UNIVERSITÄT BONN MATTER

RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN

TRANSDISCIPLINARY RESEARCH AREA

The quest for the lost mode: Detecting Mode-Collapse in Flow-Based Sampling for Lattice Field Theories

Kim A. Nicoli

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

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

Preliminaries: LQFT and HMC

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

The field configuration $\phi(x)$ is a **random variable** sampled with **MCMC** to estimate computed over a Boltzmann-like density:

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

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

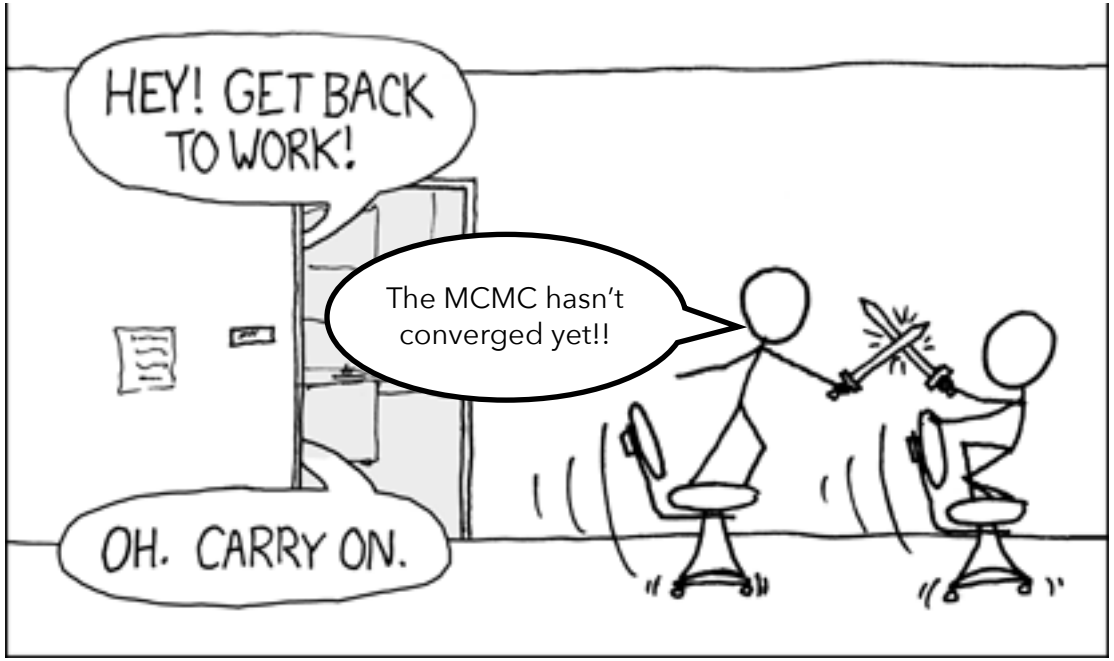
HMC and its flaws

UNIVERSITÄT BONN TRA MATTER
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
TRANSDISCIPLINARY RESEARCH AREA

MCMC: sequentially proposes new 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** $\Longrightarrow$ Phase transitions.
- 👎 Long-range autocorrelations $\Longrightarrow$ **large statistical errors.**
- 👎 The partition function Z **is unknown**.
- 👎 No direct estimation of **thermodynamic observables.**



(Adapted from:)

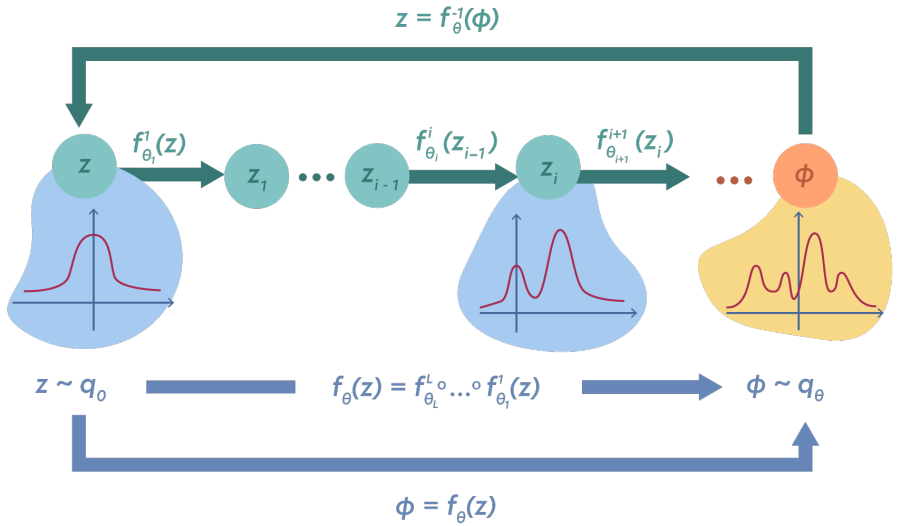
What about flow-based sampling?

