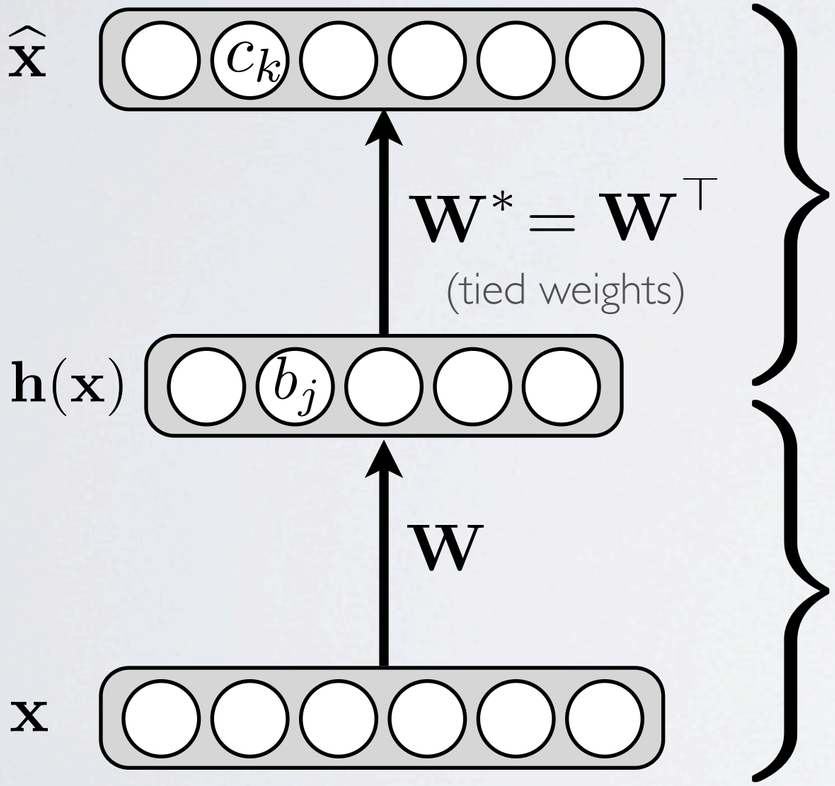
Neural networks

Autoencoder - linear autoencoder

AUTOENCODER

Topics: autoencoder, encoder, decoder, tied weights

- Feed-forward neural network trained to reproduce its input at the output layer



Decoder

$$\begin{aligned}\hat{\mathbf{x}} &= o(\hat{\mathbf{a}}(\mathbf{x})) \\ &= \underbrace{\mathrm{sigm}}_{\text{for binary inputs}}(\mathbf{c} + \mathbf{W}^*\mathbf{h}(\mathbf{x}))\end{aligned}$$

Encoder

$$\begin{aligned}\mathbf{h}(\mathbf{x}) &= g(\mathbf{a}(\mathbf{x})) \\ &= \mathrm{sigm}(\mathbf{b} + \mathbf{W}\mathbf{x})\end{aligned}$$

AUTOENCODER

Topics: optimality of a linear autoencoder

- To do the proof, we need the following theorem:
  - let $\mathbf{A}$ be any matrix, with singular value decomposition $\mathbf{A} = \mathbf{U}\,\Sigma\,\mathbf{V}^\top$
    - Σ is a diagonal matrix
    - $\mathbf{U}, \mathbf{V}$ are orthonormal matrices (columns/rows are orthonormal vectors)
  - let $\mathbf{U}_{\cdot,\leq k}\,\Sigma_{\leq k,\leq k}\,\mathbf{V}^\top_{\cdot,\leq k}$ be the decomposition where we keep only the k largest singular values
  - then, the matrix $\mathbf{B}$ of rank k that is closest to $\mathbf{A}$:

$$\mathbf{B}^* = \underset{\mathbf{B}\text{ s.t. }\mathrm{rank}(\mathbf{B})=k}{\arg\min} ||\mathbf{A}-\mathbf{B}||_F$$

  is $\mathbf{B}^* = \mathbf{U}_{\cdot,\leq k}\,\Sigma_{\leq k,\leq k}\,\mathbf{V}^\top_{\cdot,\leq k}$

$$\min_{\theta} \sum_{t} \frac{1}{2} \sum_{i} (x_i^{(t)} - \widehat{x}_i^{(t)})^2$$

$$\min_{\theta} \sum_{t} \frac{1}{2} \sum_{i} (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2$$

$$\min_{\theta} \sum_{t} \frac{1}{2} \sum_{i} (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2$$

