SCALING OF ANNEALING METHODS FOR LATTICE FIELD THEORY SAMPLING

Satria Widyanto

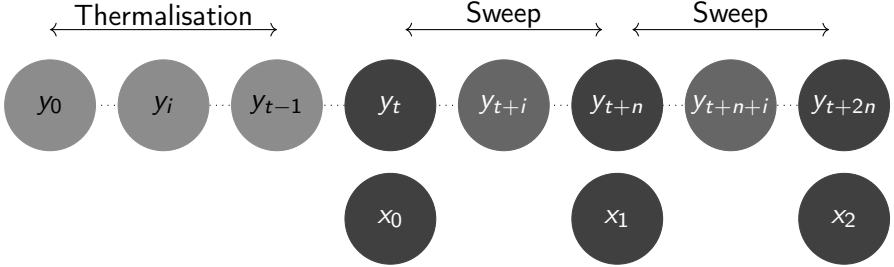
Gurtej Kanwar, Elia Cellini

May 18, 2026

Topics

1. Introduction and Motivations
2. Annealed Importance Sampling
3. Non Equilibrium Transport Sampler
4. Scaling Studies

- Objective: Generating x_i such that $x_i \sim P(x_i)$, with $P(x_i) \propto e^{-S(x_i)}$.
- Typical means: Markov Chain Monte Carlo (MCMC).



- Problem: Autocorrelation grows as we go to finer lattice. $\tau_{\text{int}} \propto \xi^z$.
- For topological observables, could get to $\tau_{\text{int}} \propto \exp(a\xi^z)$ (Del Debbio et al. 2004, hep-lat/0403001) or ξ^z with $z > 2$ (Schaefer et al. 2011, 1009.5228).

- Open Boundary Condition (Luscher and Schaefer 2011, 1105.4749) + Parallel Tempering (Hasenbusch 2018, 1709.09460)
- Annealing (Bonanno et al. 2024, 2402.06561; Syed et al. 2025, 2408.12057)
- Normalizing Flows (M. S. Albergo et al. 2019, 1904.12072; Caselle et al. 2023, 2210.03139)

- Multiple importance sampling + Markov chain

$$w(x^N, x^0) = \frac{P_N(x^N)}{P_{N-1}(x^{N-1})} \cdots \frac{P_2(x^2)}{P_1(x^1)} \frac{P_1(x^1)}{P_0(x^0)}, \qquad x^k = T^k(x^{k-1})$$
$$\Delta F = -\ln(\langle w \rangle)$$

- Choose P_k such that it is close to P_{k-1}, and T_k such that it can closely transport between the two distributions.

- Can add neural network in between $\implies$ Stochastic Normalizing Flow (Wu et al. 2020, 2002.06707; Caselle et al. 2023, 2210.03139).

- Consider $x_i \sim P_0(x) = \exp(-U_0(x))$. Performing 'continous' AIS with Langevin noise

$$dx = -\varepsilon \partial_x U \; dt + \sqrt{2\varepsilon} \; d\eta \qquad ; \qquad x(t) \sim w(t)^{-1} \exp(-U_t)$$
$$dW = \partial_t U \; dt \qquad ; \qquad w(t) = \exp(-W(t))$$

- Consider $x_i \sim P_0(x) = \exp(-U_0(x))$. Performing 'continous' AIS with Langevin noise

$$dx = -\varepsilon \partial_x U \; dt + \sqrt{2\varepsilon} \; d\eta + b_x \; dt \; ; \qquad x(t) \sim w(t)^{-1} \exp(-U_t)$$
$$dW = \partial_t U \; dt + b_x \partial_x U \; dt - \partial_x b_x \; dt; \qquad w(t) = \exp(-W(t))$$

- Add drift term b_x to reduce variance in W (Vaikuntanathan and Jarzynski 2008, 0804.3055).
- Learn the drift using neural network (Vargas et al. 2025, 2307.01050; Michael S. Albergo and Vanden-Eijnden 2025, 2410.02711)

- Numerical integration

$$\Delta x^k = -\varepsilon \Delta t \ \partial_x U^k + \sqrt{2\varepsilon \Delta t}\ \eta^k + \Delta t \ b_x^k$$
$$\Delta W^k = \Delta t \ \partial_t U^k + \Delta t \ b_x^k \partial_x U^k - \Delta t \ \partial_x b_x^k$$

