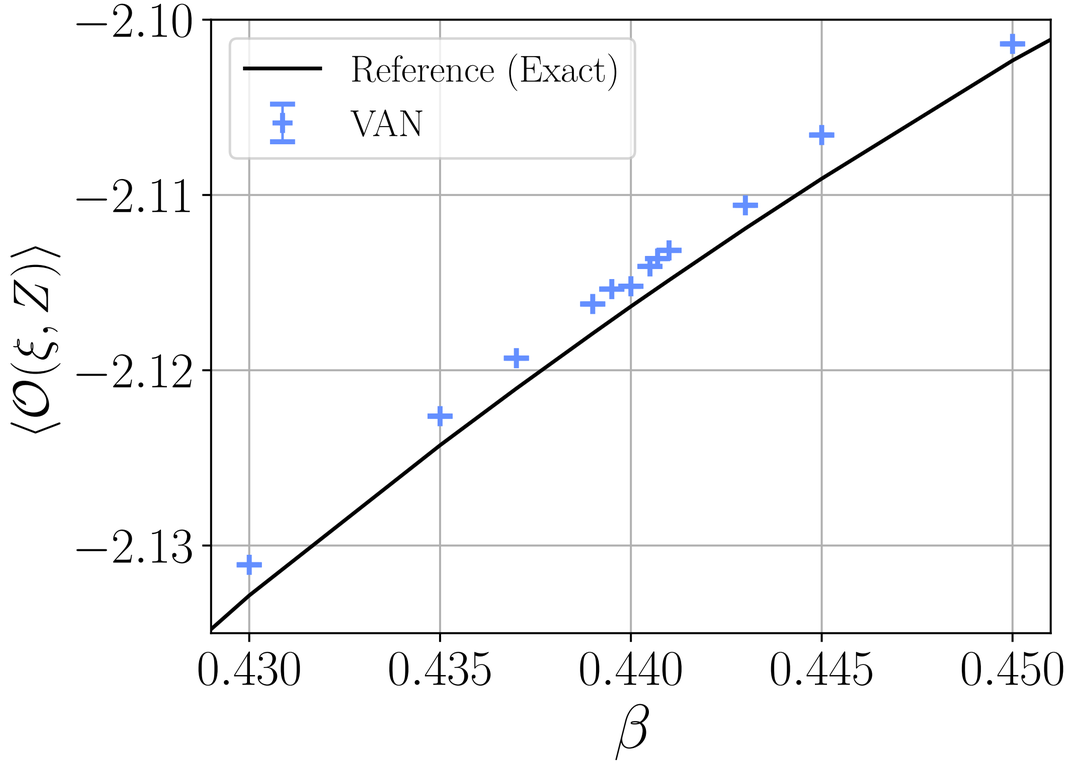
Continuous RV
e.g., QFT

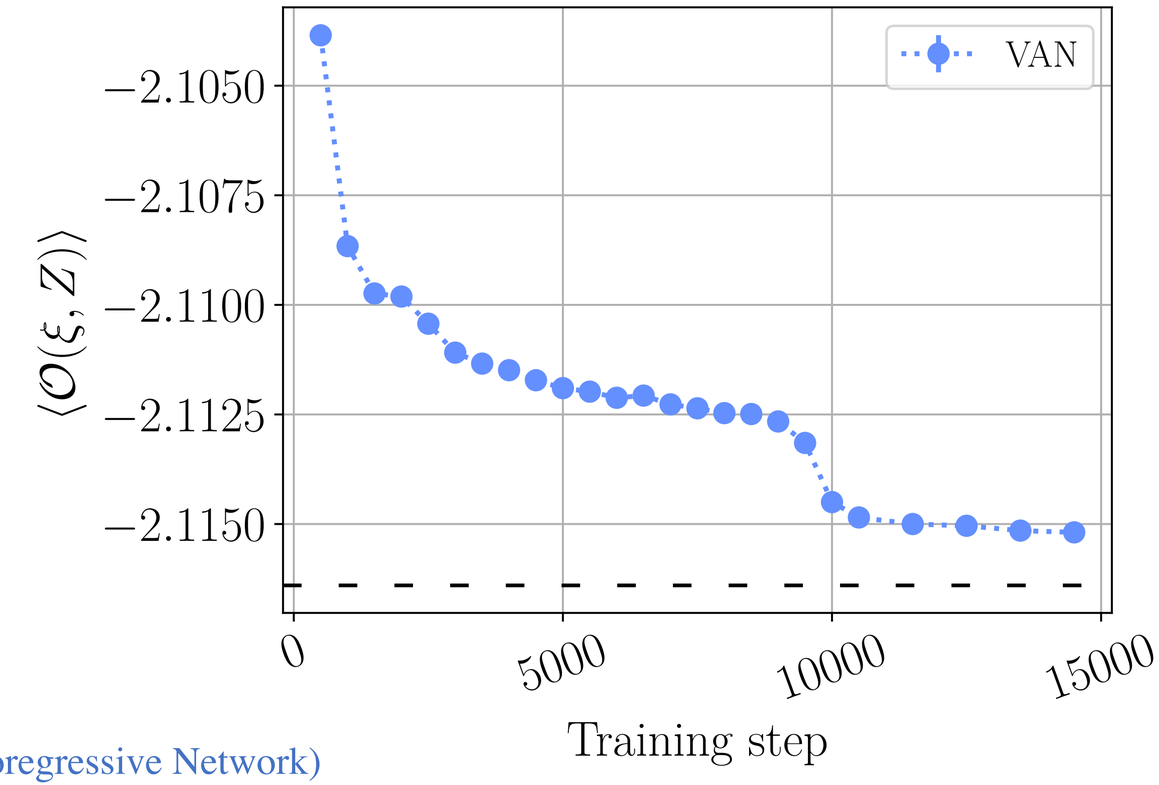
Normalizing Flows
Rezende and Mohamed (2015)

Minimize Cost Function

The 2D Ising Model

$$H(\xi) = -J \sum_{<i,j>} \xi_i \xi_j \qquad p(\xi) = \frac{e^{-\beta H(\xi)}}{Z}, \quad \beta \propto \frac{1}{T} \qquad \langle \mathcal{O} \rangle_{q_\theta} \neq \langle \mathcal{O} \rangle_p$$



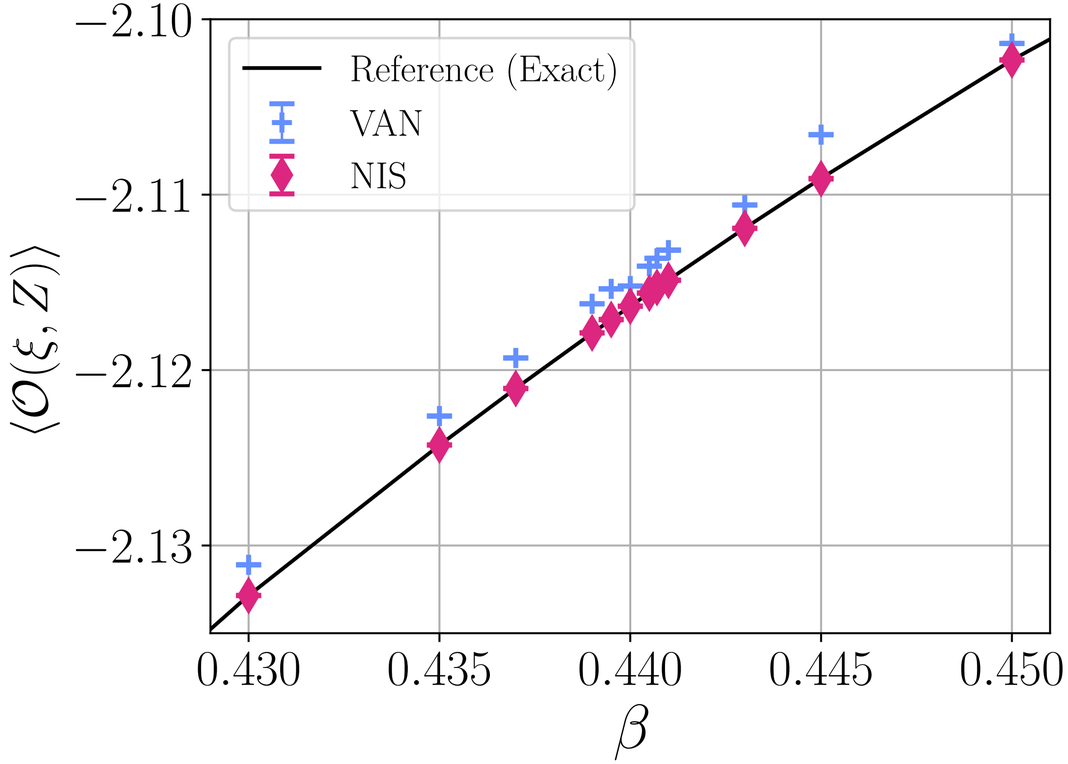


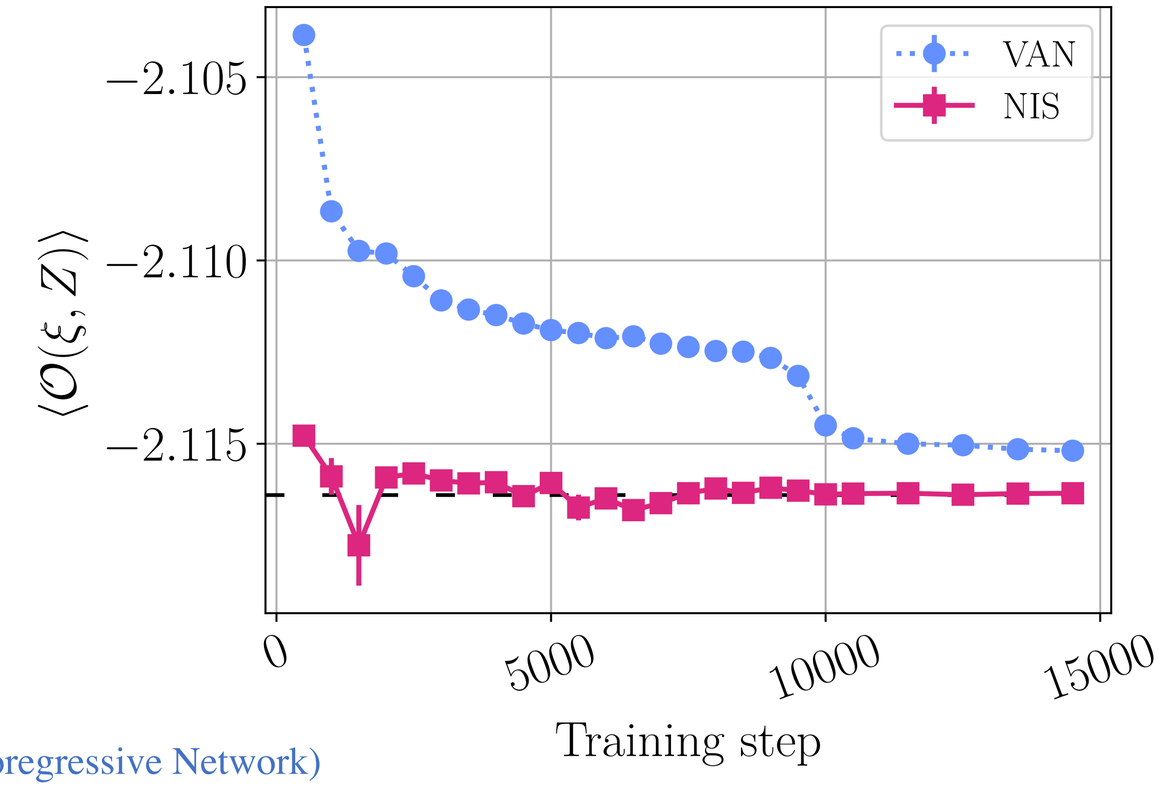
Wu, Wang, Zhang, Phys. Rev. Lett. (2019) $\Longrightarrow$ **VAN** (Variational Autoregressive Network)

The 2D Ising Model

$$H(\xi) = -J \sum_{<i,j>} \xi_i \xi_j \qquad p(\xi) = \frac{e^{-\beta H(\xi)}}{Z}$$

$$\langle \mathcal{O} \rangle_{q_\theta} \longrightarrow \langle w\mathcal{O} \rangle_{q_\theta}$$





Wu, Wang, Zhang, Phys. Rev. Lett. (2019) $\Longrightarrow$ **VAN** (Variational Autoregressive Network)

Nicoli, Nakajima, Strodthoff, Samek, Müller, and Kessel, Phys. Rev. E (2020) $\Longrightarrow$ **NIS** (Neural Importance Sampling)

Application to lattice quantum field theory

Sample $\phi \sim q_\theta \approx p$

$p(\phi)$

Minimize Cost Function

Autoregressive Models
(Pixel CNN)
Van der Oord et al. (2016)

Discrete RV
e.g., Ising

$q_\theta(\phi)$

Exact Likelihood
Density Estimation

Continuous RV
e.g., QFT

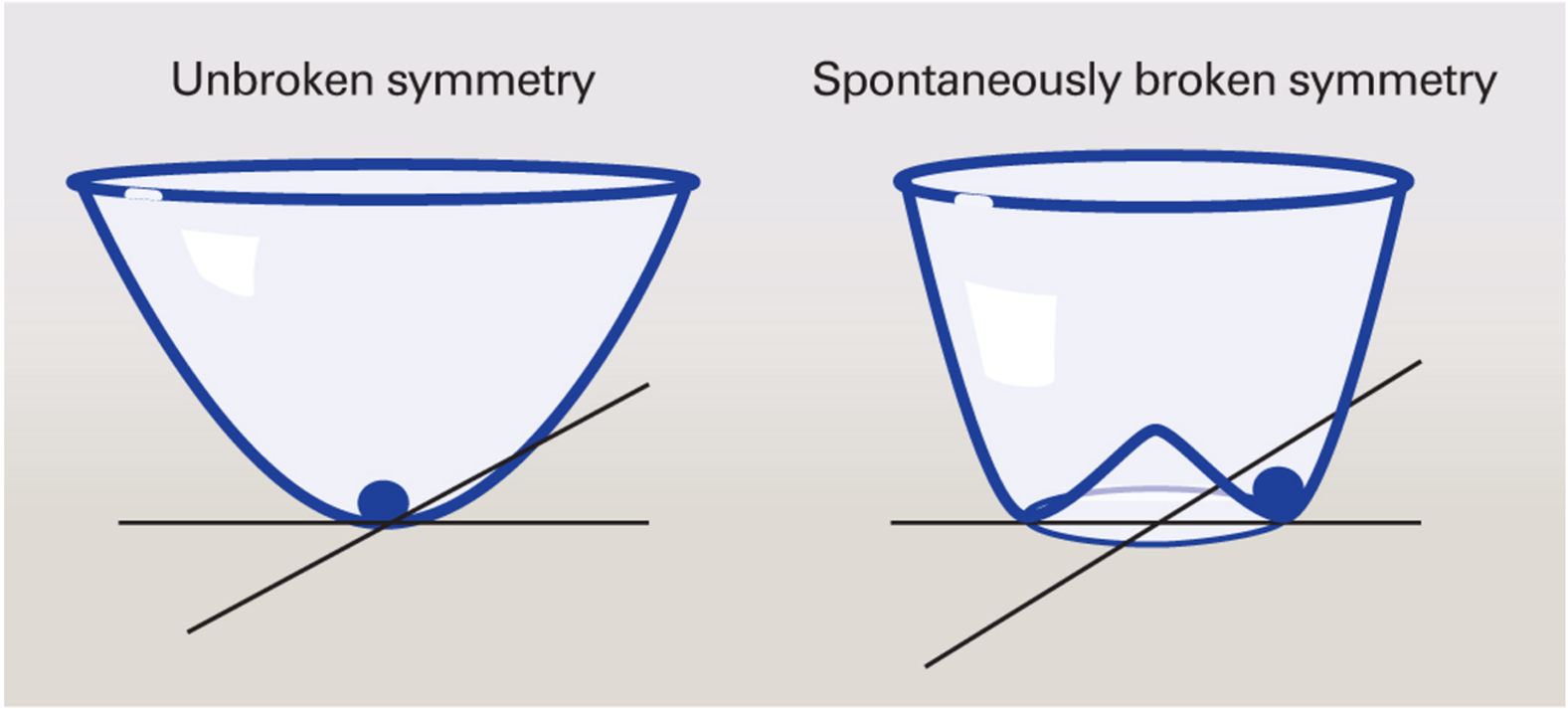
Normalizing Flows
Rezende and Mohamed (2015)

Real Scalar ϕ^4-Theory in (1+1) D

$$S(\phi) = \sum_{x \in \Lambda} \left\{ -2\kappa \sum_{\mu=1}^{d} \phi(x)\phi(x+\hat{\mu}) + (1-2\lambda)\phi(x)^2 + \lambda\phi(x)^4 \right\}$$

Real Scalar ϕ^4-Theory in (1+1) D

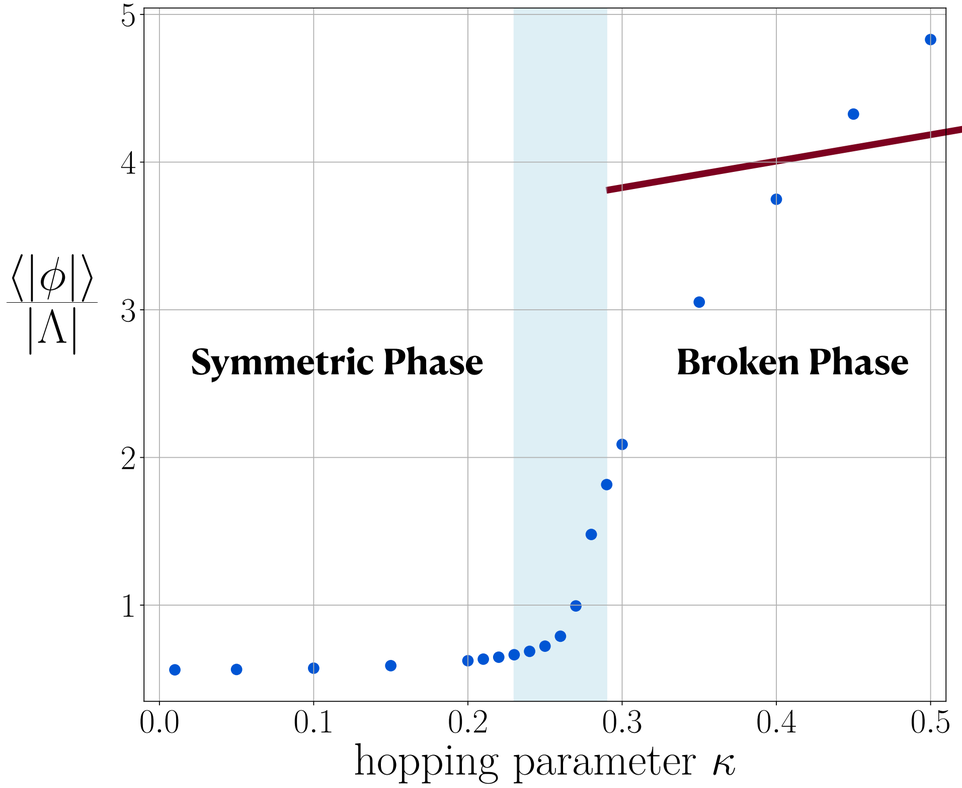
$$S(\phi) = \sum_{x \in \Lambda} \left\{ -2\kappa \sum_{\mu=1}^{d} \phi(x)\phi(x+\hat{\mu}) + (1-2\lambda)\phi(x)^2 + \lambda\phi(x)^4 \right\}$$



(Image credits: Science)

Spontaneous Breaking of $\mathbb{Z}_2$ Symmetry

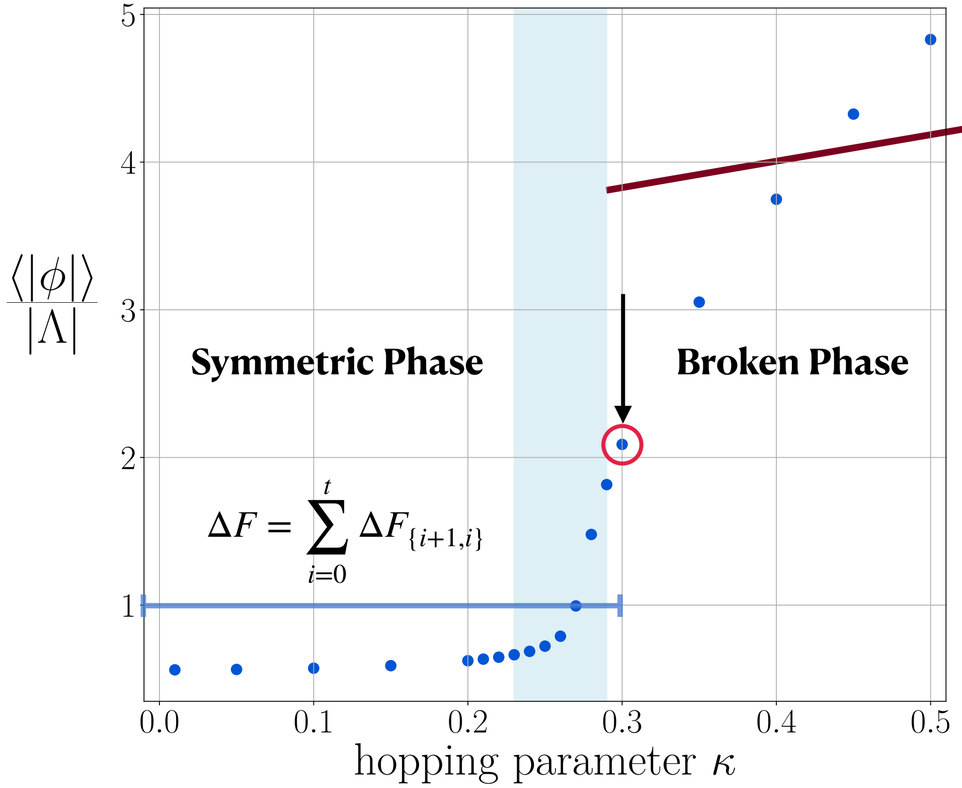
$$S(\phi) = \sum_{x \in \Lambda} \left\{ -2\kappa \sum_{\mu=1}^{d} \phi(x)\phi(x+\hat{\mu}) + (1-2\lambda)\phi(x)^2 + \lambda\phi(x)^4 \right\}$$



PHASE TRANSITION

Spontaneous Breaking of $\mathbb{Z}_2$ Symmetry

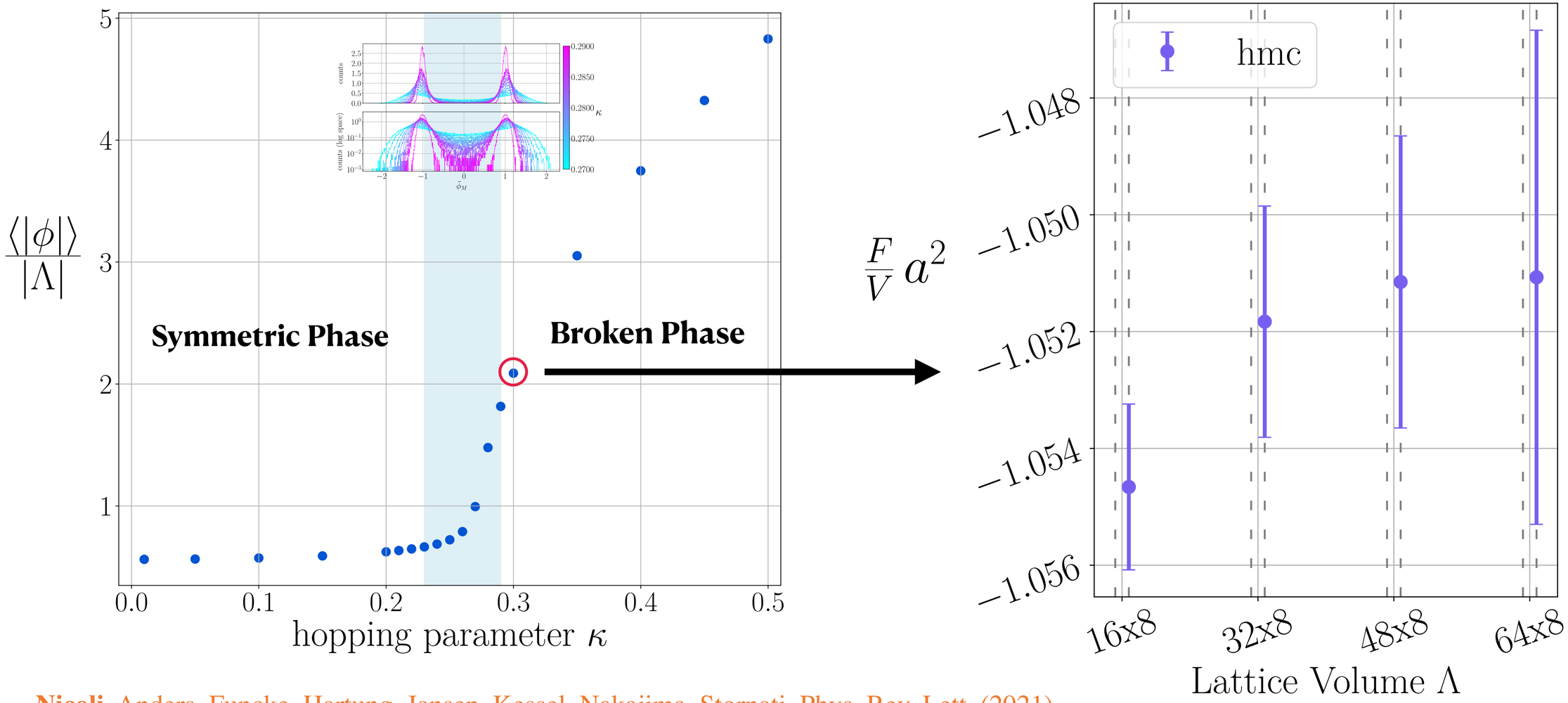
$$S(\phi) = \sum_{x \in \Lambda} \left\{ -2\kappa \sum_{\mu=1}^{d} \phi(x)\phi(x+\hat{\mu}) + (1-2\lambda)\phi(x)^2 + \lambda\phi(x)^4 \right\}$$



PHASE TRANSITION

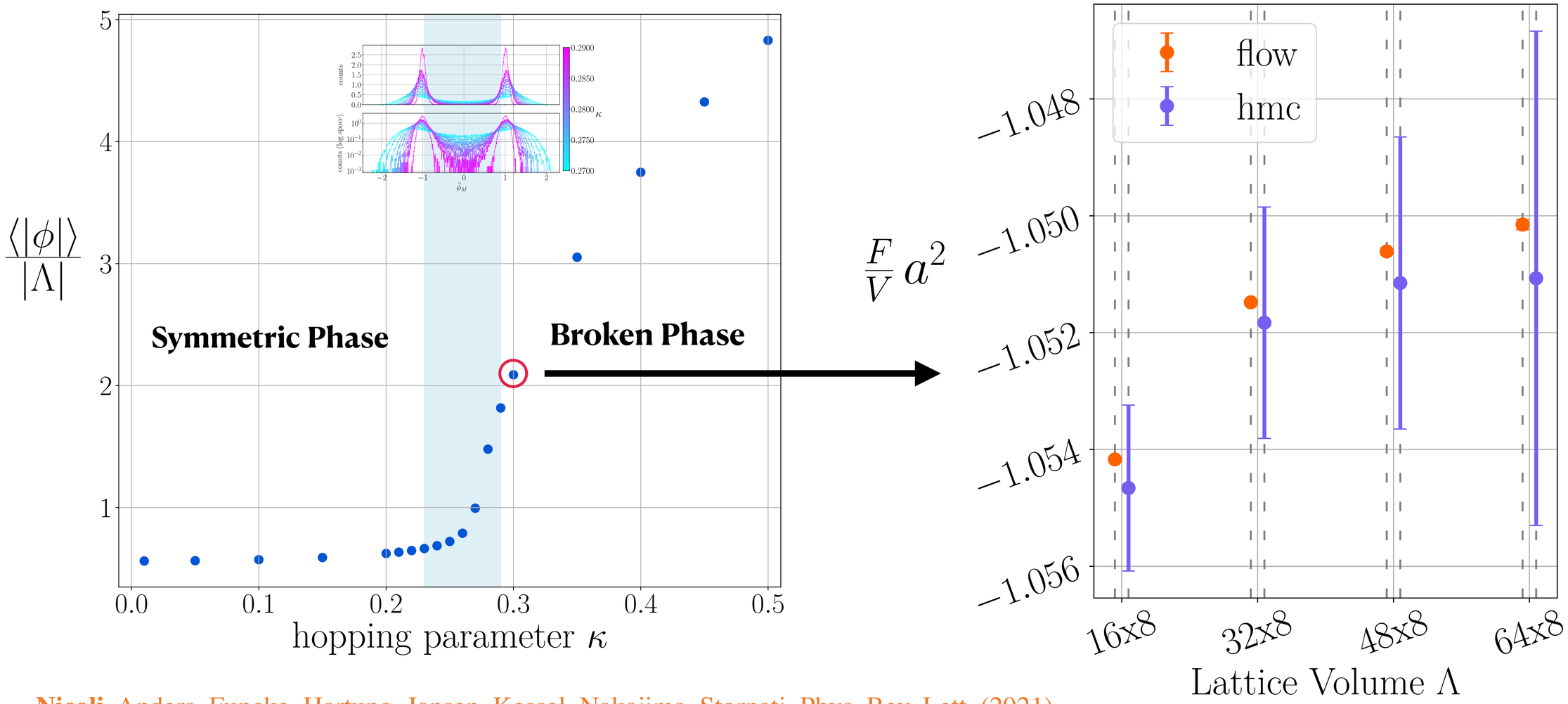
- $\Delta F = F_t - F_0 = -T \ln Z_t/Z_0$
- F_0 is needed to compute F_{target}
- Integration through phase space (MCMC)
- Intermediate steps $\Delta F_{\{i+1,i\}} = -T \ln \frac{Z_{i+1}}{Z_i}$

Estimation of Thermodynamic Observables



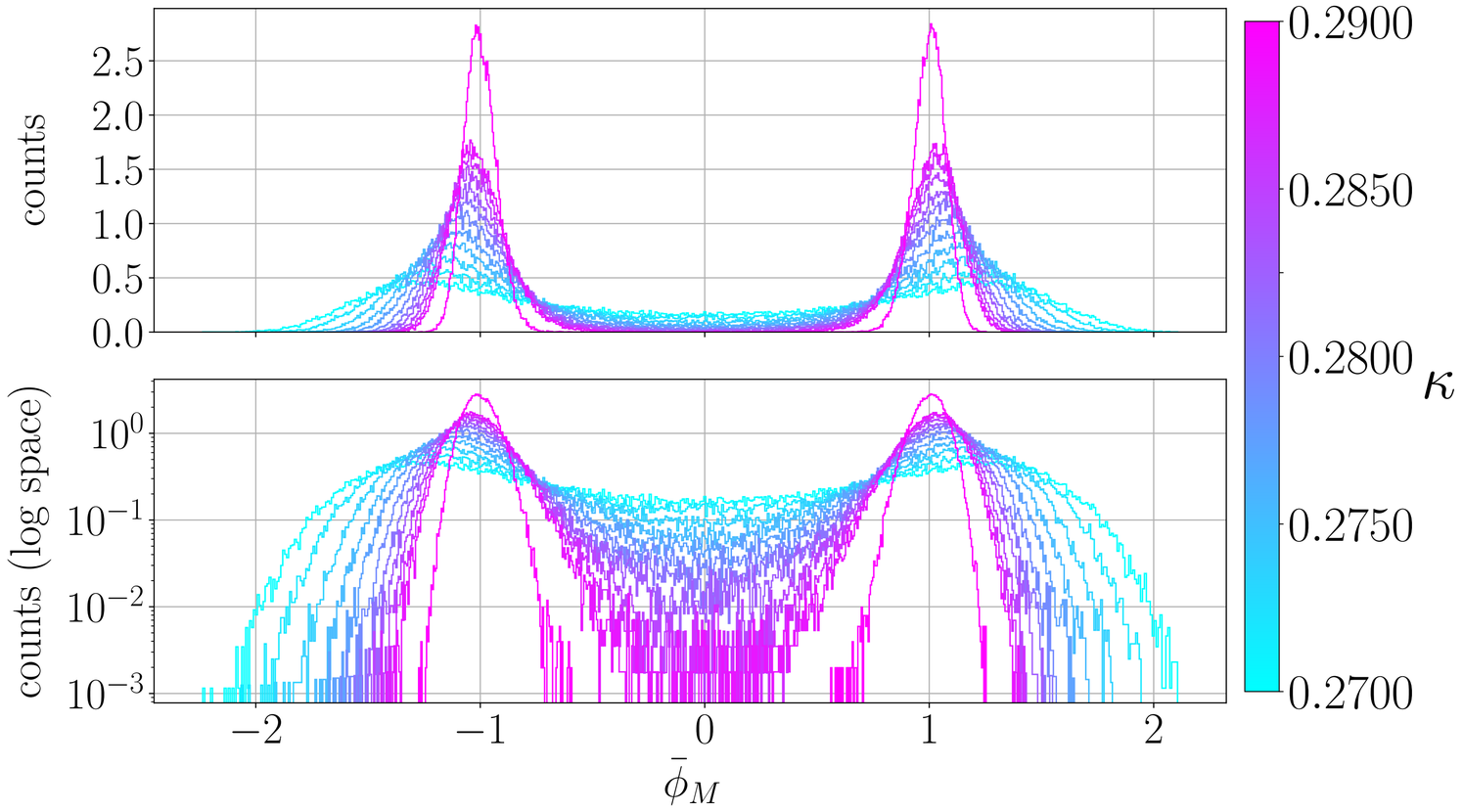
Nicoli, Anders, Funcke, Hartung, Jansen, Kessel, Nakajima, Stornati, Phys. Rev. Lett. (2021)

Estimation of Thermodynamic Observables



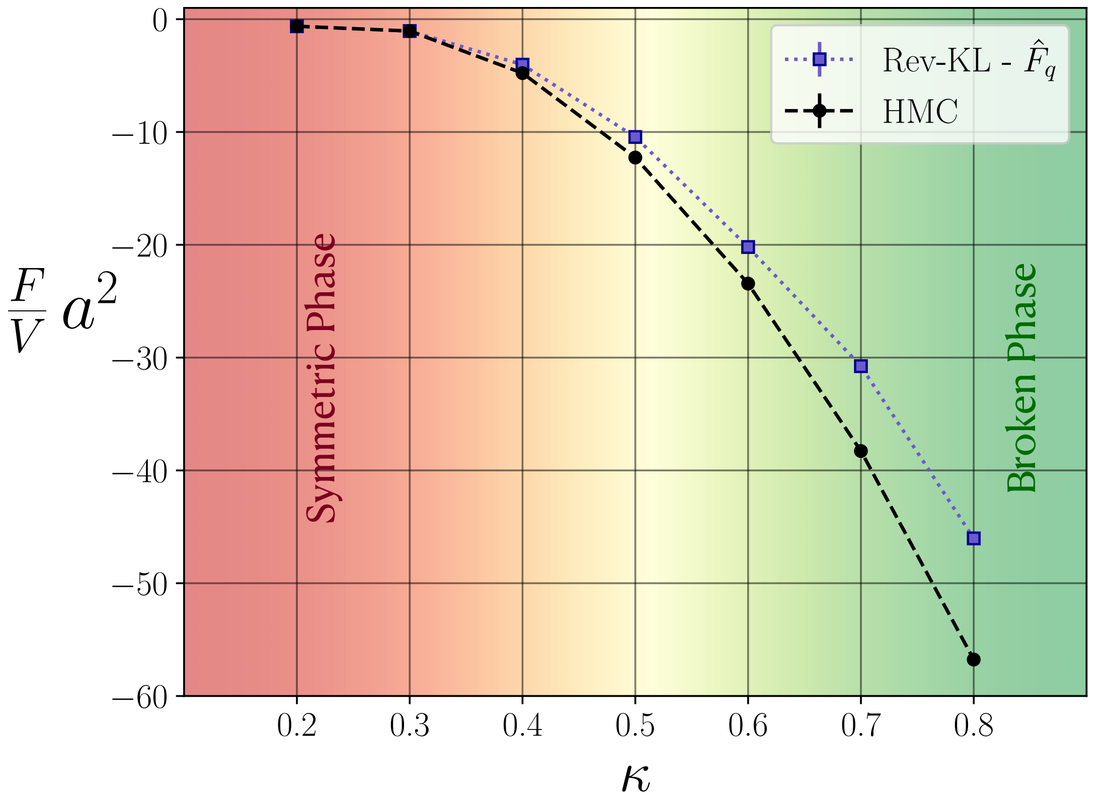
Nicoli, Anders, Funcke, Hartung, Jansen, Kessel, Nakajima, Stornati, Phys. Rev. Lett. (2021)

