Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 4: Unconstrained, smooth minimization II

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2019)
License Information for Mathematics of Data Slides

- This work is released under a Creative Commons License with the following terms:
  - **Attribution**
    - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.
  - **Non-Commercial**
    - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor's permission.
  - **Share Alike**
    - The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor’s work.
  - **Full Text of the License**
Outline

▶ This lecture
  1. Gradient and accelerated gradient descent methods

▶ Next lecture
  1. Adaptive gradient methods
  2. Second-order optimization methods
Recommended reading

Overview

This lecture covers the basics of numerical methods for *unconstrained* and *smooth* convex minimization.
Recall: Convex, unconstrained, smooth minimization

**Problem (Mathematical formulation)**

\[
F^* := \min_{x \in \mathbb{R}^p} \{ F(x) := f(x) \} 
\]

where \( f \) is convex and twice differentiable.

Note that (1) is unconstrained.

How do we design efficient optimization algorithms with accuracy-computation tradeoffs for this class of functions?
Basic principles of descent methods

Template for iterative descent methods

1. Let $x^0 \in \text{dom}(f)$ be a starting point.
2. Generate a sequence of vectors $x^1, x^2, \ldots \in \text{dom}(f)$ so that we have descent:
   \[ f(x^{k+1}) < f(x^k), \quad \text{for all } k = 0, 1, \ldots \]
   until $x_k$ is $\epsilon$-optimal.

Such a sequence $\{x^k\}_{k \geq 0}$ can be generated as:
\[ x^{k+1} = x^k + \alpha_k p^k \]
where $p^k$ is a descent direction and $\alpha_k > 0$ a step-size.

Remarks

- Iterative algorithms can use various oracle information from the objective, such as its value, gradient, or Hessian, in different ways to obtain $\alpha_k$ and $p^k$.
- These choices determine the overall convergence rate and complexity.
- The type of oracle information used becomes a defining characteristic.
A condition for local descent directions

The iterates are given as:

\[ x^{k+1} = x^k + \alpha_k p^k \]

By Taylor's theorem, we have

\[ f(x^{k+1}) = f(x^k) + \alpha_k \langle \nabla f(x^k), p^k \rangle + O(\alpha_k^2 \|p\|^2_2). \]

For \( \alpha_k \) small enough, the term \( \alpha_k \langle \nabla f(x^k), p^k \rangle \) dominates \( O(\alpha_k^2) \) for a fixed \( p^k \).

Therefore, in order to have \( f(x^{k+1}) < f(x^k) \), we require

\[ \langle \nabla f(x^k), p^k \rangle < 0 \]
Basic principles of descent methods

Local steepest descent direction

Since

\[ \langle \nabla f(x^k), p^k \rangle = \| \nabla f(x^k) \| \| p^k \| \cos \theta , \]

where \( \theta \) is the angle between \( \nabla f(x^k) \) and \( p^k \), we have that

\[ p^k := -\nabla f(x^k) \]

is the local steepest descent direction.

\[ \nabla f(x^k) \]

\[ x^k + D(f, x^k) \]

\[ p^k \]

\[ \text{level sets} \]

Figure: Descent directions in 2D should be an element of the cone of descent directions \( D(f, \cdot) \).
A reminder on notation

Important notation used throughout the whole lecture:

- $\mathcal{F}_L^{l,m}$: Functions that are $l$-times differentiable with $m$-th order Lipschitz property
  In this lecture, $m = 1$, and $l \in \{1, 2, \infty\}$

- $\mathcal{F}_{L,\mu}^{l,m}$: Subset of $\mathcal{F}_L^{l,m}$ also satisfying $\mu$-strong convexity
Gradient descent methods

Gradient descent (GD) algorithm

The gradient method we discussed before indeed use the local steepest direction:

\[ p^k = -\nabla f(x^k) \]

so that

\[ x^{k+1} = x^k - \alpha_k \nabla f(x^k). \]

Key question: How do we choose \( \alpha_k \) so that we are guaranteed to successfully descend? (ideally as fast as possible)
Gradient descent methods

Gradient descent (GD) algorithm

The gradient method we discussed before indeed use the local steepest direction:

$$p^k = -\nabla f(x^k)$$

so that

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k).$$

**Key question:** How do we choose $\alpha_k$ so that we are guaranteed to successfully descend? (ideally as fast as possible)

**Answer:** By exploiting the structures within the convex function

When $f \in \mathcal{F}_{L}^{2,1}$, we can use $\alpha_k = 1/L$ so that $x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k)$ is contractive.

