Notes on Relative Entropy

Here I provide a general introduction to Relative Entropy. The whole theory is derived starting from the assumption that we have a very large (for good statistics) set of target configurations $\{\boldsymbol{r}^{(i)}\}$ (where $1 \leq i \leq M \gg 1$) that we will use to "train" (using terminology common to the machine learning community) the simplified system with. Each target configuration $\boldsymbol{r}^{(i)}$ is the full set of coordinates of the N particles $\boldsymbol{r}^{(i)} \equiv \left\{\boldsymbol{r}_\alpha^{(i)}\right\}$ where $\boldsymbol{r}_\alpha^{(i)}$ is the coordinate of particle α in configuration i. Such configurations can be generated in whatever way imaginable. For example, one might be interested in trying to find a pair interaction that in thermal equilibrium would approximately sample the configurations generated by a non-thermal (not governed by Boltzmann statistics) process. As another example, one could use a very complex set of interactions to force particles to assume certain shapes in thermal equilibrium, as done for the pores and clusters, and then find a pair potential that can again approximately sample the same configurations in thermal equilibrium (this can be thought of projecting the many body potential onto a smaller dimensional subspace of only pair interactions). Regardless of what the target is, we assume that the target configurations are Independently Identically Distributed (IID); this basically means that we only store snapshots separated enough from one another so as to be uncorrelated.

Up till now we have only assumed that we have a data set, $\{\boldsymbol{r}^{(i)}\}$, to work with such that we can try to find simple particle interactions that can approximately sample them at thermal equilibrium. Now we assume a thermal model with an arbitrary potential energy function (not yet written as pair potentials for generality). The probability for this model to sample a single configuration $\boldsymbol{r}^{(i)}$ is governed by the Boltzmann weight

$$P\left(\boldsymbol{r}^{(i)}|\boldsymbol{\theta}\right)\delta V^N = \frac{e^{-\beta U\left(\boldsymbol{r}^{(i)}|\boldsymbol{\theta}\right)}\delta V^N}{Z(\boldsymbol{\theta})}$$

where δV is a volume element to non-dimensionalize since the Boltzmann weight is a probability distribution (not a probability), $U\left(\boldsymbol{r}^{(i)}|\boldsymbol{\theta}\right)$ is the energy of configuration $\mathbf{r}^{(i)}$ with tunable potential parameters $\boldsymbol{\theta} = \{\theta_m\}$ and the configurational partition function

$$Z(\boldsymbol{\theta}) \equiv \int d\boldsymbol{R} e^{-\beta U(\boldsymbol{R}|\boldsymbol{\theta})}$$

Since the configurations are IID one can write the probability to sample all of the configurations as a product of the independent probabilities

$$P\left(\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(2)}, \ldots, \boldsymbol{r}^{(M)}|\boldsymbol{\theta}\right)\delta V^{NM} = \prod_{i=1}^{M} \frac{e^{-\beta U\left(\boldsymbol{r}^{(i)}|\boldsymbol{\theta}\right)}\delta V^N}{Z(\boldsymbol{\theta})}$$

For convenience I will use the notation $P(\boldsymbol{r}|\boldsymbol{\theta}) \equiv P\left(\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(2)}, \ldots, \boldsymbol{r}^{(M)}|\boldsymbol{\theta}\right)$ from here on out.

In the world of data science and machine learning, the next goal is to maximize the likelihood of reproducing the data, $P(\boldsymbol{r}|\boldsymbol{\theta})$, or more typically the log-likelihood

$$\ln[P(\boldsymbol{r}|\boldsymbol{\theta})\delta V^{NM}] = -\sum_{i=1}^{M}\beta U\big(\boldsymbol{r}^{(i)}|\boldsymbol{\theta}\big) - M\ln[Z(\boldsymbol{\theta})/\delta V^{N}]$$

Optimization can use gradient ascent via differentiation with respect to the various parameters in $\boldsymbol{\theta}$ as

$$\frac{\partial\ln[P(\boldsymbol{r}|\boldsymbol{\theta})\delta V^{NM}]}{\partial\theta_m} = -\sum_{i=1}^{M}\beta\frac{\partial U\big(\boldsymbol{r}^{(i)}|\boldsymbol{\theta}\big)}{\partial\theta_m} - M\frac{\partial\ln[Z(\boldsymbol{\theta})/\delta V^{N}]}{\partial\theta_m}$$

