Supplementary Material for "Scoring and Classifying with Gated Auto-encoders"

Daniel Jiwoong Im, and Graham W. Taylor

School of Engineering
University of Guelph
Guelph, On, Canada
{imj,gwtaylor}@uoguelph.ca

1 Gated Auto-encoder Scoring

1.1 Vector field representation

To check that the vector field can be written as the derivative of a scalar field, we can submit to Poincaré's integrability criterion: For some open, simple connected set $\mathcal{U}$, a continuously differentiable function $F : \mathcal{U} \rightarrow \Re^m$ defines a gradient field if and only if

$$\frac{\partial F_i(\mathbf{y})}{\partial y_j} = \frac{\partial F_j(\mathbf{y})}{\partial y_i}, \ \forall i, j = 1 \cdots n.$$

Considering the GAE, note that i^{th} component of the decoder $r_i(\mathbf{y}|\mathbf{x})$ can be rewritten as

$$r_i(\mathbf{y}|\mathbf{x}) = (W^Y_{.i})^T(W^X\mathbf{x} \odot (W^H)^T h(\mathbf{y}, \mathbf{x})) = (W^Y_{.i} \odot W^X\mathbf{x})^T(W^H)^T h(\mathbf{y}, \mathbf{x}).$$

The derivatives of $r_i(\mathbf{y}|\mathbf{x}) - y_i$ with respect to y_j are

$$\begin{aligned}
\frac{\partial r_i(\mathbf{y}|\mathbf{x})}{\partial y_j} =& (W^Y_{.i} \odot W^X\mathbf{x})^T(W^H)^T \frac{\partial h(\mathbf{x}, \mathbf{y})}{\partial y_j} = \frac{\partial r_j(\mathbf{y}|\mathbf{x})}{\partial y_i} \\
\frac{\partial h(\mathbf{y}, \mathbf{x})}{\partial y_j} =& \frac{\partial h(\mathbf{u})}{\partial \mathbf{u}} W^H(W^Y_{.j} \odot W^X\mathbf{x}) \qquad (1)
\end{aligned}$$

where $\mathbf{u} = W^H((W^Y\mathbf{y}) \odot (W^X\mathbf{x}))$. By substituting Equation 1 into $\frac{\partial F_i}{\partial y_j}, \frac{\partial F_j}{\partial y_i}$, we have

$$\frac{\partial F_i}{\partial y_j} = \frac{\partial r_i(\mathbf{y}|\mathbf{x})}{\partial y_j} - \delta_{ij} = \frac{\partial r_j(\mathbf{y}|\mathbf{x})}{\partial y_i} - \delta_{ij} = \frac{\partial F_j}{\partial y_i}$$

where $\delta_{ij} = 1$ for $i = j$ and 0 for $i \neq j$. Similarly, the derivatives of $r_i(\mathbf{y}|\mathbf{x}) - y_i$ with respect to x_j are

$$\begin{aligned}
\frac{\partial r_i(\mathbf{y}|\mathbf{x})}{\partial x_j} =& (W^Y_{.i} \odot W^X_{.j})^T(W^H)^T h(\mathbf{x}, \mathbf{y}) + (W^Y_{.i} \odot W^X\mathbf{x})(W^H)^T \frac{\partial h}{\partial x_j} = \frac{\partial r_j(\mathbf{y}|\mathbf{x})}{\partial x_i}, \\
\frac{\partial h(\mathbf{y}, \mathbf{x})}{\partial x_j} =& \frac{\partial h(\mathbf{u})}{\partial \mathbf{u}} W^H(W^Y_{.j} \odot W^X\mathbf{x}). \qquad (2)
\end{aligned}$$

By substituting Equation 2 into $\frac{\partial F_i}{\partial x_j}, \frac{\partial F_j}{\partial x_i}$, this yields

$$\frac{\partial F_i}{\partial x_j} = \frac{\partial r_i(\mathbf{x}|\mathbf{y})}{\partial x_j} = \frac{\partial r_j(\mathbf{x}|\mathbf{y})}{\partial x_i} = \frac{\partial F_j}{\partial x_i}.$$

