Boltzmann Generators - Theory

Mohsen Sadeghi

July 1, 2019

1 Statistical mechanics of Boltzmann Generators

1.1 The training procedure

We are interested in finding a diffeomorphism pair f and f^{-1}, acting between probability distributions in configuration space and latent space (Fig. 1). Such a diffeomorphism would map the coordinates between some simple distribution $q(z)$, which usually is a multivariate Gaussian around the origin, and a distribution $\nu(x)$ which resembles a Boltzmann distribution, $\mu(x) = \frac{1}{Z_\mu} \exp\left[-\frac{u_X(x)}{kT}\right]$, as close as possible. If we apply the inverse transform to the Boltzmann distribution, we get an approximation of the Gaussian, namely $p(z)$. Thus, applying probability transformations, we have,

$$\mu(x) = p(f(x)) \left|J(f(x))\right| \tag{1}$$

$$\nu(x) = q(f(x)) \left|J(f(x))\right| \tag{2}$$

where $\left|J(f(x))\right|$ is the Jacobian determinant of the transformation f.

As we are interested in training a deep network to learn the function f, we would use the KL divergence between pairs μ and ν or p and q as part of the loss function. In each pair, one is an exact probablity distribution (μ or q), and the other (ν or p) is what is approximated by the network. Considering the "forward" divergence, we have,

$$\begin{aligned}
KL(q(z)\,||p(z)) &= \int q(z)\log(q(z))\,dz - \int q(z)\log(p(z))\,dz && (3)\\
&= -\frac{S_q}{k} - \int q(z)\log\left(\mu\left(f^{-1}(z)\right)\left|J\left(f^{-1}(z)\right)\right|\right)dz && (4)\\
&= -\frac{S_q}{k} - \int \nu\left(f^{-1}(z)\right)\left|J\left(f^{-1}(z)\right)\right|\log\left(\mu\left(f^{-1}(z)\right)\right)dz - \mathbb{E}_{z\sim q(z)}\left[\log\left(\left|J\left(f^{-1}(z)\right)\right|\right)\right] && (5)\\
&= -\frac{S_q}{k} - \int \nu(x)\log(\mu(x))\,dx - \mathbb{E}_{z\sim q(z)}\left[\log\left(\left|J\left(f^{-1}(z)\right)\right|\right)\right] && (6)\\
&= -\frac{S_q}{k} + \mathbb{E}_{x\sim\nu(x)}\left[\frac{u_X(x)}{kT}\right] + \log Z_\mu - \mathbb{E}_{z\sim q(z)}\left[\log\left(\left|J\left(f^{-1}(z)\right)\right|\right)\right] && (7)\\
&= -\frac{S_q}{k} + \frac{E_\nu}{kT} - \frac{F_\mu}{kT} - \mathbb{E}_{z\sim q(z)}\left[\log\left(\left|J\left(f^{-1}(z)\right)\right|\right)\right] && (8)\\
&= -\frac{S_q}{k} + \frac{E_\nu}{kT} - \frac{F_\mu}{kT} + \frac{S_q - S_\nu}{k} && (9)\\
&= \frac{F_\nu - F_\mu}{kT} && (10)
\end{aligned}$$

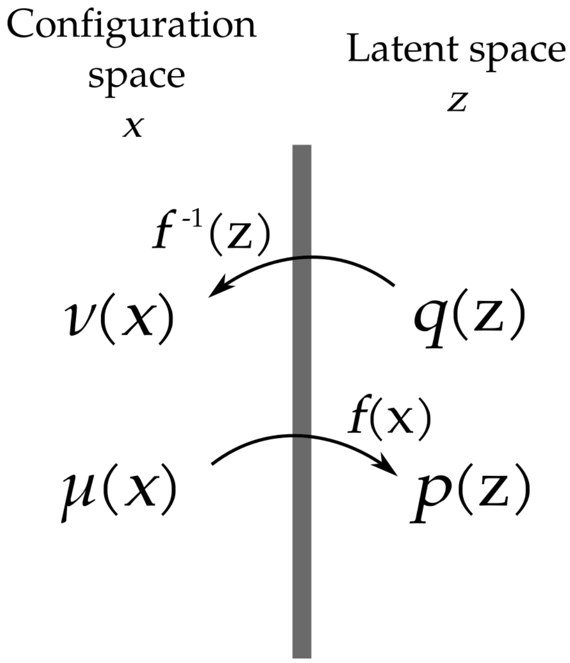

Figure 1: Transformation of probability distributions between configuration and latent space by the Boltzmann Generator

