Score Matching on the Probability Simplex

To perform score matching on the probability simplex we need the gradient of the log-logit Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi v}} \frac{1}{x(1-x)} \exp\left(-\frac{(\sigma^{-1}(x) - \mu)^2}{2v}\right)$$

where $\sigma^{-1} = \log\left(\frac{x}{1-x}\right)$

Extension to larger Domains

We will find it useful to slightly generalize this to be a distribution over any simplex of length a. For now, we will work in the 1D case.

Previously, we used $\sigma(x) = \frac{1}{1+e^{-x}}$ to map points from $\mathbb{R}$ to $(0, 1)$. We can use $\sigma_a(x) = \frac{a}{1+e^{-x}}$ where points are mapped from $\mathbb{R}$ to $(0, a)$. To get the corresponding probability distribution, we use the change of variables formula

$$p_a(x) = p(\sigma_a^{-1}(x)) \frac{\partial}{\partial x} \sigma_a^{-1}(x)$$

where is this case, $p(x)$ is the standard Gaussian. Note that $\sigma_a^{-1}(x) = \log\left[\frac{x}{a-x}\right]$
We just need to compute

$$\begin{aligned} \frac{\partial}{\partial x} \sigma_a^{-1}(x) &= \frac{\partial}{\partial x} \log\left[\frac{x}{a-x}\right] \\ &= \frac{a}{x(a-x)} \end{aligned}$$

meaning that we can write a slightly more general logit-normal distribution as:

$$p_a(x) = \frac{1}{\sqrt{2\pi v}} \frac{a}{x(a-x)} \exp\left(-\frac{(\sigma_a^{-1}(x) - \mu)^2}{2v}\right)$$

Score derivation in 1D

We are interested in:
$\nabla_x \log\, p(x)$
or for the time being:
$\frac{\partial}{\partial x} \log\, p(x)$

After working in 1D, we will then show the general case.

First we deal with the log prob:

$$\begin{aligned}\frac{\partial}{\partial x}\log\, p_a(x) &= \log\, \left[\frac{1}{\sqrt{2\pi v}}\frac{a}{x(a-x)}\exp\, \left(-\frac{(\sigma_a^{-1}(x)-\mu)^2}{2v}\right)\right] \\ &= C + \log\, \left[\frac{a}{x(a-x)}\right] - \frac{(\sigma_a^{-1}(x)-\mu)^2}{2v}\end{aligned}$$

where $C = \log\, [\frac{1}{\sqrt{2\pi v}}]$

Next, we can then differentiate each of the components seperately

The first can be solved as

$$\begin{aligned}\frac{\partial}{\partial x}\log\, \left[\frac{a}{x(a-x)}\right] &= -\frac{\partial}{\partial x}\log\, [x] - \frac{\partial}{\partial x}\log\, [a-x] \\ &= -\frac{1}{x} + \frac{1}{a-x}\end{aligned}$$

and the second can be solved as

$$\begin{aligned}\frac{\partial}{\partial x} - \frac{(\sigma_a^{-1}(x)-\mu)^2}{2v} &= -\frac{1}{2v}\frac{\partial}{\partial x}\left(\sigma_a^{-1}(x)-\mu\right)^2 \\ &= -\frac{\sigma_a^{-1}(x)-\mu}{v}\frac{\partial}{\partial x}\left(\log\, \left[\frac{x}{a-x}\right] - \mu\right) \\ &= -\frac{\sigma_a^{-1}(x)-\mu}{v}\frac{a}{x(a-x)} \\ &= -\frac{a\sigma_a^{-1}(x) - a\mu}{vx(a-x)}\end{aligned}$$

Putting it all together, we get:

$$\begin{aligned}\frac{\partial}{\partial x}\log\, p_a(x) &= \frac{1}{a-x} - \frac{a}{x} - \frac{a\sigma_a^{-1}(x) - a\mu}{vx(a-x)} \\ &= -\frac{a\sigma_a^{-1}(x) - 2vx - a\mu + av}{vx(a-x)} \\ &= \frac{2vx + a\mu - a\sigma_a^{-1}(x) - av}{vx(a-x)}\end{aligned}$$

General Case

The general logit-Guassian distribution can be written as:

$$p(x) = \frac{1}{Z}\frac{1}{\prod_{i=1}^{d} x_i}\exp\, \left(-\frac{\|\log\, \left[\frac{\bar{x}_d}{x_d}\right] - \mu\|_2^2}{2v}\right)$$

where $x \in \mathcal{S}^d$ and $\bar{x}_d = [x_1, \ldots, x_{d-1}]$.
We assume that the Gaussian has covariance $\Sigma = \sqrt{v}I$