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

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

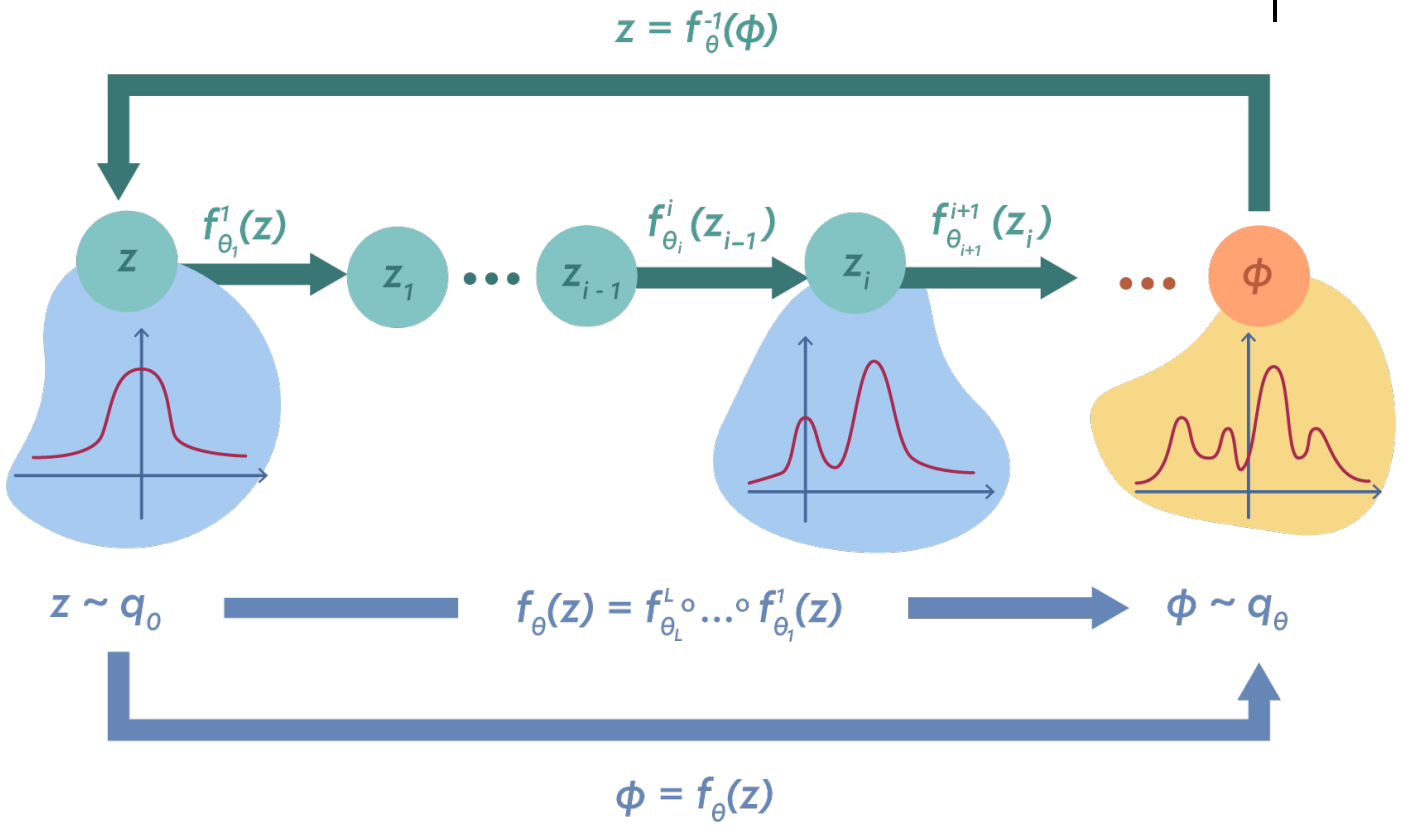
The parametric function needs to fulfill certain criteria:

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



What's flow-based sampling?

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



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?

UNIVERSITÄT BONN TRA MATTER
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
TRANSDISCIPLINARY RESEARCH AREA

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 $p(\phi) = Z^{-1}\exp\{-S(\phi)\}$

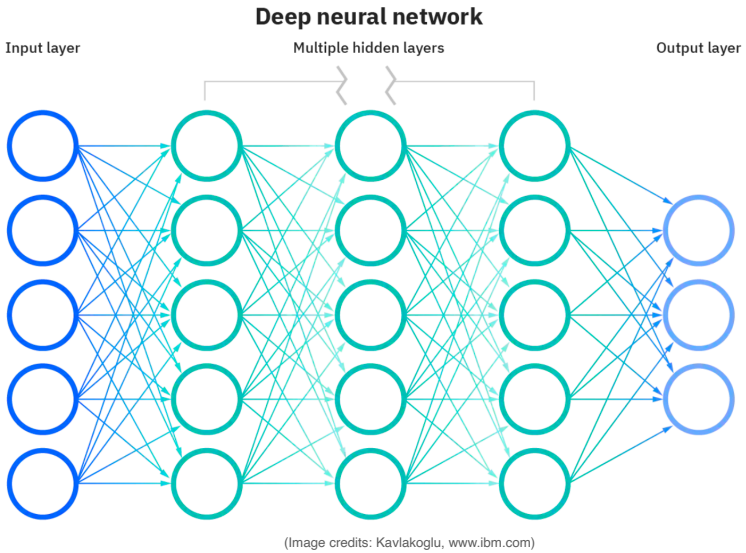
$$\nabla_\theta KL(q_\theta||p) = \mathbb{E}_{q_\theta}\left[\nabla_\theta \ln q_\theta(\phi) + \nabla_\theta S(\phi) + \cancel{\ln Z}\right]$$

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

How do we train a generative model?

