Final Lab Rotation Report
Solmaz Azimi
Prof. Emilio Gallicchio

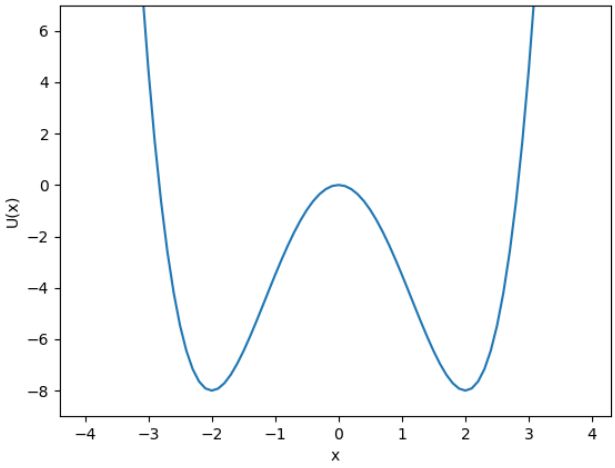
Maximum Likelihood Inference of the Symmetric Double-Well Potential

Introduction

Biomolecules undergo continuous conformational changes. Evaluating their equilibrium conformational states and obtaining accurate estimates of their free energy landscapes can elucidate important insights into their mechanisms. One approach to obtaining free energy estimates is to apply a Gaussian mixture model on a dataset reflecting multidimensional reaction coordinates[1-3].

This project aimed to estimate the probability distribution of the double-well potential fitted to a Gaussian mixture model by maximum likelihood inference. The double-well potential serves as an optimal, one-dimensional model for exploring physical phenomena (Equation 1). The function consists of a quadratic with two local minima, each of which characterize a free energy basin that is relevant to a simple study of free energy landscapes.

$$U(x) = -\frac{1}{4}x^2h^4 + \frac{1}{2}c^2x^4 \qquad \text{Eqn. (1)}$$



We sampled hypothetical datasets of the double-well potential based on the Boltzmann distribution shown in Equation 2, where k_B is the Boltzmann constant, and T is the absolute temperature.

$$P(x) = \frac{e^{\frac{-(U(x))}{K_BT}}}{\int e^{\frac{-(U(x))}{K_BT}} dx} \qquad \text{Eqn. 2}$$

The probability distribution of any phenomena is described in Equation 3, where a variable set of parameters θ returns the probability of observing the dataset x_n. In our case, θ would return a set of x values that correspond to the double-well potential.

$$P(x|\{\theta\}) \qquad \text{Eqn 3.}$$

If we assume a statistical model for x from Equation 3, we will need to evaluate the parameters that maximize the probability of observing the Boltzmann distribution. **Maximum likelihood** is a statistical method that estimates the optimal parameters to maximize the likelihood of observing a model probability distribution for a given dataset. It is a particularly powerful tool for large datasets[4].

The probability model of a sampled dataset x_n, wherein each data point is independent of the other and defined by a set of parameters θ, is given below, where c is a factor that does not affect the arrangement of the data points.

$$P(\{x_n\}|\{\theta\}) = P(x_1, x_2, \dots x_n|\{\theta\}) = \ c \cdot P(x_1|\{\theta\}) \cdot P(x_2|\{\theta\}) \cdot \dots P(x_n|\{\theta\}) \qquad \text{Eqn. 4,}$$

For a given set of parameters θ, Equation 4 predicts the outcome for the set of data x_n.

The **likelihood function** is the probability of obtaining a particular dataset given a probability distribution model defined by a set of parameters θ. The likelihood L for $P(x)$ with the given dataset x_n is

$$L(\{\theta\}|x_1, x_2, \dots x_n) = \prod_{i=1}^{n} p(x_i|\{\theta\}) \qquad \text{Eqn. 5}$$

There is an important distinction to be made between *likelihood* and *probability* in the context of statistics. **Probability** is defined as the predicted outcome of a dataset. Equation 4 returns the probability of observing the data set x_n for a given model parameter, such that

$$\{\theta\} \rightarrow \{x_n\}.$$

Likelihood is defined as an assessment of how likely a given model is able to describe a dataset[2,4]. For a given set of observed data x_n, Equation 5 provides an approximation of the model parameter, such that

$$\{x_n\} \rightarrow \{\theta\}.$$

Moreover, because probability is a fixed functional form, it is able to be integrated and there is an absolute measure of probability, whereas there is no absolute measure of likelihood[4].