Real Scalar ϕ^4-Theory in (1+1) D



Nicoli, Anders, Hartung, et al., Phys. Rev. D (2023)

Real Scalar ϕ^4-Theory in (1+1) D



q- estimator for the free energy

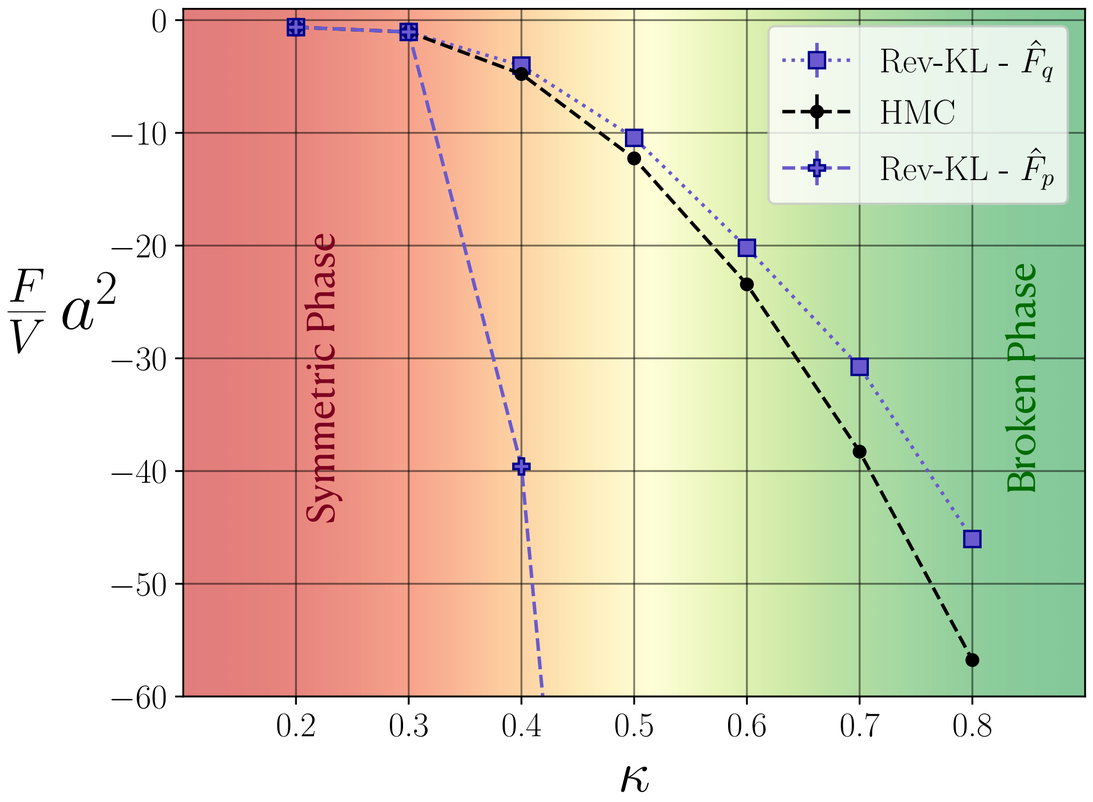
$$Z = \mathbb{E}_{\phi\sim q_\theta}\left[\frac{e^{-S(\phi)}}{q_\theta(\phi)}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{e^{-S(\phi_i)}}{q_\theta(\phi_i)} \equiv \hat{Z}_{q_\theta}$$

$$\phi_i \sim q_\theta$$

$$\hat{F}_q = -\,T\log(\hat{Z}_{q_\theta})$$

Nicoli, Anders, Hartung, et al., Phys. Rev. D (2023)

Real Scalar ϕ^4-Theory in (1+1) D



Nicoli, Anders, Hartung, et al., Phys. Rev. D (2023)

q- estimator for the free energy

$$Z = \mathbb{E}_{\phi\sim q_\theta}\left[\frac{e^{-S(\phi)}}{q_\theta(\phi)}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{e^{-S(\phi_i)}}{q_\theta(\phi_i)} \equiv \hat{Z}_{q_\theta}$$

$$\phi_i \sim q_\theta$$

$$\hat{F}_q = -\,T\log(\hat{Z}_{q_\theta})$$

p- estimator for the free energy

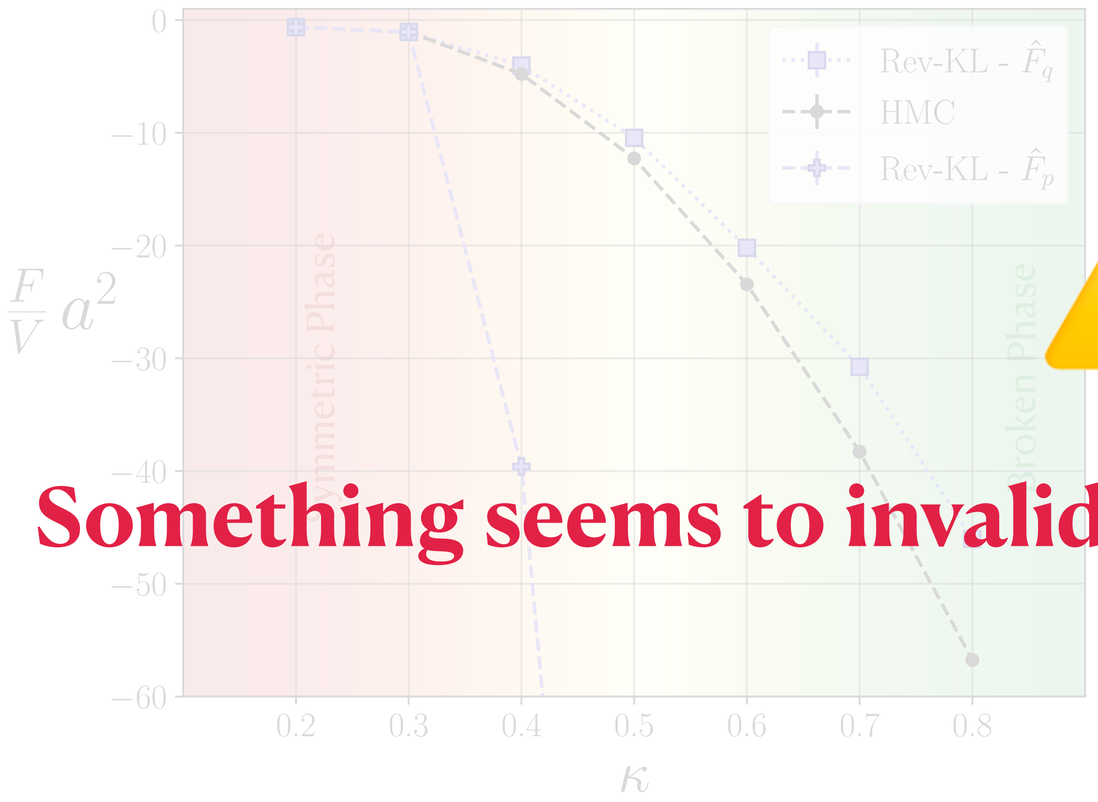
$$Z^{-1} = \mathbb{E}_{\phi\sim p}\left[\frac{q_\theta(\phi)}{e^{-S(\phi)}}\right] \approx \frac{1}{N}\sum_{j=1}^{N}\frac{q_\theta(\phi_j)}{e^{-S(\phi_j)}} \equiv \hat{Z}_p^{-1}$$

$$\phi_j \sim p$$

$$\hat{F}_p = T\log(\hat{Z}_p^{-1})$$ *

*Samples are obtained from HMC away from criticality

Real Scalar ϕ^4-Theory in (1+1) D



$$Z = \mathbb{E}_{\phi\sim q_\theta}\left[\frac{e^{-S(\phi)}}{q_\theta(\phi)}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{e^{-S(\phi_i)}}{q_\theta(\phi_i)} \equiv \hat{Z}_{q_\theta}$$

$$\phi_i \sim q_\theta$$

$$\hat{F}_q = -\,T\log(\hat{Z}_{q_\theta})$$

$$Z = \mathbb{E}_{\phi\sim p}\left[\frac{e^{-S(\phi)}}{p(\phi)}\right] \approx \frac{1}{N}\sum_{j=1}^{N}\frac{e^{-S(\phi_j)}}{p(\phi_j)} \equiv \hat{Z}_{p}$$

$$\phi_j \sim p$$

$$\hat{F}_p = T\log(\hat{Z}_p^{-1})$$

Something seems to invalidate our asymptotic guarantees!

Nicoli, Anders, Hartung, et al., Phys. Rev. D (2023)

What’s going wrong then?

Reverse-KL Div.

$$KL_R(q_\theta||p) = \int D[\phi] q_\theta(\phi) \ln \frac{q_\theta(\phi)}{p(\phi)}$$

What's going wrong then?

Reverse-KL Div.

$$KL_R(q_\theta||p) = \int D[\phi] q_\theta(\phi) \ln \frac{q_\theta(\phi)}{p(\phi)}$$

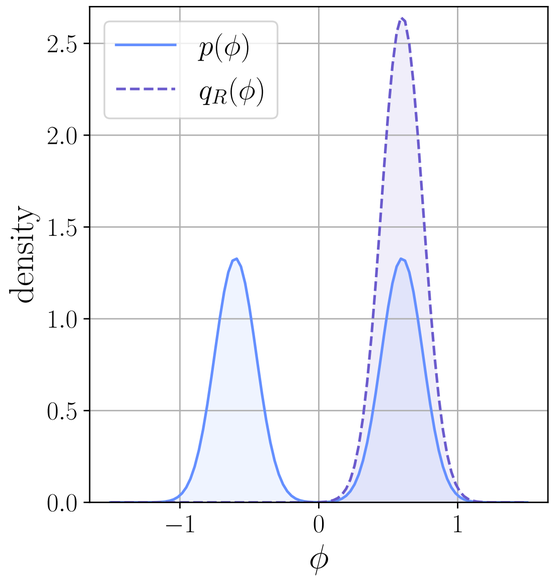
Forward-KL Div.

$$KL_F(p||q_\theta) = \int D[\phi] p(\phi) \ln \frac{p(\phi)}{q_\theta(\phi)}$$

What's going wrong then?

Reverse-KL Div.

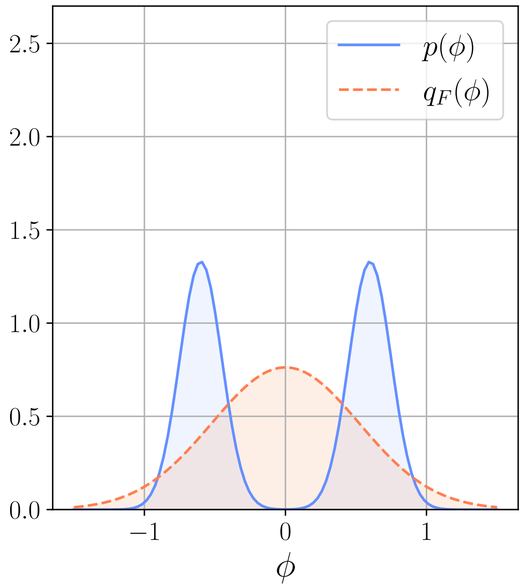
$$KL_R(q_\theta||p) = \int D[\phi] q_\theta(\phi) \ln \frac{q_\theta(\phi)}{p(\phi)}$$



- Self-Sampling (efficient).
- No need for training data.
- Mode-dropping.
- q_R flow trained with Rev. KL

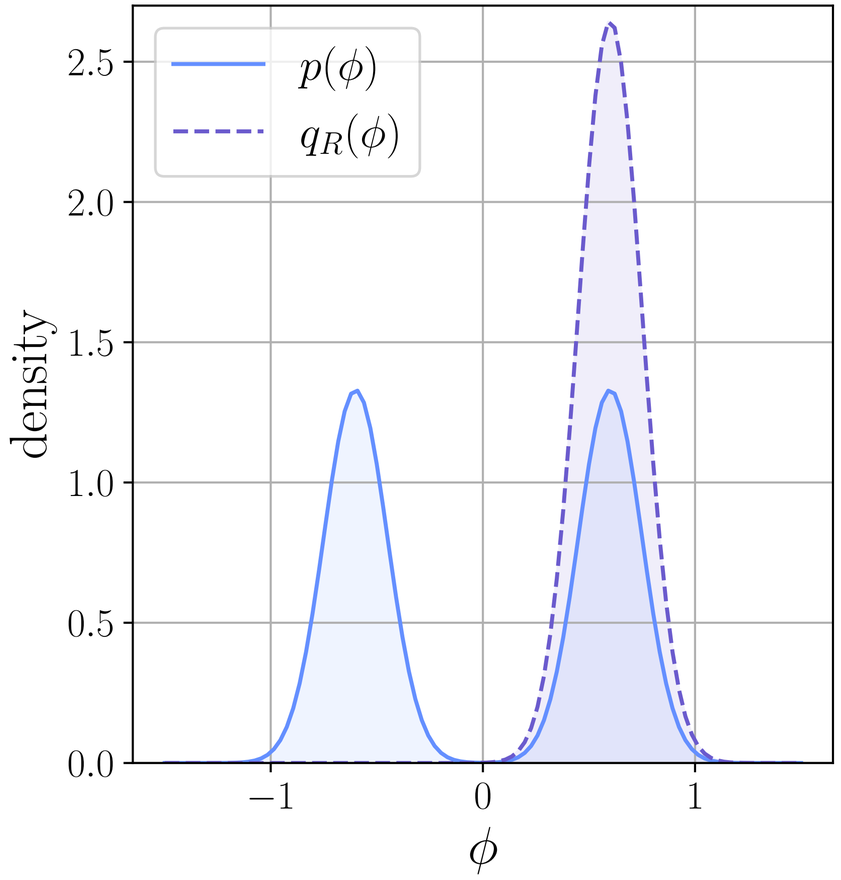
Forward-KL Div.

$$KL_F(p||q_\theta) = \int D[\phi] p(\phi) \ln \frac{p(\phi)}{q_\theta(\phi)}$$



- Maximum Likelihood.
- Requires training data.
- Fake modes.
- q_F flow trained with Fwd. KL

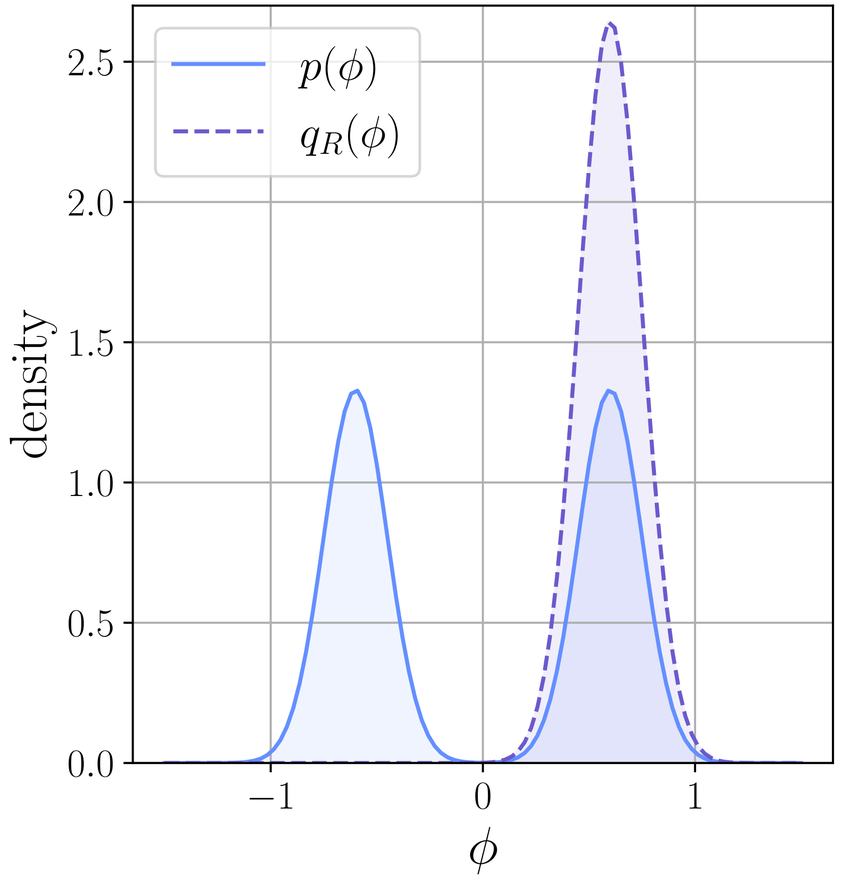
What's going wrong then?



$$\mathrm{ESR} = \frac{\mathrm{ESS}}{N} = \frac{1}{\mathbb{E}_{q_\theta}[w^2]}$$

Where $$w = \frac{p(\phi)}{q_\theta(\phi)}$$

What's going wrong then?



The Effective sampling ratio (ESR) is not a good metric!

$$\mathrm{ESR} = \frac{\mathrm{ESS}}{N} = \frac{1}{\mathbb{E}_{q_\theta}[w^2]}$$

Where $$w = \frac{p(\phi)}{q_\theta(\phi)}$$

The model is **blind** with respect to one (or more) of the modes of the target density.

The mode-dropping estimator

When q_θ has **full effective support** on the domain of p

$$w^* = \mathbb{E}_{q_\theta}\left[\frac{p(\phi)}{q_\theta(\phi)}\right] = \int_{\mathrm{supp}(q_\theta)} q_\theta(\phi)\frac{p(\phi)}{q_\theta(\phi)}\,\mathcal{D}[\phi] = \int_{\mathrm{supp}(q_\theta)} p(\phi)\,\mathcal{D}[\phi] = 1\,.$$

The mode-dropping estimator

When q_θ has **full effective support** on the domain of p

$$w^* = \mathbb{E}_{q_\theta}\left[\frac{p(\phi)}{q_\theta(\phi)}\right] = \int_{\mathrm{supp}(q_\theta)} q_\theta(\phi)\frac{p(\phi)}{q_\theta(\phi)}\,\mathcal{D}[\phi] = \int_{\mathrm{supp}(q_\theta)} p(\phi)\,\mathcal{D}[\phi] = 1\,.$$

however if q_θ is **effectively mode-dropping** this expectation value becomes

$$\bar{w} \equiv \frac{1}{Z}\mathbb{E}_{\phi\sim\widetilde{q}_\theta}\left[\frac{e^{-S(\phi)}}{q_\theta(\phi)}\right] \quad \in (0,1]$$

The mode-dropping estimator

When q_θ has **full effective support** on the domain of p

$$w^* = \mathbb{E}_{q_\theta}\left[\frac{p(\phi)}{q_\theta(\phi)}\right] = \int_{\mathrm{supp}(q_\theta)} q_\theta(\phi)\frac{p(\phi)}{q_\theta(\phi)}\,\mathcal{D}[\phi] = \int_{\mathrm{supp}(q_\theta)} p(\phi)\,\mathcal{D}[\phi] = 1\,.$$

however if q_θ is **effectively mode-dropping** this expectation value becomes

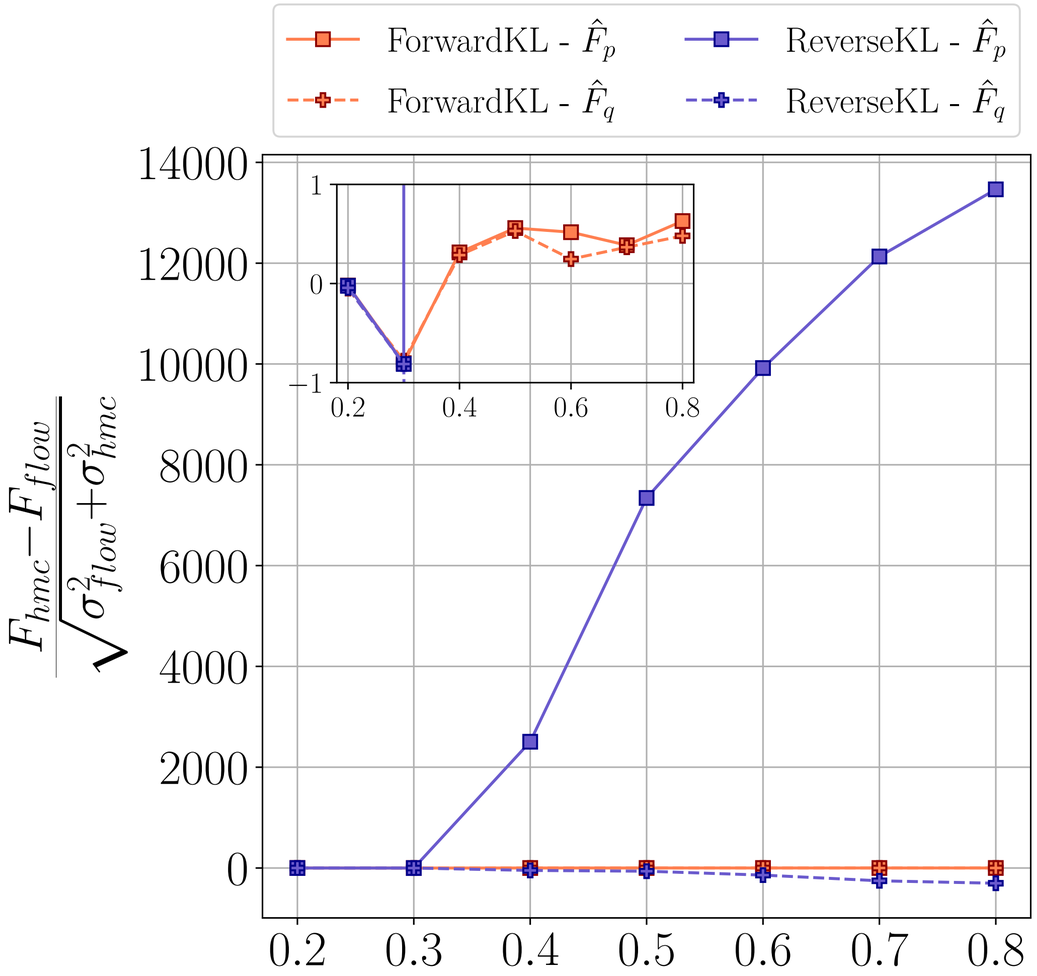
$$\bar{w} \equiv \frac{1}{Z}\mathbb{E}_{\phi\sim\tilde{q}_\theta}\left[\frac{e^{-S(\phi)}}{q_\theta(\phi)}\right] \quad \in (0,1]$$

for which we can get the corresponding Monte Carlo estimator, i.e., the **mode-dropping estimator**

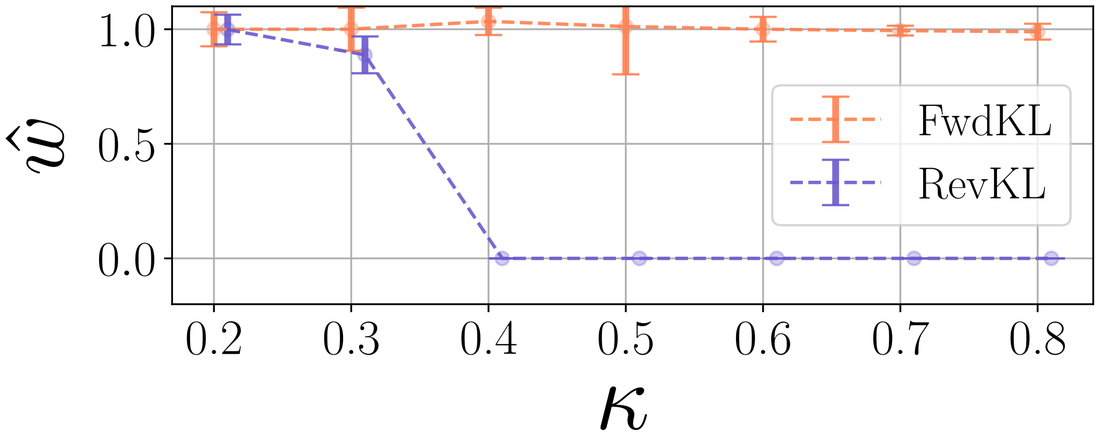
$$\bar{w} \approx \frac{1}{\hat{Z}_p}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{e^{-S(\phi_i)}}{q_\theta(\phi_i)}\right) = \left(\frac{1}{N}\sum_{j=1}^{N}\frac{q_\theta(\phi_j)}{e^{-S(\phi_j)}}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\frac{e^{-S(\phi_i)}}{q_\theta(\phi_i)}\right) \equiv \hat{w}$$

$\phi_j \sim p$ $\qquad$ $\phi_i \sim q_\theta$

Estimation of Mode Dropping Across Criticality



$$\Lambda = 64 \times 8,\ \lambda = 0.022$$



Nicoli, Anders, Hartung, et al., Phys. Rev. D (2023)

Summary and Conclusions

- **Asymptotically unbiased samplers** can be constructed from trained DGMs (NIS or NMCMC).
- **Direct** estimation of the **partition function** and **thermodynamic observables**.
- Improved DGMs can use **inductive biases**, e.g. symmetries, bootstrapping, annealing, etc.
- Sampling from DGMs is **embarrassingly parallelizable (i.i.d)** ≠ MCMC (**sequential**).
- Training with **forward-KL** leads to better models though requires training samples.
- Derivation of **mode-dropping estimator** to reliably assess the goodness of the model.
- **Mitigation** of mode-dropping using different **objectives** (FWD-KL) or **stochastic** approaches (SNFs).

Summary and Conclusions

- **Asymptotically unbiased samplers** can be constructed from trained DGMs (NIS or NMCMC).
- **Direct** estimation of the **partition function** and **thermodynamic observables**.
- Improved DGMs can use **inductive biases**, e.g. symmetries, bootstrapping, annealing, etc.
- Sampling from DGMs is **embarrassingly parallelizable** (**i.i.d**) ≠ MCMC (**sequential**).
- Training with **forward-KL** leads to better models though requires training samples.
- Derivation of **mode-dropping estimator** to reliably assess the goodness of the model.
- **Mitigation** of mode-dropping using different **objectives** (FWD-KL) or **stochastic** approaches (SNFs).

TL;DR

Deep Generative Models (DGMs) are promising candidates
for the next generation of sampling algorithms

Thank You!!

“It is nice to know that the computer understands the problem. But I would like to understand it too.”
- E. Wigner

Bonn:

UNIVERSITÄT BONN
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN

TRA
MATTER
TRANSDISCIPLINARY RESEARCH AREA

HISKP

LAMARR
INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

bctp

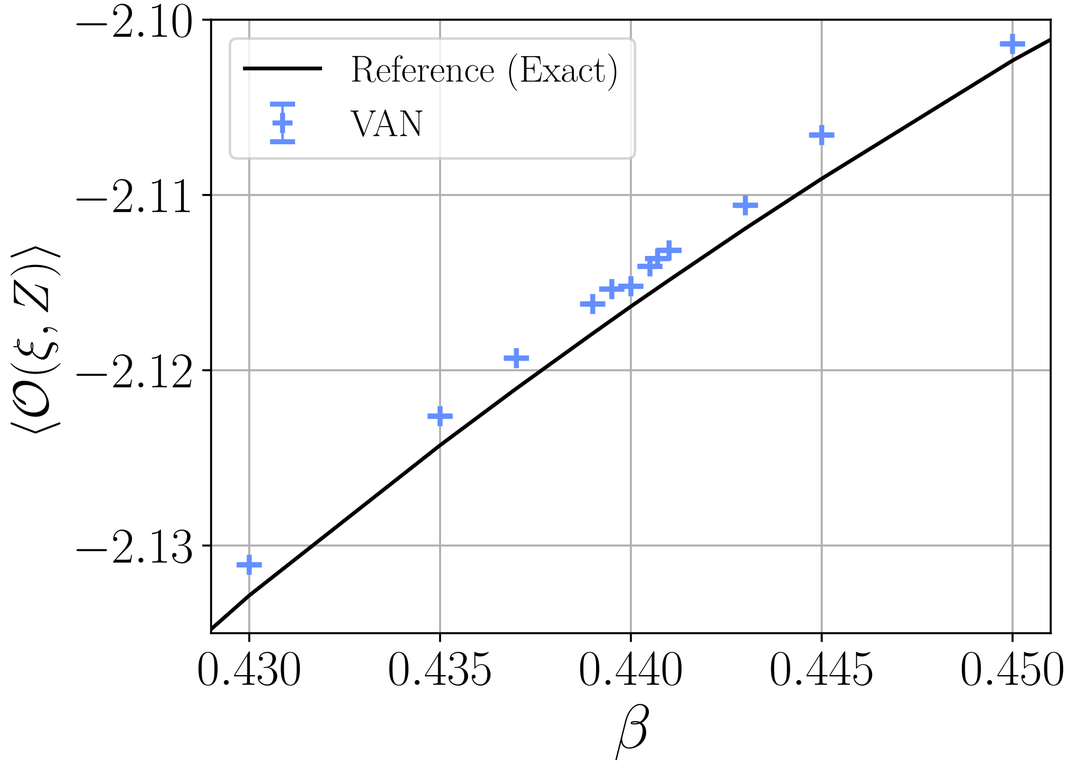
Berlin:

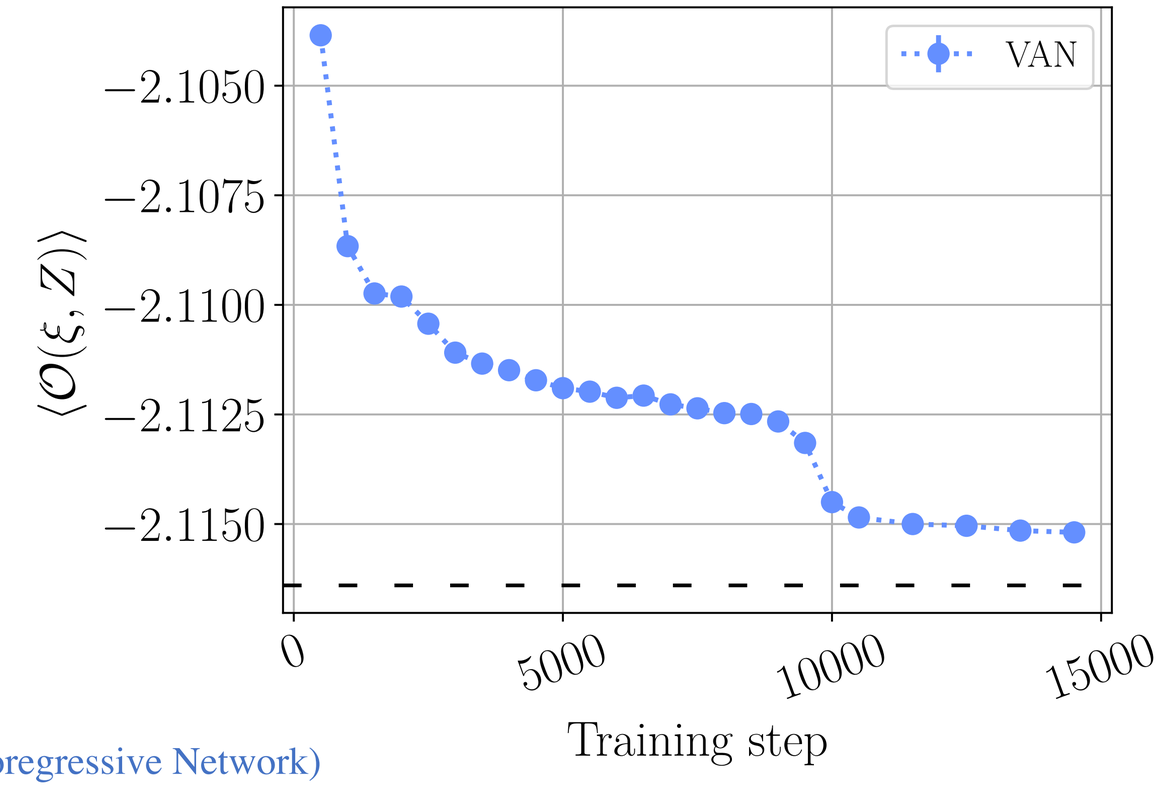
TECHNISCHE UNIVERSITÄT BERLIN

BIFOLD

The 2D Ising Model

$$H(\xi) = -J \sum_{<i,j>} \xi_i \xi_j \qquad p(\xi) = \frac{e^{-\beta H(\xi)}}{Z}, \quad \beta \propto \frac{1}{T} \qquad \langle \mathcal{O} \rangle_{q_\theta} \neq \langle \mathcal{O} \rangle_p$$





Wu, Wang, Zhang, Phys. Rev. Lett. (2019) $\Longrightarrow$ **VAN** (Variational Autoregressive Network)

Neural Markov chain Monte-Carlo (NMCMC)

$$p(\phi) \approx q_\theta \sim \phi_i$$

Neural Markov chain Monte-Carlo (NMCMC)

$$p(\phi) \approx q_\theta \sim \phi_i$$

Standard Metropolis

$$A\left(\phi', \phi_i\right) = \min\left(1, \frac{p(\phi')\, s(\phi_i \mid \phi')}{p(\phi_i)\, s(\phi' \mid \phi_i)}\right)$$

$$\xrightarrow{s\left(\phi' \mid \phi_i\right) \to q_\theta(\phi')}$$

Neural MCMC

$$A\left(\phi', \phi_i\right) = \min\left(1, \frac{q_\theta(\phi_i)\, e^{-S(\phi')}}{q_\theta(\phi')\, e^{-S(\phi_i)}}\right)$$

Neural Markov chain Monte-Carlo (NMCMC)

Standard Metropolis **Neural MCMC**

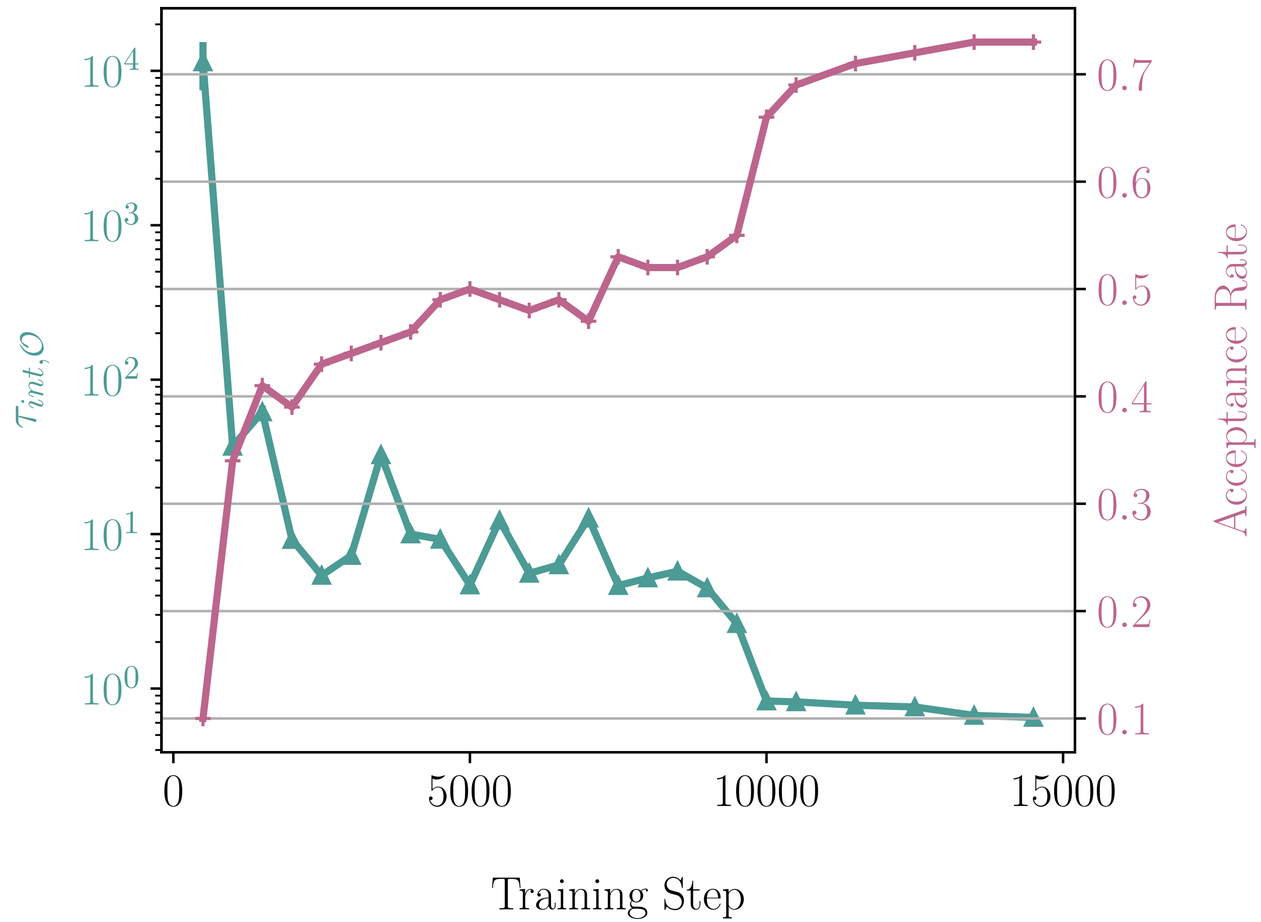
$$A\left(\phi', \phi_i\right) = \min\left(1, \frac{p(\phi')\, s(\phi_i \mid \phi')}{p(\phi_i)\, s(\phi' \mid \phi_i)}\right) \xrightarrow{s\left(\phi' \mid \phi_i\right) \to q_\theta(\phi')} A\left(\phi', \phi_i\right) = \min\left(1, \frac{q_\theta(\phi_i)\, e^{-S(\phi')}}{q_\theta(\phi')\, e^{-S(\phi_i)}}\right)$$

NMCMC allows for **independent sampling** while retaining **detail balanced** condition fulfilled

$$\begin{aligned} p(\phi' \mid \phi)\, e^{-S(\phi)} &= q_\theta(\phi')\, A\left(\phi', \phi\right) e^{-S(\phi)} \\ &= q_\theta(\phi')\, e^{-S(\phi)} \min\left(1, \frac{q_\theta(\phi)\, e^{-S(\phi')}}{q_\theta(\phi')\, e^{-S(\phi)}}\right) \\ &= q_\theta(\phi)\, e^{-S(\phi')} \min\left(\frac{q_\theta(\phi')\, e^{-S(\phi)}}{q_\theta(\phi)\, e^{-S(\phi')}}, 1\right) = p(\phi \mid \phi')\, e^{-S(\phi')} \end{aligned}$$

- min operator is symmetric.
- All factor are strictly positive.

