Boltzmann Generators - Theory

Mohsen Sadeghi

July 18, 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 $|J(f(x))|$ 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 probability distribution (μ or q), and the other (ν or p) is what is approximated by the network. Considering the "forward" divergence, we have,

$$KL(q(z)||p(z)) = \int q(z)\log(q(z))\,dz - \int q(z)\log(p(z))\,dz \tag{3}$$

$$= -\frac{S_q}{k} - \int q(z)\log\left(\mu\left(f^{-1}(z)\right)\left|J\left(f^{-1}(z)\right)\right|\right)dz \tag{4}$$

$$= -\frac{S_q}{k} - \int \nu\left(f^{-1}(z)\right)\left|J\left(f^{-1}(z)\right)\right|\log\left(\mu\left(f^{-1}(z)\right)\right)dz - \mathbb{E}_{z\sim q(z)}\left[\log\left(\left|J\left(f^{-1}(z)\right)\right|\right)\right] \tag{5}$$

$$= -\frac{S_q}{k} - \int \nu(x)\log(\mu(x))\,dx - \mathbb{E}_{z\sim q(z)}\left[\log\left(\left|J\left(f^{-1}(z)\right)\right|\right)\right] \tag{6}$$

$$= -\frac{S_q}{k} + \mathbb{E}_{x\sim\nu(x)}\left[\frac{u_X(x)}{kT}\right] + \log Z_\mu - \mathbb{E}_{z\sim q(z)}\left[\log\left(\left|J\left(f^{-1}(z)\right)\right|\right)\right] \tag{7}$$

$$= -\frac{S_q}{k} + \frac{E_\nu}{kT} - \frac{F_\mu}{kT} - \mathbb{E}_{z\sim q(z)}\left[\log\left(\left|J\left(f^{-1}(z)\right)\right|\right)\right] \tag{8}$$

$$= -\frac{S_q}{k} + \frac{E_\nu}{kT} - \frac{F_\mu}{kT} + \frac{S_q - S_\nu}{k} \tag{9}$$

$$= \frac{F_\nu - F_\mu}{kT} \tag{10}$$

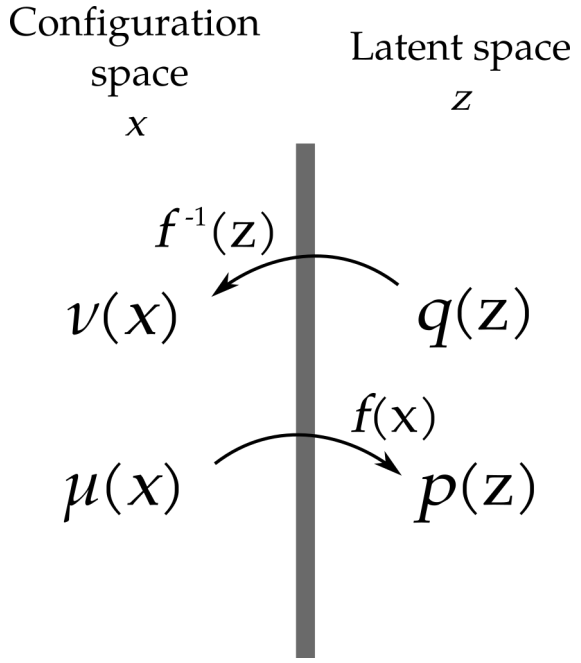

Figure 1: Transformation of probability distributions between configuration and latent space by the Boltzmann Generator

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,

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

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

1.2 Thermodynamic interpretation of probability transformations

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 re-weighting factor $w = \frac{\mu(x)}{\nu(x)}$ is in general used. We are interested in finding the thermodynamic meaning of the transformation f. We assume that both latent and configuration spaces x and z are physical configuration spaces of the same system. Therefore, we treat the Gaussian distribution $q(z)$ as a Boltzmann distribution of a system in a harmonic well with the energy $u_Z(z) = \sum \frac{kT}{2\sigma_i^2}\left(z_i^2\right)$. It is to be noted that because now we are treating both latent and real representations as belonging to physical configuration spaces, the potential energies need to have the same reference value. Adding a constant to $u_X(x)$ should also affect $u_Z(z)$, and more importantly, the Gaussian normalization factor Z_q, which we treat as a partition function.