Sketch of proof

matrix where columns are $\mathbf{x}^{(t)}$

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2$$

Sketch of proof

matrix where columns are $\mathbf{x}^{(t)}$

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

Sketch of proof

matrix where columns are $\mathbf{x}^{(t)}$

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2$$

Sketch of proof

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix where columns are } \mathbf{x}^{(t)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k} \, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U} \, \Sigma \, \mathbf{V}^\top$ and k is the hidden layer size

Sketch of proof

matrix where columns are $\mathbf{x}^{(t)}$

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\hat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix of all hidden layers (could be any encoder)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k}\, \Sigma_{\leq k, \leq k},\ \ \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U}\, \Sigma\, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

Sketch of proof

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix where columns are } \mathbf{x}^{(t)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k}\, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U}\, \Sigma\, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

$$\mathbf{h}(\mathbf{X}) \quad = \quad \mathbf{V}_{\cdot, \leq k}^\top$$

Sketch of proof

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix where columns are } \mathbf{x}^{(t)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k} \, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U} \, \Sigma \, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

$$\begin{aligned} \mathbf{h}(\mathbf{X}) &= \mathbf{V}_{\cdot, \leq k}^\top \\ &= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{X}^\top \, \mathbf{X})^{-1} \, (\mathbf{X}^\top \, \mathbf{X}) \end{aligned}$$

← multiplying by identity

Sketch of proof

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix where columns are } \mathbf{x}^{(t)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k} \, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U} \, \Sigma \, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

$$\begin{aligned}
\mathbf{h}(\mathbf{X}) &= \mathbf{V}_{\cdot, \leq k}^\top \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{X}^\top \, \mathbf{X})^{-1} \, (\mathbf{X}^\top \, \mathbf{X}) && \leftarrow \text{multiplying by identity} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{U} \, \Sigma \, \mathbf{V}^\top)^{-1} \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X}) && \leftarrow \text{replace with SVD}
\end{aligned}$$

Sketch of proof

matrix where columns are $\mathbf{x}^{(t)}$

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\hat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix of all hidden layers (could be any encoder)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k} \, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U} \, \Sigma \, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

$$\begin{aligned}
\mathbf{h}(\mathbf{X}) &= \mathbf{V}_{\cdot, \leq k}^\top \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{X}^\top \, \mathbf{X})^{-1} \, (\mathbf{X}^\top \, \mathbf{X}) && \leftarrow \text{multiplying by identity} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{U} \, \Sigma \, \mathbf{V}^\top)^{-1} \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X}) && \leftarrow \text{replace with SVD} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \Sigma \, \mathbf{V}^\top)^{-1} \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{U}^\top \, \mathbf{U} = \mathbf{I} \text{ (orthonormal)}
\end{aligned}$$

Sketch of proof

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\hat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix where columns are } \mathbf{x}^{(t)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k} \, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U} \, \Sigma \, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

$$\begin{aligned}
\mathbf{h}(\mathbf{X}) &= \mathbf{V}_{\cdot, \leq k}^\top \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{X}^\top \, \mathbf{X})^{-1} \, (\mathbf{X}^\top \, \mathbf{X}) && \leftarrow \text{multiplying by identity} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{U} \, \Sigma \, \mathbf{V}^\top)^{-1} \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X}) && \leftarrow \text{replace with SVD} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \Sigma \, \mathbf{V}^\top)^{-1} \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{U}^\top \, \mathbf{U} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \mathbf{V}^\top \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}(\Sigma^\top\Sigma)^{-1}\mathbf{V}^\top \, \mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top = \mathbf{I}
\end{aligned}$$

Sketch of proof

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix where columns are } \mathbf{x}^{(t)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k} \, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U} \, \Sigma \, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

$$\begin{aligned}
\mathbf{h}(\mathbf{X}) &= \mathbf{V}_{\cdot, \leq k}^\top \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{X}^\top \, \mathbf{X})^{-1} \, (\mathbf{X}^\top \, \mathbf{X}) && \leftarrow \text{multiplying by identity} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{U} \, \Sigma \, \mathbf{V}^\top)^{-1} \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X}) && \leftarrow \text{replace with SVD} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \Sigma \, \mathbf{V}^\top)^{-1} \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{U}^\top \, \mathbf{U} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \mathbf{V}^\top \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}(\Sigma^\top\Sigma)^{-1}\mathbf{V}^\top \, \mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top = \mathbf{I} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}^\top \, \mathbf{V} = \mathbf{I} \text{ (orthonormal)}
\end{aligned}$$