Neural Markov chain Monte-Carlo (NMCMC)



Nicoli, Nakajima, Strodthoff, Samek, Müller, and Kessel, Phys. Rev. E (2020)

The effect of mode-dropping

Definition. *The effective support of the variational density q_θ relative to p*

$$\widetilde{\mathrm{supp}}_{p,\epsilon}(q_\theta) = \{\phi \in \mathrm{supp}(q_\theta);\ q_\theta(\phi) > \epsilon p(\phi)\}$$

for a given numerical threshold ϵ. The mode dropping set is then given by

$$\mathcal{S} := \mathrm{supp}(p) \setminus \widetilde{\mathrm{supp}}_{p,\epsilon}(q_\theta)$$

The effect of mode-dropping

Definition. *The effective support of the variational density q_θ relative to p*

$$\widetilde{\mathrm{supp}}_{p,\epsilon}(q_\theta) = \{\phi \in \mathrm{supp}(q_\theta);\ q_\theta(\phi) > \epsilon p(\phi)\}$$

for a given numerical threshold ϵ. The mode dropping set is then given by

$$\mathcal{S} := \mathrm{supp}(p) \setminus \widetilde{\mathrm{supp}}_{p,\epsilon}(q_\theta)$$

if the flow is **effectively mode-dropping**, the importance-weighted estimator, with a finite number of samples N, will miss a contribution from the mass $\int_{\mathcal{S}} p(\phi)d\phi$ with approximately the probability $1 - \epsilon N \int_{\mathcal{S}} p(\phi)d\phi$.

The effect of mode-dropping

Definition. *We define the effective sampler distribution*

$$\widetilde{q}_\theta(\phi) = \begin{cases} q_\theta(\phi)/\zeta & \text{if } \phi \in \widetilde{\text{supp}}_{p,\epsilon}(q_\theta) \\ 0 & \text{otherwise,} \end{cases} \qquad \textit{where} \qquad \zeta = \int_{\widetilde{supp}_{p,\epsilon}} \mathcal{D}[\phi] q_\theta(\phi) \leq 1$$

is the multiplicative renormalization factor necessary to guarantee the normalization of $\widetilde{q}_\theta(\phi)$.

The effect of mode-dropping

Definition. *We define the effective sampler distribution*

$$\widetilde{q}_\theta(\phi) = \begin{cases} q_\theta(\phi)/\zeta & \text{if } \phi \in \widetilde{\text{supp}}_{p,\epsilon}(q_\theta) \\ 0 & \text{otherwise,} \end{cases} \qquad \textit{where} \qquad \zeta = \int_{\widetilde{supp}_{p,\epsilon}} \mathcal{D}[\phi] q_\theta(\phi) \leq 1$$

is the multiplicative renormalization factor necessary to guarantee the normalization of $\widetilde{q}_\theta(\phi)$.

It follows that the importance-weighted estimator misses the contribution from the **mode-dropping set** $\mathcal{S}$

$$\hat{\mathcal{O}} \equiv \frac{1}{N} \sum_{i=1}^{N} \frac{p(\phi_i)}{q_\theta(\phi_i)} \mathcal{O}(\phi_i) \approx \mathbb{E}_{\phi \sim \widetilde{q}_\theta} \left[\frac{p(\phi)}{q_\theta(\phi)} \mathcal{O}(\phi) \right] \equiv \bar{\mathcal{O}}$$

The effect of mode-dropping

Definition. *We define the effective sampler distribution*

$$\widetilde{q}_\theta(\phi) = \begin{cases} q_\theta(\phi)/\zeta & \text{if } \phi \in \widetilde{\text{supp}}_{p,\epsilon}(q_\theta) \\ 0 & \text{otherwise,} \end{cases} \qquad \textit{where} \qquad \zeta = \int_{\widetilde{supp}_{p,\epsilon}} \mathcal{D}[\phi] q_\theta(\phi) \leq 1$$

is the multiplicative renormalization factor necessary to guarantee the normalization of $\widetilde{q}_\theta(\phi)$.

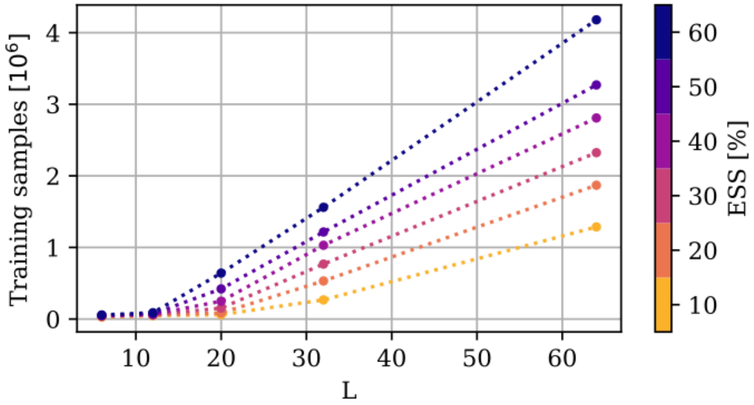
It follows that the importance-weighted estimator misses the contribution from the **mode-dropping set** $\mathcal{S}$

$$\hat{\mathcal{O}} \equiv \frac{1}{N}\sum_{i=1}^{N} \frac{p(\phi_i)}{q_\theta(\phi_i)}\mathcal{O}(\phi_i) \approx \mathbb{E}_{\phi\sim\widetilde{q}_\theta}\left[\frac{p(\phi)}{q_\theta(\phi)}\mathcal{O}(\phi)\right] \equiv \bar{\mathcal{O}}$$

the typical values of the estimator $\hat{\mathcal{O}} \approx \bar{\mathcal{O}}$ can be **significantly different** from the true expectation value! 👎

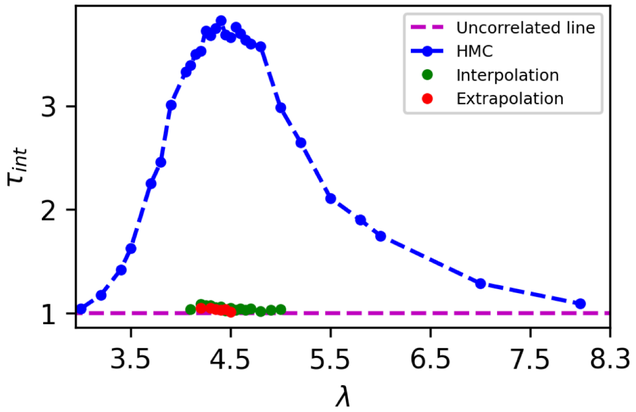
Can We Do Better?

Bootstrapping



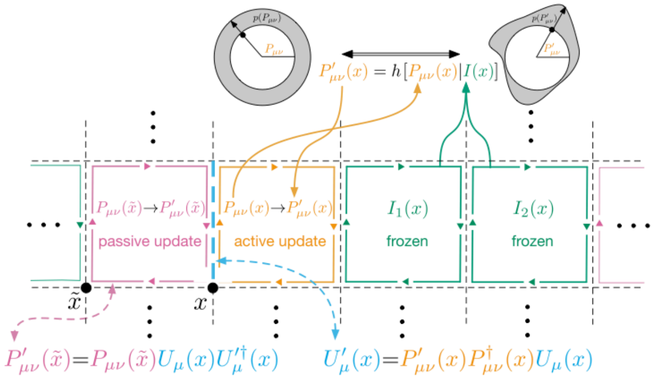
Kanwar et al., Phys. Rev. Lett. (2020)
Gerdes et al., arXiv (2022)

Conditioning



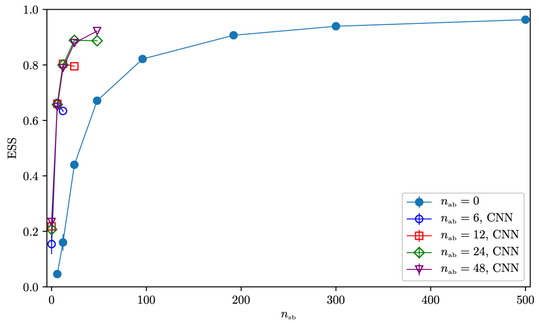
Singha et al., Phys. Rev. D (2023)
Gerdes et al., arXiv (2022)

Symmetries



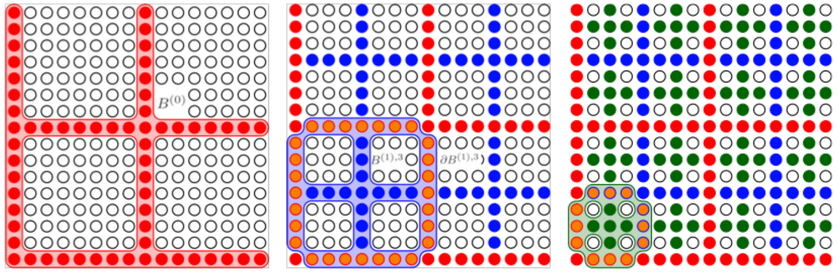
Kanwar et al., Phys. Rev. Lett. (2020)
De Haan et al., arXiv (2021)

Annealing Schemes and SNF



Caselle et al., JHEP. (2022)
Caselle et al., LATTICE22 (2022)
Matthews et al., ICML (2022) ~ **CRAFT**
Wu et al., (2022) ~ **SNF**
Midgley et al., ICLR (2023) ~ **FAB**

Hierarchical scaling



Bialas et al., CPC (2022)

and much more…

Estimation of Thermodynamic Observables

	Training	**Sampling***
FLOW	20+ hrs	<1 min
HMC	None	~1.8 hrs x 14 = 25 hrs

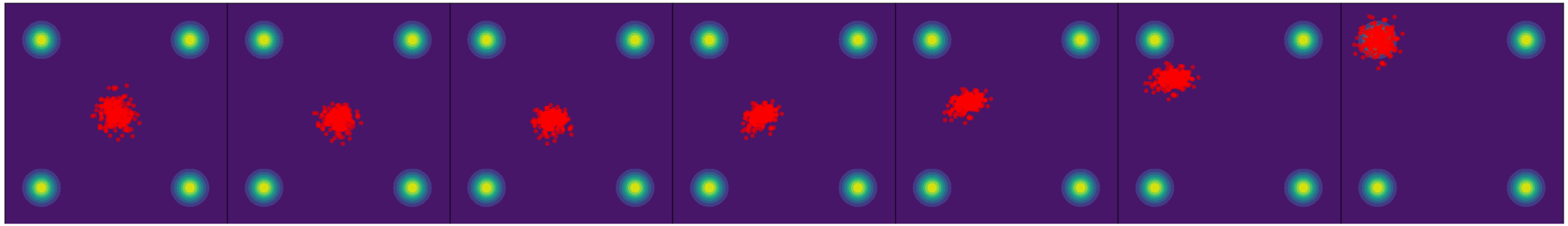
***sampling refers to draw the same number of 5.6 million configurations**

⚠ Variance of the NIS estimator scales with $O(N^{-1})$ ⟶ The flow is a much more efficient sampler!

Outlook Overview

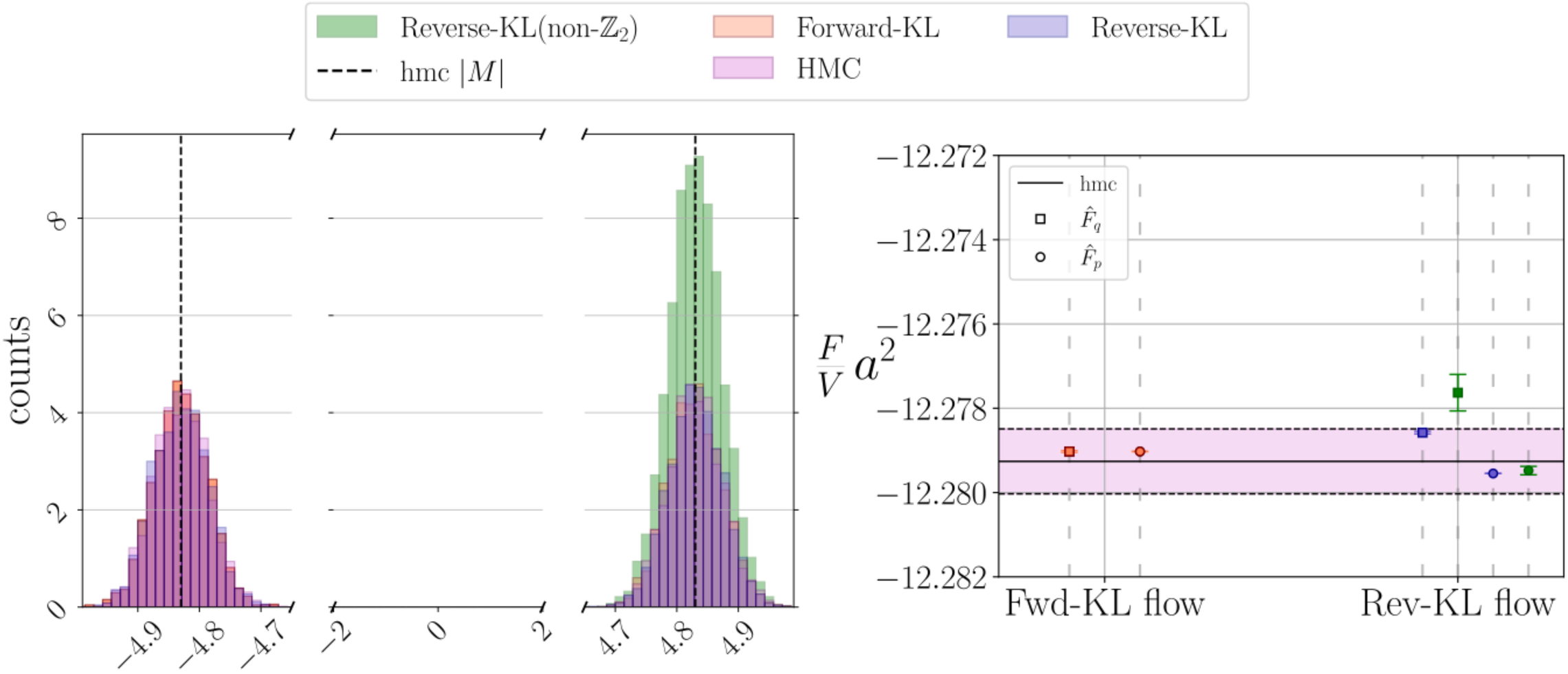
- Scaling to larger volumes (Abbott et al., 2022).
- Leveraging better training schemes (i.e. different divergences and modified objectives).
- Develop physics informed coupling layers to leverage known symmetries.
- Use approximated density to study the change of topology in the presence of a phase transition.

Multimodality Makes Things Harder

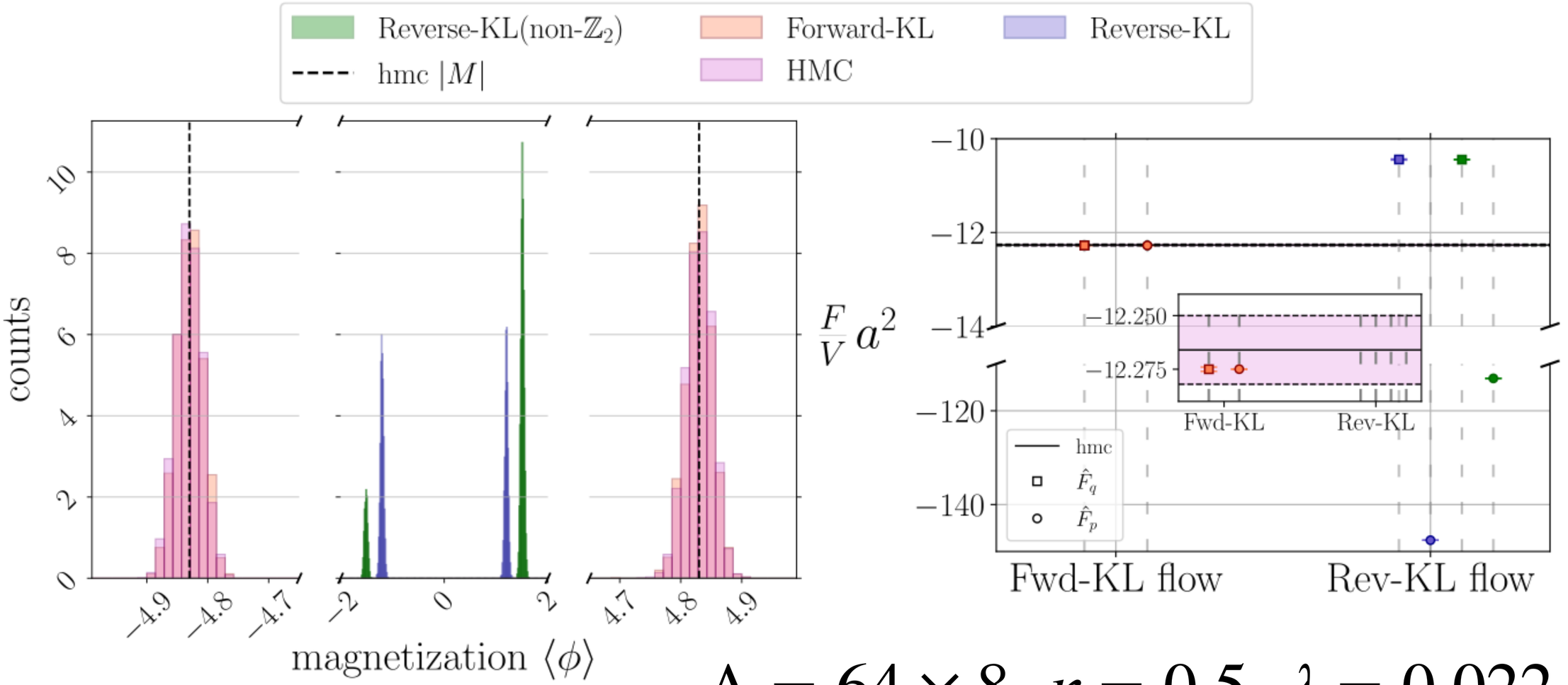


Maté B and Fleuret F., arXiv:2301.07388 (2023)

Intentionally Mode Dropping Flow

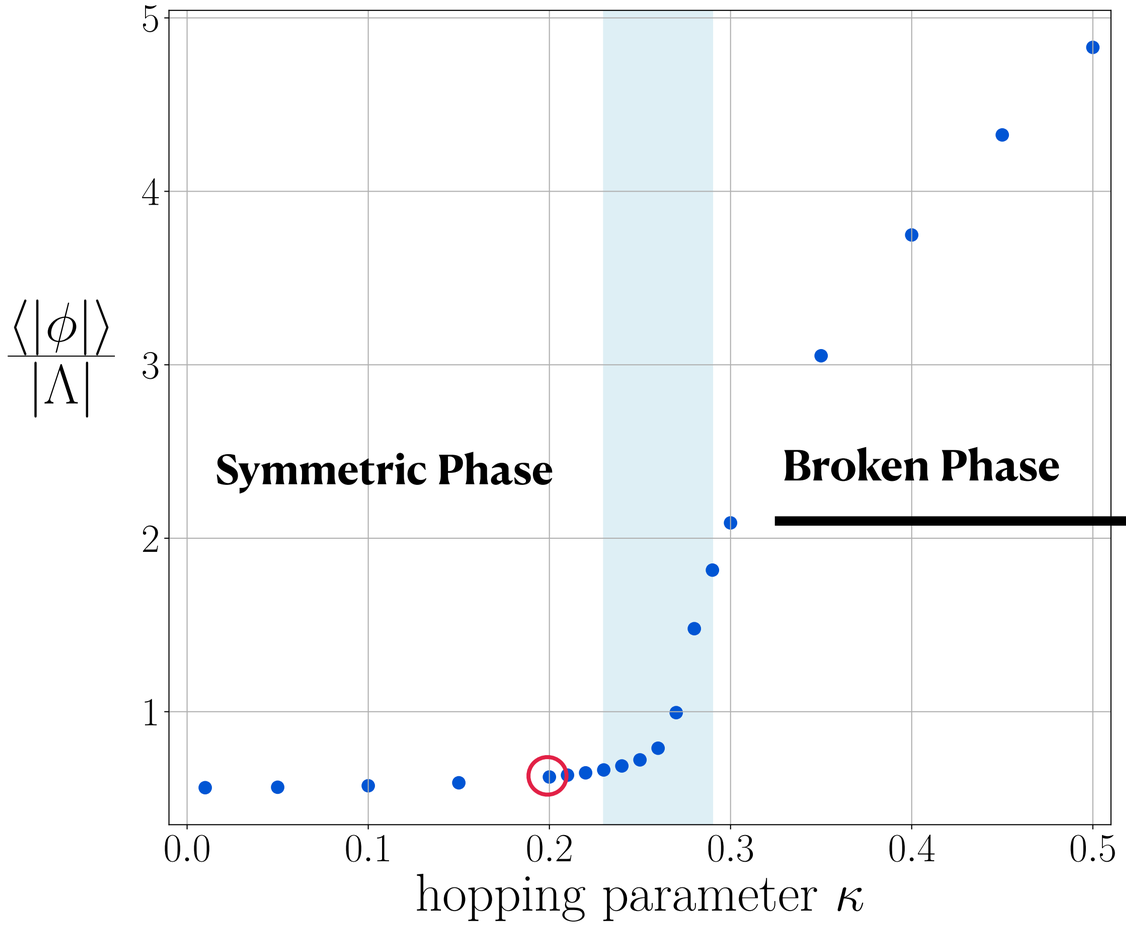


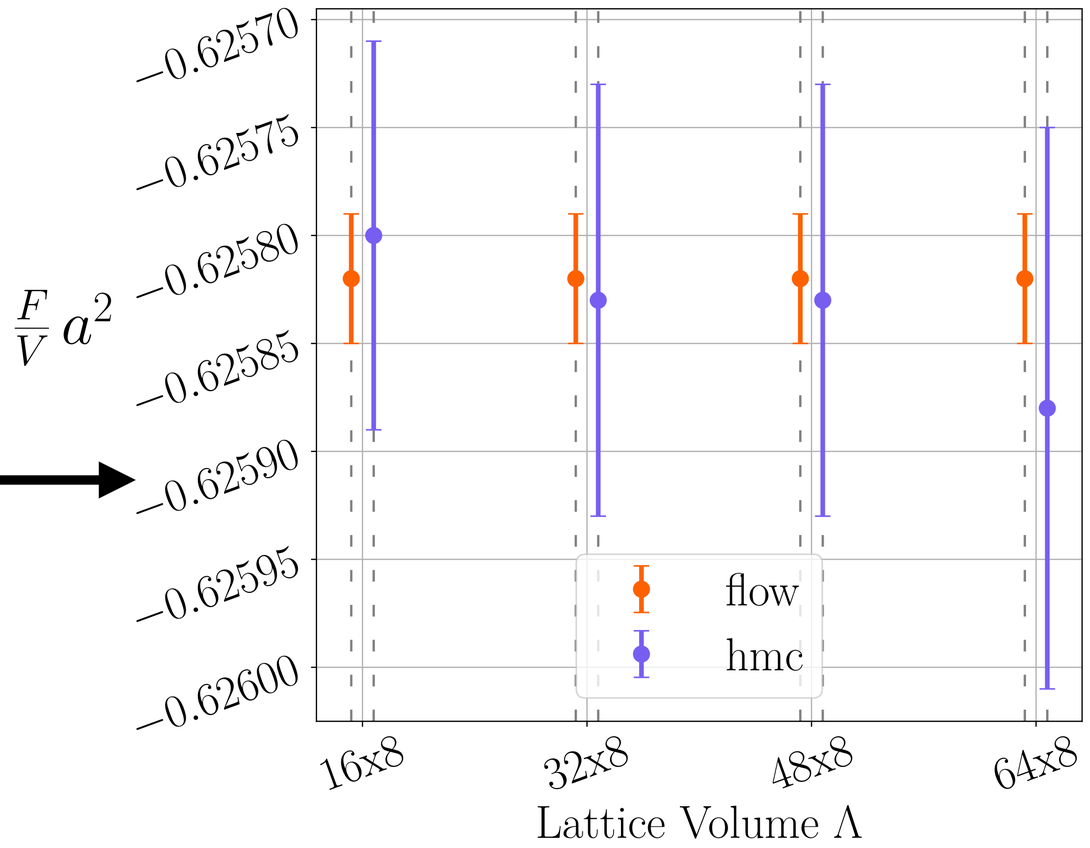
Bounds on the Free Energy Estimator



$$\Lambda = 64 \times 8,\ \kappa = 0.5,\ \lambda = 0.022$$

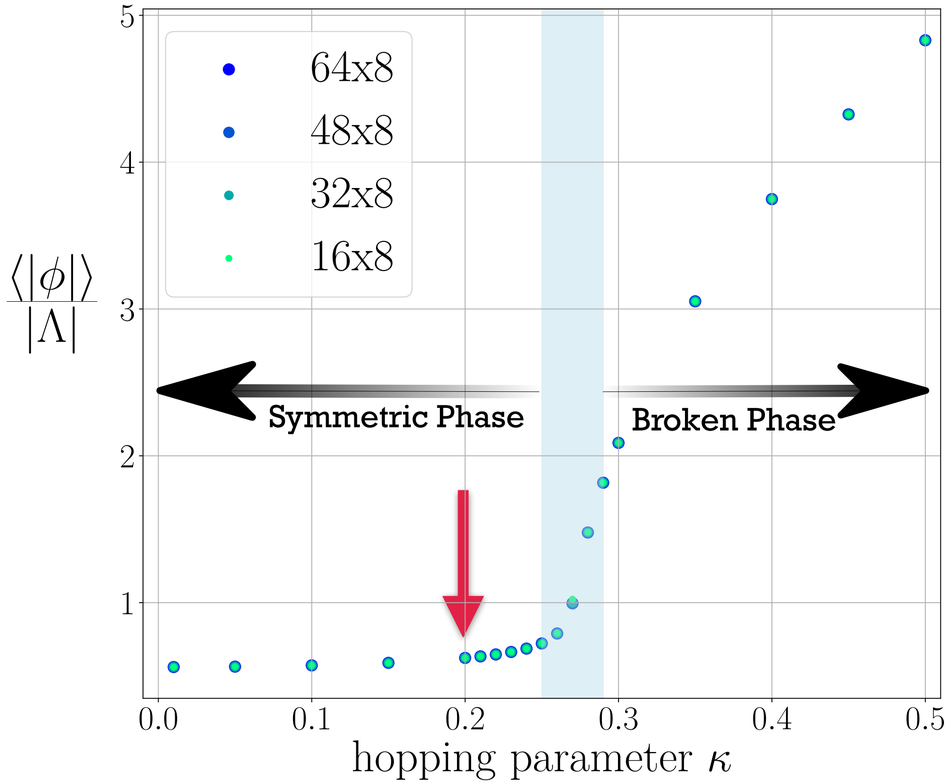
Estimation of Thermodynamic Observables

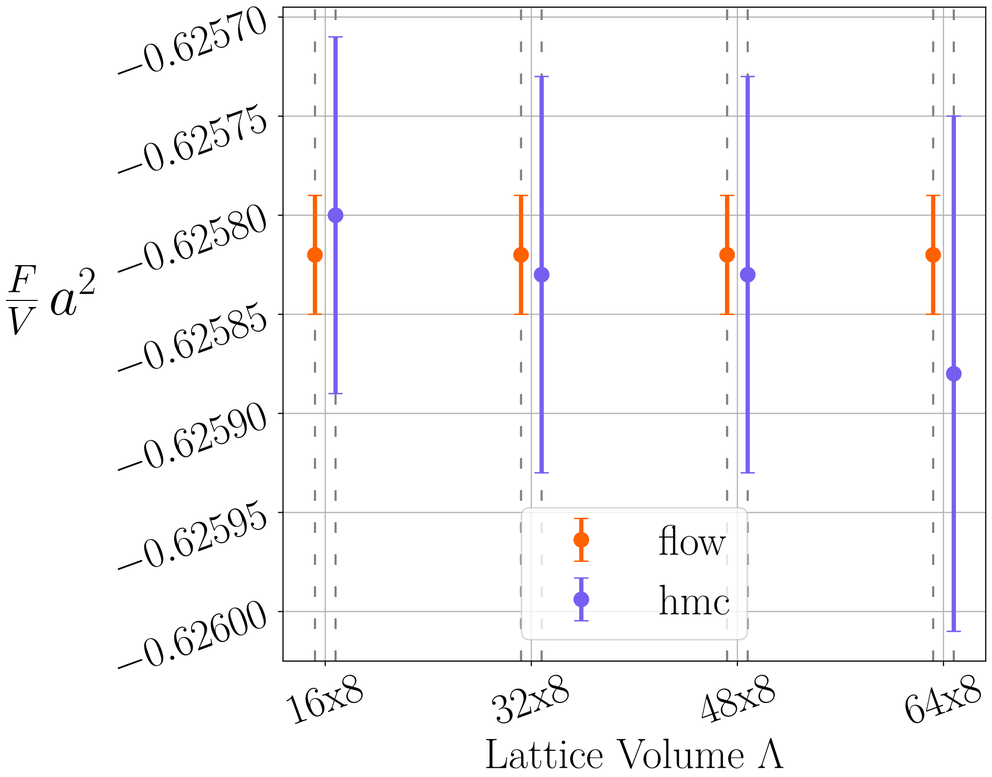




NKA, Anders, Funcke, Hartung, Jansen, Kessel, Nakajima, Stornati, Phys. Rev. Lett. (2021)

Estimation of Thermodynamic Observables

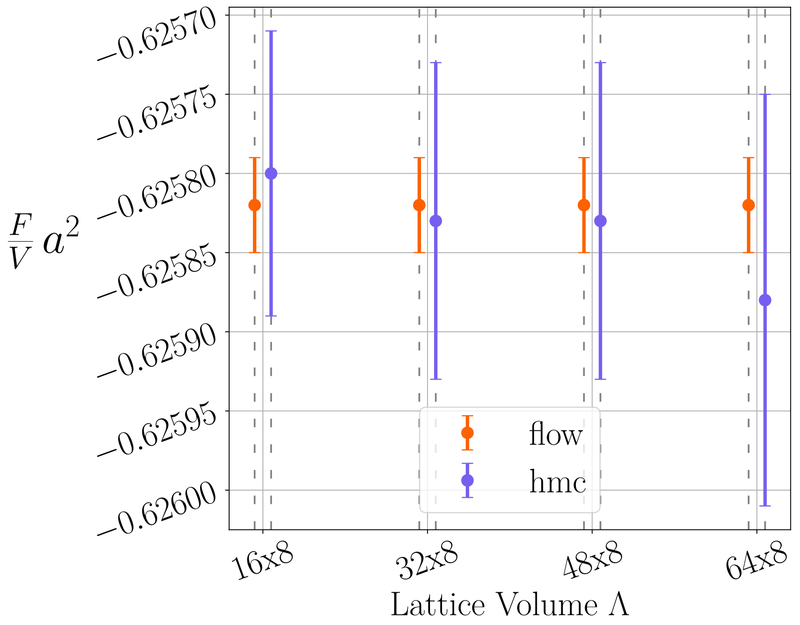




- $\lambda = 0.022, \kappa = 0.2$
- HMC: 4×10^5 samples per chain (4 chains)
- NIS: same number of samples

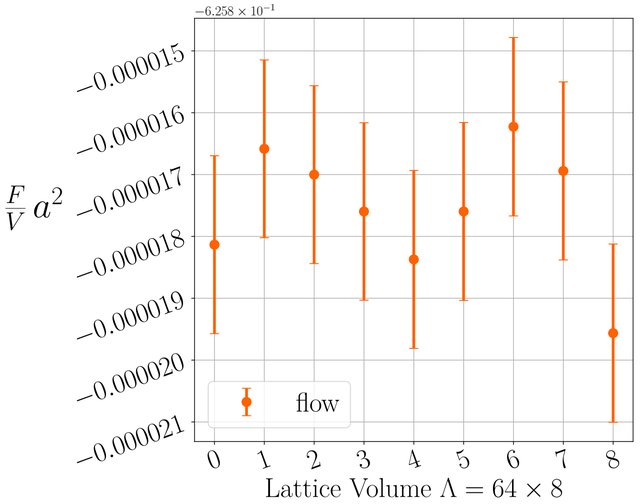
KAN, Anders, Funcke, Hartung, Jansen, Kessel, Nakajima, Stornati, Phys. Rev. Lett. (2021)

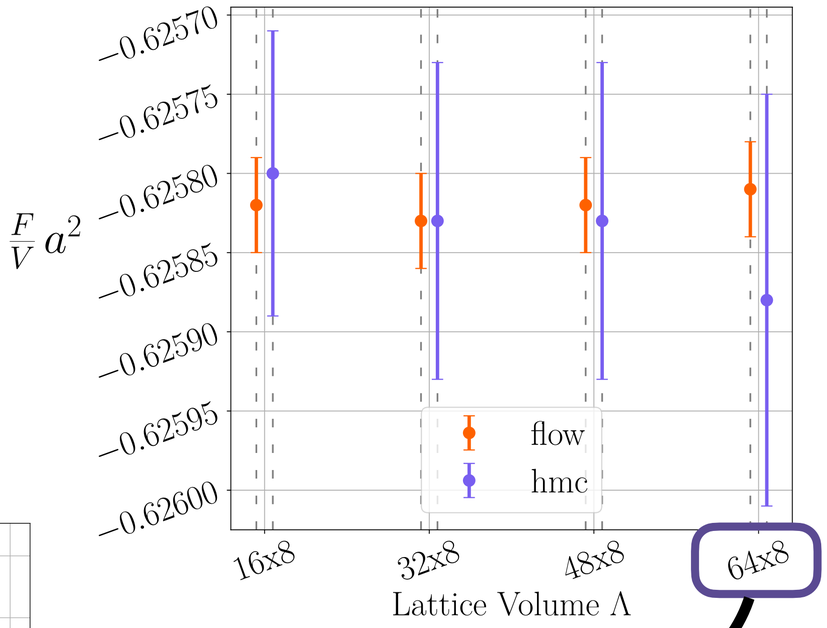
Fluctuations of the Estimator for $\kappa = 0.2$



- $\lambda = 0.022, \kappa = 0.2$
- HMC: 4×10^5 samples per chain (4 chains)
- NIS: same number of samples (1.6 M)

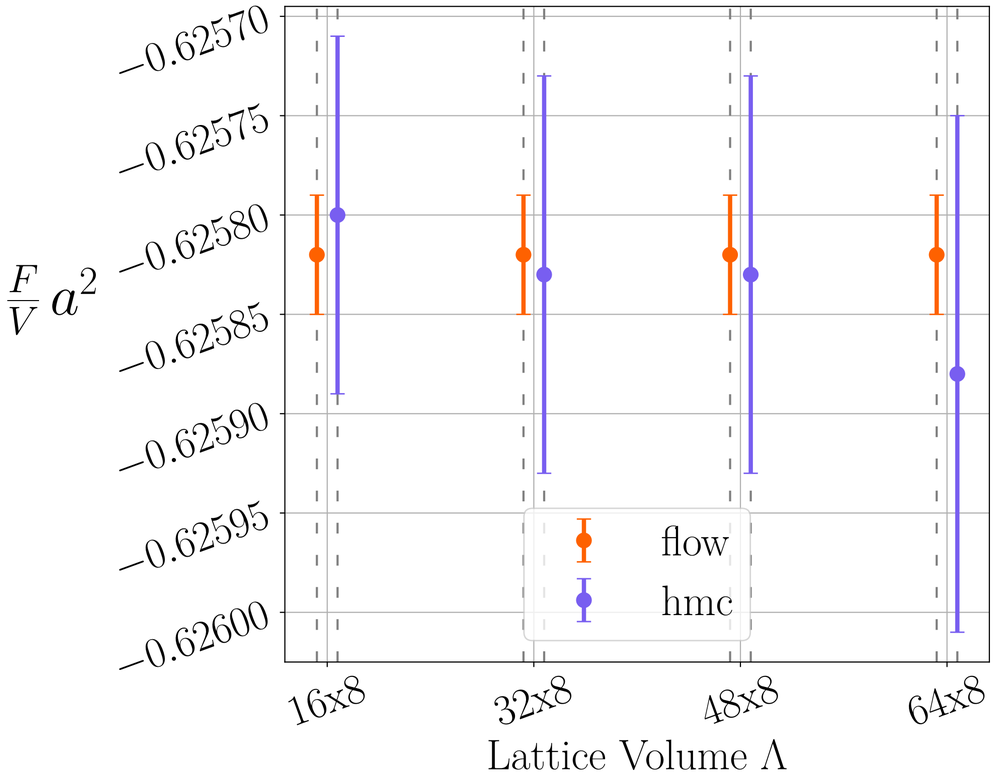
Repeating the experiment slightly changes the points (is a random process).





