The material presented herein is based on a variety of sources, but most specifically on:

- "Approximation accuracy, gradient methods, and error bound for structured convex optimization" by Paul Tseng, *Mathematical Programming B* (2010), 125, pp. 263-295, and

1 Smooth Method in Simple Settings

1.1 Problem Setting and Basics for Smooth Method

Our problem of interest is

\[ P : \text{minimize}_{x} \ f(x) \]

\[ \text{s.t.} \quad x \in \mathbb{R}^n, \tag{1} \]

where \( f(\cdot) \) is a differentiable convex function defined on \( \mathbb{R}^n \). We denote the optimal objective value by \( f^* \) and let \( x^* \) denote an optimal solution of \( P \), when such a solution exists.

Let \( \nabla f(x) \) denote the gradient of \( f(\cdot) \) at \( x \), and recall the gradient inequality:

\[ f(y) \geq f(x) + \nabla f(x)^T(y - x) \quad \text{for all} \ y \in \mathbb{R}^n. \tag{2} \]

We will measure distances between points using the Euclidean norm \( \| \cdot \| := \| \cdot \|_2 \) where \( \| \cdot \|_2 := \sqrt{x^T x} \).
We assume that \( f(\cdot) \) has a *Lipschitz gradient*. That is, there is a scalar \( L \) for which

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

(3)

Let \( B(c, R) \) denote the ball centered at \( c \) with radius \( R \), namely:

\[
B(c, R) := \{ x \in \mathbb{R}^n : \|x - c\| \leq R \}.
\]

Algorithm 1 presents a simple gradient scheme for solving (P).

**Algorithm 1** Simple Gradient Scheme

**Initialize.** Initialize with \( x^0 \) and \( i \leftarrow 0 \).

At iteration \( i \):

1. **Compute New Point.** Compute \( \nabla f(x^i) \), and then compute:

\[
x^{i+1} \leftarrow x^i - \frac{1}{L} \nabla f(x^i)
\]

1.2 **Analysis and Complexity of the Simple Gradient Scheme**

We begin with a basic property of Lipschitz gradients:

**Proposition 1.1** If \( f(\cdot) \) has Lipschitz gradient with constant \( L \), then

\[
f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2 \quad \text{for all } x, y \in \mathbb{R}^n.
\]
Proof: We use the fundamental theorem of calculus applied in a multivariate setting. We have

\[ f(y) = f(x) + \int_0^1 \nabla f(x + t(y - x))^T (y - x) dt \]

\[ = f(x) + \nabla f(x)^T (y - x) + \int_0^1 [\nabla f(x + t(y - x)) - \nabla f(x)]^T (y - x) dt \]

\[ \leq f(x) + \nabla f(x)^T (y - x) + \int_0^1 \| \nabla f(x + t(y - x)) - \nabla f(x) \| \| (y - x) \| dt \]

\[ \leq f(x) + \nabla f(x)^T (y - x) + \int_0^1 t \| (y - x) \|^2 dt \]

\[ = f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \| (y - x) \|^2 . \]

We also will use the following arithmetic result:

Proposition 1.2 For any \( x \) the following holds:

\[ \nabla f(x^i)^T (x - x^i) + \frac{L}{2} \| x - x^i \|^2 - \frac{L}{2} \| x - x^{i+1} \|^2 = \nabla f(x^i)^T (x^{i+1} - x^i) + \frac{L}{2} \| x^{i+1} - x^i \|^2 . \]

Proof: This is just rearrangement of terms using the fact that \( x^{i+1} = x^i - \frac{1}{L} \nabla f(x^i) \). Expanding the left side of the proposition we have:

\[ \nabla f(x^i)^T (x - x^i) + \frac{L}{2} \| x - x^i \|^2 - \frac{L}{2} \| x - x^{i+1} \|^2 \]

\[ = L \left[ (x^i - x^{i+1})^T (x - x^i) + \frac{1}{2} \| x \|^2 - x^T x^i + \frac{1}{2} \| x^i \|^2 - \frac{1}{2} \| x \|^2 + x^T x^{i+1} - \frac{1}{2} \| x^{i+1} \|^2 \right] \]

\[ = L \left[ (x^{i+1})^T (x^i) - \| x^i \|^2 + \frac{1}{2} \| x^i \|^2 - \frac{1}{2} \| x^{i+1} \|^2 \right] \]

\[ = L \left[ -\frac{1}{2} \| x^i \|^2 + (x^{i+1})^T (x^i) - \frac{1}{2} \| x^{i+1} \|^2 \right] \]

\[ = -\frac{L}{2} \| x^{i+1} - x^i \|^2 \]

\[ = -L(x^{i+1} - x^i)^T (x^{i+1} - x^i) + \frac{L}{2} \| x^{i+1} - x^i \|^2 \]

\[ = \nabla f(x^i)^T (x^{i+1} - x^i) + \frac{L}{2} \| x^{i+1} - x^i \|^2 . \]

Our main algorithmic analysis result for the simple gradient scheme is:
Theorem 1.1 After $k$ iterations of the simple gradient scheme, the following holds:

$$f(x^k) \leq f(x) + \frac{L\|x - x^0\|^2}{2k} \quad \text{for all } x \in \mathbb{R}^n .$$

Corollary 1.1 Suppose our region of interest is $B(x^0, R)$, which may or may not contain an optimal solution $x^*$. Then:

$$f(x^k) \leq \min_{x \in B(x^0, R)} f(x) + \frac{R^2L}{2k} .$$

Let $x^*$ be some optimal solution, let $\varepsilon > 0$, and let

$$k := \left\lceil \frac{\|x^0 - x^*\|^2 L}{2\varepsilon} \right\rceil .$$

Then $f(x^k) \leq f^* + \varepsilon .