1.2 Deriving an Energy Function

Integrating out the GAE's trajectory, we have

$$\begin{aligned} E(\mathbf{y}|\mathbf{x}) &= \int_{\mathcal{C}} (r(\mathbf{y}|\mathbf{x}) - \mathbf{y}) d\mathbf{y} \\ &= \int W^Y \left(\left(W^X \mathbf{x} \right) \odot W^H h(\mathbf{u}) \right) d\mathbf{y} - \int \mathbf{y} d\mathbf{y} \\ &= W^Y \left(\left(W^X \mathbf{x} \right) \odot W^H \int h\left(\mathbf{u} \right) d\mathbf{u} \right) - \int \mathbf{y} d\mathbf{y}, \end{aligned} \tag{3}$$

where $\mathbf{u}$ is an auxiliary variable such that $\mathbf{u} = W^H((W^Y\mathbf{y}) \odot (W^X\mathbf{x}))$ and $\frac{d\mathbf{u}}{d\mathbf{y}} = W^H(W^Y \odot (W^X\mathbf{x} \otimes \mathbf{1}_D))$, where $\otimes$ is the Kronecker product. Consider the symmetric objective function, which is defined in Equation **??**. Then we have to also consider the vector field system where both symmetric cases $\mathbf{x}|\mathbf{y}$ and $\mathbf{y}|\mathbf{x}$ are valid. As mentioned in Section 3.1, let $\xi = [\mathbf{x}; \mathbf{y}]$ and $\gamma = [\mathbf{y}; \mathbf{x}]$. As well, let $W^\xi = \text{diag}(W^X, W^Y)$ and $W^\gamma = \text{diag}(W^Y, W^X)$ where they are block diagonal matrices. Consequently, the vector field becomes

$$F(\boldsymbol{\xi}|\boldsymbol{\gamma}) = r(\boldsymbol{\xi}|\boldsymbol{\gamma}) - \boldsymbol{\xi}, \tag{4}$$

and the energy function becomes

$$\begin{aligned} E(\boldsymbol{\xi}|\boldsymbol{\gamma}) &= \int (r(\boldsymbol{\xi}|\boldsymbol{\gamma}) - \boldsymbol{\xi}) d\boldsymbol{\xi} \\ &= \int (W^\xi)^T ((W^\gamma \boldsymbol{\gamma}) \odot (W^H)^T h(\boldsymbol{u})) d\boldsymbol{\xi} - \int \boldsymbol{\xi} d\boldsymbol{\xi} \\ &= (W^\xi)^T ((W^\gamma \boldsymbol{\gamma}) \odot (W^H)^T \int h(\boldsymbol{u}) d\boldsymbol{u}) - \int \boldsymbol{\xi} d\boldsymbol{\xi} \end{aligned}$$

where $\boldsymbol{u}$ is an auxiliary variable such that $\boldsymbol{u} = W^H \left((W^\xi \boldsymbol{\xi}) \odot (W^\gamma \boldsymbol{\gamma}) \right)$. Then

$$\frac{d\boldsymbol{u}}{d\boldsymbol{\xi}} = W^H \left(W^\xi \odot (W^\gamma \boldsymbol{\gamma} \otimes \mathbf{1}_D) \right).$$

Moreover, note that the decoder can be re-formulated as

$$\begin{aligned} r(\boldsymbol{\xi}|\boldsymbol{\gamma}) &= (W^\xi)^T (W^\gamma \boldsymbol{\gamma} \odot (W^H)^T h(\boldsymbol{\xi}, \boldsymbol{\gamma})) \\ &= \left((W^\xi)^T \odot (W^\gamma \boldsymbol{\gamma} \otimes \mathbf{1}_D) \right) (W^H)^T h(\boldsymbol{\xi}, \boldsymbol{\gamma}). \end{aligned}$$

