3. Optimal gradient methods

- lower complexity bounds
- estimate sequence
- optimal gradient methods
Lower complexity bound for smooth convex optimization

computational model

- problem formulation: \( \min_{x \in \mathbb{R}^n} f(x) \)
- problem class: \( f \) is convex and \( | \nabla f(x) - \nabla f(y) | \leq L \| x - y \|_2 \)
- oracle: first-order local black box
- approximate solution: find \( \bar{x} \) such that \( f(\bar{x}) - f^* \leq \epsilon \)

assumption: iterative algorithm generates a sequence \( \{x^{(k)}\} \) such that

\[
x^{(k)} \in x^{(0)} + \text{span} \left\{ \nabla f(x^{(0)}), \nabla f(x^{(1)}), \ldots, \nabla f(x^{(k-1)}) \right\}
\]

theorem (Nesterov): for any integer \( k \leq (n - 1)/2 \) and any \( x^{(0)} \), there exist a function in the problem class such that

\[
f(x^{(k)}) - f^* \geq \frac{3L \| x^{(0)} - x^* \|_2^2}{32(k + 1)^2}
\]
**proof:** consider the quadratic function

\[
f(x) = \frac{L}{4} \left( \frac{1}{2} \left( x_1^2 + \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 + x_n^2 \right) - x_1 \right)
\]

which can be expressed as \( f(x) = \frac{L}{4} \left( \frac{1}{2} x^T A x - e_1^T x \right) \), where

\[
A = \begin{bmatrix}
    2 & -1 & 0 \\
    -1 & 2 & -1 & 0 \\
    0 & -1 & 2 & -1 \\
    & & & \ddots & \ddots & \ddots & \ddots \\
    & & & & & & -1 & 2 & -1 \\
    & & & & & & 0 & -1 & 2
\end{bmatrix}, \quad e_1 = \begin{bmatrix}
    1 \\
    0 \\
    0 \\
    \vdots \\
    0 \\
    0
\end{bmatrix}
\]

- \( 0 \leq \nabla^2 f(x) \leq L \implies f \) is convex and \( \nabla f(x) \) is \( L \)-Lipschitz continuous
- optimal solution \( x_i^* = 1 - \frac{i}{n+1} \) for \( i = 1, \ldots, n \) (by solving \( A x^* = e_1 \))

\[
\|x^*\|_2^2 = \frac{1}{(n+1)^2} (n^2 + \cdots + 1^2) \leq \frac{1}{3} (n+1)
\]

- optimal value: \( f(x^*) = \frac{L}{4} \left( \frac{1}{2} x^*^T A x^* - e_1^T x^* \right) = -\frac{L}{8} e_1^T x^* = -\frac{L}{8} \frac{n}{n+1} \)
without loss of generality, let \( x^{(0)} = 0 \); by the tri-diagonal form of \( A \),

\[
\nabla f(x^{(0)}) = -\frac{L}{4} e_1 \implies x^{(1)} \in \text{span}\{e_1\}
\]

\[
\implies \nabla f(x^{(1)}) \in \text{span}\{e_1, e_2\} \implies x^{(2)} \in \text{span}\{e_1, e_2\}
\]

\[
\vdots \implies x^{(k)} \in \text{span}\{e_1, \ldots, e_k\}
\]

therefore

\[
f(x^{(k)}) \geq \inf_{x^{(k+1)} = \ldots = x^{(n)} = 0} f(x) = -\frac{L}{8} \frac{k}{k+1}
\]

for \( k \approx n/2 \) or \( n = 2k + 1 \)

\[
f(x^{(k)}) - f^* \geq -\frac{L}{8} \frac{k}{k+1} + \frac{L}{8} \frac{n}{n+1} \geq \frac{L}{16(k+1)}
\]

finally

\[
\frac{f(x^{(k)}) - f^*}{\|x^{(0)} - x^*\|_2^2} \geq \frac{L}{16(k+1)} \left/ \frac{2k + 2}{3} \right. = \frac{3L}{32(k+1)^2}
\]