- Not necessarily unbiased w.r.t. $w = \exp(-W)$.
- To get unbiased estimate of w (Crooks 1999, cond-mat/9901352)

$$\frac{w^{k+1}}{w^k} = \frac{\exp\left(-U^{k+1}(x^{k+1})\right)}{\exp(-U^k(x^k))} \exp\left(\frac{1}{2}|\eta^k|^2 - \frac{1}{2}|\tilde{\eta}^k|^2\right)$$
$$\Delta W^k = U^{k+1}(x^{k+1}) - U^k(x^k) - \frac{1}{2}|\eta^k|^2 + \frac{1}{2}|\tilde{\eta}^k|^2$$
$$\tilde{\eta}^k = \sqrt{\frac{\Delta t}{2\varepsilon}}[b_x^k(x^{k+1}) - b_x^k(x^k)] + \sqrt{2\varepsilon \Delta t}[\partial_x U^k(x^{k+1}) + \partial_x U^k(x^k)] - \eta^k$$

Loss Objective

- Objective: Minimise variance on W.
- One natural choice of loss: Physics-informed Neural Network (PINN) loss

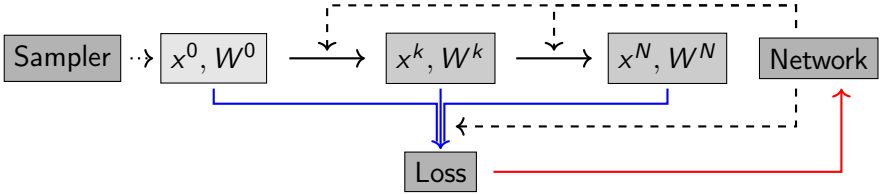
$$\mathcal{L}_{\text{PINN}} = \int dt \ \left\langle |d_t W|^2 \right\rangle$$

- $\exp(-\langle W \rangle) \leq \langle \exp(-W) \rangle \implies \langle W \rangle \geq -\ln(\langle \exp(-W) \rangle)$.
- Another choice: $\mathcal{L}_W = \langle W \rangle$

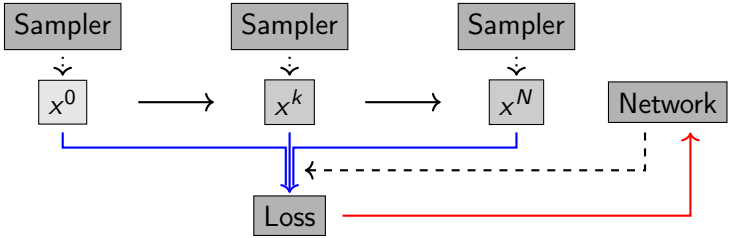
Algorithmic Comparison

Discrete NF (1904.12072)	Continuous NF (2207.00283)	AIS	Stochastic NF (2210.03139)	NETS (2410.02711)
Deterministic	Deterministic	Stochastic	Stochastic	Stochastic
Learnable Path	Learnable Path	Fixed Path	Fixed Path	Fixed Path
Include NN	Include NN	No NN	Include NN	Include NN
Discrete	Continuous	Discrete	Discrete	Continuous
Det(J)	Divergence	Work	Work + Det(J)	Work /+ Det(J)

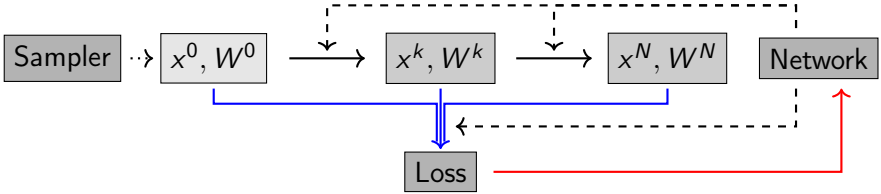
- Forward training



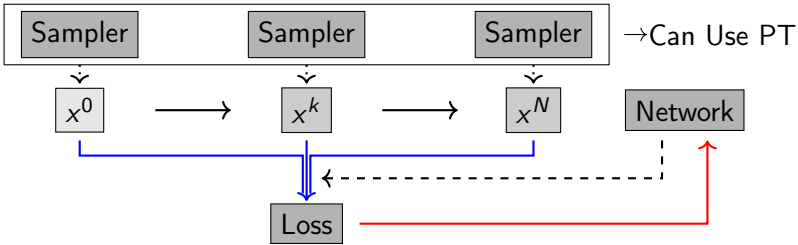
- Supervised learning



- Forward training



- Supervised learning



- Two tunable parameters: Δt (or in general, annealing schedule) and ε.
- Main proxies

