Coordinate Minimization

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

November 19, 2019

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

3. Examples
   - Linear equations
   - Logistic regression
Given a function $f : \mathbb{R}^n \to \mathbb{R}$, consider the unconstrained optimization problem

$$\text{minimize } f(x)$$

(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 randomly
  - 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 $\text{span}\{e_{i(k)}\}$. 

Notes
Algorithm 2 Coordinate minimization with cyclic order and exact minimization.

1: Choose $x_0 \in \mathbb{R}^n$ and set $k \leftarrow 0$.
2: 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: 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
  \[ \|\nabla f(x_k)\|_2 \leq 10^{-6} \max\{1, \|\nabla f(x_0)\|_2\} \]

An interesting example introduced by Powell [5, formula 2] is
\[ \text{minimize } f(x_1, x_2, x_3) := -(x_1 x_2 + x_2 x_3 + x_1 x_3) + \sum_{i=1}^{3} (|x_i| - 1)^2 \]
- $f$ 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.

Figure: Three dimensional example given above. It shows the possible lack of convergence of a coordinate descent method with exact minimization. This example and others in [5] show that we cannot expect a general convergence result for nonconvex functions similar to that for full-gradient descent. This picture was taken from [7].

Notes

Notes
Theorem 2.1 (see [6, Theorem 5.32])

Assume that the following hold:

- $f$ is continuously differentiable;
- the level set $L_0 := \{ x \in \mathbb{R}^n : f(x) \leq f(x_0) \}$ is bounded; and
- for every $x \in L_0$ 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 $\{x_k\}_{k=0}^\infty \subset L_0$. Since $L_0$ is bounded that $\{x_k\}_{k=0}^\infty$ has at least one limit point; let $x^*$ be any such limit point. Thus, there exists a subsequence $K \subset \mathbb{N}$ satisfying
\[
\lim_{k \in K} x_k = x^*. \tag{2}
\]

Combining this with monotonicity of $\{ f(x_k) \}$ and continuity of $f$ also shows that
\[
\lim_{k \to \infty} f(x_k) = f(x^*) \quad \text{and} \quad f(x_k) \geq f(x^*) \quad \text{for all} \quad k \in \mathbb{N}. \tag{3}
\]

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

First, consider the subsets $K_i \subset K, i = 0, \ldots, n - 1$ defined as
\[
K_i := \{ k \in K : k \equiv i \mod n \}.
\]

Since $K$ is an infinite subsequence of the natural numbers, one of the $K_i$ must be an infinite set. Without loss of generality, we assume it is $K_0$ (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
\[
y^* := x^*_n, \quad \text{with} \quad x^*_n := x^* + \sum_{j=1}^\ell [r^*] e_j \quad \text{for all} \quad \ell = 1, \ldots, n
\]
and note that since $\nabla f(x^*) \neq 0$ by assumption, we must have
\[
f(y^*) < f(x^*). \quad \text{(why?)} \tag{4}
\]

NOTE: If $K_i$ 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
\[ x_{k+\ell} = x_k + \sum_{j=1}^{\ell} \tau_{k+j-1} e_j \text{ for all } k \in \mathcal{K}_0 \text{ and } 1 \leq \ell \leq n, \]  
(5)
where \( \tau_i \) is the step length at iteration \( i \). Therefore,
\[ \| x_{k+\ell} - x_k \| = \left\| \begin{pmatrix} \tau_k \\ \tau_{k+1} \\ \vdots \\ \tau_{k+\ell-1} \end{pmatrix} \right\| \leq 2 \max\{\|x\| : x \in \mathcal{L}_0\} < \infty \]
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 \( \{ (\tau_k, \tau_{k+1}, \ldots, \tau_{k+n-1})^T \}_{k \in \mathcal{K}_0} \) is bounded, we may pass to a subsequence \( \mathcal{K}' \subseteq \mathcal{K}_0 \) with
\[ \lim_{k \in \mathcal{K}'} \begin{pmatrix} \tau_k \\ \tau_{k+1} \\ \vdots \\ \tau_{k+n-1} \end{pmatrix} = \tau^* \text{ for some } \tau^* \in \mathbb{R}^n. \]  
(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
\[ \lim_{k \in \mathcal{K}'} x_{k+\ell} = x^* + \sum_{j=1}^{\ell} [\tau^*]^j e_j \text{ for each } 1 \leq \ell \leq n. \]  
(7)

