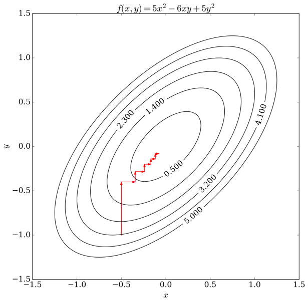
Optimization in Machine Learning

Coordinate descent

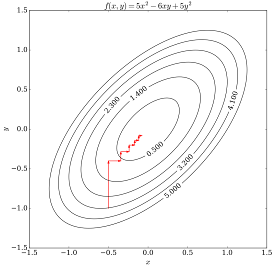


Learning goals

- Axes as descent direction
- CD on linear model and LASSO
- Soft thresholding

COORDINATE DESCENT

- **Assumption:** Objective function not differentiable
- **Idea:** Instead of gradient, use coordinate directions for descent
- First: Select starting point $\boldsymbol{x}^{[0]} = (x_1^{[0]}, \ldots, x_d^{[0]})$
- Step t: Minimize f along x_i for each dimension i for fixed $x_1^{[t]}, \ldots, x_{i-1}^{[t]}$ and $x_{i+1}^{[t-1]}, \ldots, x_d^{[t-1]}$:



Source: Wikipedia (Coordinate descent)

COORDINATE DESCENT

- Minimum is determined with (exact / inexact) line search
- Order of dimensions can be any permutation of $\{1, 2, \ldots, d\}$
- **Convergence:**
  - f convex differentiable
  - f sum of convex differentiable and *convex separable* function:

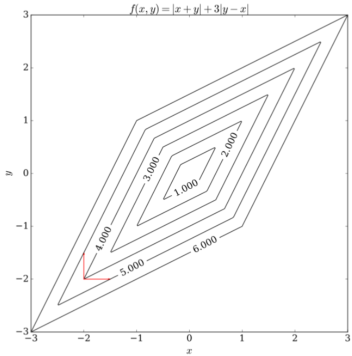
$$f(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^{d} h_i(x_i),$$

where g convex differentiable and h_i convex

COORDINATE DESCENT

Not convergence in general for convex functions.

Counterexample:



Source: Wikipedia (Coordinate descent)

EXAMPLE: LINEAR REGRESSION

Minimize LM with L2-loss via CD:

$$\min g(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)} \right)^2 = \min_{\boldsymbol{\theta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times d}$ with columns $\mathbf{x}_1, \dots, \mathbf{x}_d \in \mathbb{R}^n$.

Assume: Scaled data, i.e., $\mathbf{X}^\top \mathbf{X} = I_d$ (just to get intuition)

Then:

$$\begin{aligned} g(\boldsymbol{\theta}) &= \frac{1}{2}\mathbf{y}^\top \mathbf{y} + \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta} - \mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} \\ &\overset{(*)}{=} \frac{1}{2}\mathbf{y}^\top \mathbf{y} + \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta} - \mathbf{y}^\top \sum_{k=1}^{p} \mathbf{x}_k \theta_k \end{aligned}$$

$^{(*)}$ $\mathbf{X}\boldsymbol{\theta} = \mathbf{x}_1\theta_1 + \mathbf{x}_2\theta_2 + \dots + \mathbf{x}_d\theta_d = \sum_{k=1}^{d} \mathbf{x}_k \theta_k$

EXAMPLE: LINEAR REGRESSION

- Exact CD update in direction j:

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_j} = \theta_j - \mathbf{y}^\top \mathbf{x}_j$$

- By solving $\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_j} = 0$, we get

$$\theta_j^* = \mathbf{y}^\top \mathbf{x}_j$$

- **Repeat** this update for all θ_j

SOFT THRESHOLDING

Minimize LM with L2-loss and L1 regularization via CD:

$$\min_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1$$

Note that $h(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^\top\mathbf{y} + \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta} - \sum_{k=1}^{d}(\mathbf{y}^\top\mathbf{x}_k\theta_k + \lambda|\theta_k|)$

Assume (again): $\mathbf{X}^\top\mathbf{X} = I_d$.
Since $|\cdot|$ is not differentiable, distinguish three cases:

- **Case 1:** $\theta_j > 0$. Then $|\theta_j| = \theta_j$ and

$$0 = \frac{\partial h(\boldsymbol{\theta})}{\partial \theta_j} = \theta_j - \mathbf{y}^\top\mathbf{x}_j + \lambda \qquad \Leftrightarrow \qquad \theta^*_{j,\text{LASSO}} = \theta^*_j - \lambda$$

- **Case 2:** $\theta_j < 0$. Then $|\theta_j| = -\theta_j$ and

$$0 = \frac{\partial h(\boldsymbol{\theta})}{\partial \theta_j} = \theta_j - \mathbf{y}^\top\mathbf{x}_j - \lambda \qquad \Leftrightarrow \qquad \theta^*_{j,\text{LASSO}} = \theta^*_j + \lambda$$

- **Case 3:** $\theta_j = 0$

SOFT THRESHOLDING

We can write the solution as:

$$\theta^*_{j,\text{LASSO}} = \begin{cases} \theta^*_j - \lambda & \text{if } \theta^*_j > \lambda \\ \theta^*_j + \lambda & \text{if } \theta^*_j < -\lambda \\ 0 & \text{if } \theta^*_j \in [-\lambda, \lambda], \end{cases}$$

This operation is called **soft thresholding**.

Coefficients for which the solution to the unregularized problem are smaller than a threshold, $|\boldsymbol{\theta}^*_j| < \lambda$, are shrinked to zero.

Note: Derivation of soft thresholding operator not trivial (subgradients)

CD FOR STATISTICS AND ML

Why is it being used?

- Easy to implement
- Scalable: no storage/operations on large objects, just current point
  $\Rightarrow$ Good implementation can achieve state-of-the-art performance
- Applicable for non-differentiable (but convex separable) objectives

Examples:

- Lasso regression, Lasso GLM, graphical Lasso
- Support Vector Machines
- Regression with non-convex penalties