Importance weights distribution in neural samplers

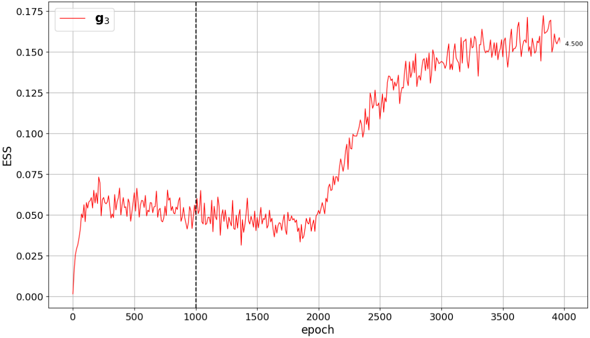
Piotr Białas

Institute of Applied Computer Science
Jagiellonian University, Kraków, Poland

Machine Learning Based Sampling in LFT and Quantum Chemistry
22 October 2024, Bonn

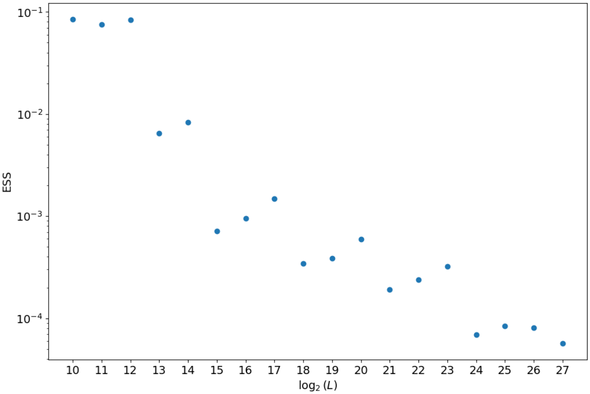
with P. Korcyl, T. Stebel and D. Zapolski

Motivation



$\mathbf{g}_3$
4.500
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
ESS
0
500
1000
1500
2000
2500
3000
3500
4000
epoch

Motivation



10^{-1}
10^{-2}
10^{-3}
10^{-4}
ESS
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
$\log_2(L)$

Neural samplers

$$p(\phi) \qquad P(\phi) = Z \cdot p(\phi)$$

$$q(\phi|\theta) \approx p(\phi)$$

$$w(\phi) = \frac{P(\phi)}{q(\phi)}$$

$$\langle w \rangle_q = \int \mathrm{d}\phi q(\phi) w(\phi) = Z \int \mathrm{d}\phi p(\phi) = Z$$

Partition function

$$Z \approx \hat{Z} = \frac{1}{N} \sum_{i=1}^{N} w(\phi_i), \quad \phi_i \sim q(\phi_i)$$

$$\text{var}\left[\hat{Z}\right]_q = \frac{1}{N} \text{var}\left[w\right]_q$$

Importance sampling

$$\langle h \rangle_p = \int \mathrm{d}\phi p(\phi) h(\phi) = \int \mathrm{d}\phi q(\phi) \overline{w}(\phi) h(\phi) = \langle \overline{w} \cdot h \rangle_q$$

$$\overline{w}(\phi) = \frac{w(\phi)}{\langle w \rangle_q}$$

$$\langle h \rangle_p \approx \hat{h} \equiv \frac{\sum_{i=1}^{N} w(\phi_i) h(\phi_i)}{\sum_{i=1}^{N} w(\phi_i)} \qquad \phi_i \sim q(\phi_i)$$

Nicoli, K. A., et al. "Asymptotically unbiased estimation of physical observables with neural samplers." Physical Review E, (2019) 101(2)

Errors

$$\mathrm{var}\left[\hat{h}\right] \approx \frac{1}{N}\left(\frac{\left\langle w^2\right\rangle_q}{\left\langle w\right\rangle_q^2}\,\mathrm{var}\left[h\right]_p + \frac{\left\langle \delta w \delta^2 h\right\rangle_p}{\left\langle w\right\rangle_q}\right)$$

$$\frac{1}{N}\,\mathrm{var}\left[\overline{w}\right] \ll 1$$

$$\mathrm{var}\left[\hat{h}\right]_q \approx \frac{\mathrm{var}\left[h\right]_p}{N \cdot \mathrm{ESS}}$$

A. Kong. A note on importance sampling using standarized weights. University of Chicago Technical Reports, 1992. Jun S. Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. Statistics and Computing, 6: 113–119, 1996.

