1.

Fisrt change $F(A,B)$ to the function of y_n and z_n.

Then we can represent $F(A,B)$ by p.

$$F(A,B) = \frac{1}{N}\sum_{n=1}^{N} \ln\big(1+\exp\big(-y_n(A(w_{SVM}^T\phi(x_n)+b_{SVM})+B)\big)\big)$$

$$= \frac{1}{N}\sum_{n=1}^{N} \ln\big(1+\exp\big(-y_n(Az_n+B)\big)\big) = -\frac{1}{N}\sum_{n=1}^{N} \ln\left(\frac{1}{1+\exp\big(-y_n(Az_n+B)\big)}\right)$$

$$= -\frac{1}{N}\sum_{n=1}^{N} \ln\left(1-\frac{\exp\big(-y_n(Az_n+B)\big)}{1+\exp\big(-y_n(Az_n+B)\big)}\right) = -\frac{1}{N}\sum_{n=1}^{N} \ln(1-p_n)$$

We first have the gradient if the logistuc function p.

$$\begin{bmatrix} \frac{\partial p}{\partial A} \\ \frac{\partial p}{\partial B} \end{bmatrix} = \nabla\theta\big(-y_n(Az_n+B)\big) = p(1-p)\begin{bmatrix} -y_n z_n \\ -y_n \end{bmatrix} = \begin{bmatrix} -p(1-p)y_n z_n \\ -p(1-p)y_n \end{bmatrix}$$

Then we calculate the gradient of F by chain rule.

$$\nabla F = \frac{\partial F}{\partial P}\begin{bmatrix} \frac{\partial p}{\partial A} \\ \frac{\partial p}{\partial B} \end{bmatrix} = -\frac{1}{N}\sum_{n=1}^{N}(1-p_n)^{-1}\,(-1)(-p)(1-p)\begin{bmatrix} -y_n z_n \\ -y_n \end{bmatrix} = -\frac{1}{N}\sum_{n=1}^{N}\begin{bmatrix} p_n y_n z_n \\ p_n y_n \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{1}{N}\sum_{n=1}^{N} p_n y_n z_n \\ -\frac{1}{N}\sum_{n=1}^{N} p_n y_n \end{bmatrix}$$

2.

To have the Hessian matrix we further calculate the derivatives below.

$$H = \begin{bmatrix} \frac{\partial F}{\partial A \partial A} & \frac{\partial F}{\partial A \partial B} \\ \frac{\partial F}{\partial B \partial A} & \frac{\partial F}{\partial B \partial B} \end{bmatrix}$$

$$\frac{\partial F}{\partial A \partial A} = \frac{\partial}{\partial A}\left(\frac{\partial F}{\partial A}\right) = \frac{\partial}{\partial A}\left(-\frac{1}{N}\sum_{n=1}^{N} p_n y_n z_n\right) = -\frac{1}{N}\sum_{n=1}^{N} y_n z_n \frac{\partial p_n}{\partial A} = \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n^2 p_n(1-p_n)$$

$$= \frac{1}{N}\sum_{n=1}^{N} z_n^2 p_n(1-p_n)$$

$$\frac{\partial F}{\partial A \partial B} = \frac{\partial}{\partial A}\left(\frac{\partial F}{\partial B}\right) = \frac{\partial}{\partial A}\left(-\frac{1}{N}\sum_{n=1}^{N} p_n y_n\right) = -\frac{1}{N}\sum_{n=1}^{N} z_n \frac{\partial p_n}{\partial A} = \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n p_n(1-p_n)$$

$$= \frac{1}{N}\sum_{n=1}^{N} z_n p_n(1-p_n)$$

$$\frac{\partial F}{\partial B \partial B} = \frac{\partial}{\partial B}\left(\frac{\partial F}{\partial B}\right) = \frac{\partial}{\partial B}\left(-\frac{1}{N}\sum_{n=1}^{N} p_n y_n\right) = -\frac{1}{N}\sum_{n=1}^{N} z_n \frac{\partial p_n}{\partial B} = \frac{1}{N}\sum_{n=1}^{N} y_n^2 p_n(1-p_n)$$

$$= \frac{1}{N}\sum_{n=1}^{N} p_n(1-p_n)$$

3.