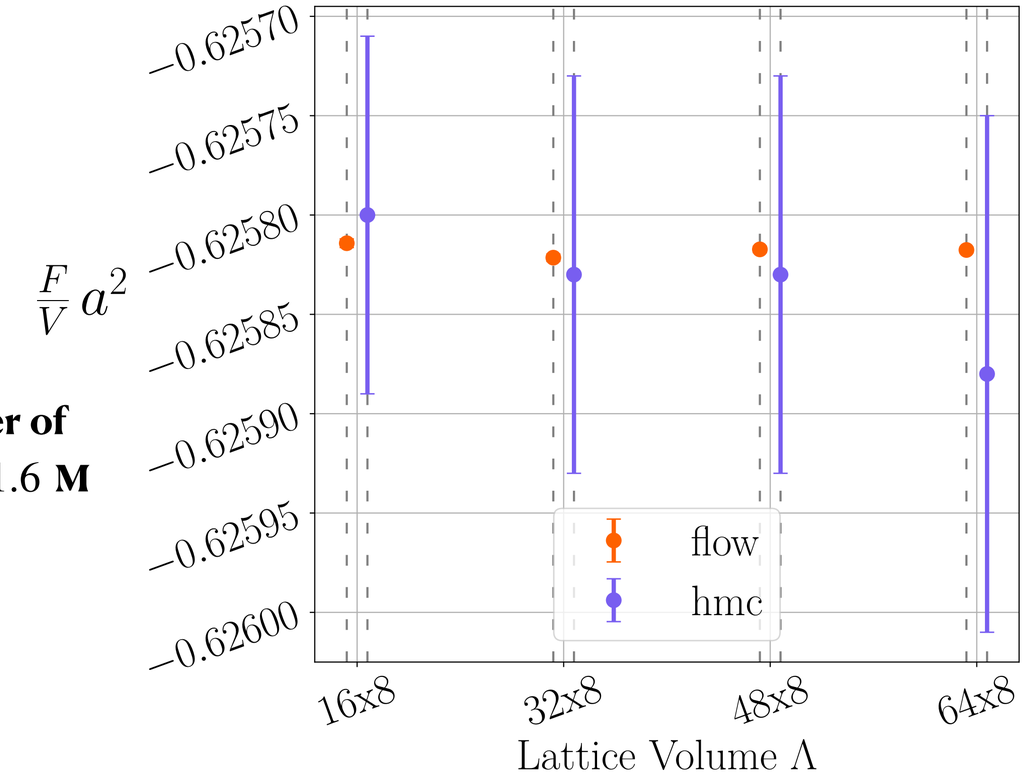
Confirmed by fixing the volume and repeating the experiment 9 times. The mean fluctuates leading to statistically consistent results.

Better Model, Smaller Uncertainties



Model trained for 300k steps

Same number of samples $N = 1.6$ M



Model trained for 700k steps

The ϕ^4-theory explained

$$S(\varphi) = \sum_{x \in \Lambda} a^2 \frac{1}{2} \sum_{\hat{\mu}=1}^{2} \frac{(\varphi(x + a\hat{\mu}) - \varphi(x))^2}{a^2} + \frac{m_0^2}{2}\varphi^2(x) + \frac{g_0}{4!}\varphi^4(x)$$

$$\varphi = (2\kappa)^{\frac{1}{2}}\phi$$

$$(am_0)^2 = \frac{1 - 2\lambda}{\kappa} - 4$$

$$a^2 g_0 = \frac{6\lambda}{\kappa^2}$$

$$S(\phi) = \sum_{x \in \Lambda} - 2\kappa \sum_{\hat{\mu}=1}^{2} \phi(x)\phi(x + \hat{\mu}) + (1 - 2\lambda)\phi(x)^2 + \lambda\,\phi(x)^4$$

The ϕ^4-theory explained

$m_0^2 = -m^2 + T^2$ is the **effective mass**.

$m^2 \ll T^2$

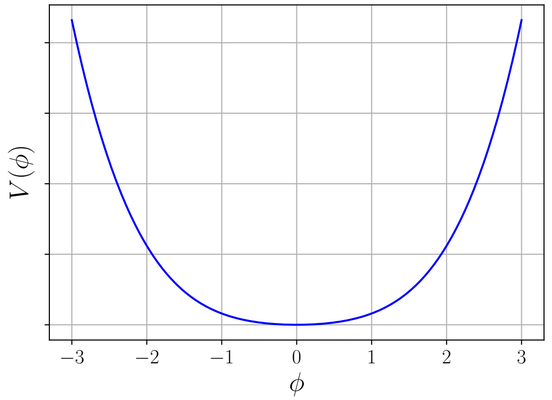
Early Universe

The temperature term dominates and the effective mass ends up **positive**

$m^2 \gg T^2$

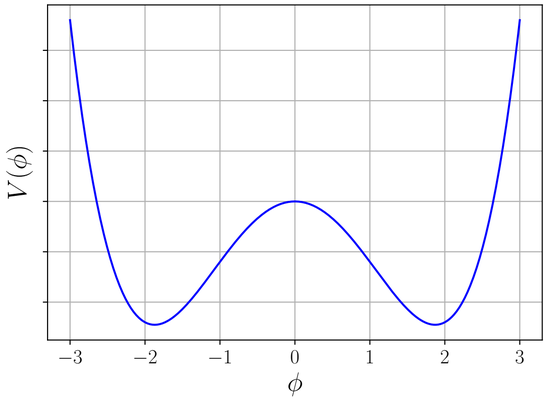
Our Universe

The mass term dominates and the effective mass ends up **negative**



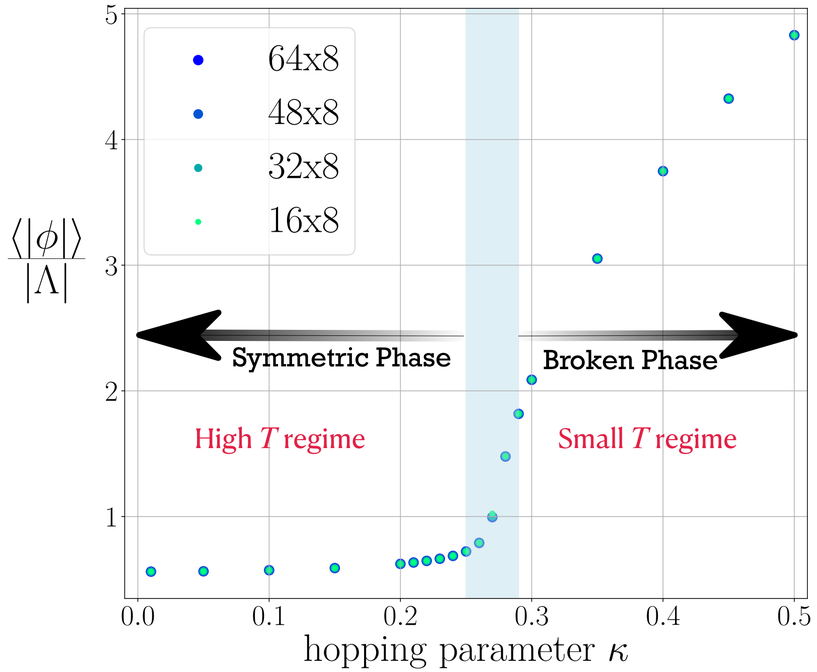
Spontaneous Symmetry Breaking

Temperature **Decreases**



The ϕ^4-theory explained

$$(am_0)^2 = \frac{1-2\lambda}{\kappa} - 4$$



$\kappa > \kappa_c$ The **second term** dominates and the reparametrized effective mass is **negative**

$\kappa < \kappa_c$ The **first term** dominates and the reparametrized effective mass is **positive**

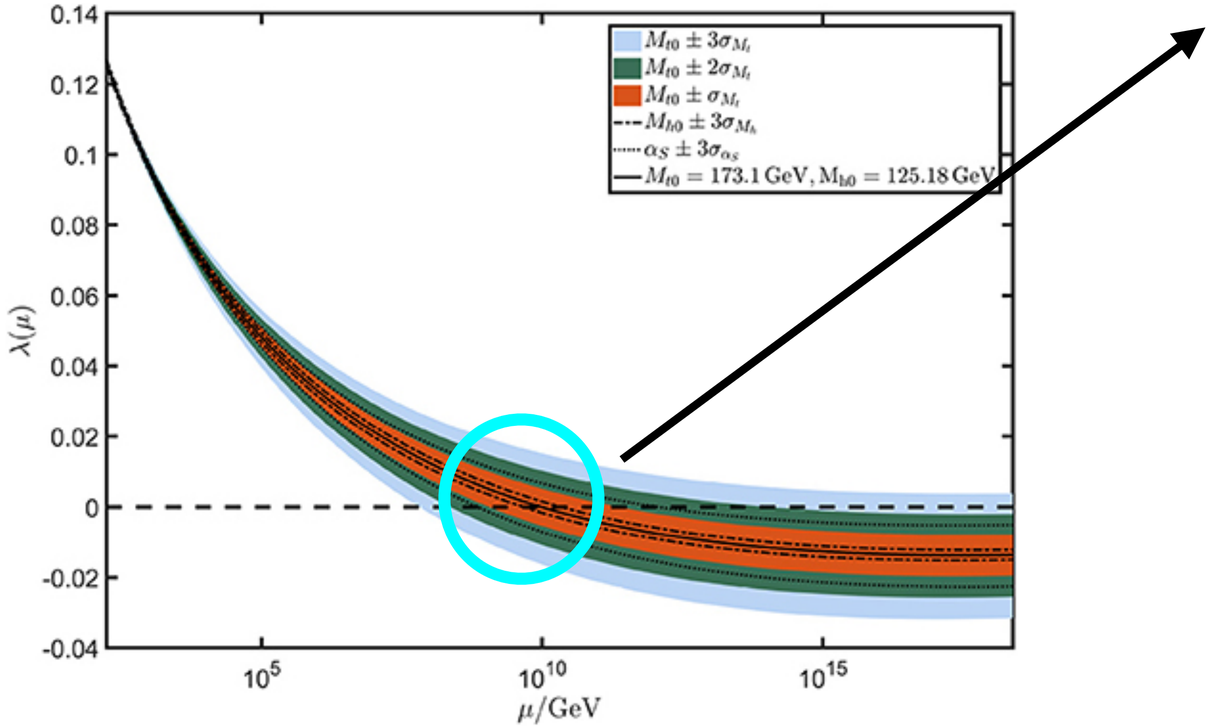
Lower κ $\Longrightarrow$ higher T

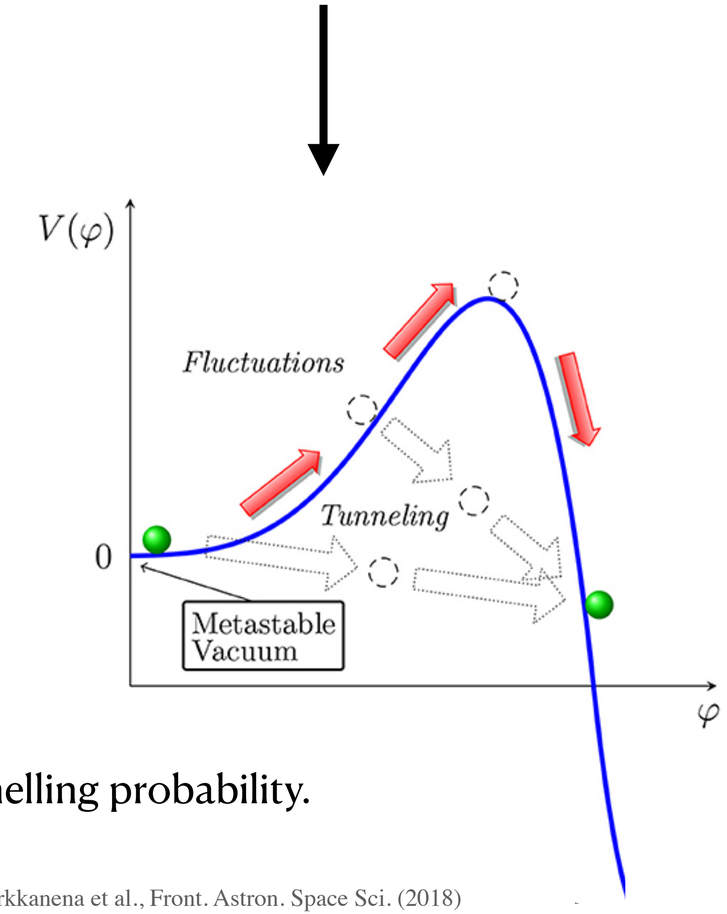
Cool down of the universe leads to **spontaneous symmetry breaking in the Higgs**

Higher κ $\Longrightarrow$ lower T

Triviality of ϕ^4-theory

Turning point of $\lambda(\mu)$ becoming negative as the energy increases.





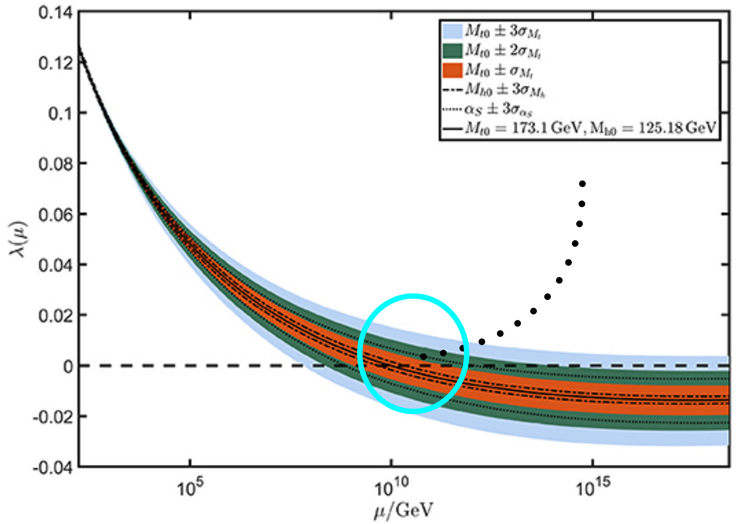
- In the Lagrangian this **flips the potential** thus making a non vanishing tunnelling probability.
- We live in a **metastable state of the vacuum**.

Markkanena et al., Front. Astron. Space Sci. (2018)

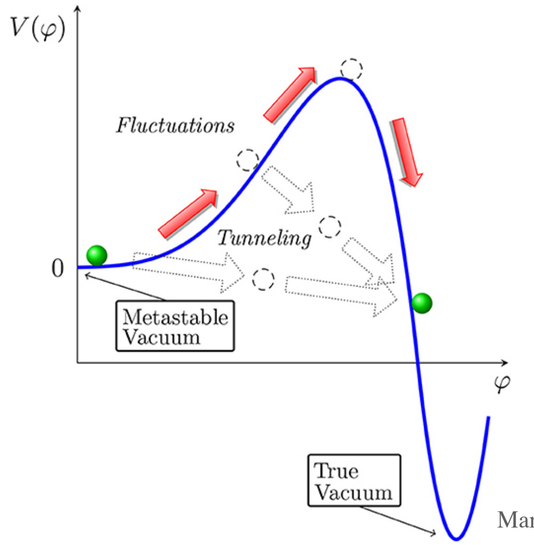
Triviality of ϕ^4-theory

The theory is **well defined** as long as it is bounded from below. When λ becomes negative this is not true anymore.

Two directions to argue against the possibility of facing the end of the universe:



- As soon as the λ reaches its zero value the theory becomes **ill-defined**.
- At this point is reached, the theory becomes **non interacting** (e.g. a free theory) and therefore is **trivialised**.
- Theories beyond the standard should prevent $\lambda < 0$ while $\lambda = 0$ is a measure zero point



Markkanena et al., Front. Astron. Space Sci. (2018)

- There should be **physics beyond the standard model** that prevents the potential to go to -infinity (e.g. unknown particles.)
- Even though we might tunnel through the metastable state there's a **true vacuum** at a **lower energy** introduced by this new physics.

Annealed Importance Sampling

Similar lines of work leveraging Annealed Importance Sampling

- SNF (Wu et al.)
- CRAFT (Matthews et al.)
- **FAB (Midgley et al.)**

} Flows + stochastic sampling steps

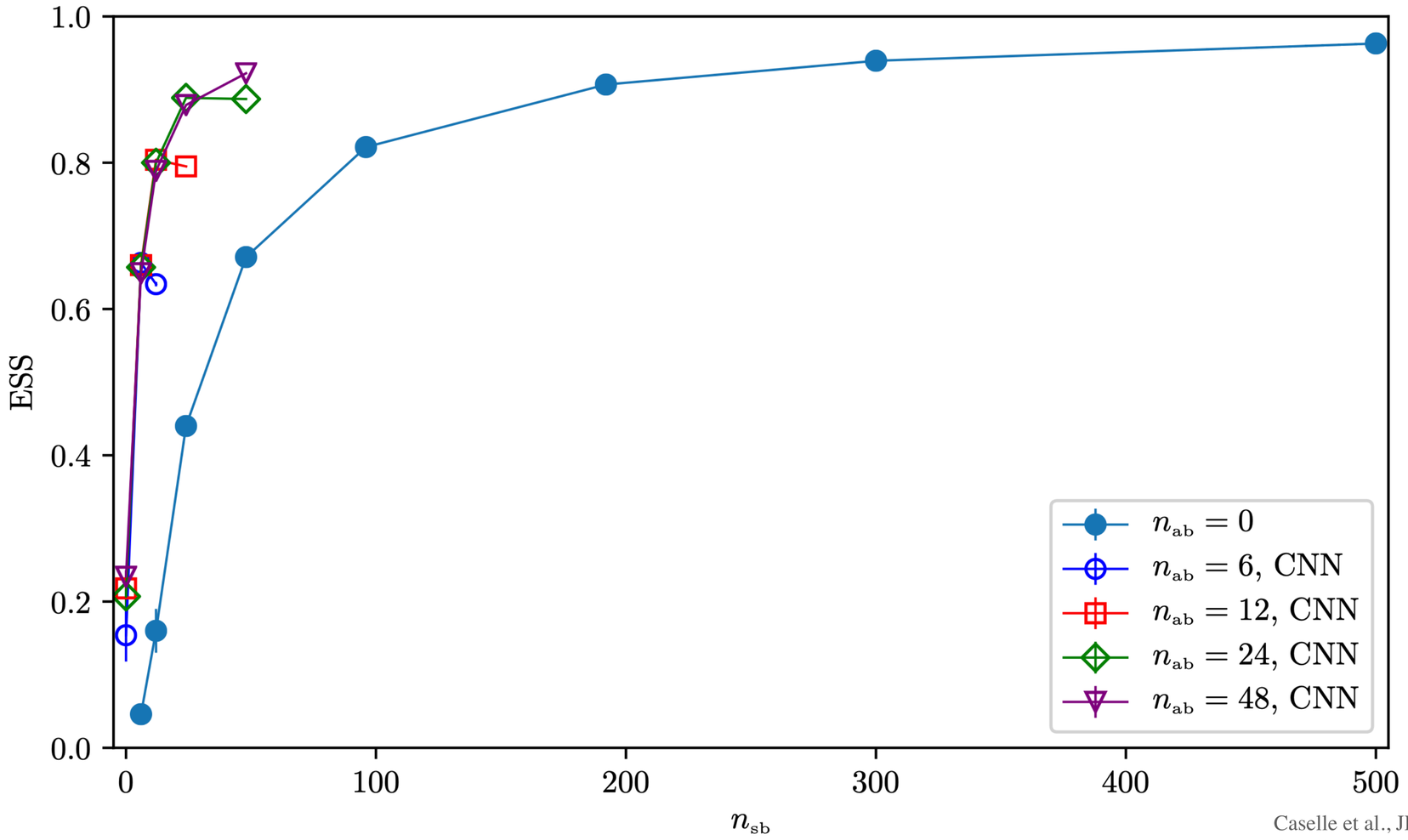
$$D_\alpha(p\|q) = -\frac{\int_x p(\mathbf{x})^\alpha q(\mathbf{x})^{1-\alpha} \mathrm{d}\mathbf{x}}{\alpha(1-\alpha)}$$

Minimising α-divergence for $\alpha = 2$ corresponds to minimise the variance of importance weights

$$D_{\alpha=2}(p\|q_\theta) \propto \mathcal{L}(\theta) = \int \frac{p(\mathbf{x})^2}{q_\theta(\mathbf{x})} \mathrm{d}\mathbf{x} = \mathrm{E}_{g(\mathbf{x})}\left[\frac{p(\mathbf{x})^2}{q_\theta(\mathbf{x}) g(\mathbf{x})}\right]$$

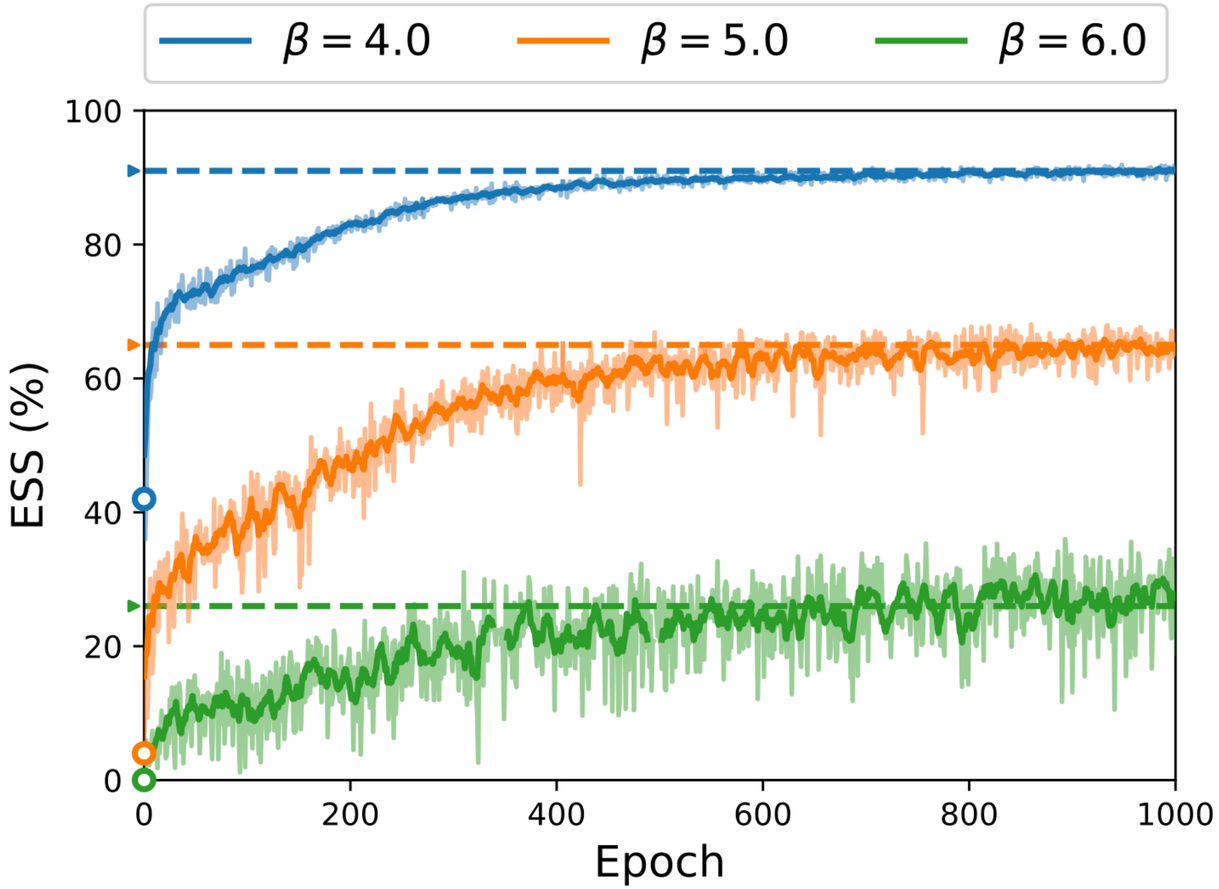
Target $\propto p^2/q_\theta$

SNFs for Lattice Field Theory



Caselle et al., JHEP. (2022)
Caselle et al., LATTICE22 (2022)

Learning Trivialising Maps



Ref.		N_{params}	ESS at β 4.0	5.0	6.0
Lüscher, NL [3]		8 non-zero values	42%	4%	<1%
This work	**A**	$14 \equiv 2_t \times 7_W$	91%	65%	26%
	B	$420 \equiv 10_t \times 42_W$	98%	88%	70%
Boyda et al. [8]		$\mathcal{O}(10^6)$ *estimated*	88%	75%	48%

[A, B] Bacchio et al., arXiv. (2022)

[8] Boyda et al., Phys. Rev. D (2021)

Incorporating Symmetries

- Permutations

  $p(\ldots, x_i, \ldots, x_j, \ldots) = p(\ldots, x_j, \ldots, x_i, \ldots)$ Symmetric permutation

  $p(\ldots, x_i, \ldots, x_j, \ldots) = -p(\ldots, x_j, \ldots, x_i, \ldots)$ Antisymmetric permutation

- Translations, Rotations

  Special Euclidean Group SE(3)

  Octahedral Symmetries

- Gauge Symmetries

  Special Unitary Group SU(N)

  Unitary Group U(N)

$$(\Omega \cdot U)_\mu(x) = \Omega(x) U_\mu(x) \Omega^\dagger(x + \hat{\mu})$$

Incorporating Symmetries

Knowing the symmetries of the system one can build a flow which is **invariant** or **equivariant** under some group action.

Gauge Equivariant Flow

$$Y_\mu(x) = f(U_\mu(x); \theta)$$

$$U_\mu(x) \to \Omega(x) U_\mu(x) \Omega(x+\hat{\mu})^\dagger$$

$$Y_\mu(x) \to \Omega(x) Y_\mu(x) \Omega(x+\hat{\mu})^\dagger$$

Kanwar et al., Phys. Rev. Lett. (2020)
Boyda et al., Phys. Rev. D (2021)

$\mathbb{Z}_2$ Invariant Flow

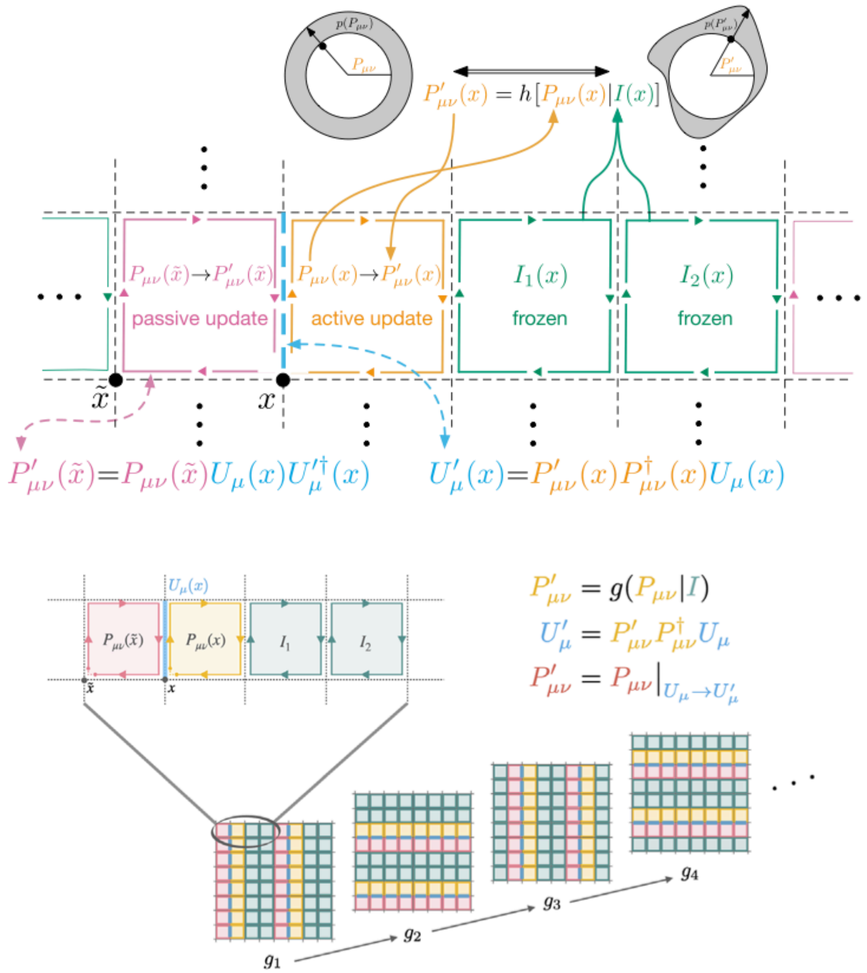
$$S(\phi) = S(-\phi)$$

$$\downarrow$$

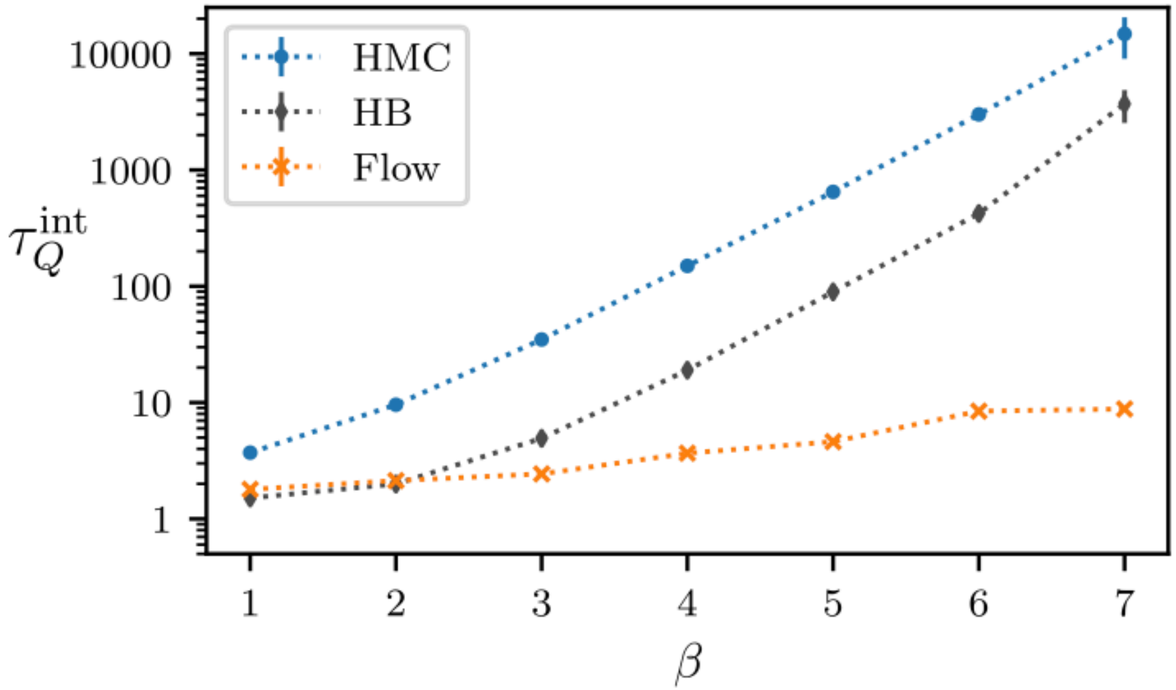
$$S(f(z;\theta)) = S(f(-z;\theta))$$

Nicoli et al., Phys. Rev. Lett. (2021)