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

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

We know from the coordinate minimization that
\[ f(x_{k+1}) \leq f(x_k + \tau e_1) \text{ for all } k \text{ and } \tau \in \mathbb{R}. \]
Taking limits over \( k \in \mathcal{K}' \subseteq \mathcal{K} \) and using continuity of \( f \), (7) with \( \ell = 1 \), and (2) yields
\[ f(x^* + [\tau^*]_1 e_1) = f(\lim_{k \in \mathcal{K}'} x_{k+1}) = \lim_{k \in \mathcal{K}'} f(x_{k+1}) \leq \lim_{k \in \mathcal{K}'} f(x_k + \tau e_1) \]
\[ = f(\lim_{k \in \mathcal{K}'} x_k + \tau e_1) = f(x^* + \tau e_1) \text{ for all } \tau \in \mathbb{R}. \]

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
\[ \lim_{k \in \mathcal{K}'} x_{k+1} = x^* + [\tau^*]_1 e_1 = x^* + [\tau^*]_1 e_1 \equiv x^*_1, \]
which completes the base case.
**Induction step:** assume that (8) and (9) hold for \(1 \leq p \leq \bar{p} - 1\).

We know from the coordinate minimization that
\[
f(x_{k+\bar{p}+1}) \leq f(x_{k+\bar{p}} + \tau e_{\bar{p}+1})
\]
for all \(k \in \mathbb{N}\) and \(\tau \in \mathbb{R}\).

Taking the limit over \(k \in K'\), continuity of \(f\), (7) with \(\ell = \bar{p} + 1\), and (9) give
\[
f\left(x^* + \sum_{j=1}^{\bar{p}+1} [\tau^*]_j e_j\right) = f(\lim_{k \in K'} x_{k+\bar{p}+1}) = \lim_{k \in K'} f(x_{k+\bar{p}} + \tau e_{\bar{p}+1})
\]
\[
= f(\lim_{k \in K'} x_{k+\bar{p}} + \tau e_{\bar{p}+1}) = f(x^* + \tau e_{\bar{p}+1})
\]
for all \(\tau \in \mathbb{R}\).

Thus, the definition of \(x^*_\bar{p}\), and the fact that (8) holds for all \(1 \leq p \leq \bar{p}\) show that
\[
f(x^*_\bar{p} + \sum_{j=1}^{\bar{p}} [\tau^*]_j e_j) \leq f(x^* + \sum_{j=1}^{\bar{p}+1} [\tau^*]_j e_j)
\]
\[
= f(x^* + \sum_{j=1}^{\bar{p}+1} [\tau^*]_j e_j)
\]
\[
= f(x^* + \sum_{j=1}^{\bar{p}+1} [\tau^*]_j e_j) \leq f(x^* + \tau e_{\bar{p}+1})
\]
for all \(\tau \in \mathbb{R}\).

Uniqueness of the minimizer implies \([\tau^*]_{\bar{p}+1} = [\tau^*]_{\bar{p}+1}\) and combining with (7) gives
\[
\lim_{k \in K'} x_{k+\bar{p}+1} = x^* + \sum_{j=1}^{\bar{p}+1} [\tau^*]_j e_j = x^* + \sum_{j=1}^{\bar{p}+1} [\tau^*]_j e_j \equiv x^*_{\bar{p}+1},
\]
which completes the proof by induction.