We have the Hessian matrix form the result above.

$$H = \begin{bmatrix} \sum_{n=1}^{N} z_n^2 p_n(1-p_n) & \sum_{n=1}^{N} z_n p_n(1-p_n) \\ \sum_{n=1}^{N} z_n p_n(1-p_n) & \sum_{n=1}^{N} p_n(1-p_n) \end{bmatrix}$$

Since $0 < p < 1$, we can know that the diagonal elements in the H are greater than 0

$$\sum_{n=1}^{N} z_n^2 p_n(1-p_n) > 0 \ and \ \sum_{n=1}^{N} p_n(1-p_n) > 0$$

and based on the eigenvalue formula,

$$\left(\sum_{n=1}^{N} z_n^2 p_n(1-p_n) - \lambda\right)\left(\sum_{n=1}^{N} p_n(1-p_n) - \lambda\right) = \left(\sum_{n=1}^{N} z_n p_n(1-p_n)\right)^2$$

if the below inequality holds, we can know that $\forall \lambda_i \ in \ \lambda, \ \lambda_i \geq 0$

$$\left(\sum_{n=1}^{N} z_n^2 p_n(1-p_n)\right)\left(\sum_{n=1}^{N} p_n(1-p_n)\right) \geq \left(\sum_{n=1}^{N} z_n p_n(1-p_n)\right)^2$$

By substracting the left side by the right side of the inequality, we prove that the inequality holds.

$$\sum_{i=1}^{N}\sum_{j=1}^{N} z_i^2 p_i(1-p_i)\, p_j(1-p_j) - \sum_{i=1}^{N}\sum_{j=1}^{N} z_i z_j p_i(1-p_i)\, p_j(1-p_j)$$

$$= \sum_{i=1}^{N}\sum_{j!=i}^{N} z_i^2 p_i(1-p_i)\, p_j(1-p_j) - \sum_{i=1}^{N}\sum_{j!=i}^{N} z_i z_j p_i(1-p_i)\, p_j(1-p_j)$$

$$= \sum_{i=1}^{N}\sum_{j>i}^{N} (z_i^2 + z_j^2) p_i(1-p_i)\, p_j(1-p_j) - 2\sum_{i=1}^{N}\sum_{j>i}^{N} z_i z_j p_i(1-p_i)\, p_j(1-p_j)$$

$$= \sum_{i=1}^{N}\sum_{j>i}^{N} \left(z_i + z_j\right)^2 p_i(1-p_i)\, p_j(1-p_j) \geq 0$$

Since we prove that $\forall \lambda_i\ in\ \lambda,\ \lambda_i \geq 0$, therefore H is a $p.s.d.$ matrix.

4.

$w_0 = d - 0.5,\ w_i = -1\ for\ i\ in\ \{2, \dots, d\}$

By setting wieghts like this, $\sum_{i=0}^{d} w_i x_i$ can remain positive except that all the input is false (in this case $\sum_{i=0}^{d} w_i x_i = -0.5$).

5.

We define the error functionas below.

$$e_n = (y - \tanh\,(s_1^L))^2 = \left(y - \tanh\left(\sum_{i=0}^{d^{(L-1)}} w_{i1}^{(L)} x_i^{(L-1)} \right) \right)^2$$

Start from the output layer:

$$\frac{\partial e_n}{\partial s_1^L} = 2(y - \tanh(s_1^L))\big(-tanh'(s_1^L)\big),$$

Since all $w_i = 0,\ s_1^L = 0$, and thus $-tanh'(s_1^L) \neq 0$.

Hence $\frac{\partial e_n}{\partial s_1^L} \neq 0$.

Then for the hidden layer:

$$\delta_j^{(l)} = \frac{\partial e_n}{\partial s_j^{(l)}} = \sum_{k=1}^{d^{(l+1)}} \frac{\partial e_n}{\partial s_k^{(l+1)}} \frac{\partial s_k^{(l+1)}}{\partial x_j^{(l+1)}} \frac{\partial x_j^{(l+1)}}{\partial s_j^{(l)}} = \sum_k \left(\delta_k^{(l+1)}\right)\left(w_{jk}^{(l+1)}\right)\left(tanh'\left(s_j^{(l)}\right)\right)$$

Since all $w_i = 0$, all $\delta_j^{(l)} = 0$.

