Coordinate Minimization

Daniel P. Robinson
Department of Applied Mathematics and Statistics
Johns Hopkins University

November 12, 2020

Outline

1. Introduction
2. Algorithms
   - Cyclic order with exact minimization
   - Cyclic order with fixed step size
   - Steepest direction (Gauss-Southwell rule) with fixed step size
   - Alternatives
   - Summary

Given a function $f : \mathbb{R}^n \to \mathbb{R}$, consider the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \quad (1)$$

- We will consider various assumptions on $f$:
  - nonconvex and differentiable $f$
  - convex and differentiable $f$
  - strongly convex and differentiable $f$
- We will not consider general non-smooth $f$, because we can not prove anything.
- We will briefly consider structured non-smooth problems, i.e., problems that use an additional (separable) regularizer.

**Notation:** $f_k := f(x_k)$ and $g_k := \nabla f(x_k)$

**Basic idea (coordinate minimization):** Compute the next iterative using the update

$$x_{k+1} = x_k - \alpha_k e_{i(k)}$$

**Algorithm 1** General coordinate minimization framework.

1. Choose $x_0 \in \mathbb{R}^n$ and set $k \leftarrow 0$.
2. **loop**
   3. Choose $i(k) \in \{1, 2, \ldots, n\}$.
   4. Choose $\alpha_k > 0$
   5. Set $x_{k+1} \leftarrow x_k - \alpha_k e_{i(k)}$
   6. Set $k \leftarrow k + 1$.
   7. **end loop**

- $\alpha_k$ is the step size. Options include:
  - fixed, but sufficiently small
  - inexact linesearch
  - exact linesearch
- $i(k) \in \{1, 2, \ldots, n\}$ has to be chosen. Options include:
  - cycle through the entire set
  - choose it uniformly at random
  - choose it based on which element of $\nabla f(x_k)$ is the largest in absolute value
- $e_{i(k)}$ is the $i(k)$-th coordinate vector
- this update seeks better points in $\text{span}\{e_{i(k)}\}$.
Algorithm 2 Coordinate minimization with cyclic order and exact minimization.

1. Choose \( x_0 \in \mathbb{R}^n \) and set \( k \leftarrow 0 \).
2. \textbf{loop}
3. Choose \( i(k) = \text{mod}(k, n) + 1 \).
4. Calculate the exact coordinate minimizer:
   \[
   \alpha_k \leftarrow \arg\min_{\alpha \in \mathbb{R}} f(x_k - \alpha e_{i(k)})
   \]
5. Set \( x_{k+1} \leftarrow x_k - \alpha_k e_{i(k)} \).
6. Set \( k \leftarrow k + 1 \).
7. \textbf{end loop}

Comments:
- This algorithm assumes that the exact minimizers exist and that they are unique.
- A reasonable stopping condition should be incorporated, such as
  \[
  \|\nabla f(x_k)\|_2 \leq 10^{-6} \max\{1, \|\nabla f(x_0)\|_2\}
  \]

Theorem 2.1 (see [6, Theorem 5.32])

Assume that the following hold:
- \( f \) is continuously differentiable;
- the level set \( \mathcal{L}_0 := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\} \) is bounded; and
- for every \( x \in \mathcal{L}_0 \) and all \( j \in \{1, 2, \ldots, n\} \), the optimization problem
  \[
  \min_{\zeta \in \mathbb{R}} f(x + \zeta e_j)
  \]
  has a unique minimizer.