Proof of Theorem 1.1: We have:

$$f(x^{i+1}) \leq f(x^i) + \nabla f(x^i)^T(x^{i+1} - x^i) + \frac{L}{2}\|x^{i+1} - x^i\|^2 \quad \text{ (from Proposition 1.1)}$$

$$= f(x^i) - L\|x^{i+1} - x^i\|^2 + \frac{L}{2}\|x^{i+1} - x^i\|^2$$

$$= f(x^i) - \frac{L}{2}\|x^{i+1} - x^i\|^2$$

$$\leq f(x^i) ,$$

whereby we see that the function values $f(x^i)$ are decreasing in $i$. Restating the first inequality above again for a different purpose, we have for all $x$:

$$f(x^{i+1}) \leq f(x^i) + \nabla f(x^i)^T(x^{i+1} - x^i) + \frac{L}{2}\|x^{i+1} - x^i\|^2 \quad \text{ (from Proposition 1.1)}$$

$$= f(x^i) + \nabla f(x^i)^T(x - x^i) + \frac{L}{2}\|x - x^i\|^2 - \frac{L}{2}\|x - x^{i+1}\|^2 \quad \text{ (from Proposition 1.2)}$$

$$\leq f(x) + \frac{L}{2}\|x - x^i\|^2 - \frac{L}{2}\|x - x^{i+1}\|^2 . \quad \text{(from (2))}$$
Summing up and recalling from the start of the proof that the sequence \( f(x^i) \) is nonincreasing, we have:

\[
k f(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1}) \leq kf(x) + \frac{L}{2} \|x - x^0\|^2 - \frac{L}{2} \|x - x^k\|^2.
\]

Dividing by \( k \) and noting that the final subtractand above is nonnegative, we obtain:

\[
f(x^k) \leq f(x) + \frac{L\|x - x^0\|^2}{2k}.
\]

\[\blacksquare\]

1.3 Comments and Extensions

1. It is important for future extensions and analysis to observe that \( x^{i+1} \)
   solves the following elementary optimization problem:

\[
x^{i+1} = \arg \min_x f(x^i) + \nabla f(x^i)^T (x - x^i) + \frac{L}{2} \|x - x^i\|^2.
\]

2. How might the scheme and the analysis be modified if \( L \) is not explicitly known?

3. How might the scheme and the analysis be modified if one can efficiently do a linesearch to compute:

\[
t_i := \arg \min_t \{f(x^i - t\nabla f(x^i))\},
\]

instead of (implicitly) assigning the value \( t_i \leftarrow 1/L \) at each iteration?

2 Non-smooth Method in Simple Settings

2.1 Problem Setting and Basics for Non-smooth Method

Our problem of interest still is

\[
P : \text{ minimize}_x \quad f(x)
\]

s.t. \( x \in \mathbb{R}^n \).
where $f(\cdot)$ is a convex function defined on $\mathbb{R}^n$. As before, we denote the optimal objective value by $f^*$.

We no longer assume that $f(\cdot)$ is differentiable. Let $\partial f(x)$ denote the set of subgradients of $f(\cdot)$ at $x$, namely $g \in \partial f(x)$ if and only if:

$$
  f(y) \geq f(x) + g^T(y - x) \quad \text{for all } y \in \mathbb{R}^n.
$$

(4)

Recall that $\partial f(x)$ is nonempty, and we assume that computing a subgradient of $f(\cdot)$ at any given $x$ can be done efficiently.

We will continue to measure distances between points using the Euclidean norm $\|\cdot\| := \|\cdot\|_2$ where $\|\cdot\|_2 := \sqrt{x^T x}$.

We assume that $f(\cdot)$ has Lipschitz function values. That is, there is a scalar $L$ for which

$$
  |f(y) - f(x)| \leq L\|y - x\| \quad \text{for all } x, y \in \mathbb{R}^n.
$$

(5)

Also, recall that $B(c, R)$ denotes the ball centered at $c$ with radius $R$, namely:

$$
  B(c, R) := \{x \in \mathbb{R}^n : \|x - c\| \leq R\}.
$$

Algorithm 2 presents a simple subgradient scheme.

2.2 Analysis and Complexity of the Simple Subgradient Scheme

We begin with a simple relationship between the Lipschitz constant and the norm of any subgradient:

**Proposition 2.1** If $g \in \partial f(x)$, then $\|g\| \leq L$.

**Proof:** Let $g \in \partial f(x)$ be given. If $g = 0$, the result is trivial. If $g \neq 0$, then since $f(x + g) \geq f(x) + g^T(x + g - x) = f(x) + g^T g$, it follows that

$$
  \|g\|^2 = g^T g \leq f(x + g) - f(x) \leq |f(x + g) - f(x)| \leq L\|g\|,
$$

7
Algorithm 2 Simple Subgradient Scheme

**Initialize.** Define step-sizes $t_i$, $i = 0, 1, 2, \ldots$

Initialize with $x^0$ and $i \leftarrow 0$.

At iteration $i$:

1. **Compute New Point.** Compute $g_i \in \partial f(x^i)$, and then compute:

   $$x^{i+1} \leftarrow x^i - t_i g_i .$$

   which upon dividing by $\|g\|$ yields $\|g\| \leq L$.

Our main algorithmic analysis result is:

**Theorem 2.1** After $k$ iterations of the simple subgradient scheme, the following holds:

$$\min_{0 \leq i \leq k} f(x^i) \leq f(x) + \frac{L^2 \sum_{i=0}^k t_i^2 + \|x - x^0\|^2}{2 \sum_{i=0}^k t_i} \quad \text{for all } x \in \mathbb{R}^n .$$

**Corollary 2.1** Suppose we wish to run Algorithm 2 for $k$ iterations, and our region of interest is $B(x^0, R)$, which may or may not contain an optimal solution $x^*$. If we set the step-sizes to be:

$$t_i := \bar{t} := \frac{R}{L\sqrt{k + 1}}$$

we obtain:

$$\min_{0 \leq i \leq k} f(x^i) \leq \min_{x \in B(x^0, R)} f(x) + \frac{RL}{\sqrt{k + 1}} .$$
If we are given $\varepsilon > 0$ and we wish to compute a point whose objective function value is within $\varepsilon$ of an optimal solution $x^*$, then we can do so within

$$k := \left\lceil \frac{\|x^0 - x^*\|^2 L^2}{\varepsilon^2} \right\rceil - 1$$

iterations using the constant step-size

$$t_i = \bar{t} = \frac{\varepsilon}{L^2}.$$ 

**Proof of Theorem 2.1:** If we have just computed $x^{i+1}$, we have for any $x$:

$$\|x^{i+1} - x\|^2 = \|x^i - t_i g_i - x\|^2$$

$$= \|x^i - x\|^2 + t_i^2 \|g_i\|^2 + 2t_i g_i^T (x - x^i)$$

$$\leq \|x^i - x\|^2 + t_i^2 \|g_i\|^2 + 2t_i (f(x) - f(x^i)) \quad \text{(subgradient inequality)}$$

$$\leq \|x^i - x\|^2 + t_i^2 L^2 + 2t_i (f(x) - f(x^i)) \quad \text{(Proposition 2.1)}$$

which rearranges to:

$$2t_i f(x^i) \leq 2t_i f(x) + t_i^2 L^2 + \|x^i - x\|^2 - \|x^{i+1} - x\|^2.$$ 

Denote:

$$f^*_k := \min_{0 \leq i \leq k} f(x^i).$$

Now sum up the above inequality over $i$:

$$2 \sum_{i=0}^{k} t_i f^*_k \leq 2 \sum_{i=0}^{k} t_i f(x^i) \leq 2 \sum_{i=0}^{k} t_i f(x) + \sum_{i=0}^{k} t_i^2 L^2 + \|x^0 - x\|^2 - \|x^{k+1} - x\|^2.$$ 