We selected a Gaussian mixture model in Equation 6 to represent the sum of the random distribution obtained by the double-well potential function.

$$P_G(x) = \sum_i w_i \frac{e^{\frac{-(x-a_i)^2}{2\sigma_i^2}}}{\sqrt{2\pi\sigma_i}} = \frac{1}{2}\frac{e^{\frac{-(x-a_1)^2}{2\sigma_1^2}}}{\sqrt{2\pi\sigma_1}} + \frac{1}{2}\frac{e^{\frac{-(x-a_2)^2}{2\sigma_2^2}}}{\sqrt{2\pi\sigma_2}} \quad \text{Eqn. 6}$$

The Gaussian mixture is the sum of two Gaussian functions with four distinct parameters, a_1, a_2, σ_1, and σ_2 that represent θ. This component will be estimated by maximum likelihood inference to determine if $P_G(x)$ is consistent with the sampled data.

In higher dimension analyses, the overall free energy landscape is determined by Equation 7, where *G(x)* represents the Gaussian mixture density at *x* with the incorporated optimal parameters.

$$G(x) = -k_B T \ln P_G(x) \quad \text{Eqn. 7}$$

Procedure

1. Sampling Simulation

Hypothetical datasets from the potential function were generated using a smart-darting Monte-Carlo simulation (SDMC) under the conventional Metropolis acceptance criterion in Python 2.0. For the darting component of the simulation to be effective, the minima of the function in question must be known. Upon differentiating *U(x)*, the values 2.0 and 1.0 were designated for the variables *h* and *c* respectively. The generated datasets were compared to normalized Boltzmann Distribution functions (BD) with the respective partition functions at two temperatures, which were calculated by integration with WolframAlpha.

2. Maximum Likelihood Estimate

In order to obtain the parameters that maximize the likelihood, the procedure essentially involves a search guided by the gradient of the log form of the likelihood function,

$$\log L(\{\theta\}|x_n) = \log c + \prod_{i=1}^{n} \log p(x_i|\{\theta\}) \quad \text{Eqn. 8}$$

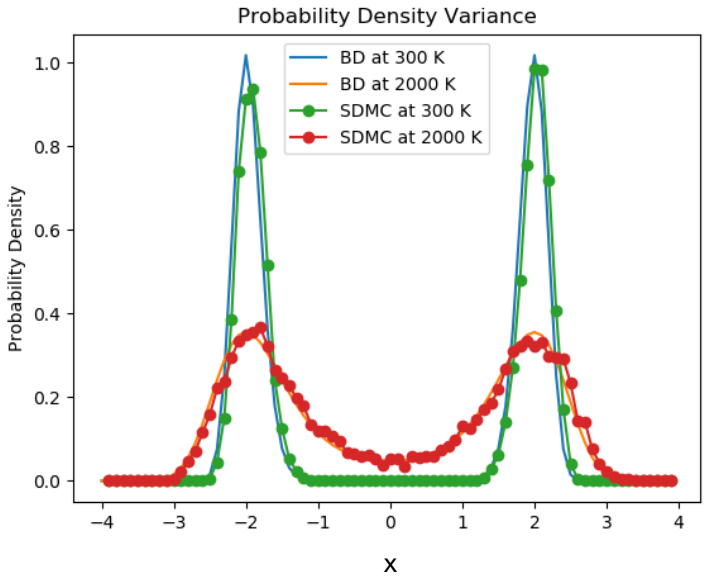
where c is a constant that does not affect the maximum.

For this particular step, TensorFlow (TF) was utilized on Jupyter Notebook in Python 3.0. Because the likelihood is the probability of numerous quantities, the log-likelihood was given to TF, in addition to the datasets obtained from SDMC simulations. In the logarithmic form (Equation 8), the method can be applicable to datasets that not only have any number of variants, but that also display either discontinuous or continuous distributions[4]. When the parameters are equally distributed maximum likelihood can achieve accuracy in larger sample sizes.

Results

The datasets that were obtained by SDMC simulations demonstrated close resemblance to the normalized Boltzmann Distributions at two different temperatures. At higher

temperatures, both potential functions broaden which reflect higher average energies, as well as a higher probability of observing these energy transitions.



A major challenge in studying free energy surfaces is sampling efficiency. In order to transition between two energy minima where the conventional Metropolis acceptance criterion apply, a larger displacement must be implemented in Monte-Carlo sampling[5]. A larger displacement, however, yields a lower acceptance rate and thus demonstrates lower efficiency. For example, a displacement of 1.0 results in an acceptance rate of 31%, whereas a displacement of 0.5 results in a 55% acceptance rate. Although a Monte-Carlo approach can accomplish displacement from one free energy basin to the next, the simulation does so at the expense of accuracy as the acceptance rate for the simulation deviates from the optimal 50%. We implemented a larger displacement (dart) chosen uniformly at random in addition to the traditional Monte-Carlo approach that implements small displacement steps to be able to accomplish this energy transition in 300 K. At higher temperatures, darting is not necessary because the system is in a state that is able to accommodate substantial energy transitions.