- GD method only uses the gradient information, hence called a **first-order method**.
- **First-order methods** employ only the value of $f$ and $\nabla f$ at specific points.
- **Second-order methods** also use the Hessian $\nabla^2 f$. 
Recall: Gradient descent methods - a geometrical intuition

\[ f(x) \]

\[ x^* \quad x^k \]
Recall: Gradient descent methods - a geometrical intuition

\[ f(x) \geq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle \]

Structure in optimization:
Recall: Gradient descent methods - a geometrical intuition

Majorize:
\[ f(x) \leq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \| x - x^k \|^2 := Q_L(x, x^k) \]

Minimize:
\[ x^{k+1} = \arg \min_x Q_L(x, x^k) \]
\[ = \arg \min_x \left\| x - \left( x^k - \frac{1}{L} \nabla f(x^k) \right) \right\|^2 \]
\[ = x^k - \frac{1}{L} \nabla f(x^k) \]

Structure in optimization:

(1) \[ f(x) \geq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle \]

(2) \[ f(x) \leq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \| x - x^k \|^2 \]
Recall: Gradient descent methods - a geometrical intuition

Majorize:
\[ f(x) \leq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L'}{2} \| x - x^k \|_2^2 := Q_{L'}(x, x^k) \]

Minimize:
\[ x^{k+1} = \arg \min_x Q_{L'}(x, x^k) \]
\[ = \arg \min_x \left\| x - \left( x^k - \frac{1}{L'} \nabla f(x^k) \right) \right\|^2 \]
\[ = x^k - \frac{1}{L'} \nabla f(x^k) \]

Structure in optimization:
(1) \[ f(x) \geq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle \]
(2) \[ f(x) \leq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \| x - x^k \|_2^2 \]
Recall: Gradient descent methods - a geometrical intuition

Majorize:
\[ f(x) \leq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \| x - x^k \|_2^2 := Q_L(x, x^k) \]

Minimize:
\[ x^{k+1} = \arg \min_x Q_L(x, x^k) \]
\[ = \arg \min_x \| x - \left( x^k - \frac{1}{L} \nabla f(x^k) \right) \|^2 \]
\[ = x^k - \frac{1}{L} \nabla f(x^k) \]

Structure in optimization:
(1) \[ f(x) \geq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle \]
(2) \[ f(x) \leq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \| x - x^k \|_2^2 \]
(3) \[ f(x) \geq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{\mu}{2} \| x - x^k \|_2^2 \]
Convergence rate of gradient descent

**Theorem**

Let the starting point for GD be $x^0 \in \text{dom}(f)$.

- If $f \in \mathcal{F}^{2,1}_L$, with the choice $\alpha = \frac{1}{L}$, the iterates of GD satisfy
  \[
  f(x^k) - f(x^*) \leq \frac{2L}{k+4} \|x^0 - x^*\|_2^2
  \]

- If $f \in \mathcal{F}^{2,1}_{L,\mu}$, with the choice $\alpha = \frac{2}{L+\mu}$, the iterates of GD satisfy
  \[
  \|x^k - x^*\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|x^0 - x^*\|_2
  \]

- If $f \in \mathcal{F}^{2,1}_{L,\mu}$, with the choice $\alpha = \frac{1}{L}$, the iterates of GD satisfy
  \[
  \|x^k - x^*\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^{\frac{k}{2}} \|x^0 - x^*\|_2
  \]
Proof of convergence rates of gradient descent - part I (self-study)

- We first need to prove a basic result about functions in $\mathcal{F}^1_L$

**Lemma**

Let $f \in \mathcal{F}^1_L$. Then it holds that

$$\frac{1}{L} \| \nabla f(x) - \nabla f(y) \|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle$$

(2)

**Proof (Advanced material).**

First, recall the following result about Lipschitz gradient functions $h \in \mathcal{F}^1_L$

$$h(x) \leq h(y) + \langle \nabla h(y), x - y \rangle + \frac{L}{2} \| x - y \|^2.$$  

(3)

To prove the result, let $\phi(y) := f(y) - \langle \nabla f(x), y \rangle$, with $\nabla \phi(y) = \nabla f(y) - \nabla f(x)$. Clearly, $\phi(y)$ attains its minimum value at $y^* = x$. Hence, and by also applying (3) with $h = \phi$ and $x = y - \frac{1}{L} \nabla \phi(y)$, we get

$$\phi(x) \leq \phi \left( y - \frac{1}{L} \nabla \phi(y) \right) \leq \phi(y) - \frac{1}{2L} \| \nabla \phi(y) \|^2.$$  

Substituting the above definitions into the left and right hand sides gives

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \| \nabla f(x) - \nabla f(y) \|^2 \leq f(y)$$

(4)

By adding two copies of (4) with each other, with $x$ and $y$ swapped, we obtain (2).
Proof of convergence rates of gradient descent - part II (self-study)

**Theorem**

If \( f \in \mathcal{F}_{L,1}^2 \), with the choice \( \alpha = \frac{1}{L} \), the iterates of GD satisfy

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

**(5)**

**Proof**

- Consider the constant step-size iteration \( x^{k+1} = x^k - \alpha \nabla f(x^k) \).
- Let \( r_k := \|x^k - x^*\| \), where \( x^* \) denotes a minimizer. Show \( r_k \leq r_0 \).

\[
r_{k+1}^2 := \|x^{k+1} - x^*\|^2 = \|x^k - x^* - \alpha \nabla f(x^k)\|^2
\]

\[
= \|x^k - x^*\|^2 - 2\alpha \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle + \alpha^2 \|\nabla f(x^k)\|^2
\]

\[
\leq r_k^2 - \alpha(2/L - \alpha)\|\nabla f(x^k)\|^2 \quad \text{(by (2))}
\]

\[
\leq r_k^2, \quad \forall \alpha < 2/L.
\]

Hence, the gradient iterations are contractive when \( \alpha < 2/L \) for all \( k \geq 0 \).

