Boltzmann Generators - Theory

May 25, 2019

1 Boltzmann Generators

1.1 Short summary of setting

The goal of Boltzmann Generators is to sample efficiently from a Boltzmann distribution. In contrast to established trajectory-based sampling methods, Boltzmann Generators are able to produce statistically independent samples from stationary distributions. This is achieved using Deep Learning.

The learning problem can be stated as follows: We are interested in learning an invertible transformation between a known prior distributions $q_Z(z)$ and some distribution $p_X(x)$, such that $p_X(x)$ is as close as possible to another distribution $\mu_X(x)$, e.g. a Boltzmann distribution. Another formulation of the same problem is that we are interested in learning an invertible transformation between the known distributions $\mu_X(x)$ and some distribution $p_Z(z)$, such that $p_Z(z)$ is as close as possible to the prior distribution $q_Z(z)$. If we have such a transformation it is possible to sample from $q_Z(z)$ and transform the samples with the learned transformation to obtain samples which follow nearly $\mu_X(x)$. In practice a multivariate Gaussian is used for the latent prior distribution $q_Z(z)$ and we want to find an invertible transformation $f(x)$, such that $p_X(x)$ is as close as possible to our desired distribution, e.g. a Boltzmann distribution $\mu_X(x)$. In theory it should be possible to match the distributions $\mu_X(x)$ and $p_X(x)$ exactly if there are no restrictions to f. However, because of the training process and the restrictions to f this is not possible in reality. Therefore, reweighting is used In the sampling process to obtain the true Boltzmann distribution $\mu_X(x)$. The reweighting factor is given by

$$w(x) = \frac{\mu_X(x)}{p_X(x)} \tag{1}$$

The transformation f can be modeled by an invertible generative network which contains e.g. NICER or RealNVP transformations. The training of Boltzmann Generators is done by minimizing the KL-divergence between $p_X(x)$ and $\mu_X(x)$ (or between $q_Z(z)$ and $p_Z(z)$ which yields the same loss function). However, the KL-divergence is not symmetric and, therefore, the forward and reverse KL-divergence yield different loss functions. Both can be used to train the network.

2 Non linear transformations and entropy

2.1 Substitution in probability theory

Consider two random variables X and Z with probability density $\mu_X(x)$ and $p_Z(z)$, respectively. The probability that Z takes a value in the arbitrary subset S is given by

$$P(Z \in S) = \int_S p_Z(z)dz \tag{2}$$

Now, we introduce an invertible transformation $z = f(x)$ such that Z takes a value in S whenever X takes a value in $f^{-1}(S)$. Thus

$$P(Z \in S) = \int_{f^{-1}(S)} \mu_X(x) dx \tag{3}$$

Performing the substitution $x = f^{-1}(z)$ yields

$$\int_{f^{-1}(S)} \mu_X(x) dx = \int_S \mu_X(f^{-1}(z)) \left| \frac{df^{-1}}{dz} \right| dz \tag{4}$$

Combining the above equations, we obtain

$$\int_S p_Z(z) dz = \int_S \mu_X(f^{-1}(z)) \left| \frac{df^{-1}}{dz} \right| dz \tag{5}$$

As this holds for arbitrary subsets S, we conclude

$$p_Z(z) = \mu_X(f^{-1}(z)) \left| \frac{df^{-1}(z)}{dz} \right| \tag{6}$$

or equally

$$\mu_X(x) = p_Z(f(x)) \left| \frac{df(x)}{dx} \right| \tag{7}$$

For multiple dimensions, this generalizes to

$$\mu_X(x) = p_Z(f(x)) \left| \det J(f(x)) \right| \tag{8}$$

where $\det J(f(x))$ is the determinant of the Jacobian Matrix.

2.2 Forward KL-Divergence in latent space

Assume we have an invertible transformation $z = f(x)$ between configuration space and latent space with distributions $\mu_X(x)$ and $p_Z(z)$, respectively. The relation between the two distributions is defined in equation 7. In the case of Boltzmann Generators, one is interested in learning the transformation f such that $p_Z(z)$ is close to a prior distribution $q_Z(z)$, e.g. a multivariate Gaussian.