Then, for any limit point \( x^* \) of the sequence \( \{x_k\} \) generated by Algorithm 2 satisfies
\[
\nabla f(x^*) = 0.
\]

Proof: Since \( f(x_{k+1}) \leq f(x_k) \) for all \( k \in \mathbb{N} \), we know that the sequence \( \{x_k\}_{k=0}^\infty \subseteq \mathcal{L}_0 \).
Since \( \mathcal{L}_0 \) is bounded, \( \{x_k\}_{k=0}^\infty \) has at least one limit point; let \( x^* \) be any such limit point. Thus, there exists a subsequence \( \mathcal{K} \subseteq \mathbb{N} \) satisfying
\[
\lim_{k \to \infty} x_k = x^*.
\]
Combining this with monotonicity of \( f(x_k) \) and continuity of \( f \) also shows that
\[
\lim_{k \to \infty} f(x_k) = f(x^*) \quad \text{and} \quad f(x_k) \geq f(x^*) \quad \text{for all} \quad k \in \mathbb{N}.
\]

An interesting example introduced by Powell [5, formula 2] is
\[
\min f(x_1, x_2, x_3) := -(x_1 x_2 + x_2 x_3 + x_1 x_3) + \sum_{i=1}^3 (|x_i| - 1)^2
\]
- \( f \) is continuously differentiable and nonconvex
- \( f \) has minimizers at \((-1, -1, -1)\) and \((1, 1, 1)\) of the unit cube.
- Coordinate descent with exact minimization started just outside the unit cube near any nonoptimal vertex cycles around neighborhoods of all 6 non-optimal vertices.
- Powell shows that the cyclic nonconvergence behavior is special and is destroyed by small perturbations on this particular example.

We assume \( \nabla f(x^*) \neq 0 \), and then reach a contradiction.

First, consider the subsets \( \mathcal{K}_i \subseteq \mathcal{K}, i = 0, \ldots, n-1 \) defined as
\[
\mathcal{K}_i := \{k \in \mathcal{K} : k \equiv i \mod n\}.
\]
Since \( \mathcal{K} \) is an infinite subsequence of the natural numbers, one of the \( \mathcal{K}_i \) must be an infinite set. Without loss of generality, we assume it is \( \mathcal{K}_0 \) (the argument is very similar for any other \( i \), because we are using cyclic order).

Let us perform a hypothetical "sweep" of coordinate minimization starting from \( x^* \), so that we would obtain
\[
y^* := x^*, \quad \text{with} \quad x^*_\ell := x^* + \sum_{j=1}^\ell [r^*] e_j \quad \text{for all} \quad \ell = 1, \ldots, n
\]
and note that since \( \nabla f(x^*) \neq 0 \) by assumption, we must have
\[
f(r^*) < f(x^*). \quad \text{(why?)}
\]

NOTE: If \( \mathcal{K}_i \) was infinite for some \( i \neq 0 \), then we would do above "sweep" at \( x^* \) starting with coordinate \( i \) and going in cyclic order to cover all \( n \) coordinates.
Next, notice that by construction of the coordinate minimization scheme, that
\[
x_{k+\ell} = x_k + \sum_{j=1}^{\ell} \tau_{j} e_j \quad \text{for all } k \in \mathcal{K}_0 \text{ and } 1 \leq \ell \leq n, \tag{5}
\]
where \(\tau_i\) is the step length at iteration \(i\). Therefore,
\[
\|x_{k+\ell} - x_k\| = \left\| \begin{pmatrix} \tau_1 \\ \tau_{n+1} \\ \vdots \\ \tau_{n+\ell-1} \end{pmatrix} \right\| \leq 2 \max\{\|x\|: x \in \mathcal{L}_0\} < \infty
\]
for all \(k \in \mathcal{K}_0\) and \(1 \leq \ell \leq n\). We used the assumption that \(\mathcal{L}_0\) is bounded. Since this shows that the set \(\{(\tau_1, \tau_{n+1}, \ldots, \tau_{n+\ell-1})\}\) is bounded, we may pass to a subsequence \(\mathcal{K}' \subseteq \mathcal{K}_0\) with
\[
\lim_{k \in \mathcal{K}'} \tau_{n+\ell-1} = \tau^\ell \quad \text{for some } \tau^\ell \in \mathbb{R}^n. \tag{6}
\]
Taking the limit of (5) over \(k \in \mathcal{K}' \subseteq \mathcal{K}_0\) for each \(\ell\), and using (2) and (6) we find that
\[
\lim_{k \in \mathcal{K}'} x_{k+\ell} = x^* + \sum_{j=1}^{\ell} [\tau^\ell] e_j \quad \text{for each } 1 \leq \ell \leq n. \tag{7}
\]