- **An auxiliary result:** Let \( \Delta_k := f(x^k) - f^* \). Show \( \Delta_k \leq r_0 \|\nabla f(x^k)\| \).

\[
\Delta_k \leq \langle \nabla f(x^k), x^k - x^* \rangle \leq \|\nabla f(x^k)\| \|x^k - x^*\| = r_k \|\nabla f(x^k)\| \leq r_0 \|\nabla f(x^k)\|.
\]
Proof of convergence rates of gradient descent - part III
(self-study)

Proof (continued)

- We can establish **convergence** along with the auxiliary result above:

\[
\begin{align*}
    f(x_{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x_{k+1} - x^k \rangle + \frac{L}{2} \|x_{k+1} - x^k\|^2 \\
    &= f(x^k) - \omega_k \|\nabla f(x^k)\|^2, \quad \omega_k := \alpha(1 - L\alpha/2).
\end{align*}
\]

Subtract \(f^*\) from both sides and apply the last inequality of the previous slide to get

\[
\Delta_{k+1} \leq \Delta_k - (\omega_k/r_0^2)\Delta_k^2.
\]

Thus, dividing by \(\Delta_{k+1}\Delta_k\)

\[
\Delta_{k+1}^{-1} \geq \Delta_k^{-1} + (\omega_k/r_0^2)\Delta_k/\Delta_{k+1} \geq \Delta_k^{-1} + (\omega_k/r_0^2).
\]

By induction, we have \(\Delta_{k+1}^{-1} \geq \Delta_0^{-1} + (\omega_k/r_0^2)(k + 1)\). Then, taking \((\cdot)^{-1}\) of both sides (and hence replacing \(\geq\) by \(\leq\)) and substituting all of the definitions gives

\[
f(x^k) - f(x^*) \leq \frac{2(f(x_0) - f(x^*))\|x_0 - x^*\|^2}{2\|x_0 - x^*\|^2 + k\alpha(2 - \alpha L)(f(x_0) - f^*)},
\]

- In order to choose the **optimal** step-size, we maximize the function \(\phi(\alpha) = \alpha(2 - \alpha L)\).

  Hence, the optimal step size for the gradient method for \(f \in F^{1,1}_L\) is given by \(\alpha = \frac{1}{L}\).

- Finally, since \(f(x_0) \leq f^* + \nabla f(x^*)^T(x_0 - x^*) + (L/2)\|x_0 - x^*\|^2 = f^* + (L/2)r_0^2\),

  we obtain (5).
Proof of convergence rates of gradient descent - part IV
(self-study)

**Theorem**

- If $f \in \mathcal{F}_{L, \mu}^{2,1}$, with the choice $\alpha = \frac{2}{L+\mu}$, the iterates of GD satisfy

  $$\|x^k - x^*\|_2 \leq \left(\frac{L - \mu}{L + \mu}\right)^k \|x^0 - x^*\|_2$$

  (6)

- If $f \in \mathcal{F}_{L, \mu}^{2,1}$, with the choice $\alpha = \frac{1}{L}$, the iterates of GD satisfy

  $$\|x^k - x^*\|_2 \leq \left(\frac{L - \mu}{L + \mu}\right)^{\frac{k}{2}} \|x^0 - x^*\|_2$$

  (7)

Before proving the convergence rate, we first need a result about functions in $\mathcal{F}_{L, \mu}^{1,1}$. It is proved similarly to (2).

**Theorem**

If $f \in \mathcal{F}_{L, \mu}^{1,1}$, then for any $x$ and $y$, we have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2.$$  

(8)
Proof of (6) and (7)

Let \( r_k = \|x^k - x^*\| \). Then, using (8) and the fact that \( \nabla f(x^*) = 0 \), we have

\[
\begin{align*}
    r_{k+1}^2 &= \|x_{k+1} - x^* - \alpha \nabla f(x^k)\|^2 \\
    &= r_k^2 - 2\alpha \langle \nabla f(x^k), x^k - x^* \rangle + \alpha^2 \|\nabla f(x^k)\|^2 \\
    &\leq \left(1 - \frac{2\alpha \mu L}{\mu + L}\right) r_k^2 + \alpha \left(\alpha - \frac{2}{\mu + L}\right) \|\nabla f(x^k)\|^2
\end{align*}
\]

Since \( \mu \leq L \), we have \( \alpha \leq \frac{2}{\mu + L} \) in both the cases \( \alpha = \frac{1}{L} \) or \( \alpha = \frac{2}{\mu + L} \). So the last term in the previous inequality is negative, and hence

\[
r_{k+1}^2 \leq \left(1 - \frac{2\alpha \mu L}{\mu + L}\right) r_0^2
\]

Plugging \( \alpha = \frac{1}{L} \) and \( \alpha = \frac{2}{\mu + L} \), we obtain the rates as advertised.