where S, E and F respectively denote entropy, internal energy, and free energy of the distribution given by the subscript, and u_X denotes the potential energy. Essentially, what the network minimizes when forward KL divergence is used as the loss function, is the free energy difference between the approximate distribution $\nu(x)$, and the exact Boltzmann distribution $\mu(x)$. This can also be written as the ratio of partition functions,

$$KL\left(q\left(z\right)||p\left(z\right)\right)=\log\frac{Z_{\mu}}{Z\nu} \tag{11}$$

It can also be shown that $KL\left(q\left(z\right)||p\left(z\right)\right)=KL\left(\nu\left(x\right)||\mu\left(x\right)\right)$, i.e. the KL divergence is the same if considered in either configuration or latent space.

1.2 Thermodynamic interpretation of the probability transformation

Now, let's assume that the network has converged to a transformation. Because the approximate distributions will not have matched the expected ideal ones ($\mu_X\neq\nu_X$ and $q_Z\neq p_Z$), a reweighting factor $w=\frac{\mu(x)}{\nu(x)}$ is in general used. We are interested in finding the thermodynamic meaning of the transformation f. We make the assumption that f is a volume preserving diffeomorphism, which basically describes a Hamiltonian flow, if both spaces x and z are considered as physical configuration spaces. We can treat the Gaussian distribution $q(z)$ as a Boltzmann distribution of a system in a harmonic well with the energy $u_Z(z)=\sum\frac{kT}{2\sigma_i^2}\left(z_i^2\right)$.

One problem that immediately becomes apparent is that states connected by a BG transformation cannot both be considered in thermodynamic equilibriun. For example, if we consider the transformation from $q(z)$ to $\nu(x)$, while $q(z)$ can be considered a Boltzmann distribution, we cannot make such a claim for the $\nu(x)$, because it only "resembles" the Boltzmann distribution, $\mu(x)$. Keeping that in mind, we consider the paths defining the Hamiltonian flow to be given by the field $x=\phi\left(t;x_0,t_0\right)$. We have $\phi\left(-\tau;z,-\tau\right)=z$ and $\phi\left(+\tau;z,-\tau\right)=x=f^{-1}\left(z\right)$, where time t has been considered to change in the symmetric interval $[-\tau,\tau]$. In order to discuss the free energy change under this transformation, we first introduce the Jarzynski equality.

1.2.1 The Jarzynski equality

The Jarzynski equality states that if a system in thermodynamic equilibrium state A, of temperature T, is transformed to another state B (which does not necessarily need to be an equilibrium state),

$$e^{-\frac{\Delta F}{kT}} = \left\langle e^{-\frac{W}{kT}} \right\rangle \tag{12}$$

in which W is the work done on the system in getting it from state A to state B, and $\langle \cdots \rangle$ designate the ensemble average for all different paths through which this transformation can happen.

1.2.2 Free energy difference

Now, assume two macrostates A and B, which encompass ensembles of microstates in the starting and finishing configurations, i.e. if the system is in state A before the application of the transformation, it will end up in state B under the flow ϕ. We can use the Jarzynski equality between these two states,

$$\exp\left(-\frac{F_B - F_A}{kT}\right) = \left\langle \exp\left(-\frac{W\left[z \rightarrow x\right]}{kT}\right)\right\rangle_{z\in A, x\in B} \tag{13}$$

$$= \left\langle \exp\left(-\frac{u_X\left(x\right) - u_Z\left(z\right) - Q\left[z \rightarrow x\right]}{kT}\right)\right\rangle_{z\in A, x\in B} \tag{14}$$