**Induction step:** assume that (8) and (9) hold for \(1 \leq p \leq \bar{p} \leq n - 1\).

We know from the coordinate minimization that
\[
f(x_{k+p+1}) \leq f(x_{k+p} + \rho e_p) \quad \text{for all } k \in \mathbb{N} \text{ and } \rho \in \mathbb{R}.
\]
Taking the limit over \(k \in \mathcal{K}'\), continuity of \(f\), (7) with \(\ell = \bar{p} + 1\), and (9) give
\[
f(x^* + \sum_{j=1}^{\bar{p}+1} [\tau^\ell] e_j) = f(\lim_{k \in \mathcal{K}'} x_{k+p+1}) = \lim_{k \in \mathcal{K}'} f(x_{k+p+1}) \leq \lim_{k \in \mathcal{K}'} f(x_{k+p} + \rho e_p) = f(x^* + \rho e_p) \quad \text{for all } \rho \in \mathbb{R}.
\]
Thus, the definition of \(x_{p+1}^*\), and the fact that (8) holds for all \(1 \leq p \leq \bar{p}\) show that
\[
f(x_{\bar{p}+1}^* + [\tau^\ell] e_{\bar{p}+1}) = f(x^* + \sum_{j=1}^{\bar{p}} [\tau^\ell] e_j + [\tau^\ell] e_{\bar{p}+1})
\]
\[
= f(x^* + \sum_{j=1}^{\bar{p}} [\tau^\ell] e_j + [\tau^\ell] e_{\bar{p}+1})
\]
\[
= f(x^* + \sum_{j=1}^{\bar{p}} [\tau^\ell] e_j + [\tau^\ell] e_{\bar{p}+1}) \quad \text{for all } \tau \in \mathbb{R}.
\]
Uniqueness of the minimizer implies \([\tau^\ell] e_{\bar{p}+1} = [\tau^*] e_{\bar{p}+1}\) and combining with (7) gives
\[
\lim_{k \in \mathcal{K}'} x_{k+p+1} = x^* + \sum_{j=1}^{\bar{p}+1} [\tau^*] e_j = x^* + \sum_{j=1}^{\bar{p}+1} [\tau^*] e_j \equiv x_{\bar{p}+1}^*,
\]
which completes the proof by induction.

We next claim the following, which we will prove by induction:
\[
[\tau^*]_p = [\tau^*]_p \quad \text{for all } 1 \leq p \leq n, \tag{8}
\]
\[
\lim_{k \in \mathcal{K}'} x_{k+p} = x_p^* \quad \text{for all } 1 \leq p \leq n. \tag{9}
\]

**Base case:** \(p = 1\).

We know from the coordinate minimization that
\[
f(x_{k+1}) \leq f(x_k + \tau e_1) \quad \text{for all } k \in \mathbb{N} \text{ and } \tau \in \mathbb{R}.
\]
Taking limits over \(k \in \mathcal{K}' \subseteq \mathcal{K}\) and using continuity of \(f\), (7) with \(\ell = 1\), and (2) yields
\[
f(x^* + [\tau^*] e_1) = f(\lim_{k \in \mathcal{K}'} x_{k+1}) = \lim_{k \in \mathcal{K}'} f(x_{k+1}) \leq \lim_{k \in \mathcal{K}'} f(x_k + \tau e_1)
\]
\[
= f(\lim_{k \in \mathcal{K}'} x_k + \tau e_1) = f(x^* + \tau e_1) \quad \text{for all } \tau \in \mathbb{R}.
\]
Since the minimizations in coordinate directions are unique by assumption, we know that \([\tau^*]_1 = [\tau^*]_1\), which is the first desired result. Also, combining it with (7) gives
\[
\lim_{k \in \mathcal{K}'} x_{k+1} = x^* + [\tau^*] e_1 = x^* + [\tau^*] e_1 \equiv x_1^*,
\]
which completes the base base.
Notation:
- Let $L_i$ denote the $j$th component Lipschitz constant, i.e., it satisfies $|\nabla_i f(x + te_j) - \nabla_i f(x)| \leq L_i |t|$ for all $x \in \mathbb{R}^n$ and $t$.
- Let $L_{\max}$ denote the coordinate Lipschitz constant, i.e., it satisfies $L_{\max} := \max_{1 \leq i \leq \dim x} L_i$.
- Let $L$ denote the Lipschitz constant for $\nabla f$.

