1.1.1

We first show that $0 < \sigma'(x) \leq \frac{1}{4} \quad \forall x$.

$$\sigma'(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) = -(1+e^{-x})^{-2}(-e^{-x})$$

$$= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \sigma(x)(1-\sigma(x))$$

Since $\sigma(x) \in (0,1)$, then $\sigma(x)(1-\sigma(x)) > 0$

and $\sigma(x)(1-\sigma(x)) \leq \frac{[\sigma(x)+1-\sigma(x)]^2}{4} = \frac{1}{4}$

by basic inequality. □

Denote $f_n(x)$ as the output of x after n layers.

We then prove by induction that $0 < f_n'(x) \leq (\frac{1}{4})^n \quad \forall n \in \mathbb{N}^+$

- Base Case:

$f_1(x) = \sigma(x+b) \Rightarrow f'(x) = \sigma'(x+b) \in (0, \frac{1}{4}]$

- Induction Step:

Fix $k \in \mathbb{N}$. Assume $f_k(x) \in (0, (\frac{1}{4})^k]$.

Given $f_{k+1}(x) = \sigma(f_k(x) + b_k)$, where b_k is a constant bias.

there is $f'_{k+1}(x) = \underbrace{\sigma'(f_k(x)+b_k)}_{\in (0, \frac{1}{4}]} \underbrace{f'_k(x)}_{\in (0, (\frac{1}{4})^k]}$

thus $f'_{k+1}(x) \in (0, (\frac{1}{4})^{k+1}]$.

Thus, $f_n'(x) \in (0, (\frac{1}{4})^n] \ \forall n \in \mathbb{N}^+$

and $\left|\frac{\partial f_n(x)}{\partial x}\right| \in [0, (\frac{1}{4})^n] \ \forall n \in \mathbb{N}^+$.

$\Rightarrow \lim_{n\to\infty} \left|\frac{\partial f_n(x)}{\partial x}\right| = 0$ by squeeze theorem.

There would be gradient vanishing issue if activation function is Sigmoid.

1.1.2

$$\tanh'(x) = \operatorname{sech}^2(x) = \frac{1}{\cosh^2(x)} = \left(\frac{2}{e^x + e^{-x}}\right)^2$$

Let $g(x) = e^x + e^{-x}$, then $g'(x) = e^x - e^{-x}$ which is positive for $x > 0$ and negative for $x < 0$.

Thus $g(x)$ has a minimum of 2 at $x = 0$.

i.e. $g(x) \geq 2 \Rightarrow \quad 0 < \tanh'(x) \leq 1$

Similar to 1.1.1, we can show that $0 < f_n'(x) \leq 1$ which doesn't necessarily inform anything.

But notice that $f_n'(x)$ can be expressed as the product of n $\tanh'$ functions multiplied

together, and it's highly probable that many values among these tanh' functions are smaller than 1 as $n \to \infty$, causing the gradient vanishing problem.

1.2.1

We first state that the largest singular value of the Jacobian Matrix of the elementwise function $\tanh(x): R^k \to R^k$ does not exceed 1. (1)

Proof:

The Jacobian Matrix of $\tanh(x_1, \dots, x_k)$ $= (\tanh(x_1), \dots, \tanh(x_k))$ is

$$J = \begin{pmatrix} \tanh'(x_1) & 0 & \cdots & 0 \\ 0 & \tanh'(x_2) & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \tanh'(x_k) \end{pmatrix}.$$

In 1.1.2 we proved that $\tanh'(x) \leq 1 \; \forall x$.

Since J is a diagonal matrix, its diagonal elements are its singular values. Thus, the largest singular value of J is no more than 1, no matter the input.

Denote this as $\sigma_{max}^{\tanh} \leq 1$. ∎

Given that $x_{t+1} = g(x_t) = \tanh(Wx_t)$, by chain rule we have $J_g(x_t) = J_{\tanh}(Wx_t) \times W$

where J is the Jacobian matrix.

Therefore, $\sigma_g \leq \sigma_{\tanh} \cdot \sigma_W \leq \frac{1}{2}$,

no matter the input

Notice that $x_n = g^{(n)}(x_1)$, using chain rule again there is $J_{g^{(n)}}$

$$= J_g(x_{n-1})\, J_g(x_{n-2}) \cdots J_g(x_1)$$

$$\Rightarrow \delta_{g^{(n)}} \leq (\delta_g)^n = \frac{1}{2}. \square$$

1.3.1

Consider the scenario of multiplying A with B where $A \in \mathbb{M}^{m \times n}$, $B \in \mathbb{M}^{n \times k}$.