1.2.1 Reversibility and volume-preservation

While the diffeomorphism f, by the virtue of reversible layers used in the BG, describes a reversible flow in the configuration space, we note that in general, we cannot make the assumption that f is

also volume preserving ($|J(f(x))| \neq 1$). Thus, quite generally, it describes a flow in phase space resulting from a "reversible non-Hamiltonian dynamics". This is not surprising, if we aim to treat both distributions $q(z)$ and $\nu(x)$ at the same temperature, T. The transformation between these two thermodynamic states needs to be "thermostatted", and in general, through this process, there is a net energy transfer between the system and the environment in the form of heat. This can be translated into fluctuations in the volume elements of the phase space.

1.2.2 Thermodynamic equilibrium

One problem that immediately becomes apparent is that states connected by a BG transformation cannot both be considered in thermodynamic equilibrium. 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, $\mu(x)$, while the equilibrium Boltzmann distribution at temperature T is unique. Keeping that in mind, we consider the paths defining the non-Hamiltonian flow to be given by the field $x = \phi(t; x_0, t_0)$. We have $\phi(-\tau; z, -\tau) = z$ and $\phi(+\tau; z, -\tau) = x = f^{-1}(z)$, 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, which in general belongs to the realm of non-equilibrium thermodynamics, we first introduce the Jarzynski equality.

1.2.3 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), [2]

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

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.

1.2.4 Entropy production

As we mentioned, for the non-equilibrium transformation between q and ν, we need to calculate the energy transfer in the form of heat. Here, we consider two assumptions for the process taking place between q and ν.

In the first case, we assume this transformation to be a Markovian stochastic process. It happens in memory-less steps, and the choice of destination probability distribution in each step is stochastic. Reversibility has only meaning as the detailed-balance holding in these steps. This obviously is different from what the BG does. But it helps elucidate further results.

In the second case, we assume ϕ to represent a non-Hamiltonian flow. We make the assumption that this process can be modeled as equivalent to a Nosé-Hoover type thermostat.

1.2.5 Case 1: stochastic Markovian process

For this case, the "microscopically reversibility" of process, is equivalent to [1],

$$\exp\left(-\frac{Q[z \to x]}{kT}\right) = \frac{P[z \to x \mid \phi]}{P\left[x \to z \mid \bar{\phi}\right]} \tag{13}$$

where $P[z \to x \mid \phi]$ denotes the probability that, under the flow ϕ, we take the path from the microstate z to the microstate x. In reverse, $P\left[x \to z \mid \bar{\phi}\right]$ denotes the probability of the reverse path under the reverse flow. If we insist to apply this interpretation to the action of BG, we need to consider the detailed balance as,

$$q(z)\,P[z \to x \mid \phi] = \nu(x)\,P\left[x \to z \mid \bar{\phi}\right] \tag{14}$$

which results in,

$$Q[z \to x] = -kT \log\left(\frac{\nu(x)}{q(z)}\right) \tag{15}$$

It is interesting to observe that,

$$\Delta S[z \to x] = \frac{Q[z \to x]}{T} = -k \log\left(\frac{q(f(x))\,|J(f(x))|}{q(z)}\right) = -k \log|J(f(x))| \tag{16}$$

1.2.6 Case 2: Thermostatted non-Hamiltonian flow

As we are now considering non-Hamiltonian dynamics, we need to account for entropy production due to volume element fluctuations. In order to do so in a reliable manner, we start by connecting the Jacobian determinant to the metric of the phase space,

$$|J(\phi(t; z, -\tau))| = \frac{\sqrt{g(\phi(t; z, -\tau))}}{\sqrt{g(z)}} \tag{17}$$

where the $\sqrt{g}$ is the metric determinant factor. With this definition, we can write the generalized Liouville equation [3],

$$\frac{\partial\left(\rho\sqrt{g}\right)}{\partial t} + \nabla \cdot \left(\rho\sqrt{g}\dot{\phi}\right) = 0 \tag{18}$$

where $\rho(t)$ is the time-dependent non-equilibrium probability distribution in the phase space.