Gauge-Equivariant Flows for LGT

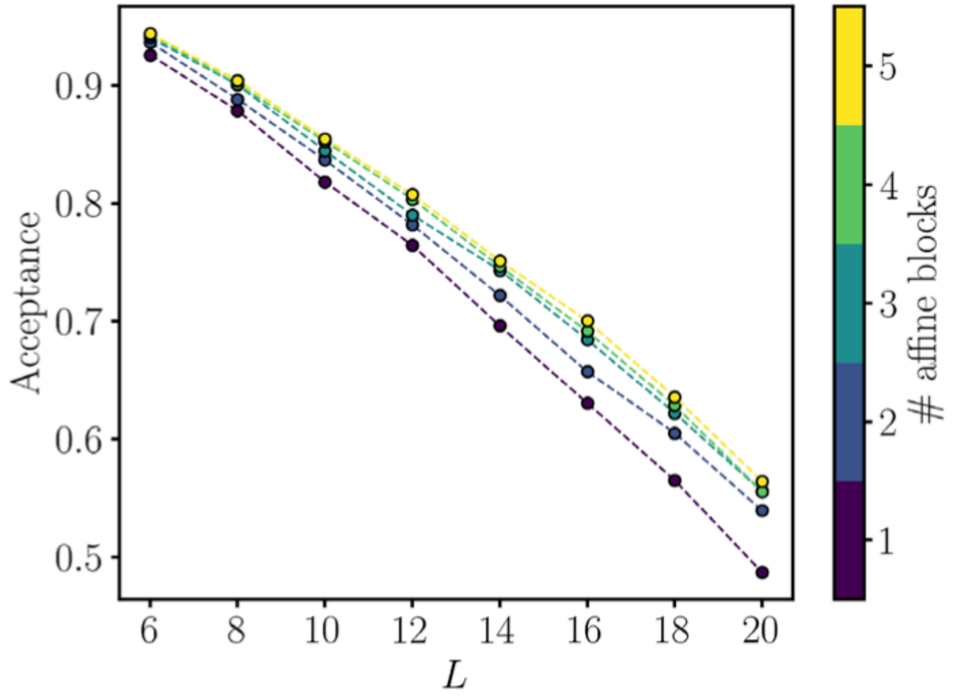


$U(1)$ Gauge Theory on 16×16 lattice

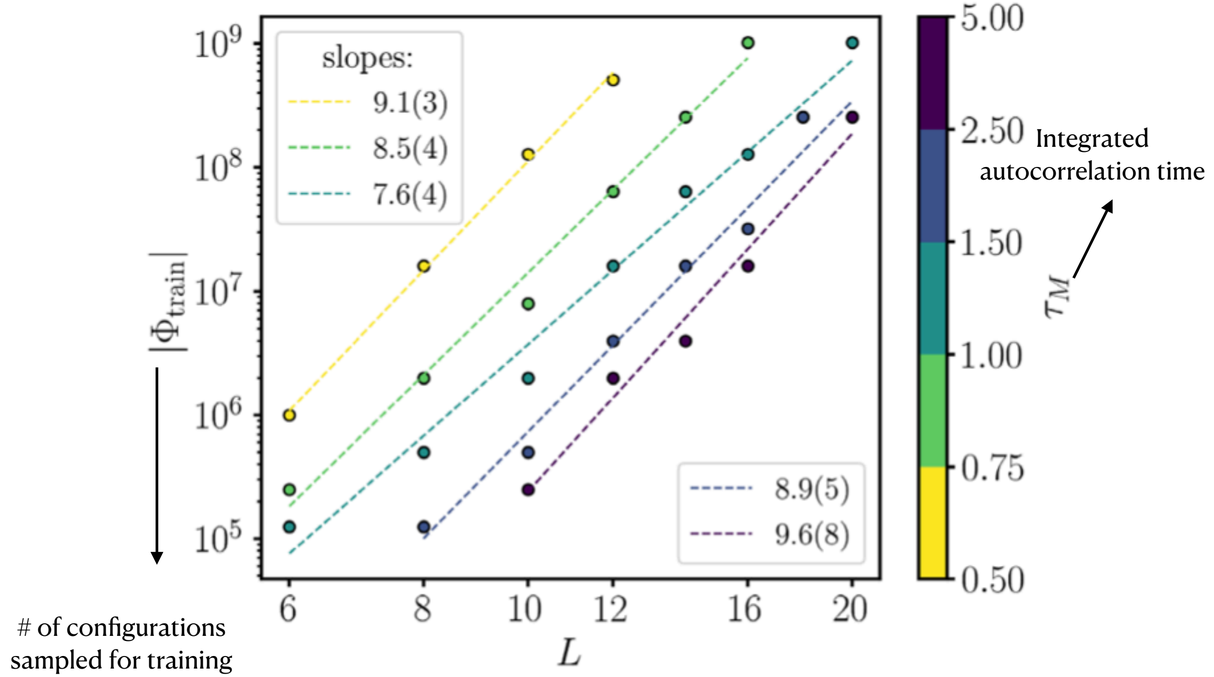


Kanwar et al., Phys. Rev. Lett. (2020)
Boyda et al., Phys. Rev. D (2021)

Scaling to Larger Lattices



Acceptance rate drastically decays as the lattice shape increases



Exponential increase in the training cost as the lattice shape increases

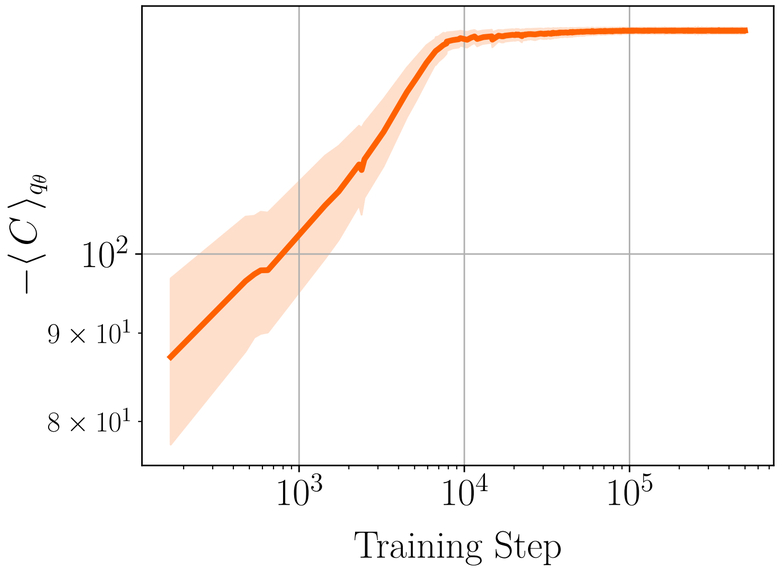
Del Debbio et al., Phys. Rev. D (2021)

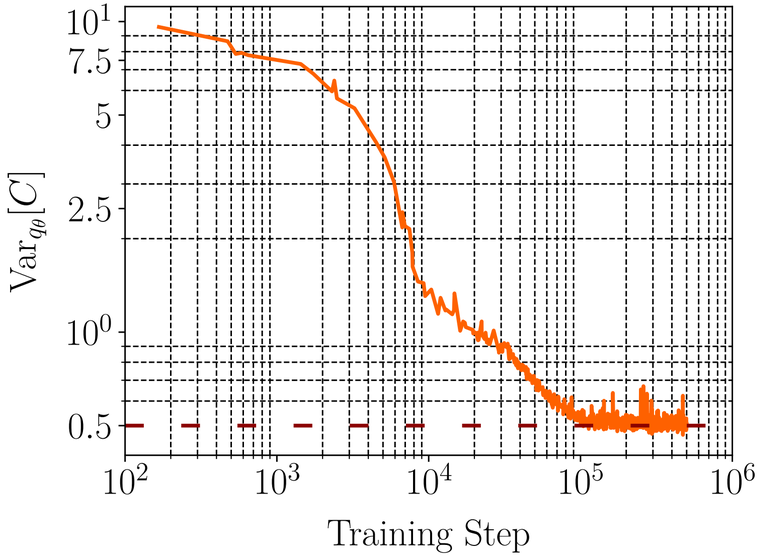
KL is Related to the Variance of Importance Weights

Theorem *Let $C(\phi) = S(\phi) + \ln q_\theta(\phi)$. The following relation between the KL divergence and the variance of the random variable C holds:*

$$KL(q_\theta \,||\, p) = \tfrac{1}{2} \operatorname{Var}_{q_\theta}[C] + \mathcal{O}\left(\mathbb{E}_{q_\theta}\left[|w-1|^3\right]\right),$$

where $w(\phi) = \frac{p(\phi)}{q_\theta(\phi)}$ is the normalized importance weight.





Relation to Gibbs Inequality

The **Gibbs Inequality** states that given a (discrete) probability distribution $P = \{p_1, \ldots, p_N\}$, then for any other probability distribution $Q = \{q_1, \ldots, q_N\}$ the following inequality holds

$$-\sum_{i=1}^{n} p_i \ln p_i \leq -\sum_{i=1}^{n} p_i \ln q_i$$

Which implies that **the information entropy of a distribution P is always smaller or equal to its cross-entropy with respect to another distribution Q.**

Rearranging this inequality

$$+\sum_{i=1}^{n} p_i \ln p_i - \sum_{i=1}^{n} p_i \ln q_i \leq 0 \implies \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} = \mathrm{KL}(P\,||\,Q) \geq 0$$

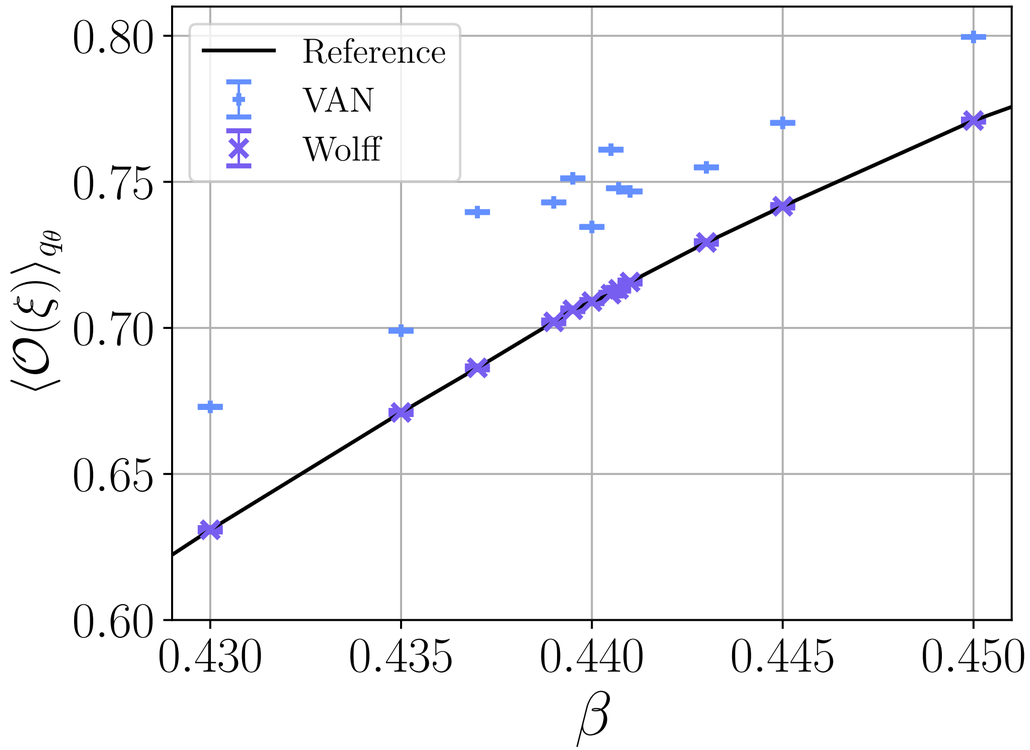
Which tells us that the Kullback-Leibler divergence is always **positive**!

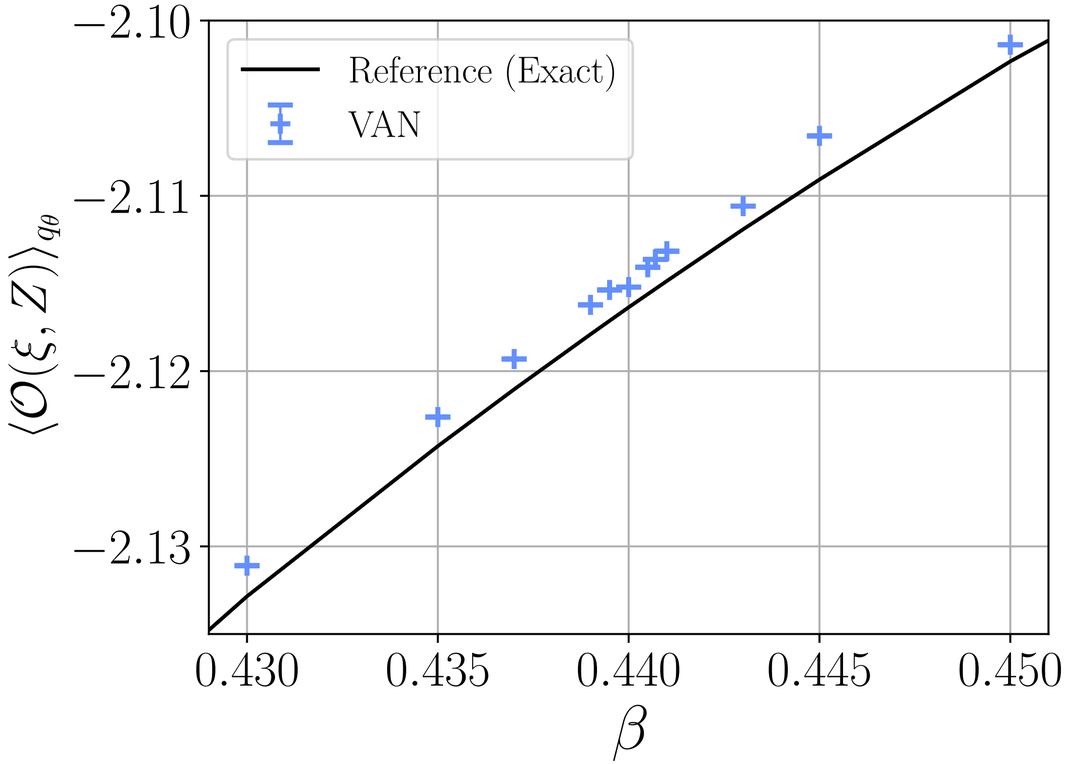
$$KL(q_\theta\,||\,p) = \beta(F_q - F) \geq 0 \implies \text{One can minimise } F_q \text{ instead of } F_q - F$$

Application: the 2D Ising Model

$$H(\xi) = -J \sum_{<i,j>} \xi_i \xi_j \qquad p(\xi) = \frac{e^{-\beta H(\xi)}}{Z}$$

$$\langle \mathcal{O} \rangle_{q_\theta} \neq \langle \mathcal{O} \rangle_p$$



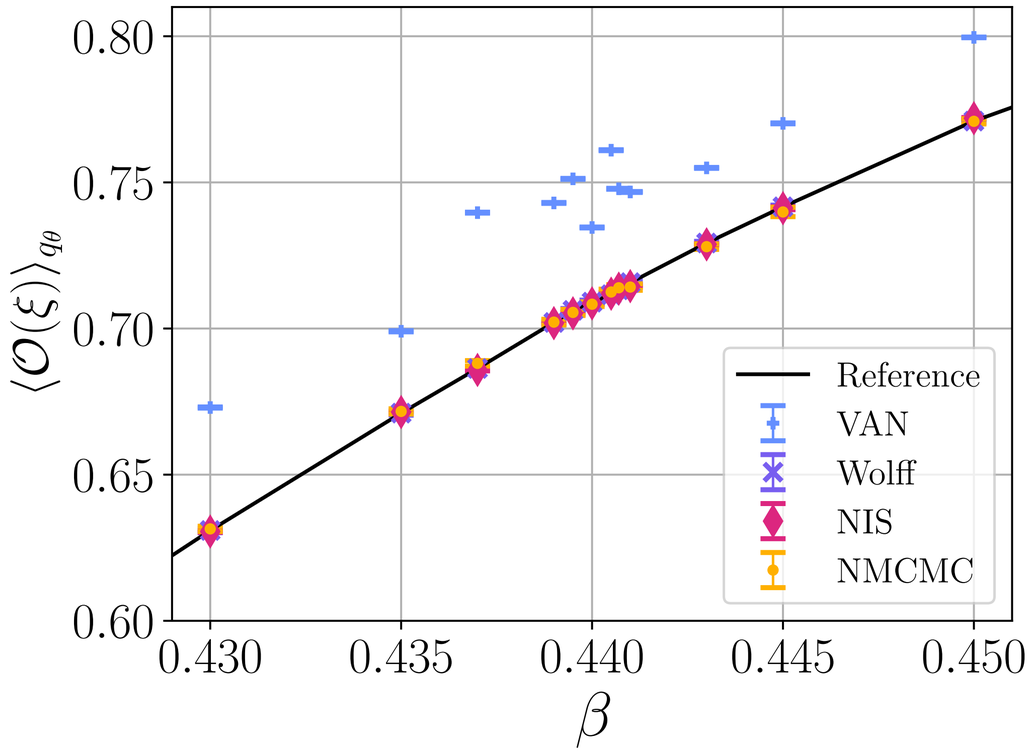


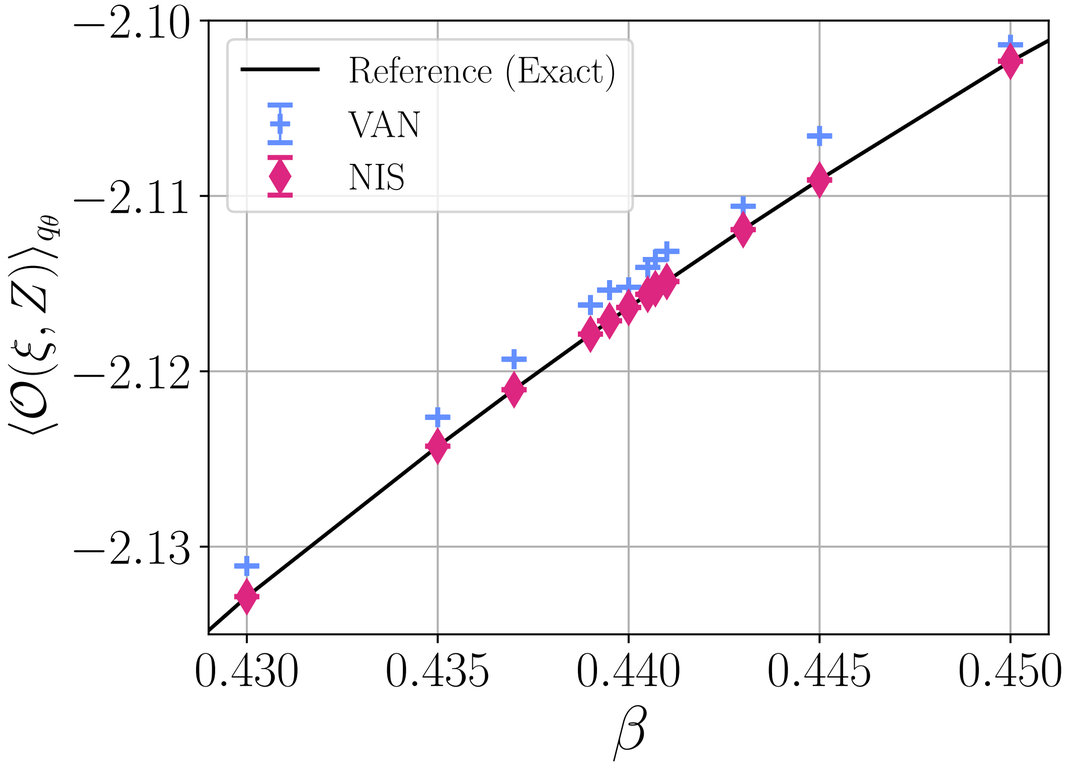
Wu, Wang, Zhang, Phys. Rev. Lett. (2019). $\Longrightarrow$ **VAN** (Variational Autoregressive Network)

Application: the 2D Ising Model

$$H(\xi) = -J \sum_{<i,j>} \xi_i \xi_j \qquad p(\xi) = \frac{e^{-\beta H(\xi)}}{Z}$$

$$\langle \mathcal{O} \rangle_{q_\theta} \longrightarrow \langle w\mathcal{O} \rangle_{q_\theta}$$



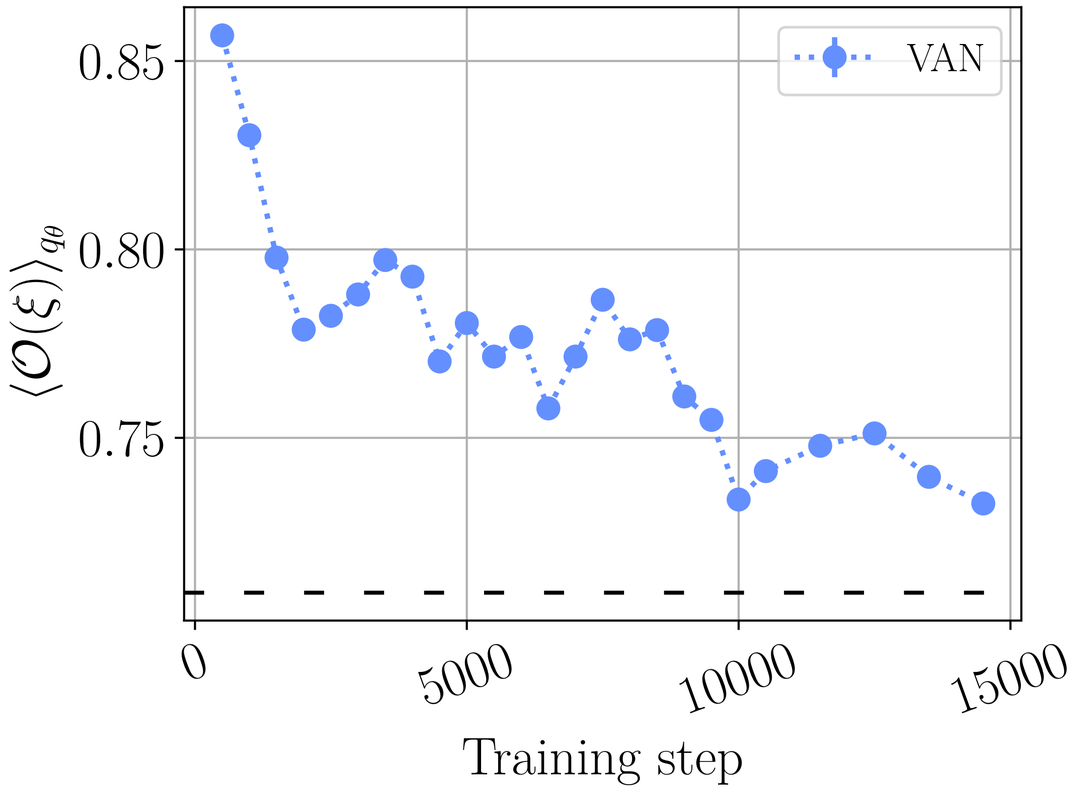


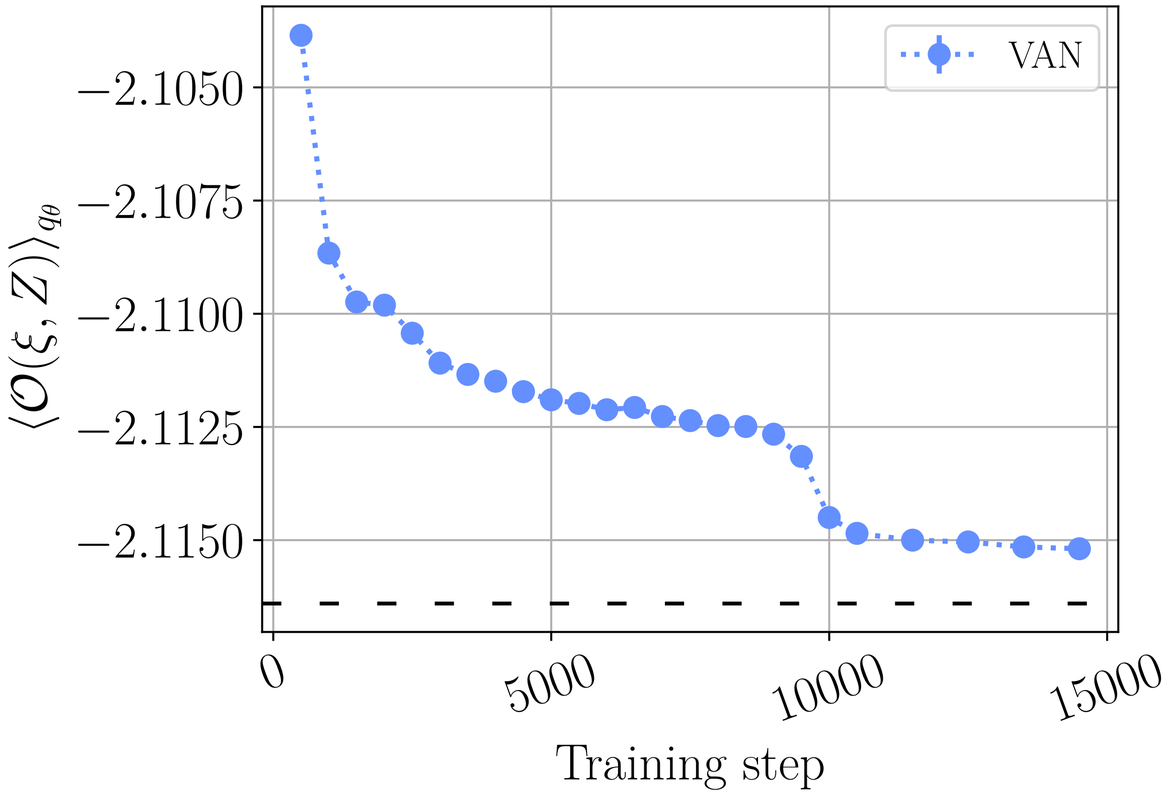
KAN, Nakajima, Strodthoff, Samek, Müller, and Kessel, Phys. Rev. E (2020).
Wu, Wang, Zhang, Phys. Rev. Lett. (2019).

Application: the 2D Ising Model

$$H(\xi) = -J \sum_{<i,j>} \xi_i \xi_j \qquad p(\xi) = \frac{e^{-\beta H(\xi)}}{Z}$$

$$\langle \mathcal{O} \rangle_{q_\theta} \neq \langle \mathcal{O} \rangle_p$$



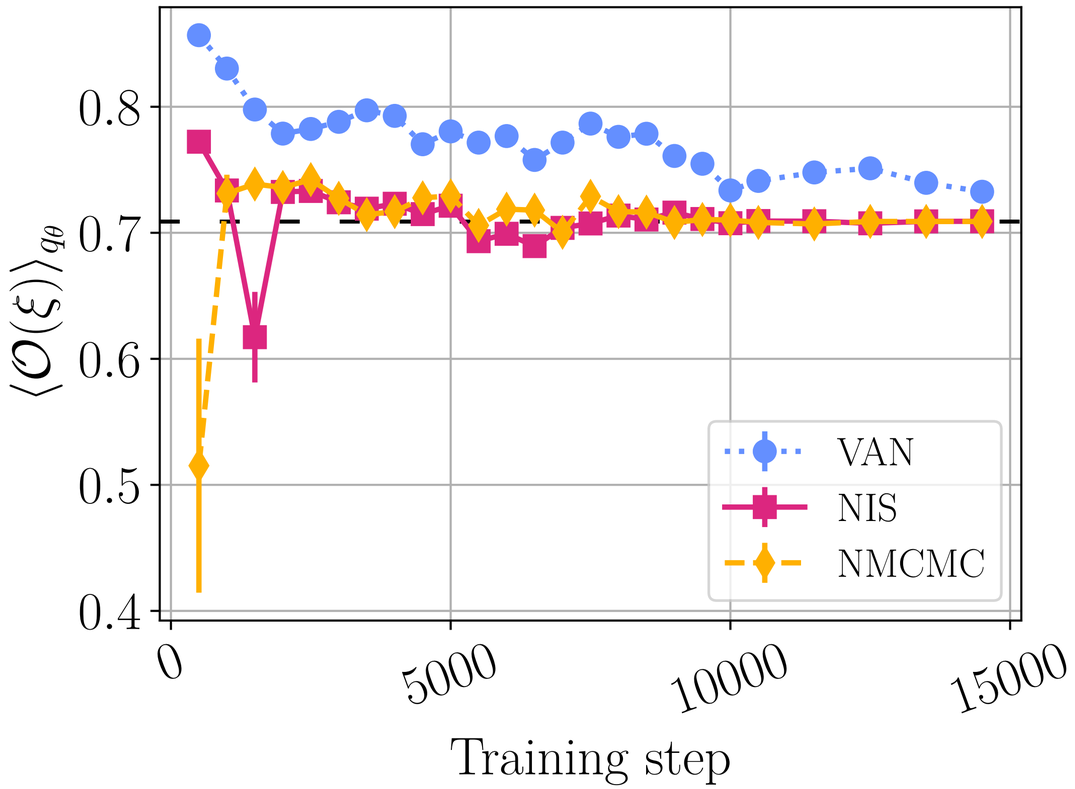


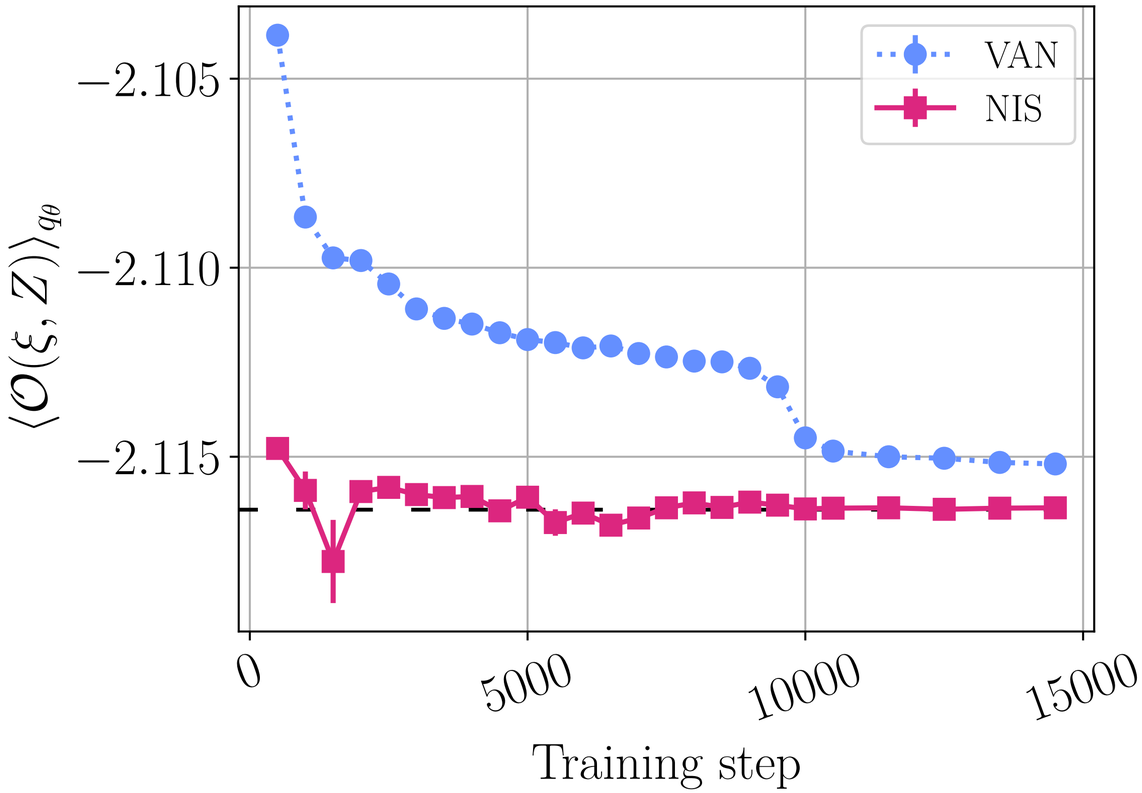
Wu, Wang, Zhang, Phys. Rev. Lett. (2019).

Application: the 2D Ising Model

$$H(\xi) = -J \sum_{<i,j>} \xi_i \xi_j \qquad p(\xi) = \frac{e^{-\beta H(\xi)}}{Z}$$

$$\langle \mathcal{O} \rangle_{q_\theta} \longrightarrow \langle w\mathcal{O} \rangle_{q_\theta}$$





KAN, Nakajima, Strodthoff, Samek, Müller, and Kessel, Phys. Rev. E (2020).
Wu, Wang, Zhang, Phys. Rev. Lett. (2019).

Delta Method of Moments

Theorem 2. *Let $\hat{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ be the sample mean of independent and identically distributed random variables X_i with $\mathbb{E}\left[X_i^{2k+2}\right] < \infty$ for $k \in \{0, 1\}$. Let h be a real-valued function with uniformly bounded derivatives. It then holds that*

$$\mathbb{E}\left[h(\hat{X}_N)\right] = c_0 + \frac{c_1}{N} + O\left(\frac{1}{N^2}\right), \tag{4.10}$$

where

$$c_0 = h(\mu), \qquad c_1 = h''(\mu)\frac{\sigma^2}{2}, \tag{4.11}$$

with $\sigma^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ and $\mu = \mathbb{E}[X]$.

There's a subtlety here. The delta method requires the derivatives of the function h. For a generic LQFT, this will not be the case for $h(x) = \ln(x)$ since its derivatives diverge for $x \to 0^+$.

To overcome this problem one may require that the action of the considered lattice quantum field theory is bounded. For example, this can be ensured by putting the field theory in a box potential. Since only very high energy configurations (very unlikely to be sampled) are affected by this modification, there's no practical effect on the numerical experiments.