Sample $\hat{\phi} \sim q_\theta \approx p$

$p(\phi)$



$q_\theta(\phi)$

Minimize the reverse-KL

How do we train a generative model?

$$p(\phi) \overset{?}{=} q_\theta(\phi)$$

How do we train a generative model?

$$p(\phi) \neq q_\theta(\phi)$$

How do we train a generative model?

UNIVERSITÄT BONN TRA MATTER
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
TRANSDISCIPLINARY RESEARCH AREA

$$p(\phi) \neq q_\theta(\phi)$$

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

Neural Importance Sampling (NIS)

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

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

Neural Importance Sampling (NIS)

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

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

Neural Importance Sampling (NIS)

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

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

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

Neural Importance Sampling (NIS)

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

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

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

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

Asymptotically Unbiased Estimators

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

Asymptotically Unbiased Estimators

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

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

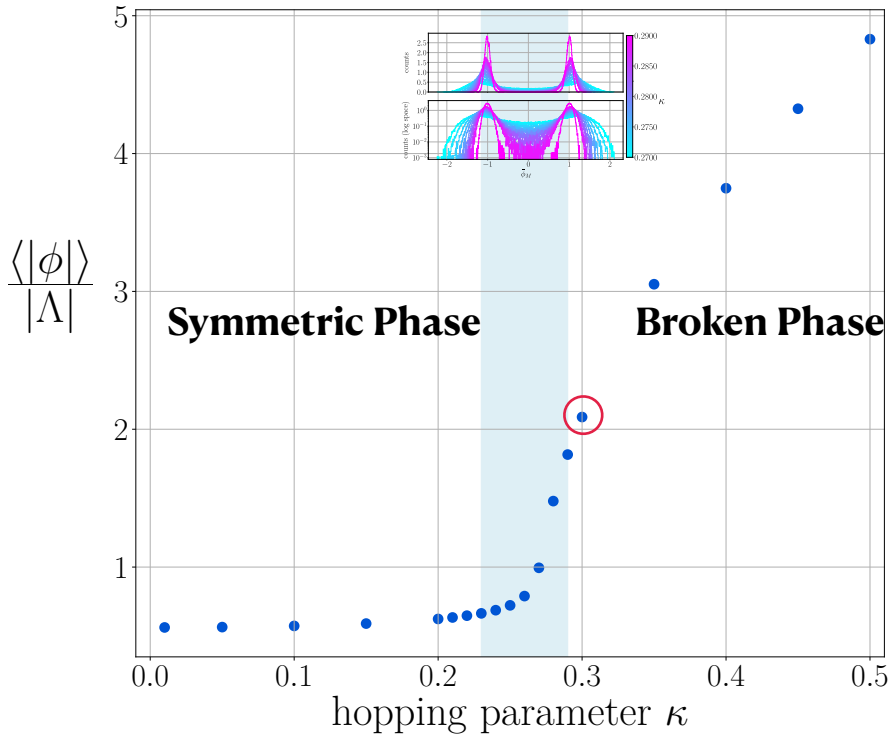
Asymptotically Unbiased Estimators

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

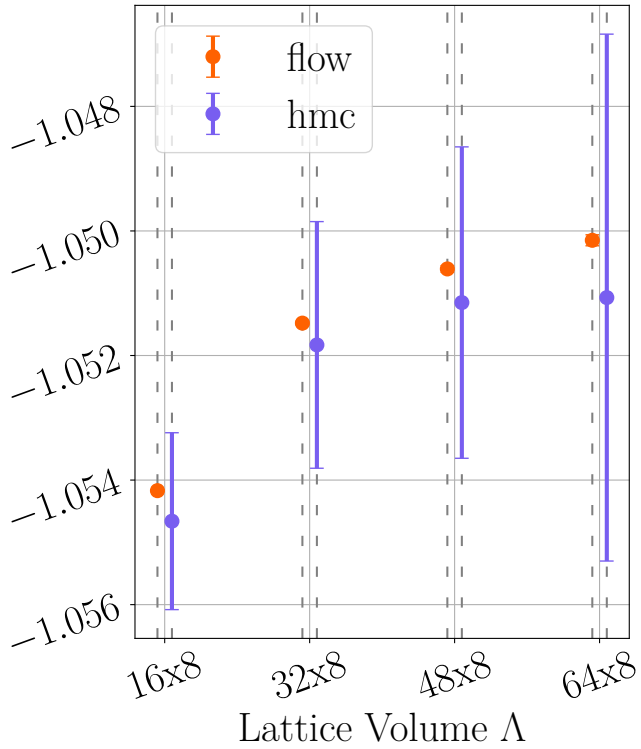
$$w(\phi) = \frac{p(\phi)}{q_\theta(\phi)}$$