Effective sample size - ESS

$$ESS = \frac{\langle w \rangle_q^2}{\langle w^2 \rangle_q} = \frac{\langle w \rangle_q^2}{\mathrm{var}\left[w\right]_q + \langle w \rangle_q^2} = \frac{1}{\mathrm{var}\left[\overline{w}\right]_q + 1}$$

$$ESS = \frac{1}{\mathrm{var}\left[\bar{w}\right]_q + 1}$$

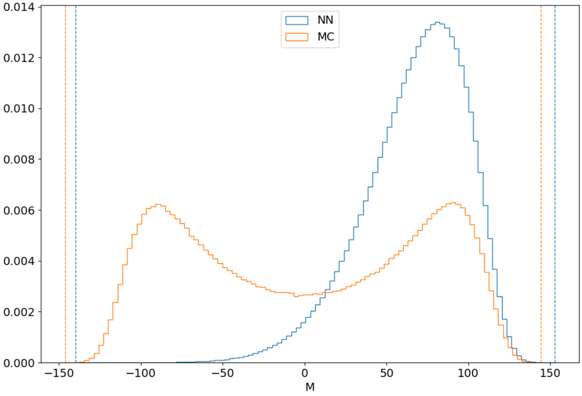
ϕ^4 model

$$S_A(\phi|\mu^2, \lambda, \kappa) = -\sum_{i,j=0}^{L-1} (\phi_{i+1,j}\phi_{i,j} + \phi_{i,j+1}\phi_{i,j}) + \sum_{i,j=0}^{L-1} \left(\frac{\mu^2 + 4}{2}\phi_{i,j}^2 + \frac{\lambda}{4!}\phi_{i,j}^4 \right)$$

