Coordinate descent converges faster with the Gauss-Southwell rule than random selection

Julie Nutini (UBC), Mark Schmidt (UBC), Issam Laradji (UBC), Michael Friedlander (UC Davis) and Hoyt Koepke (Dato)

OVERVIEW: Revisiting the Gauss-Southwell Rule

- Nesterov [2012] shows random selection has same rate as Gauss-Southwell (GS) rule.
- Empirically, if costs are similar, GS is faster.

In this work, we present:

- new analysis of GS (can be much faster than random);
- improved GS rate with exact coordinate optimization;
- faster rule: Gauss-Southwell-Lipschitz;
- analysis for approximate GS rules; and
- analysis for proximal-gradient GS rules.

Problems for Coordinate Descent and Gauss-Southwell

Coordinate descent is faster than gradient descent when coordinate update is n faster than gradient calculation. Key problem classes:

$$h_1(x) := f(Ax) + \sum_{i=1}^{n} g_i(x_i), \text{ or } h_2(x) := \sum_{i\in V} g_i(x_i) + \sum_{(i,j)\in E} f_{ij}(x_{ij}),$$

where f is smooth and cheap, f_{ij} are smooth, g_i are convex, $\{V, E\}$ is a graph, A is a matrix.

- h_1 includes least squares, logistic regression, lasso, and SVMs.
  - → Often solvable in $O(cr \log n)$ with c and r non-zeros per column/row.
  - → Or can formulate as a maximum inner-product search (MIPS).
- h_2 includes graph-based label propagation and graphical models.
  - → GS efficient if maximum degree similar to average degree.
  - → E.g., lattice-structured graphs and complete graphs.

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

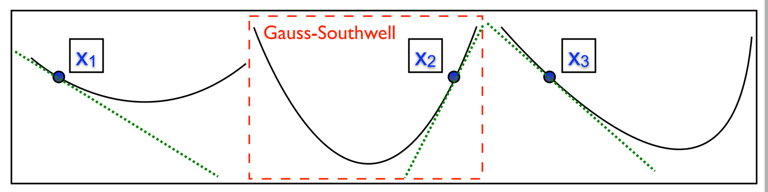
$$|\nabla_i f(x+\alpha e_i) - \nabla_i f(x)| \le L|\alpha|, \quad \forall x \in \mathbb{R}^n \text{ and } \alpha \in \mathbb{R}.$$

We consider coordinate descent with a constant step-size,

$$x^{k+1} = x^k - \frac{1}{L}\nabla_{i_k} f(x^k) e_{i_k}.$$

GS chooses the coordinate with largest directional derivative:

$$i_k = \operatorname{argmax}_i |\nabla_i f(x^k)|$$



Under any rule, we have the following upper bound on progress,

$$\begin{aligned} f(x^{k+1}) &\le f(x^k) + \nabla_{i_k} f(x^k)(x^{k+1}-x^k)_{i_k} + \frac{L}{2}(x^{k+1}-x^k)^2_{i_k} \\ &= f(x^k) - \frac{1}{L}(\nabla_{i_k} f(x^k))^2 + \frac{L}{2}\left[\frac{1}{L}\nabla_{i_k} f(x^k)\right]^2 \\ &= f(x^k) - \frac{1}{2L}[\nabla_{i_k} f(x^k)]^2. \end{aligned} \tag{1}$$

We also assume f is strongly convex with constant μ,

$$f(y) \ge f(x) + \langle \nabla f(x), y-x\rangle + \frac{\mu}{2}\|y-x\|^2, \quad \forall x, y \in \mathbb{R}^n,$$

which minimizing both sides in terms of y gives the lower bound

$$f(x^*) \ge f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2. \tag{2}$$

Convergence Analysis Randomized Coordinate Descent

Expectation of (1) when choosing i_k with uniform sampling gives

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

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

$$\mathbb{E}[f(x^{k+1})] - f(x^*) \le \left(1 - \frac{\mu}{Ln}\right)[f(x^k) - f(x^*)].$$

Classic Convergence Analysis of Gauss-Southwell

Choosing i_k using GS rule. Using $(\nabla_{i_k} f(x^k))^2 = \|\nabla f(x^k)\|^2_\infty$ in (1) we have

$$f(x^{k+1}) \le f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2_\infty. \tag{3}$$

Now use that

$$\|\nabla f(x^k)\|^2_\infty \ge \frac{1}{n}\|\nabla f(x^k)\|^2, \tag{4}$$

which together with (2) implies the same rate as random,

$$f(x^{k+1}) - f(x^*) \le \left(1 - \frac{\mu}{Ln}\right)[f(x^k) - f(x^*)].$$

Refined Convergence Analysis of Gauss-Southwell

Avoid using (4) by measuring strong-convexity in ℓ_1-norm, i.e.,

$$f(y) \ge f(x) + \langle \nabla f(x), y-x\rangle + \frac{\mu_1}{2}\|y-x\|_1^2.$$