6.

Our goal is to maximize the objective function

$$f(d^1, d^2, \dots, d^{L-1}) = 12 \times d^1 + (d^1+1)d^2 + \cdots + (d^{L-2}+1)d^{L-1} + (d^{L-1}+1)$$
$$= 12 \times d^1 + (d^1 d^2 + \cdots + d^{L-2}d^{L-1}) + (d^2 + \cdots + d^{L-1}) + (d^{L-1}+1)$$

with the constraint

$$\sum_{l=1}^{L-1}(d^l+1) = \sum_{l=1}^{L-1} d^l + (L-1) = 48$$

We can see that in $f(d^1, d^2, \dots, d^{L-1})$, the sum of quadratic terms $(d^1 d^2 + \cdots + d^{L-2}d^{L-1})$ will shrink drastically as $L-1$ increase. Also, the linear terms $12d^1 + (d^2 + \cdots + d^{L-2}) + 2d^{L-1}$ shows the importance of $d^1$and d^{L-1}. Hence, it's reasonable that we solve the objective function with $L-1=2$, which minimize $L-1$ but retain a quadratic term and important linear terms at the same time.

Hence our objective function can be simplified as

$$f(d^1, d^2) = 12 \times d^1 + (d^1+1)d^2 + (d^2+1) = d^1 d^2 + 12d^1 + 2d^2 + 1$$

with the constraint

$$\sum_{l=1}^{2}(d^l+1) = d^1 + d^2 + 2 = 48$$

Let $d^2 = 46 - d^1$, we can modify our objective function as

$$f(d^1) = d^1(46-d^1) + 12d^1 + 2(46-d^1) + 1 = -(d^1)^2 + 56d^1 + 93$$

and maximize $f(d^1)$ with $d^1_{max} = argmax\big(f(d^1)\big) = \frac{56}{2} = 28$

$$\max\big(f(d^1)\big) = -28^2 + 56 \times 28 + 93 = 877$$

7.

We first expand the error function.

$$err_n(w) = \left||x_n - ww^T x_n\right||^2 = (x_n - ww^T x_n)^T (x_n - ww^T x_n) = (x_n^T - x_n^T ww^T)(x_n - ww^T x_n)$$

$$= x^T x - x_n^T ww^T x_n - x_n^T ww^T x_n + x_n^T ww^T ww^T x_n$$

$$= x^T x - 2(w^T x_n)^2 + (w^T x_n)^2 (w^T w)$$

Then have the derivative for w_i.

$$\frac{\partial err_n(w)}{\partial w_i} = 0 - 2tr[2(w^T x_n)(e_i^T x_n)] + tr[2(w^T x_n)(e_i^T x_n)](w^T w) + (w^T x_n)^2 (e_i^T w + w^T e_i)$$

$$= -4w^T x_n x_i + 2(w^T w) w^T x_n x_i + 2(w^T x_n)^2 w_i$$

Last, combine the result of all w_i together.

$$\frac{\partial err_n(w)}{\partial w} = 4w^T x_n x_n + 2(w^T w) w^T x_n x_n + 2(w^T x_n)^2 w_n$$

8.

$$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}\left|\left|x_n - ww^T(x_n + \epsilon_n)\right|\right|^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}(x_n - ww^T x_n - ww^T \epsilon_n)^T (x_n - ww^T x_n - ww^T \epsilon_n)$$

$$= \frac{1}{N}\sum_{n=1}^{N} x_n^T x_n - x_n^T ww^T x_n - x_n^T ww^T \epsilon_n - x_n^T ww^T x_n + x_n^T ww^T ww^T x_n$$

$$+ x_n^T ww^T ww^T \epsilon_n - \epsilon_n^T ww^T x_n + \epsilon_n^T ww^T ww^T x_n + \epsilon_n^T ww^T ww^T \epsilon_n$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left|\left|x_n - ww^T x_n\right|\right|^2 + \frac{1}{N}\sum_{n=1}^{N} -x_n^T ww^T \epsilon_n + x_n^T ww^T ww^T \epsilon_n - \epsilon_n^T ww^T x_n$$