Next we can divide through by the number of target configurations we have, M, and take the derivative in the last term yielding

$$\frac{1}{M}\frac{\partial\ln[P(\boldsymbol{r}|\boldsymbol{\theta})\delta V^{NM}]}{\partial\theta_m} = -\left|\frac{\partial\beta U(\boldsymbol{R}|\boldsymbol{\theta})}{\partial\theta_m}\right|_{\{\mathbf{r}^{(i)}\}} + \langle\frac{\partial\beta U(\boldsymbol{R}|\boldsymbol{\theta})}{\partial\theta_m}\rangle_{U(\boldsymbol{R}|\boldsymbol{\theta})}$$

where

$$\langle(\cdots)\rangle_{\beta U(\boldsymbol{R}|\boldsymbol{\theta})} \equiv \frac{1}{Z(\boldsymbol{\theta})}\int d\boldsymbol{R}(\cdots)e^{-\beta U(\boldsymbol{R}|\boldsymbol{\theta})}$$

is the ensemble average with particles interacting via the dimensionless potential $\beta U(\boldsymbol{R}|\boldsymbol{\theta})$ and

$$|(\cdots)|_{\{\boldsymbol{r}^{(i)}\}} \equiv \frac{1}{M}\sum_{i=1}^{M}\int d\boldsymbol{R}(\cdots)\delta\big(\boldsymbol{R}-\boldsymbol{r}^{(i)}\big)$$

is an average over various target configurations with $\delta\big(\boldsymbol{R}-\boldsymbol{r}^{(i)}\big)$ a high dimensional delta function. As mentioned before, the last average, $|(\cdots)|_{\{\boldsymbol{r}^{(i)}\}}$, can encode non-equilibrium processes that generated the configurations or it could be a fully equilibrium process with a many body potential $\beta W(\boldsymbol{R})$ as used for pores and clusters--in this case $|(\cdots)|_{\{\boldsymbol{r}^{(i)}\}} = \langle(\cdots)\rangle_{\beta W(\boldsymbol{R})}$ and we are effectively finding a best mapping of $W(\boldsymbol{R}) \rightarrow U(\boldsymbol{R}|\boldsymbol{\theta})$ by tuning parameters $\boldsymbol{\theta}$.

Both averages are easy to compute if we only use pair potentials, $U(\boldsymbol{R}|\boldsymbol{\theta}) \equiv \sum_{i}^{N}\sum_{j>i}^{N}u\big(R_{i,j}|\boldsymbol{\theta}\big)$ since the averages only require two-point distributions, or in other words, radial distribution functions. This is correct regardless of the process used to generate the target configurations. In doing so we can write

$$\langle\frac{\partial\beta U(\boldsymbol{R}|\boldsymbol{\theta})}{\partial\theta_m}\rangle_{U(\boldsymbol{R}|\boldsymbol{\theta})} = \int d\boldsymbol{r}'\int d\boldsymbol{r}\rho^2 g(|\boldsymbol{r}-\boldsymbol{r}'||\boldsymbol{\theta})\frac{\partial\beta u(|\boldsymbol{r}-\boldsymbol{r}'||\boldsymbol{\theta})}{\partial\theta_m}$$

$$\left|\frac{\partial\beta U(\boldsymbol{R}|\boldsymbol{\theta})}{\partial\theta_m}\right|_{\{\mathbf{r}^{(i)}\}} = \int d\boldsymbol{r}'\int d\boldsymbol{r}\rho^2 g_{tgt}(|\boldsymbol{r}-\boldsymbol{r}'|)\frac{\partial\beta u(|\boldsymbol{r}-\boldsymbol{r}'||\boldsymbol{\theta})}{\partial\theta_m}$$

Shifting coordinates we obtain the final expression which tells us how to update the system according to a gradient descent scheme

$$\frac{1}{M}\frac{\partial \ln[P(\boldsymbol{r}|\boldsymbol{\theta})\delta V^{NM}]}{\partial \theta_m} = 4\pi\rho^2 V \int drr^2\big[g(r|\boldsymbol{\theta}) - g_{tgt}(r)\big]\frac{\partial \beta u(r|\boldsymbol{\theta})}{\partial \theta_m}$$

where V is the volume as a result of coordinate shifting and integrating. The left hand side expresses the gradient of the log of the sampling probability while the right hand side says how to actually compute it. In a simple gradient descent scheme this means applying updates to the parameters that are in the direction of the gradient ala (we can ignore the constant multiplicative factors of $4\pi\rho^2 V$ since they do not modify the direction)

$$\theta_m^{(i+1)} = \theta_m^{(i)} + \alpha \int drr^2\big[g\big(r|\boldsymbol{\theta}^{(i)}\big) - g_{tgt}(r)\big]\left[\frac{\partial \beta u(r|\boldsymbol{\theta})}{\partial \theta_m}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}}$$

where α is just some constant multiple that we can set to ensure the scheme does not apply to large of updates so as to fail and I have now labeled which step the parameters are on. This scheme can be improved a little by using gradient boosting which amounts to including a little bit of each prior update to give the thing “momentum” so to speak (impetus to keep moving in an averaged direction using knowledge of all prior steps). Thus, one can see how we can move downhill on the landscape for $\ln[P(\boldsymbol{r}|\boldsymbol{\theta})\delta V^{NM}]$ without actually needing to be able to ever compute it which would require a Free energy calculation [since we need $Z(\boldsymbol{\theta})$].