Minimizing both sides with respect to y we get

$$\begin{aligned} f(x^*) &\ge f(x) - \sup_y\{\langle -\nabla f(x), y-x\rangle - \frac{\mu_1}{2}\|y-x\|_1^2\} \\ &= f(x) - \left(\frac{\mu_1}{2}\|\cdot\|_1^2\right)^*(-\nabla f(x)) \\ &= f(x) - \frac{1}{2\mu_1}\|\nabla f(x)\|^2_\infty. \end{aligned}$$

Combining this with (3),

$$f(x^{k+1}) - f(x^*) \le \left(1 - \frac{\mu_1}{L}\right)[f(x^k) - f(x^*)]. \tag{5}$$

Using norm inequalities we can show that

$$\frac{\mu}{n} \le \mu_1 \le \mu.$$

Separable Quadratic: μ vs. μ_1

Consider a quadratic f with diagonal Hessian:

$$\mu = \min_i \lambda_i, \quad \text{and} \quad \mu_1 = \left(\sum_{i=1}^{n}\frac{1}{\lambda_i}\right)^{-1}.$$

Constant μ_1 is the harmonic mean of λ_i divided by n:

- All λ_i equal: GS and random have same rates.
- One large λ_i: GS only slightly faster than random.
- One small λ_i: GS almost n times faster than random.

'Time need when working together' is μ_1 (dominated by smallest).

Gauss-Southwell with Different Lipschitz Constants

With a different Lipschitz constant L_i for each coordinate, we have

$$x^{k+1} = x^k - \frac{1}{L_{i_k}}\nabla_{i_k} f(x^k) e_{i_k}.$$

This gives a rate of

$$\mathbb{E}[f(x^k)] - f(x^*) \le \left[\prod_{j=1}^{k}\left(1 - \frac{\mu_1}{L_{i_j}}\right)\right][f(x^0) - f(x^*)].$$

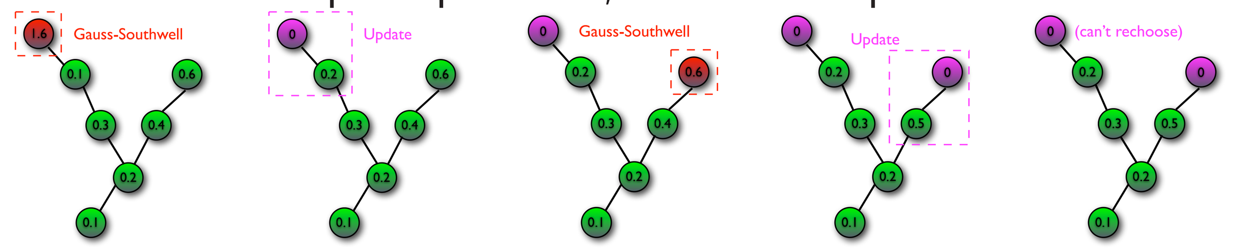
- As $L = \max_i L_i$, this is faster if $L_{i_k} < L$ for any i_k.

Gauss-Southwell with Exact Coordinate Optimization

Rates for randomized and GS still hold with exact optimization as

$$f(x^{k+1}) = \min_\alpha\{f(x^k - \alpha\nabla_{i_k} f(x^k) e_{i_k}\} \le f(x^k) - \frac{1}{2L_{i_k}}[\nabla_{i_k} f(x^k)]^2.$$

Faster rates for sparse problems, since exact update restricts order:



GS with exact optimization under a chain-structured graph has rate

$$f(x^k) - f(x^*) \le O\left(\max\{\rho_2^G, \rho_3^G\}^k\right)[f(x^0) - f(x^*)],$$

- ρ_2^G maximizes $\sqrt{(1-\mu_1/L_i)(1-\mu_1/L_j)}$ among neighbours;
- ρ_3^G maximizes $\sqrt{(1-\mu_1/L_i)(1-\mu_1/L_j)(1-\mu_1/L_k)}$, when i is neighbour of j and j is neighbour of k.

This is much faster if the large L_i are not neighbours.

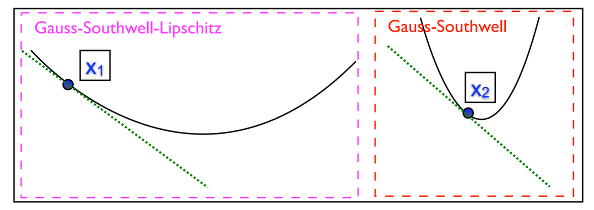
Rules Depending on Lipschitz Constants and GSL Rule

Nesterov showed that sampling proportional to L_i yields:

$$\mathbb{E}[f(x^{k+1})] - f(x^*) \le \left(1 - \frac{\mu}{n\bar{L}}\right)[f(x^k) - f(x^*)].$$

We propose a Gauss-Southwell-Lipschitz (GSL) rule using the L_i:

$$i_k = \operatorname{argmax}_i \frac{|\nabla_i f(x^k)|}{\sqrt{L_i}}$$



For this rule we have

$$f(x^{k+1}) - f(x^*) \le (1 - \mu_L)[f(x^k) - f(x^*)],$$

where strong-convexity constant μ_L for $\|x\|_L = \sum_{i=1}^n \sqrt{L_i}|x_i|$ has

$$\max\left\{\frac{\mu}{n\bar{L}}, \frac{\mu_1}{L}\right\} \le \mu_L \le \frac{\mu_1}{\min_i\{L_i\}}.$$

This also yields a tighter bound on 'maximum improvement' rule.