where $Q\left[z \rightarrow x\right]$ is the heat supplied to the system from the heat bath during such a transformation. For a "microscopically reversible" process, which presumably happens under the volume-preserving transformaiton learned by the Boltzmann Generator, we have,

$$\exp\left(-\frac{Q\left[z \rightarrow x\right]}{kT}\right) = \frac{P\left[z \rightarrow x \mid \phi\right]}{P\left[x \rightarrow z \mid \bar{\phi}\right]} \tag{15}$$

where $P\left[z \rightarrow x \mid \phi\right]$ denotes the probablity that, under the flow ϕ, we take the path from the microstate z to the microstate x. In reverse, $P\left[x \rightarrow z \mid \bar{\phi}\right]$ denotes the probability of the reverse path under the reverse flow. It is to be noted that because the number of particles and the temperature are considered constant here, a change in the volume is the only explanation possible for the probability changes. Because the action of the Boltzmann generator is deterministic, this condition simplifies to,

$$Q\left[z \rightarrow x\right] = -kT \log\left(\frac{q\left(z\right)\delta\left(\phi\left(+\tau; z, -\tau\right) - x\right)}{\nu\left(x\right)\delta\left(\bar{\phi}\left(-\tau; x, +\tau\right) - z\right)}\right) \tag{16}$$

Substituting in 14, we get,

$$\exp\left(-\frac{F_B - F_A}{kT}\right) = \int \mathbf{1}_A(z)\, q(z) \exp\left(-\frac{u_X\left(f^{-1}(z)\right) - u_Z(z)}{kT}\right) \exp\left(-\frac{Q[z \to x]}{kT}\right) dz \tag{17}$$

$$= \int \mathbf{1}_A(z)\, q(z) \exp\left(-\frac{u_X\left(f^{-1}(z)\right) - u_Z(z)}{kT}\right) \frac{q(z)}{\nu\left(f^{-1}(z)\right)} dz \tag{18}$$

$$= \int \mathbf{1}_A(z)\, q(z) \frac{\exp\left(\frac{u_Z(z)}{kT}\right) \frac{1}{Z_q} \exp\left(\frac{-u_Z(z)}{kT}\right) \exp\left(\frac{-u_X\left(f^{-1}(z)\right)}{kT}\right)}{\nu\left(f^{-1}(z)\right)} dz \tag{19}$$

$$= \frac{1}{Z_q} \int \mathbf{1}_A(z)\, q(z) \frac{\exp\left(\frac{-u_X\left(f^{-1}(z)\right)}{kT}\right)}{\nu\left(f^{-1}(z)\right)} dz \tag{20}$$

$$= \frac{1}{Z_q} \int \mathbf{1}_B(x)\, q(f(x)) \frac{\exp\left(\frac{-u_X(x)}{kT}\right)}{\nu(x)} \left|J(f(x))\right| dx \tag{21}$$

$$= \frac{1}{Z_q} \int \mathbf{1}_B(x)\, \nu(x) \frac{\exp\left(\frac{-u_X(x)}{kT}\right)}{\nu(x)} dx \tag{22}$$

$$= \frac{1}{Z_q} \int \mathbf{1}_B(x) \exp\left(\frac{-u_X(x)}{kT}\right) dx \tag{23}$$

where we have used $\mathbf{1}_A(z)$ and $\mathbf{1}_B(x)$ to denote sets of microstates z and x respectively belonging to macrostates A and B, with $\mathbf{1}_A(f(x)) = \mathbf{1}_B(x)$.

As a test of this result, we can consider the extreme case where the macrostates cover the whole configuration spaces. In that case, we get the identity $F_B - F_A = -kT \log\left(\frac{Z_\mu}{Z_q}\right) = F_\mu - F_q$. In general, it is interesting to see that this result does not explictly depend on the approximate distribution $\nu(x)$. Also, Z_q is just the normalization factor of the Gaussian distribution, and is analytically available. Thus, this result provides a tool for calculating free energy differences between any two macrostates in the original configuration space, even if different Boltzmann Generators have been used.