$$+ \epsilon_n^T ww^T ww^T x_n + \epsilon_n^T ww^T ww^T \epsilon_n$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left|\left|x_n - ww^T x_n\right|\right|^2 + \frac{1}{N}\sum_{n=1}^{N} \epsilon_n^T ww^T ww^T \epsilon_n$$

$$\phi(w) = E\left(\frac{1}{N}\sum_{n=1}^{N} \epsilon_n^T ww^T ww^T \epsilon_n\right) = \frac{1}{N}\sum_{n=1}^{N} E(\epsilon_n^T ww^T ww^T \epsilon_n) = \frac{1}{N}\sum_{n=1}^{N} tr[E(\epsilon_n^T ww^T ww^T \epsilon_n)]$$

$$= \frac{1}{N}\sum_{n=1}^{N} E[tr(\epsilon_n^T ww^T ww^T \epsilon_n)] = \frac{1}{N}\sum_{n=1}^{N} E[tr(w^T w \epsilon_n^T \epsilon_n w^T w)]$$

$$= \frac{1}{N}\sum_{n=1}^{N} tr(w^T w E(\epsilon_n^T \epsilon_n) w^T w) = \frac{1}{N}\sum_{n=1}^{N} tr(w^T w w^T w) = (w^T w)^2$$

9.

Let tanh() be a element wise operator

$$E_9(u) = \sum_{i=1}^{d} (g_i(x) - x_i)^2 = \sum_{i=1}^{d} \left(u_i\left(tanh(u^T x)\right) - x_i\right)^2 = \sum_{i=1}^{d} \left(u_i \left(\begin{bmatrix} tanh(u_1^T x) \\ \vdots \\ tanh(u_j^T x) \\ \vdots \\ tanh(u_{\tilde{d}}^T x) \end{bmatrix} \right) - x_i \right)^2$$

$$= \sum_{i=1}^{d} \left(\sum_{j=1}^{\tilde{d}} u_{ij} \tanh\left(u_j^T x\right) - x_i \right)^2 = \sum_{i_1=1}^{d} \left(\sum_{j=1}^{\tilde{d}} u_{i_1 j} \tanh\left(\sum_{i_2=1}^{d} u_{i_2 j} x_{i_2} \right) - x_{i_1} \right)^2$$

10.

$$E_{10}(w) = \sum_{i=1}^{d} (g_i(x) - x_i)^2 = \sum_{i=1}^{d} \left(w_i^{(2)^T} \left(tanh\left(w^{(1)^T} x \right) \right) - x_i \right)^2$$

$$= \sum_{i=1}^{d} \left(w_i^{(2)^T} \left(\begin{bmatrix} tanh(w^1\_1 \;\; x) \\ \vdots \\ tanh\left(w^{(1)}{}_j^T x \right) \\ \vdots \\ tanh\left(w^{(1)}{}_{\tilde{d}}^T x \right) \end{bmatrix} \right) - x_i \right)^2$$

$$= \sum_{i=1}^{d} \left(\sum_{j=1}^{\tilde{d}} w_{ji}^{(2)} \tanh\left(w^{(1)}{}_j^T x \right) - x_i \right)^2$$

$$= \sum_{i_1=1}^{d} \left(\sum_{j=1}^{\tilde{d}} w_{ji_1}^{(2)} \tanh\left(\sum_{i_1=1}^{d} w_{i_1 j}^{(1)} x_{i_1} \right) - x_{i_1} \right)^2$$

$$\frac{\partial E_{10}(w)}{\partial w_{i'j'}^{(1)}} = \sum_{i_1=1}^{d} 2 \left(\sum_{j=1}^{\tilde{d}} w_{ji_1}^{(2)} \tanh\left(\sum_{i_2=1}^{d} w_{i_2 j}^{(1)} x_i \right) - x_{i_1} \right) \left(w_{j'i}^{(2)} \tanh'\left(\sum_{i_2=1}^{d} w_{i_2 j'}^{(1)} x_{i_2} \right) \right) w_{i'j'}^{(1)}$$

$$\frac{\partial E_{10}(w)}{\partial w_{j'i'}^{(2)}} = 2 \left(\sum_{j=1}^{\tilde{d}} w_{ji'}^{(2)} \tanh\left(\sum_{i_2=1}^{d} w_{i_2 j}^{(1)} x_i \right) - x_{i_1} \right) \left(\tanh\left(\sum_{i_2=1}^{d} w_{i_2 j}^{(1)} x_i \right) \right)$$