Gauss-Southwell-Lipschitz as Nearest Neighbour

If h_1 has no g_i functions, GS rule has the form: $\operatorname{argmax}_i |a_i^T r(x^k)|$. Dhillon et al. [2011] approximate GS as nearest neighbour,

$$\operatorname{argmin}_i \|r(x^k) - a_i\| = \operatorname{argmin}_i \left\{|\nabla_i f(x^k)| - \frac{1}{2}\|a_i\|^2\right\}.$$

When $L_i = \gamma\|a_i\|^2$, exact GSL is a nearest neighbour problem,

$$\operatorname{argmin}_i \left\|r(x^k) - \frac{a_i}{\|a_i\|}\right\| = \operatorname{argmin}_i \left\{\frac{|\nabla_i f(x^k)|}{\sqrt{L_i}}\right\}.$$

Approximate Gauss-Southwell

- For multiplicative error $|\nabla_{i_k} f(x^k)| \ge \|\nabla f(x^k)\|_\infty(1-\epsilon_k)$,

$$f(x^{k+1}) - f(x^*) \le \left[\prod_{i=1}^{k}\left(1 - \frac{\mu_1(1-\epsilon_k)^2}{L}\right)\right][f(x^0) - f(x^*)],$$

  and we do not need $\epsilon_k \to 0$.
- For additive error $|\nabla_{i_k} f(x^k)| \ge \|\nabla f(x^k)\|_\infty - \epsilon_k$,

$$f(x^{k+1}) - f(x^*) \le \left(1 - \frac{\mu_1}{L}\right)^k [f(x^0) - f(x^*) + A_k],$$

  where A_k depends on ϵ_k, and rate depends on how fast

Proximal Gauss-Southwell

An important application of coordinate descent is for problems

$$\min_{x\in\mathbb{R}^n} F(x) \equiv f(x) + \sum_i g_i(x_i),$$

where f is smooth, but g_i may be non-smooth.
Examples include bound-constraints and ℓ_1-regularization.
We can use a proximal-gradient style update,

$$x^{k+1} = \mathbf{prox}_{\frac{1}{L}g_{i_k}}\left[x^k - \frac{1}{L}\nabla_{i_k} f(x^k) e_{i_k}\right],$$

where

$$\mathbf{prox}_{\alpha g}[y] = \operatorname{argmin}_{x\in\mathbb{R}^n} \frac{1}{2}\|x-y\|^2 + \alpha g(x).$$

Three Proximal Generalizations of the GS Rule

- GS-s: Minimize directional derivative,

$$i_k = \operatorname{argmax}_i \left\{\min_{s\in\partial g_i} |\nabla_i f(x^k) + s|\right\}$$

→ Commonly-used for ℓ_1-regularization, $\|x^{k+1} - x^k\|$ could be tiny.

- GS-r: Maximize how far we move,

$$i_k = \operatorname{argmax}_i \left\{\left|x_i^k - \mathbf{prox}_{\frac{1}{L}g_{i_k}}\left[x_i^k - \frac{1}{L}\nabla_{i_k} f(x^k)\right]\right|\right\}$$

→ Effective for bound constraints, but ignores $g_i(x_i^{k+1}) - g_i(x_i^k)$.

- GS-q: Maximize progress under quadratic approximation of f.

$$i_k = \operatorname{argmin}_i \left\{\min_d f(x^k) + \nabla_i f(x^k) d + \frac{L}{2}d^2 + g_i(x_i^k + d) - g_i(x_i^k)\right\}$$

→ Least intuitive, but has the best theoretical properties.
→ Generalizes GSL if you use L_i instead of L (not true of GS-r).

Proximal GS-q Convergence Rate

Richtárik and Takáč [2014] show for randomized i_k selection that

$$\mathbb{E}[F(x^{k+1})] - F(x^*) \le \left(1 - \frac{\mu}{Ln}\right)[F(x^k) - F(x^*)].$$

For the GS-q rule, we show a rate of

$$F(x^{k+1}) - F(x^*) \le \min\left\{\left(1 - \frac{\mu}{Ln}\right)[F(x^k) - F(x^*)], \left(1 - \frac{\mu_1}{L}\right)[F(x^k) - F(x^*)] + \epsilon_k\right\},$$

where $\epsilon_k \to 0$ measures non-linearity of g_i that are not updated.

Experiments for Instances of Problem h_1