Optimal gradient methods
Lower complexity bound for $S_{\mu,L}(\mathbb{R}^\infty)$

computational model

- formulation: minimize $x \in \ell_2 f(x)$, where $\ell_2 = \{x \in \mathbb{R}^\infty | \sum_{i=1}^{\infty} x_i^2 \leq \infty\}$
- problem class: $f$ is $\mu$-strongly convex and $|\nabla f(x) - \nabla f(y)| \leq L \|x - y\|_2$
- oracle: first-order local black box
- approximate solution: find $\bar{x}$ such that $f(\bar{x}) - f^* \leq \epsilon$

assumption: iterative algorithm generates a sequence $\{x^{(k)}\}$ such that

$$x^{(k)} \in x^{(0)} + \text{span}\left\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \ldots, \nabla f(x^{(k-1)})\right\}$$

theorem (Nesterov): for any constants $\mu > 0$ and $\kappa \triangleq L/\mu > 1$, and any $x^{(0)} \in \ell_2$, there exist a function in the problem class such that

$$f(x^{(k)}) - f^* \geq \frac{\mu}{2} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{2k} \|x^{(0)} - x^*\|_2^2$$
proof: consider the quadratic function

\[ f(x) = \frac{\mu(\kappa - 1)}{4} \left( \frac{1}{2} \left( x_1^2 + \sum_{i=1}^{\infty} (x_i - x_{i+1})^2 \right) - x_1 \right) + \frac{\mu}{2} \|x\|^2 \]

which can be expressed as \( f(x) = \frac{\mu(\kappa - 1)}{4} \left( \frac{1}{2} x^T A x - e_1^T x \right) + \frac{\mu}{2} \|x\|^2 \), where

\[
A = \begin{bmatrix}
2 & -1 & 0 \\
-1 & 2 & -1 & 0 \\
0 & -1 & 2 & -1 & 0 \\
& & & & \\
& & & & \\
& & & & \\
\end{bmatrix}, \quad e_1 = \begin{bmatrix}
1 \\
0 \\
0 \\
0 \\
\vdots
\end{bmatrix}
\]

- \( 0 \preceq A \preceq 4I \implies \mu I \preceq \nabla^2 f(x) \preceq LI \)
- first-order optimality condition: \( \nabla f(x^*) = 0 \implies \left( A + \frac{4}{\kappa - 1} \right) x^* = e_1 \)

\[
x_i^* = q^i, \quad i = 1, 2, \ldots \quad \text{where} \quad q = \frac{\sqrt{\kappa - 1}}{\sqrt{\kappa + 1}}
\]

therefore

\[
\|x^*\|^2 = \sum_{i=1}^{\infty} x_i^{*2} = \sum_{i=1}^{\infty} q^{2i} = \frac{q^2}{1 - q^2}
\]
without loss of generality, let \( x^{(0)} = 0 \); by the tri-diagonal form of \( A \),

\[
\nabla f(x^{(0)}) = -\frac{L}{4}e_1 \quad \implies \quad x^{(1)} \in \text{span}\{e_1\}
\]

\[
\implies \nabla f(x^{(1)}) \in \text{span}\{e_1, e_2\} \quad \implies \quad x^{(2)} \in \text{span}\{e_1, e_2\}
\]

\[
\cdots \quad \implies \quad x^{(k)} \in \text{span}\{e_1, \ldots, e_k\}
\]

therefore

\[
\|x^{(k)} - x^*\|^2 \geq \sum_{i=k+1}^{\infty} x_i^2 = \sum_{i=k+1}^{\infty} q^{2i} = \frac{q^{2(k+1)}}{1 - q^2} = q^{2k}\|x^{(0)} - x^*\|^2
\]

by strong convexity with parameter \( \mu \),

\[
f(x^{(k)}) - f^* \geq \frac{\mu}{2}\|x^{(k)} - x^*\|^2 \geq \frac{\mu}{2}q^{2k}\|x^{(0)} - x^*\|^2
\]
Complexity of the gradient method

gradient method is not optimal

- for smooth convex functions \((L\text{-Lipschitz gradient})\)
  \[
  f(x^{(k)}) - f^* \leq \frac{L}{2k} \|x^{(0)} - x^*\|^2
  \]