Re-writing the first term of Equation 3 in terms of the auxiliary variable $\boldsymbol{u}$, the energy reduces to

$$\begin{aligned}
E(\boldsymbol{\xi}|\boldsymbol{\gamma}) &= \left((W^{\xi})^T \odot (W^{\gamma}\boldsymbol{\gamma} \otimes \mathbf{1}_D)\right)(W^H)^T \int h(\boldsymbol{u}) \left(W^H(W^{\xi} \odot (W^{\gamma}\boldsymbol{\gamma} \otimes \mathbf{1}_D))\right)^{-1} d\boldsymbol{u} - \int \boldsymbol{\xi} d\boldsymbol{\xi} \\
&= \left((W^{\xi})^T \odot (W^{\gamma}\boldsymbol{\gamma} \otimes \mathbf{1}_D)\right)(W^H)^T \left((W^{\xi} \odot (W^{\gamma}\boldsymbol{\gamma} \otimes \mathbf{1}_D))W^H\right)^{-T} \int h(\boldsymbol{u}) d\boldsymbol{u} - \int \boldsymbol{\xi} d\boldsymbol{\xi} \\
&= \int h(\boldsymbol{u}) d\boldsymbol{u} - \int \boldsymbol{\xi} d\boldsymbol{\xi} \\
&= \int h(\boldsymbol{u}) d\boldsymbol{u} - \frac{1}{2}\boldsymbol{\xi}^2 + \text{const.}
\end{aligned}$$

2 Relation to other types of Restricted Boltzmann Machines

2.1 Gated Auto-encoder and Factored Gated Conditional Restricted Boltzmann Machines

Suppose that the hidden activation function is a sigmoid. Moreover, we define our Gated Auto-encoder to consists of an encoder $h(\cdot)$ and decoder $r(\cdot)$ such that

$$\begin{aligned}
h(\mathbf{x}, \mathbf{y}) &= h(W^H((W^X\mathbf{x}) \odot (W^Y\mathbf{y})) + \mathbf{b}) \\
r(\mathbf{x}|\mathbf{y}, h) &= (W^X)^T((W^Y\mathbf{y}) \odot (W^H)^T h(\mathbf{x}, \mathbf{y})) + \mathbf{a},
\end{aligned}$$

where $\theta = \{W^H, W^X, W^Y, \mathbf{b}\}$ is the parameters of the model. Note that the weights are not tied in this case. The energy function for the Gated Auto-encoder will be:

$$\begin{aligned}
E_{\sigma}(\mathbf{x}|\mathbf{y}) &= \int (1 + \exp\left(-W^H(W^X\mathbf{x}) \odot (W^Y\mathbf{y}) - \mathbf{b}\right))^{-1} d\mathbf{u} - \frac{\mathbf{x}^2}{2} + \mathbf{a}\mathbf{x} + \text{const} \\
&= \sum_k \log\left(1 + \exp\left(-W^H_{k\cdot}(W^X\mathbf{x}) \odot (W^Y\mathbf{y}) - b_k\right)\right) - \frac{\mathbf{x}^2}{2} + \mathbf{a}\mathbf{x} + \text{const.}
\end{aligned}$$

Now consider the free energy of a Factored Gated Conditional Restricted Boltzmann Machine (FCRBM).

The energy function of a FCRBM with Gaussian visible units and Bernoulli hidden units is defined by

$$E(\mathbf{x}, \mathbf{h}|\mathbf{y}) = \frac{(\mathbf{a} - \mathbf{x})^2}{2\sigma^2} - \mathbf{b}\mathbf{h} - \sum_f W^X_{f\cdot}\mathbf{x} \odot W^Y_{f\cdot}\mathbf{y} \odot W^H_{f\cdot}\mathbf{h}.$$