$$\frac{\partial E_9(u)}{\partial u_{i'j'}} = \sum_{i_1=1, i\neq i'}^{d} 2\left(\sum_{j=1}^{\tilde{d}} u_{i_1 j} \tanh\left(\sum_{i_2=1}^{d} u_{i_2 j} x_{i_2}\right) - x_{i_1}\right)\left(u_{i_1 j'} \tanh'\left(\sum_{i_2=1}^{d} u_{i_2 j} x_{i_2}\right)\right) x_{i'}$$

$$+ 2\left(\sum_{j=1}^{\tilde{d}} u_{i'j} \tanh\left(\sum_{i_2=1}^{d} u_{i_2 j} x_{i_2}\right) - x_{i'}\right)\left(\tanh\left(\sum_{i_2=1}^{d} u_{i_2 j'} x_{i_2}\right)\right.$$

$$\left. + u_{i'j'} \tanh\left(\sum_{i_2=1}^{d} u_{i_2 j'} x_{i_2}\right) x_{i'}\right)$$

$$= \sum_{i_1=1}^{d} 2\left(\sum_{j=1}^{\tilde{d}} u_{i_1 j} \tanh\left(\sum_{i_2=1}^{d} u_{i_2 j} x_{i_2}\right) - x_{i_1}\right)\left(u_{i_1 j'} \tanh'\left(\sum_{i_2=1}^{d} u_{i_2 j} x_{i_2}\right)\right) x_{i'}$$

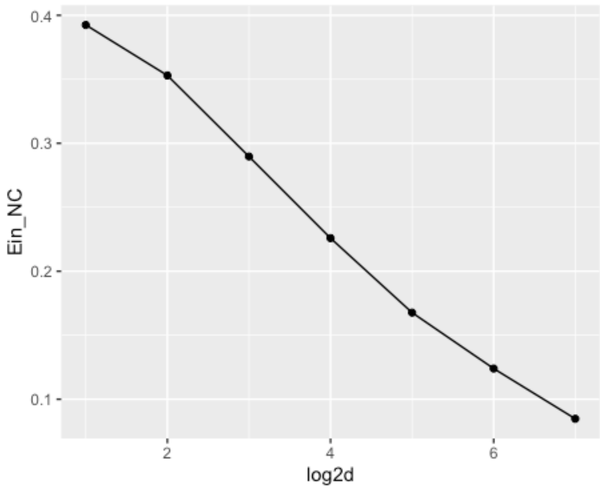
$$+ 2\left(\sum_{j=1}^{\tilde{d}} u_{i'j} \tanh\left(\sum_{i_2=1}^{d} u_{i_2 j} x_{i_2}\right) - x_{i'}\right)\left(\tanh\left(\sum_{i_2=1}^{d} u_{i_2 j'} x_{i_2}\right)\right)$$

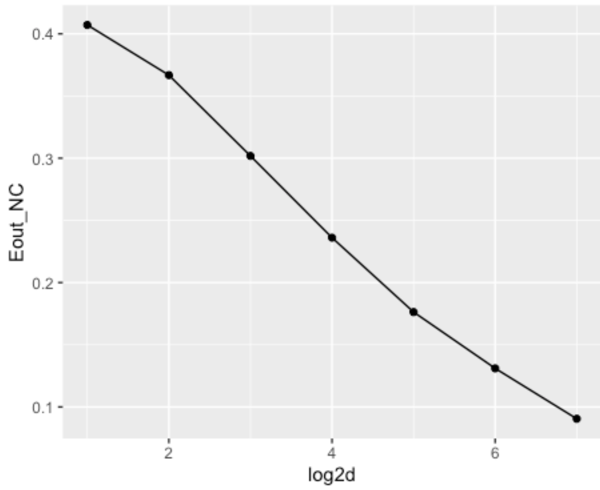
$$\frac{\partial E_9(u)}{\partial u_{i'j'}} = \sum_{i_1=1}^{d} 2\left(\sum_{j=1}^{\tilde{d}} w_{j i_1}^{(2)} \tanh\left(\sum_{i_2=1}^{d} w_{i_2 j}^{(1)} x_i\right) - x_{i_1}\right)\left(w_{j'i}^{(2)} \tanh'\left(\sum_{i_2=1}^{d} w_{i_2 j'}^{(1)} x_{i_2}\right)\right) w_{i'j'}^{(1)}$$