For \( f \in \mathcal{F}_{L,\mu}^{1,1} \), the optimal step-size is given by \( \alpha = \frac{2}{\mu + L} \) (i.e., it optimizes the worst case bound).
Convergence rate of gradient descent

\begin{align*}
  f \in \mathcal{F}_{L}^{2,1}, \quad \alpha = \frac{1}{L} & \quad f(x^k) - f(x^*) \leq \frac{2L}{k+4} \|x^0 - x^*\|_2^2 \\
  f \in \mathcal{F}_{L,\mu}^{2,1}, \quad \alpha = \frac{2}{L+\mu} & \quad \|x^k - x^*\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|x^0 - x^*\|_2 \\
  f \in \mathcal{F}_{L,\mu}^{2,1}, \quad \alpha = \frac{1}{L} & \quad \|x^k - x^*\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^{k/2} \|x^0 - x^*\|_2
\end{align*}

Remarks

- **Assumption**: Lipschitz gradient. **Result**: convergence rate in **objective values**.
- **Assumption**: Strong convexity. **Result**: convergence rate in **sequence** of the iterates and in **objective values**.
- Note that the suboptimal step-size choice \( \alpha = \frac{1}{L} \) **adapts** to the strongly convex case (i.e., it features a linear rate vs. the standard sublinear rate).
Example: Ridge regression

### Optimization formulation

- Let $A \in \mathbb{R}^{n \times p}$ and $b \in \mathbb{R}^n$ given by the model $b = Ax^\dagger + w$, where $w \in \mathbb{R}^n$ is some noise.
- We can try to estimate $x^\dagger$ by solving the Tikhonov regularized least squares
  \[
  \min_{x \in \mathbb{R}^p} f(x) := \frac{1}{2} \|b - Ax\|_2^2 + \frac{\rho}{2} \|x\|_2^2.
  \]
  where $\rho \geq 0$ is a regularization parameter.

### Remarks

- $f \in \mathcal{F}_{L,\mu}^{2,1}$ with:
  - $L = \lambda_p (A^T A) + \rho$;
  - $\mu = \lambda_1 (A^T A) + \rho$;
  - where $\lambda_1 (A^T A) \leq \ldots \leq \lambda_p (A^T A)$ are the eigenvalues of $A^T A$.
- The ratio $\frac{L}{\mu}$ decreases as $\rho$ increases, leading to faster linear convergence.
- Note that if $n < p$ and $\rho = 0$, we have $\mu = 0$, hence $f \in \mathcal{F}_L^{2,1}$ and we can expect only $O(1/k)$ convergence from the gradient descent method.
Example: Ridge regression

Case 1:
\[ n = 500, \ p = 2000, \ \rho = 0 \]

Case 2:
\[ n = 500, \ p = 2000, \ \rho = 0.01\lambda_p(A^T A) \]
Information theoretic lower bounds [5]

What is the **best** achievable rate for a **first-order** method (one using gradient information but not higher-order quantities)?

\[
f \in \mathcal{F}_{L}^{\infty,1}: \text{Smooth and Lipschitz-gradient}\]

It is possible to construct a function in \( \mathcal{F}_{L}^{\infty,1} \), for which **any** first order method must satisfy

\[
f(x^k) - f(x^*) \geq \frac{3L}{32(k+1)^2} \|x^0 - x^*\|_2^2 \quad \text{for all } k \leq \frac{p-1}{2}
\]

\[
f \in \mathcal{F}_{L,\mu}^{\infty,1}: \text{Smooth and strongly convex}\]

It is possible to construct a function in \( \mathcal{F}_{L,\mu}^{\infty,1} \), for which **any** first order method must satisfy

\[
\|x^k - x^*\|_2 \geq \left( \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^k \|x^0 - x^*\|_2
\]

**Gradient descent is** \( O(1/k) \) **for** \( \mathcal{F}_{L}^{\infty,1} \) **and it is slower for** \( \mathcal{F}_{L,\mu}^{\infty,1} \), **hence it does not achieve the lower bounds!**
Problem

Is it possible to design optimal first-order methods with convergence rates matching the theoretical lower bounds?

NOTE:
AG is not monotone, but the cost-per-iteration is essentially the same as GD.
### Problem

Is it possible to design optimal first-order methods with convergence rates matching the theoretical lower bounds?

### Solution [Nesterov’s accelerated scheme]

Accelerated Gradient (AG) methods achieve optimal convergence rates at a negligible increase in the computational cost.
Accelerated gradient descent algorithm

Problem

Is it possible to design optimal first-order methods with convergence rates matching the theoretical lower bounds?