$$\begin{aligned}\mathrm{var}(W) &= \left\langle W^2 \right\rangle - \langle W \rangle^2 \\ D_{KL} &= \langle W \rangle + \ln\left(\left\langle e^{-W} \right\rangle\right) = \langle W_d \rangle \\ \mathrm{ESS} &= \left\langle e^{-W} \right\rangle^2 / \left\langle e^{-2W} \right\rangle\end{aligned}$$

- How do they scale with lattice size?

- Position space action $S = \sum_x \left[(2D + m^2)\phi_x^2 - \sum_\mu 2\phi_x\phi_{x+\mu}\right]$.
- Momentum space action
  $S = \sum_k \left[m^2 + 4\sum_\mu \sin^2\left(\frac{\pi}{L}k_\mu\right)\right]\tilde{\phi}_k\tilde{\phi}_{-k}; \qquad \tilde{\phi}_{-k} \equiv \tilde{\phi}_{L-k} = \tilde{\phi}_k^*.$
- $\implies$ Independent Gaussian distributions $\implies$ Can be solved analytically.

$$\Delta\tilde{\phi}_k^i = \Delta t\left[\tilde{b}_{\tilde{\phi}_k} - 2\varepsilon M_k^i\tilde{\phi}_k\right] + \tilde{\eta}_k\sqrt{2\varepsilon\Delta t}$$
$$\Delta W_k^i = \Delta t\left[2\tilde{b}_{\tilde{\phi}_k}M_k^i\tilde{\phi}_{-k} + \partial_t m_i^2\tilde{\phi}_k\tilde{\phi}_{-k} - \partial_{\tilde{\phi}_k}\tilde{b}_{\tilde{\phi}_k}\right]$$
$$M_k^i = m_i^2 + 4\sum_\mu \sin^2\left(\frac{\pi}{L}k_\mu\right)$$

- Set $\tilde{b} = 0,\ m^2(t) = (1-t)\ m_0^2 + t\ m_1^2$.
- Set $\varepsilon = \text{const.},\ \Delta t \equiv 1/N \to 0$.

$$\Delta\tilde{\phi}_k^i = -2\varepsilon\Delta t\ M_k^i\tilde{\phi}_k + \tilde{\eta}_k\sqrt{2\varepsilon\Delta t} \implies \tilde{\phi}_k^i \simeq \sum_{a=0}^{i}\tilde{\eta}_k^a f_{a,i}(\varepsilon,k)$$

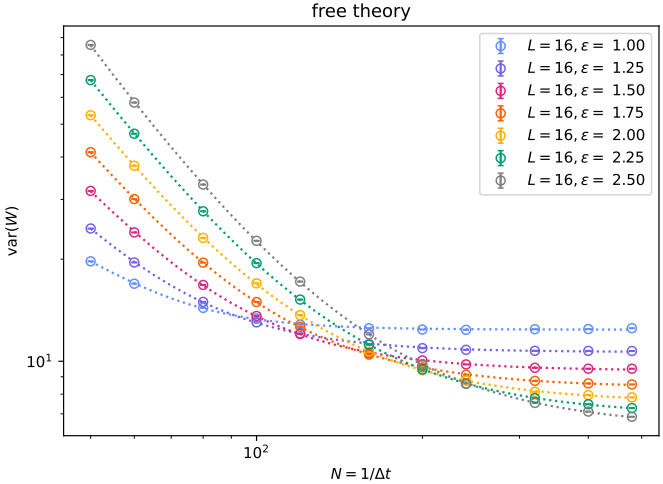
$$\Delta W_k^i = -\frac{(m_0^2-m_1^2)}{N}\tilde{\phi}_k^i\tilde{\phi}_{-k}^i \qquad \implies W_k \simeq \frac{(m_0^2-m_1^2)}{N}\sum_i\sum_{a,b}\tilde{\eta}_k^a\tilde{\eta}_k^b\ f_{a,i}(\varepsilon,k)f_{b,i}(\varepsilon,k)$$