$\mu^2 = -4$, $\lambda = 24.0 - 36.0$

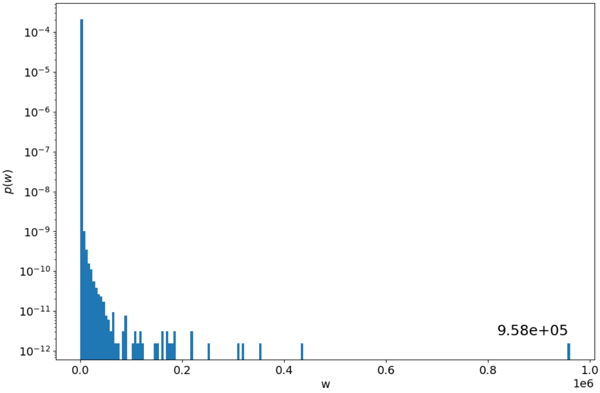
Poorly trained network

Magnetization $\lambda = 27.0$



ESS $= 0.006\%$

Weights distribution $\lambda = 27.0$



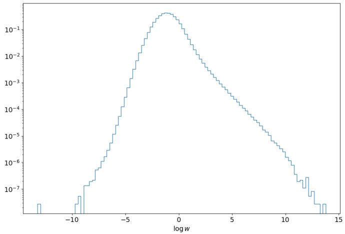
Pareto and exponential distribution

$$p(w) = a^b \frac{b}{w^{b+1}}, \quad w > a$$

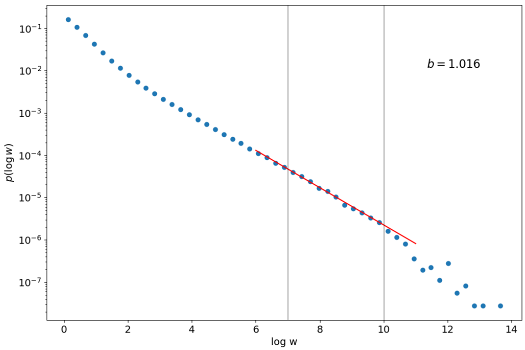
$$x = \log w$$

$$p(x) = a^b b e^{-bx}, \quad x \geq \log a$$

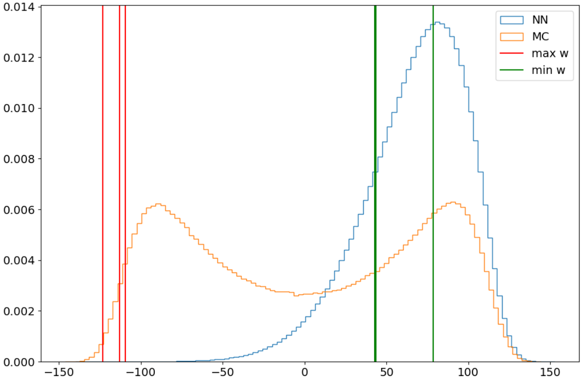
$\log w$ distribution $\lambda = 27.0$



log w distribution $\lambda = 27.0$

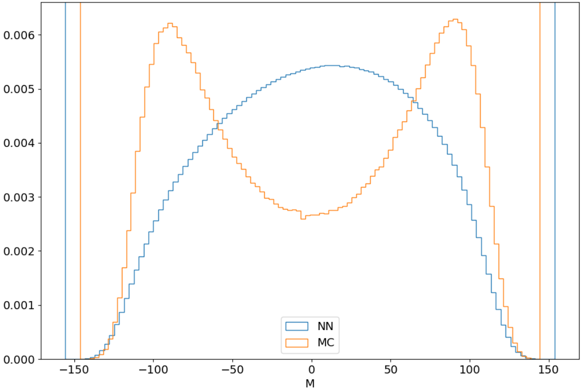


Magnetisation