From our induction proof, we have that
\[
\tau^* = \tau^\ell.
\]

Combining this with (7) and the definition of \(y^*\) gives
\[
\lim_{k \in K'} x_{k+n} = x^* + \sum_{j=1}^{n} \tau^*_j e_j = x^* + \sum_{j=1}^{n} \tau^*_j e_j \equiv x^*_{n} \equiv y^*.
\]

(10)

Finally, combining (3), continuity of \(f\), (10), and (4) shows that
\[
f(x^*) = \lim_{k \in K'} f(x_{k+n}) = f(\lim_{k \in K'} x_{k+n}) = f(y^*) < f(x^*),
\]
which is a contradiction. This completes the proof. \(\blacksquare\)
Notation:

- Let \( L_j \) denote the \( j \)th component Lipschitz constant, i.e., it satisfies
  \[
  \left| \nabla_j f(x + te_j) - \nabla_j f(x) \right| \leq L_j |t| \]
  for all \( x \in \mathbb{R}^n \) and \( t \).

- Let \( L_{\max} \) denote the coordinate Lipschitz constant, i.e., it satisfies
  \[
  L_{\max} := \max_{1 \leq i \leq n} L_i.
  \]

- 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: \textbf{loop}
4: \hspace{0.5cm} Choose \( i(k) = \text{mod}(k, n) + 1 \).
5: \hspace{0.5cm} Set \( x_{k+1} \leftarrow x_k - \alpha \nabla_{i(k)} f(x_k) e_{i(k)} \).
6: \hspace{0.5cm} Set \( k \leftarrow k + 1 \).
7: \textbf{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 \( x^* := \min_{x \in \mathbb{R}^n} f(x) \)
- 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

\[
\min_{k=0,\ldots,T} \|\nabla f(x_k)\| \leq \sqrt{\frac{4nL_{\max}(1 + nL^2/L_{\max})f(x_0) - f^*}{T + 1}} \quad \forall T \in \{n, 2n, 3n, \ldots\} \tag{11}
\]

If \( f \) is convex, then

\[
f(x_T) - f^* \leq \frac{4nL_{\max}(1 + nL^2/L_{\max})R_0^2}{T + 8} \quad \forall T \in \{n, 2n, 3n, \ldots\}. \tag{12}
\]

If \( f \) is \( \mu \)-strongly convex, then

\[
f(x_T) - f^* \leq \left(1 - \frac{\mu}{2L_{\max}(1 + nL^2/L_{\max})}\right)^{T/n} (f(x_0) - f^*) \quad \forall T \in \{n, 2n, 3n, \ldots\}
\]
Proof: See [1, Lemma 3.3, Theorem 3.6 and Theorem 3.9] and use (i) each iteration “k” in [1] is a cycle of \( n \) iterations; (ii) choose in [1] the values \( \bar{L}_i = L_{\text{max}} \) for all \( i \); (iii) in [1] we have \( p = 1 \) since our blocks of variables are singletons, i.e., coordinate descent.

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
  \[
  |\nabla f(x + te_j) - \nabla f(x)| \leq ||\nabla f(x + te_j) - \nabla f(x)||_2 \leq L|t|
  \]
  holds for all \( j, t, \) and \( x \) that \( L_j \leq L \).
- By combining the previous two bullet points, we find that
  \[
  L_{\text{max}} \equiv \max_j L_j \leq L \leq \sum_{j=1}^n L_j \leq nL_{\text{max}}
  \]
  so that
  \[
  1 \leq \frac{L}{L_{\text{max}}} \leq n
  \]
- Roughly speaking, \( L/L_{\text{max}} \) 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 $L_j$ denote the $j$th component Lipschitz constant, i.e., it satisfies 
  $$\| \nabla f(x + te_j) - \nabla f(x) \| \leq L_j |t| \text{ for all } x \in \mathbb{R}^n \text{ and } t.$$ 
- Let $L_{\text{max}}$ denote the coordinate Lipschitz constant, i.e., it satisfies 
  $$L_{\text{max}} := \max_{1 \leq i \leq n} L_i.$$

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

1: Choose $\alpha \in (0, 1/L_{\text{max}}]$.
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 $x_{k+1} \leftarrow x_k - \alpha \nabla_i f(x_k)e_{i(k)}$.
6: Set $k \leftarrow k + 1$.
7: end loop

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

Theorem 2.3
Suppose that $\alpha = 1/L_{\text{max}}$ 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 $K_0$ such that the diameter of the level set 
  $$\{ x \in \mathbb{R}^n : f(x) \leq f(x_0) \}$$ 
  is bounded by $K_0$.