$$+ 2\left(\sum_{j=1}^{\tilde{d}} w_{j i'}^{(2)} \tanh\left(\sum_{i_2=1}^{d} w_{i_2 j}^{(1)} x_i\right) - x_{i_1}\right)\left(\tanh\left(\sum_{i_2=1}^{d} w_{i_2 j}^{(1)} x_i\right)\right)$$

$$= \frac{\partial E_{10}(w)}{\partial w_{j'i'}^{(2)}} + \frac{\partial E_{10}(w)}{\partial w_{i'j'}^{(1)}}$$

11.





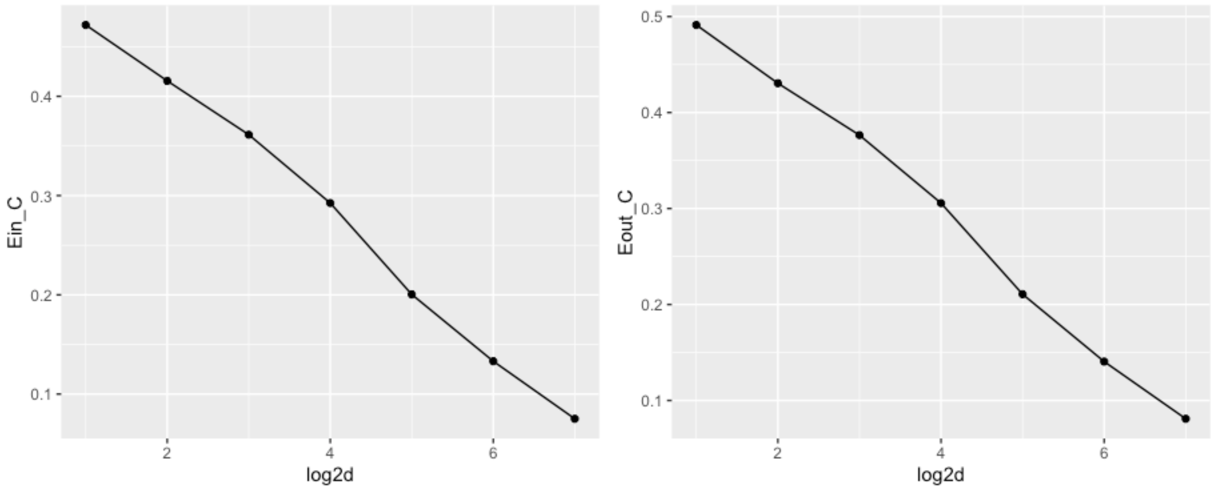
Ein decrease from around 0.39 to around 0.08 as the encoder dimension $\tilde{d}$ increase from 2^1 to 2^7.

Eout decrease from around 0.40 to around 0.09 as the encoder dimension $\tilde{d}$ increase from 2^1 to 2^7. In average, Eout is slightly greater than Ein with small gap.

12.

(with 11.)

13.



Ein decrease from around 0.47 to around 0.07 as the encoder dimension $\tilde{d}$ increase from 2^1 to 2^7.

Eout decrease from around 0.49 to around 0.08 as the encoder dimension $\tilde{d}$ increase from 2^1 to 2^7. In average, Eout is slightly greater than Ein with small gap.