Algorithm 3 Coordinate minimization with cyclic order and a fixed step size.

1. Choose $\alpha \in (0, 1/L_{\max}]$.
2. Choose $x_0 \in \mathbb{R}^n$ and set $k \leftarrow 0$.
3. loop
   4. Choose $i(k) = \text{mod}(k, n) + 1$.
   5. Set $x_{k+1} \leftarrow x_k - \alpha \nabla_i f(x_k) e_{i(k)}$.
   6. Set $k \leftarrow k + 1$.
7. end loop

Comments:
- A reasonable stopping condition should be incorporated, such as $\|\nabla f(x_k)\|_2 \leq 10^{-6} \max\{1, \|\nabla f(x_0)\|_2\}$.
- A maximum number of allowed iterations should be included in practice.

Theorem 2.2 (see [1, Theorem 3.6, Theorem 3.9] and [7, Theorem 3])

Suppose that $\alpha = 1/L_{\max}$ and let the following assumptions hold:
- $\nabla f$ is globally Lipschitz continuous
- $f$ has a minimizer $x^*$ and $f^* := f(x^*) = \min_{x \in \mathbb{R}^n} f(x)$
- there exists a scalar $R_0$ such that the diameter of the level set $\{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ is bounded by $R_0$.

Then, the iterate sequence $\{x_k\}$ of Algorithm 3 satisfies

$$
\min_{k=0, \ldots, T} \|\nabla f(x_k)\| \leq \frac{4nL_{\max}(1 + nL^2/L_{\max}^2)(f(x_0) - f^*)}{T + 1} \quad \forall T \in \{n, 2n, 3n, \ldots\}
$$

(11)

If $f$ is convex, then

$$
f(x_T) - f^* \leq \frac{4nL_{\max}(1 + nL^2/L_{\max}^2)R_0^2}{T + 8} \quad \forall T \in \{n, 2n, 3n, \ldots\}.
$$

(12)

If $f$ is $\mu$-strongly convex, then

$$
f(x_T) - f^* \leq \left(1 - \frac{\mu}{2L_{\max}(1 + nL^2/L_{\max}^2)}\right)^{T/n} (f(x_0) - f^*) \quad \forall T \in \{n, 2n, 3n, \ldots\}
$$

Comments on Theorem 2.2:
- The numerator in (11) and (12) is $O(n^2)$, while the numerator in the analogous result for the random coordinate choice with fixed step size is $O(n)$ (see SGD notes and HW 4). But Theorem 2.2 is a deterministic result, while the result for random coordinate choice is in expectation.
- Recall also with the full gradient the iteration complexity has no dependence on $n$. But every iteration itself is $n$ times more expensive than cyclic or random choice.
- It can be shown that $L \leq \sum_{j=1}^n L_j$. (see [3, Lemma 2 with $\alpha = 1$])
- It follows from the fact that $|\nabla_j f(x + te_j) - \nabla_j f(x)| \leq \|\nabla f(x + te_j) - \nabla f(x)\|_2 \leq L_j |t|$ holds for all $j$, $t$, and $x$ such that $L_j \leq L$.
- By combining the previous two bullet points, we find that $L_{\max} \equiv \max_{1 \leq j \leq \dim x} L_j \leq L \leq \sum_{j=1}^n L_j \leq nL_{\max}$
  so that $1 \leq L_{\max} \leq n$.
- Roughly speaking, $L/L_{\max}$ is closer to 1 when the coordinates are "more decoupled". In light of (12), the complexity result for coordinate descent becomes better as the variables become more decoupled. This makes sense!
Notation:
- Let $L_i$ denote the $i$th component Lipschitz constant, i.e., it satisfies $|\nabla_i f(x + t e_i) - \nabla_i f(x)| \leq L_i |t|$ for all $x \in \mathbb{R}^n$ and $t$.
- Let $L_{\max}$ denote the coordinate Lipschitz constant, i.e., it satisfies $L_{\max} := \max_{1 \leq i \leq n} L_i.$