Since the last subtractand is nonnegative, it can be eliminated, and dividing by $2 \sum_{i=0}^{k} t_i$ yields:

$$f^*_k \leq f(x) + \frac{L^2 \sum_{i=0}^{k} t_i^2 + \|x - x^0\|^2}{2 \sum_{i=0}^{k} t_i}.$$ 

$\blacksquare$
2.3 Comments and Extensions

1. The first result in Corollary 2.1 presumes that we know in advance how many iterations we wish to run the method. This enables a constant step-size $\bar{t}$ to be determined that is optimal for the inequality of the corollary. What if one wishes to run the method simply for “a while”? How might one proceed with an analysis of, say, choosing $t_i := \frac{C}{\sqrt{i+1}}$ for some suitably chosen constant $C$?

2. How might the scheme and the analysis be modified if $L$ is not explicitly known?

3. Suppose that our optimization problem includes a feasibility set $S$ that is computationally efficient to work with, in the sense that computing the (Euclidean) projection $P_S(x)$ of a point $x$ onto the set $S$ is easy to do. (This is the case when $S$ is a box, a ball, or the standard simplex $Q := \{ x \in \mathbb{R}^n : x \geq 0, e^T x = 1 \}$ where $e = (1, \ldots, 1)^T$ denotes the column vector of 1’s.) Then it is a simple exercise to modify the main algorithm step to:

$$x^{i+1} \leftarrow P_S(x^i - t_i g_i).$$

The proof of Theorem 2.1 needs to be suitably modified; this can easily be done if one relies on the following contraction property of projection operators:

**Proposition 2.2** If $S$ is a convex set, then for all $x, y \in \mathbb{R}^n$ it holds that $\|P_S(x) - P_S(y)\| \leq \|x - y\|$.

4. A slightly modified step-size rule is to incorporate the norm of $g_i$ into the step-size itself. If we set $t_i = \frac{\alpha_i}{\|g_i\|}$ for some $\alpha_i$, then $\|x^{i+1} - x^i\| = t_i \|g_i\| = \alpha_i$, so that $\alpha_i$ is literally the length of the step. It is a useful exercise to explore how the results of Theorem 2.1 and Corollary 2.1 change with this step-size rule.
3 An Accelerated Smooth Method (with Optimal Iteration Complexity) in Simple Setting

In this section we reconsider the case when \( f(\cdot) \) is smooth (namely, differentiable.) We present an “accelerated” version of the gradient method of Section 1.1 for solving (1), that has superior theoretical and practical performance over the gradient scheme presented in Section 1.1. Recall the gradient inequality (2). As in Section 1.1, we will consider the case when \( f(\cdot) \) has a Lipschitz gradient, see (3), and recall Proposition 1.1 which states a useful inequality for functions whose gradient is Lipschitz.

Algorithm 3 presents an accelerated simple gradient scheme.

**Algorithm 3** Accelerated Simple Gradient Scheme

**Initialize.**

Define step-size parameters \( \theta_i \in (0, 1] \) for \( i = 0, 1, \ldots \).

Initialize with \( x^0 \) and \( z^0 := x^0 \), and \( i \leftarrow 0 \).

At iteration \( i \):

1. **Compute New Points.**

   Define \( y^i \leftarrow (1 - \theta_i)x^i + \theta_iz^i \), and compute \( \nabla f(y^i) \).

   \[
   z^{i+1} \leftarrow z^i - \frac{1}{\theta_i} \nabla f(y^i).
   \]

   \[
   x^{i+1} \leftarrow (1 - \theta_i)x^i + \theta_iz^{i+1}.
   \]

Note that Algorithm 3 generates and keeps track of three points at each iteration: \( x^i, y^i, \) and \( z^i \). Observe that \( y^i \) is used only to compute the gradient \( \nabla f(y^i) \), which is then used to determine the value of \( z^{i+1} \) in Step 1, and then \( z^{i+1} \) is used to determine the value of \( x^{i+1} \) in Step 1. As the information about \( y^i \) is not used directly in subsequent iterations, the
explicit computation or storage of $y^i$ can be eliminated from the algorithm description, and indeed the algorithm could be re-stated only with the pair $x^i, z^i$ at each iteration.

The algorithm requires a sequence of step-size parameters $\theta_i$ that satisfies $\theta_0 = 1, \theta_i \in [0, 1]$ for $i = 0, 1, \ldots$. We will also require the values of $\theta_i$ to satisfy the following inequality that prevents the sequence from decreasing too quickly:

$$\frac{1 - \theta_{i+1}}{\theta_{i+1}^2} \leq \frac{1}{\theta_i^2} \quad \text{for } i = 0, 1, \ldots. \quad (6)$$

The following proposition presents and proves properties of two different step-size sequences that satisfy (6). Its proof is deferred to later.

**Proposition 3.1**

1. Consider the sequence

   $$\theta_i = \frac{2}{i + 2}, \quad i = 0, 1, \ldots. \quad (7)$$

   This sequence satisfies the step-size inequality (6) strictly.

2. Consider the sequence

   $$\theta_0 = 1, \quad \theta_{i+1} = \frac{2}{1 + \sqrt{1 + \frac{4}{\theta_i^2}}}, \quad i = 0, 1, \ldots. \quad (8)$$

   This sequence satisfies $\theta_i < \frac{2}{i+2}, \quad i = 0, 1, \ldots$ and satisfies the step-size inequality (6) at equality.