Solution [Nesterov’s accelerated scheme]

Accelerated Gradient (AG) methods achieve optimal convergence rates at a negligible increase in the computational cost.

Accelerated Gradient algorithm for $F_L^{1,1}$ (AG-L)

1. Set $x^0 = y^0 \in \text{dom} (f)$ and $t_0 := 1$.
2. For $k = 0, 1, \ldots$, iterate

\[
\begin{cases}
x^{k+1} = y^k - \frac{1}{L} \nabla f(y^k) \\
t_{k+1} = (1 + \sqrt{4t_k^2 + 1})/2 \\
y^{k+1} = x^{k+1} + \frac{(t_k - 1)}{t_{k+1}} (x^{k+1} - x^k)
\end{cases}
\]

NOTE: AG is not monotone, but the cost-per-iteration is essentially the same as GD.
Accelerated gradient descent algorithm

Problem

Is it possible to design optimal first-order methods with convergence rates matching the theoretical lower bounds?

Solution [Nesterov’s accelerated scheme]

Accelerated Gradient (AG) methods achieve optimal convergence rates at a negligible increase in the computational cost.

### Accelerated Gradient algorithm for $\mathcal{F}_{L}^{1,1}$ (AG-L)

1. Set $x^0 = y^0 \in \text{dom}(f)$ and $t_0 := 1$.
2. For $k = 0, 1, \ldots$, iterate

$$
\begin{align*}
x^{k+1} &= y^k - \frac{1}{L} \nabla f(y^k) \\
t_{k+1} &= (1 + \sqrt{4t_k^2 + 1})/2 \\
y^{k+1} &= x^{k+1} + \frac{(t_k - 1)}{t_{k+1}} (x^{k+1} - x^k)
\end{align*}
$$

### Accelerated Gradient algorithm for $\mathcal{F}_{L,\mu}^{1,1}$ (AG-$\mu$L)

1. Choose $x^0 = y^0 \in \text{dom}(f)$
2. For $k = 0, 1, \ldots$, iterate

$$
\begin{align*}
x^{k+1} &= y^k - \frac{1}{L} \nabla f(y^k) \\
y^{k+1} &= x^{k+1} + \gamma (x^{k+1} - x^k)
\end{align*}
$$

where $\gamma = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$. 

NOTE: AG is not monotone, but the cost-per-iteration is essentially the same as GD.
Accelerated gradient descent algorithm

**Problem**

*Is it possible to design optimal first-order methods with convergence rates matching the theoretical lower bounds?*

**Solution [Nesterov’s accelerated scheme]**

Accelerated Gradient (AG) methods achieve optimal convergence rates at a negligible increase in the computational cost.

<table>
<thead>
<tr>
<th>Accelerated Gradient algorithm for $\mathcal{F}^{1,1}_L$ (AG-L)</th>
<th>Accelerated Gradient algorithm for $\mathcal{F}^{1,1}_{L,\mu}$ (AG-$\mu$L)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.</strong> Set $x^0 = y^0 \in \text{dom} ,(f)$ and $t_0 := 1$.</td>
<td><strong>1.</strong> Choose $x^0 = y^0 \in \text{dom} ,(f)$</td>
</tr>
<tr>
<td><strong>2.</strong> For $k = 0, 1, \ldots$, iterate</td>
<td><strong>2.</strong> For $k = 0, 1, \ldots$, iterate</td>
</tr>
<tr>
<td>[ x^{k+1} = y^k - \frac{1}{L} \nabla f(y^k) ] [ t_{k+1} = (1 + \sqrt{4t_k^2 + 1})/2 ] [ y^{k+1} = x^{k+1} + \frac{(t_k-1)}{t_{k+1}}(x^{k+1} - x^k) ]</td>
<td>[ x^{k+1} = y^k - \frac{1}{L} \nabla f(y^k) ] [ y^{k+1} = x^{k+1} + \gamma(x^{k+1} - x^k) ] where $\gamma = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$.</td>
</tr>
</tbody>
</table>

**NOTE:** AG is not monotone, but the cost-per-iteration is essentially the same as GD.
Global convergence of AGD [5]

Theorem ($f$ is convex with Lipschitz gradient)

If $f \in \mathcal{F}^{1,1}_L$ or $\mathcal{F}^{1,1}_{L,\mu}$, the sequence $\{x^k\}_{k \geq 0}$ generated by AGD-L satisfies

$$f(x^k) - f^* \leq \frac{4L}{(k+2)^2} \|x^0 - x^*\|_2^2, \ \forall k \geq 0.$$  \hfill (9)
Global convergence of AGD [5]

**Theorem (\(f\) is convex with Lipschitz gradient)**

If \(f \in \mathcal{F}_{L,1}^1\) or \(\mathcal{F}_{L,\mu}^1\), the sequence \(\{x^k\}_{k \geq 0}\) generated by AGD-L satisfies

\[
f(x^k) - f^* \leq \frac{4L}{(k+2)^2} \|x^0 - x^*\|_2^2, \quad \forall k \geq 0.
\]

(9)

AGD-L is optimal for \(\mathcal{F}_{L,1}^1\) but NOT for \(\mathcal{F}_{L,\mu}^1\)!

