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 not faster than gradient calculation. Key problem classes:
- $h(x) = f(Ax) = \sum_i g_i(x_i)$, or $h(x) = \sum_{i=1}^n f_i(x_i) + \sum_{i\neq j} f_{ij}(x_i, x_j)$, where $f$ is smooth and cheap, $g_i$ are smooth, $g$ is convex, $(V, E)$ is a graph, $A$ is a matrix.
- $h_i$ includes least squares, logistic regression, lasso, and SVMs.
- Often scalable in $O(r\log n)$ with $r$ and $n$ non-zero per column/row.
- Or can formulate as a maximum inner-product search (MIPS).
- $h_i$ includes graph-based labelling propagation and graphical models.
- GS efficient if degree much larger than 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

$$\nabla f(x + \alpha v) \leq \nabla f(x) + L\alpha v$$

for all $x, v \in \mathbb{R}^n$. $\mathcal{V} \subseteq \mathbb{R}^n$ and $\alpha \in \mathbb{R}$.

We consider coordinate descent with a constant step-size,

$$x^{k+1} = x^k - \frac{\mu}{L} \nabla f_i(x^k),$$

GS chooses the coordinate with largest directional derivative:

$$i = \arg\max_{i \in \mathcal{V}} |\nabla f_i(x)|.$$  

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

$$f(x^{k+1}) - f(x^k) \leq \nabla f_i(x^k)^T (x^k - x^{k+1}) = \frac{\mu}{L} \nabla f_i(x^k) \nabla f_i(x^k) = \frac{\mu}{L} \nabla f_i(x^k)^2.$$  

Gauss-Southwell with Exact Coordinate Optimization

Rules for randomized and GS still hold with exact optimization as

$$f(x^{k+1}) = \arg\min_{x} (f(x) + \frac{\mu}{L} \nabla f_i(x) \nabla f_i(x)),$$

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

$$f(x^k) - f(x^{k+1}) \leq O\left(\sqrt{\frac{\mu}{L} \|\nabla f_i(x)\|^2}\right).$$

Gauss-Southwell with Exact Coordinate Optimization

Nesterov showed that sampling proportional to $L$ yields:

$$f(x^{k+1}) - f(x^k) \leq \left(1 - \frac{\mu}{L}\right) f(x^k) - f(x^{k+1}).$$

For this rule we have

$$f(x^{k+1}) - f(x^k) \leq \left(1 - \frac{\mu}{L}\right) f(x^k) - f(x^{k+1}),$$

where strong convexity constant $\mu$ for $f_i$ on $\mathbb{R}^n$ has

$$\left(\frac{\mu}{L}\right) \leq 2 \sum_{i \in \mathcal{V}} \nabla f_i(x) \nabla f_i(x) \leq \mu \leq 2 \sum_{i \in \mathcal{V}} \nabla f_i(x) \nabla f_i(x).$$

This also yields a tighter bound on ‘maximum improvement’ rule.

Gauss-Southwell-Lipschitz as Nearest Neighbour

If $h_i$ has no $g_i$ functions, GS rule has the form: argmax $|\nabla f_i(x)\|$. Dhillon et al. [2011] approximate GS as nearest neighbour,

$$\arg\min_{x \in \mathcal{V}} \|x - a\|_2^2.$$  

When $L = \gamma |\nabla f_i(x)|$, exact GS is a nearest neighbour problem,

$$\arg\min_{x \in \mathcal{V}} \|x - a\|_2^2.$$  

Approximate Gauss-Southwell with Different Lipschitz Constants

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

$$x^{k+1} = x^k - \frac{\mu_i}{L_i} \nabla f_i(x^k).$$

This gives a rate of

$$\|f(x^k) - f(x^{k+1})\| \leq \prod_{i \in \mathcal{V}} \left(1 - \frac{\mu_i}{L_i}\right) f(x^k) - f(x^{k+1}).$$

As $L = \max_{i \in \mathcal{V}} L_i$, this is faster if $L_i < L$ for any $i$.

Proximal Gauss-Southwell

An important application of coordinate descent is for problems

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

where $f$ is smooth, but $g_i$ may be non-smooth.

Examples include bound-constraints and $\ell_1$-regularization.

We can use a proximal-gradient style update,

$$x^{k+1} = \prox_{\lambda f_i} (x^k - \frac{\mu_i}{L_i} \nabla f_i(x^k)),$$

where

$$\prox_{\lambda f_i} (y) = \arg\min_{z \in \mathbb{R}^n} \frac{1}{2} \|z - y\|^2 + \lambda f_i(z).$$

Proximal Gauss-Southwell

Typically used for $\ell_1$-regularization, $\|x^{k+1} - x^k\|$ could be tiny.

- GS: Maximize how far we move,

$$\prox_{\lambda f_i} (y) = \arg\min_{z \in \mathbb{R}^n} \frac{1}{2} \|z - y\|^2 + \lambda f_i(z)$$

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

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

Three Proximal Generalizations of the GS Rule

- GS-S: Minimize directional derivative,

$$i = \arg\max_{i \in \mathcal{V}} \|\nabla f_i(x^k + \mu_i \nabla f_i(x^k))\|.$$

- GS-S: Maximize how far we move,

$$\prox_{\lambda f_i} (x^k - \frac{\mu_i}{L_i} \nabla f_i(x^k))$$

- Least intuitive, but has the best theoretical properties.

- Generalizes GS if $x^k$ is $L$ instead of $L_i$ (not true of GS-r).

Proximal GS-$\eta$ Convergence Rate

Richtárik and Takáč [2014] show for randomized $i$ selection that

$$\|f(x^{k+1}) - f(x^k)\| \leq \frac{1}{\eta} \|f(x^k) - f^\star\|.$$  

For the GS-$\eta$ rule, we show a rate of

$$\|f(x^{k+1}) - f(x^k)\| \leq \min \left\{ \frac{1}{\eta} \|f(x^k) - f^\star\|, \right\}$$  

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

Experiments for Instances of Problem $h_1$