The prior distribution $q_Z(z)$ transforms similar to eq. 8 via

$$p_X(x) = q_Z(f(x)) \left| \det J(f(x)) \right| \tag{9}$$

to the distribution $p_X(x)$. A suitable loss function for training purposes is given by the KL-Divergence in latent space, namely

$$\begin{aligned} KL\left(q_Z(z) || p_Z(z)\right) &= \int q_Z(z) \log\left(q_Z(z)\right) dz - \int q_Z(z) \log\left(p_Z(z)\right) dz & (10) \\ &= -H_q - \int q_Z(z) \log\left(p_Z(z)\right) dz & (11) \end{aligned}$$

where H_q denotes the (information) entropy of the prior distribution. Using equation 8, this can be rewritten as

$$\begin{aligned} KL\left(q_Z(z) || p_Z(z)\right) &= -H_q - \int q_Z(z) \log\left(\mu_X(f^{-1}(z)) \left|\det J\left(f^{-1}(z)\right)\right|\right) dz & (12) \\ &= -H_q - \int q_Z(z) \log\left(\mu_X(f^{-1}(z))\right) dz - \int q_Z(z) \log\left(\left|\det J\left(f^{-1}(z)\right)\right|\right) dz & (13) \\ &= -H_q + \mathbb{E}_{z \sim q_Z(z)}\left[-\log\left(\mu_X(f^{-1}(z))\right) - \log\left(\left|\det J\left(f^{-1}(z)\right)\right|\right)\right] & (14) \end{aligned}$$

The same result can be obtained by a KL divergence in configuration space.

$$KL\left(p_X\left(x\right)||\mu_X\left(x\right)\right) = \int p_X\left(x\right)\log\left(p_X\left(x\right)\right)dx - \int p_X\left(x\right)\log\left(\mu_X\left(x\right)\right)dx \tag{15}$$

$$= \int q_Z\left(f\left(x\right)\right)\left|\det J\left(f\left(x\right)\right)\right|\log\left(q_Z\left(f\left(x\right)\right)\left|\det J\left(f\left(x\right)\right)\right|\right)dx \tag{16}$$

$$- \int q_Z\left(f\left(x\right)\right)\left|\det J\left(f\left(x\right)\right)\right|\log\left(\mu_X\left(x\right)\right)dx \tag{17}$$

$$= \int q_Z\left(z\right)\log\left(q_Z\left(z\right)\left|\det J\left(f^{-1}\left(z\right)\right)\right|^{-1}\right)dz \tag{18}$$

$$- \int q_Z\left(z\right)\log\left(\mu_X\left(f^{-1}\left(z\right)\right)\right)dz \tag{19}$$

$$= \int q_Z\left(z\right)\log\left(q_Z\left(z\right)\right)dz - \int q_Z\left(z\right)\log\left(\left|\det J\left(f^{-1}\left(z\right)\right)\right|\right)dz \tag{20}$$

$$- \int q_Z\left(z\right)\log\left(\mu_X\left(f^{-1}\left(z\right)\right)\right)dz \tag{21}$$

$$= KL\left(q_Z(z)||p_Z(z)\right) \tag{22}$$

where substitution and the inverse function theorem $\det J\left(f^{-1}\left(z\right)\right) = \left[\det J\left(f\left(x\right)\right)\right]^{-1}$ were used. Therefore, the distributions can either be compared in the latent or in the configuration space.

Because we are dealing with real systems, we are usually interested in sampling from a Boltzmann distribution $\mu_X(x) = \frac{1}{Z_X}\exp\left\{-\frac{u(x)}{k_B T}\right\}$ in configuration space with the partition function Z_X and the Hamiltonian u (which is in our case just the potential energy) of the investigated system. A common choice for the prior distribution is a multivariate Gaussian. This simplifies the KL-Divergence to

$$KL\left(q_Z(z)||p_Z(z)\right) = -H_q + \log\left(Z_X\right) + \mathbb{E}_{z\sim q_Z(z)}\left[\left(\frac{u(f^{-1}(z))}{k_B T}\right) - \log\left(\left|\det J\left(f^{-1}(z)\right)\right|\right)\right] \tag{23}$$