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

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

If $f$ is convex, then

$$f(x_T) - f^* \leq \frac{2nL_{\text{max}}K_0^2}{T} \quad \forall T \geq 1.$$ 

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

$$f(x_T) - f^* \leq \left(1 - \frac{\mu}{nL_{\text{max}}} \right)^k (f(x_0) - f^*) \quad \forall T \geq 1.$$
Proof: We recall our fundamental inequality (see the “Stochastic Gradient Descent” lecture notes on random coordinate choice)

\[ f(x_{k+1}) \leq f(x_k) - \frac{1}{2L_{\text{max}}} (\nabla_i(f(x_k)))^2. \]

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

\[
\begin{align*}
f(x_{k+1}) & \leq f(x_k) - \frac{1}{2L_{\text{max}}} (\nabla_i(f(x_k)))^2 \\
& = f(x_k) - \frac{1}{2L_{\text{max}}} \|\nabla f(x_k)\|_\infty^2 \quad (13) \\
& \leq f(x_k) - \frac{1}{2nL_{\text{max}}} \|\nabla f(x_k)\|_2^2. \quad (14)
\end{align*}
\]

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)

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

Using the notation \( \Delta_k = f(x_k) - f^* \), this is equivalent to

\[ \Delta_{k+1} \leq \Delta_k - \frac{1}{2nL_{\text{max}}R_0^2} \Delta_k^2. \]

From the previous slide, we have

\[ \Delta_{k+1} \leq \Delta_k - \frac{1}{2nL_{\text{max}}R_0^2} \Delta_k^2 \]

which is exactly the same as the inequality (2.2) from the “Smooth Convex Optimization” lecture notes, except with an extra factor of the dimension \( n \) in the denominator of the last term. Then, as shown in that proof, we have

\[ f(x_k) - f^* = \Delta_k \leq \frac{2nL_{\text{max}}R_0^2}{k} \]

which is the desired result for convex \( f \).
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:

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

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

\[
\begin{align*}
    f(x_{k+1}) - f^* &\leq f(x_k) - f^* - \frac{1}{2nL_{\max}} \|\nabla f(x_k)\|^2 \\
    &\leq f(x_k) - f^* - \frac{\mu}{nL_{\max}} (f(x_k) - f^*) = \left(1 - \frac{\mu}{nL_{\max}}\right) (f(x_k) - f^*)
\end{align*}
\]

so that

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

which is the last desired result.

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

- $\nabla f$ is globally Lipschitz continuous

- the minimum value of $f$ is obtained

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

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

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

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

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

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

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

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

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

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
\[ \| \nabla f(x_k) \|_\infty \geq 2 \mu_1 (f(x_k) - f^*). \]
Subtracting \( f^* \) from both sides of (13) and using the previous inequality shows that
\[
\begin{align*}
 f(x_{k+1}) - f^* & \leq f(x_k) - f^* - \frac{1}{2L_{\text{max}}} \| \nabla f(x_k) \|_\infty^2 \\
 & \leq f(x_k) - f^* - \frac{\mu_1}{L_{\text{max}}} (f(x_k) - f^*) \\
 & = \left(1 - \frac{\mu_1}{L_{\text{max}}} \right) (f(x_k) - f^*). 
\end{align*}
\]
Applying this inequality recursively gives
\[
 f(x_k) - f^* \leq \left(1 - \frac{\mu_1}{L_{\text{max}}} \right)^k (f(x_0) - f^*)
\]
which is the desired result. 