3.1 Analysis and Complexity of the Accelerated Simple Gradient Scheme

Our main algorithmic analysis result is:
Theorem 3.1 Suppose that the sequence \( \{\theta_i\}_{i=0}^{\infty} \) is given by (7) or (8). Then after \( k \) iterations of the accelerated simple gradient scheme, the following holds:

\[
\min_{0 \leq i \leq k} f(x^i) \leq f(x) + \frac{2L\|x - x^0\|^2}{(k + 1)^2} \text{ for all } x \in X.
\]

\[\square\]

Corollary 3.1 Suppose our region of interest is \( S_R := \{x : \|x - x^0\| \leq R\} \) for some \( R \), which may or may not contain an optimal solution \( x^\ast \). Then:

\[
\min_{0 \leq i \leq k} f(x^i) \leq \min_{x \in S_R} f(x) + \frac{2LR^2}{(k + 1)^2}.
\]

If we wish to compute a point whose objective function value is within \( \varepsilon \) of the optimal value \( f^\ast \), then we can do so within

\[k := \left\lceil \sqrt{\frac{2L\|x^\ast - x^0\|^2}{\varepsilon}} \right\rceil - 1\]

iterations. \[\square\]

In order to facilitate the proof of Theorem 3.1, we will utilize the following arithmetic result:

Proposition 3.2 For all \( x \) the following holds:

\[
\nabla f(y^i)^T(x - y^i) + \frac{\theta_i L}{2} \|x - z^i\|^2 - \frac{\theta_i L}{2} \|x - z^{i+1}\|^2 = \nabla f(y^i)^T(z^{i+1} - y^i) + \frac{\theta_i L}{2} \|z^{i+1} - z^i\|^2.
\]

Proof: This is just rearrangement of terms using the definition of \( z^{i+1} \):

\[z^{i+1} = z^i - \frac{1}{\theta_i L} \nabla f(y^i) .\]
Expanding the left side of the proposition we have:
\[
\nabla f(y^i)^T (x - y^i) + \frac{\theta_i}{2} \|x - z^i\|^2 - \frac{\theta_i L}{2} \|x - z^{i+1}\|^2
\]
\[
= -\nabla f(y^i) y^i + \theta_i L \left[ (z^i - z^{i+1})^T x + \frac{1}{2} \|x\|^2 - x^T z^i + \frac{1}{2} \|z^i\|^2 - \frac{1}{2} \|x\|^2 + x^T z^{i+1} - \frac{1}{2} \|z^{i+1}\|^2 \right]
\]
\[
= -\nabla f(y^i) y^i + \theta_i L \left[ \frac{1}{2} \|z^i\|^2 - \frac{1}{2} \|z^{i+1}\|^2 \right]
\]
\[
= \nabla f(y^i)^T (z^{i+1} - y^i) + \theta_i L \left[ (z^{i+1} - z^i)^T z^{i+1} + \frac{1}{2} \|z^i\|^2 - \frac{1}{2} \|z^{i+1}\|^2 \right]
\]
\[
= \nabla f(y^i)^T (z^{i+1} - y^i) + \theta_i L \left[ \|z^{i+1}\|^2 - (z^i)^T z^{i+1} + \frac{1}{2} \|z^i\|^2 - \frac{1}{2} \|z^{i+1}\|^2 \right]
\]
\[
= \nabla f(y^i)^T (z^{i+1} - y^i) + \frac{\theta_i L}{2} \|z^{i+1} - z^i\|^2 .
\]

We now introduce some facilitating notation. Let \( f_l(x, y) \) denote the first-order linear expansion of \( f(\cdot) \) about the point \( y \) evaluated at \( x \), namely:
\[
f_l(x, y) := f(y) + \nabla f(y)^T (x - y) . \tag{9}
\]

With this notation, the conclusion of Proposition 3.2 can be written equivalently as:
\[
f_l(x, y^i) + \frac{\theta_i L}{2} \|x - z^i\|^2 - \frac{\theta_i L}{2} \|x - z^{i+1}\|^2 = f_l(z^{i+1}, y^i) + \frac{\theta_i L}{2} \|z^{i+1} - z^i\|^2 . \tag{10}
\]

The next result will be used in the proof of Theorem 3.1.

**Lemma 3.1** If \( f(x) \leq f(x^{i+1}) \) or (6) holds at equality, then
\[
\frac{1 - \theta_{i+1}}{\theta_{i+1}} (f(x^{i+1}) - f(x)) + \frac{L}{2} \|x - z^{i+1}\|^2 \leq \frac{1 - \theta_i}{\theta_i} (f(x^i) - f(x)) + \frac{L}{2} \|x - z^i\|^2 .
\]
Proof: We have for all $x \in X$:
\[
\begin{align*}
f(x^{i+1}) &\leq f(y^i) + \nabla f(y^i)^T(x^{i+1} - y^i) + \frac{L}{2} \|x^{i+1} - y^i\|^2 \\
&= f(x^{i+1}, y^i) + \frac{L}{2} \|x^{i+1} - y^i\|^2 \\
&= f((1 - \theta_i)x^i + \theta_i z^{i+1}, y^i) + \frac{L}{2} \|(1 - \theta_i)x^i + \theta_i z^{i+1} - y^i\|^2 \\
&= (1 - \theta_i)f(x^i, y^i) + \theta_i f(z^{i+1}, y^i) + \frac{L}{2} \|(1 - \theta_i)x^i + \theta_i z^{i+1} - y^i\|^2 \\
&= (1 - \theta_i)f(x^i, y^i) + \theta_i f(z^{i+1}, y^i) + \frac{\theta^2 L}{2} \|z^{i+1} - z^i\|^2 \\
&= (1 - \theta_i)f(x^i, y^i) + \theta_i [f(x, y^i) + \theta_i \frac{L}{2} \|x - z^i\|^2 - \theta_i \frac{L}{2} \|x - z^{i+1}\|^2] \\
&\leq (1 - \theta_i)f(x^i) + \theta_i [f(x) + \theta_i \frac{L}{2} \|x - z^i\|^2 - \theta_i \frac{L}{2} \|x - z^{i+1}\|^2].
\end{align*}
\]
(Grad.-Ineq.)