Real Scalar ϕ^4-Theory in (1+1) D

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

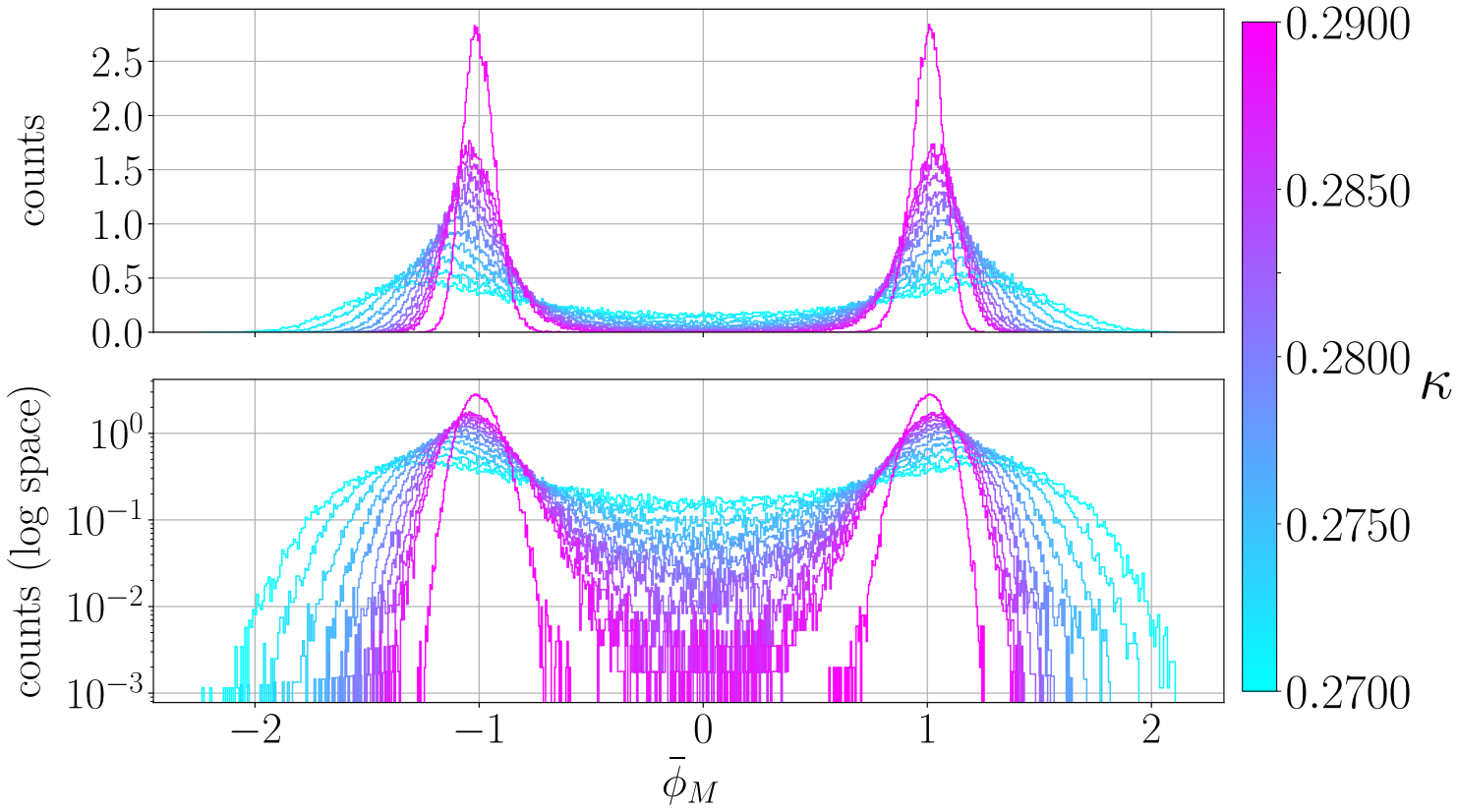


$\frac{F}{V}a^2$



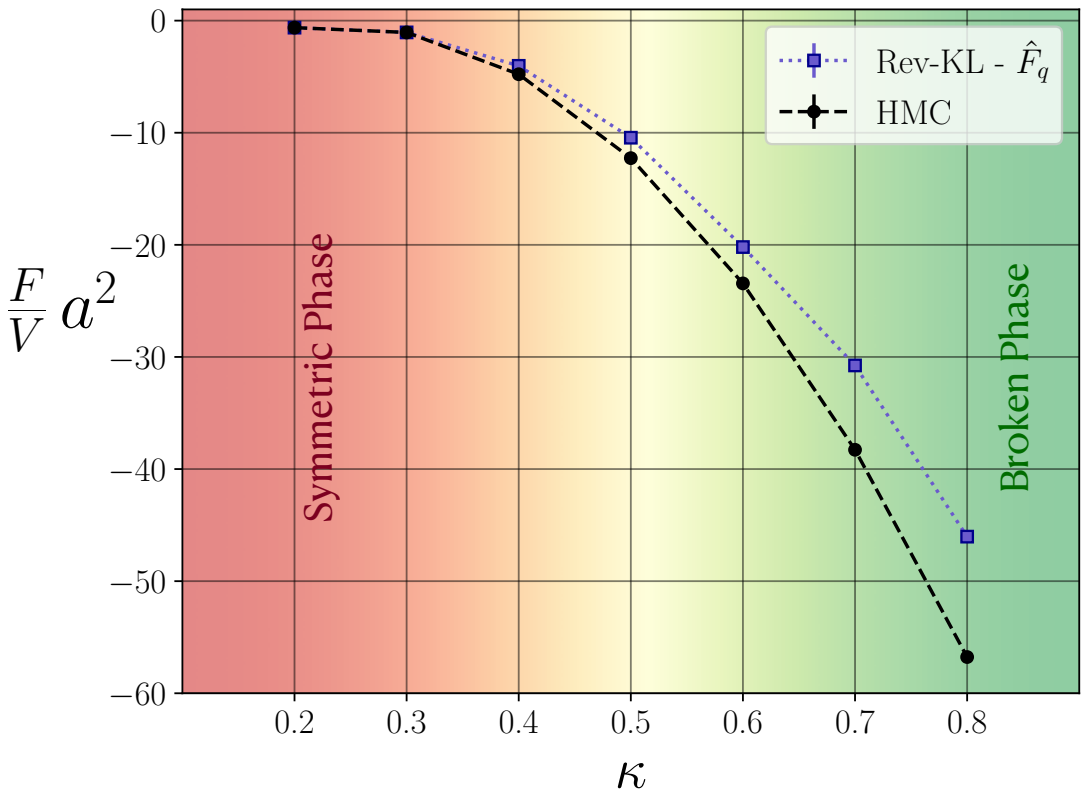
KAN, Anders, Funcke, et al., Phys. Rev. Lett. (2021)