There are in total $mk(2n-1)$ multiplication and additions. This is because there are mk elements in AB, and each element is the result of the dot product between two n-dimensional vectors. Each dot product consists of $2n-1$ multplications and additions. (1)

Softmax function over a k-dimensional vector costs k exponentiations, k-1 additions

and k divisions, in total $3k-1$ steps. (2)

Applying these lemma to the original question, multiplying Q by K^T costs $n^2(2d-1)$ steps, dividing by $\sqrt{d_k}$ takes n^2 steps, softmax along each column of a n-by-n matrix takes $n(3n-1)$ steps, multiplying a n-by-n matrix with V takes $nd(2n-1)$ steps.

All these operations are bounded by $O(n^2)$, thus attention is $O(n^2)$.

1.3.2

The rank of P being k means P has k non-zero singular values. $P = U\Sigma W^T$ where Σ has only k non-zero entries. We compute $PV = U\Sigma W^T V$ in the following order:

① ΣW^T: Considering the zero entries, it can be regarded as multiplying $k \times k$ by $k \times n$ matrices. Cost is $O(k^2 n)$. There are only k non-zero rows in ΣW^T.

② $(\Sigma W^T)V$: It can be regarded as multiplying $k \times n$ by $n \times d$ matrices. Cost is $O(knd)$. There are only k non-zero rows in the resulting matrix.

③ $U(\Sigma W^T V)$: This can be regarded as multiplying $n \times d$ by $d \times k$ matrices. Cost is $O(ndk)$.

Thus, the total cost is $O(k^2 n) + O(knd) + O(knd) = O(knd)$.

1.3.3

We use compact SVD to decompose P as $M\Sigma N$ where M is $n\times k$, Σ is $k\times k$, N is $k\times N$, and M, N are semi-unitary matrices i.e. $M^TM = I_k$ and $NN^T = I_k$.

We take $C = D = N$, then

$$Q(CK)^TDV = (QK^T)(C^TD)V = PN^TNV$$

$$= M\Sigma(NN^T)NV = M\Sigma NV = PV.$$

We compute CK $(k\times n)\times(n\times d)$ first,
then we compute DV $(k\times n)\times(n\times d)$,
then $Q(CK)^T$ $(n\times d)\times(d\times k)$,
finally $Q(CK)^T(DV)$ $(n\times k)\times(k\times d)$.

Each operation takes $O(nkd)$ time, so in total $O(nkd)$ time

2-1

$$E_{q(z|x)}[\log p(z)] + E_{q(z|x)}\left[\log \frac{p(x|z)}{q(z|x)}\right] + E_{q(z|x)}\left[\log \frac{q(z|x)}{p(z|x)}\right]$$

$$= \int \left[\log p(z) + \log \frac{p(x|z)}{q(z|x)} + \log \frac{q(z|x)}{p(z|x)}\right] q(z|x)\, dz$$

$$= \int \Big[\cancel{\log p(z)} + \log p(x,z) - \cancel{\log p(z)} - \cancel{\log q(z,x)} + \cancel{\log q(x)} + \cancel{\log q(z,x)} - \cancel{\log q(x)} - \cancel{\log p(z,x)} + \log p(x)\Big] q(z|x)\, dz$$

$$= \log p(x) \underbrace{\int q(z|x)\, dz}_{=1} = \log p(x) \quad \blacksquare$$

2.2

Given that $\delta(g(z)) = \left|\det \frac{\partial g(z)}{\partial z}\right|^{-1}_{z=z_0} \delta(z-z_0)$,

there is $\delta(x-f(z)) = \left|\det\left(-\frac{\partial f(z)}{\partial z}\right)\right|^{-1}_{z=z_0} \delta(z-z_0)$

where $z_0 = f^{-1}(x)$.

$$\Rightarrow \frac{p(x|z)}{p(z|x)} = \frac{\delta(x-f(z))}{\delta(z-f^{-1}(x))} = \left|\det \frac{\partial f(z)}{\partial z}\right|^{-1}_{z=f^{-1}(x)}$$

Thus,

$$\log p(x) = E_{q(z|x)}[\log p(z)] + E_{q(z|x)}\left[\log \frac{p(x|z)}{q(z|x)}\right] + E_{q(z|x)}\left[\log \frac{q(z|x)}{p(z|x)}\right]$$

$$= E_{p(z|x)}[\log p(z)] + E_{p(z|x)}\left[\log \frac{p(x|z)}{p(z|x)}\right] + \underbrace{E_{p(z|x)}\left[\log \frac{p(z|x)}{p(z|x)}\right]}_{=0}$$

$$= \log p(f^{-1}(x)) + E_{p(z|x)}\left[\log\left(\left|\det \frac{\partial f(z)}{\partial z}\right|^{-1}_{z=f^{-1}(x)}\right)\right]$$

constant w.r.t z

$$= \log p(z) + \log\left|\det \frac{\partial f(z)}{\partial z}\right|^{-1}$$

∎

2.3

$\log p(f(z)) \geq \log p(z) + V$ ← lower bound gap

For bijective and smooth functions f, lower bound gap

$V(z) = \log \left| \det \frac{\partial f(z)}{\partial z} \right|^{-1}$.