To bring x from the simplex back to $\mathbb{R}^{d-1}$ we can use:

$$y_i = \log \left[\frac{x_i}{x_d}\right], i \in \{1, \ldots, d-1\}$$

The inverse transformation of this is:

$$x_i = \frac{e^{y_i}}{1 + \sum_{k=1}^{d-1} e^{y_k}}, i \in \{1, \ldots, d-1\}$$

$$x_d = \frac{1}{1 + \sum_{k=1}^{d-1} e^{y_k}} = 1 - \sum_{i=1}^{d-1} x_i$$

Extension to Larger Domain

Yet again, we are interested in extending this definition to work for large domains.

First we must find the function to map from $\mathbb{R}^d$ to $\mathcal{S}_a^d$, which is the simplex of length a. To begin, we define

$$x_i = \sigma_a(y_i) = a\sigma(y_i), i \in \{1, \ldots, d-1\}$$
$$x_d = a - \sum_{i=1}^{d-1} x_i$$

where σ is the previous tranformation from $\mathbb{R}^d$ to $\mathcal{S}_a^d$. The inverse can easily be seen to be the same as the previous example, meaning that:

$$\sigma_a^{-1}(\mathbf{x}) = \sigma^{-1}(\mathbf{x})$$

Working in more dimensions, the change of variables formula is:

$$p_a(\mathbf{x}) = p(\sigma_a^{-1}(\mathbf{x}))\left|\det \frac{\partial}{\partial \mathbf{x}} \sigma_a^{-1}(\mathbf{x})\right|$$

Thus, our next step is to find this log det term. We shall do so by first getting the Jacobian matrix into a convenient form.

$$\begin{aligned}
\frac{\partial}{\partial x_j}\sigma_a^{-1}(x)_i &= \frac{\partial}{\partial x_j}\log\ \left[\frac{x_i}{x_d}\right] \\
&= \frac{\partial}{\partial x_j}\log\ [x_i] - \frac{\partial}{\partial x_j}\log\ \left[a - \sum_{i=1}^{d-1} x_i\right] \\
&= \delta_{ij}\frac{1}{x_i} - \frac{\partial}{\partial u}\log\ [u]\frac{\partial}{\partial x_j}u, u = a - \sum_{i=1}^{d-1} x_i \\
&= \delta_{ij}\frac{1}{x_i} + \frac{1}{x_d}
\end{aligned}$$

Now that we have the Jacobian, we can calculate the determinant by using the fact that a diagonal matrix D plus a constant matrix C has the following determinant:

$$\det\ (D + C) = \left(1 + c\sum_{i=1}^{n} d_i^{-1}\right)\prod_{i=1}^{n} d_i$$

In our case we have the following:

$$\begin{aligned}
\left|\det\ \frac{\partial}{\partial \mathbf{x}}\sigma_a^{-1}(\mathbf{x})\right| &= \left(1 + \frac{1}{x_d}\sum_{i=1}^{d-1} x_i\right)\prod_{i=1}^{d-1}\frac{1}{x_i} \\
&= \left(1 + \frac{1}{x_d}(a - x_d)\right)\prod_{i=1}^{d-1}\frac{1}{x_i} \\
&= \prod_{i=1}^{d}\frac{a}{x_i}
\end{aligned}$$

We now have all the ingredients to make our slightly more general logit-Gaussian distribution:

$$p_a(x) = \frac{1}{(2\pi)^{(d-1)/2}}\frac{a}{\prod_{i-1}^{d} x_i}\exp\ \left(-\frac{1}{2v}\left\|\log\ \left[\frac{\bar{x}_d}{x_d}\right] - \mu\right\|_2^2\right)$$

Derivation of Score

Overall, we want to calculate:
$\nabla_x \log\ p_a(x)$

Following the same process as the 1D case:

$$\log\, p_a(x) = -\log\,[Z] - \log\,\left[\prod_{i-1}^{d} x_i\right] - \frac{1}{2v}\left\|\log\,\left[\frac{\bar{x}_d}{x_d}\right] - \mu\right\|_2^2$$

We deal with the gradients, starting with the second term (the first one has no gradient).

$$\begin{aligned}
g &:= -\nabla_x \log\,\left[\prod_{i=1}^{d} x_i\right] \\
g_i &= -\frac{\partial}{\partial x_i}\left(\sum_{i=1}^{d-1}\log\,[x_i] + \log\,\left[a - \sum_{k=1}^{d-1} x_k\right]\right) \\
&= -\frac{1}{x_i} + \frac{1}{a - \sum_{k=1}^{d-1} x_k} \\
&= \frac{1}{x_d} - \frac{1}{x_i} \\
&= \frac{x_i - x_d}{x_i x_d}
\end{aligned}$$