Real Scalar ϕ^4-Theory in (1+1) D



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

Real Scalar ϕ^4-Theory in (1+1) D



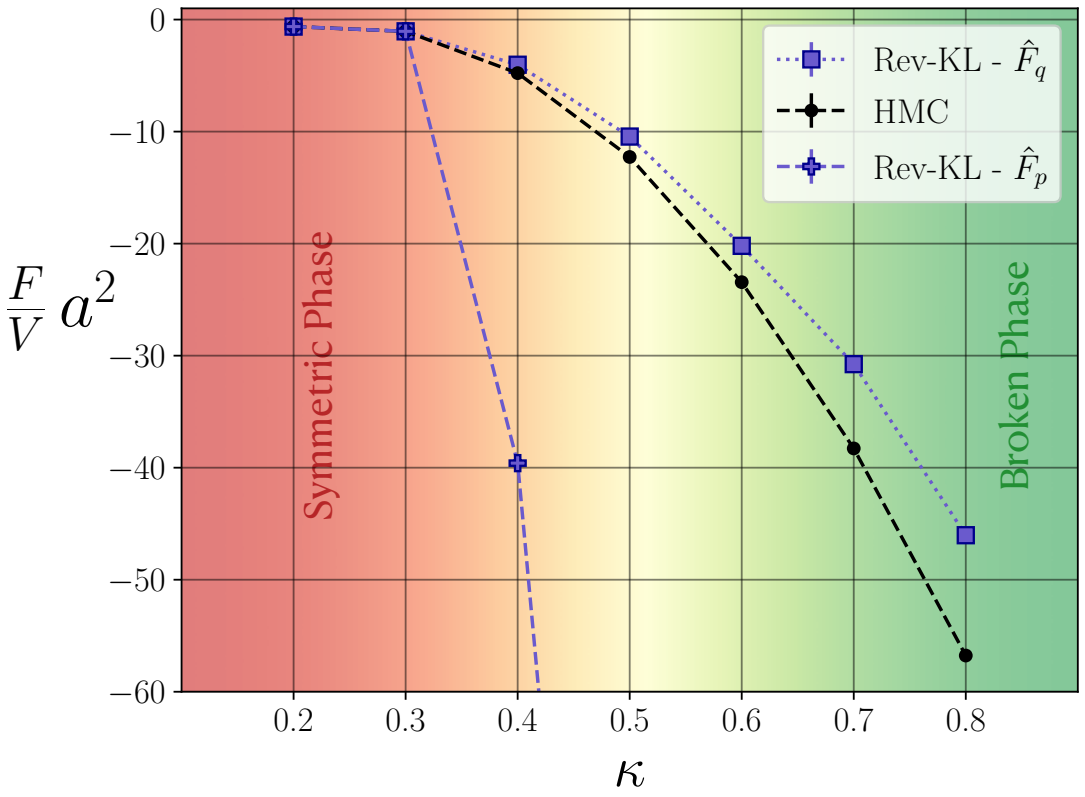
$$Z = \mathbb{E}_{\phi\sim q_\theta}\left[\frac{e^{-S(\phi)}}{q_\theta(\phi)}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{e^{-S(\phi_i)}}{q_\theta(\phi_i)} \equiv \hat{Z}_{q_\theta}$$