Bias and Variance of the Free Energy Estimator

Theorem 3. *The bias of the estimator* $\hat{F}$ *is given by*

$$\mathrm{Bias}\left[-\beta\hat{F}\right] = -\frac{1}{2\,N}\frac{\mathbb{E}_{q_\theta}\left[(\tilde{w}-\mathbb{E}_{q_\theta}\left[\tilde{w}\right])^2\right]}{\left(\mathbb{E}_{q_\theta}\left[\tilde{w}\right]\right)^2} + O(N^{-2})\,,$$

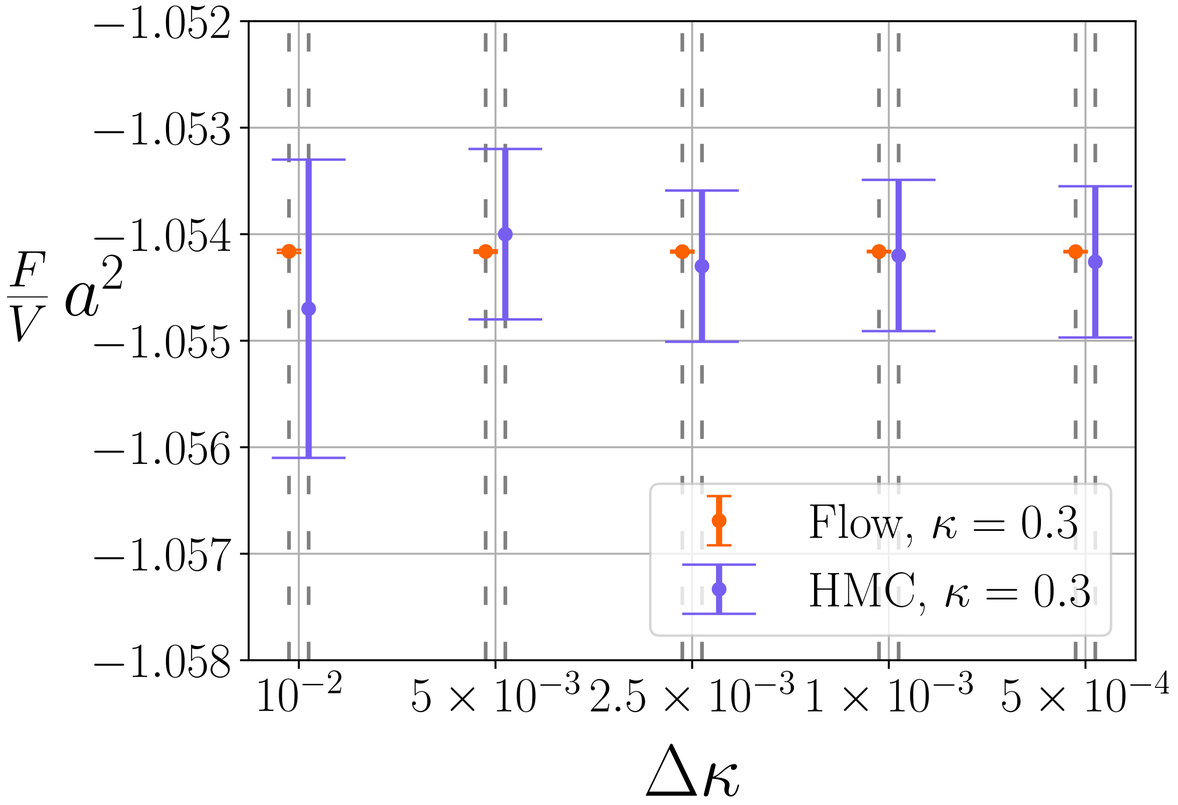
assuming that $\mathbb{E}_q[\tilde{w}^{2k+2}] < \infty$ *for* $k \in \{0,1\}$.

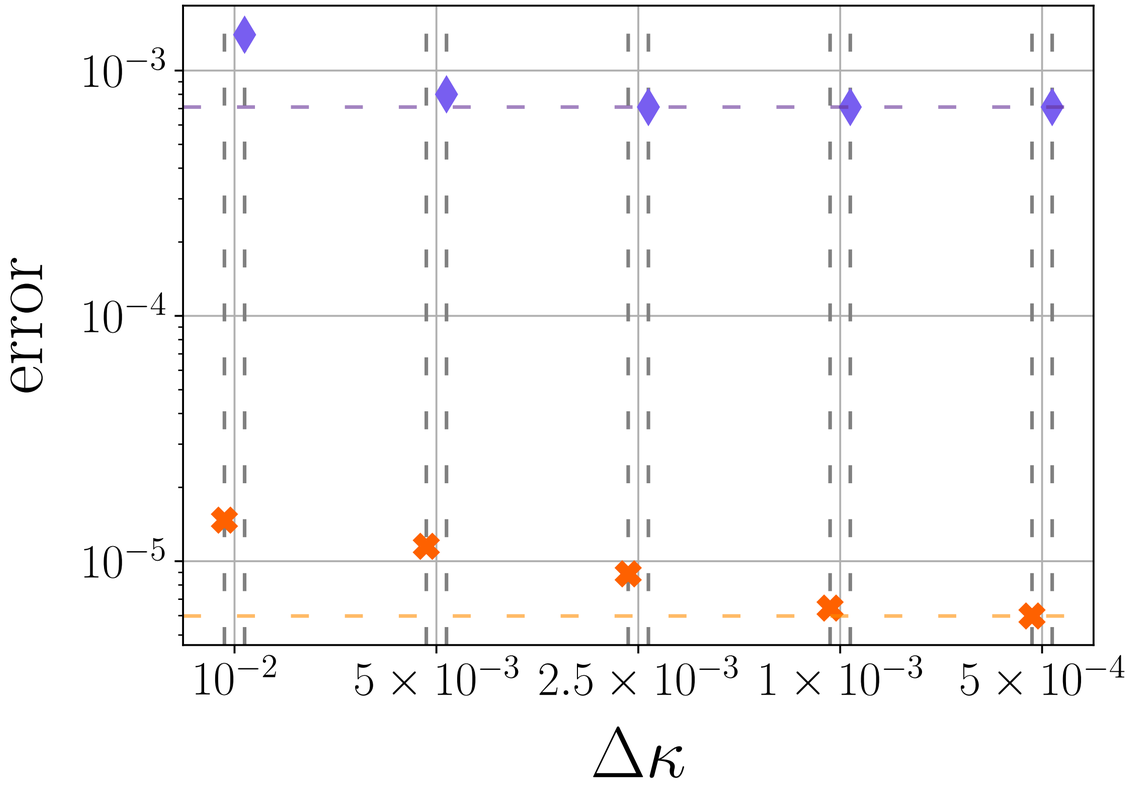
Theorem 4. *The variance of the estimator* $\hat{F}$ *is given by*

$$\mathrm{Var}\left[-\beta\hat{F}\right] = \frac{1}{N}\frac{\mathbb{E}_{q_\theta}\left[(\tilde{w}-\mathbb{E}_{q_\theta}\left[\tilde{w}\right])^2\right]}{\left(\mathbb{E}_{q_\theta}\left[\tilde{w}\right]\right)^2} + O(N^{-2})\,,$$

assuming that $\mathbb{E}_q[\tilde{w}^{2k+2}] < \infty$ *for* $k \in \{0,1\}$.

Error Scaling of Free Energy

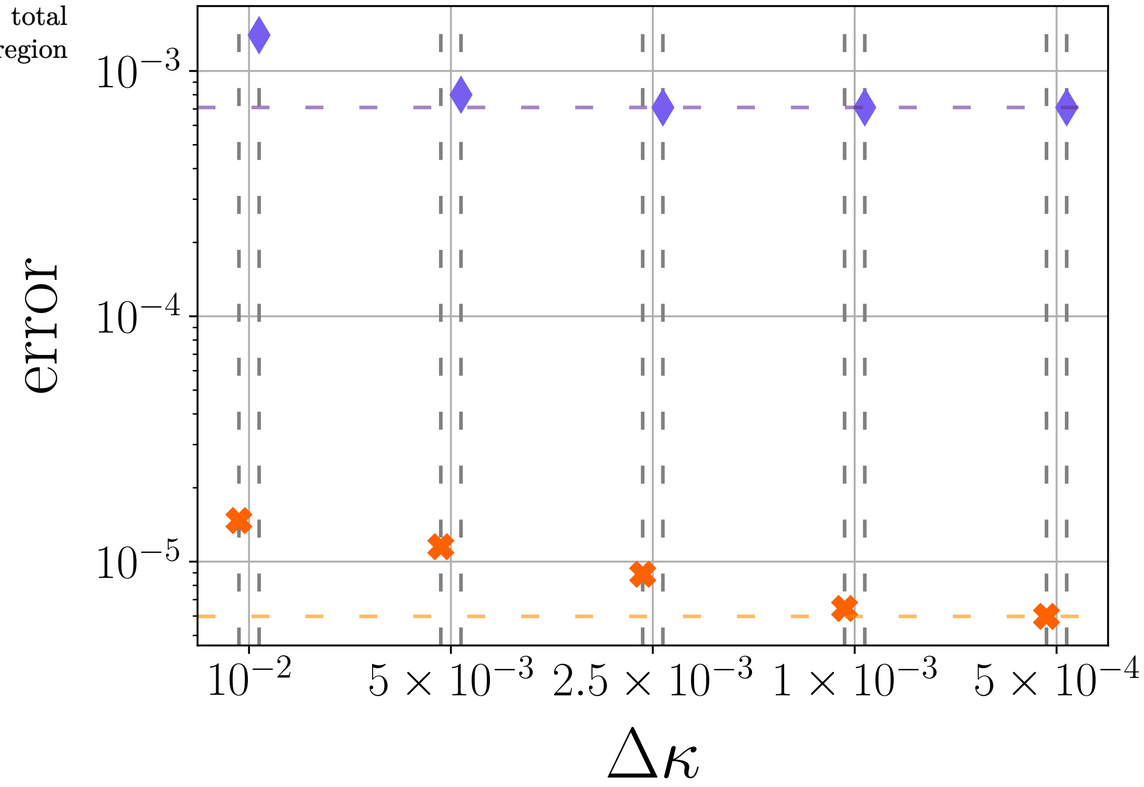




Error Scaling of Free Energy

Table 4.1. **Details on the refinement analysis for smaller $\Delta\kappa$**: in each refinement stage, we take steps $\Delta\kappa$ in a certain sub-region of the hopping parameter κ trajectory (see the last column). The focus is around the critical regime where the biggest error contribution for HMC, see fig. 4.5, is caused by critical slowing down phenomena. The step size taken in this region is shown in the first column. Outside this critical region, the same step sizes for all refinements are taken. Therefore, the total number of samples (third column) grows inversely proportional to the step size of the critical region (first column). Adapted from Nicoli *et al.* (2021).

$\Delta\kappa$	# chains	# samples	κ refined region
0.01	14	5.6 M	0.20-0.30
0.005	24	9.6 M	0.20-0.30
0.0025	40	16 M	0.22-0.30
0.001	76	30.4 M	0.24-0.30
0.0005	88	35.2 M	0.267-0.279



The Effective Sampling Size

$$N_{\text{eff}} = \frac{N}{1 + \text{Var}_{q_\theta}[w]}$$

Since

$$\text{Var}_{q_\theta}[w] = \mathbb{E}_{q_\theta}\left[w^2\right] - \mathbb{E}_{q_\theta}[w]^2 = \mathbb{E}_{q_\theta}\left[w^2\right] - 1$$

one recovers the definition of effective sample ratio (ESR)

$$N_{\text{eff}} = \frac{N}{\mathbb{E}_{q_\theta}[w^2]} \implies \text{ESR} = \frac{N_{\text{eff}}}{N} = \frac{1}{\mathbb{E}_{q_\theta}[w^2]}\,.$$

We refer to Kong (1992) for a mathematical derivation of the ESS from the variance of the estimator.

Is Reweigthing All You Need?

- Our estimator is **consistent** which implies the bias scales with $O(N^{-1})$

- This might have **potentially large** (or infinite) **variance** depending on the **goodness** of our variational approximation q.

- We can lower the bias by taking N very large **BUT** our estimator might still have prohibitively high variance if q is bad.

$$\text{Bias}[\hat{Z}] \approx O(N^{-1})$$

$$\text{Std}[\hat{Z}] \approx O(1/\sqrt{N}) \longrightarrow \text{Dominant term.}$$

- The variance of the importance weights, also relates to the effective sampling ratio (ESR)

$$\text{ESR} = \frac{N_{eff}}{N} = \frac{1}{\mathbb{E}_q[w^2]} \text{ relates to ESR} = \frac{N_{eff}}{N} = \frac{1}{2\tau_{int,\mathcal{O}}}$$

the goodness of the sampler relates to the degree of iid sampling from the generative model

Example: the Delta Function

$$Z = \int dx\ \tilde{p}(x) = 1 \longrightarrow \text{ where } \tilde{p}(x) \text{ is the Delta function centred in } x = 0$$

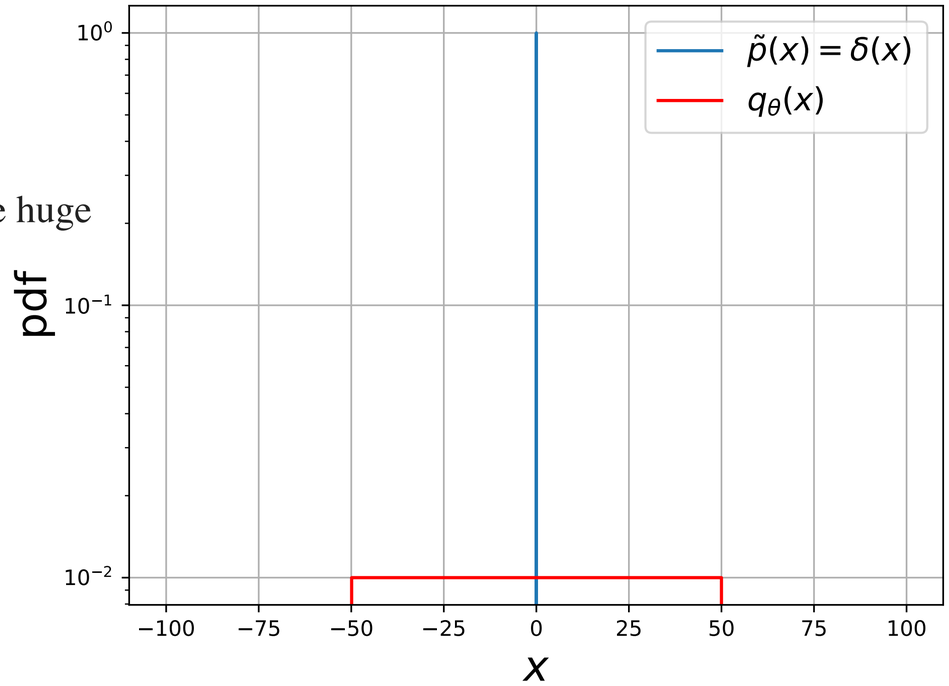
$$Z \approx N^{-1} \sum_n \tilde{p}(x_n) \qquad x_n \sim q_\theta = \text{Uniform}(-a, a)$$

x_n rarely hits the low energy valley of $\tilde{p}(x) \Longrightarrow$ for finite N, the bias and variance are huge

$$Z \approx \hat{Z} = N^{-1} \sum_n \tilde{p}(x_n) \approx 0 \qquad p(x_n) = 0 \quad \forall \ x_n \sim q$$

$$\text{Bias}[Z, \hat{Z}] = Z - \mathbb{E}_q[Z] = Z - \hat{Z} \approx Z = 1$$

$$\text{Var}[\hat{Z}] = (\mathbb{E}_q[Z]^2 - \mathbb{E}_q[Z^2]) = 0$$



- sampling **finite** number of samples: we hit the **high probability region** of $p(x)$ with very **low probability**. Hence $\tilde{p}(x_n) \approx 0 \rightarrow \hat{Z} \approx 0$
- sampling **Infinite** number of samples: we have a scaling $O(N^{-1})$. When $p(x) = \delta(x)$ **additional care is needed** for mathematical rigorousness around $x = 0$.

Flow-Based and HMC-Based sampling

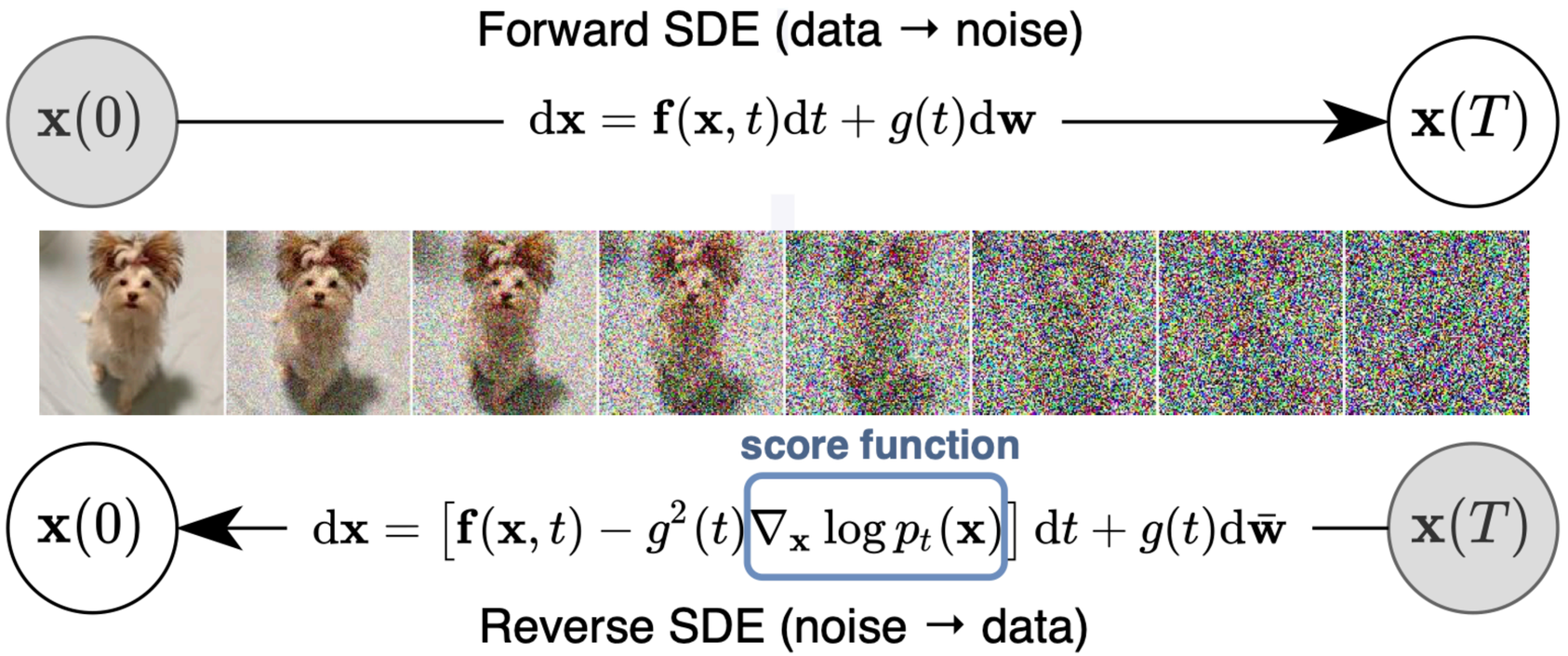
	PRO	CONS
HMC	• No training required • Asymptotically unbiased	• Sequential Sampling • Struggles at criticality • Critical Slowing down • No direct access to Z
FLOW	• Parallel sampling • Asymptotically unbiased • Direct access to Free Energy • Point-wise estimates • Smaller statistical errors • Can leverage inductive biases	• Up-front expensive training • Bias from imperfect training

Generalisation to Other Deep Generative Models

Table 3.1. **Applicability of NMCMC and NIS to different families of DGMs:** h refers to the explicit dependence of the operator $\mathcal{O}(\phi, Z)$ on the partition function Z, see eq. (3.25). The observations from the main text are summarized using the following notation about our methods: ✓: applicable and the estimator is *asymptotically unbiased*; (✓): applicable but the estimator is still *biased*; X: not applicable. Adapted from Nicoli *et al.* (2020).

DGM	Sampling probability	NMCMC, NIS($h \equiv 0$)	NIS($h \neq 0$)
GAN	none	X	X
RBMs	approximate, unnormalized	(✓)	X
VAE	approximate, normalized	(✓)	(✓)
ARNN, NF	exact, normalized	✓	✓

From Normalizing Flows to Diffusion Models

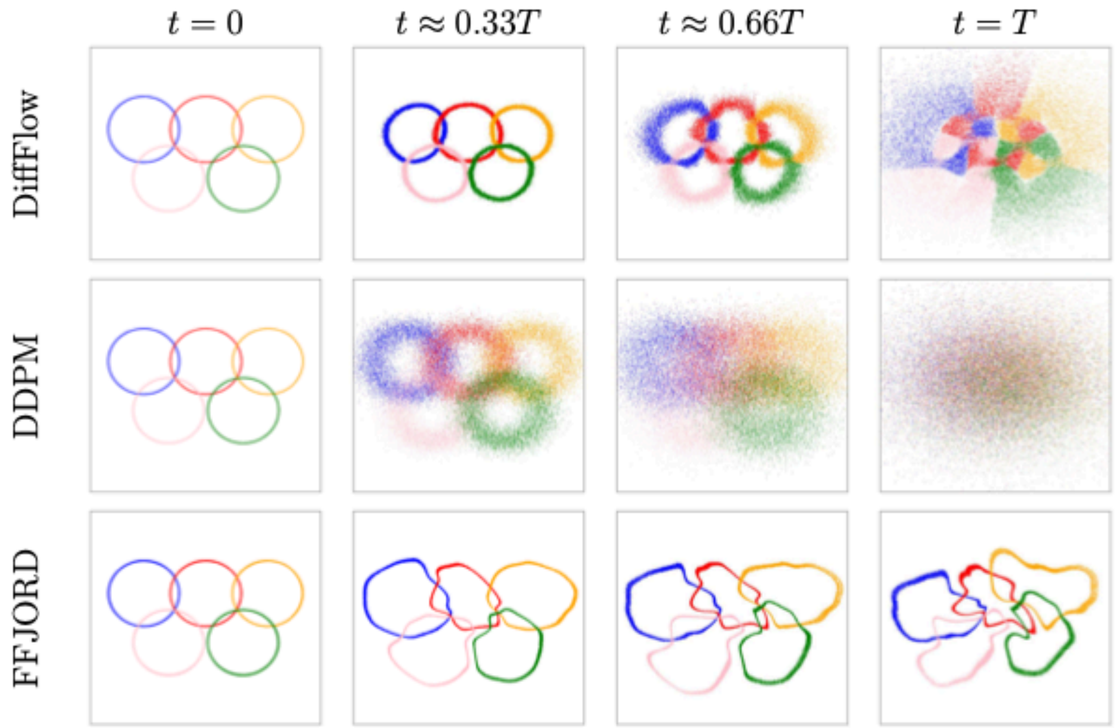


Y. Song et al., ICLR (2021)

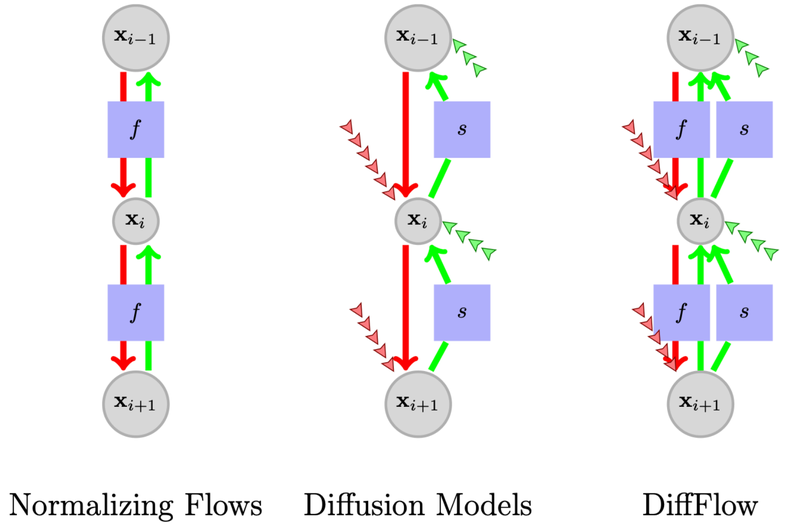
In principle one can compute the **exact likelihood** on any input data leveraging instantaneous change of variable transformation from **Neural ODEs**.

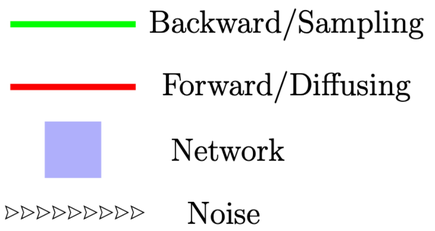
However, Diffusion Models: **I)** are slow to train and sample from, **II)** require data, **III)** are memory intensive **IV)** need a lot of intermediate steps.

DiffFlow: Diffusion Normalizing Flows



Q. Zhang et al., NeurIPS (2021)





	Forward Process (Diffusion)	Reverse Process (Denoising)
DDPM	$\mathrm{d}\mathbf{x} = f(\mathbf{x}, t)\,\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$	$\mathrm{d}\mathbf{x} = [f(\mathbf{x}, t) - g^2(t)\,\mathbf{s}(\mathbf{x}, t)]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$
DiffFlow	$\mathrm{d}\mathbf{x} = f(\mathbf{x}, t, \theta)\,\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$	$\mathrm{d}\mathbf{x} = [f(\mathbf{x}, t, \theta) - g^2(t)\,\mathbf{s}(\mathbf{x}, t, \theta)]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$

→ **Parametrised Drift Term**

How Big is Enough?

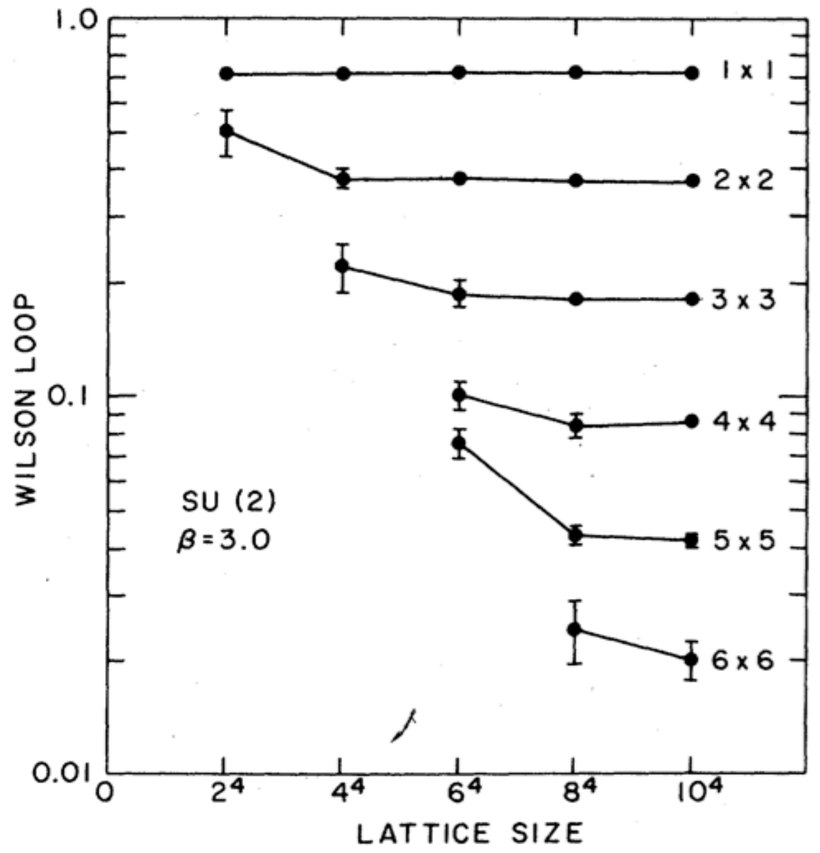


FIG. 3. Wilson loops at $\beta = 3$ as a function of lattice size.

M. Creutz, Phys. Rev. D (1980)

64×8 is not small !!!

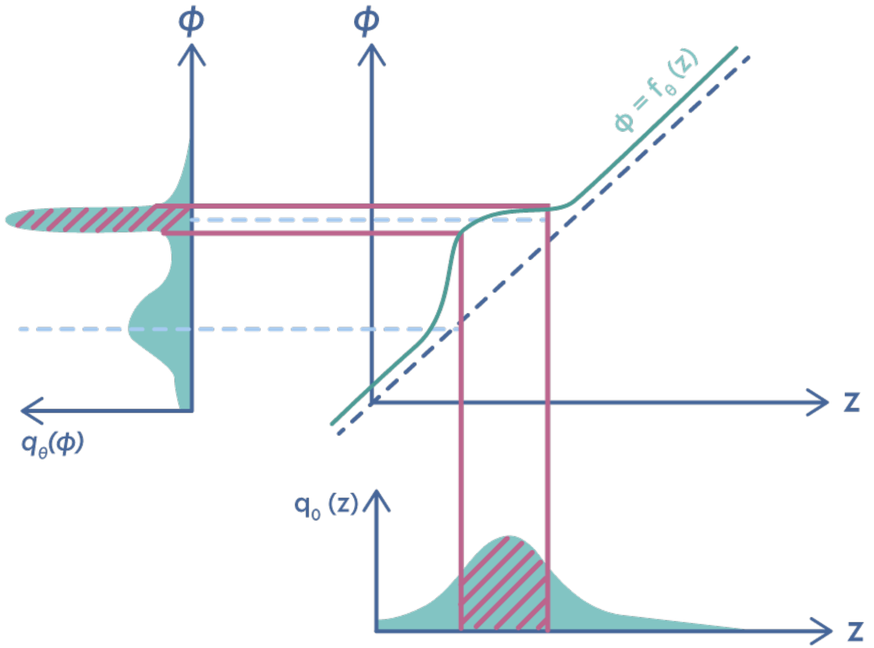
- $\sim$ 40 years were needed to scale up LQCD.
- To start small is totally justified.
- Large room for improvement.
- Initial experiments as proof of concepts.

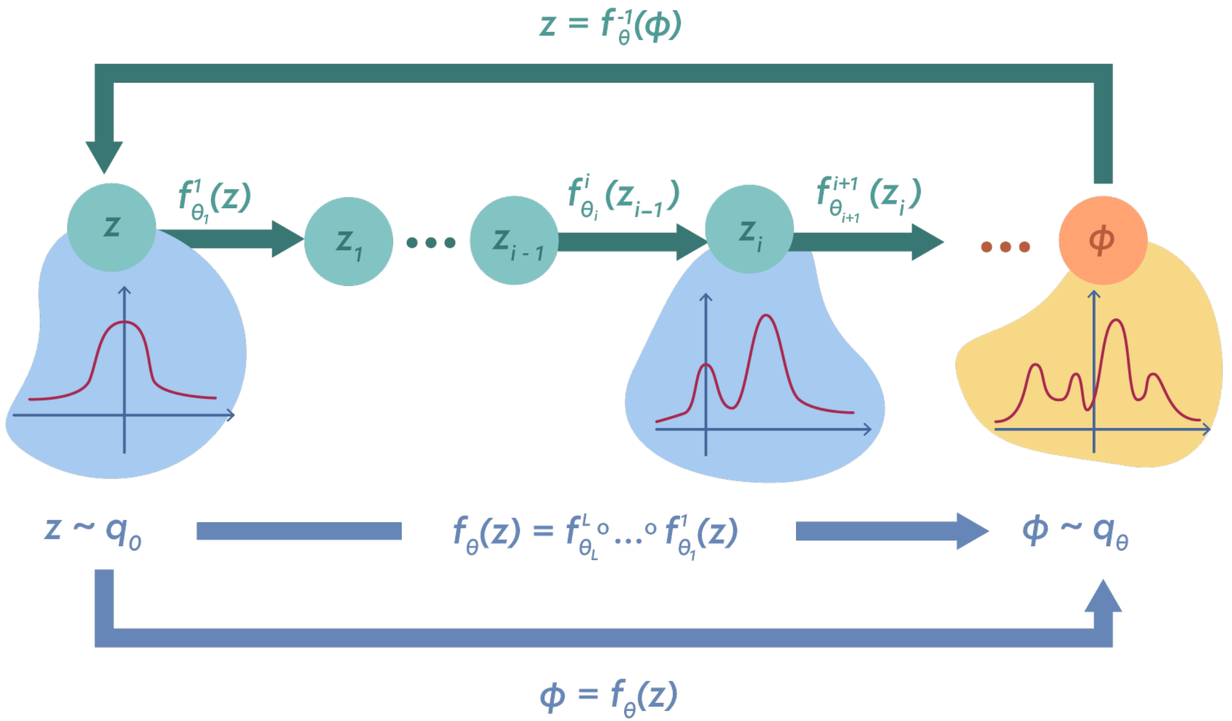
Temporal extent — SU(3) dimensionality

Full QCD goal: $64^3 \times 128 \times 4 \times 8 \times 2 \approx 10^9$

Spatial extent — Links from vertex — Complex valued

Normalizing Flows





- Bijective transformation $\phi = f_\theta(z)$
- Invertible and differentiable*
- Tractable Jacobian

$$q_\theta(\phi) = q_0(f_\theta^{-1}(\phi)) \left| \det \left(\frac{\partial f_\theta}{\partial z} \right) \right|^{-1}$$

*Enough to be **a diffeomorphism** and **differentiable almost everywhere** wrt the Lebesgue measure of the base space

Normalizing Flows: a different intuition

Let be $z \in \mathbb{R}^D$ be a **random variable** sampled from a **base** distribution $p_z : \mathbb{R}^D \to \mathbb{R}$

Let be $f : \mathbb{R}^D \to \mathbb{R}^D$ be a **diffeomorphism** on $\mathbb{R}^D$ such that the base r.v. transforms like $\phi = f(z)$.

Using the change of variable formula one can compute the probability density of the transformed variable ϕ

$$q_\theta(\phi) = q_0(f_\theta^{-1}(\phi)) \left| \det \left(\frac{\partial f_\theta}{\partial z} \right) \right|^{-1}$$

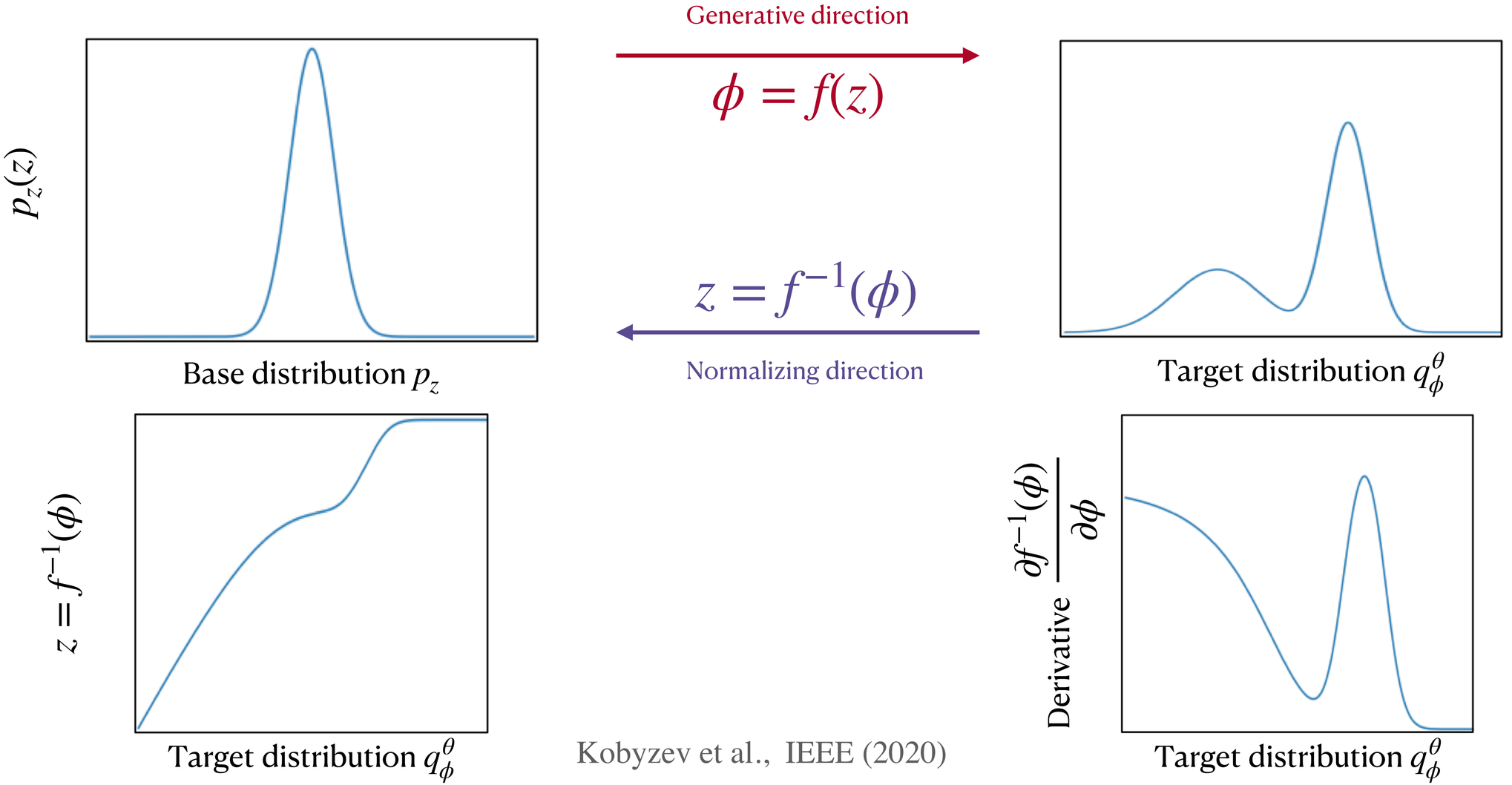
This new density is called the ***pushforward*** of the base density p_z by the function f and can be denoted by $f_* p_z$

Generative Direction: the function f pushes forward the base density p_z (noise) to a more complicated one q_ϕ^θ parametrised by θ .