Next, we deal with the exponential term:

$$\begin{aligned}
h &:= -\frac{1}{2v}\nabla_x \left\|\log\,\left[\frac{\bar{x}_d}{x_d}\right] - \mu\right\|_2^2 \\
h_i &= -\frac{1}{2v}\frac{\partial}{\partial x_i}\left(\sum_{k=1}^{d-1}\left(\log\,\left[\frac{x_k}{x_d}\right] - \mu\right)^2\right) \\
&= -\frac{1}{2v}\sum_{k=1}^{d-1}\left(\frac{\partial}{\partial u}u^2\frac{\partial}{\partial x_i}u\right), u = \log\,\left[\frac{x_k}{x_d}\right] - \mu
\end{aligned}$$

We can just focus on $\beta := \frac{\partial}{\partial u}u^2\frac{\partial}{\partial x_i}u$ for now

$$\begin{aligned}
\beta &:= \frac{\partial}{\partial u}u^2\frac{\partial}{\partial x_i}u \\
&= 2u\left(\frac{\partial}{\partial x_i}\log\,[x_k] - \frac{\partial}{\partial x_i}\log\left[a - \sum_{k=1}^{d-1} x_k\right]\right) \\
&= 2u\left(\delta_{ik}\frac{1}{x_i} + \frac{1}{x_d}\right)
\end{aligned}$$

Combining terms we get:

$$
\begin{aligned}
h_i &= -\frac{1}{v}\sum_{k=1}^{d-1}\left(\delta_{ik}\frac{1}{x_i}+\frac{1}{x_d}\right)\left(\log\left[\frac{\bar{x}_d}{x_d}\right]-\mu\right) \\
&= -\frac{1}{vx_d}\sum_{k=1}^{d-1}\left(\log\left[\frac{x_k}{x_d}\right]-\mu\right)-\frac{1}{vx_i}\left(\log\left[\frac{x_i}{x_d}\right]-\mu\right) \\
&= -\frac{1}{vx_d}\sum_{k=1}^{d-1}\gamma_\mu^k(\mathbf{x})-\frac{1}{vx_i}\gamma_\mu^i(\mathbf{x})
\end{aligned}
$$

where we write $\gamma_\mu^i(\mathbf{x}) = \log\left[\frac{x_i}{x_d}\right]-\mu$

For the final results, we must combine the h and g terms together to get:

$$
\nabla_x \log\, p_a(x)_i = -\frac{1}{vx_d}\sum_{k=1}^{d-1}\gamma_\mu^k(\mathbf{x})-\frac{1}{vx_i}\gamma_\mu^i(\mathbf{x})+\frac{x_i-x_d}{x_ix_d}
$$

Sampling and Ito's Lemma

We are working with an OU process of the following form:

$$
d\mathbf{X}_t = -\theta\mathbf{X}_t dt + d\mathbf{B}_t
$$

with a corresponding process:

$$
\mathbf{S}_t = \sigma^a(\mathbf{X}_t)
$$

To keep this section self-contained the definition of σ_a is:

$$
\sigma_i^a(\mathbf{y}) = \frac{ae^{y_i}}{1+\sum_{k=1}^{d-1}e^{y_k}}, i \in \{1,\ldots,d-1\}
$$

To sample from our model, we must write S_t in a form where $S_t = f(x,t)dt + g(x,t)dB_t$. This can be done via Ito's Lemma:

$$
dS_i = -\theta(\nabla_X\sigma_i^a(\mathbf{X}))^\top\mathbf{X}dt + \frac{1}{2}\mathrm{Tr}[H_X\sigma_i^a(\mathbf{X})]dt + \nabla_X\sigma_i^a(\mathbf{X})^\top d\mathbf{B}
$$

Where H_X is the Hessian matrix and we drop the time dependence of $\mathbf{S}_t$ and $\mathbf{X}_t$ for notational simplicity.

