Neural networks

Sparse coding - inference (ISTA algorithm)

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 much as possible
- More formally:

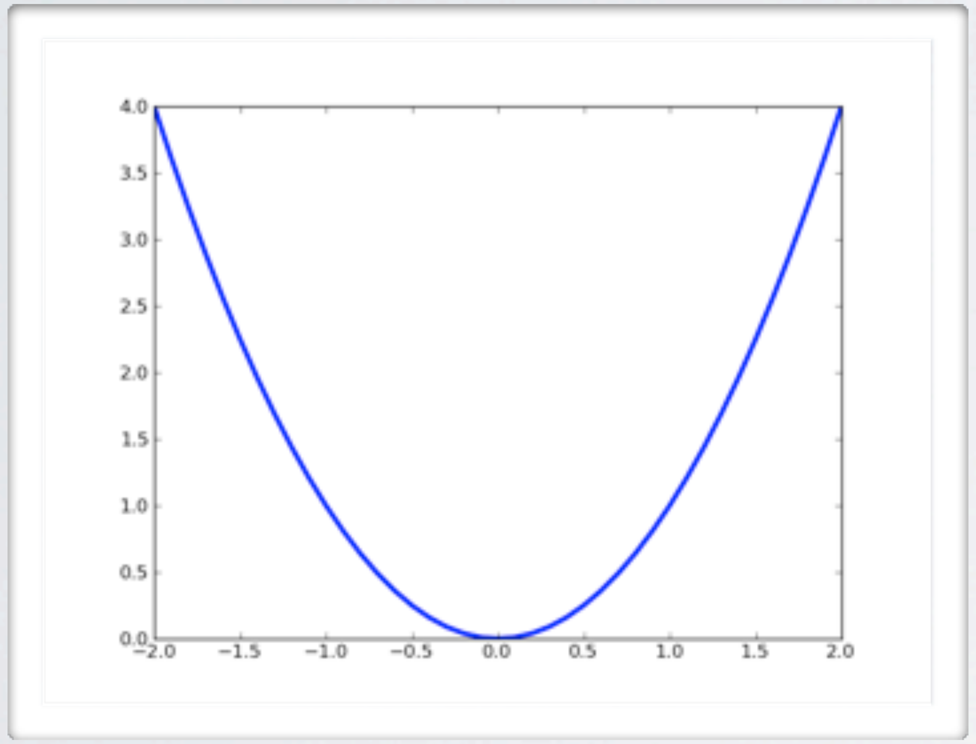
$$\min_{\mathbf{D}} \frac{1}{T} \sum_{t=1}^{T} \min_{\mathbf{h}^{(t)}} \frac{1}{2} \overbrace{||\mathbf{x}^{(t)} - \underbrace{\mathbf{D}\,\mathbf{h}^{(t)}}_{\text{reconstruction } \widehat{\mathbf{x}}^{(t)}}||_2^2}^{\text{reconstruction error}} + \underbrace{\lambda}_{\text{reconstruction vs. sparsity control}} \overbrace{||\mathbf{h}^{(t)}||_1}^{\text{sparsity penalty}}$$