Because both the entropy of the latent distribution H_q and the partition function Z of the Boltzmann distribution are constants, we can omit these terms. This leaves us with

$$KL\left(q_Z(z)||p_Z(z)\right) \sim \mathbb{E}_{z\sim q_Z(z)}\left[\left(\frac{u(f^{-1}(z))}{k_B T}\right) - \log\left(\left|\det J\left(f^{-1}(z)\right)\right|\right)\right] \tag{24}$$

The first term can be seen as the internal energy of the system as it is the ensemble average of the Hamiltonian. The second term is related to the entropy difference between the distributions $p_X(x)$ and $q_Z(z)$ as we will see in the next section.

The same loss function can be obtained by minimizing the negative average of the reweighting weights $w\left(x\right)$

$$-\int p_X\left(x\right)\log\left(w\left(x\right)\right)dx = \int p_X\left(x\right)\log\left(\frac{p_X\left(x\right)}{\mu_X\left(x\right)}\right)dx \tag{25}$$

$$= KL\left(p_X\left(x\right)||\mu_X\left(x\right)\right) \tag{26}$$

2.3 Invertible transformations and entropy

Again, we consider the same invertible transformation between the two random variables X and Z. This time, we are interested in the entropy of the distributions and their relation. The information entropy H_{P_X} of the distribution $p_X(x)$ is defined as

$$H_{p_X} = -\int_\Omega p_X(x)\log\left(p_X(x)\right)dx \tag{27}$$

Note, that this is nearly equivalent to the entropy expression in statistical mechanics, which only requires an additional factor of k_B. As the distributions are related by equation 8, substitution is again possible. This yields for the entropy

$$\begin{aligned} H_{p_X} &= -\int_{\Omega} p_X(x) \log\left(q_Z(f(x)) \left|\det J(f(x))\right|\right) dx && (28) \\ &= -\int_{\Omega} p_X(x) \log\left(q_Z(f(x))\right) dx - \int_{\Omega} p_X(x) \log\left(\left|\det J(f(x))\right|\right) dx && (29) \end{aligned}$$

The second term is just the average of $\log\left(\left|\det \frac{df(x)}{dx}\right|\right)$ with respect to the distribution $p_X(x)$, which we will take care of later. The first term is equal to the entropy H_Z of the prior distribution $q_Z(z)$. To show this, we start again by using equation 8

$$-\int_{\Omega} p_X(x) \log\left(q_Z(f(x))\right) dx = -\int_{\Omega} q_Z(f(x)) \left|\det J(f(x))\right| \log\left(q_Z(f(x))\right) dx \qquad (30)$$

Now, we are in a position to perform the change of variables with $z = f(x)$. The differentials transform as

$$dz = \left|\det J(f(x))\right| dx \qquad (31)$$

As this is the same Jacobian as in the transformation of the distributions these two terms cancel each other

$$\begin{aligned} -\int_{\Omega} q_Z(f(x)) \left|\det J(f(x))\right| \log\left(q_Z(f(x))\right) dx &= -\int_{f(\Omega)} q_Z(z) \left|\det J(f(x))\right| \log\left(q_Z(z)\right) \left|\det J(f(x))\right|^{-1} dz && (32) \\ &= -\int_{f(\Omega)} q_Z(z) \log\left(q_Z(z)\right) dz = H_q && (33) \end{aligned}$$

which is the information entropy of the distribution $q_Z(z)$. Collecting the terms, we obtain

$$H_{p_X} = H_q - \mathbb{E}_{x \sim p_X(x)} \log\left(\left|\det J(f(x))\right|\right) \qquad (34)$$

or equally

$$H_q = H_{p_X} - \mathbb{E}_{z \sim q_Z(z)} \log\left(\left|\det J\left(f^{-1}(z)\right)\right|\right) \qquad (35)$$

Therefore, the entropy difference between the two distributions connected by the transformation $f(x)$ is given by

$$\begin{aligned} H_q - H_{p_X} &= \mathbb{E}_{x \sim p_X(x)} \log\left(\left|\det J(f(x))\right|\right) && (36) \\ &= -\mathbb{E}_{z \sim q_Z(z)} \log\left(\left|\det J\left(f^{-1}(z)\right)\right|\right) && (37) \end{aligned}$$

3 KL-divergence and free energy

3.1 Helmholtz free energy

Helmholtz free energy F is defined as

$$F = U - TS \qquad (38)$$

as usual U denotes the internal energy, T the temperature and S the entropy of the system. Helmholtz free energy is the defining potential of the canonical ensemble.