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

Consider the problem

\[
\min_{x \in \mathbb{R}^n} g^T x + \frac{1}{2} x^T H x
\]

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

\[
\mu = \min\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \quad \text{and} \quad \mu_1 = \left( \sum_{i=1}^n \frac{1}{\lambda_i} \right)^{-1}
\]

**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

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

Thus, the convergence constants are:

(random selection) : \( 1 - \frac{\mu}{nL_{\max}} = 1 - \frac{\alpha}{nL_{\max}} \)

(Gauss-Southwell selection) : \( 1 - \frac{\mu_1}{L_{\max}} = 1 - \frac{\alpha}{nL_{\max}} \)

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

**Case 2:** For this other extreme case, let us suppose that

\[
\lambda_1 = \beta \quad \text{and} \quad \lambda_2 = \lambda_3 = \cdots = \lambda_n = \alpha
\]

with \( \alpha \geq \beta \). For this case, it can be shown that

\[
\mu = \beta \quad \text{and} \quad \mu_1 = \frac{\beta \alpha^{n-1}}{\alpha^{n-1} + (n-1)\beta \alpha^{n-2}} = \frac{\beta \alpha}{\alpha + (n-1)\beta}
\]

If we now take the limit as \( \alpha \to \infty \) we find that

\[
\mu = \beta \quad \text{and} \quad \mu_1 \to \beta = \mu
\]

Thus, the convergence constants (in the limit) are:

(random selection) : \( 1 - \frac{\mu}{nL_{\max}} = 1 - \frac{\beta}{nL_{\max}} \)

(Gauss-Southwell selection) : \( 1 - \frac{\mu_1}{L_{\max}} = 1 - \frac{\beta}{L_{\max}} \)

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

\[ x_{k+1} = x_k + \frac{1}{L_i(k)} \nabla_i f(x_k) e_i(k) \]

- Using a similar analysis as before, it can be shown

\[ f(x_k) - f^* \leq \left( \prod_{j=1}^{k} \left( 1 - \frac{\mu_1}{L_j} \right) \right) (f(x_0) - f^*) \]

- Better decrease than prior analysis since (see (15))

\[ \text{new rate} = \left( \prod_{j=1}^{k} \left( 1 - \frac{\mu_1}{L_j} \right) \right) \leq \left( 1 - \frac{\mu_1}{L_{\text{max}}} \right)^k \text{ previous rate} \]

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

Alternative 2 (strongly convex): Lipschitz sampling.

Use a random coordinate direction chosen using a non-uniform probability distribution:

\[ P(i(k) = j) = \frac{L_j}{\sum_{\ell=1}^{n} L_\ell} \text{ for all } j \in \{1, 2, \ldots, n\} \]

- Using an analysis similar to the previous one, but using the new probability distribution when computing the expectation, it can be shown that

\[ \mathbb{E}[f(x_{k+1})] - f^* \leq \left( 1 - \frac{\mu}{nL} \right) (\mathbb{E}[f(x_k)] - f^*) \]

with \( \bar{L} \) being the average component Lipschitz constant, i.e.,

\[ \bar{L} := \frac{1}{n} \sum_{i=1}^{n} L_i \]

- The analysis was first performed in [2].

- This rate is faster than uniform random sampling if not all of the component Lipschitz constants are the same.
Alternative 3 (strongly convex): Gauss-Southwell-Lipschitz rule.

Choose $i(k)$ according to the rule

$$i(k) \leftarrow \max_{1 \leq i \leq n} \left( \frac{\nabla_i f(x_k)^2}{L_i} \right)$$  \hspace{1cm} (16)

- We recall our fundamental inequality for coordinate descent with step size
  $$\alpha_k = \frac{1}{L_{\text{max}}}$$
  $$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L_i(i)} \left( \nabla_i f(x_k) \right)^2$$  \hspace{1cm} (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
  $$f(x_{k+1}) - f^* \leq (1 - \mu_L) (f(x_k) - f^*)$$
  where $\mu_L$ is the strong convexity parameter with respect to $\|v\|_L := \sum_{i=1}^{n} \sqrt{L_i} |v_i|$.
- It is shown in [4, Appendix 6.2] that
  $$\max \left\{ \frac{\mu}{nL}, \frac{\mu_1}{L_{\text{max}}} \right\} \leq \mu_L \leq \frac{\mu_1}{\min_{1 \leq i \leq n} \{L_i\}}$$
- 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, $L_{\text{max}}$)
- Gauss-Southwell ($L_{\text{max}}$)
- Gauss-Southwell with $\{L_i\}$
- random (Lipschitz sampling, $\{L_i\}$)
- Gauss-Southwell-Lipschitz

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., $L_{\text{max}}$. 
Linear Equations

Let \( m \leq n, b \in \mathbb{R}^m, \) and \( A^T = (a_1, \ldots, a_n) \in \mathbb{R}^{n \times m} \) with \( \|a_i\|_2 = 1 \) for all \( i \).