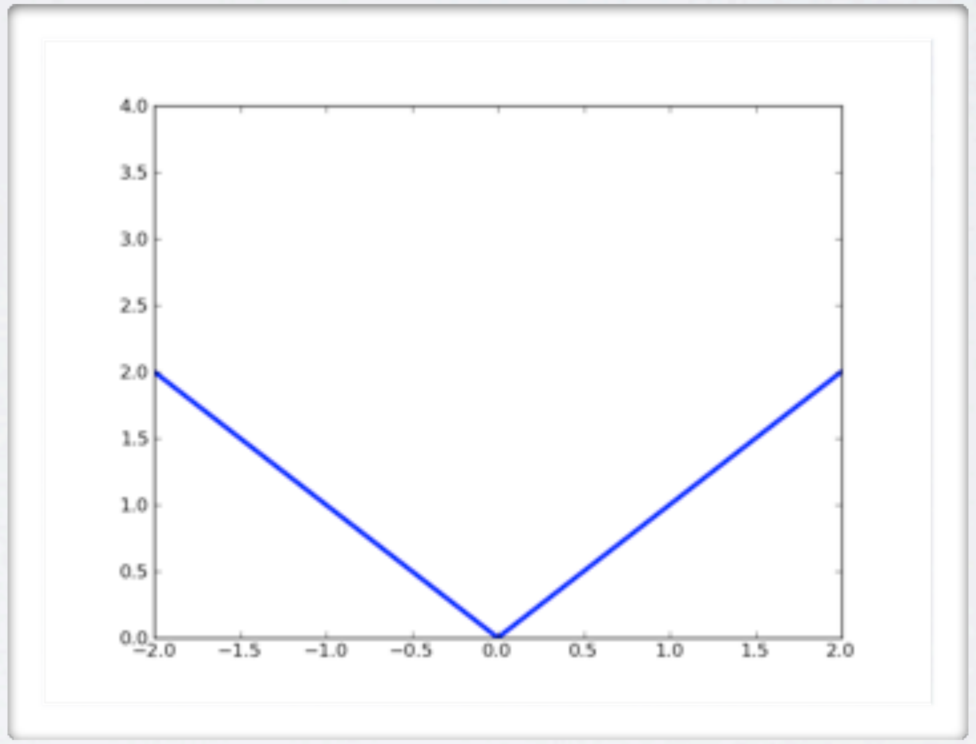
  - $\mathbf{D}$ is equivalent to the autoencoder output weight matrix
  - however, $\mathbf{h}(\mathbf{x}^{(t)})$ is now a complicated function of $\mathbf{x}^{(t)}$
    - encoder is the minimization $\mathbf{h}(\mathbf{x}^{(t)}) = \arg\min_{\mathbf{h}^{(t)}} \frac{1}{2}||\mathbf{x}^{(t)} - \mathbf{D}\,\mathbf{h}^{(t)}||_2^2 + \lambda||\mathbf{h}^{(t)}||_1$

SPARSE CODING

Topics: inference of sparse codes

- Given $\mathbf{D}$, how do we compute $\mathbf{h}(\mathbf{x}^{(t)})$
  - we want to optimize $l(\mathbf{x}^{(t)}) = \frac{1}{2}||\mathbf{x}^{(t)} - \mathbf{D}\,\mathbf{h}^{(t)}||_2^2 + \lambda||\mathbf{h}^{(t)}||_1$ w.r.t. $\mathbf{h}^{(t)}$





  - we could use a gradient descent method:

$$\nabla_{\mathbf{h}^{(t)}} l(\mathbf{x}^{(t)}) = \mathbf{D}^\top(\mathbf{D}\,\mathbf{h}^{(t)} - \mathbf{x}^{(t)}) + \lambda\,\mathrm{sign}(\mathbf{h}^{(t)})$$

SPARSE CODING

Topics: inference of sparse codes

- For a single hidden unit:

$$\frac{\partial}{\partial h_k^{(t)}} l(\mathbf{x}^{(t)}) = (\mathbf{D}_{\cdot,k})^\top (\mathbf{D}\,\mathbf{h}^{(t)} - \mathbf{x}^{(t)}) + \lambda\,\mathrm{sign}(h_k^{(t)})$$

  - issue: L1 norm not differentiable at 0
    - very unlikely for gradient descent to "land" on $h_k^{(t)} = 0$ (even if it's the solution)
  - solution: if $h_k^{(t)}$ changes sign because of L1 norm gradient, clamp to 0
  - each hidden unit update would be performed as follows:
    - $h_k^{(t)} \Longleftarrow h_k^{(t)} - \alpha (\mathbf{D}_{\cdot,k})^\top (\mathbf{D}\;\mathbf{h}^{(t)} - \mathbf{x}^{(t)})$
    - if $\mathrm{sign}(h_k^{(t)}) \neq \mathrm{sign}(h_k^{(t)} - \alpha\,\lambda\,\mathrm{sign}(h_k^{(t)}))$ then: $h_k^{(t)} \Longleftarrow 0$
    - else: $h_k^{(t)} \Longleftarrow h_k^{(t)} - \alpha\,\lambda\,\mathrm{sign}(h_k^{(t)})$