Algorithm 4 Coordinate minimization with Gauss-Southwell Rule and a fixed step size.

1. Choose $\alpha \in (0, 1/L_{\max}].$
2. Choose $x_0 \in \mathbb{R}^n$ and set $k \leftarrow 0$.
3. Loop
   4. Calculate $i(k)$ as the steepest coordinate direction, i.e.,
      $$i(k) \leftarrow \arg\max_{1 \leq i \leq n} |\nabla_i f(x_k)|$$
   5. Set $x_{k+1} \leftarrow x_k - \alpha \nabla_{i(k)} f(x_k) e_{i(k)}.$
   6. Set $k \leftarrow k + 1.$
7. End loop

Comments:
- A reasonable stopping condition should be incorporated, such as $\|\nabla f(x_k)\|_2 \leq 10^{-6} \max \{1, \|\nabla f(x_0)\|_2\}$

Proof: We recall our fundamental inequality (see the "Stochastic Gradient Descent" lecture notes on random coordinate choice)

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L_{\max}} (\nabla_{i(k)} f(x_k))^2.$$

Combining this with the choice $i(k) \leftarrow \arg\max_{1 \leq i \leq n} |\nabla_i f(x_k)|$ and the standard norm inequality $|v|_2 \leq \sqrt{n} |v|_\infty$, it holds that

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L_{\max}} |\nabla f(x_k)|_\infty^2$$

$$\leq f(x_k) - \frac{1}{2L_{\max}} |\nabla f(x_k)|_2^2.$$  \hspace{1cm} (13)

Recall equation (2.1) from the "Smooth Convex Optimization" lecture notes:

$$f(x_k) - f^* \leq R_0 \|\nabla f(x_k)\|_2.$$

Substituting the above into (14)

$$f(x_{k+1}) - f^* \leq f(x_k) - f^* - \frac{1}{2nL_{\max}} |\nabla f(x_k)|_2^2 \leq f(x_k) - f^* - \frac{1}{2nL_{\max} R_0^2} (f(x_k) - f^*)^2.$$

Using the notation $\Delta_t = f(x_t) - f^*$, this is equivalent to

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{2nL_{\max} R_0^2} \Delta_t^2.$$  \hspace{1cm} (14)

Theorem 2.3

Suppose that $\alpha = 1/L_{\max}$ and let the following assumptions hold:
- $\nabla f$ is globally Lipschitz continuous
- $f$ has a minimizer $x^*$ and $f^* := f(x^*) = \min_{x \in \mathbb{R}^n} f(x)$
- there exists a scalar $R_0$ such that the diameter of the level set $\{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ is bounded by $R_0$.

Then, the iterate sequence $\{x_k\}$ computed from Algorithm 4 satisfies

$$\min_{k=0, \ldots, T} \|\nabla f(x_k)\| \leq \sqrt{\frac{2n L_{\max} (f(x_0) - f^*)}{T + 1}} \quad \forall T \geq 1.$$

If $f$ is convex, then

$$f(x_T) - f^* \leq \frac{2n L_{\max} R_0^2}{T} \quad \forall T \geq 1.$$

If $f$ is $\mu$-strongly convex, then

$$f(x_T) - f^* \leq \left(1 - \frac{\mu}{n L_{\max}}\right) \Delta_0 (f(x_0) - f^*) \quad \forall T \geq 1.$$
Next, assume that \( f \) is \( \mu \)-strongly convex, and recall the following inequality we established in the proof of Theorem 2.3 of the "Smooth Convex Optimization" lecture notes:

\[
    f^* \geq f(x_2) - \frac{1}{2\mu} \|\nabla f(x_2)\|^2.
\]