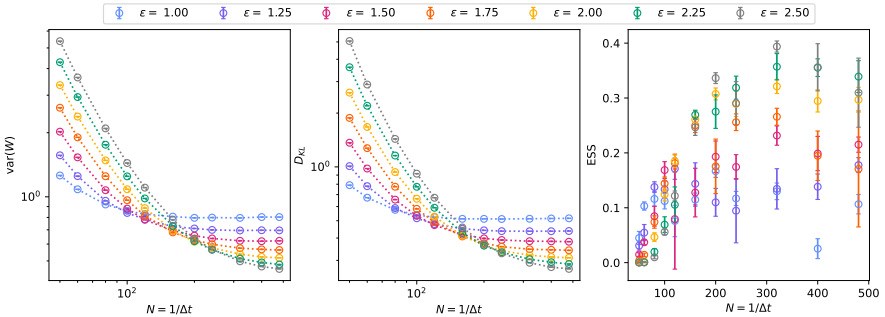
$$\text{var}(W_k) \simeq \frac{(m_0^2-m_1^2)}{N}F(\varepsilon)\sum_i G(i,k)$$

$$\text{var}(W) \simeq F(\varepsilon)\int_{m_0^2}^{m_1^2} dm^2\left(\frac{L^D}{\pi^D}\prod_i^D\int_0^\pi dk_i\right)\ G(m^2,k_1,k_2)$$

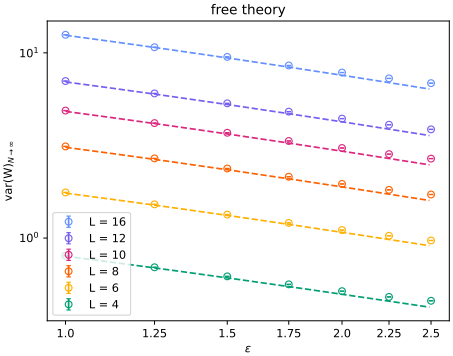
Scaling Studies

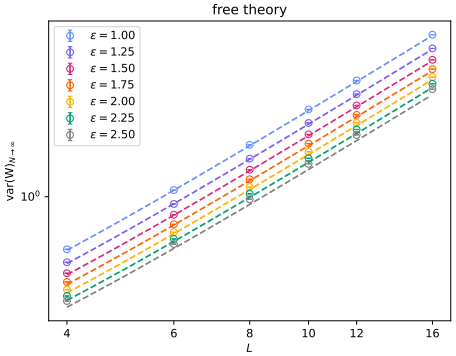
Free Scalar Theory, $D = 2, m_0^2 = 10.0, m_1^2 = 1.0$





Metrics comparison for $L = 4$





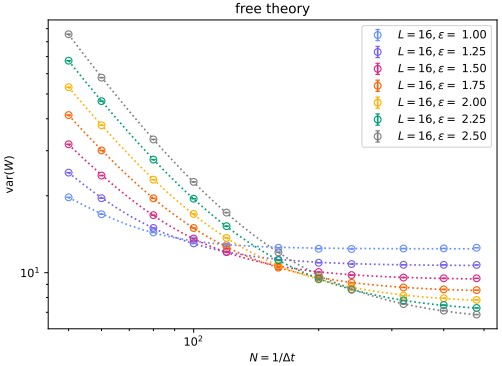
Scaling of variance as $N \to \infty$ w.r.t ε and L. We obtain scaling of $\sim L^2$.