SPARSE CODING

Topics: inference of sparse codes

- For a single hidden unit:

$$\frac{\partial}{\partial h_k^{(t)}} l(\mathbf{x}^{(t)}) = (\mathbf{D}_{\cdot,k})^\top (\mathbf{D}\,\mathbf{h}^{(t)} - \mathbf{x}^{(t)}) + \lambda\,\mathrm{sign}(h_k^{(t)})$$

  - issue: L1 norm not differentiable at 0
    - very unlikely for gradient descent to "land" on $h_k^{(t)} = 0$ (even if it's the solution)
  - solution: if $h_k^{(t)}$ changes sign because of L1 norm gradient, clamp to 0
  - each hidden unit update would be performed as follows:
    - $h_k^{(t)} \Longleftarrow h_k^{(t)} - \alpha(\mathbf{D}_{\cdot,k})^\top (\mathbf{D}\,\mathbf{h}^{(t)} - \mathbf{x}^{(t)})$ } update from reconstruction
    - if $\mathrm{sign}(h_k^{(t)}) \neq \mathrm{sign}(h_k^{(t)} - \alpha\,\lambda\,\mathrm{sign}(h_k^{(t)}))$ then: $h_k^{(t)} \Longleftarrow 0$
    - else: $h_k^{(t)} \Longleftarrow h_k^{(t)} - \alpha\,\lambda\,\mathrm{sign}(h_k^{(t)})$

SPARSE CODING

Topics: inference of sparse codes

- For a single hidden unit:

$$\frac{\partial}{\partial h_k^{(t)}} l(\mathbf{x}^{(t)}) = (\mathbf{D}_{\cdot,k})^\top (\mathbf{D}\,\mathbf{h}^{(t)} - \mathbf{x}^{(t)}) + \lambda \,\mathrm{sign}(h_k^{(t)})$$

  - issue: L1 norm not differentiable at 0
    - very unlikely for gradient descent to "land" on $h_k^{(t)} = 0$ (even if it's the solution)
  - solution: if $h_k^{(t)}$ changes sign because of L1 norm gradient, clamp to 0
  - each hidden unit update would be performed as follows:
    - $h_k^{(t)} \Longleftarrow h_k^{(t)} - \alpha (\mathbf{D}_{\cdot,k})^\top (\mathbf{D}\,\mathbf{h}^{(t)} - \mathbf{x}^{(t)})$ } update from reconstruction
    - if $\mathrm{sign}(h_k^{(t)}) \neq \mathrm{sign}(h_k^{(t)} - \alpha\,\lambda\,\mathrm{sign}(h_k^{(t)}))$ then: $h_k^{(t)} \Longleftarrow 0$
    - else: $h_k^{(t)} \Longleftarrow h_k^{(t)} - \alpha\,\lambda\,\mathrm{sign}(h_k^{(t)})$

    } update from sparsity

SPARSE CODING

Topics: ISTA (Iterative Shrinkage and Thresholding Algorithm)

- This process corresponds to the ISTA algorithm:

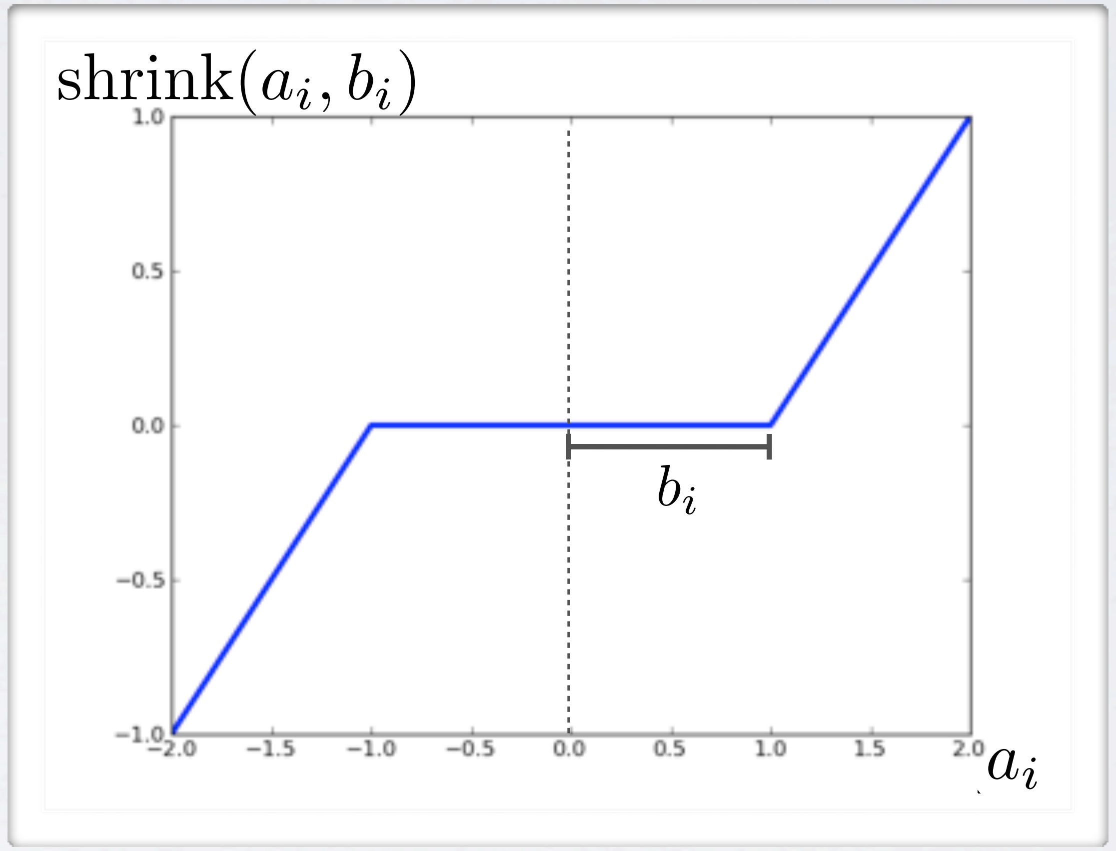
  - initialize $\mathbf{h}^{(t)}$ (for instance to 0)
  - while $\mathbf{h}^{(t)}$ has not converged
    - $\mathbf{h}^{(t)} \Longleftarrow \mathbf{h}^{(t)} - \alpha\, \mathbf{D}^\top(\mathbf{D}\,\mathbf{h}^{(t)} - \mathbf{x}^{(t)})$
    - $\mathbf{h}^{(t)} \Longleftarrow \mathrm{shrink}(\mathbf{h}^{(t)}, \alpha\, \lambda)$
  - return $\mathbf{h}^{(t)}$

- Here $\mathrm{shrink}(\mathbf{a}, \mathbf{b}) = [\ldots, \mathrm{sign}(a_i)\ \max(|a_i| - b_i, 0), \ldots]$
- Will converge if $\frac{1}{\alpha}$ is bigger than the largest eigenvalue of $\mathbf{D}^\top\mathbf{D}$

SPARSE CODING

Topics: ISTA (Iterative Shrinkage and Thresholding Algorithm)

- This process corresponds to the ISTA algorithm:
  - initialize $\mathbf{h}^{(t)}$ (for instance to 0)
  - while $\mathbf{h}^{(t)}$ has not converged
    - $\mathbf{h}^{(t)} \Longleftarrow \mathbf{h}^{(t)} - \alpha\, \mathbf{D}^\top (\mathbf{D}\, \mathbf{h}^{(t)} - \mathbf{x}^{(t)})$
    - $\mathbf{h}^{(t)} \Longleftarrow \text{shrink}(\mathbf{h}^{(t)}, \alpha\, \lambda)$
  - return $\mathbf{h}^{(t)}$



- Here $\text{shrink}(\mathbf{a}, \mathbf{b}) = [\ldots, \text{sign}(a_i)\ \max(|a_i| - b_i, 0), \ldots]$
- Will converge if $\frac{1}{\alpha}$ is bigger than the largest eigenvalue of $\mathbf{D}^\top \mathbf{D}$