Subtracting \( f^* \) from each side of (14) and then using the previous inequality shows that

\[
    f(x_{k+1}) - f^* \leq f(x_k) - f^* - \frac{1}{2nL_{\text{max}}} \|\nabla f(x_k)\|^2
\]

\[
    \leq f(x_k) - f^* - \frac{\mu}{nL_{\text{max}}} (f(x_k) - f^*)
\]

so that

\[
    f(x_k) - f^* \leq \left( 1 - \frac{\mu}{nL_{\text{max}}} \right)^k (f(x_0) - f^*)
\]

which is the last desired result. \( \blacksquare \)

**Theorem 2.4**

Suppose that \( \alpha = 1/L_{\text{max}} \) and let the following assumptions hold:

- \( f \) is \( \ell_1 \)-strongly convex, i.e., there exists \( \mu_1 > 0 \) such that
  \[
  f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\mu_1}{2} \|y - x\|_1 \quad \text{for all } \{x, y\} \subset \mathbb{R}^n
  \]

- \( \nabla f \) is globally Lipschitz continuous
- the minimum value of \( f \) is obtained

Then, the iterate sequence \( \{x_k\} \) computed from Algorithm 4 satisfies

\[
    f(x_k) - f^* \leq \left( 1 - \frac{\mu_1}{L_{\text{max}}} \right)^k (f(x_0) - f^*)
\]

**Comments so far for fixed step size:**

- Cyclic has the worst dependence on \( n \):
  - Cyclic: \( O(n^2) \)
  - Random and Gauss-Southwell: \( O(n) \)
  - Random is a rate in expectation.
  - Gauss-Southwell is a deterministic rate.

- But Gauss-Southwell has \( O(n) \) complexity for each iteration (similar to full gradient descent), whereas cyclic and random choice have \( O(1) \) complexity for each iteration.

- There is a better analysis for Gauss-Southwell when we assume that \( f \) is strongly convex that changes the above comment! (See [4]). We show this next.

**Proof (see [4]):** Using \( \ell_1 \)-strong convexity means that

\[
    f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\mu_1}{2} \|y - x\|_1^2 \quad \text{for all } \{x, y\} \subset \mathbb{R}^n
\]

for the \( \ell_1 \)-strong convexity parameter \( \mu_1 \). If we now minimize both sides with respect to \( y \) and replace \( x \) by \( x_k \), we find that

\[
    f^* = \min_{y \in \mathbb{R}^n} f(y)
\]

\[
    \geq \min_{y \in \mathbb{R}^n} f(x_k) + \nabla f(x_k) \cdot (y - x_k) + \frac{\mu_1}{2} \|y - x_k\|_1^2
\]

\[
    = f(x_k) + \nabla f(x_k) \cdot (y_k^* - x_k) + \frac{\mu_1}{2} \|y_k^* - x_k\|_1^2 \quad \text{(why?) exercise}
\]

\[
    = f(x_k) - \frac{1}{2\mu_1} \|\nabla f(x_k)\|_\infty^2
\]

where \( y_k^* := x_k + z_k^* \) with

\[
    [z_k]_i := \begin{cases} 0 & \text{if } i \neq \ell \\ -\frac{\nabla f(x_k)}{\mu_1} & \text{if } i = \ell \end{cases}
\]

and \( \ell \) any index satisfying

\[
    \ell \in \{ j : \|\nabla f(x_k)\| = \|\nabla f(x_k)\|_\infty \}.
\]