For VAE transformation f (or $q(z|x)$) with decoder $p(x|z)$, lower bound gap

$$V(z) = E_{q(z|x)}\left[\log \frac{p(x|z)}{q(z|x)}\right] + E_{q(z|x)}\left[\log \frac{q(z|x)}{p(z|x)}\right]$$

$$= E_{q(z|x)}\left[\log \frac{p(x|z)}{p(z|x)}\right]$$ which can be estimated through Monte-Carlo.

Note that although $p(z|x)$ is not directly given, it can be computed as $\frac{p(x|z)\,p(z)}{p(x)}$.

Thus, given $x_t = f_t \circ (x_{t-1})$, we can compute or estimate value of lower bound gap $V_t(x_{t-1})$ depending on the type of transformation

3.1

a) $T^{\pi}V_1(s) - T^{\pi}V_2(s) = r^{\pi}(s) + \gamma\int P(s'|s,a)\pi(a|s)V_1(s') - r^{\pi}(s) - \gamma\int P(s'|s,a)\pi(a|s)V_2(s')$

$= \gamma\int P(s'|s,a)\pi(a|s)(V_1 - V_2)(s') \leq 0$ ∎

$>0 \quad \geq 0 \quad \geq 0 \quad < 0$

b) $\| T^{\pi}(Q_1)(s,a) - T^{\pi}(Q_2)(s,a) \|_{\infty}$

$= \| \gamma\int\int P(s'|s,a)\pi(a'|s')[Q_1(s',a') - Q_2(s',a')]ds'da' \|_{\infty}$

$= \gamma \sup_{s,a} \left| \int\int P(s'|s,a)\pi(a'|s')[Q_1(s',a') - Q_2(s',a')]ds'da' \right|$

$P(x)$ $f(x)$

$\leq \gamma \sup_{s,a} \sup_{s',a'} [Q_1(s',a') - Q_2(s',a')]$ by given lemma

$= \gamma \sup_{s',a'} [Q_1(s',a') - Q_2(s',a')]$

$= \gamma \| Q_1(s,a) - Q_2(s,a) \|_{\infty}$ ∎

where $\int\int p(x)\,dx$

$$= \int\int \underbrace{\pi(a'|s')\,da'}_{=1}\, p(s'|s,a)\,ds'$$

$$= \int p(s'|s,a)\,ds' = 1$$

c) do we have access to this term?

1. $V_*(s) = \int \pi(a|s)\, q_*(s,a)\,da$

2. $q_*(s,a) = r(s,a) + \gamma \int p(s'|s,a)\, V_*(s')\,ds'$

3. $a_* = \underset{a}{\text{argmax}}\, q_*(s,a)$

4. $a_* = \text{argmax}\left[r(s,a) + \gamma \int p(s'|s,a)\, V_*(s')\,ds' \right]$

- From Equilibrium Q function to Equilibrium V function:

$$Q_*(s,a) = r(s,a) + \gamma \iint p(s'|s,a)\,\pi(a'|s')\,Q_*(s',a')\,ds'da'$$

$$\Rightarrow \int Q_*(s,a)\,\pi(a|s)\,da = \int r(s,a)\,\pi(a|s)\,da$$

$$V_*(s)$$

$$+ \gamma \iint p(s'|s,a)\,\pi(a|s)\left[\int Q_*(s',a')\,\pi(a'|s')\,da'\right]ds'da$$

$$V_*(s')$$

$$\Rightarrow V_*(s) = \int r(s,a)\,\pi(a|s)\,da + \gamma \iint p(s'|s,a)\,\pi(a|s)\,V_*(s')\,ds'da$$

3.2.1

In this question we're only concerned with b_k, so denote $\nabla_{\theta_k} \log \pi_\theta(a_t^{(i)} | s_t^{(i)})$ as m_t, $R(s_{t'}^{(i)}, a_{t'}^{(i)})$ as $n_{t'}$, then m_t and $n_{t'}$ are random variables.