First we deal with the gradient term of the equation. We will use $\gamma(\mathbf{X}) = 1 + \sum_{k=1}^{d-1} e^{X_k}$ to keep notation smaller.

$$\nabla_X \sigma_i^a(\mathbf{X}) := \mathbf{g} = \nabla_X \frac{ae^{X_i}}{\gamma(\mathbf{X})}$$
$$g_j = \frac{\partial}{\partial X_j} \frac{ae^{X_i}}{\gamma(\mathbf{X})}$$

We deal with the case when when $j = i$ below

$$\begin{aligned}
g_i &= \frac{\partial}{\partial X_i} \frac{ae^{X_i}}{\gamma(\mathbf{X})} \\
&= \gamma(\mathbf{X})^{-2} \left[\gamma(\mathbf{X}) \frac{\partial}{\partial X_i} ae^{X_i} - ae^{X_i} \frac{\partial}{\partial X_i} \gamma(\mathbf{X}) \right] \\
&= \gamma(\mathbf{X})^{-2} \left[ae^{X_i} \gamma(\mathbf{X}) - ae^{2X_i} \right] \\
&= a\sigma_i(\mathbf{X}) \gamma(\mathbf{X})^{-1} [\gamma(\mathbf{X}) - e^{X_i}] \\
&= a\sigma_i(\mathbf{X})(1 - \sigma_i(\mathbf{X}))
\end{aligned}$$

and the case when $j \neq i$:

$$\begin{aligned}
g_j &= \frac{\partial}{\partial X_j} \frac{ae^{X_i}}{\gamma(\mathbf{X})} \\
&= -\frac{ae^{X_i} e^{X_j}}{\gamma(X)^2} \\
&= -a\sigma_i(\mathbf{X})\sigma_j(\mathbf{X})
\end{aligned}$$

Next we deal with the trace Hessian term:

$$\mathrm{Tr}[H_X \sigma_i^a(\mathbf{X})] = \sum_{j=1}^{d-1} \frac{\partial^2}{\partial X_j^2} \sigma_i^a(\mathbf{X})$$

which again can be split into two cases. First we deal with the case when $j = i$

$$\begin{aligned}
\frac{\partial^2}{\partial X_i^2} \sigma_i^a(\mathbf{X}) &= a \frac{\partial}{\partial X_i} \sigma_i(\mathbf{X})(1 - \sigma_i(\mathbf{X})) \\
&= a\sigma_i(\mathbf{X})(1 - \sigma_i(\mathbf{X}))(1 - 2\sigma_i(\mathbf{X}))
\end{aligned}$$

Then the case where $j \neq i$

$$\begin{aligned}
\frac{\partial^2}{\partial X_j^2} \sigma_i^a(\mathbf{X}) &= -a \frac{\partial}{\partial X_j} \sigma_i(\mathbf{X})\sigma_j(\mathbf{X}) \\
&= -a^2 \sigma_i(\mathbf{X})\sigma_j(\mathbf{X})(1 - 2\sigma_j(\mathbf{X}))
\end{aligned}$$

Vectorized Implimentation

We would like a vectorized implimentation in the following form:

$$d\mathbf{S}_t = -\theta J\mathbf{X}_t dt + \frac{1}{2}hdt + J\mathbf{B}_t$$

where $J_i = \nabla_X \sigma_i^a(\mathbf{X}_t)$ is the vectorized gradient (Jacobian), and $h_i = \mathrm{Tr}[H_X \sigma_i^a(\mathbf{X_t})]$

Beginning with J we have:

$$J_{ii} = a\sigma_i(\mathbf{X})(1-\sigma_i(\mathbf{X})), i = j$$
$$J_{ij} = -a\sigma_i(\mathbf{X})\sigma_j(\mathbf{X}), i \neq j$$

Leaving us with a vectorized version:

$$J = a\mathbf{S}_t\mathbf{1}^\top \odot \mathrm{diag}_1\left([1-\mathbf{S}_t]\mathbf{1}^\top\right) \odot \mathrm{diag}^1\left(\mathbf{1}^\top\mathbf{S}_t\right)$$

where $\mathrm{diag}_1()$ take a matrix and sets all values besides the diagonal to 1 while keeping other elemets the same. Similarily, $\mathrm{diag}^1()$ takes a matrix and sets the diagonals to 1 and keeps everything else the same.

For the Hessian term h we have do the following:

$$h_i = a\sigma_i(\mathbf{X})(1-\sigma_i(\mathbf{X}))(1-2\sigma_i(\mathbf{X})) + \sum_{j\neq i}^{d-1} -a^2\sigma_i(\mathbf{X})\sigma_j(\mathbf{X})(1-2\sigma_j(\mathbf{X}))$$

and can vectorize this to:

$$h = a\mathbf{S}_t(1-\mathbf{S}_t)(1-2\mathbf{S}_t) + \mathrm{diag}^0(\mathbf{S}_t\mathbf{1}^\top)(\mathbf{S}_t[1-2\mathbf{S}_t])$$

where $\mathrm{diag}^0()$ takes a matrix and sets all values in the the diagonal to 0 while keeping other elemets the same.