One can also define a ***pullback*** $f_*^{-1} q_\phi^\theta$ which maps the non-trivial density back into the base one**.**

Normalizing Direction: the inverse function f^{-1} flows backwards pulling back the complicated density towards the more regular form of the base measure p_z.

Normalizing Flows: a different intuition



Kobyzev et al., IEEE (2020)

Normalizing Flows: a different intuition

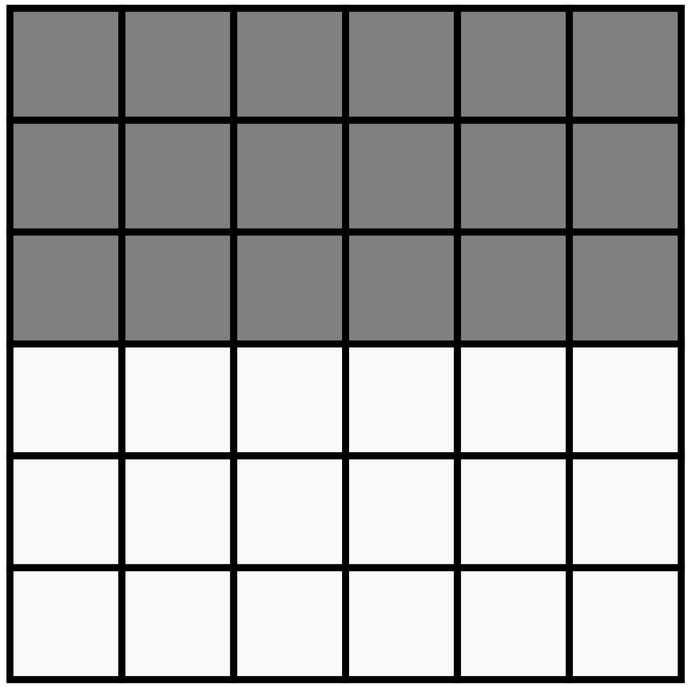
This view is what gives rise to the name **"normalizing flows"** as f is "normalizing" the data distribution.

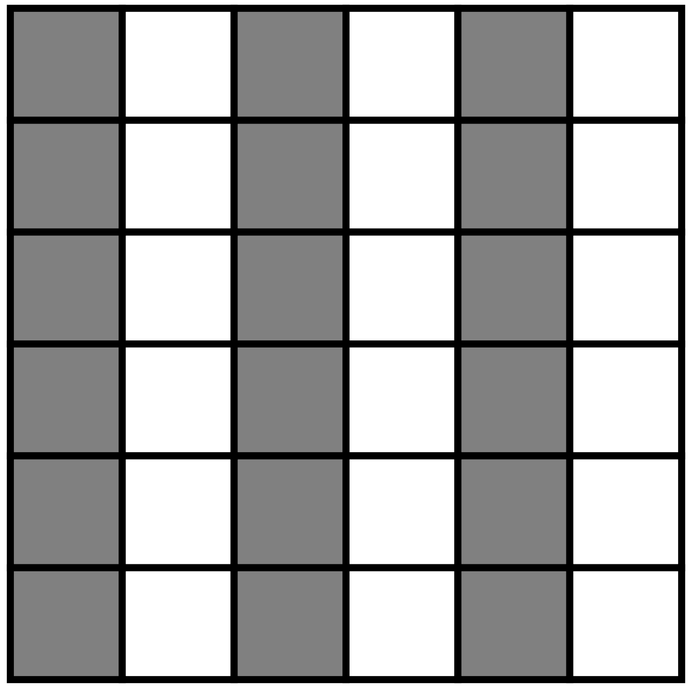
Normalizing Flows $\Longrightarrow$ ~~bijections~~ ⟶ Formally incorrect

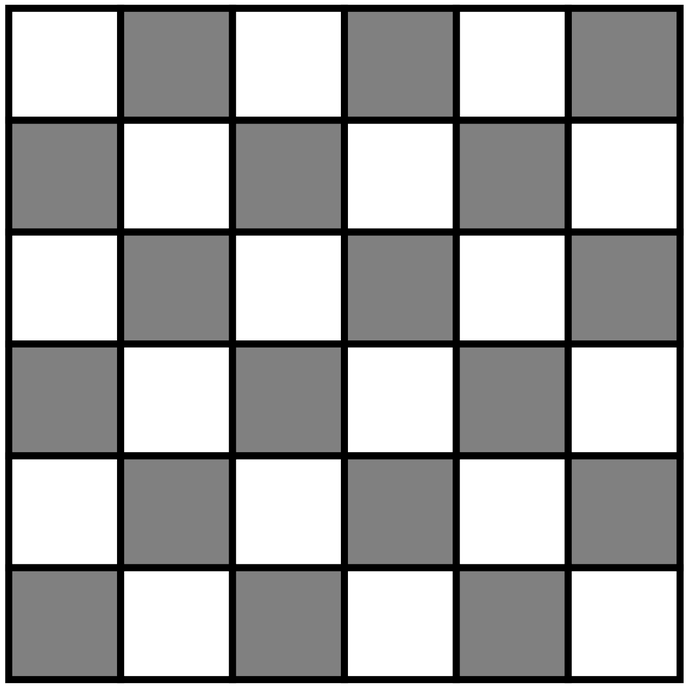
Normalizing Flows $\Longrightarrow$ **diffeomorphism** differentiable **almost everywhere** with respect to the Lebesgue measure on $\mathbb{R}^D$.

This allows for piecewise differentiable functions to be used to construct the transformation f.

NICE Splitting Strategies







Training a Flow with Reverse-KL

- **Minimize** the Reverse KL divergence

$$KL(q_\theta||p) = \int D[\phi] q_\theta(\phi) \ln \frac{q_\theta(\phi)}{p(\phi)} = \mathbb{E}_{q_\theta}\left[\ln q_\theta(\phi) - \ln p(\phi)\right] \geq 0$$

- **Maximize** the variational lower bound on the free energy F_q

$$KL(q_\theta||p) = \underbrace{\mathbb{E}_{q_\theta}\left[\ln q_\theta(\phi) + S(\phi)\right]}_{\beta F_q} + \underbrace{\ln Z}_{F=-\beta \ln Z} = \beta(F - F_q)$$

Training a Flow with Forward-KL

Input:

- prior density, e.g., $q_z \sim \mathcal{N}(0, \mathbb{1})$
- parametric model with parameters θ
- parametric action $S(\phi, \kappa, \lambda)$ with fixed coupling parameters $\{\lambda, \kappa\}$
- empty tensor for storing a batch of B configurations $\mathbf{\Phi} \in \mathbb{R}^{B \times N_S \times N_T}$

Result:

- learned bijective transformation f_θ s.t. $\phi_i = f_\theta(z_i)$
- exact likelihood function q_θ

begin
  `/* Generate training samples from HMC */`
  for c *in* $1, \ldots, C_{\text{max}}$ **do**
    $\phi =$ `sampleHMC` (λ, κ) ; `// sample configurations` $\boldsymbol{\phi} \in \mathbb{R}^{n \times N_S \times N_T}$
    $\mathbf{\Phi} = \text{concatenate}(\mathbf{\Phi}, \boldsymbol{\phi})$; `// concatenate configurations`
  end
  `/* Flow training using the generated dataset` $\mathbf{\Phi}$ `*/`
  for t *in* $1, \ldots, T_{\text{max}}$ **do**
    `/* Iterate over dataset to collect batches of configurations */`
    draw samples $\boldsymbol{\phi} = \{\phi_i\}_{i=1}^m$ from $\mathbf{\Phi}$ where $\phi_i \in \mathbb{R}^{N_S \times N_T}$ $\forall i \in \{1, \ldots, m\}$
    use $\boldsymbol{\phi}$ to evaluate $\frac{1}{m} \sum_{i=1}^m \ln q_\theta(\phi_i)$;
    $\Delta\theta = \underset{\theta}{\text{argmin}}\, \mathbb{E}_p \left[\ln q_\theta(\phi)\right]$;
    update f_θ with $\theta \leftarrow \theta + \eta\, \Delta\theta$;
  end
end
return f_θ, q_θ

Two Sides of the Same Coin

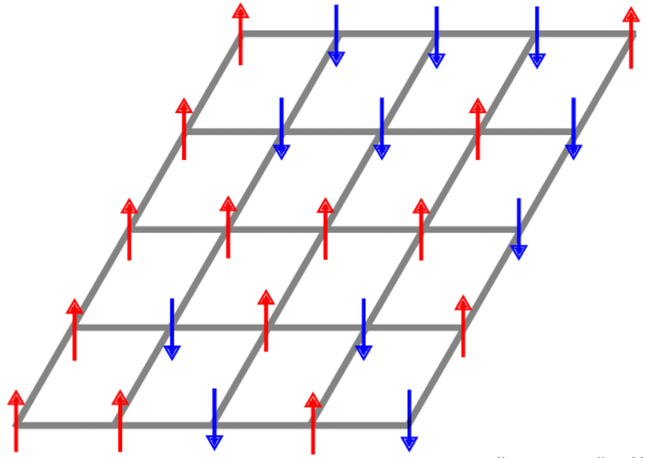
Learning from **data**

(Image credits: medium.com)

- Maximum likelihood (needs data)
- Minimising **Forward KL**
- Target: generating distribution

$$\mathscr{L} = \mathbb{E}_{\phi\sim p}[\ln q_\theta(\phi)]$$

Learning from **physics**



(Image credits: Wald 2017)

- Self-sampling (needs action/hamilt.)
- Minimising **Reverse KL**
- Target: samples, normalisation Z

$$\mathscr{L} = \mathbb{E}_{\phi\sim q_\theta}\left[\ln q_\theta(\phi) + S(\phi) + \ln Z\right]$$

(Slide inspired by Lei Wang lecture notes)

The Free Energy as a Natural Cost Function

$$F = \underbrace{E}_{\text{Energy}} - \underbrace{TS}_{\text{Entropy}}$$

Free Energy is a **Cost Function** given by nature!

$$KL(q_\theta||p) = \int D[\phi] q_\theta(\phi) \ln \frac{q_\theta(\phi)}{p(\phi)} = \mathbb{E}_{q_\theta}\left[-\ln \overbrace{p(\phi)}^{\approx \mathbb{E}_{q_\theta}[S(\phi)]} + \ln \overbrace{q_\theta(\phi)}^{\mathbb{E}_{q_\theta}[\ln q_\theta(\phi)]}\right] \simeq F_q \geq 0$$

The optimisation problem translates into minimising a variational form of the free energy F_q

(Slide inspired by Lei Wang lecture notes)

HMC Estimate of Free Energies

$$\Delta F_{e\,b} = F_e - F_b \implies F_e = \Delta F_{e\,b} + F_b$$

In the limit of vanishing hopping parameter:

$$F_0(\lambda) = -\,|\,\Lambda\,|\,T\,\ln z_0(\lambda) = -\,T\,\ln Z_0(\lambda) \qquad \text{with} \qquad z_0(\lambda) = \sqrt{\frac{1-2\lambda}{4\lambda}}\,\exp\left(\frac{(1-2\lambda)^2}{8\lambda}\right) K_{\frac{1}{4}}\left(\frac{(1-2\lambda)^2}{8\lambda}\right)$$

Difference of free energies can also be understood in terms of ratios of partition functions:

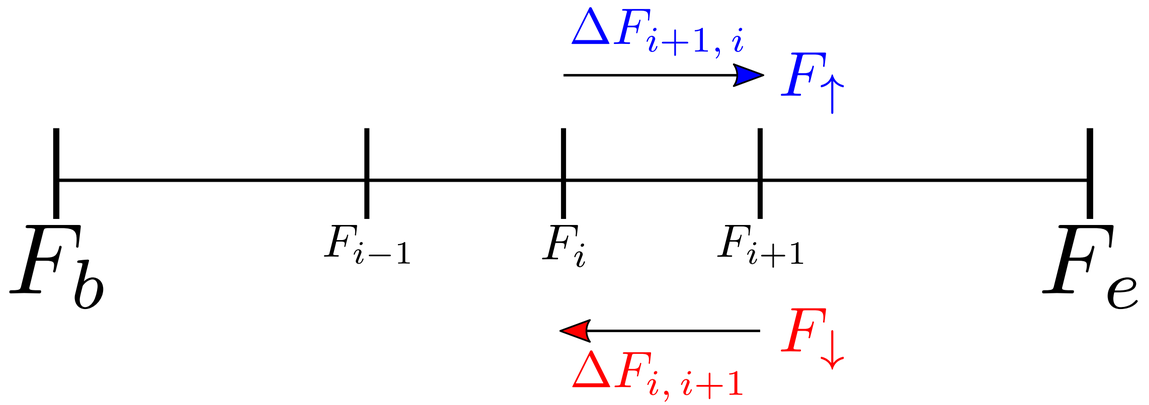
$$\Delta F_{i+1,\,i} = F_{i+1} - F_i = -\,T\ln\frac{Z_{i+1}}{Z_i}$$

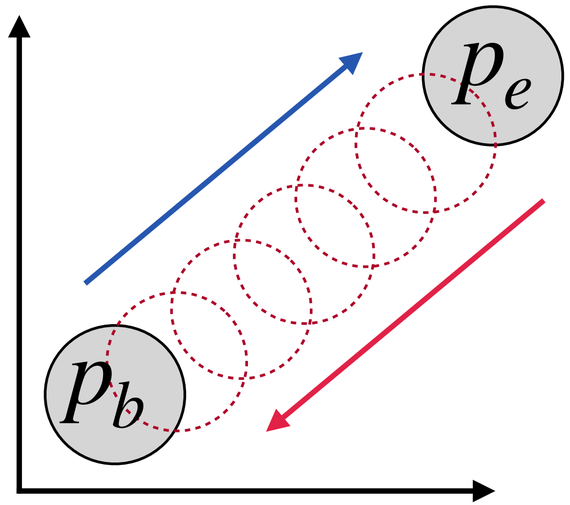
⚠ Actions should be similar to ensure small variance!

Which can be estimated by MCMC sampling:

$$\mathbb{E}_{p_i}\left[\frac{\exp(-S_{i+1})}{\exp(-S_i)}\right] = \frac{1}{Z_i}\int \mathcal{D}[\phi]\, e^{-S_i(\phi)}\,\overbrace{\frac{e^{-S_{i+1}(\phi)}}{e^{-S_i(\phi)}}} = \frac{Z_{i+1}}{Z_i}$$

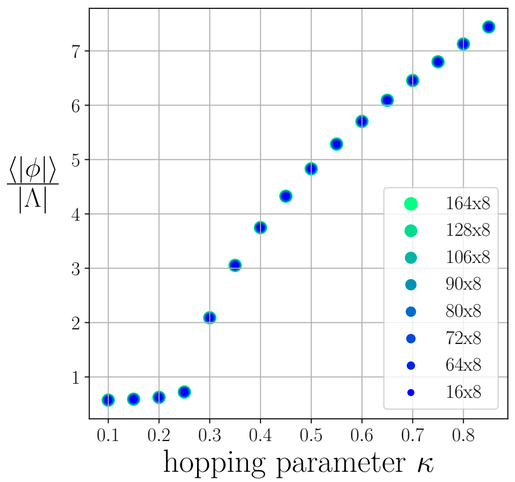
HMC Estimate of Free Energies

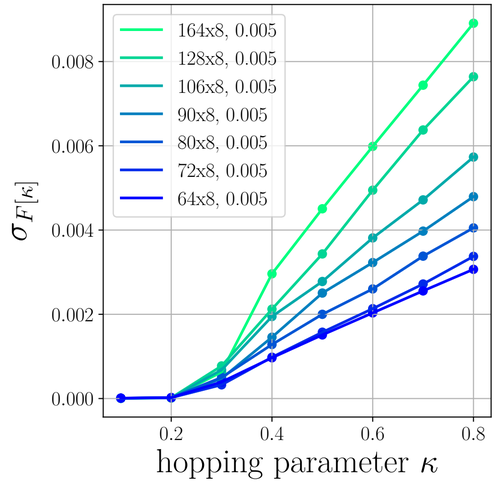


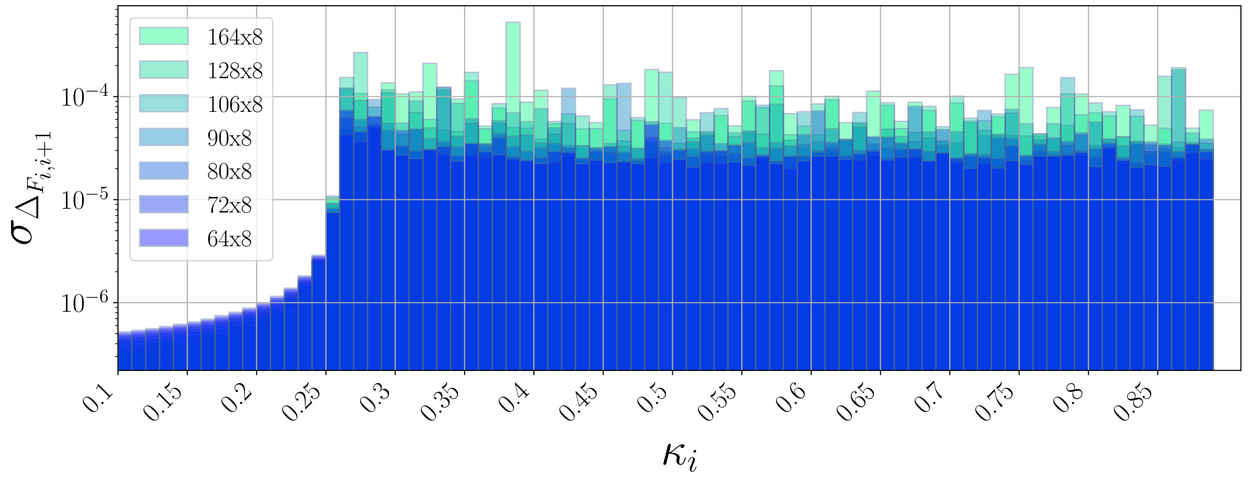


(Image inspired by Weinsberger et al. (2020))

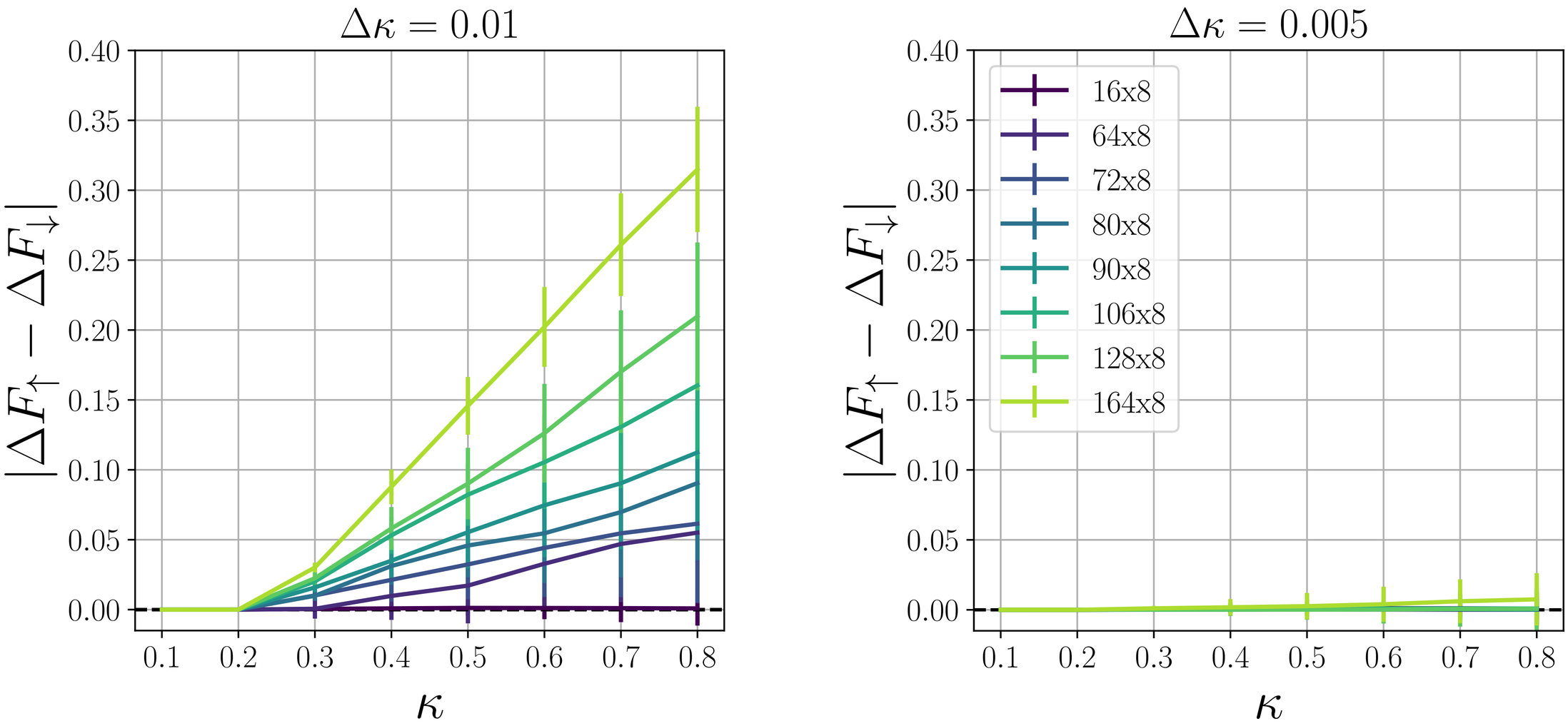
HMC Estimate of Free Energies



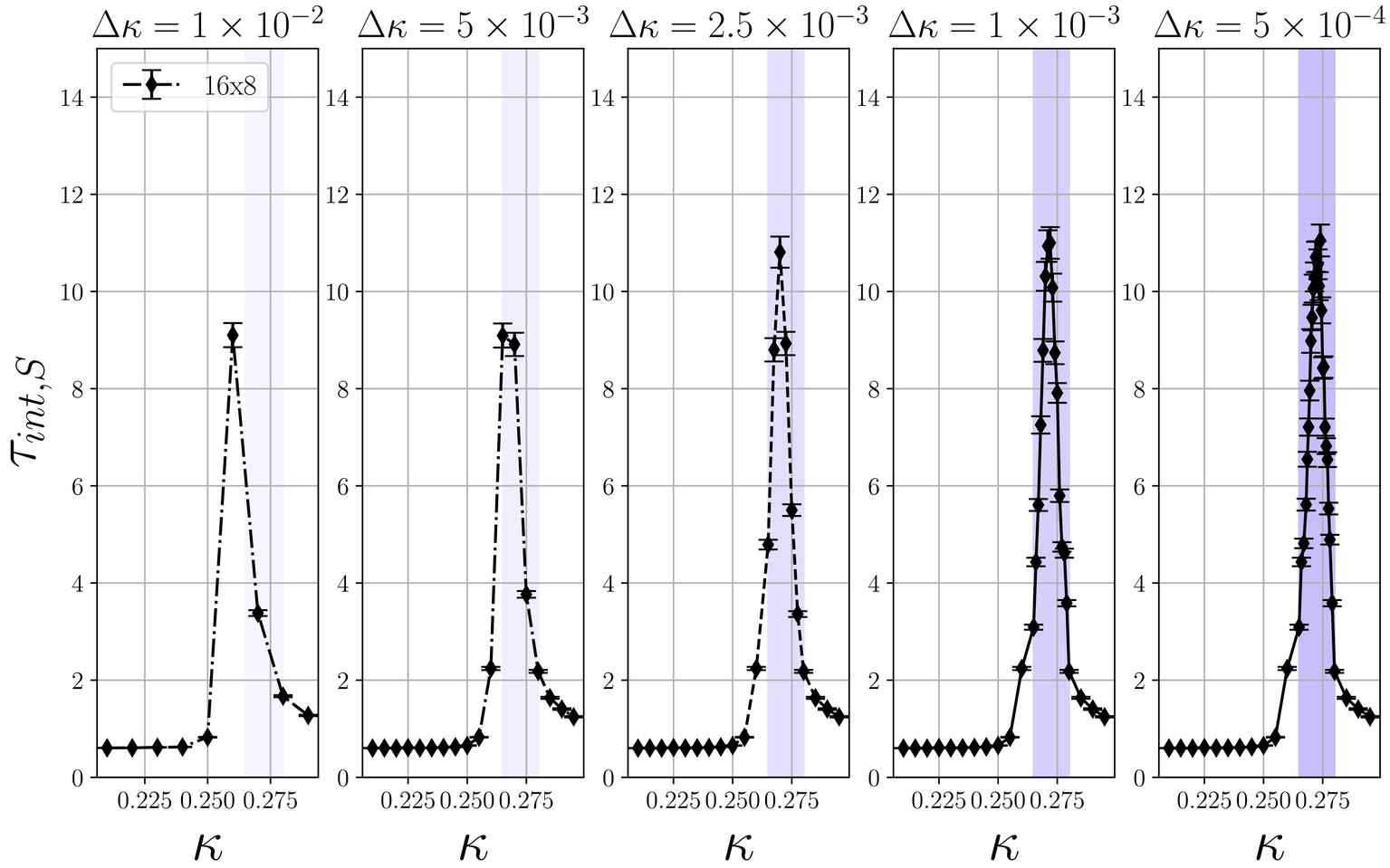




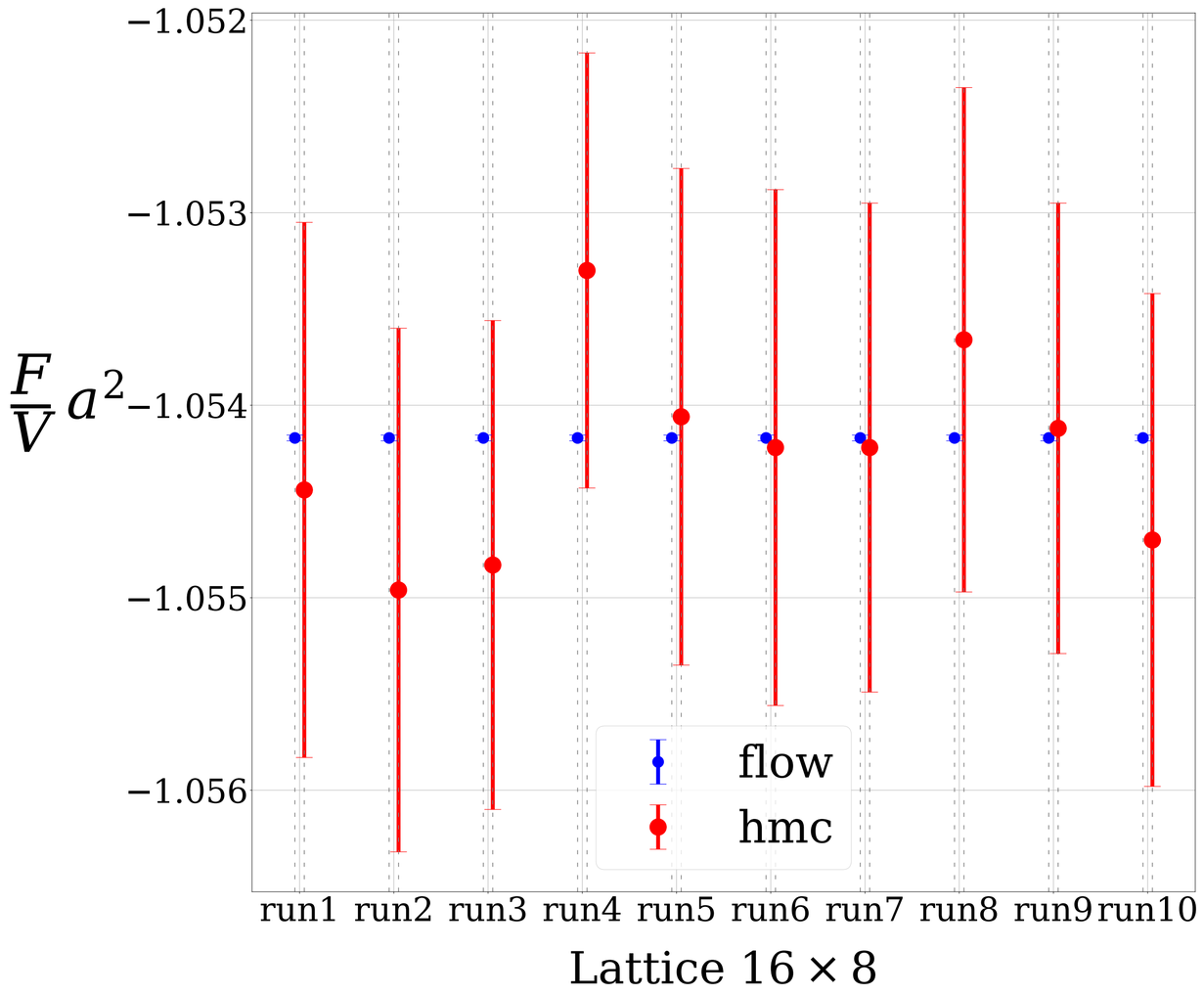
HMC Estimate of Free Energies



Integrated Autocorrelation



Stability of the Estimator



KAN, et al., Phys. Rev. Lett. (2021)

Reparametrization Trick

$$\mathrm{KL}(q_\theta||p) = \mathbb{E}_{x\sim q_\theta}\left[S(x) + \log q_\theta(x)\right] + \text{const.}$$

Reparametrization Trick $\implies$ $x \sim q_\theta \to z \sim q_z$

$$\boxed{\mathbb{E}_{x\sim q_\theta}[f(x)] = \mathbb{E}_{z\sim q_z}[f(g_\theta(z))]}$$

Compute Gradient $\frac{d}{d\theta}$

$$\frac{d}{d\theta}\mathrm{KL}(q_\theta||p) = \mathbb{E}_{z\sim q_Z}\left[\frac{d}{d\theta}S(g_\theta(z)) + \frac{d}{d\theta}\log q_\theta(g_\theta(z))\right] \underbrace{\approx}_{MC} G_{Rep}$$

Total Gradient Estimator

$$G_{RepQP} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{d}{d\theta}S(g_\theta(z_i)) + \frac{d}{d\theta}\log q_\theta(g_\theta(z_i))\right), \quad z_i \sim q_Z$$

Path derivative

$$\blacktriangledown_\theta f(g_\theta(z),\theta) = \frac{\partial f(g_\theta(z),\theta)}{\partial g_\theta(z)}\frac{\partial g_\theta(z)}{\partial \theta}$$

Total derivative **Partial derivative**

$$\frac{d}{d\theta}f(g_\theta(z),\theta) = \blacktriangledown_\theta f(g_\theta(z),\theta) + \frac{\partial}{\partial\theta}f(x,\theta)\Big|_{x=g_\theta(z)}$$

$$\frac{d}{d\theta}\mathrm{KL}(q_\theta||p) = \mathbb{E}_{z\sim q_Z}\left[\blacktriangledown_\theta S(g_\theta(z)) + \blacktriangledown_\theta \log q_\theta(g_\theta(z))\right] + \mathbb{E}_{z\sim q_Z}\left[\frac{\partial}{\partial\theta}\log q_\theta(x)\Big|_{x=g_\theta(z)}\right]$$

Path and Score Terms

$$G_{RepQP} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{d}{d\theta}S(g_\theta(z_i)) + \frac{d}{d\theta}\log q_\theta(g_\theta(z_i))\right), \quad z_i \sim q_Z$$

Path derivative

$$\blacktriangledown_\theta f(g_\theta(z),\theta) = \frac{\partial f(g_\theta(z),\theta)}{\partial g_\theta(z)}\frac{\partial g_\theta(z)}{\partial \theta}$$

Total derivative

$$\frac{d}{d\theta}f(g_\theta(z),\theta) = \blacktriangledown_\theta f(g_\theta(z),\theta) + \frac{\partial}{\partial\theta}f(x,\theta)\Big|_{x=g_\theta(z)}$$

Partial derivative

$$\frac{d}{d\theta}\mathrm{KL}(q_\theta||p) = G_{RepQP} = G_{PathQP} + G_{Score}$$

Path and Score Terms

$$G_{RepQP} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{d}{d\theta}S(g_\theta(z_i)) + \frac{d}{d\theta}\log q_\theta(g_\theta(z_i))\right), \quad z_i \sim q_Z$$

Path derivative

$$\blacktriangledown_\theta f(g_\theta(z),\theta) = \frac{\partial f(g_\theta(z),\theta)}{\partial g_\theta(z)}\frac{\partial g_\theta(z)}{\partial \theta}$$

Total derivative

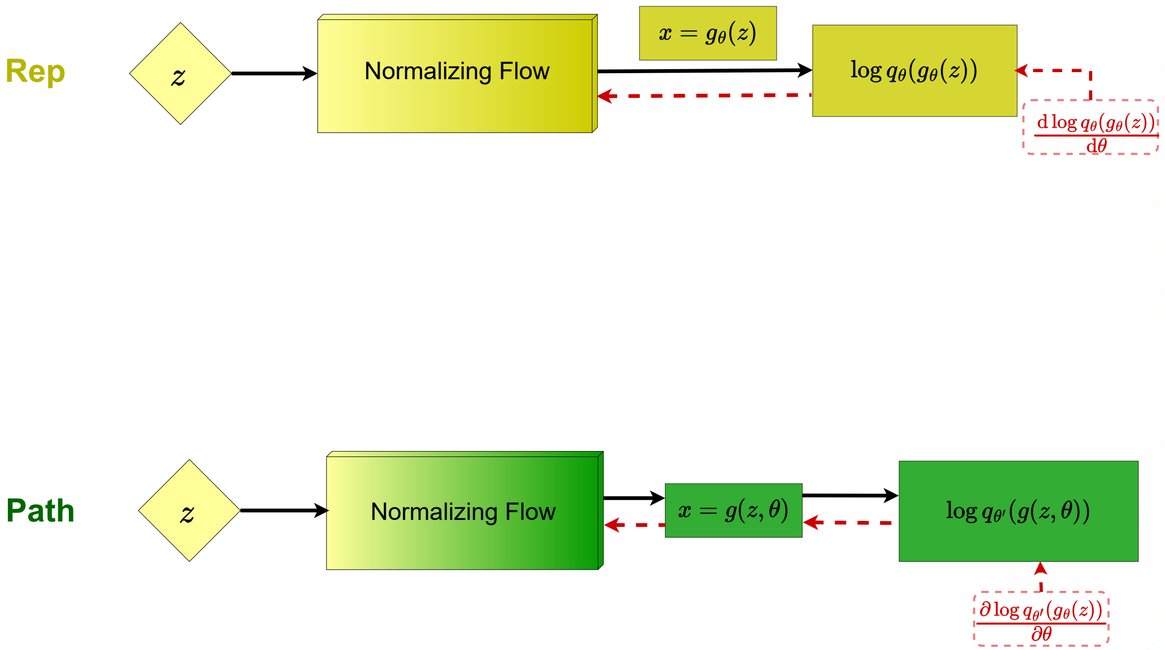
$$\frac{d}{d\theta}f(g_\theta(z),\theta) = \blacktriangledown_\theta f(g_\theta(z),\theta) + \frac{\partial}{\partial\theta}f(x,\theta)\Big|_{x=g_\theta(z)}$$

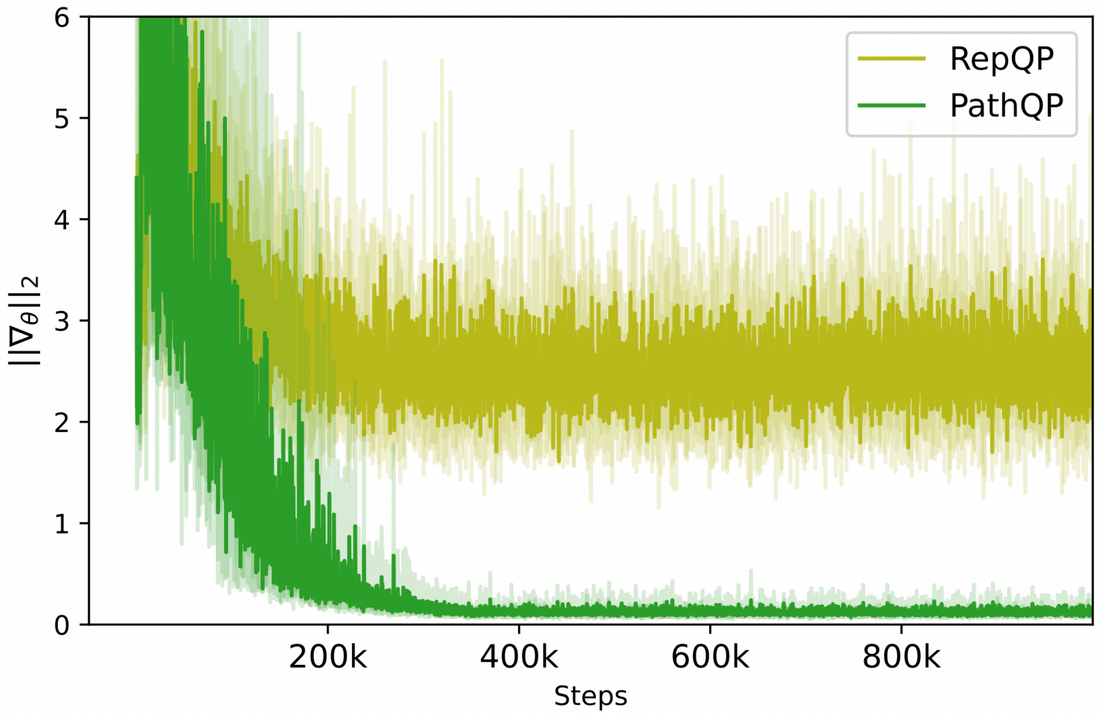
Partial derivative

$$G_{RepQP} \approx G_{PathQP} + \cancel{G_{Score}}$$

Roeder, G. et al., NeurIPS (2017)

Stable Gradients Mean Better Convergence





Stable Gradients Mean Better Convergence

Algorithm 1: Path gradient $\blacktriangledown_\theta \log q_\theta(g_\theta(z))$.