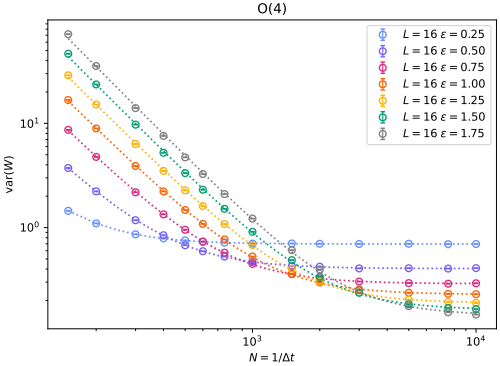
- Not all interesting stuff happens on fields with $\mathbb{R}^D$.
- Obvious example: QCD gauge fields live on SU(3) manifold.
- Other non-trivial manifolds: O(N), CP^{N-1}.
- We can construct NETS to suit these non-trivial manifold dynamics.
- Let us pick O(N) as an example, with $N = 4$.

$$S = -\beta \sum_{<xy>} s_x^a s_y^a; \qquad s_x^a s_x^a = 1$$

- We evolve β from $\beta_0 = 1.0$ to $\beta_1 = 1.1$.

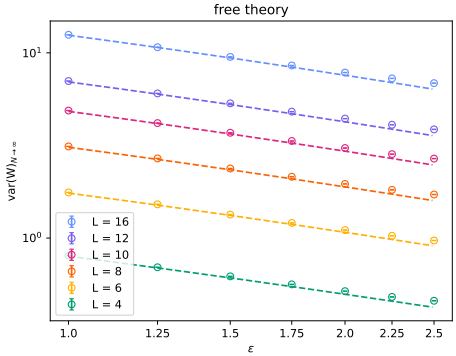
Scaling Studies
Free Scalar vs O(4)

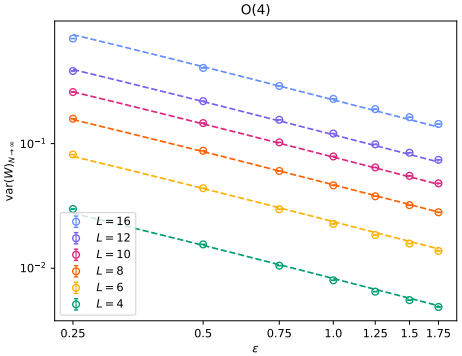


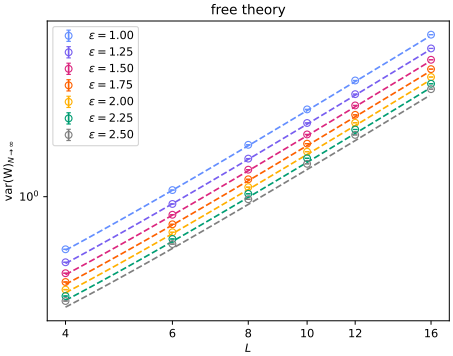


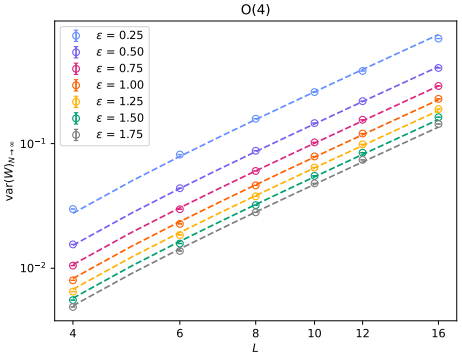
Both theory's metric converges as $N \to \infty$

Scaling Studies
Free Scalar vs O(4)







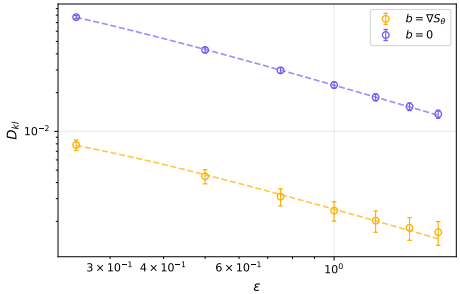


$\sim L^{2.0}$ for free theory, and $\sim L^{2.2}$ for O(4).

Summary

- NETS mixes stochasticity, continuity, and neural network into one framework, allowing unbiased and learnable approach for sampling lattice configurations.
- NETS can be extended to non-Euclidean manifold.
- Some scaling relations can be studied analytically on free scalar theory, and the numerical results is carried over in O(N) model.