3.2 Putting things together

Coming back to equation 23, we can now use equation 37 to make sense of the term with the determinant of the Jacobian

$$KL\left(q_Z(z)||p_Z(z)\right) = -H_q + \log\left(Z_X\right) + \mathbb{E}_{z\sim q_Z(z)}\left[\left(\frac{u(f^{-1}(z))}{k_BT}\right) - \log\left(\left|\det J\left(f^{-1}(z)\right)\right|\right)\right] \tag{39}$$

$$= -H_q + \log\left(Z_X\right) + \mathbb{E}_{z\sim q_Z(z)}\left[\left(\frac{u(f^{-1}(z))}{k_BT}\right)\right] + H_q - H_{p_X} \tag{40}$$

The entropy of the prior distribution cancels out and up to the constant partition function Z_X this expression looks like Helmholtz free energy divided by k_BT. And hence

$$KL\left(q_Z(z)||p_Z(z)\right) \sim \mathbb{E}_{z\sim q_Z(z)}\left[\left(\frac{u(f^{-1}(z))}{k_BT}\right)\right] - H_{p_X} \tag{41}$$

$$= \mathbb{E}_{x\sim p_X(x)}\left[\left(\frac{u(x)}{k_BT}\right)\right] - H_{p_X} \tag{42}$$

$$= \frac{U}{k_BT} - \frac{S}{k_B} = \frac{F}{k_BT} \tag{43}$$

The first term is the ensemble average of the potential energy which is the internal energy U of the system. The second term is simply the information entropy of the system which is the same as the entropy in statistical mechanics multiplied by k_B. Therefore, learning a transformation using the KL-divergence as a loss function is equivalent to minimizing the free energy F of the system, which is described by the distribution $p_x(x)$. In other words, the transformation f is optimized such that $p_x(x)$ has the minimal free energy (stationary distribution). The distribution with the minimal free energy in the canonical ensemble is the Boltzmann distribution. Accordingly, minimizing the free energy equals minimizing the difference between $p_X(x)$ and the Boltzmann distribution $\mu_X(x)$.

4 Training at different temperatures (not working atm)

4.1 Idea

Simple MCMC sampling can be improved by running simulations of the same system at different temperatures simultaneously and exchanging the configurations based on the Metropolis criterion depending on the temperatures. This is called Parallel tempering or replica exchange and improves the exploration of less likely regions.

Therefore, it is desirable to be able to do a similar thing using Boltzmann generators. In the best case, we would like to use the same transformation (and therefore also the same network) and a different prior distribution in the latent space, e.g. a Gaussian with a different variance, for the same Boltzmann distribution at different temperatures. In the following, the possibilities and limitations of this approach are discussed for different generating network architectures.

The key idea is to use the same transformation f for the training at different temperatures. Otherwise this would result in different transformations for different temperatures and nothing is gained. However, this is only possible by changing the prior distributions, i.e. using modified Boltzmann and Gaussian distributions for training at different temperatures. Currently, the loss function for different temperatures does not make sense. This is always marked with red in the following. Accordingly, we have to find a different loss function which is useful for this problem or the error in the calculations/assumptions..