We make the assumption that a Nosé-Hoover type non-Hamiltonian dynamics governs the flow ϕ in phase space. Thus,

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i} \tag{19}$$

$$\dot{\mathbf{p}}_i = \mathbf{F}_i - \frac{p_\eta}{Q}\mathbf{p}_i \tag{20}$$

$$\dot{\eta} = \frac{p_\eta}{Q} \tag{21}$$

$$\dot{p}_\eta = \sum \frac{\mathbf{p}_i^2}{m_i} - 3NkT \tag{22}$$

where η and its conjugate momentum, p_η, are the additional degrees of freedom, allowing for the thermostatting. The average kinetic energy of the system in this setup fluctuates around the equilibrium value of $\frac{3}{2}NkT$. While the dynamics is non-Hamiltonian in the original degrees of freedom, it conserves energy in the extended system, i.e. [3]

$$H' = H(\mathbf{p}, \mathbf{q}) + \frac{p_\eta^2}{2Q} + 3NkT\eta = const. \tag{23}$$

On the other hand, the metric determinant factor for this extended dynamics is [3],

$$\sqrt{g} = \exp(3N\eta) \tag{24}$$

which suggests that,

$$|J(\phi(t;z,-\tau))| = \exp[3N(\eta(-\tau)-\eta(t))] \tag{25}$$

and considering the endpoints of the transformation,

$$\left|J\left(f^{-1}(z)\right)\right| = \exp[3N(\eta(-\tau)-\eta(\tau))] \tag{26}$$

If we neglect the amount of energy attributed to the momentum of the extended variable (the $\frac{p_\eta^2}{2Q}$ term in Eq. 23), we can come up with an estimate of the energy transfer to the system due to thermostatting, which is the heat transferred to the system. This would give us,

$$\begin{aligned} Q[z\to x] &= \left(H-H'\right)_{z\to x} && (27)\\ &\approx 3NkT(\eta(-\tau)-\eta(\tau)) && (28)\\ &= kT\log\left|J\left(f^{-1}(z)\right)\right| && (29)\\ &= -kT\log|J(f(x))| && (30) \end{aligned}$$

and, for the entropy production,

$$\begin{aligned} \Delta S[z\to x] &= \frac{Q[z\to x]}{T} && (31)\\ &= -k\log|J(f(x))| && (32) \end{aligned}$$

Interestingly, the results from the two cases match. Though the second case is more realistic, we have neglected the effect due to the term $\frac{p_\eta^2}{2Q}$ in this case.

1.2.7 Free energy difference

Now, assume two macrostates A and B, which encompass ensembles of microstates in the starting and final 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,

$$\begin{aligned} \exp\left(-\frac{F_B-F_A}{kT}\right) &= \left\langle \exp\left(-\frac{W[z\to x]}{kT}\right)\right\rangle_{z\in A, x\in B} && (33)\\ &= \left\langle \exp\left(-\frac{u_X(x)-u_Z(z)-Q[z\to x]}{kT}\right)\right\rangle_{z\in A, x\in B} && (34) \end{aligned}$$

where $Q[z\to x]$ is the heat supplied to the system from the heat bath during such a transformation, and can be substituted based on either of the methods described above.

Substituting in 34, we get,

$$\exp\left(-\frac{F_B - F_A}{kT}\right) = \int \mathbf{1}_A(z)\, q(z) \exp\left(-\frac{u_X\left(f^{-1}(z)\right) - u_Z(z)}{kT}\right) \exp\left(\frac{Q[z \to x]}{kT}\right) dz \tag{35}$$

$$= \int \mathbf{1}_A(z)\, q(z) \exp\left(-\frac{u_X\left(f^{-1}(z)\right) - u_Z(z)}{kT}\right) \left|J\left(f^{-1}(z)\right)\right| dz \tag{36}$$

$$= \int \mathbf{1}_A(z)\, \frac{1}{Z_q} \exp\left(\frac{-u_Z(z)}{kT}\right) \exp\left(\frac{u_Z(z)}{kT}\right) \exp\left(\frac{-u_X\left(f^{-1}(z)\right)}{kT}\right) \left|J\left(f^{-1}(z)\right)\right| dz \tag{37}$$

$$= \frac{1}{Z_q} \int \mathbf{1}_A(z) \exp\left(\frac{-u_X\left(f^{-1}(z)\right)}{kT}\right) \left|J\left(f^{-1}(z)\right)\right| dz \tag{38}$$

$$= \frac{1}{Z_q} \int \mathbf{1}_B(x) \exp\left(\frac{-u_X(x)}{kT}\right) dx \tag{39}$$

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 explicitly 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.

References

[1] Gavin E. Crooks. Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems. *J. Stat. Phys.*, 90(5/6):1481–1487, 1998.

[2] C Jarzynski. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.*, 78(14):2690–2693, 1997.

[3] Mark E Tuckerman, Yi Liu, Giovanni Ciccotti, and Glenn J. Martyna. Non-Hamiltonian molecular dynamics: Generalizing Hamiltonian phase space principles to non-Hamiltonian systems. *J. Chem. Phys.*, 115(4):1678–1702, 2001.