The SDMC datasets were then given to TensorFlow with the likelihood function. In addition, an initial set of parameters were defined for TF to determine the most optimal parameters. This presented a unique opportunity to probe the solutions that TF returned. Our trials demonstrate that TF is able to return a set of optimal parameters for the Gaussian Mixture model when a variety of initial parameters are defined. The factors that impact this procedure include the value of the Adam Optimizer, which we termed "maximum displacement" (d_{max}), and the number of steps this displacement involves (nsteps).

Our trials indicate that a d_{max} value of 0.02 returns more accurate parameters by TF regardless of the number of steps incorporated in the optimization. We applied the parameters returned under this criterion at various d_{max} values to confirm this. For the systems at both 300 K and 2000 K, the most optimal parameters obtained by TF

demonstrated a more accurate fit to the Gaussian mixture model at a larger maximum displacement value (Table 1).

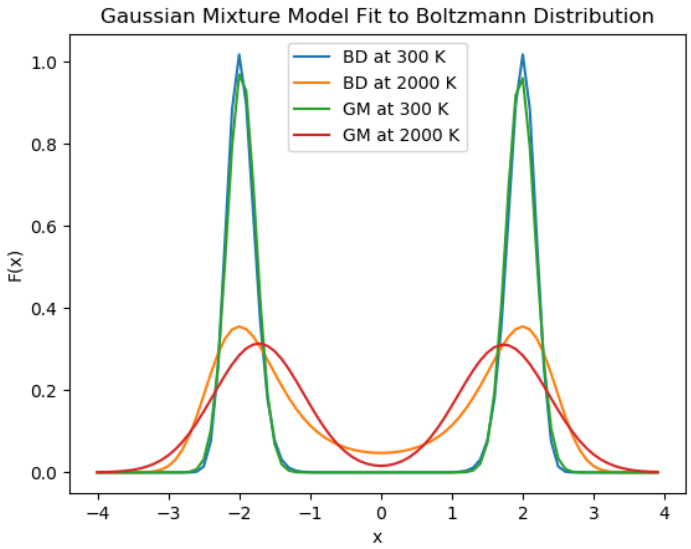


Table 1. The most optimal parameters estimated by TF for a system at two temperatures with a d_{max} of 0.02 and nsteps = 1000.

T	a_1	a_2	σ_1	σ_2
300 K	-1.967	1.969	0.2032	0.2054
2000 K	-1.722	1.731	0.6353	0.6407

Conclusions

Two very important lessons were learned in this project. First, escaping free energy minima can be accomplished by efficient sampling. These simulations can vary depending on the context of the problem, as we demonstrated with the implications of temperature. Since most biomolecular systems are at relatively lower temperatures, sampling simulations must be implemented to minimize errors that are associated to these realities in order to obtain an accurate free energy profile. Second, although maximum likelihood inference can estimate an underlying distribution for a hypothetical dataset, the method has its limitations. In our case, given a Gaussian mixture model, TF approximated a set of parameters that resemble the Boltzmann distribution of the double-well potential at 300 K, however the method is unable to do so with similar accuracy for conditions in higher temperatures. Maximum likelihood inference restricted to a Gaussian mixture model will never optimally fit the double-well potential because the model we selected essentially represents a distribution for the potential.

References

(1) Westerlund, A. M.; Delemotte, L. InfleCS: Clustering Free Energy Landscapes with Gaussian Mixtures. *Journal of Chemical Theory and Computation.* **0***, 0*, null.

(2) Lee, T.-S.; Radak, B. K.; Pabis, A.; York, D. M. A New Maximum Likelihood Approach for Free Energy Profile Construction from Molecular Simulations. *Journal of Chemical Theory and Computation.* **2013***, 9*, 153-164.

(3) Maragakis, P.; van der Vaart, A.; Karplus, M. Gaussian-Mixture Umbrella Sampling. *The Journal of Physical Chemistry B*. **2009***, 113*, 4664-4673.

(4) Fisher, R. A.; Russell, E. J. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*. *Series A, Containing Papers of a Mathematical or Physical Character*. **1922***, 222*, 309-368.

(5) Andricioaei, I.; Straub, J. E.; Voter, A. F. Smart Darting Monte Carlo. *The Journal of Chemical Physics*. **2001***, 114*, 6994-7000.

(6) Chen, Y.; Roux, B. Generalized Metropolis acceptance criterion for hybrid non-equilibrium molecular dynamics—Monte Carlo simulations. *The Journal of Chemical Physics*. **2015***, 142*, 024101.

(7) Sminchisescu, C.; Welling, M. Generalized darting Monte Carlo. *Pattern Recognition*. **2011***, 44*, 2738 – 2748.