4.2 General approach for arbitrary transformations

The probability density of a system is dependent on the temperature as it appears in the Boltzmann distribution μ_X, which is given by

$$\mu_X(x) = \frac{1}{Z_X} \exp\left\{-\frac{u(x)}{k_B T}\right\} \tag{44}$$

with partition function Z_X and the potential energy $u\left(x\right)$. In the latent space we have a multivariate standard Gaussian

$$q_Z\left(z\right) = \frac{1}{Z_Z} \exp\left\{-\frac{1}{2}\left\|z\right\|^{\mathbf{2}}\right\} \tag{45}$$

Samples from this distribution are transformed to the configuration space with the transformation $f^{-1}\left(z\right)$. The resulting distribution $p_X\left(x\right)$ is given due to equation 9 by

$$q_Z(z) = p_X(f^{-1}\left(z\right)) \left|\det J\left(f^{-1}\left(z\right)\right)\right| \tag{46}$$

We can rewrite equation 46 utilizing multiplicative inverse factor of the reweighting weights $\Delta(x) = w^{-1}(x)$

$$q_Z(z) = \mu_X\left(f^{-1}\left(z\right)\right) \Delta\left(f^{-1}\left(z\right)\right) \left|\det J\left(f^{-1}\left(z\right)\right)\right| \tag{47}$$

In the case of a perfectly trained network we have $\Delta(x) = 1$, but this might not be possible depending on the restrictions to f. Taking both sides of the equation to the power of $\frac{1}{\tau_k}$ yields

$$\left(q_Z(z)\right)^{\frac{1}{\tau_k}} = \left(\mu_X\left(f^{-1}\left(z\right)\right)\right)^{\frac{1}{\tau_k}} \left(\Delta\left(f^{-1}\left(z\right)\right)\right)^{\frac{1}{\tau_k}} \left|\det J\left(f^{-1}\left(z\right)\right)\right|^{\frac{1}{\tau_k}} \tag{48}$$

This operation changes the variance of each independent component of the Gaussian by τ_k, which are no longer normalized correctly. Moreover, the temperature of the Boltzmann distribution is also changed by the factor τ_k. However, even a perfectly trained network ($\Delta(x) = 1$) would not sample the correct Boltzmann distribution at a different temperature, but a modified one with the scaling factor $\left|\det J\left(f^{-1}\left(z\right)\right)\right|^{\frac{1-\tau_k}{\tau_k}}$. Nevertheless, there will always be an error $\Delta(x)$ (due to the training and the constraints to f) in a real setup. Accordingly, it is important to take a closer look at this factor. Writing down the previous equation explicitly yields

$$\frac{1}{Z_Z^{\frac{1}{\tau_k}}} \exp\left\{-\frac{1}{2}\frac{\left\|z\right\|^{\mathbf{2}}}{\tau_k}\right\} = \frac{1}{Z_X^{\frac{1}{\tau_k}}} \exp\left\{-\frac{u(f^{-1}\left(z\right))}{k_B \tau_k T}\right\} \frac{\left(p_X(f^{-1}\left(z\right))\right)^{\frac{1}{\tau_k}}}{\frac{1}{Z_X^{\frac{1}{\tau_k}}} \exp\left\{-\frac{u(f^{-1}(z))}{k_B \tau_k T}\right\}} \left|\det J\left(f^{-1}\left(z\right)\right)\right|^{\frac{1}{\tau_k}} \tag{49}$$

$$\frac{1}{Z_Z^{\frac{1}{\tau_k}}} \frac{Z_Z^{(\tau_k)}}{Z_Z^{(\tau_k)}} \exp\left\{-\frac{1}{2}\frac{\left\|z\right\|^{\mathbf{2}}}{\tau_k}\right\} = \frac{1}{Z_X^{\frac{1}{\tau_k}}} \exp\left\{-\frac{u(f^{-1}\left(z\right))}{k_B \tau_k T}\right\} \frac{\left(p_X(f^{-1}\left(z\right))\right)^{\frac{1}{\tau_k}}}{\frac{1}{Z_X^{\frac{1}{\tau_k}}} \exp\left\{-\frac{u(f^{-1}(z))}{k_B \tau_k T}\right\}} \left|\det J\left(f^{-1}\left(z\right)\right)\right|^{\frac{1-\tau_k}{\tau_k}} \left|\det J\left(f^{-1}\left(z\right)\right)\right. \tag{50}$$

$$q_Z^{(\tau_k)}(z) = \frac{Z_X^{(\tau_k)}}{Z_X^{\frac{1}{\tau_k}}} \frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}} \mu_X^{(\tau_k)}\left(f^{-1}\left(z\right)\right) \left|\det J\left(f^{-1}\left(z\right)\right)\right|^{\frac{1-\tau_k}{\tau_k}} \Delta^{(\tau_k)}\left(f^{-1}\left(z\right)\right) \left|\det J\left(f^{-1}\left(z\right)\right)\right| \tag{51}$$

where the upper index (τ_k) denotes the different temperature or variance, e.g. $q_Z^{(\tau_k)}(z)$ is a normalized multivariate Gaussian with variance τ_k and $\mu_X^{(\tau_k)}$ is a normalized Boltzmann distribution at temperature