Therefore, we have that

\[
    \|\nabla f(x_k)\|_\infty^2 \geq 2\mu_1 (f(x_k) - f^*).
\]
From the previous slide, we showed that
$$\|\nabla f(x_i)\|_\infty^2 \geq 2\mu_1(f(x_i) - f^*) .$$
Subtracting $f^*$ from both sides of (13) and using the previous inequality shows that
$$f(x_{i+1}) - f^* \leq f(x_i) - f^* - \frac{1}{2L_{\max}}\|\nabla f(x_i)\|_\infty^2$$
$$\leq f(x_i) - f^* - \frac{\mu_1}{L_{\max}}(f(x_i) - f^*)$$
$$= \left(1 - \frac{\mu_1}{L_{\max}}\right)(f(x_i) - f^*) .$$
Applying this inequality recursively gives
$$f(x_i) - f^* \leq \left(1 - \frac{\mu_1}{L_{\max}}\right)^i(f(x_0) - f^*)$$
which is the desired result. \(\blacksquare\)

---

**Example: A Simple Diagonal Quadratic Function**

Consider the problem
$$\text{minimize}_{x \in \mathbb{R}^n} g^T x + \frac{1}{2} x^T H x$$
where
$$H = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$
with $\lambda_i > 0$ for all $i \in \{1, 2, \ldots, n\}$. For this problem, we know that
$$\mu = \min\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \text{ and } \mu_1 = \left(\sum_{i=1}^n \frac{1}{\lambda_i}\right)^{-1}$$

**Case 1:** For $\lambda_1 = \alpha$ for some $\alpha > 0$, the minimum value for $\mu_1$ occurs when $\alpha = \lambda_1 = \lambda_2 = \cdots = \lambda_n$, which gives
$$\mu = \alpha \text{ and } \mu_1 = \frac{\alpha}{n} .$$
Thus, the convergence constants are:
- (random selection): $\left(1 - \frac{\mu}{nL_{\max}}\right) = \left(1 - \frac{\alpha}{nL_{\max}}\right)$
- (Gauss-Southwell selection): $\left(1 - \frac{\mu_1}{L_{\max}}\right) = \left(1 - \frac{\alpha}{nL_{\max}}\right)$

so the convergence constants are the same; this is the worst case for Gauss-Southwell.

---

For **strongly convex functions**:
- Random coordinate choice has the expected rate of
$$\mathbb{E}[f(x_i)] - f^* \leq \left(1 - \frac{\mu}{nL_{\max}}\right)^i(f(x_0) - f^*) .$$
- Gauss-Southwell coordinate choice has the deterministic rate of
$$f(x_i) - f^* \leq \left(1 - \frac{\mu_1}{L_{\max}}\right)^i(f(x_0) - f^*) \quad (15)$$
- The bound for Gauss-Southwell is better since
$$\frac{\mu}{n} \leq \mu_1 \leq \mu$$
so that
$$\mu_1 \geq \frac{\mu}{n} \iff \frac{\mu_1}{L_{\max}} \geq \frac{\mu}{nL_{\max}} \iff \left(1 - \frac{\mu_1}{L_{\max}}\right) \leq \left(1 - \frac{\mu}{nL_{\max}}\right)$$

**Case 2:** For this other extreme case, let us suppose that
$$\lambda_1 = \beta \text{ and } \lambda_2 = \lambda_3 = \cdots = \lambda_n = \alpha$$
with $\alpha \geq \beta$. For this case, it can be shown that
$$\mu = \beta \text{ and } \mu_1 = \frac{\beta\alpha^{n-1}}{\alpha^{n-1} + (n-1)\beta\alpha^{n-2}} = \frac{\beta\alpha}{\alpha + (n-1)\beta} .$$
If we now take the limit as $\alpha \to \infty$ we find that
$$\mu = \beta \text{ and } \mu_1 \to \beta = \mu$$
Thus, the convergence constants (in the limit) are:
- (random selection): $\left(1 - \frac{\mu}{nL_{\max}}\right) = \left(1 - \frac{\beta}{nL_{\max}}\right)$
- (Gauss-Southwell selection): $\left(1 - \frac{\mu_1}{L_{\max}}\right) = \left(1 - \frac{\beta}{L_{\max}}\right)$
so that Gauss-Southwell is a factor $n$ faster than using a random coordinate selection.
Alternative 1 (strongly convex): individual coordinate Lipschitz constants.
The iteration update is
\[ x_{k+1} = x_k + \frac{1}{L_{i(k)}} \nabla_{i(k)} f(x_k) e_{i(k)} \]

