Neural networks
Sparse coding - definition
Topics: unsupervised learning

- Unsupervised learning: only use the inputs $x^{(t)}$ for learning
  - automatically extract meaningful features for your data
  - leverage the availability of unlabeled data
  - add a data-dependent regularizer to trainings

- We will see 3 neural networks for unsupervised learning
  - restricted Boltzmann machines
  - autoencoders
  - sparse coding model
Topics: sparse coding

- For each \( x^{(t)} \) find a latent representation \( h^{(t)} \) such that:
  - it is sparse: the vector \( h^{(t)} \) has many zeros
  - we can reconstruct the original input \( x^{(t)} \) as well as possible
- More formally:

\[
\min_D \frac{1}{T} \sum_{t=1}^{T} \min_{h^{(t)}} \frac{1}{2} \|x^{(t)} - Dh^{(t)}\|_2^2 + \lambda \|h^{(t)}\|_1
\]

- we also constrain the columns of \( D \) to be of norm 1
  - otherwise, \( D \) could grow big while \( h^{(t)} \) becomes small to satisfy the prior
- sometimes the columns are constrained to be no greater than 1
**SPARSE CODING**

**Topics:** sparse coding

- For each $x^{(t)}$ find a latent representation $h^{(t)}$ such that:
  - it is sparse: the vector $h^{(t)}$ has many zeros
  - we can reconstruct the original input $x^{(t)}$ as well as possible

- More formally:
  
  $$
  \min_D \frac{1}{T} \sum_{t=1}^{T} \min_{h^{(t)}} \frac{1}{2}||x^{(t)} - Dh^{(t)}||_2^2 + \lambda||h^{(t)}||_1
  $$

- we also constrain the columns of $D$ to be of norm $1$
  - otherwise, $D$ could grow big while $h^{(t)}$ becomes small to satisfy the prior
- sometimes the columns are constrained to be no greater than 1
**SPARSE CODING**

**Topics:** sparse coding

- For each $\mathbf{x}^{(t)}$ find a latent representation $\mathbf{h}^{(t)}$ such that:
  - it is sparse: the vector $\mathbf{h}^{(t)}$ has many zeros
  - we can reconstruct the original input $\mathbf{x}^{(t)}$ as well as possible

- More formally:

$$
\min_{\mathbf{x}} \frac{1}{T} \sum_{t=1}^{T} \min_{\mathbf{h}^{(t)}} \frac{1}{2} \| \mathbf{x}^{(t)} - \mathbf{D} \mathbf{h}^{(t)} \|_2^2 + \lambda \| \mathbf{h}^{(t)} \|_1
$$

  - we also constrain the columns of $\mathbf{D}$ to be of norm 1
    - otherwise, $\mathbf{D}$ could grow big while $\mathbf{h}^{(t)}$ becomes small to satisfy the prior
  - sometimes the columns are constrained to be no greater than 1
Topics: sparse coding

• For each \( \mathbf{x}^{(t)} \) find a latent representation \( \mathbf{h}^{(t)} \) such that:
  
  ‣ it is sparse: the vector \( \mathbf{h}^{(t)} \) has many zeros
  
  ‣ we can reconstruct the original input \( \mathbf{x}^{(t)} \) as well as possible

• More formally:

\[
\min_{\mathbf{D}} \frac{1}{T} \sum_{t=1}^{T} \min_{\mathbf{h}^{(t)}} \frac{1}{2} \| \mathbf{x}^{(t)} - \mathbf{D} \mathbf{h}^{(t)} \|_2^2 + \lambda \| \mathbf{h}^{(t)} \|_1
\]

  ‣ we also constrain the columns of \( \mathbf{D} \) to be of norm 1
    
    - otherwise, \( \mathbf{D} \) could grow big while \( \mathbf{h}^{(t)} \) becomes small to satisfy the prior
  
  ‣ sometimes the columns are constrained to be no greater than 1
SPARSE CODING

Topics: sparse coding

• For each $x^{(t)}$ find a latent representation $h^{(t)}$ such that:
  ‣ it is sparse: the vector $h^{(t)}$ has many zeros
  ‣ we can reconstruct the original input $x^{(t)}$ as well as possible

• More formally:

$$\min_{D} \frac{1}{T} \sum_{t=1}^{T} \min_{h^{(t)}} \frac{1}{2} \|x^{(t)} - Dh^{(t)}\|_2^2 + \lambda \|h^{(t)}\|_1$$

  ‣ we also constrain the columns of $D$ to be of norm 1
    - otherwise, $D$ could grow big while $h^{(t)}$ becomes small to satisfy the prior
  ‣ sometimes the columns are constrained to be no greater than 1
**Topics:** sparse coding

- For each \( x^{(t)} \) find a latent representation \( h^{(t)} \) such that:
  - it is sparse: the vector \( h^{(t)} \) has many zeros
  - we can reconstruct the original input \( x^{(t)} \) as well as possible

- More formally:

\[
\min_D \frac{1}{T} \sum_{t=1}^T \min_{h^{(t)}} \frac{1}{2} \| x^{(t)} - D h^{(t)} \|^2 + \lambda \| h^{(t)} \|_1
\]

- \( D \) is equivalent to the autoencoder output weight matrix
- however, \( h(x^{(t)}) \) is now a complicated function of \( x^{(t)} \)
  - encoder is the minimization \( h(x^{(t)}) = \arg\min_{h^{(t)}} \frac{1}{2} \| x^{(t)} - D h^{(t)} \|^2 + \lambda \| h^{(t)} \|_1 \)
**SPARSE CODING**

**Topics:** dictionary

- Can also write \( \hat{x}(t) = D \cdot h(x(t)) = \sum_{k \text{ s.t. } h(x(t))_k \neq 0} D_{.,k} h(x(t))_k \)

\[
\begin{align*}
\begin{array}{c|c|c|c|c|c|c}
7 & + & 1 & + & 1 & + & 1 \\
\hline
+ & 1 & 1 & + & 0.8 & + & 0.8 \\
\end{array}
\end{align*}
\]

- we also refer to \( D \) as the dictionary
  - in certain applications, we know what dictionary matrix to use
  - often however, we have to learn it
Topics: dictionary

- Can also write \( \hat{x}^{(t)} = D h(x^{(t)}) = \sum_{k \text{ s.t. } h(x^{(t)})_k \neq 0} D_{:,k} h(x^{(t)})_k \)

- we also refer to \( D \) as the dictionary
  - in certain applications, we know what dictionary matrix to use
  - often however, we have to learn it
SPARSE CODING

Topics: dictionary

- Can also write \( \hat{x}^{(t)} = D \ h(x^{(t)}) = \sum_{k \text{ s.t. } h(x^{(t)})_k \neq 0} D_{.,k} \ h(x^{(t)})_k \)

\[
\begin{align*}
7 &= 1 \cdot \text{image 1} + 1 \cdot \text{image 2} + 1 \cdot \text{image 3} + 1 \cdot \text{image 4} + 1 \cdot \text{image 5} \\
&\quad + 1 \cdot \text{image 6} + 1 \cdot \text{image 7} + 0.8 \cdot \text{image 8} + 0.8 \cdot \text{image 9}
\end{align*}
\]

- we also refer to \( D \) as the dictionary
  - in certain applications, we know what dictionary matrix to use
  - often however, we have to learn it