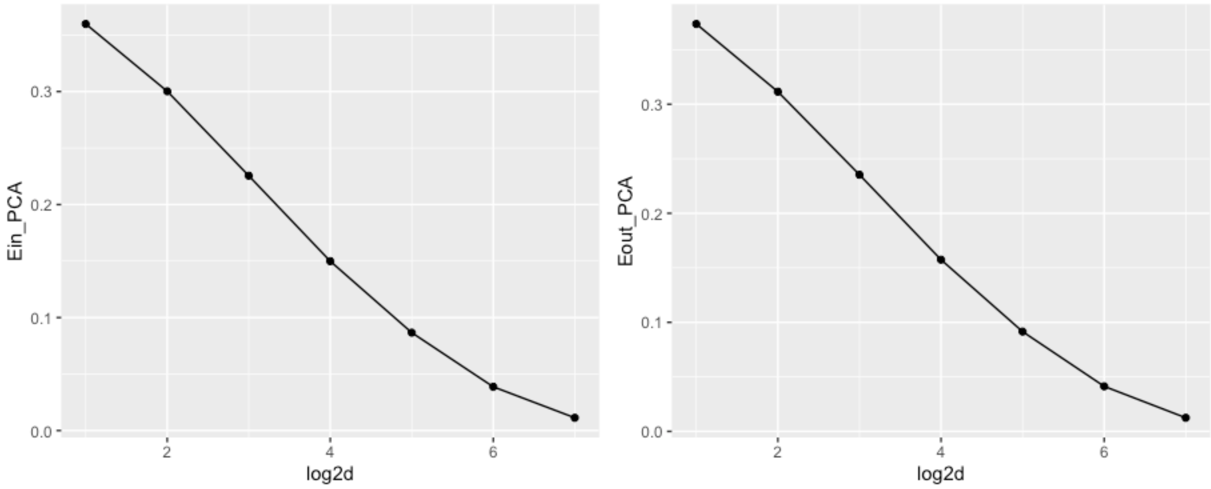
The auto encoder with constraints is less complex than that without constraint, thus it is reasonable that its Ein and Eout are higher in this case.

However, the difference vanishes as $\tilde{d}$ increase, it might imply that the symmetric weights is a good property for a large $\tilde{d}$ auto encoder.

14.

(with 13)

15.



Ein decrease from around 0.36 to around 0.01 as the PC number $\tilde{d}$ increase from 2^1 to 2^7.

Eout decrease from around 0.47 to around 0.01 as the PC number $\tilde{d}$ increase from 2^1 to 2^7. In average, Eout is slightly greater than Ein with small gap.

The linear auto encoder (PCA) is much faster than both non-linear auto encoders.
Also, Ein and Eout of the linear auto encoder is significantly lower than those of non-linear auto encoders.

Considering that our non-linear auto encoders has only one hidden layer, where restrict the non-linear property to fit a complex model, the result is reasonable.

16.

(with 15)

17.

Given $N \geq 3\Delta \log_2 \Delta$, we can have $2^N \geq \Delta^{3\Delta}$.

If we can proove $\Delta^{3\Delta} > N^\Delta + 1$, then we can have the result $2^N > \Delta^{3\Delta} > N^\Delta + 1$,

which is the inequality we want to prove.

For $N^\Delta + 1$, the max value happens when Δ is the biggest one given a fixed N, that is, when the $N = 3\Delta \log_2 \Delta$. Hence, we can let $N^\Delta + 1$ be $(3\Delta \log_2 \Delta)^\Delta + 1$ as an extreme case for the comparison between $\Delta^{3\Delta}$ and $N^\Delta + 1$.

Let $\Delta = 2$, $\Delta^{3\Delta} = 64$ and $(3\Delta \log_2 \Delta)^\Delta + 1 = 37$, the inequality $(\Delta^{3\Delta} > N^\Delta + 1)$ holds.
Then we can check whether $\Delta^{3\Delta}$ increase faster than $N^\Delta + 1$ to make sure
the inequality $(\Delta^{3\Delta} > N^\Delta + 1)$ still holds when $\Delta > 2$.

Since $\Delta > 2$, we can simplify the comparison from $(\Delta^3)^\Delta \; and \; (3\Delta \log_2 \Delta)^\Delta \rightarrow \Delta^2 \; and \; 3 \log_2 \Delta$

Since Δ^2 is a convex and $3 \log_2 \Delta$ is a concave, we can check whether the point they have the same derivatives happens before 2 (after that point a convex must increase faster than a concave).

By solving $2\Delta = \frac{3}{\Delta ln2}$, we have $\Delta = \sqrt{\frac{3}{2ln2}} \sim 1.04 < 2$.

Hence, we can conclude that the inequality $(\Delta^{3\Delta} > N^\Delta + 1)$ still holds when $\Delta > 2$.

Thus, we prove that $2^N > \Delta^{3\Delta} > N^\Delta + 1$ when $\Delta \geq 2$.