Given $\mathbf{y}$, the conditional probability density assigned by the FCRBM to data point $\mathbf{x}$ is

$$p(\mathbf{x}|\mathbf{y}) = \frac{\sum_{\mathbf{h}} \exp -(E(\mathbf{x}, \mathbf{h}|\mathbf{y}))}{Z(\mathbf{y})} = \frac{\exp\left(-F(\mathbf{x}|\mathbf{y})\right)}{Z(\mathbf{y})}$$
$$-F(\mathbf{x}|\mathbf{y}) = \log\left(\sum_{\mathbf{h}} \exp\left(-E(\mathbf{x}, \mathbf{h}|\mathbf{y})\right)\right)$$

where $Z(\mathbf{y}) = \sum_{\mathbf{x},\mathbf{h}} \exp\left(E(\mathbf{x}, \mathbf{h}|\mathbf{y})\right)$ is the partition function and $F(\mathbf{x}|\mathbf{y})$ is the free energy function. Expanding the free energy function, we get

$$\begin{aligned}
-F(\mathbf{x}|\mathbf{y}) &= \log \sum_{\mathbf{h}} \exp\left(-E(\mathbf{x}, \mathbf{h}|\mathbf{y})\right) \\
&= \log \sum_{\mathbf{h}} \exp\left(\frac{-(\mathbf{a}-\mathbf{x})^2}{2\sigma^2} + \mathbf{bh} + \sum_f W^X_{f\cdot}\mathbf{x} \odot W^Y_{f\cdot}\mathbf{y} \odot W^H_{f\cdot}\mathbf{h}\right) \\
&= -\frac{(\mathbf{a}-\mathbf{x})^2}{2\sigma^2} + \log\left(\sum_{\mathbf{h}} \exp\left(\mathbf{bh} + \sum_f W^X_{f\cdot}\mathbf{x} \odot W^Y_{f\cdot}\mathbf{y} \odot W^H_{f\cdot}\mathbf{h}\right)\right) \\
&= -\frac{(\mathbf{a}-\mathbf{x})^2}{2\sigma^2} + \log\left(\sum_{\mathbf{h}} \prod_k \exp\left(b_k h_k + \sum_f (W^X_{f\cdot}\mathbf{x} \odot W^Y_{f\cdot}\mathbf{y}) \odot W^H_{fk} h_k\right)\right) \\
&= -\frac{(\mathbf{a}-\mathbf{x})^2}{2\sigma^2} + \sum_k \log\left(1 + \exp\left(b_k + \sum_f \left((W^H_{fk})^T (W^X\mathbf{x} \odot W^Y\mathbf{y})\right)\right)\right).
\end{aligned}$$

Note that we can center the data by subtracting mean of $\mathbf{x}$ and dividing by its standard deviation, and therefore assume that $\sigma^2 = 1$. Substituting, we have

$$\begin{aligned}
-F(\mathbf{x}|\mathbf{y}) &= -\frac{(\mathbf{a}-\mathbf{x})^2}{2} + \sum_k \log\left(1 + \exp\left(-b_k - \sum_f (W^H_{fk})^T (W^X\mathbf{x} \odot W^Y\mathbf{y})\right)\right) \\
&= \sum_k \log\left(1 + \exp\left(b_k + \sum_f (W^H_{fk})^T (W^X\mathbf{x} \odot W^Y\mathbf{y})\right)\right) - \mathbf{a}^2 + \mathbf{ax} - \frac{\mathbf{x}^2}{2} \\
&= \sum_k \log\left(1 + \exp\left(b_k + \sum_f (W^H_{fk})^T (W^X\mathbf{x} \odot W^Y\mathbf{y})\right)\right) + \mathbf{ax} - \frac{\mathbf{x}^2}{2} + \text{const}
\end{aligned}$$

Letting $W^H = (W^H)^T$, we get

$$= \sum_k \log\left(1 + \exp\left(b_k + \sum_f W^H_{kf} (W^X\mathbf{x} \odot W^Y\mathbf{y})\right)\right) + \mathbf{ax} - \frac{\mathbf{x}^2}{2} + \text{const}$$

Hence, the Conditional Gated Auto-encoder and the FCRBM are equal up to a constant.