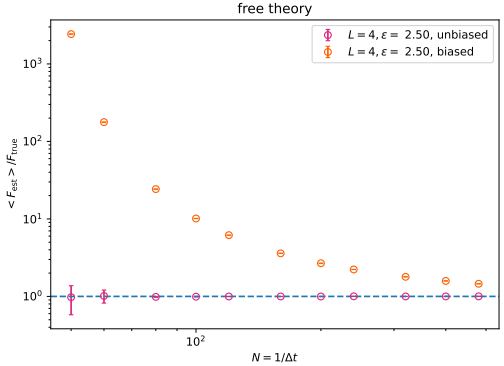
- Turning on learnable flow!
- Exploring various network architectures: energy-based network, CNN, Unet, etc.
- Large scale study: parallel tempering, long run, more scaling, etc.
- Finding empirical balance between memory and time complexity, hyperparameters, etc.

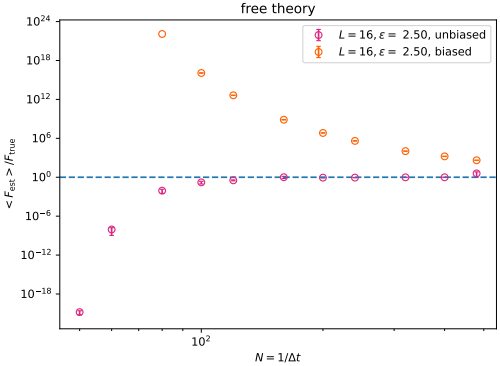


END
THANK YOU!

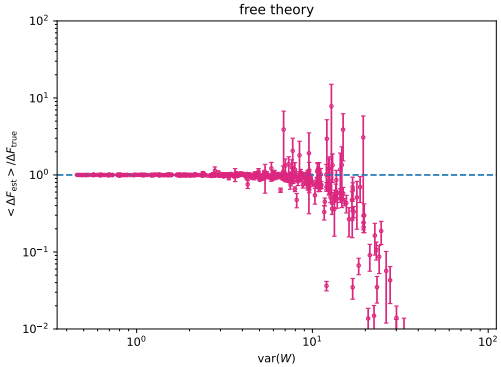
Backup Slides

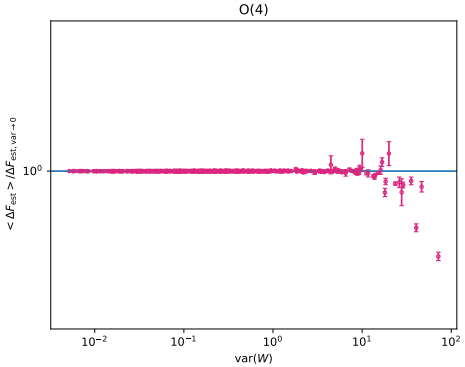
Biased vs Unbiased Integration





Unbiased Integrator is unbiased, provided that ...

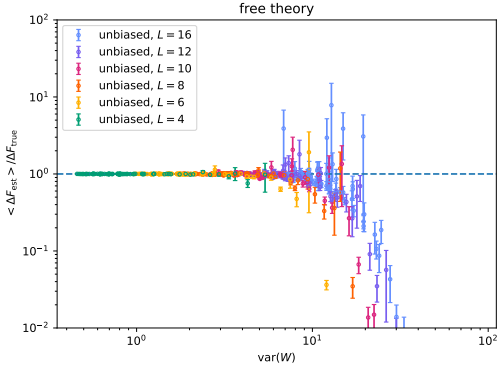


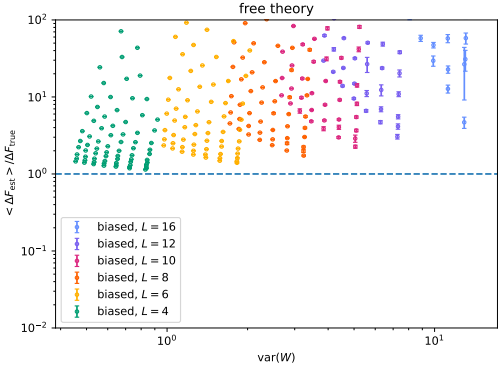


Left: free scalar, right: O(4)