Sketch of proof

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix where columns are } \mathbf{x}^{(t)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k} \, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U} \, \Sigma \, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

$$\begin{aligned}
\mathbf{h}(\mathbf{X}) &= \mathbf{V}_{\cdot, \leq k}^\top \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{X}^\top \, \mathbf{X})^{-1} \, (\mathbf{X}^\top \, \mathbf{X}) && \leftarrow \text{multiplying by identity} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{U} \, \Sigma \, \mathbf{V}^\top)^{-1} \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X}) && \leftarrow \text{replace with SVD} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \Sigma \, \mathbf{V}^\top)^{-1} \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{U}^\top \, \mathbf{U} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \mathbf{V}^\top \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}(\Sigma^\top\Sigma)^{-1}\mathbf{V}^\top \, \mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top = \mathbf{I} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}^\top \, \mathbf{V} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{I}_{\leq k, \cdot} \, (\Sigma^\top \, \Sigma)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \text{idem}
\end{aligned}$$

Sketch of proof

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\hat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix where columns are } \mathbf{x}^{(t)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k} \, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U} \, \Sigma \, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

$$\begin{aligned}
\mathbf{h}(\mathbf{X}) &= \mathbf{V}_{\cdot, \leq k}^\top \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{X}^\top \, \mathbf{X})^{-1} \, (\mathbf{X}^\top \, \mathbf{X}) && \leftarrow \text{multiplying by identity} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{U} \, \Sigma \, \mathbf{V}^\top)^{-1} \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X}) && \leftarrow \text{replace with SVD} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \Sigma \, \mathbf{V}^\top)^{-1} \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{U}^\top \, \mathbf{U} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \mathbf{V}^\top \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}(\Sigma^\top\Sigma)^{-1}\mathbf{V}^\top \, \mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top = \mathbf{I} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}^\top \, \mathbf{V} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{I}_{\leq k, \cdot} \, (\Sigma^\top \, \Sigma)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \text{idem} \\
&= \mathbf{I}_{\leq k, \cdot} \, \Sigma^{-1} \, (\Sigma^\top)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow (\Sigma^\top\Sigma)^{-1} = \Sigma^{-1}(\Sigma^\top)^{-1}
\end{aligned}$$

Sketch of proof

$$\min_\theta \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\hat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix where columns are } \mathbf{x}^{(t)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k} \, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U} \, \Sigma \, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

$$\begin{aligned}
\mathbf{h}(\mathbf{X}) &= \mathbf{V}_{\cdot, \leq k}^\top \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{X}^\top \, \mathbf{X})^{-1} \, (\mathbf{X}^\top \, \mathbf{X}) && \leftarrow \text{multiplying by identity} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{U} \, \Sigma \, \mathbf{V}^\top)^{-1} \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X}) && \leftarrow \text{replace with SVD} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \Sigma \, \mathbf{V}^\top)^{-1} \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{U}^\top \, \mathbf{U} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \mathbf{V}^\top \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}(\Sigma^\top\Sigma)^{-1}\mathbf{V}^\top \, \mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top = \mathbf{I} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}^\top \, \mathbf{V} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{I}_{\leq k, \cdot} \, (\Sigma^\top \, \Sigma)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \text{idem} \\
&= \mathbf{I}_{\leq k, \cdot} \, \Sigma^{-1} \, (\Sigma^\top)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow (\Sigma^\top\Sigma)^{-1} = \Sigma^{-1}(\Sigma^\top)^{-1} \\
&= \mathbf{I}_{\leq k, \cdot} \, \Sigma^{-1} \, \mathbf{U}^\top \, \mathbf{X}
\end{aligned}$$

Sketch of proof

$$\min_{\theta} \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\hat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix where columns are } \mathbf{x}^{(t)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k} \, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U} \, \Sigma \, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

$$\begin{aligned}
\mathbf{h}(\mathbf{X}) &= \mathbf{V}_{\cdot, \leq k}^\top \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{X}^\top \, \mathbf{X})^{-1} \, (\mathbf{X}^\top \, \mathbf{X}) && \leftarrow \text{multiplying by identity} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{U} \, \Sigma \, \mathbf{V}^\top)^{-1} \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X}) && \leftarrow \text{replace with SVD} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \Sigma \, \mathbf{V}^\top)^{-1} \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{U}^\top \, \mathbf{U} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \mathbf{V}^\top \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}(\Sigma^\top\Sigma)^{-1}\mathbf{V}^\top \, \mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top = \mathbf{I} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}^\top \, \mathbf{V} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{I}_{\leq k, \cdot} \, (\Sigma^\top \, \Sigma)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \text{idem} \\
&= \mathbf{I}_{\leq k, \cdot} \, \Sigma^{-1} \, (\Sigma^\top)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow (\Sigma^\top\Sigma)^{-1} = \Sigma^{-1}(\Sigma^\top)^{-1} \\
&= \mathbf{I}_{\leq k, \cdot} \, \Sigma^{-1} \, \mathbf{U}^\top \, \mathbf{X} \\
&= \Sigma_{\leq k, \leq k}^{-1} \, (\mathbf{U}_{\cdot, \leq k})^\top \, \mathbf{X} && \leftarrow \text{multiplying by } \mathbf{I}_{\leq k, \cdot} \text{ selects the } k \text{ first rows}
\end{aligned}$$