$\tau_k T$. The error is now given by

$$\Delta^{(\tau_k)}\left(f^{-1}(z)\right) = \frac{\frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}}\left(p_X(f^{-1}(z))\right)^{\frac{1}{\tau_k}}\left|\det J\left(f^{-1}(z)\right)\right|^{\frac{1-\tau_k}{\tau_k}}}{\frac{1}{Z_X^{\frac{1}{\tau_k}}}\frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}}\exp\left\{-\frac{u(f^{-1}(z))}{k_B\tau_k T}\right\}\left|\det J\left(f^{-1}(z)\right)\right|^{\frac{1-\tau_k}{\tau_k}}} \tag{52}$$

$$= \frac{p_X^{(\tau_k)}(f^{-1}(z))}{\frac{Z_X^{(\tau_k)}}{Z_X^{\frac{1}{\tau_k}}}\frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}}\mu_X^{(\tau_k)}\left(f^{-1}(z)\right)\left|\det J\left(f^{-1}(z)\right)\right|^{\frac{1-\tau_k}{\tau_k}}} \tag{53}$$

The distribution in the numerator can be obtained by transforming samples from $q_Z^{(\tau_k)}(z)$ using the transformation f. This transformation is still the same as for the original temperature. In that way the same transformation f can be used for different temperatures, allowing simultaneous training. This is the reason for the equation to be written like that. Accordingly, the learning goal for different temperatures has slightly changed: Instead of matching $p_X^{(\tau_k)}(x)$ and $\mu_X^{(\tau_k)}(x)$ directly, $p_X^{(\tau_k)}(x)$ is matched with a modified Boltzmann dist

$$\tilde{\mu}_X^{(\tau_k)}(x) = \frac{Z_X^{(\tau_k)}}{Z_X^{\frac{1}{\tau_k}}}\frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}}\mu_X^{(\tau_k)}(x)\left|\det J(x)\right|^{\frac{1-\tau_k}{\tau_k}} \tag{54}$$

such that $\left(p_X(f^{-1}(z))\right)^{\frac{1}{\tau_k}}$ is close to $\mu_X^{(\tau_k)}(x)$ and the same transformation f can be used. However, depending on the determinant, $p_X^{(\tau_k)}(x)$ might not be a Boltzmann distribution.

As shown in eq. 26 minimizing the average of the error $\Delta^{(\tau_k)}(x)$ is equivalent to minimizing the KL-divergence. Therefore, the loss function for the modified temperature is given by

$$\int q_Z^{(\tau_k)}(z)\log\left(\Delta^{(\tau_k)}\left(f^{-1}(z)\right)\right)dz = KL\left(p_X^{(\tau_k)}(x)||\frac{Z_X^{(\tau_k)}}{Z_X^{\frac{1}{\tau_k}}}\frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}}\mu_X^{(\tau_k)}(x)\left|\det J\left(f^{-1}(z)\right)\right|^{\frac{1-\tau_k}{\tau_k}}\right) \tag{55}$$

$$\approx \mathbb{E}_{z\sim q_Z^{(\tau_k)}(z)}\left[-\log\left(\frac{Z_X^{(\tau_k)}}{Z_X^{\frac{1}{\tau_k}}}\frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}}\mu_X^{(\tau_k)}(f^{-1}(z))\left|\det J\left(f^{-1}(z)\right)\right|^{\frac{1-\tau_k}{\tau_k}}\right)\right] \tag{56}$$

$$+\mathbb{E}_{z\sim q_Z^{(\tau_k)}(z)}\left[-\log\left(\left|\det J\left(f^{-1}(z)\right)\right|\right)\right] \tag{57}$$

