Notes on Rotskoff and Vanden-Eijnden (2020)

Apostolos Psaros

1 Some notions from statistical mechanics

The state of a physical system is defined by its microstate (e.g., position and velocity of every molecule that composes it) and its macrostate (e.g., temperature). Every microstate $\boldsymbol{s}$ has some energy $H(\boldsymbol{s})$ and we want to infer, based on the macrostate, what the possible underlying microstates are.

If the system is isolated, the energy is fixed at H_0 and it is typically assumed that all microstates with energy equal to H_0 are equally likely whereas all microstates with a different energy are impossible. This defines a probability distribution with respect to the microstates.

On the other hand, if the system is not isolated but exchanges energy with a large reservoir at a constant temperature (i.e., is in thermal equilibrium with a heat bath at a fixed temperature) then each microstate $\boldsymbol{s}$ has a different energy and the probability density of the system being at $\boldsymbol{s}$ is given by

$$p(\boldsymbol{s}) = \frac{1}{Z}\exp(-H(\boldsymbol{s})/T) \tag{1}$$

where T is the temperature and Z is called partition function. The distribution defined by Eq. (1) is called canonical or Gibbs or Boltzmann distribution over microstates. Some microstates are more probable than others based on their energy $H(\boldsymbol{s})$.

2 Stochastic dynamical system

Consider the physical system with coordinates $(\boldsymbol{x}, \boldsymbol{v})$ governed by

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{v} \\ \dot{\boldsymbol{v}} &= -\nabla V(\boldsymbol{x}) - \gamma\boldsymbol{v} + \sqrt{2\beta^{-1}}\boldsymbol{\eta} \end{aligned} \tag{2}$$

where $V(\boldsymbol{x})$ is a potential energy function and β controls the magnitude of the fluctuations and is related with the temperature discussed above. The total energy of this system is given by the Hamiltonian $H(\boldsymbol{x}, \boldsymbol{v}) = \frac{1}{2}\|\boldsymbol{v}\|^2 + V(\boldsymbol{x})$. We therefore have a statistical ensemble for the coordinates $(\boldsymbol{x}, \boldsymbol{v})$ defined by the probability distribution

of Eq. (1) with energy given by the Hamiltonian. As a result, the expectation of any function g of the coordinates $(\boldsymbol{x}, \boldsymbol{v})$ is given by

$$\mathbb{E}[g(\boldsymbol{x}, \boldsymbol{v})] = \int g(\boldsymbol{x}, \boldsymbol{v}) \exp(-\beta H(\boldsymbol{x}, \boldsymbol{v})) d\boldsymbol{x} d\boldsymbol{v} \tag{3}$$

3 Backward Kolmogorov equation

We want to calculate the probability of the system leaving a configuration in phase space $\hat{A}$ to a distinct state $\hat{B}$. One way to do this is by quantifying the probability that a trajectory starting at $(\boldsymbol{x}, \boldsymbol{v})$ will first reach $\hat{B}$ before $\hat{A}$. This is denoted as $\hat{q}(\boldsymbol{x}, \boldsymbol{v})$ and follows the backward Kolmogorov equation; i.e. a PDE given by Eq. (4) in Rotskoff and Vanden-Eijnden (2020). This PDE is defined by an operator L acting on $\hat{q}(\boldsymbol{x}, \boldsymbol{v})$; i.e., $L\hat{q}(\boldsymbol{x}, \boldsymbol{v}) = 0$ for all the points $(\boldsymbol{x}, \boldsymbol{v})$ except for the ones in the BCs (defined by $\hat{A}$ and $\hat{B}$).

This PDE can be solved by defining a loss $\hat{C}(\hat{q})$ given by the average of $|L\hat{q}|^2$ (this is the "disagreement" with the PDE). This is very similar with the loss in PINNs. By integrating out the momenta we end up with a loss $C(q)$ and by adding the BCs as well we end up with the loss

$$C_\lambda(q) = \int |\nabla q(\boldsymbol{x})|^2 d\mu(\boldsymbol{x}) + Loss(BCs) \tag{4}$$

where

$$d\mu(\boldsymbol{x}) = \frac{1}{Z} \exp(-\beta V(\boldsymbol{x})) d\boldsymbol{x} \tag{5}$$

Thus, the distribution with respect which we average depends now only on the potential $V(\boldsymbol{x})$.

4 Fitting a neural network

We can use a NN parametrized by $\boldsymbol{\theta}$ for solving the PDE by minimizing the loss of Eq. (4). However we do not have access to the true loss. We can only sample $\boldsymbol{x}$ values and compute the average loss on the samples. This is called empirical loss. But how to sample such a high-dimensional space? There may be points that we have not sampled which incur a high loss. If we sample according to the canonical distribution defined by $\exp(-\beta V(\boldsymbol{x}))$ (as we should in general according to Eq. (5)), there may be low-probability points with high loss that dominate the average. As noted in Rotskoff and Vanden-Eijnden (2020), this is true for this problem. For example, consider a double

well potential $V(\boldsymbol{x})$. The points on the barrier have low probability according to the distribution defined by $\exp(-\beta V(\boldsymbol{x}))$. However, they are the ones incurring high loss: from going from one well to the other we have large gradients of q (the probability to move from $\hat{A}$ to $\hat{B}$ increases or decreases depending on the direction) and high gradients produce high loss according to Eq. (4). This is also depicted in Fig. 1 (notice how the lines of equal probability change with respect to the barrier).

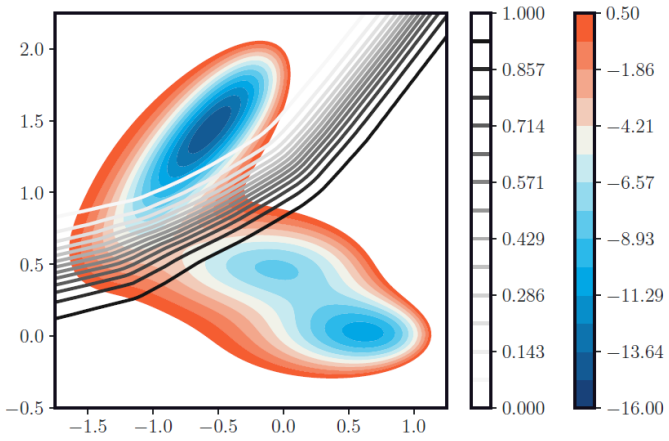

Fig. 1. Figure 1 from Rotskoff and Vanden-Eijnden (2020): The potential function is shown as a contour plot. The isocommittor lines are shown from white to black. Notably, level set $q = 0.5$ coincides with the saddle, as expected.

We thus need to use an importance sampling technique; i.e., sample from another distribution and take into account this change of measure in the expectation computation. For more information on the importance sampling technique used see Thiede et al. (2016). Finally, the algorithm used in Rotskoff and Vanden-Eijnden (2020) is very similar with the one used in PINNs. The difference lies in the points $\boldsymbol{x}$ that the loss is evaluated on in each iteration of the algorithm. In the algorithm of Thiede et al. (2016) for each iteration random locations are sampled according to a distribution that depends on both the potential $V(\boldsymbol{x})$ (the probability distribution) and the importance sampling functions.

References

Rotskoff, G. M. and Vanden-Eijnden, E. (2020). "Learning with Rare Data: Using Active Importance Sampling to Optimize Objectives Dominated by Rare Events". *arXiv preprint arXiv:2008.06334*.

Thiede, E. H., Van Koten, B., Weare, J., and Dinner, A. R. (2016). "Eigenvector Method for Umbrella Sampling Enables Error Analysis". *The Journal of chemical physics* 145, p. 084115.