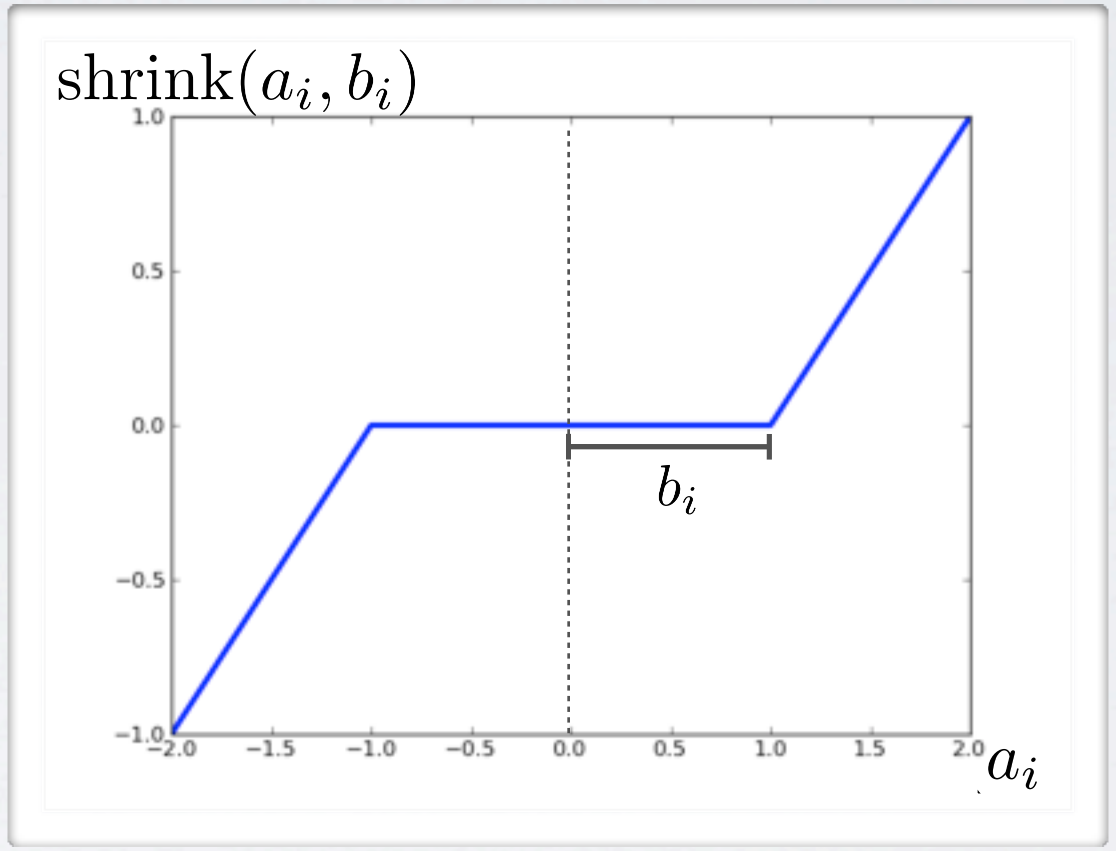
SPARSE CODING

Topics: ISTA (Iterative Shrinkage and Thresholding Algorithm)

- This process corresponds to the ISTA algorithm:

  - initialize $\mathbf{h}^{(t)}$ (for instance to 0)
  - while $\mathbf{h}^{(t)}$ has not converged
    - $\mathbf{h}^{(t)} \Longleftarrow \mathbf{h}^{(t)} - \alpha\, \mathbf{D}^\top (\mathbf{D}\, \mathbf{h}^{(t)} - \mathbf{x}^{(t)})$
    - $\mathbf{h}^{(t)} \Longleftarrow \mathrm{shrink}(\mathbf{h}^{(t)}, \alpha\, \lambda)$
  - return $\mathbf{h}^{(t)}$

this is $\mathbf{h}(\mathbf{x}^{(t)})$



- Here $\mathrm{shrink}(\mathbf{a}, \mathbf{b}) = [\ldots, \mathrm{sign}(a_i)\ \max(|a_i| - b_i, 0), \ldots]$
- Will converge if $\frac{1}{\alpha}$ is bigger than the largest eigenvalue of $\mathbf{D}^\top \mathbf{D}$

SPARSE CODING

Topics: coordinate descent for sparse coding inference

- ISTA updates all hidden units simultaneously
  - this is wasteful if many hidden units have already converged

- Idea: update only the "most promising" hidden unit
  - see coordinate descent algorithm in
    - Learning Fast Approximations of Sparse Coding.
      Gregor and Lecun, 2010.
  - this algorithm has the advantage of not requiring a learning rate α