- for strongly convex and smooth functions
  \[
  f(x^{(k)}) - f^* \leq \frac{L}{2} \left(\frac{L - \mu}{L + \mu}\right)^{2k} \|x^{(0)} - x^*\|^2
  \]

Nesterov’s comments:

- gradient method relied on decreasing objective values (“relaxation”):
  \[
  f(x^{(k+1)}) \leq f(x^{(k)})
  \]

- optimal methods: don’t rely on relaxation (too “microscopic” of a property); use some global topological properties of convex functions
Estimate sequence (Nesterov)

A pair of sequences \( \{\lambda_k, \phi_k(x)\}_{k=0}^{\infty} \) is called estimate sequence of \( f(x) \) if

- \( \lambda_k \to 0 \)
- \( \phi_k(x) \leq (1 - \lambda_k)f(x) + \lambda_k\phi_0(x) \) for any \( x \in \mathbb{R}^n \) and all \( k > 0 \)

**Lemma:** If a sequence \( \{x^{(k)}\} \) satisfies \( f(x^{(k)}) \leq \min_{x \in \mathbb{R}^n} \phi_k(x) \), then

\[
f(x^{(k)}) - f^* \leq \lambda_k (\phi_0(x^*) - f^*) \to 0
\]

**Proof:**

\[
f(x^{(k)}) \leq \min_{x \in \mathbb{R}^n} \phi_k(x) \leq \min_{x \in \mathbb{R}^n} \{(1 - \lambda_k)f(x) + \lambda_k\phi_0(x)\}
\leq (1 - \lambda_k)f(x^*) + \lambda_k\phi_0(x^*)
= f(x^*) + \lambda_k (\phi_0(x^*) - f(x^*))
\]
**estimate sequence:** pair of sequences \( \{ \lambda_k, \phi_k(x) \}_{k=0}^{\infty} \) such that

- \( \lambda_k \to 0 \)
- \( \phi_k(x) \leq (1 - \lambda_k)f(x) + \lambda_k \phi_0(x) \) for any \( x \in \mathbb{R}^n \) and all \( k > 0 \)

**questions:**

- how to form the estimate sequence?
- how can we ensure \( f(x^{(k)}) \leq \phi_k^* \triangleq \min_{x \in \mathbb{R}^n} \phi_k(x) \)?
**Lemma:** Suppose $f \in S_{\mu,L}(\mathbb{R}^n)$, then for any function $\phi_0(x)$, any sequence $\{y^{(k)}\}_{k=1}^{\infty}$, and $\{\alpha_k\}_{k=0}^{\infty}$ that satisfies

$$\alpha_k \in (0, 1), \quad \sum_{k=1}^{\infty} \alpha_k = \infty$$

the following pair is an estimate sequence

$$\lambda_{k+1} = (1 - \alpha_k) \lambda_k, \quad \text{with} \quad \lambda_0 = 1$$

$$\phi_{k+1}(x) = (1 - \alpha_k) \phi_k(x) + \alpha_k \left( f(y^{(k)}) + \left\langle \nabla f(y^{(k)}), x - y^{(k)} \right\rangle + \frac{\mu}{2} \|x - y^{(k)}\|^2 \right)$$