Subtract $f(x)$ from each side, divide by $\theta_i^2$, and rearrange to yield:
\[
\frac{1}{\theta_i^2} (f(x^{i+1}) - f(x)) + \frac{L}{2} \|x - z^{i+1}\|^2 \leq \frac{1 - \theta_i}{\theta_i^2} (f(x^i) - f(x)) + \frac{L}{2} \|x - z^i\|^2. \quad (11)
\]

If $f(x) \leq f(x^{i+1})$, then (6) implies that
\[
\frac{1 - \theta_{i+1}}{\theta_i^2} (f(x^{i+1}) - f(x)) \leq \frac{1}{\theta_i^2} (f(x^{i+1}) - f(x))
\]
which combined with (11) proves the result.

If (6) is satisfied at equality, then
\[
\frac{1 - \theta_{i+1}}{\theta_i^2} = \frac{1}{\theta_i^2},
\]
which combined with (11) proves the result. □

Proof of Theorem 3.1: Let $x \in X$ be given. If $f(x) \geq \min_{0 \leq i \leq k} f(x^i)$ then the conclusion of the theorem follows trivially. Otherwise, $f(x) < \min_{0 \leq i \leq k} f(x^i)$, in which case from Lemma 3.1 we have:
\[
\frac{1 - \theta_{i+1}}{\theta_i^2} (f(x^{i+1}) - f(x)) + \frac{L}{2} \|x - z^{i+1}\|^2 \leq \frac{1 - \theta_i}{\theta_i^2} (f(x^i) - f(x)) + \frac{L}{2} \|x - z^i\|^2, \quad i = 0, \ldots, k-1.
\]

15
This cascading chain of inequalities can be combined for \( i = 1, \ldots, k - 2 \) to yield:

\[
\frac{1 - \theta_{k-1}}{\theta_{k-1}} \left( f(x^{k-1}) - f(x) \right) + \frac{L}{2} \|x - z^{k-1}\|^2 \leq \cdots \leq \frac{1 - \theta_0}{\theta_0} \left( f(x^0) - f(x) \right) + \frac{L}{2} \|x - z^0\|^2.
\]

Noting that \( \theta_0 = 1 \) and \( z^0 = x^0 \) yields:

\[
\frac{1 - \theta_{k-1}}{\theta_{k-1}} \left( f(x^{k-1}) - f(x) \right) + \frac{L}{2} \|x - z^{k-1}\|^2 \leq \frac{L}{2} \|x - x^0\|^2.
\]

Invoking (11) for \( i = k - 1 \) yields:

\[
\frac{1 - \theta_{k-1}}{\theta_{k-1}} \left( f(x^k) - f(x) \right) + \frac{L}{2} \|x - z^k\|^2 \leq \frac{1 - \theta_{k-1}}{\theta_{k-1}} \left( f(x^{k-1}) - f(x) \right) + \frac{L}{2} \|x - z^{k-1}\|^2,
\]

which when combined with the previous inequality yields:

\[
\frac{1}{\theta_{k-1}} \left( f(x^k) - f(x) \right) + \frac{L}{2} \|x - z^k\|^2 \leq \frac{L}{2} \|x - x^0\|^2.
\]

Lastly note from Proposition 3.1 that \( \theta_{k-1} \leq \frac{2}{k+1} \) and also noting that \( \|x - z^k\| \geq 0 \), we obtain:

\[
f(x^k) \leq f(x) + \theta_{k-1}^2 \frac{L}{2} \|x - x^0\|^2 \leq f(x) + \frac{4L}{(k+1)^2} \|x - x^0\|^2.
\]

**Proof of Proposition 3.1:** We first prove (1.). For the sequence (7) it follows trivially that \( \theta_0 = 1 \). For a given \( i \) we have

\[
1 - \theta_{i+1} = \frac{(3 + i)^2}{4} \left( 1 - \frac{2}{3 + i} \right) = \frac{i^2 + 4i + 3}{4} < \frac{i^2 + 4i + 4}{4} = \frac{(i + 2)^2}{2^2} = \frac{1}{\theta_i^2},
\]

which proves (1.). To prove (2.), we show that the formula for \( \theta_{i+1} \) in (8) is simply an application of the quadratic formula applied to (6) at equality. Indeed, given \( \theta_i \), then the equality version of (6) is

\[
\left( \frac{1}{\theta_{i+1}} \right)^2 - \left( \frac{1}{\theta_{i+1}} \right) - \left( \frac{1}{\theta_i} \right)^2 = 0,
\]

which is quadratic in \( \beta := \frac{1}{\theta_{i+1}} \). Invoking the quadratic formula to solve this equation yields precisely (8). Note also from (8) that \( \theta_{i+1} \) is monotone
increasing in $\theta_i$. By definition we have $\theta_0 = 1$ satisfies $\theta_i \leq \frac{2}{i+1}$ at $i = 0$. Suppose that $\theta_i \leq \frac{2}{i+1}$ holds for some $i$. Then from monotonicity we have

$$\theta_{i+1} = \frac{2}{1 + \sqrt{1 + \frac{4}{\theta_i^2}}} \leq \frac{2}{1 + \sqrt{1 + \frac{4}{(\frac{2}{i+1})^2}}} < \frac{2}{1 + \sqrt{\frac{4}{(\frac{2}{i+1})^2}}} = \frac{2}{2 + i + 1},$$

and the proof is completed by induction. ■

3.2 Comments and Extensions

1. How might the scheme and the analysis be modified if $L$ is not explicitly known?

2. One might ask whether the complexity bound in Theorem 3.1 can be improved. In a rather deep result, it was shown by Nemirovskii and Yudin that this bound is “order optimal” in the sense that no algorithm that only relies on first-order information can have a better worst-case complexity bound except perhaps by an absolute constant. In this sense the accelerated simple gradient method is an optimal algorithm.

4 Greedy Coordinate Descent Method for Smooth Optimization

4.1 Problem Setting and Basics for Coordinate Descent Method

As in Section 1, our problem of interest is

$$P: \quad \text{minimize}_x \quad f(x)$$

$$\text{s.t.} \quad x \in \mathbb{R}^n,$$

where $f(\cdot)$ is a differentiable convex function defined on $\mathbb{R}^n$. We denote the optimal objective value of $P$ by $f^*$, and let $x^*$ denote an optimal solution of $P$, when such a solution exists.
Let $\nabla f(x)$ denote the gradient of $f(\cdot)$ at $x$, and recall the gradient inequality:

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) \quad \text{for all } x, y \in \mathbb{R}^n .$$

(13)

We will measure distances between points using the $\ell_1$ norm $\|x\|_1 := \sum_{j=1}^{n} |x_j|$.