Backup Slides

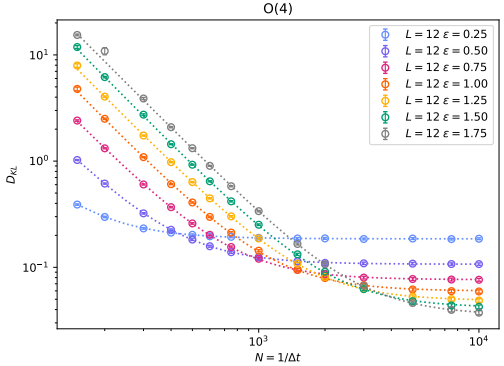
Let us compare with the biased integrator

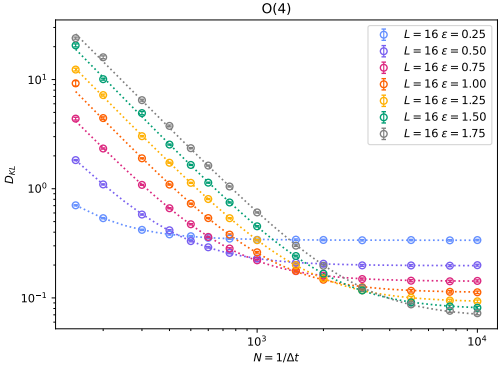




Backup Slides

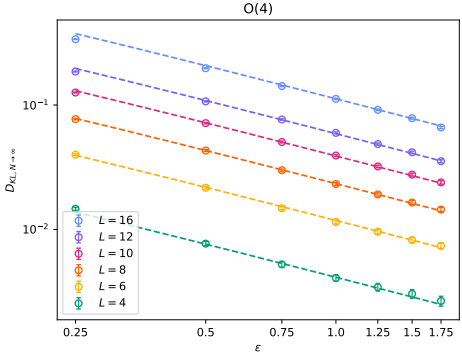
O(4) D_{KL} plots

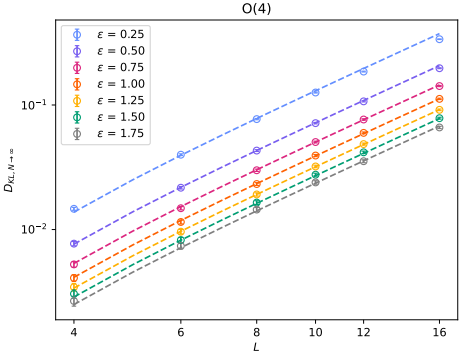




Backup Slides

O(4) D_{KL} plots





Fit result for variance vs ε and L

$$\mathrm{var}(W)_{\mathrm{free}} \simeq 0.064 \times \frac{L^{2.03} + 0.481}{\varepsilon^{0.94} + 0.426} + 0.032$$

$$\mathrm{var}(W)_{\mathrm{O}(4)} \simeq 5.13 \times 10^{-4} \times \frac{L^{2.21} - 4.720}{\varepsilon^{0.92} + 0.029}$$

$$\langle \exp(-W) \rangle = 1 - \langle W \rangle + \frac{1}{2} \langle W^2 \rangle + ...$$

$$\ln(\langle \exp(-W) \rangle) = \ln\left(1 - \langle W \rangle + \frac{1}{2} \langle W^2 \rangle + ...\right)$$

$$= - \langle W \rangle + \frac{1}{2} \langle W^2 \rangle + ... - \frac{1}{2} \left(- \langle W \rangle + \frac{1}{2} \langle W^2 \rangle + ...\right)^2 + ...$$

$$= - \langle W \rangle + \frac{1}{2} \langle W^2 \rangle - \frac{1}{2} \langle W \rangle^2 + ...$$

$$\langle W \rangle + \ln(\langle \exp(-W) \rangle) = \frac{1}{2} \mathrm{var}(W) + ...$$

- In general

$$\tilde{\phi}^i = \sum_{a=0}^{i} \tilde{\eta}^a c^a \prod_{r=a}^{i} \left(1 - \frac{2\varepsilon}{N} M^r\right); \quad c^0 = \frac{1}{2M^0}, c^{a\neq 0} = \sqrt{\frac{2\varepsilon}{N}}$$

