Problems where we can apply Coordinate Descent and Gauss-Southwell

- For certain convex problems, coordinate descent is faster than gradient descent.
- Requires that coordinate update is n times faster than gradient calculation.
- This is true in linear-prediction and graph-based problems:
  \[ h_i(x) = f(Ax) + \sum_{j \in V} g_j(x), \quad \text{or} \quad h_i(x) = \sum_{j \in V} g_j(x) + \sum_{j \in G} g_j(x), \]

where \( f \) is smooth and cheap, \( h \) includes least squares, logistic regression, lasso, and SVMs [Hsieh et al., 2008].

- \( h_i \) includes graph-based label propagation [Bengio et al., 2006] and graphical models.
- \( h_i \) is also used in graph-based label propagation [Bengio et al., 2006] and graphical models.
- \( h_i \) is used in Mark Schmidt (UBC) and Michael Friedlander (UC Davis).
- \( h_i \) is used in Relaxed Strong Convexity [Luo & Tseng, 1993].

Gauss-Southwell chooses the coordinate with largest directional derivative,

\[ \arg \max_i \| \nabla f(x^i) \|^2. \]

Assumptions, Algorithm, and Basic Bounds

- We consider the convex optimization problem

\[ \min_{x \in \mathbb{R}^n} f(x), \]

where \( f \) is coordinate-wise \( L \)-Lipschitz continuous

\[ \| \nabla f(x) + \nabla f(y) \| \leq L \| x - y \|, \quad \forall x, y \in \mathbb{R}^n \quad \text{and} \quad \alpha \in \mathbb{R}. \]

- We consider coordinate descent with a constant step-size,

\[ x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k). \]

Gauss-Southwell chooses the coordinate with largest directional derivative,

\[ i_k = \arg \max_i \| \nabla f(x^i) \|^2. \]

- Under any rule, we have a following bound on progress,

\[ f(x^{k+1}) \leq f(x^k) + \frac{1}{2L} \| x^{k+1} - x^k \|^2 \]

which gives

\[ f(x^* \geq f(x) - \frac{1}{2L} \| \nabla f(x) \|^2. \]

Convergence Analysis of Randomized Coordinate Descent

- If we choose \( i_k \) with uniform sampling, taking expectation of (1) gives

\[ \mathbb{E}[f(x^{k+1})] \leq f(x^k) - \frac{1}{2L} \mathbb{E}[\| \nabla f(x) \|^2]. \]

Using (2) and subtracting \( f(x^*) - f(x) \) from both sides we get

\[ \mathbb{E}[f(x^{k+1})] - f(x^*) \leq (1 - \frac{1}{2L}) \mathbb{E}[f(x^k) - f(x^*)]. \]

Existing Convergence Analysis of Gauss-Southwell

- If we choose \( i_k \) using Gauss-Southwell, using \( \| \nabla f(x) \|^2 = \mathbb{E}[\| \nabla f(x) \|^2], \) in (1) we have

\[ f(x^{k+1}) \leq f(x^k) - \frac{1}{2L} \| \nabla f(x^k) \|^2. \]

Now use that

\[ \| \nabla f(x^k) \|^2 \geq \frac{1}{\mu} \| \nabla f(x^k) \|^2, \]

which together with (2) implies the same rate,

\[ f(x^{k+1}) - f(x^*) \leq (1 - \frac{1}{2L}) \mathbb{E}[f(x^k) - f(x^*)]. \]

Overview: Fixing and improving the rate of Gauss-Southwell

- Recent explosion of interest in randomized coordinate descent.
- Nesterov [2010] shows random selection has same rate as (greedy) Gauss-Southwell rule.
- Disagrees with empirical behaviour: if cost is similar, Gauss-Southwell converges faster.
- We give a simple analysis of Gauss-Southwell, showing it can be much faster.
- Under extra assumptions, we give a faster refined Gauss-Southwell rule.

Example of Worst and Best Case

- For a function with constant diagonal Hessian, \( \nabla^2 f(x) = \text{diag}(\lambda_1, \ldots, \lambda_n) \), we have

\[ \mu = \min_i \lambda_i, \quad \text{and} \quad \mu_1 = \prod_{i=1}^n \lambda_i. \]

We can obtain the worst case is if all \( \lambda_i \) are equal,

\[ \mu = \alpha \quad \text{and} \quad \mu_1 = \alpha^n. \]

here all coordinate change the same rate and Gauss-Southwell has no advantage.

- We can obtain the best case if \( \lambda_1 = \beta \) and all other \( \lambda_i > \beta \),

\[ \mu = \beta, \quad \text{and} \quad \mu_1 = \frac{\beta^n \alpha - \beta}{\alpha - (n - 1) \beta}. \]

Comparison to Lipschitz Sampling

- If we have a different Lipschitz constant \( L_i \) with respect to each coordinate we can use

\[ x^{k+1} = x^k - \frac{1}{L_i} \nabla f(x^i). \]

- If we sample proportional to \( L_i \), we get [Nesterov, 2010],

\[ \mathbb{E}[f(x^{k+1})] - f(x^*) \leq (1 - \frac{1}{\mu \sum_i L_i}) \mathbb{E}[f(x^k) - f(x^*)], \]

which is faster than the rate (3) for uniform sampling (for one \( L_i \) different).

Example: Gauss-Southwell

- We can get a faster rate by using a modified Gauss-Southwell rule,

\[ i_k = \arg \max_i \frac{\| \nabla f(x^i) \|^2}{L_i}. \]

which gives a rate of

\[ f(x^{k+1}) - f(x^*) \leq (1 - \mu) \mathbb{E}[f(x^k) - f(x^*)], \]

where \( \mu \) is the strong-convexity constant with respect to \( \| x \|^2 = \sum_{i=1}^n \| \nabla x \|^2. \)

Improved Gauss-Southwell Rule

- This is the same or faster than both Gauss-Southwell and Lipschitz sampling.