**Theorem (\(f\) is strongly convex with Lipschitz gradient)**

If \(f \in \mathcal{F}_{L,\mu}^1\), the sequence \(\{x^k\}_{k \geq 0}\) generated by AGD-\(\mu\)L satisfies

\[
f(x^k) - f^* \leq L \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \|x^0 - x^*\|_2^2, \quad \forall k \geq 0
\]

(10)

\[
\|x^k - x^*\|_2 \leq \sqrt{\frac{2L}{\mu}} \left(1 - \sqrt{\frac{\mu}{L}}\right)^{k/2} \|x^0 - x^*\|_2, \quad \forall k \geq 0.
\]

(11)

- AGD-L’s iterates are not guaranteed to converge.
- AGD-L does not have a linear convergence rate for \(\mathcal{F}_{L,\mu}^1\).
- AGD-\(\mu\)L does, but needs to know \(\mu\).

**AGD achieves the iteration lowerbound within a constant!**
Example: Ridge regression

Case 1:

\( n = 500, \ p = 2000, \ \rho = 0 \)

Case 2:

\( n = 500, \ p = 2000, \ \rho = 0.01\lambda_p(A^T A) \)
Enhancements

Two enhancements

1. Line-search for estimating $L$ for both GD and AGD.
2. Restart strategies for AGD.
Enhancements

Two enhancements

1. Line-search for estimating $L$ for both GD and AGD.
2. Restart strategies for AGD.

When do we need a line-search procedure?

We can use a line-search procedure for both GD and AGD when

- $L$ is known but it is expensive to evaluate;
- The global constant $L$ usually does not capture the local behavior of $f$ or it is unknown;
Enhancements

Two enhancements

1. Line-search for estimating $L$ for both GD and AGD.
2. Restart strategies for AGD.

When do we need a line-search procedure?

We can use a line-search procedure for both GD and AGD when

- $L$ is known but it is expensive to evaluate;
- The global constant $L$ usually does not capture the local behavior of $f$ or it is unknown;

**Line-search**

At each iteration, we try to find a constant $L_k$ that satisfies:

$$f(x^{k+1}) \leq Q_{L_k}(x^{k+1}, y^k) := f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k \rangle + \frac{L_k}{2} \|x^{k+1} - y^k\|_2^2.$$ 

Here: $L_0 > 0$ is given (e.g., $L_0 := c \frac{\|\nabla f(x^1) - \nabla f(x^0)\|_2}{\|x^1 - x^0\|_2}$) for $c \in (0, 1]$. 
How can we better adapt to the local geometry?

\[ f(x) = \min_{x} \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \| x - x^k \|^2 \right\} \]

\[ \| \nabla f(x) - \nabla f(y) \| \leq L \| y - x \| \]

L is a global worst-case constant

\[ Q_L(x, x^k) \] is a global quadratic upper bound

\[ x^{k+1} = \text{arg min}_x \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2} \| x - x^k \|^2 \right\} \]
How can we better adapt to the local geometry?

\[ f(x) \]

Local quadratic upper bound

\[ Q_{L_k}(x, x^k) \]

\[ x^{k+1} = \arg\min_x \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{L_k}{2} \| x - x^k \|_2^2 \right\} \]

\[ \| \nabla f(x) - \nabla f(y) \| \leq L \| y - x \| \]

\( L \) is a global worst-case constant

\[ f(x) \leq f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{L_k}{2} \| x - x^k \|_2^2 \]

applies only locally
Enhancements

Why do we need a restart strategy?

- AG-$\mu L$ requires knowledge of $\mu$ and AG-$L$ does not have optimal convergence for strongly convex $f$.
- AG is non-monotonic (i.e., $f(x^{k+1}) \leq f(x^k)$ is not always satisfied).
- AG has a periodic behavior, where the momentum depends on the local condition number $\kappa = \frac{L}{\mu}$.
- A restart strategy tries to reset this momentum whenever we observe high periodic behavior. We often use function values but other strategies are possible.

Restart strategies

1. **O’Donoghue - Candes’s strategy [7]**: There are at least three options: Restart with fixed number of iterations, restart based on objective values, and restart based on a gradient condition.
2. **Giselsson-Boyd’s strategy [4]**: Do not require $t_k = 1$ and do not necessary require function evaluations.
3. **Fercoq-Qu’s strategy [3]**: Unconditional periodic restart for strongly convex functions. Do not require the strong convexity parameter.
Oscillatory behavior of AGD

- Minimize a quadratic function $f(x) = x^T \Phi x$, with $p = 200$ and $\kappa(\Phi) = L/\mu = 2.4 \times 10^4$

- Use stepsize $\alpha = 1/L$ and update $x^{k+1} + \gamma_{k+1}(x^{k+1} - x^k)$ where
  - $\gamma_{k+1} = \theta_k(1 - \theta_k)/(\theta_k^2 + \theta_{k+1})$
  - $\theta_{k+1}$ solves $\theta_{k+1}^2 = (1 - \theta_{k+1})\theta_k^2 + q\theta_{k+1}$.