We will measure the size of the gradients with the $\ell_\infty$ norm $\|g\|_\infty := \max_j \{|g_j|\}$.

We assume that $f(\cdot)$ has a Lipschitz gradient. Given our use of the above norms, this means that there is a constant $L$ for which:

$$\|\nabla f(y) - \nabla f(x)\|_\infty \leq L\|y - x\|_1 \quad \text{for all } x, y \in \mathbb{R}^n .$$

(14)

Let $B_1(c, R)$ denote the $\ell_1$ ball centered at $c$ with radius $R$, namely:

$$B_1(c, R) := \{ x \in \mathbb{R}^n : \|x - c\|_1 \leq R \} .$$

Algorithm 4 presents the greedy coordinate descent scheme for solving (P).

4.2 Analysis and Complexity of the Greedy Coordinate Descent Scheme

The following is a simple variant of Proposition 1.1 adapted here for our different choice of norm:

**Proposition 4.1** If $f(\cdot)$ has Lipschitz gradient with constant $L$, then

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|_1^2 \quad \text{for all } x, y \in \mathbb{R}^n .$$

18
Algorithm 4 Greedy Coordinate Descent Scheme

Initialize. Initialize at $x^0$, $i ← 0$

At iteration $i$:

1. **Compute Gradient.** Compute $\nabla f(x^i)$

2. **Compute Coordinate.** Compute:

   $$j_i ∈ \arg \max_{j ∈ \{1,...,n\}} \{|\nabla f(x^k)_j|\},$$

   and

   $$s_i ← \text{sgn}(\nabla f(x^i)_{j_i})$$

3. **Compute Step-Size and New Point.** Compute:

   $$\alpha_i ← -s_i \frac{\nabla f(x^i)_{j_i}}{L}$$

   $$x^{i+1} ← x^i + \alpha_i e_{j_i}$$
Proof: We use the fundamental theorem of calculus applied in a multivariate setting. We have
\[
    f(y) = f(x) + \int_0^1 \nabla f(x + t(y - x))^T (y - x) dt
\]
\[
    = f(x) + \nabla f(x)^T (y - x) + \int_0^1 [\nabla f(x + t(y - x)) - \nabla f(x)]^T (y - x) dt
\]
\[
    \leq f(x) + \nabla f(x)^T (y - x) + \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\|_\infty \|(y - x)\|_1 dt
\]
\[
    \leq f(x) + \nabla f(x)^T (y - x) + \int_0^1 tL \|(y - x)\|_1^2 dt
\]
\[
    = f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|(y - x)\|_1^2 . \]

We will also use the following property of a nonnegative series, whose proof is given at the end of this subsection.

Proposition 4.2 Suppose there is a constant \( C > 0 \) for which the nonnegative series \( \{a_i\} \) satisfies
\[
a_{i+1} \leq a_i - \frac{a_i^2}{C} \quad \text{for all } i \geq 0 .
\]
Then it holds that
\[
a_k \leq \frac{1}{\frac{1}{a_0} + \frac{k}{C}} \quad \text{for all } k \geq 0 .
\]

Let \( S_0 \) denote the set of all \( x \) whose objective value is at most \( f(x^0) \), namely
\[
    S_0 := \{ x \in \mathbb{R}^n : f(x) \leq f(x^0) \},
\]
and let \( S^* \) denote the set of optimal solutions of (12), i.e., \( S^* := \{ x \in \mathbb{R}^n : f(x) = f^* \} \). Now let Dist\(_0\) denote the largest distance of points in \( S_0 \) to an optimal solution:
\[
    \text{Dist}_0 := \max_{x \in S_0} \left\{ \min_{x^* \in S^*} \| x - x^* \|_1 \right\} . \quad (15)
\]
Our main algorithmic analysis result for the greedy coordinate descent scheme is:
Theorem 4.1 Let \( \{x^k\} \) be generated according to the greedy coordinate descent scheme (Algorithm 4). Then for all \( k \geq 0 \) it holds that:

\[
f(x^k) - f^* \leq \frac{1}{\frac{1}{f(x^0) - f^*} + \frac{k}{2L(Dist_0)^2}} \leq \frac{2L(Dist_0)^2}{k} .
\] (16)

Proof: From Proposition 4.1 we have for each \( i \geq 0 \):

\[
f(x^{i+1}) \leq f(x^i) + \nabla f(x^i)^T (x^{i+1} - x^i) + \frac{L}{2} \|x^{i+1} - x^i\|_1^2
\]

\[
= f(x^i) - \alpha_i |\nabla f(x^i)|_j + \frac{L}{2} \alpha_i^2
\]

\[
= f(x^i) - \alpha_i \|\nabla f(x^i)\|_{\infty} + \frac{L}{2} \alpha_i^2
\]

\[
= f(x^i) - \frac{1}{2L} \|\nabla f(x^i)\|_{\infty}^2 .
\] (17)

This shows that the values \( f(x^i) \) are decreasing and in particular \( f(x^i) \leq f(x^0) \), whereby \( x^i \in S_0 \) for all \( i \geq 0 \). Therefore for any \( i \geq 0 \) there exists \( x^* \in S^* \) for which \( \|x^i - x^*\|_1 \leq Dist_0 \), and from the gradient inequality for the convex function \( f(\cdot) \) it holds that

\[
f^* = f(x^*) \geq f(x^i) + \nabla f(x^i)^T (x^* - x^i) \quad \text{(from Gradient Inequality)}
\]

\[
\geq f(x^i) - \|\nabla f(x^i)\|_{\infty} \|x^* - x^i\|_1 \quad \text{(from Norm Inequality)}
\]

\[
\geq f(x^i) - \|\nabla f(x^i)\|_{\infty} Dist_0 , \quad \text{(from (15))}
\]

and rearranging the above yields \( \|\nabla f(x^i)\|_{\infty} \geq \frac{f(x^i) - f^*}{f(x^i) - f^*} \). Substituting this inequality into (17) and subtracting \( f^* \) from both sides yields:

\[
f(x^{i+1}) - f^* \leq f(x^i) - f^* - \frac{(f(x^i) - f^*)^2}{2LDist_0^2} .
\]