Furthermore, suppose that \( A^T \) has full column rank, meaning that the linear system

\[
Aw = b
\]

has infinitely many solutions. To seek the least-length solution, we wish to solve

\[
\text{minimize } \frac{1}{2} ||w||_2^2 \text{ subject to } Aw = b.
\]

The Lagrangian dual problem is

\[
\text{minimize } f(x) := \frac{1}{2} ||A^T x||_2^2 - b^T x,
\]

where we note that \( \nabla f(x) = A A^T x - b \) and \( \nabla_i f(x) = a_i^T A^T x - b_i \). The solutions to the primal and dual are related via \( w^* = A^T x^* \). Coordinate descent gives

\[
w_{k+1} = w_k - \alpha (a_i^T A^T x_k - b_i) a_i.
\]

If we maintain an estimate \( x_k = A^T w_k \), then we see that

\[
w_{k+1} = A^T x_{k+1} = A^T (x_k - \alpha (a_i^T A^T x_k - b_i) e_i)
\]

so that the \( i \)-th equation is satisfied exactly.

Summary

- Coordinate minimization for solving the dual problem associated with linear equations along the direction \( e_i \) with \( \alpha = 1 \) satisfies the \( i \)-th linear equation exactly.
- Sometimes called the method of successive projections.
- Update: \( w_{k+1} = w_k - \alpha (a_i^T w_k - b_i) a_i \)
  - \( (n+1) \) addition/subtractions
  - \( (2n+1) \) multiplications
  - \( (3n+2) \) total floating-point operations
- Computing \( \nabla f(x) \) requires a multiplication with \( A \), which is much more expensive.
Logistic Regression

Give data \( \{d_j\}_{j=1}^N \subset \mathbb{R}^n \) and labels \( \{y_j\}_{j=1}^N \subset \{-1, 1\} \) associated with the data, solve

\[
\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{N} \sum_{j=1}^N \log \left( 1 + e^{-y_j d_j^T x} \right).
\]

If we define the data matrix \( D \) such that

\[
D = \begin{pmatrix} d_1^T \\ \vdots \\ d_N^T \end{pmatrix},
\]

then it follows that

\[
\nabla_i f(x) = -\frac{1}{N} \sum_{j=1}^N \frac{e^{-y_j d_j^T x}}{1 + e^{-y_j d_j^T x}} y_j d_{ji}.
\]

Consider the coordinate minimization update

\[
x_{k+1} = x_k - \alpha \nabla_i f(x) e_{i(k)}
\]

for some \( i(k) \in \{1, 2, \ldots, n\} \) and \( \alpha \in \mathbb{R} \).

For efficiency, we store and update the required quantities \( \{Dx_k\} \) using

\[
Dx_{k+1} = D(x_k + \alpha \nabla_i f(x) e_{i(k)}) = Dx_k + \alpha \nabla_i f(x) De_{i(k)} = Dx_k + \alpha \nabla_i f(x) D(:, i(k)),
\]

where \( D(:, i(k)) \) denotes the \( i(k) \)-th column of \( D \); if \( x_0 \leftarrow 0 \), then we can set \( Dx_0 \leftarrow 0 \).

Logistic Regression

Summary

- Update to obtain \( Dx_{k+1} \) requires a single vector-vector add.
- Computing \( \nabla f(x_k) \) only requires accessing a single column of the data matrix \( D \).
- Computing \( \nabla f(x) \) requires accessing the entire data matrix \( D \).