- The parameter $q$ should be equal to the reciprocal of condition number $q^* = \mu/L$.
- A different choice of $q$ might lead to oscillatory behavior.

---

Figure 1: Convergence of Algorithm 1 with different estimates of $q$.

Interpretation. The optimal momentum depends on the condition number of the function; specifically, higher momentum is required when the function has a higher condition number. Underestimating the amount of momentum required leads to slower convergence. However, we are more often in the other regime, that of overestimated momentum, because generally $\beta_k \uparrow 1$, which corresponds to high momentum and rippling behavior, as we see in Figure 1. This can be visually understood in Figure (2), which shows the trajectories of sequences generated by Algorithm 1 minimizing a positive definite quadratic in two dimensions, under $q = q^*$, the optimal choice of $q$, and $q = 0$. The high momentum causes the trajectories to overshoot the minimum and oscillate around it. This causes a rippling in the function values along the trajectory. Later we shall demonstrate that the period of these ripples is proportional to the square root of the (local) condition number of the function.

Lastly we mention that the condition number is a global parameter; the sequence generated by an accelerated scheme may enter regions that are locally better conditioned, say, near the optimum. In these cases the choice of $q = q^*$ is appropriate outside of this region, but once we enter it we expect the rippling behavior associated with high momentum to emerge, despite the optimal choice of $q$. 
Example: Ridge regression

Case 1:

\[ n = 500, p = 2000, \rho = 0 \]

Case 2:

\[ n = 500, p = 2000, \rho = 0.01 \lambda_p (A^T A) \]
The (special) quadratic case – Step-size

Consider the minimization of a quadratic function

$$\min_x f(x) := \frac{1}{2} \langle x, Ax \rangle - \langle b, x \rangle$$

where $A$ is a $p \times p$ symmetric positive definite matrix, i.e., $A = A^T \succ 0$.

### Gradient Descent

$$\alpha_k = \frac{1}{L} \quad \text{with} \quad L = \|A\|$$

### Steepest descent

$$\alpha_k = \frac{\|\nabla f(x^k)\|^2}{\langle \nabla f(x^k), A\nabla f(x^k) \rangle} \quad (12)$$

### Barzilai-Borwein

$$\alpha_k = \frac{\|\nabla f(x^{k-1})\|^2}{\langle \nabla f(x^{k-1}), A\nabla f(x^{k-1}) \rangle} \quad (13)$$
The (special) quadratic case – convergence rates

For $f(x) = \frac{1}{2} \langle x, Ax \rangle - \langle b, x \rangle$, we have $L = \|A\| = \lambda_p$ and $\mu = \lambda_1$, where $0 < \lambda_1 \leq \lambda_2 \leq \cdots \lambda_p$ are the eigenvalues of $A$.

**Theorem (Gradient Descent)**

$$\|x^k - x^*\|_2 \leq \left(1 - \frac{\lambda_1}{\lambda_p}\right)^k \|x^0 - x^*\|_2$$

**Theorem (Steepest Descent)**

$$\|x^{k+1} - x^*\|_A \leq \left(\frac{\lambda_p - \lambda_1}{\lambda_p + \lambda_1}\right)^k \|x^0 - x^*\|_A$$

**Theorem (Barzilai-Borwein)**

*Under the condition $\lambda_p < 2\lambda_1$

$$\|x^{k+1} - x^*\|_2 \leq \left(\frac{\lambda_p - \lambda_1}{\lambda_1}\right)^k \|x^0 - x^*\|_2$$
Example: Quadratic function

Case 1: $n = p = 100, \kappa(A) = 10$

Case 1: $n = p = 100, \kappa(A) = 100$
The gradient method for non-convex optimization

- Gradient descent does not match lower bounds in convex setting.
- How about non-convex problems?

Lower bounds for non-convex problems

Consider $f \in F_L^{1,1}$ and $f$ is non-convex. Then any first-order method must satisfy,

$$\|\nabla f(x^t)\| = \Omega\left(\frac{1}{\sqrt{t}}\right)$$

As a corollary,

$$T = \Omega\left(\epsilon^{-2}\right) [2]$$

Gradient descent is optimal for non-convex problems, up to some constant factor!
The gradient method for non-convex optimization

- Gradient descent does not match lower bounds in convex setting.
- How about non-convex problems?