$$\approx \mathbb{E}_{z\sim q_Z^{(\tau_k)}(z)}\left[-\log\left(\mu_X^{(\tau_k)}(f^{-1}(z))\right) - \frac{1}{\tau_k}\log\left(\left|\det J\left(f^{-1}(z)\right)\right|\right)\right] \tag{58}$$

$$= \mathbb{E}_{z\sim q_Z^{(\tau_k)}(z)}\left[\left(\frac{u(f^{-1}(z))}{k_B T\tau_k}\right) - \frac{1}{\tau_k}\log\left(\left|\det J\left(f^{-1}(z)\right)\right|\right)\right] \tag{59}$$

$$= \frac{1}{\tau_k}\mathbb{E}_{z\sim q_Z^{(\tau_k)}(z)}\left[\left(\frac{u(f^{-1}(z))}{k_B T}\right) - \log\left(\left|\det J\left(f^{-1}(z)\right)\right|\right)\right] \tag{60}$$

where the same derivation as in section 2.2 is used and constant terms are ignored as they are not relevant in the optimization process. Minimization of this loss yields by construction the same transformation f for arbitrary temperatures $\tau_k T$. However, it obviously does not. The loss function is the same as for $\tau_k = 1$ with the exception that we sample from a different Gaussian. The minimum will be different and therefore also f!

Because the learned distributions $p_X^{(\tau_k)}(x)$ can be quite different from the Boltzmann distribution at the respective temperature, sampling will be problematic. Nevertheless, we have access to the appropriate reweighting factor. The reweighing factor is given by

$$w^{(\tau_k)}(x) = \frac{\mu_X^{(\tau_k)}(x)}{p_X^{(\tau_k)}(x)} \tag{61}$$

but because we are not trying to match these two distributions, these weights might become very large during training, resulting in bad samples.

4.3 Volume preserving transformations (not very useful)

In the case of volume preserving transformations there is a direct connection between the temperature of the system and the variance in the latent space. A volume preserving transformation can be achieved using a NICER network without additional rescaling. In this case the determinant of the Jacobian is one. This is quite a big restriction to f so it wont be possible to learn certain Boltzmann distributions. (Actually, it seems to me that only harmonic oscillators can be learned exactly.) However, this simplifies the equations quite a lot. The change of variables equation becomes

$$q_Z^{(\tau_k)}(z) = \frac{Z_X^{(\tau_k)}}{Z_X^{\frac{1}{\tau_k}}} \frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}} \mu_X^{(\tau_k)}\left(f^{-1}(z)\right) \Delta^{(\tau_k)}\left(f^{-1}(z)\right) \tag{62}$$

Therefore, a perfectly trained network (i.e. $\Delta(x) = 1$) would be able to produce samples at arbitrary temperatures by simply changing the variance of each component of the latent Gaussian distribution. This makes sense for Harmonic oscillators, but not for more complicated potential energy functions (remember that it is probably not possible to learn them anyways). The loss function for the modified temperature is now given by

$$KL\left(p_X^{(\tau_k)}(x) || \frac{Z_X^{(\tau_k)}}{Z_X^{\frac{1}{\tau_k}}} \frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}} \mu_X^{(\tau_k)}(x)\right) \approx \mathbb{E}_{z \sim q_Z^{(\tau_k)}(z)}\left[-\log\left(\mu_X^{(\tau_k)}(f^{-1}(z))\right)\right] \tag{63}$$

$$= \mathbb{E}_{z \sim q_Z^{(\tau_k)}(z)}\left[\left(\frac{u(f^{-1}(z))}{k_B T \tau_k}\right)\right] \tag{64}$$

Same problem as before...