$$\phi_i \sim q_\theta$$

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

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

Real Scalar ϕ^4-Theory in (1+1) D



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

$\phi_i \sim q_\theta$

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

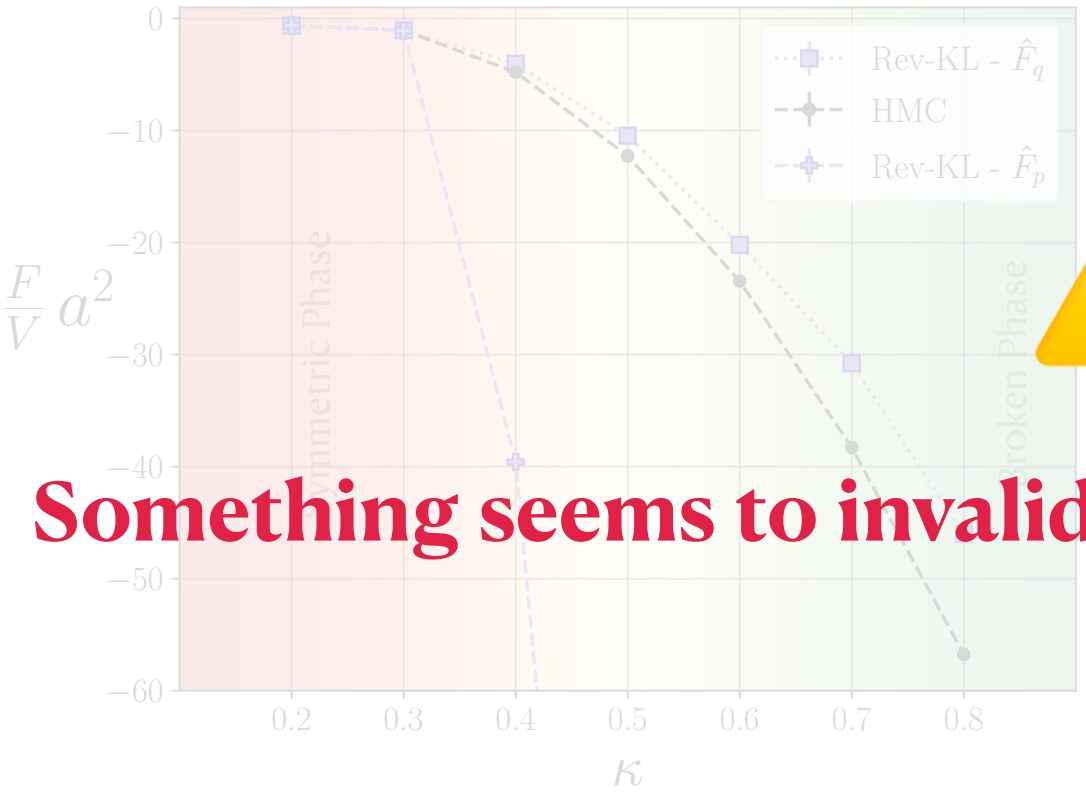
$$Z^{-1} = \mathbb{E}_{\phi\sim p}\left[\frac{q_\theta(\phi)}{e^{-S(\phi)}}\right] \approx \frac{1}{N}\sum_{j=1}^{N}\frac{q_\theta(\phi_j)}{e^{-S(\phi_j)}} \equiv \hat{Z}_p^{-1}$$

$\phi_j \sim p$

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

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

Real Scalar ϕ^4-Theory in (1+1) D



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

$$\phi_i \sim q_\theta$$

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

$$\left[\frac{e^{-S(\phi)}}{q_\theta(\phi)}\right] \approx \frac{1}{N}\sum_{j=1}^{N}\frac{e^{-S(\phi_j)}}{q_\theta(\phi_j)}$$

$$\phi_j \sim p$$

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

Something seems to invalidate our asymptotic guarantees!

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

What's going wrong then?

UNIVERSITÄT BONN TRA MATTER
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
TRANSDISCIPLINARY RESEARCH AREA

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?

UNIVERSITÄT BONN TRA MATTER
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
TRANSDISCIPLINARY RESEARCH AREA

Reverse-KL Div.

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

Forward-KL Div.

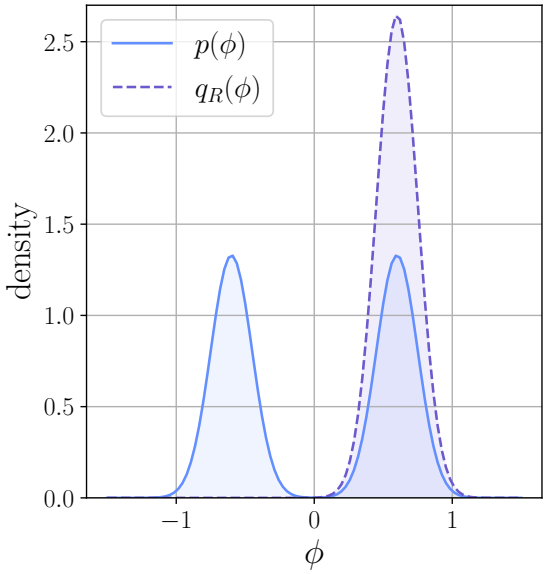
$$KL_F(p||q_\theta) = \int D[\phi] p(\phi) \ln \frac{p(\phi)}{q_\theta(\phi)}$$

What's going wrong then?

UNIVERSITÄT BONN TRA MATTER
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
TRANSDISCIPLINARY RESEARCH AREA

Reverse-KL Div.

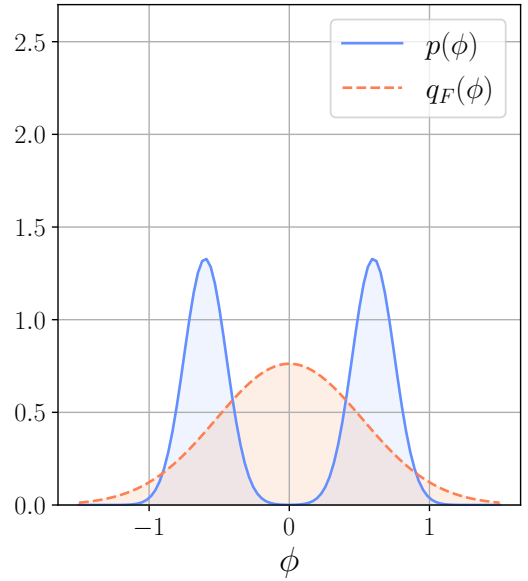
$$KL_R(q_\theta||p) = \int D[\phi] q_\theta(\phi) \ln \frac{q_\theta(\phi)}{p(\phi)}$$



- Self-Sampling (efficient).
- No need for training data.
- Mode-dropping.