**Proof:** Note $\phi_0(x) \leq (1 - \lambda_0) f(x) + \lambda_0 \phi_0(x) = \phi_0(x)$; use induction

$$\phi_{k+1}(x) \leq (1 - \alpha_k) \phi_k(x) + \alpha_k f(x)$$

$$= (1 - (1 - \alpha_k) \lambda_k) f(x) + (1 - \alpha_k) \left( \phi_k(x) - (1 - \lambda_k) f(x) \right)$$

$$\leq (1 - (1 - \alpha_k) \lambda_k) f(x) + (1 - \alpha_k) \lambda_k \phi_0(x)$$

$$= (1 - \lambda_{k+1}) f(x) + \lambda_{k+1} \phi_0(x)$$

Optimal gradient methods
\[ \lambda_k = \lambda_0 \prod_{i=0}^{k} (1 - \alpha_i) \to 0 \text{ due to the fact} \]

\[ \alpha_k \in (0, 1), \quad \sum_{k=1}^{\infty} \alpha_k = \infty \quad \implies \quad \prod_{k=0}^{\infty} (1 - \alpha_k) \to 0 \]

proof:

- \( \{\lambda_k\}_{k=0}^{\infty} \) monotone decreasing and bounded below, so has limit

- suppose \( \lambda_k \to c > 0 \)

- rewrite iteration as \( \lambda_k - \lambda_{k+1} = \alpha_k \lambda_k \), and sum over \( k = 0, \ldots, N \)

\[ \lambda_0 - \lambda_{N+1} = \sum_{k=1}^{N} \alpha_k \lambda_k \geq c \sum_{k=1}^{N} \alpha_k \]

contradiction when \( N \to \infty \), so need to have \( c = 0 \)
Update quadratic approximations

let $\phi_0(x) = \phi^*_0 + \frac{\gamma_0}{2} \|x - v_0\|^2$, then $\{\phi_k(x)\}$ on page 3–11 can be written as

$$\phi_k(x) = \phi^*_k + \frac{\gamma_k}{2} \|x - v^{(k)}\|^2,$$

where

$$\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k \mu$$
$$v^{(k+1)} = \frac{1}{\gamma_{k+1}} \left( (1 - \alpha_k)\gamma_k v^{(k)} + \alpha_k \mu y^{(k)} - \alpha_k \nabla f(y^{(k)}) \right)$$
$$\phi^*_{k+1} = (1 - \alpha_k)\phi^*_k + \alpha_k f(y^{(k)}) - \frac{\alpha_k^2}{2\gamma_{k+1}} \|\nabla f(y^{(k)})\|^2$$
$$+ \frac{\alpha_k(1 - \alpha_k)\gamma_k}{\gamma_{k+1}} \left( \langle \nabla f(y^{(k)}), v^{(k)} - y^{(k)} \rangle + \frac{\mu}{2} \|y^{(k)} - v^{(k)}\|^2 \right)$$

(manipulations of simple quadratic functions)
assume we already have $\phi_k^* \geq f(x^{(k)})$, then

$$\phi_{k+1}^* \geq (1 - \alpha_k) f(x^{(k)}) + \alpha_k f(y^{(k)}) - \frac{\alpha_k^2}{2\gamma_{k+1}} \|\nabla f(y^{(k)})\|^2$$

$$+ \frac{\alpha_k (1 - \alpha_k) \gamma_k}{\gamma_{k+1}} \langle \nabla f(y^{(k)}), u^{(k)} - y^{(k)} \rangle$$

by convexity, $f(x^{(k)}) \geq f(y^{(k)}) + \langle \nabla f(y^{(k)}), x^{(k)} - y^{(k)} \rangle$,

$$\phi_{k+1}^* \geq f(y^{(k)}) - \frac{\alpha_k^2}{2\gamma_{k+1}} \|\nabla f(y^{(k)})\|^2$$

$$+ (1 - \alpha_k) \left\langle \nabla f(y^{(k)}), \frac{\alpha_k \gamma_k}{\gamma_{k+1}} (v^{(k)} - y^{(k)}) + x^{(k)} - y^{(k)} \right\rangle$$

finally, in order to make $\phi_{k+1}^* \geq f(x^{(k+1)})$,

- choose $x^{(k+1)}$ such that $f(x^{(k+1)}) \leq f(y^{(k)}) - \frac{\alpha_k^2}{2\gamma_{k+1}} \|\nabla f(y^{(k)})\|^2$