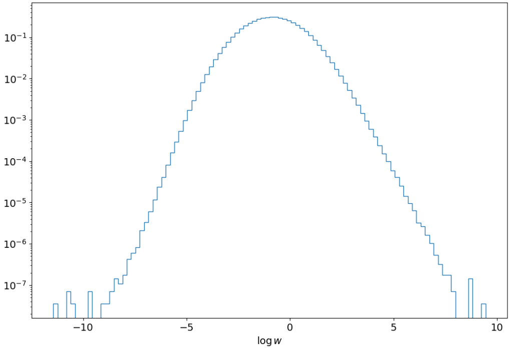
Better trained network

Magnetization $\lambda = 27.0$

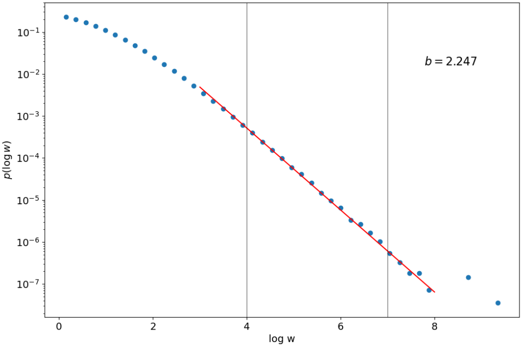


ESS $= 9\%$

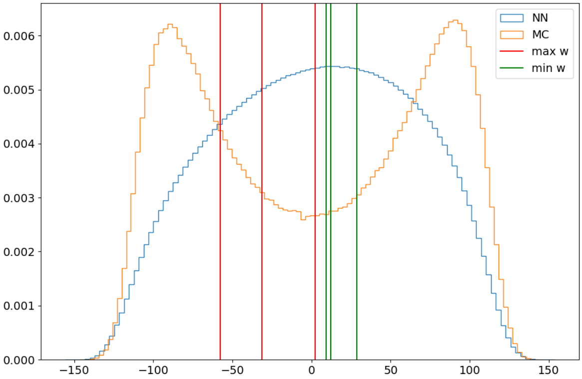
$\log w$ distribution $\lambda = 27.0$



log w distribution $\lambda = 27.0$

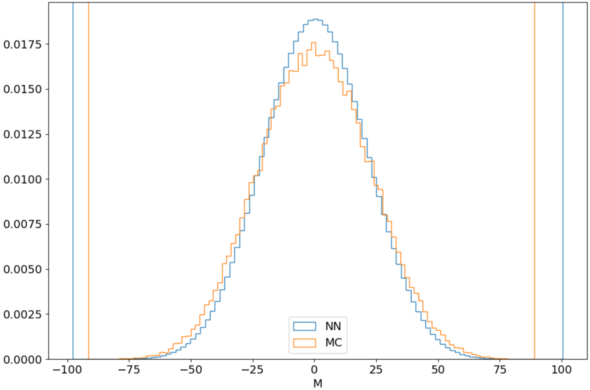


Magnetisation



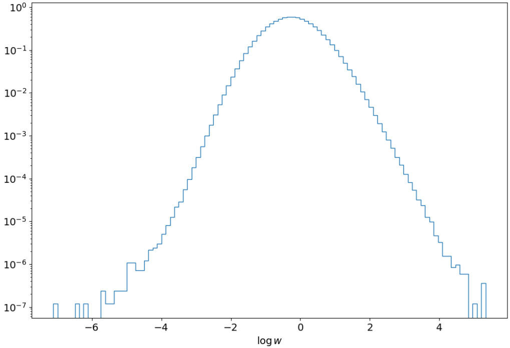
Disordered phase

Magnetization $\lambda = 36.0$

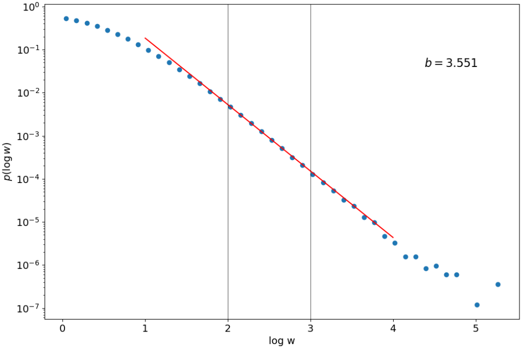


ESS $= 57\%$

log w distribution $\lambda = 36.0$



log w distribution $\lambda = 36.0$



Quotients (uncorrelated variables)