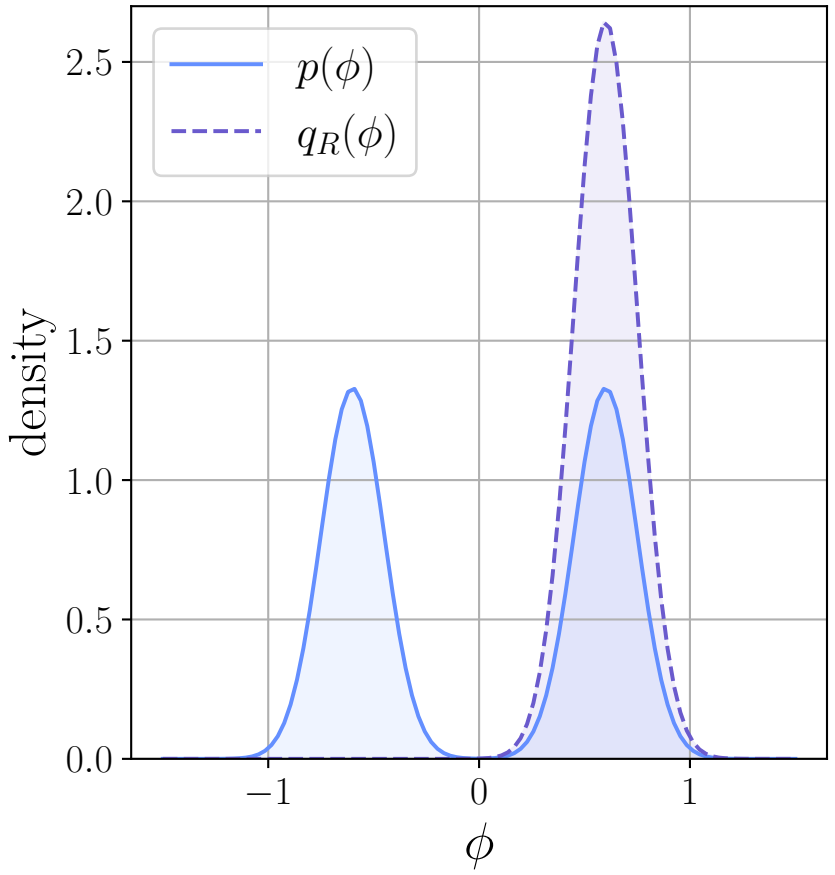
Forward-KL Div.

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



- Maximum Likelihood.
- Requires training data.
- Fake modes.

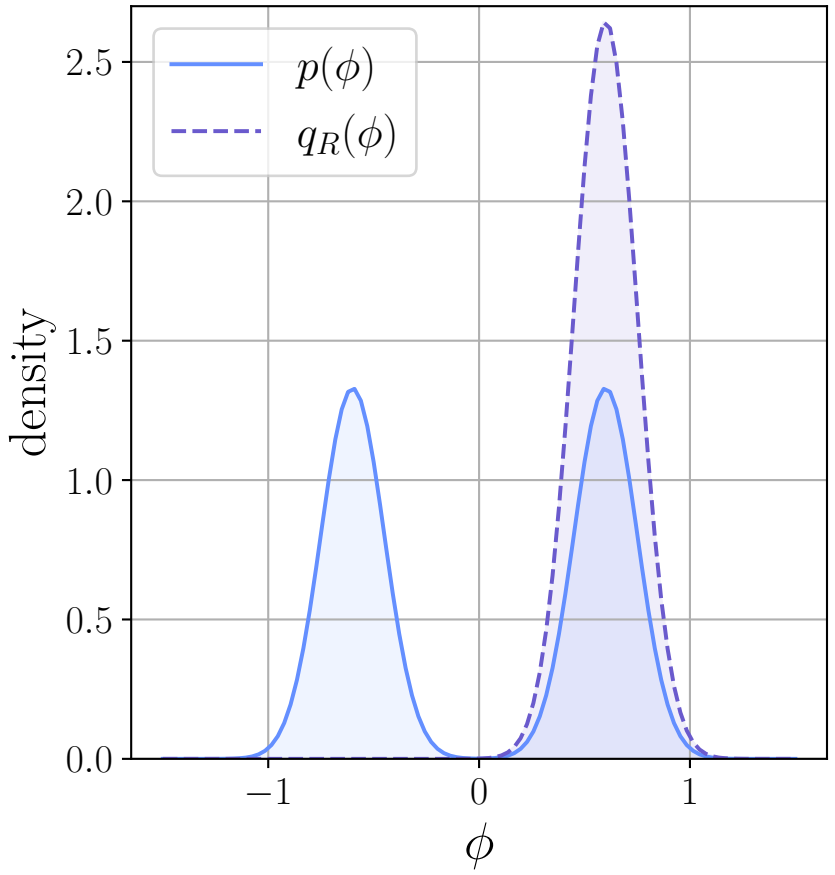
What's going wrong then?



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

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

What's going wrong then?



ESR is not a good metric!