Define \( a_i := f(x^i) - f^* \), and it follows that the nonnegative series \( \{a_i\} \) satisfies \( a_{i+1} \leq a_i - \frac{a_i^2}{2LDist_0^2} \). It then follows from Proposition 4.2 using \( C = 2L(Dist_0)^2 \) that

\[
a_k \leq \frac{1}{a_0} + \frac{k}{2LDist_0^2} ,
\]
which establishes that:

\[ f(x^k) - f^* \leq \frac{1}{f(x^0) - f^*} + \frac{k}{2L(Dist_0)^2}. \]

\[ \square \]

**Proof of Proposition 4.2** Notice that the conclusion of the proposition holds for \( k = 0 \). By induction, suppose that the conclusion is true for some \( k \geq 0 \). Then:

\[ a_{k+1} \leq a_k - \frac{a_k^2}{C} \]
\[ \leq a_k - \frac{a_k a_{k+1}}{C} , \]

and collecting terms and rearranging yields:

\[ a_{k+1} \left(1 + \frac{a_k}{C}\right) \leq a_k . \]

Taking reciprocals and rearranging yields:

\[ \frac{1}{a_{k+1}} \geq \frac{1}{a_k} \left(1 + \frac{a_k}{C}\right) \]
\[ = \frac{1}{a_k} + \frac{1}{C} \]
\[ \geq \frac{1}{a_0} + \frac{k}{C} + \frac{1}{C} \]
\[ = \frac{1}{a_0} + \frac{k + 1}{C} , \]

which rearranges to yield:

\[ a_{k+1} \leq \frac{1}{\frac{1}{a_0} + \frac{k+1}{C}} , \]

completing the proof. \( \square \)
4.3 Comments and Extensions

1. How might the scheme and the analysis be modified if \( L \) is not explicitly known?

2. How might the scheme and the analysis be modified if one can efficiently do a linesearch to compute:

\[
\alpha_i := \arg \min_{\alpha} \{ f(x^i + \alpha e^j) \},
\]

instead of assigning the value \( \alpha_i \leftarrow -\frac{\nabla f(x^i)_{ji}}{L} (\text{sgn}(\nabla f(x^k)_{ji})) \) at each iteration?

5 The Frank-Wolfe Method for Constrained Optimization

Let us now consider the following constrained optimization problem:

\[
P: \quad \text{minimize}_x \ f(x)
\]

\[
s.t. \quad x \in S,
\]

where \( f(x) \) is a differentiable convex function on \( S \), and \( S \) is a closed and bounded convex set. Let \( f^* \) denote the optimal objective value of \( P \). We assume that \( f(\cdot) \) has a Lipschitz gradient on \( S \), which means there exists \( L \) for which:

\[
\| \nabla f(y) - \nabla f(x) \| \leq L \| y - x \| \quad \text{for all} \ x, y \in S. \quad (18)
\]

(Throughout this Section the norm is the Euclidean norm \( \| v \| = \sqrt{v^T v} \).)

We now describe the Frank-Wolfe method for solving \( P \). This method is based on the premise that the set \( S \) is well-suited for linear optimization. This means that either \( S \) is itself a system of linear inequalities
$S = \{ x \mid Ax \leq b \}$ over which linear optimization is an easy task, or more generally that the problem:

\[
LO_c : \quad \text{minimize} \quad c^T x \\
\text{s.t.} \quad x \in S
\]

is easy to solve for any given objective function vector $c$. This being the case, suppose that we have a given iterate value $x^k \in S$. The linearization of the function $f(x)$ at $x = x^k$ is:

\[
z_1(x) := f(x^k) + \nabla f(x^k)^T (x - x^k),
\]

which is the first-order Taylor expansion of $f(\cdot)$ at $x^k$. Since we can easily do linear optimization on $S$, let us solve:

\[
LP : \quad \text{minimize} \quad f(x^k) + \nabla f(x^k)^T (x - x^k) \\
\text{s.t.} \quad x \in S,
\]

which of course is equivalent to:

\[
LP : \quad \text{minimize} \quad \nabla f(x^k)^T x \\
\text{s.t.} \quad x \in S.
\]

Let $\tilde{x}^k$ denote the optimal solution to this problem. The Frank-Wolfe method proceeds by choosing the next iterate as $x^{k+1} \leftarrow x^k + \alpha (\tilde{x}^k - x^k)$ for some step-size $\alpha$. Since $S$ is a convex set and both $x^k$ and $\tilde{x}^k$ are contained in $S$, then $x^k + \alpha (\tilde{x}^k - x^k) \in S$ for all $\alpha \in [0, 1]$. Let $\bar{\alpha}_k$ denote the step-size at iteration $k$ of the method. We have:

\[
x^{k+1} \leftarrow x^k + \bar{\alpha}_k (\tilde{x}^k - x^k) \quad \text{for some } \bar{\alpha}_k \in [0, 1].
\]
The value of $\bar{\alpha}_k$ can be determined in a number of different ways. One way to set the step-size $\bar{\alpha}_k$ is according to some pre-determined rule, such as the following often-used rule:

$$\bar{\alpha}_k = \frac{2}{k + 2} .$$

(We will see shortly that this rule leads to a very good computational bound on the convergence of the Frank-Wolfe method.) Another way is to choose $\bar{\alpha}_k$ by performing a line-search of $f(\cdot)$ over the interval $\alpha \in [0, 1]$. That is, we might determine $\bar{\alpha}_k$ as the optimal solution to the following line-search problem:

$$\bar{\alpha}_k \leftarrow \arg\min_{\alpha} f(x^k + \alpha(\tilde{x}^k - x^k))$$

s.t. $0 \leq \alpha \leq 1$.

Algorithm 5 presents a formal statement of the Frank-Wolfe method just described.

**Algorithm 5** Frank-Wolfe Method for minimizing $f(x)$ over $x \in S$

Initialize at $x^0 \in S$, $k \leftarrow 0$.

At iteration $k$:

1. Compute $\nabla f(x^k)$, and then solve linear optimization problem:

$$\tilde{x}^k \leftarrow \arg\min_{x \in S} \{ f(x^k) + \nabla f(x^k)^T (x - x^k) \} .$$

2. Set $x^{k+1} \leftarrow x^k + \bar{\alpha}_k (\tilde{x}^k - x^k)$, where $\bar{\alpha}_k \in [0, 1]$.
5.1 Analysis and Complexity of the Frank-Wolfe Method

Before stating the main algorithmic analysis result for the Frank-Wolfe method, we first present two important preliminary results. The first result states that at each iteration \( k \) of the Frank-Wolfe method, the solution of the linear optimization problem leads to a valid lower bound on the optimal value \( f^* \) of the optimization problem \( P \).

