Projects

Other ideas
- nonlinear extensions of PCA
  - kernel PCA
  - locally linear embeddings
  - non-negative matrix factorization
  - autoencoder NN's

Other ideas
- sparse weights or small # of hidden nodes?
- data augmentation
- linear activation?
  (autoencoder NN with linear activation = PCA)

- LASSO
  - group LASSO
  - elastic net (LASSO + Ridge)
  - multi-task learning.

- Low-rank tensors
minimize \( w \bullet \ell(w) + \lambda \gamma(w) \)

subject to

\[ f(w) \]

\[ \hat{w} = \arg\min_w f(w) \]

assume \( f(w) = \sum_{i=1}^n f_i(w) \)

\[ e.g. \quad f(w) = \|y - Xw\|_2^2 \]
\[ = \sum_{i} (y_i - x_i^T w)^2 \]
\[ \Rightarrow f_i(w) = (y_i - x_i^T w)^2 \]

**SGD:**

- At iteration \( k \), choose \( i_k \in \{1, \ldots, n\} \)

\[ \hat{w}^{(k+1)} = \hat{w}^{(k)} - \frac{\tau}{2} \nabla f_{i_k}^{(k)}(\hat{w}^{(k)}) \]

most commonly: \( i_k \sim \text{uniformly at random} \)

for \( 1, 2, \ldots, n \)
A. cyclic \( i_k = k \mod n \)
   
   if \( n = 3 \), \( i_k \)'s: 1, 2, 3, 1, 2, 3, ...

B. permutation

\[
i_k \text{'s: } 3, 1, 2, 1, 3, 2, 2, 1, 3
\]

epoch 1 epoch 2

C. \( i_k \)'s unit at random

\[
i_k \text{'s: } 1, 1, 1, 1, 3, 1, 2, 2, 3, 1
\]

if \( i_k \)'s are uniform at random,

\[
E[f_{i_k}] = \frac{f}{n}
\]
Ex: \( f(w) = \|y - Xw\|_2^2 + \lambda \|w\|_2^2 \)

\[
= \sum_{i=1}^{n} \left[ (y_i - x_i^T w)^2 + \frac{\lambda}{n} \|w\|_2^2 \right]
\]

\( f_i(w) \)

\( \nabla f_i(w) = -2(y_i - x_i^T w) x_i + \frac{2\lambda}{n} w \)

SGD: \( \hat{w}(k+1) = \hat{w}(k) - \frac{\beta}{4} \left[ -(y_i - x_i^T \hat{w}(k)) x_i + \frac{\lambda}{n} \hat{w}(k) \right] \)
can replace gradients with subgradients:

Recall: if $f$ is convex and differentiable:
$$f(u) \geq f(w) + (u-w)^T \nabla f(w)$$

if $f$ is convex but not differentiable, then $v$ is a subgradient of $f$ at $w$ if
$$f(u) \geq f(w) + (u-w)^T v$$

set of subgradients at $w$ is called "differential set" denoted $\partial f(w)$
write $v \in \partial f(w)$
e.g., \( r(w) = \|w\|_1 = \sum_{j=1}^{p} |w_j| \)

for \( w_j \neq 0 \), \( |w_j| \) is differentiable
derivative is \( \text{sign}(w_j) \)

for \( w_j = 0 \), then \( v_j \in [-1, +1] \)

for \( v \in \text{arg}(r(w)) \)

then \( v_j = \begin{cases} \text{sign}(w_j) & \text{if } w_j \neq 0 \\ \in [-1, +1] & \text{if } w_j = 0 \end{cases} \)

popular choice: \( v_j = \begin{cases} \text{sign}(w_j) & \text{if } w_j \neq 0 \\ 0 & \text{if } w_j = 0 \end{cases} = \text{"sign}(w)\)"
\[ E_{2} \quad \text{SGD:} \quad f(w) = \|y - Xw\|_2^2 + \lambda \|w\|_1 \quad \text{(LASSO)} \]

\[ = \sum_{i=1}^{n} \left[ (y_i - x_i^T w)^2 + \frac{\lambda}{n} \|w\|_1 \right] \]

\[ f_i(w) \]

Let \( v = -2(y_i - x_i^T w) x_i + \frac{\lambda}{n} \text{sign}(w) \)

SGD: \( \hat{w}^{(k+1)} = \hat{w}^{(k)} - \frac{\lambda}{2} (-v_i) \)

\[ = \hat{w}^{(k)} + \frac{1}{T}(y_i - x_i^T \hat{w}^{(k)}) - \frac{T \lambda}{2n} \text{sign}(\hat{w}^{(k)}) \)
$$h_m = \sigma \left( \sum_{j=1}^{P} X_j W_{m,j} \right)$$

$$y_k = \sigma \left( \sum_{m=1}^{M} h_m V_{k,m} \right)$$
\[ y = x^T w \]

\[ y = \sigma(x^T w) \]

- \( \sigma(z) = \max(0, z) = \text{ReLU} \)
- \( \sigma(z) = \text{sign}(z) \in [-1, 0, +1] \)
- \( \sigma(z) = \frac{1}{1 + e^{-z}} \in [0, 1] \)