Sketch of proof

$$\min_\theta \sum_t \frac{1}{2} \sum_i (x_i^{(t)} - \underbrace{\widehat{x}_i^{(t)}}_{\text{based on linear encoder}})^2 \quad \geq \quad \min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\overbrace{\mathbf{X}}^{\text{matrix where columns are } \mathbf{x}^{(t)}} - \mathbf{W}^* \underbrace{\mathbf{h}(\mathbf{X})}_{\text{matrix of all hidden layers (could be any encoder)}}||_F^2$$

$$\arg\min_{\mathbf{W}^*, \mathbf{h}(\mathbf{X})} \frac{1}{2} ||\mathbf{X} - \mathbf{W}^* \mathbf{h}(\mathbf{X})||_F^2 \quad = \quad \left(\mathbf{W}^* \leftarrow \mathbf{U}_{\cdot, \leq k} \, \Sigma_{\leq k, \leq k}, \;\; \mathbf{h}(\mathbf{X}) \leftarrow \mathbf{V}_{\cdot, \leq k}^\top\right)$$

based on previous theorem, where $\mathbf{X} = \mathbf{U} \, \Sigma \, \mathbf{V}^\top$ and k is the hidden layer size

Let's show $\mathbf{h}(\mathbf{X})$ is a linear encoder:

$$\begin{aligned}
\mathbf{h}(\mathbf{X}) &= \mathbf{V}_{\cdot, \leq k}^\top \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{X}^\top \, \mathbf{X})^{-1} \, (\mathbf{X}^\top \, \mathbf{X}) && \leftarrow \text{multiplying by identity} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{U} \, \Sigma \, \mathbf{V}^\top)^{-1} \, (\mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X}) && \leftarrow \text{replace with SVD} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, (\mathbf{V} \, \Sigma^\top \, \Sigma \, \mathbf{V}^\top)^{-1} \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{U}^\top \, \mathbf{U} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \mathbf{V}^\top \, \mathbf{V} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}(\Sigma^\top\Sigma)^{-1}\mathbf{V}^\top \, \mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top = \mathbf{I} \\
&= \mathbf{V}_{\cdot, \leq k}^\top \, \mathbf{V} \, (\Sigma^\top \, \Sigma)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \mathbf{V}^\top \, \mathbf{V} = \mathbf{I} \text{ (orthonormal)} \\
&= \mathbf{I}_{\leq k, \cdot} \, (\Sigma^\top \, \Sigma)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow \text{idem} \\
&= \mathbf{I}_{\leq k, \cdot} \, \Sigma^{-1} \, (\Sigma^\top)^{-1} \, \Sigma^\top \, \mathbf{U}^\top \, \mathbf{X} && \leftarrow (\Sigma^\top\Sigma)^{-1} = \Sigma^{-1}(\Sigma^\top)^{-1} \\
&= \mathbf{I}_{\leq k, \cdot} \, \Sigma^{-1} \, \mathbf{U}^\top \, \mathbf{X} \\
&= \underbrace{\Sigma_{\leq k, \leq k}^{-1} \, (\mathbf{U}_{\cdot, \leq k})^\top}_{\text{this is a linear encoder}} \, \mathbf{X} && \leftarrow \text{multiplying by } \mathbf{I}_{\leq k, \cdot} \text{ selects the } k \text{ first rows}
\end{aligned}$$

AUTOENCODER

Topics: optimality of a linear autoencoder

- So an optimal pair of encoder and decoder is

$$\mathbf{h}(\mathbf{x}) = \underbrace{\left(\Sigma^{-1}_{\leq k,\leq k} \left(\mathbf{U}_{\cdot,\leq k}\right)^\top\right)}_{\mathbf{W}} \mathbf{x} \qquad \hat{\mathbf{x}} = \underbrace{\left(\mathbf{U}_{\cdot,\leq k}\, \Sigma_{\leq k,\leq k}\right)}_{\mathbf{W}^*} \mathbf{h}(\mathbf{x})$$

  - for the sum of squared difference error
  - for an autoencoder with a linear decoder
  - where optimality means "has the lowest training reconstruction error"
- If inputs are normalized as follows: $\mathbf{x}^{(t)} \leftarrow \frac{1}{\sqrt{T}} \left(\mathbf{x}^{(t)} - \frac{1}{T} \sum_{t'=1}^{T} \mathbf{x}^{(t')}\right)$
  - encoder corresponds to Principal Component Analysis (PCA)
    - singular values and (left) vectors = the eigenvalues/vectors of covariance matrix