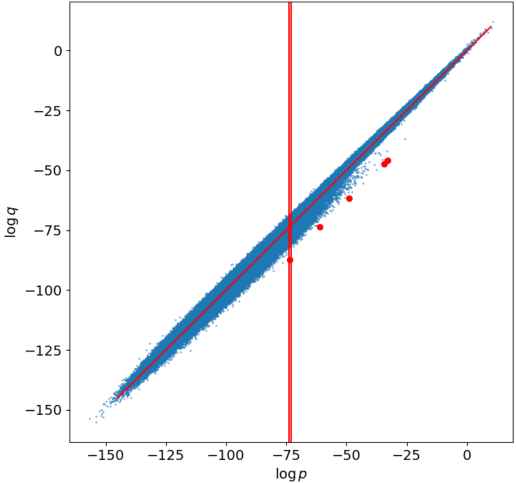
$$w(\phi) = \frac{p(\phi)}{q(\phi)}$$

$$W = \frac{X}{Y}.$$

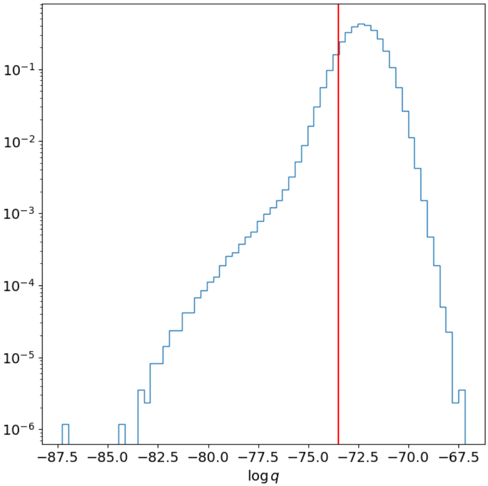
$$p_y(y) \approx a \cdot y^{b-1} \quad y \ll 1$$

$$p_W(w) \approx \frac{a''}{w^{b+1}} \qquad 1 \gg w$$

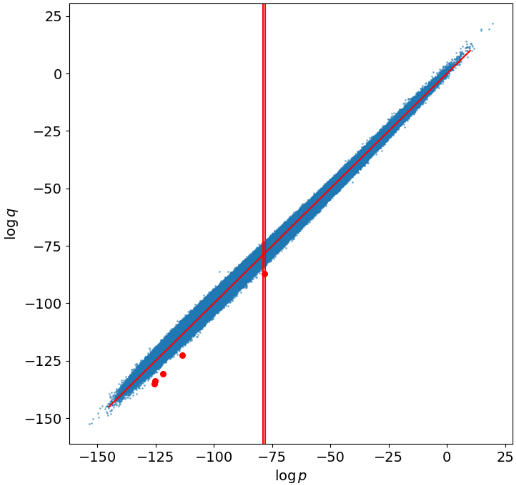
log w distribution $\lambda = 27$ (poorly trained)



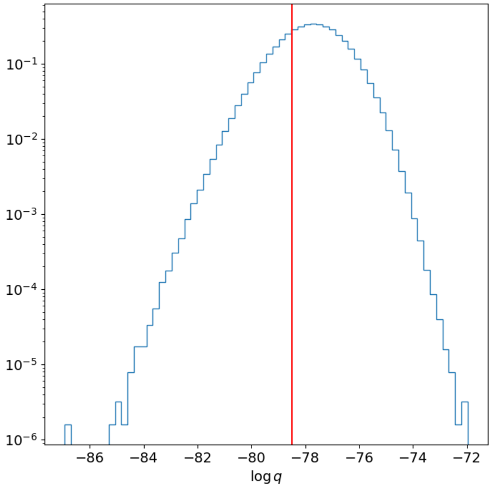
log w distribution $\lambda = 27$ (poorly trained)



$\log w$ distribution $\lambda = 27$



$\log w$ distribution $\lambda = 27$



Summary

- Importance weights seem to have a power like (Pareto) distribution.
- The power exponent depends on the training results.
- Worst case ($b < 2$) appears to be associated associated with "partial" mode collapse.
- Similar results for Ising model, U(1) and Schwinger.

Effective support

$$\begin{aligned}\text{supp}_{p,\epsilon}(q) &= \{\phi \in \text{supp}(q) : q(\phi) > \epsilon p(\phi)\} \\ &= \left\{\phi \in \text{supp}(q) : w(\phi) < \frac{1}{\epsilon}\right\}\end{aligned}$$

"Detecting and Mitigating Mode-Collapse for Flow-based Sampling of Lattice Field Theories", Kim A. Nicoli et al. arXiv:2302.14082

NeuMC

https://github.com/nmcmc/nmcmc-code

NeuMC

- Physics models (2D)
  - Scalar fields
  - Gauge fields
    - Different plaquette coupling layers
    - Different masking patterns
  - Fermions (Schwinger model)
- Gradients estimators
  - Reparameterisation trick
  - REINFORCE
  - Path gradients
- Introductory notebooks

Albergo, M. S., et al., "Introduction to Normalizing Flows for Lattice Field Theory." arxiv.2101.08176