Input: base sample $z \sim q_Z$

$x' \leftarrow \text{stop\_gradient}(g_\theta(z))$ # forward pass of z through the flow without gradients

$q_\theta(x') \leftarrow q_Z(g_\theta^{-1}(x')) \left| \frac{\partial g_\theta^{-1}(x')}{\partial x'} \right|$ # reverse pass to calculate density

$G \leftarrow \frac{\partial \log(q_\theta(x'))}{\partial x'}$ # compute gradient with respect to x'

$x \leftarrow g_\theta(z)$ # standard forward pass

return $\frac{d}{d\theta}\left(\text{stop\_gradient}(G)^T x\right)$ # for path-gradient, contract $\frac{\partial x}{\partial \theta}$ with $\partial \log q_\theta(x)/\partial x$

Quantum Mechanical Particle in Double Well

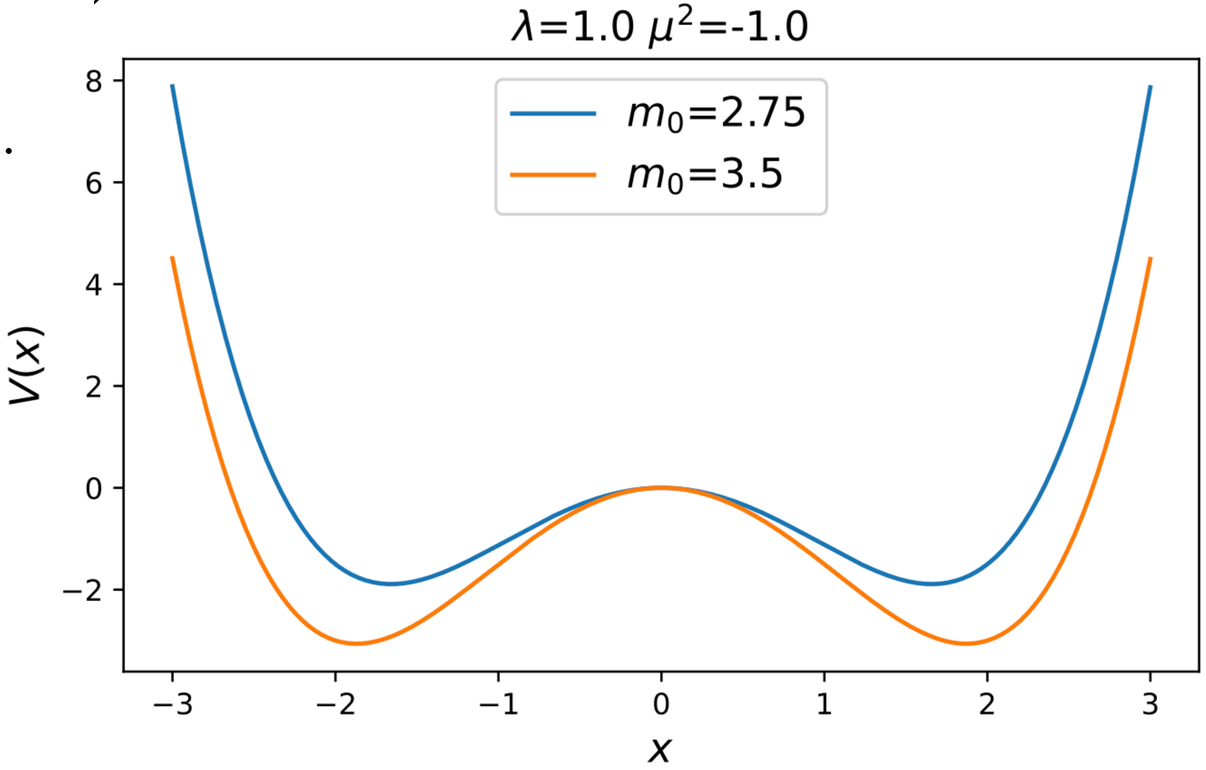
$$S(x) = a \sum_{t=0}^{T-1} \left(\frac{m_0}{2} \left(x_{t+1} - x_t \right)^2 + \frac{m_0 \mu^2}{2} x^2 + \frac{\lambda}{4} x^4 \right)$$

Where $\{m_0, \mu, \lambda\}$ control the shape of the potential.

The probability density of the path is given by

$$p(\mathbf{x}) = \frac{1}{Z} \exp(-S(\mathbf{x}))$$

for a discretised trajectory $\mathbf{x} = (x_0, x_1, \ldots, x_{T-1})$



A Few Words on the Effective Sampling Size

MCMC

$$\mathrm{ESS} = N_{eff} = \frac{N}{2\tau_{int,\mathcal{O}}}$$

Reverse

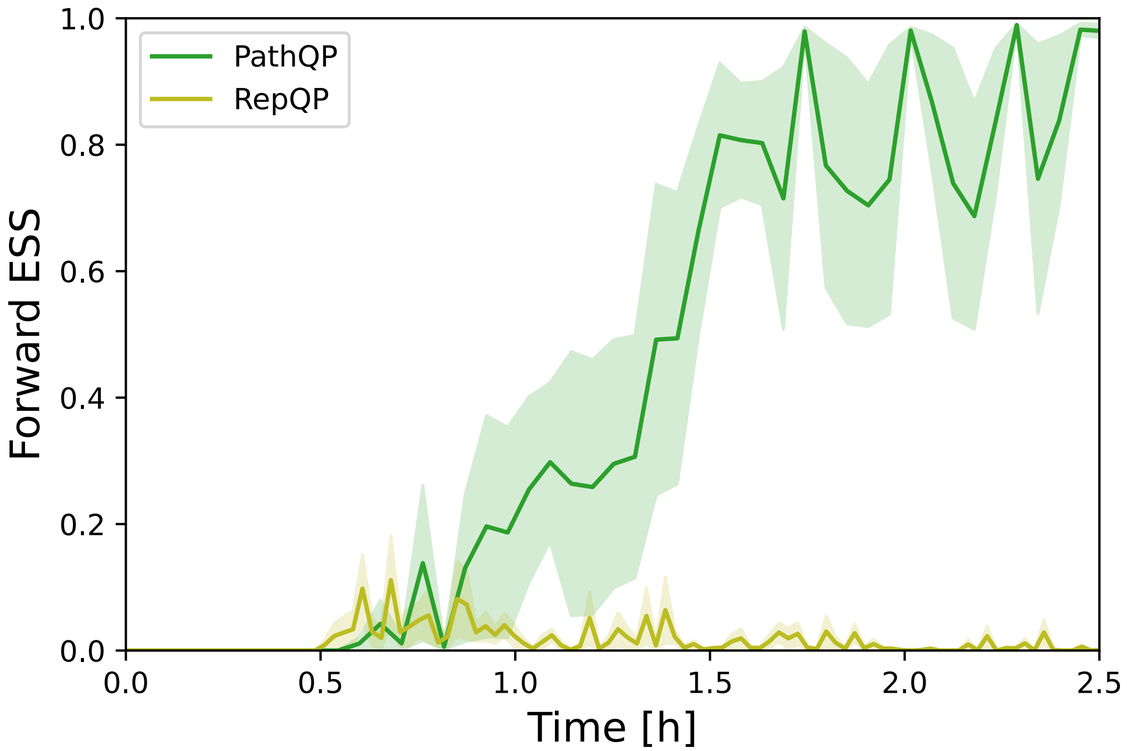
$$\mathrm{ESS} = N_{eff} = \frac{N}{\mathbb{E}_{q_\theta}[w^2]}$$

Forward

$$\mathrm{ESS} = N_{eff} = \frac{N}{\mathbb{E}_{p}[w]}$$

Forward ESS for a Double-Well

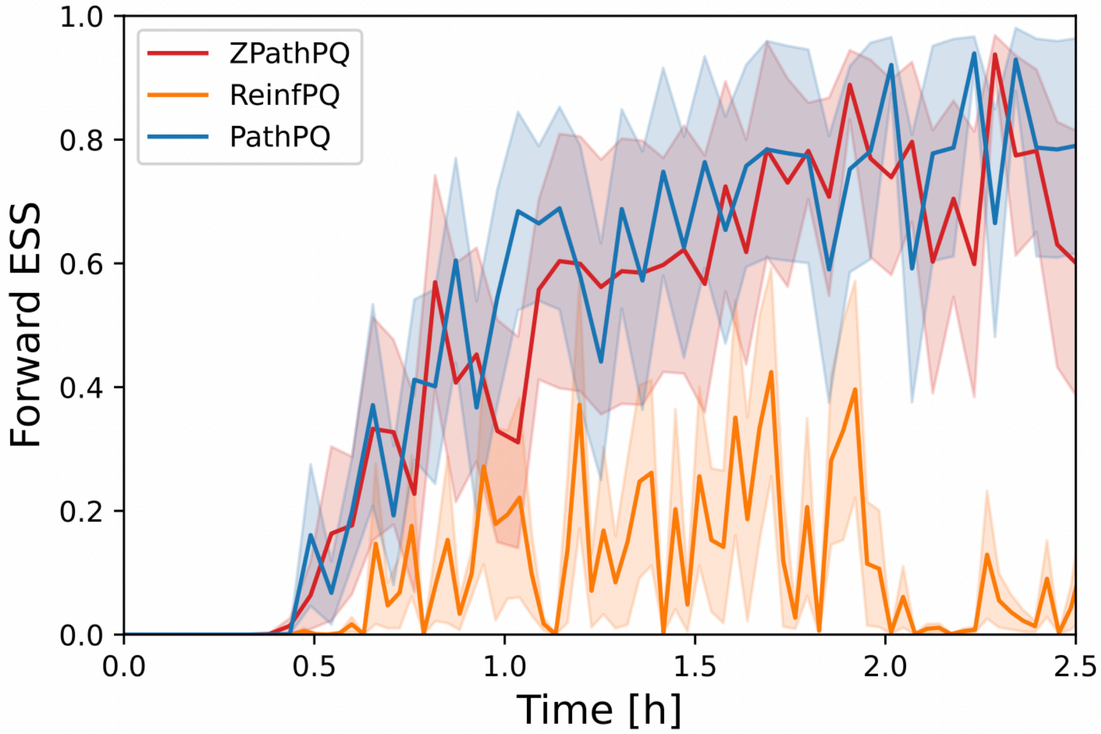
$\{m_0 = 2.75, \mu^2 = -1, \lambda = 1\}$



Unimodal target density

Reverse KL training approach

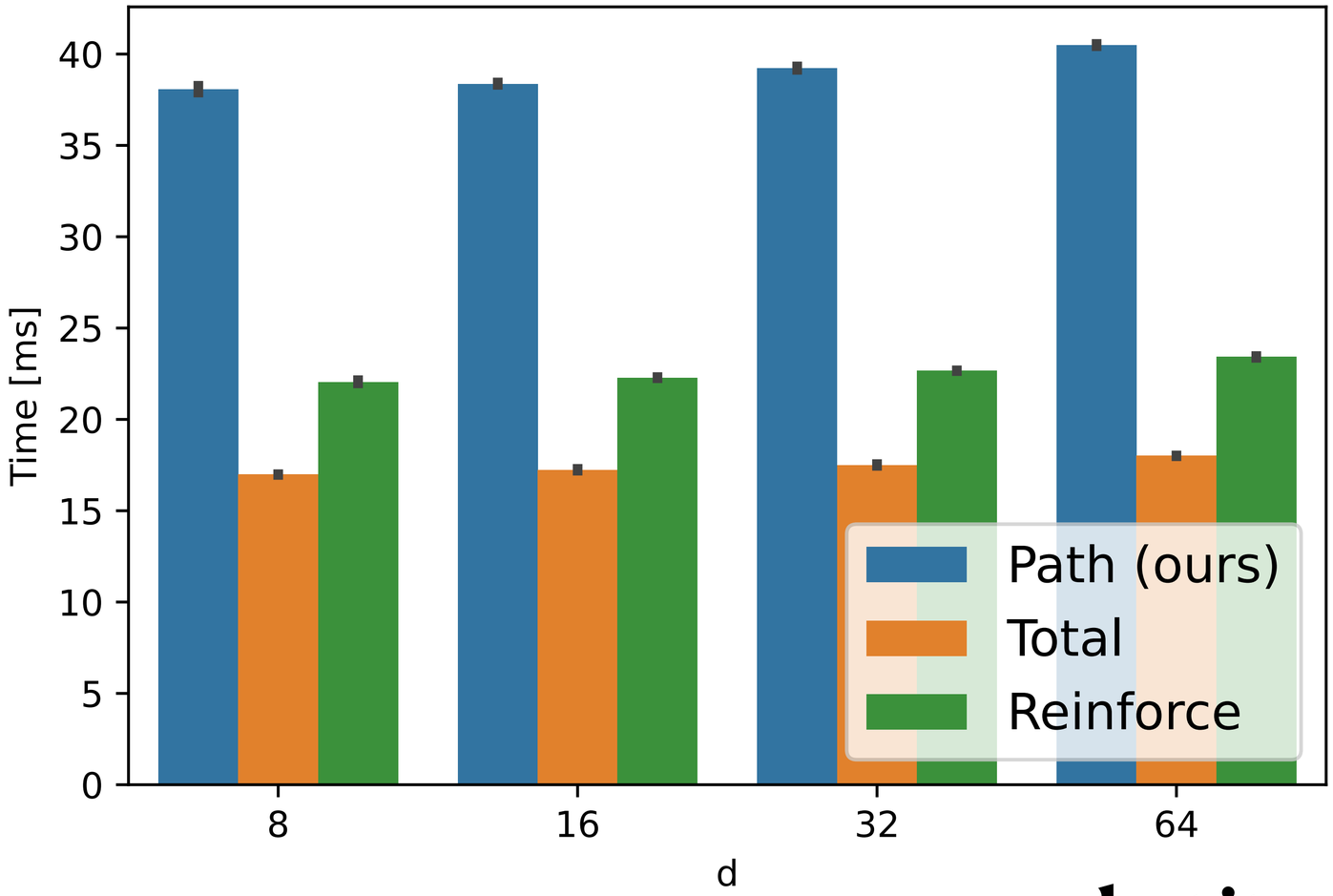
$\{m_0 = 3.0, \mu^2 = -1, \lambda = 1\}$



Multimodal target density

Forward KL training approach

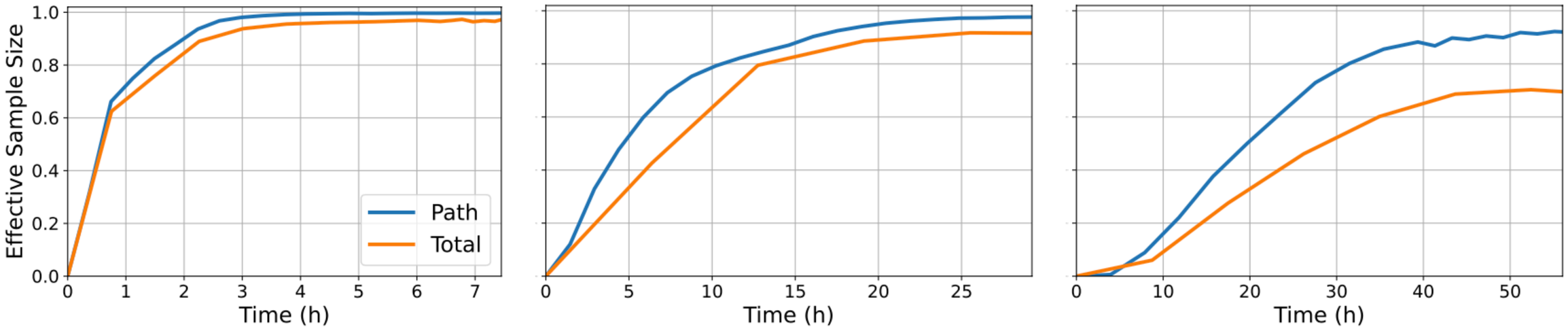
There’s no free lunch…



… but it can be cheap!

Continous Normalizing Flows for ϕ^4

12×12 20×20 32×32



$$\text{ESS} = \frac{N_{\text{eff}}}{N} \approx \frac{(\frac{1}{N}\sum_{i=1}^{N} \tilde{w}_i)^2}{\frac{1}{N}\sum_{i=1}^{N} \tilde{w}_i^2} \in [0,1]$$

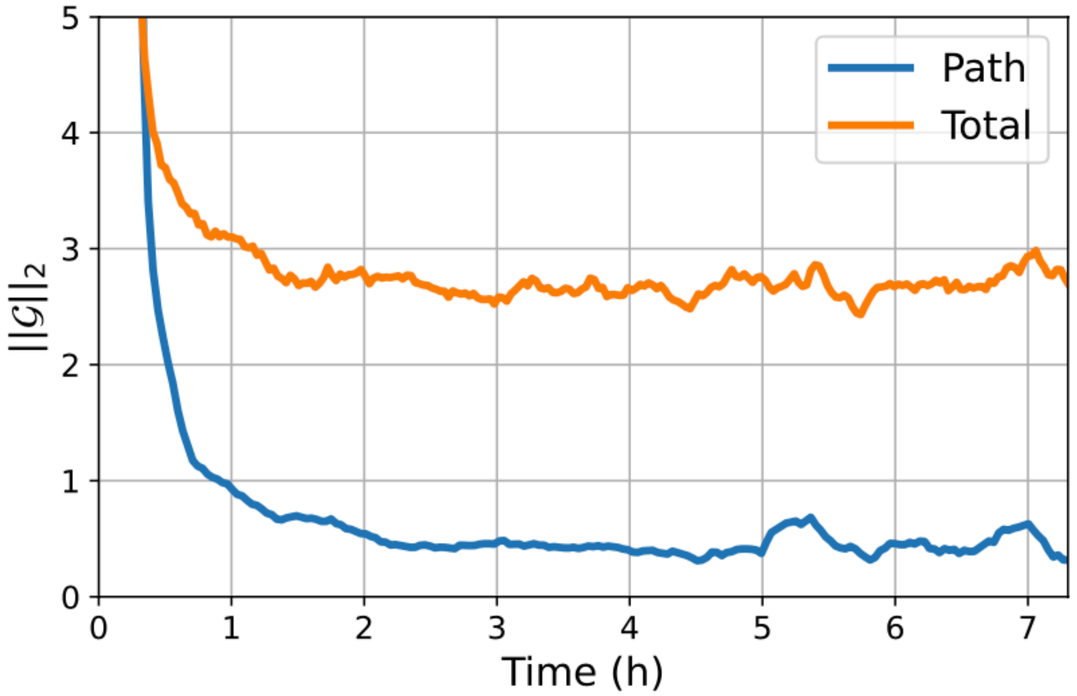
Lattice size	Path [2]	Total [1]
12x12	**99.66** $\pm$ 0.07	98.01 $\pm$ 0.44
20x20	**97.65** $\pm$ 0.14	91.56 $\pm$ 1.13
32x32†	**91.81** $\pm$ 1.32	69.53 $\pm$ 5.59

[1] de Haan et al., NeurIPS workshop (2021)

[2] Vaitl L., NKA, Nakajima S., Kessel P., ICML (2022)

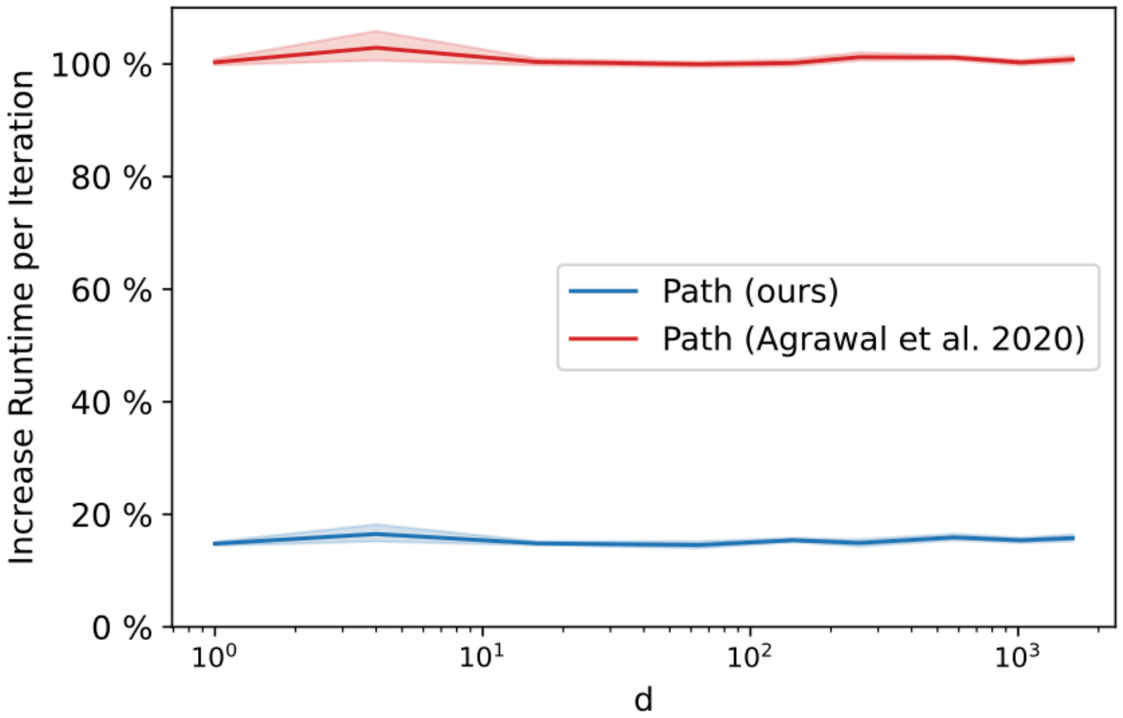
Path Gradient Estimators

Norm for the gradient estimator over training



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

Runtime comparison with naive path gradient



Lattice size	Time per Iteration
12x12	+14% $\pm$ 0
20x20	+12% $\pm$ 1
32x32	+5% $\pm$ 1

Path Gradient for CNFs: Technical Details

Continuous normalizing flows (Chen T.Q 2018) transform a sample $z_0 \sim q_Z$ from a base-density q_Z

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

Under the mild assumptions of the Picard–Lindelöf theorem, the map g_θ is bijective.

This map is implemented using a Neural Ordinary Differential Equation (NODE) of the form

$$\frac{dz_t}{dt} = f_\theta(z_t, t)$$

where:

- f_θ is a (**not necessarily** invertible) neural network with parameters θ.
- z_t is the intermediate state at time t.

Path Gradient for CNFs: Technical Details

The path gradient previously defined now becomes

$$\ln \left| \det \frac{dg_\theta}{dz_0} \right| = \int_0^T \operatorname{tr} \left(\frac{df_\theta(z_t, t)}{dz_t} \right) \mathrm{d}t$$

And the log variational density is thus given by

$$\ln q_\theta(\phi) = \ln q_Z(g_\theta^{-1}(\phi)) - \ln \left| \det \frac{dg_\theta}{dz_0} \right|_{z_0 = g_\theta^{-1}(\phi)} = \ln q_Z(z_0) - \int_0^T \operatorname{tr} \left(\frac{\partial f_\theta(z_t, t)}{\partial z_t} \right) \mathrm{d}t$$

Using the adjoint state method gradients can be computed with

- **Constant** memory requirements
- At the cost of 2 **additional** forward passes.

Path Gradient for CNFs: Technical Details

The path gradient previously introduced

$$\frac{d}{d\theta}\mathrm{KL}(q_\theta, p) = \mathbb{E}_{z\sim q_Z}\left[\blacktriangledown_\theta S(g_\theta(z)) + \blacktriangledown_\theta \log q_\theta(g_\theta(z))\right] = \mathbb{E}_{z\sim q_Z}\left[\frac{\partial}{\partial g_\theta(z)}\left(\ln \frac{q_\theta(g_\theta(z))}{p(g_\theta(z))}\right)\frac{\partial g_\theta(z)}{\partial \theta}\right]$$

Now becomes

$$\frac{d}{d\theta}\mathrm{KL}(q_\theta, p) = \left[\left(\frac{\partial}{\partial g_\theta(z)} \ln q_\theta(g_\theta(z)) - \frac{\partial}{\partial g_\theta(z)} \ln p(g_\theta(z))\right)\frac{\partial g_\theta(z)}{\partial \theta}\right]$$

$$\frac{\partial}{\partial z_T} \ln q_\theta(z_T)$$

This derivative is non-trivial to compute for NFs and requires additional care!

Path Gradient for CNFs: Technical Details

Theorem

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

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

with initial condition

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

Path Gradient Estimators

The only **non trivial term** to be computed for the path gradient is

$$\blacktriangledown_\theta \ln q_\theta(g_\theta(z)) = \frac{\partial \ln q_\theta(z_T)}{\partial z_T}^\intercal \frac{\partial z_T}{\partial \theta}$$

where $\blacktriangledown_\theta = \frac{d}{d\theta}$

and $z_T \equiv x = g_\theta(z_0)$

The **derivative of the log density** wrt the final state is computed using an ODE solved forward in time

- Compute the derivative $\frac{\partial \ln q_\theta(z_T)}{\partial z_T}$ solving **an initial value problem**
- with **initial condition** $\frac{\partial \ln q_\theta(z_0)}{\partial z_0} = \frac{\partial \ln q_Z(z_0)}{\partial z_0}$
- **Adjoint state method** (From CNF, Chen et al. 2018)

4 Forward passes

- 1 Initial value problem

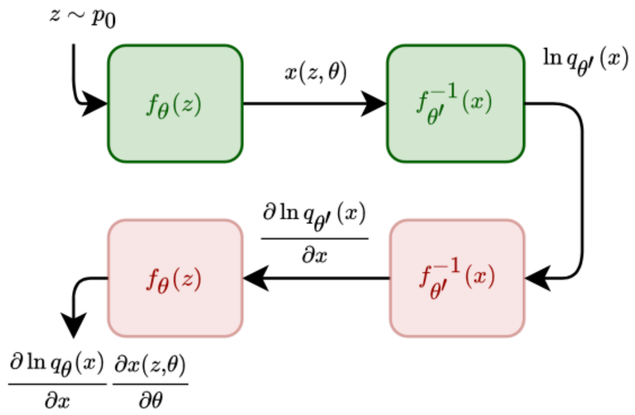
- 1 Forward integration
- 2 Adj. state method

Path Gradient Estimators

4 Forward passes

- 2 for forward mode CNF (adjoint state method for NODEs)
- 2 for Vector Jacobian product for Path gradient

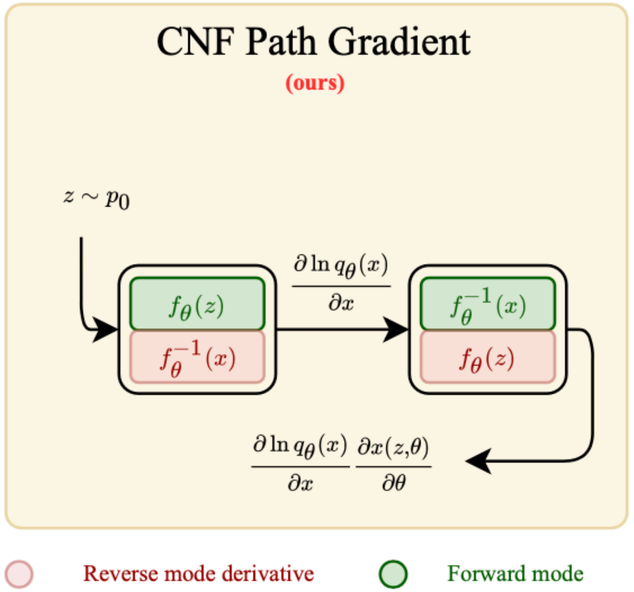
Naive Path Gradient
(Agrawal et al. 2020)



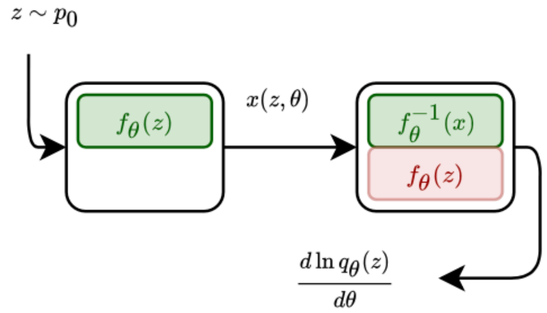
6 Forward passes

- Twice the memory footprint
- Considers 2 copy of the model for density estimation and sampling
- 2 forward passes
- 2 backward with the cost of 2 forward passes each

CNF Path Gradient
(ours)



CNF Total Gradient
(Chen et al. 2018)



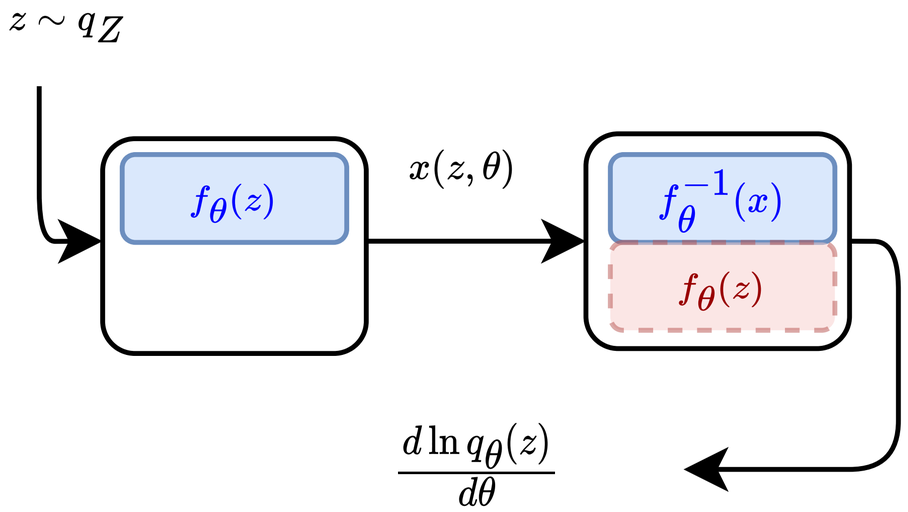
3 Forward passes

- 2 for forward mode (adjoint state method for NODEs)
- 1 for Total gradient

Path Gradient for CNFs: Technical Details

CNF Total Gradient

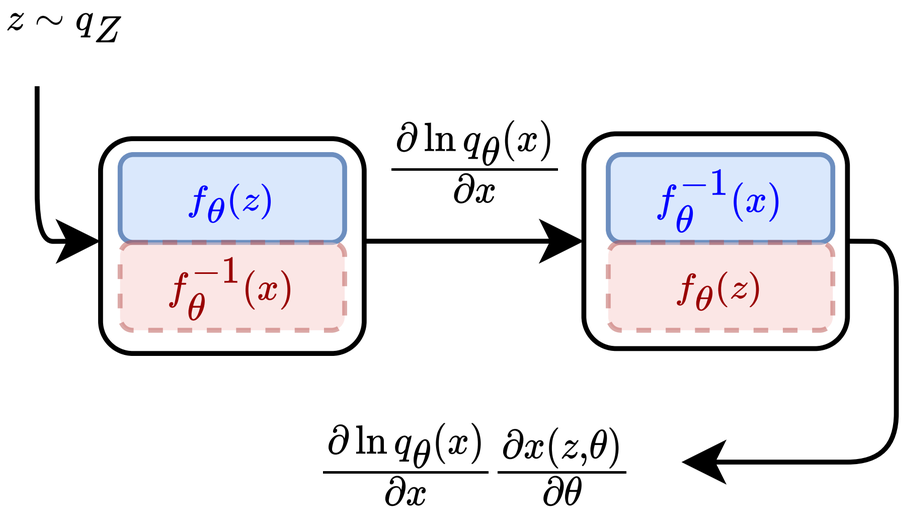
(Chen et al. 2018)



Reverse mode derivative Forward mode

CNF Path Gradient

(ours)



Reverse mode derivative Forward mode

Path Gradient Estimators

- We proposed a path-gradient estimator which can be efficiently implemented for continuous normalizing flows.
- We expect the path-gradient estimator to have slightly higher costs per iteration.
- We find that this effect is more than compensated by better convergence properties of the path-gradient.
- The difference in iteration costs decreases as the size of the lattice increases (quadratic scaling of model params with lattice size).
- A limitation of our work is that so far there is only a theoretical explanation for the lower variance of the path-gradient estimator in the final training phase.
- However, our experiments suggest that the entire training process benefits from the path-gradient estimator. Further investigation is needed.