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

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 effect of mode-dropping

UNIVERSITÄT BONN TRA MATTER
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
TRANSDISCIPLINARY RESEARCH AREA

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! 👎

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

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

The mode-dropping estimator

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

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

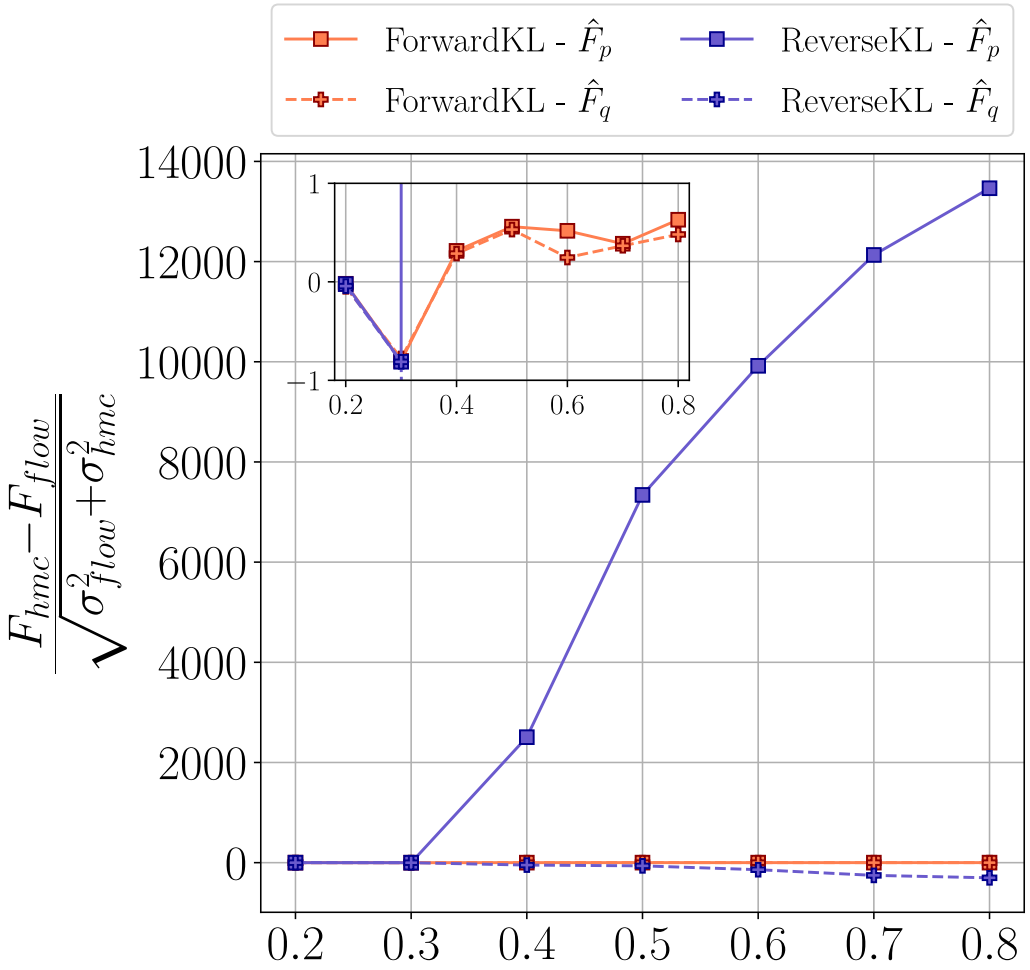
however if q_θ is **effectively mode-dropping** this expectation value becomes

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

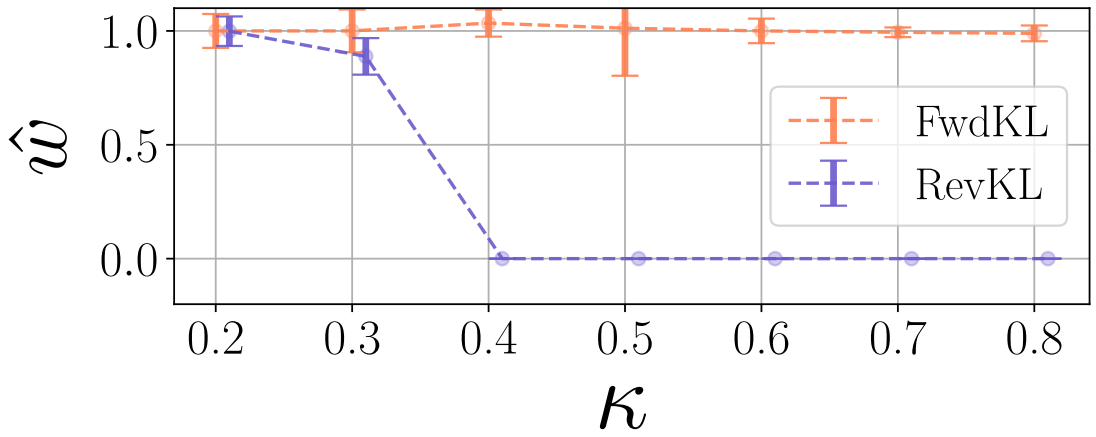
for which we can get the corresponding Monte Carlo estimator, i.e., the **mode-dropping estimator**

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

Estimation of Mode Dropping Across Criticality



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



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

Summary and Conclusions

UNIVERSITÄT BONN TRA MATTER
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
TRANSDISCIPLINARY RESEARCH AREA

i) **Asymptotically unbiased samplers** can be constructed from trained DGMs (NIS or NMCMC).

ii) **Direct** estimation of the **partition function** and **thermodynamic observables**.

iii) Sampling from DGMs is **embarrassingly parallelizable** ≠ MCMC (**sequential**).

iv) Training with **forward-KL** leads to better models though requires training samples.

v) Derivation of **mode-dropping estimator** to reliably assess the goodness of the model.

vi) **Mitigation** of mode-dropping using different **objectives** (FWD-KL) or **stochastic** approaches (SNFs).

Thank you for your attention!

(… and Merry Xmas! 🎅)