**Proposition 5.1** At iteration \( k \) of the Frank-Wolfe method (Algorithm 5), it holds that

\[
f^* \geq f(x^k) + \nabla f(x^k)^T (\tilde{x}^k - x^k).
\]

**Proof:** From the gradient inequality for convex functions one has:

\[
f(x) \geq f(x^k) + \nabla f(x^k)^T (x - x^k) \quad \text{for any } x \in S.
\]

Therefore:

\[
f^* = \min_{x \in S} f(x) \geq \min_{x \in S} f(x^k) + \nabla f(x^k)^T (x - x^k) = f(x^k) + \nabla f(x^k)^T (\tilde{x}^k - x^k).
\]

\[
\]

We will also use the following property of certain types of nonnegative series', whose proof is given at the end of this subsection.

**Proposition 5.2** Suppose there is a constant \( C > 0 \) for which the nonnegative series \( \{a_i\} \) satisfies

\[
a_{i+1} \leq a_i \left(1 - \frac{2}{i + 2}\right) + \frac{C}{(i + 2)^2} \quad \text{for all } i \geq 0.
\]

Then it holds that

\[
a_k \leq \frac{C}{k + 2} \quad \text{for all } k \geq 1.
\]
Let Diam(S) denote the largest distance between any two points in S, namely:

\[
\text{Diam}(S) := \max_{x, y \in S} \{\|x - y\|\}.
\] (19)

Our main algorithmic analysis result for the Frank-Wolfe method is:

**Theorem 5.1** Let \(\{x^k\}\) be generated according to the Frank-Wolfe method (Algorithm 5) using the step-size rule

\[
\bar{\alpha}_i = \frac{2}{i + 2}.
\]

Then for all \(k \geq 1\) it holds that:

\[
f(x^k) - f^* \leq \frac{2L(\text{Diam}(S))^2}{k + 2}.
\] (20)

**Proof:** For every iteration \(k \geq 0\) it holds that:

\[
f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^T (x^{k+1} - x^k) + \frac{L}{2} \|x^{k+1} - x^k\|^2 \quad \text{(from Prop. 1.1)}
\]

\[
= f(x^k) + \bar{\alpha}_k \nabla f(x^k)^T (\bar{x}^k - x^k) + \frac{L\bar{\alpha}_k^2}{2} \|\bar{x}^k - x^k\|^2
\]

\[
\leq f(x^k) + \bar{\alpha}_k \nabla f(x^k)^T (\bar{x}^k - x^k) + \frac{L\bar{\alpha}_k^2 (\text{Diam}(S))^2}{2}
\]

\[
= (1 - \bar{\alpha}_k) f(x^k) + \bar{\alpha}_k (f(x^k) + \nabla f(x^k)^T (x^k - x^k)) + \frac{L\bar{\alpha}_k^2 (\text{Diam}(S))^2}{2}
\]

\[
\leq (1 - \bar{\alpha}_k) f(x^k) + \bar{\alpha}_k f^* + \frac{L\bar{\alpha}_k^2 (\text{Diam}(S))^2}{2}. \quad \text{(from Prop. 5.1)}
\]

Subtracting \(f^*\) from each side and rearranging the right-hand side, one obtains:

\[
f(x^{k+1}) - f^* \leq (f(x^k) - f^*) (1 - \bar{\alpha}_k) + \frac{L\bar{\alpha}_k^2 (\text{Diam}(S))^2}{2},
\]

27
and substituting the value of $\bar{\alpha}_k = \frac{2}{k+2}$ we arrive at:

$$f(x^{k+1}) - f^* \leq (f(x^k) - f^*)(1 - \frac{2}{k+2}) + \frac{2L(Diam(S))^2}{(k+2)^2}.$$ 

Define $a_k := f(x^k) - f^*$, and define $C := 2L(Diam(S))^2$, and notice that the above inequality is:

$$a_{k+1} \leq a_k \left(1 - \frac{2}{k+2}\right) + \frac{C}{(k+2)^2} \quad \text{for all } k \geq 0.$$ 

Notice that the series $\{a_k\}$ satisfies the condition of Proposition 5.2. It therefore follows from Proposition 5.2 that $a_k \leq \frac{C}{k+2}$ for $k \geq 1$. Therefore:

$$f(x^k) - f^* = a_k \leq \frac{C}{k+2} = \frac{2L(Diam(S))^2}{k+2}.$$ 

Proof of Proposition 5.2 Notice that the conclusion of the proposition holds for $k = 1$, since using $i = 0$ we have:

$$a_1 = a_{i+1} \leq a_i \left(1 - \frac{2}{i+2}\right) + \frac{C}{(i+2)^2}$$

$$= a_0 \left(1 - \frac{2}{0+2}\right) + \frac{C}{(0+2)^2}$$

$$= \frac{C}{4} \leq \frac{C}{3} = \frac{C}{k+2}.$$ 

28
By induction, suppose that the conclusion is true for some \( k \geq 0 \), namely \( a_k \leq \frac{C}{k+2} \). Then:

\[
\begin{align*}
a_{k+1} & \leq a_k \left(1 - \frac{2}{k+2}\right) + \frac{C}{(k+2)^2} \\
& \leq \frac{C}{k+2} \left(1 - \frac{2}{k+2}\right) + \frac{C}{(k+2)^2} \\
& = \frac{C(k+1)}{(k+2)^2} \\
& < \frac{C}{(k+3)} ,
\end{align*}
\]

where the last inequality follows since \((k+1)(k+3) < (k+2)^2\). The result then follows by induction. ■

### 5.2 Comments and Extensions

1. How might the Frank-Wolfe method and its analysis be modified if \( L \) is not explicitly known?

2. How might the scheme and the analysis be modified if one can efficiently do a linesearch to compute:

\[
\bar{\alpha}_k := \arg \min_{\alpha \in [0,1]} \{ f(x^k + \alpha(x^k - x^k)) \} ,
\]

instead of assigning the value \( \alpha_k = \frac{2}{k+2} \) at each iteration?