Nevertheless, generating samples at a different temperatures is easy. As stated in equation 62, transforming the samples from a Gaussian distribution with variance τ_k will result in samples from a Boltzmann distribution (not normalized) at the temperature $T\tau_k$ with the error $\Delta^{(\tau_k)}(x)$. The error is the inverse reweighting factor and reads

$$\Delta^{(\tau_k)}(x) = \frac{(p_X(x))^{\frac{1}{\tau_k}}}{\frac{1}{Z_X^{\frac{1}{\tau_k}}} \exp\left\{-\frac{u(x)}{k_B \tau_k T}\right\}} \tag{65}$$

$$= \left(\frac{p_X(x)}{\frac{1}{Z_X} \exp\left\{-\frac{u(x)}{k_B T}\right\}}\right)^{\frac{1}{\tau_k}} \tag{66}$$

$$= \Delta^{\frac{1}{\tau_k}}(x) \tag{67}$$

Therefore, the error we are making by sampling at the different temperature depends on the error at the original temperature and will probably be much larger. As a result, having original reweighting weights $w(x)$ close to one is very important if we want to sample at different temperatures using the same transformation f.

4.4 Non-volume preserving transformations with constant Jacobian

Another special case are non-volume preserving transformations where the Jacobian does not depend on the input, i.e. $\left|\det J\left(f^{-1}(z)\right)\right| = S_{f^{-1}}$. Nevertheless, the Jacobian is still important in the minimization process, because it depends on the parameters we are optimizing for. A transformation like that can be achieved using a NICER network with rescaling. (Accordingly, only the rescaling layers contribute to the Jacobian.) The loss function, i.e. the KL-divergence, for different temperatures is similar to the general case

$$KL\left(p_X^{(\tau_k)}(x)||\frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}}\mu_X^{(\tau_k)}(x)\, S_f^{\frac{1-\tau_k}{\tau_k}}\right) \approx \mathbb{E}_{z\sim q_Z^{(\tau_k)}(z)}\left[-\log\left(\mu_X^{(\tau_k)}(f^{-1}(z))\right) - \frac{1}{\tau_k}\log\left(S_f\right)\right] \tag{68}$$

$$= \mathbb{E}_{z\sim q_Z^{(\tau_k)}(z)}\left[\left(\frac{u(f^{-1}(z))}{k_B T\tau_k}\right) - \frac{1}{\tau_k}\log\left(S_f\right)\right] \tag{69}$$

$$= \frac{1}{\tau_k}\mathbb{E}_{z\sim q_Z^{(\tau_k)}(z)}\left[\left(\frac{u(f^{-1}(z))}{k_B T}\right) - \log\left(S_f\right)\right] \tag{70}$$

Same problem as before...

Generating samples at different temperatures is again possible. The modified distributions are related by

$$q_Z^{(\tau_k)}(z) = \frac{Z_X^{(\tau_k)}}{Z_X^{\frac{1}{\tau_k}}}\frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}}\mu_X^{(\tau_k)}\left(f^{-1}(z)\right)\Delta^{(\tau_k)}\left(f^{-1}(z)\right) S_{f^{-1}}^{\frac{1-\tau_k}{\tau_k}}\, S_{f^{-1}} \tag{71}$$

Because $S_{f^{-1}}$ is only a scaling factor which does not depend on the input, the transformed samples will follow $\mu_X^{(\tau_k)}(x)$ with the error $\Delta^{(\tau_k)}(x)$. This is exactly the same as for volume preserving transformations.

4.5 Non-volume preserving transformations

In the more general case of non-volume preserving transformations sampling is more difficult. Such a transformation can be realized with RealNVP transformations or ordinary differential equation networks. The traning was already discussed (see equation 60). Same problem as before...

Sampling, however, is more difficult than before, because the determinant of the Jacobian is dependent on the input. Therefore, transformed samples of $q_Z^{(\tau_k)}(z)$ follow the distribution

$$\tilde{\mu}_X^{(\tau_k)}(x) = \frac{Z_X^{(\tau_k)}}{Z_X^{\frac{1}{\tau_k}}}\frac{Z_Z^{\frac{1}{\tau_k}}}{Z_Z^{(\tau_k)}}\mu_X^{(\tau_k)}(x)\left|\det J(x)\right|^{\frac{1-\tau_k}{\tau_k}}\Delta^{(\tau_k)}(x) \tag{72}$$

which is not normalized and even without the error not a Boltzmann distribution. Thus, reweighting is possible in theory, but wont result in good samples.