Then $\sum_{t=1}^{T} \nabla_{\theta_k} \log \pi_\theta(a_t^{(i)} | s_t^{(i)}) \left[\sum_{t'=t}^{T} R(s_{t'}^{(i)}, a_{t'}^{(i)}) - b_k \right]$

Can be written as

$$\sum_{t=1}^{T} m_t \left(\sum_{t'=t}^{T} n_{t'} - b_k \right)$$

$$= \sum_{t=1}^{T} \sum_{t'=t}^{T} m_t n_{t'} - \sum_{t=1}^{T} m_t b_k .$$

$f(b_k)$

Then,

$$E\left(f^2(b_k)\right) = E\left(\left(\sum_{t=1}^{T} \sum_{t'=t}^{T} m_t n_{t'} - \sum_{t=1}^{T} m_t b_k \right)^2\right)$$

$$= E\left(\left(\sum_{t=1}^{T} m_t b_k\right)^2 - 2\left(\sum_{t=1}^{T} \sum_{t'=t}^{T} m_t n_{t'}\right)\left(\sum_{t=1}^{T} m_t b_k\right)\right.$$

$+$ constant w.r.t to b_k)

$$= b_k^2 E\left(\left(\sum_{t=1}^{T} m_t\right)^2\right) - 2b_k E\left(\left[\sum_{t=1}^{T}\sum_{t'=t}^{T} m_t n_{t'}\right]\sum_{t=1}^{T} m_t\right)$$

$+$ constant

$$E^2(f(b_k)) = \left[\sum_{t=1}^{T}\sum_{t'=t}^{T} E(m_t n_{t'}) - \sum_{t=1}^{T} E(m_t) b_k\right]^2$$

$$= b_k^2\left[\sum_{t=1}^{T} E(m_t)\right]^2 - 2b_k\left(\sum_{t=1}^{T}\sum_{t'=t}^{T} E(m_t n_{t'})\right)\left(\sum_{t=1}^{T} E(m_t)\right)$$

$+$ constant

Thus, $\mathrm{Var}(f(b_k)) = E(f^2(b_k)) - E^2(f(b_k))$

$$= b_k^2\underbrace{\left[E\left(\left(\sum_{t=1}^{T} m_t\right)^2\right) + \left(\sum_{t=1}^{T} E(m_t)\right)^2\right]}_{= C_1}$$

$$-2b_k\underbrace{\left(E\left(\left[\sum_{t=1}^{T}\sum_{t'=t}^{T} m_t n_{t'}\right]\sum_{t=1}^{T} m_t\right) - \left(\sum_{t=1}^{T}\sum_{t'=t}^{T} E(m_t n_{t'})\right)\left(\sum_{t=1}^{T} E(m_t)\right)\right)}_{= C_2}$$

$+$ constant

$= b_k^2 \cdot c_1 - 2 b_k \cdot c_2 + \text{constant}$

Notice that c_1 must be positive.

$$\frac{d}{db_k} \text{var}(f(b_k)) = 2 b_k \cdot c_1 - 2 c_2 = 0$$

$$\Rightarrow b_k^* = \frac{c_2}{c_1}$$

and $\frac{d^2}{db_k^2} \text{var}(f(b_k)) = 2 c_1 > 0$ so b_k^* is a global min.

Therefore, $b_k^* = \frac{c_1}{c_2}$, where

$$c_1 = E\left(\left(\sum_{t=1}^{T} m_t\right)^2\right) + \left(\sum_{t=1}^{T} E(m_t)\right)^2 \text{ and}$$

$$c_2 = E\left(\left[\sum_{t=1}^{T}\sum_{t'=t}^{T} m_t n_{t'}\right]\sum_{t=1}^{T} m_t\right) - \left(\sum_{t=1}^{T}\sum_{t'=t}^{T} E(m_t n_{t'})\right)\left(\sum_{t=1}^{T} E(m_t)\right),$$

where $m_t = \nabla_{\theta_k} \log \pi_\theta(a_t^{(i)} | s_t^{(i)})$ and

$$n_{t'} = R(s_{t'}^{(i)}, a_{t'}^{(i)}),$$

where $s_1 \sim \rho_0(s)$, $a_t \sim \pi_\theta(a_t | s_t)$ and $s_{t+1} \sim p(s_{t+1} | s_t, a_t)$

3.2.2

Starting with Equation 3.3,

$$\nabla_\theta J(\theta) = \nabla_\theta \underset{b_t \sim \pi_\theta}{E} \left(\sum_{t'=1}^{T} R(s_{t'}, b_{t'}) \right)$$

$$= \underset{b_t \sim \pi_\theta}{E} \left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(b_t | s_t) \left[\sum_{t'=1}^{T} R(s_{t'}, b_{t'}) \right] \right]$$

$$= \underset{a_t \sim \pi'_\theta}{E} \left(\sum_{t=1}^{T} \frac{\pi_\theta(a_t | s_t)}{\pi'_\theta(a_t | s_t)} \nabla_\theta \log \pi_\theta(b_t | s_t) \left[\sum_{t'=1}^{T} R(s_{t'}, b_{t'}) \right] \right)$$ by Importance Sampling