- Using a similar analysis as before, it can be shown
\[ f(x_k) - f^* \leq \left[ \prod_{i=1}^k \left( 1 - \frac{\mu_1}{L_i} \right) \right] (f(x_0) - f^*) \]

- Better decrease than prior analysis since (see (15))
\[ \text{new rate} = \left[ \prod_{i=1}^k \left( 1 - \frac{\mu_1}{L_i} \right) \right] \leq \left( 1 - \frac{\mu_1}{L_{\max}} \right)^k = \text{previous rate} \]

- faster provided at least one of the used \( L_i \) satisfies \( L_i < L_{\max} \).

Choose \( i(k) \) according to the rule
\[ i(k) \leftarrow \max_{1 \leq i \leq n} \frac{(\nabla_i f(x_k))^2}{L_i} \quad (16) \]

- We recall our fundamental inequality for coordinate descent with step size
\[ \alpha_k = \frac{\mu_1}{L_{\max}} \]
\[ f(x_{k+1}) \leq f(x_k) - \frac{1}{2L_{i(k)}} (\nabla_{i(k)} f(x_k))^2 \quad (17) \]

- The update (16) is designed to choose \( i(k) \) to minimize the guaranteed decrease given by (17), which uses the component Lipschitz constants.

- It may be shown, using this update, that
\[ f(x_{k+1}) - f^* \leq (1 - \mu_k) (f(x_k) - f^*) \]

where \( \mu_k \) is the strong convexity parameter with respect to \( \|v\|_{\mu} := \sum_{i=1}^n \sqrt{L_i} |v_i| \).

- It is shown in [4, Appendix 6.2] that
\[ \max \left\{ \frac{\mu}{n L_{\max}}, \frac{\mu_1}{L_{\max}} \right\} \leq \mu_k \leq \frac{\mu_1}{\min_{1 \leq i \leq n} \{L_i\}} \]

- At least as fast as the fastest of Gauss-Southwell and Lipschitz sampling options.

Alternative 3 (strongly convex): Gauss-Southwell-Lipschitz rule.
The iteration update is
\[ x_{k+1} = x_k + \frac{1}{L_{i(k)}} \nabla_{i(k)} f(x_k) e_{i(k)} \]

- Using a similar analysis as before, it can be shown
\[ f(x_k) - f^* \leq \left[ \prod_{i=1}^k \left( 1 - \frac{\mu_1}{L_i} \right) \right] (f(x_0) - f^*) \]

- It is shown in [4, Appendix 6.2] that
\[ \max \left\{ \frac{\mu}{n L_{\max}}, \frac{\mu_1}{L_{\max}} \right\} \leq \mu_k \leq \frac{\mu_1}{\min_{1 \leq i \leq n} \{L_i\}} \]

- At least as fast as the fastest of Gauss-Southwell and Lipschitz sampling options.

Ordering of constant in linear convergence results when \( f \) is strongly convex:

- random (uniform sampling, \( L_{\max} \))
  \[ \checkmark \]
- Gauss-Southwell (\( L_{\max} \))
  \[ \checkmark \]
- Gauss-Southwell with \( \{L_i\} \)
  \[ \checkmark \]
- random (Lipschitz sampling, \( \{L_i\} \))
  \[ \checkmark \]
- Gauss-Southwell-Lipschitz \( \left( \max_{1 \leq i \leq n} \frac{(\nabla_i f(x_k))^2}{L_i} \right) \)

Comments:

- Gauss-Southwell-Lipschitz: the best rate, but is the most expensive per iteration.
- Better rates if you know and use \( \{L_i\} \) instead of just using their max, i.e., \( L_{\max} \).
References