Lower bounds for non-convex problems

Consider $f \in \mathcal{F}_{1,1}^L$ and $f$ is non-convex. Then any first-order method must satisfy,

$$\|\nabla f(x^t)\| = \Omega \left( \frac{1}{\sqrt{t}} \right)$$

As a corollary,

$$T = \Omega \left( \epsilon^{-2} \right) \ [2]$$

- In fact, we cannot talk about acceleration for $f \in \mathcal{F}_{1,1}^L$.
- Acceleration for non-convex, $L$-Lipschitz gradient functions is not as meaningful.
Example: Logistic regression with non-convex regularizer

Problem (Regularized logistic regression)

Given $A \in \{0, 1\}^{n \times p}$ and $b \in \{-1, +1\}^n$, solve:

$$ f^* := \min_{x, \beta} \left\{ f(x) := \frac{1}{n} \sum_{j=1}^{n} \log \left( 1 + \exp \left( -b_j (a_j^T x + \beta) \right) \right) + \frac{\theta}{2} \phi(x) \right\}. $$

where $\phi(x) = \sum_{i=1}^{d} \phi(x_i)$.

Definition (Smoothly clipped absolute deviation (SCAD))

$$ \phi(x_i) = \begin{cases} 
 \lambda |x_i| & |x_i| \leq \lambda, \\
 -|x_i|^2 + 2a\lambda |x_i| - \lambda^2 & \lambda < |x_i| \leq a\lambda, \\
 (1 + a)\lambda^2 / 2 & |x_i| > a\lambda
\end{cases} $$
SCAD penalty

SCAD regularizer with $\lambda = 1$, $a = 4$. 
Example: Non-convex logistic regression

Gradient descent for SCAD-regularized logistic regression
*What to remember about mirror descent?

- Approximates the optimum by lower bounding the function via hyperplanes at $x_t$

  \[
  f(x_t) \leq f(y) + \langle \nabla f(y), x_t - y \rangle
  \]

- The smaller the gradients, the better the approximation!
What to remember about mirror descent?

- Approximates the optimum by lower bounding the function via hyperplanes at $x_t$

- The smaller the gradients, the better the approximation!
*What to remember about mirror descent?

- **Approximates** the optimum by **lower bounding** the function via **hyperplanes** at $x_t$.

- The smaller the gradients, the better the approximation!
*What to remember about mirror descent?

- Approximates the optimum by lower bounding the function via hyperplanes at $x_t$

- The smaller the gradients, the better the approximation!
Gradient descent + mirror descent: Nesterov’s acceleration

- There has been quite a few interpretations of Nesterov's accelerated gradient method.
- One way to achieve acceleration is to combine gradient and mirror descent.

### Nesterov’s Optimal Scheme [6], Linear Coupling [1]

1. Set $x^0 = y^0 = z^0 \in \text{dom}(f)$.
2. Define $d_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle$.
3. For $k = 0, 1, \ldots$, iterate

$$
\left\{ \begin{array}{l}
\alpha_{k+1} = \frac{k+2}{2L}, \quad \text{and} \quad \tau_k = \frac{1}{\alpha_{k+1} L} = \frac{2}{k+2} \\
x^{k+1} = \tau_k z^k + (1 - \tau_k) y^k \\
y^{k+1} = \arg \min_y \left\langle \nabla f(x^{k+1}), y - x^{k+1} \right\rangle + \frac{L}{2} ||y - x^{k+1}||^2 \quad \text{(Gradient Descent)} \\
z^{k+1} = \arg \min_z \left\langle \nabla f(x^{k+1}), z - z^k \right\rangle + \frac{1}{\alpha_{k+1}} d_\phi(z, z^k) \quad \text{(Mirror Descent)}
\end{array} \right.
$$

Gradient descent + mirror descent

- $y^{k+1}$ update: gradient descent (recall majorization/minimization)
- $z^{k+1}$ update: mirror descent (recall bregman divergence)
Gradient descent + mirror descent: Nesterov’s acceleration

Intuition behind this alternative interpretation

- Gradient descent takes larger steps when the gradient is large, i.e., when we are far away from optimum
- Mirror descent has tighter lower bounds with respect to optimum when gradients are small, i.e., when close to optimum
- Essentially, this method exploits the complementary behavior of each method.

Convergence rates

Let $f$ have $L$-Lipschitz gradient with respect to some norm $\| \cdot \|$ and $d_\phi(\cdot, \cdot)$ be a Bregman divergence w.r.t. some 1-strongly convex function $\phi$. Define $D = \sup_x d_\phi(x^*, x)$. Then,

$$f(y^T) - f(x^*) \leq \frac{4DL}{T^2}$$  \hspace{1cm} (14)
References

Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent.

[2] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford.
Lower bounds for finding stationary points of non-convex, smooth high-dimensional functions.
2017.

Restarting accelerated gradient methods with a rough strong convexity estimate.
2016.

Monotonicity and restart in fast gradient methods.

Introductory Lectures on Convex Optimization: A Basic Course.
Smooth minimization of non-smooth functions.

Adaptive restart for accelerated gradient schemes.