- Provided N sufficiently large

$$\begin{aligned}
\tilde{\phi}^i &= \sum_{a=0}^{i} \tilde{\eta}^a c^a \prod_{r=a}^{i} \exp\left(-\frac{2\varepsilon}{N} M^r\right) \\
&= \sum_{a=0}^{i} \tilde{\eta}^a c^a \exp\left(-2\varepsilon \sum_{r=a}^{i} \frac{1}{N} M^r\right) \\
&= \sum_{a=0}^{i} \tilde{\eta}^a c^a \exp(-2\varepsilon \mathcal{M}_i^a)
\end{aligned}$$

- The work now becomes

$$W^i = -\frac{(m_0^2 - m_1^2)}{N} \sum_{j=0}^{i} \sum_{a=0}^{j} \sum_{b=0}^{j} \tilde{\eta}^a \tilde{\eta}^b c^a c^b \exp\Big(-2\varepsilon[\mathcal{M}_j^a + \mathcal{M}_j^b]\Big)$$

- Focusing only on ab terms and $a, b > 0$

$$\begin{aligned} W^i &= -2\frac{(m_0^2 - m_1^2)}{N} \sum_{a=1}^{i} \sum_{b=a}^{i} \tilde{\eta}^a \tilde{\eta}^b \sum_{j=b}^{i} \frac{2\varepsilon}{N} \exp\Big(-2\varepsilon[\mathcal{M}_j^a + \mathcal{M}_j^b]\Big) \\ &= -2\frac{(m_0^2 - m_1^2)}{N} \sum_{a=1}^{i} \sum_{b=a}^{i} \tilde{\eta}^a \tilde{\eta}^b F^{ab}(\varepsilon, \mathcal{M}) \end{aligned}$$

- We have $\text{var}(\sum_{a>b} \eta^a \eta^b c^{ab}) = \sum_{a>b} (c^{ab})^2 + 2\sum_a (c^{aa})^2$. Focusing only on ab terms

$$\begin{aligned}
\text{var}(W^i) &= -4\frac{(m_0^2 - m_1^2)^2}{N^2} \sum_{a=1}^{i} \sum_{b=a}^{i} F^{ab}(\varepsilon, \mathcal{M}) \\
&= -4\frac{(m_0^2 - m_1^2)}{N} \sum_{a=1}^{i} G(\varepsilon, \mathcal{M}) \\
&\sim \int dm^2 \; G(\varepsilon, m^2)
\end{aligned}$$

- Following (Engel and Schaefer 2011, 1102.1852), given a vector x^a on sphere, we can 'add' infinitesimal vector p^a to x^a by projecting p^a

$$\text{Proj}(p) = p^a - x^a x^b p^b$$

such that $p^a x^a = 0$.

- For a finite p^a, we can perform the following update

$$x^{a'} = x^a \cos(\alpha) + \frac{p^a}{|p|} \sin(\alpha)$$
$$p^{a'} = p^a \cos(\alpha) - |p| x^a \sin(\alpha) \quad ; \alpha = \Delta t |p|$$

such that x^a still lives on sphere and p^a still lives in the tangent space.

- Since arbitrary vector can be 'added' using projection, we can formulate 'gradient' on a sphere as

$$\overline{\partial_x^a S} = \lim_{p^a \to 0} \frac{S(x^a + \mathrm{Proj}(p^a)) - S(x^a)}{p^a} = \frac{\mathrm{Proj}(p^a)\partial_x^a S}{p^a} = \mathrm{Proj}(\partial_x^a S)$$

- We will also need to do the same for the divergence, although the formulation might be a bit more difficult.
- All in all, the NETS update equation can be rewritten as

$$dx^a = \mathrm{Proj}(-\varepsilon \partial_x^a U \; dt + \sqrt{2\varepsilon} \; d\eta^a + b_x^a \; dt)$$
$$dW = \partial_t U \; dt + b_x^a \mathrm{Proj}(\partial_x^a U \; dt) - \overline{\partial_x^a b_x^a} \; dt$$

- Note that, $\mathrm{Proj}(p^a)\mathrm{Proj}(q^a) = p^a \mathrm{Proj}(q^a) = \mathrm{Proj}(p^a) q^a$.