- choose $y^{(k)}$ so that $\frac{\alpha_k \gamma_k}{\gamma_{k+1}} (v^{(k)} - y^{(k)}) + x^{(k)} - y^{(k)} = 0$
Choose \( \{y^{(k)}\} \) and \( \{x^{(k+1)}\} \)

- choose \( y^{(k)} \) to eliminate inner-product term

\[
y^{(k)} = \frac{1}{\gamma_k + \alpha_k \mu} \left( \alpha_k \gamma_k v^{(k)} + \gamma_k + 1 x^{(k)} \right)
\]

- recall from quadratic upper bound (page 2-6):

\[
f \left( y - \frac{1}{L} \nabla f(y) \right) \leq f(y) - \frac{1}{2L} \| \nabla f(y) \|_2^2
\]

so we can let

\[
x^{(k+1)} = y^{(k)} - \frac{1}{L} \nabla f(y^{(k)})
\]

and solve for \( \alpha_k \) from the equation \( \frac{\alpha_k^2}{\gamma_{k+1}} = \frac{1}{L} \), that is,

\[
L \alpha_k^2 = (1 - \alpha_k) \gamma_k + \alpha_k \mu
\]
General scheme of optimal method (Nesterov)

- choose $x_0 \in \mathbb{R}^n$ and $\gamma_0 > 0$, and set $v_0 = x_0$
- for $k = 0, 1, 2, \ldots$, repeat
  1. find $\alpha_k \in (0, 1)$ that satisfies the equation
     \[ L\alpha_k^2 = (1 - \alpha_k)\gamma_k + \alpha_k \mu \]
     and let $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k \mu$
  2. choose
     \[ y^{(k)} = \frac{1}{\gamma_k + \alpha_k \mu} \left( \alpha_k \gamma_k v^{(k)} + \gamma_{k+1} x^{(k)} \right) \]
     and compute $f(y^{(k)})$ and $\nabla f(y^{(k)})$
  3. find $x^{(k+1)}$ such that
     \[ f(x^{(k+1)}) \leq f(y^{(k)}) - \frac{1}{2L} \| \nabla f(y^{(k)}) \|^2 \]
  4. set
     \[ v^{(k+1)} = \frac{1}{\gamma_{k+1}} \left( (1 - \alpha_k)\gamma_k v^{(k)} + \alpha_k \mu y^{(k)} - \alpha_k \nabla f(y^{(k)}) \right) \]
Bounding $\lambda_k$

**Lemma:** If $\gamma_0 \geq \mu$ in the optimal scheme on page 3–16, then

$$\lambda_k = \prod_{i=0}^{k-1} (1 - \alpha_i) \leq \min \left\{ \left( 1 - \sqrt{\frac{\mu}{L}} \right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\gamma_0})^2} \right\}$$

**Proof:**

- $\gamma_k \geq \mu$ and $\alpha_k \geq \sqrt{\mu/L}$ for all $k \geq 0$ because
  $$\gamma_{k+1} = L\alpha_k^2 = (1 - \alpha_k)\gamma_k + \alpha_k\mu \geq \mu$$

- $\gamma_k \geq \gamma_0 \lambda_k$ for all $k \geq 0$, since $\gamma_0 \geq \gamma_0 \lambda_0$ and
  $$\gamma_{k+1} \geq (1 - \alpha_k)\gamma_k \geq (1 - \alpha_k)\gamma_0 \lambda_k = \gamma_0 \lambda_{k+1}$$

- Let $a_k = \frac{1}{\sqrt{\lambda_k}}$, then $a_k \geq 1 + \frac{k}{2\sqrt{\gamma_0/L}}$ because
  $$a_{k+1} - a_k = \frac{\sqrt{\lambda_k} - \sqrt{\lambda_{k+1}}}{\sqrt{\lambda_k} \sqrt{\lambda_{k+1}}} \geq \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k \sqrt{\lambda_{k+1}}} = \frac{\alpha_k \lambda_k}{2\lambda_k \sqrt{\lambda_{k+1}}} = \frac{\alpha_k}{2\sqrt{\lambda_{k+1}}} \geq \frac{1}{2}\sqrt{\gamma_0/L}$$
Rate of convergence