2.2 Gated Auto-encoder and mean-covariance Restricted Boltzmann Machines

Theorem 1. *Consider a covariance auto-encoder with an encoder and decoder,*

$$h(\mathbf{x}, \mathbf{x}) = h(W^H((W^F\mathbf{x})^2) + \mathbf{b})$$
$$r(\mathbf{x}|\mathbf{y} = \mathbf{x}, h) = (W^F)^T(W^F\mathbf{y} \odot (W^H)^T h(\mathbf{x}, \mathbf{y})) + \mathbf{a},$$

where $\theta = \{W^F, W^H, \mathbf{a}, \mathbf{b}\}$ are the parameters of the model. Moreover, consider a covariance Restricted Boltzmann Machine with Gaussian distribution over the visibles and Bernoulli distribution over the hiddens, such that its energy function is defined by

$$E^c(\mathbf{x}, \mathbf{h}) = \frac{(\mathbf{a} - \mathbf{x})^2}{\sigma^2} - \sum_f P\mathbf{h}(C\mathbf{x})^2 - \mathbf{bh},$$

where $\theta = \{P, C, \mathbf{a}, \mathbf{b}\}$ are its parameters. Then the energy function for a covariance Auto-encoder with dynamics $r(\mathbf{x}|\mathbf{y}) - \mathbf{x}$ is equivalent to the free energy of a covariance Restricted Boltzmann Machine. The energy function of the covariance Auto-encoder is

$$E(\mathbf{x}, \mathbf{x}) = \sum_k \log(1 + \exp(W^H(W^F\mathbf{x})^2 + \mathbf{b})) - \mathbf{x}^2 + const \tag{5}$$

Proof. Note that the covariance auto-encoder is the same as a regular Gated Auto-encoder, but setting $\mathbf{y} = \mathbf{x}$ and making the factor loading matrices the same, i.e. $W^F = W^Y = W^X$. Then applying the general energy equation for GAE, Equation **??**, to the covariance auto-encoder, we get

$$\begin{aligned} E(\mathbf{x}, \mathbf{x}) &= \int h(\mathbf{u})d\mathbf{u} - \frac{1}{2}\mathbf{x}^2 + \text{const} \\ &= \sum_k \log(1 + \exp(W^H(W^F\mathbf{x})^2 + \mathbf{b})) - \mathbf{x}^2 + \mathbf{ax} + \text{const}, \end{aligned} \tag{6}$$

where $\mathbf{u} = W^H(W^F\mathbf{x})^2 + \mathbf{b}$.

Now consider the free energy of the mean-covariance Restricted Boltzmann Machine (mcRBM) with Gaussian distribution over the visible units and Bernoulli distribution over the hidden units:

$$\begin{aligned} -F(\mathbf{x}|\mathbf{y}) &= \log \sum_{\mathbf{h}} \exp\left(-E(\mathbf{x}, \mathbf{h}|\mathbf{y})\right) \\ &= \log \sum_h \exp\left(-\frac{(\mathbf{a} - \mathbf{x})^2}{\sigma^2} + (P\mathbf{h})(C\mathbf{x})^2 + \mathbf{bh}\right) \\ &= \log \sum_h \prod_k \exp\left(-\frac{(\mathbf{a} - \mathbf{x})^2}{\sigma^2} + \sum_f (P_{fk}h_k)(C\mathbf{x})^2 + b_k h_k\right) \\ &= \sum_k \log\left(1 + \exp\left(\sum_f (P_{fk}h_k)(C\mathbf{x})^2\right)\right) - \frac{(\mathbf{a} - \mathbf{x})^2}{\sigma^2}. \end{aligned}$$

As before, we can center the data by subtracting mean of $\mathbf{x}$ and dividing by its standard deviation, and therefore assume that $\sigma^2 = 1$. Substituting, we have

$$= \sum_k \log\left(1 + \exp\left(\sum_f (P_{fk} h_k)(C\mathbf{x})^2\right)\right) - (\mathbf{a} - \mathbf{x})^2. \tag{7}$$

Letting $W^H = P^T$ and $W^F = C$, we get

$$= \sum_k \log\left(1 + \exp\left(\sum_f (P_{fk} h_k)(C\mathbf{x})^2\right)\right) - \mathbf{x}^2 + \mathbf{a}\mathbf{x} + \text{const.} \tag{8}$$

Therefore, the two equations are equivalent. □