**Theorem:** Let $\gamma_0 = L$, then the method on page 3–16 generates $\{x^{(k)}\}_{k=0}^{\infty}$ such that

$$f(x^{(k)}) - f^* \leq \min \left\{ \left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4}{(k+2)^2} \right\} L\|x_0 - x^*\|^2$$

this means the method is *optimal* for functions from class $S_{\mu,L}(\mathbb{R}^n)$

**Proof:** by lemma on page 3–9,

$$f(x^{(k)}) - f^* \leq \lambda_k \left( f(x^{(0)}) - f^* + \frac{\gamma_0}{2} \|x^{(0)} - x^*\|^2 \right)$$

then use $\gamma_0 = L$ and quadratic upper bound $f(x^{(0)}) - f^* \leq \frac{L}{2}\|x^{(0)} - x^*\|^2$
Variant of optimal method

eliminate \{v^{(k)}\} and \{\gamma_k\}, and use constant step size \( t = 1/L \)

- choose \( x^{(0)} \in \mathbb{R}^n \) and \( \alpha_0 \in [\sqrt{\frac{\mu}{L}}, 1) \), set \( y^{(0)} = x^{(0)} \) and \( q = \mu/L \)
- for \( k = 0, 1, 2, \ldots \), repeat
  1. compute \( f(y^{(k)}) \) and \( \nabla f(y^{(k)}) \), use gradient step update in step 3 in page 3-16, i.e.,
     \[
     x^{(k+1)} = y^{(k)} - \frac{1}{L} \nabla f(y^{(k)})
     \]
  2. compute \( \alpha_{k+1} \in (0, 1) \) from equation
     \[
     \alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}
     \]
     and set \( \beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}} \) and
     \[
     y^{(k+1)} = x^{(k+1)} + \beta_k(x^{(k+1)} - x^{(k)})
     \]
A simpler variant

choose $\alpha_0 = \sqrt{\frac{\mu}{L}}$, then

$$\alpha_k = \sqrt{\frac{\mu}{L}}, \quad \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$$

- choose $y^{(0)} = x^{(0)} \in \mathbb{R}^n$
- for $k = 0, 1, 2, \ldots$, repeat

$$x^{(k+1)} = y^{(k)} - \frac{1}{L} \nabla f(y^{(k)})$$

$$y^{(k+1)} = x^{(k+1)} + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}(x^{(k+1)} - x^{(k)})$$

however, this scheme does not work for $\mu = 0$
A simple variant when $\mu = 0$

- choose $y^{(0)} = x^{(0)} \in \mathbb{R}^n$
- for $k = 0, 1, 2, \ldots$, repeat

\[
x^{(k+1)} = y^{(k)} - \frac{1}{L} \nabla f(y^{(k)})
\]
\[
y^{(k+1)} = x^{(k+1)} + \frac{k}{k+3} (x^{(k+1)} - x^{(k)})
\]

when $L$ is unknown, can replace first equation with line search

\[
x^{(k+1)} = y^{(k)} - t_k \nabla f(y^{(k)})
\]
Example

minimize \ \log\left(\sum_{i=1}^{m} \exp(a_i^T x + b_i)\right)

randomly generated data with \( m = 500 \) and \( n = 200 \), same fixed step size

Optimal gradient methods
Example

\[
\text{minimize } \log \left( \sum_{i=1}^{m} \exp(a_i^T x + b_i) \right)
\]

randomly generated data with \( m = 500, n = 200 \), backtracking line search

![Graph showing function values over iterations](image-url)
References


• L. Vandenberghe, *Lecture notes for EE236C - Optimization Methods for Large-Scale Systems* (Spring 2011), UCLA.

almost all materials of this lecture are taken from Nesterov’s book (2004) (except the numerical examples)