Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 5: Unconstrained, smooth minimization II

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

EE-556 (Fall 2016)

lions@epfl

MARIE CURIE ACTIONS

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

erc

EPFL

License Information for Mathematics of Data Slides

Outline

- This lecture
  1. Gradient and accelerated gradient descent methods
- Next lecture
  1. The quadratic case and conjugate gradient
  2. Other optimization methods

Recommended reading

- Chapters 2, 3, 5, 6 in Nocedal, Jorge, and Wright, Stephen J., *Numerical Optimization*, Springer, 2006.
- Chapter 9 in Boyd, Stephen, and Vandenberghe, Lieven, *Convex optimization*, Cambridge university press, 2009.
- Chapter 1 in Bertsekas, Dimitris, *Nonlinear Programming*, Athena Scientific, 1999.
- Chapters 1, 2 and 4 in Nesterov, Yurii, *Introductory Lectures on Convex Optimization: A Basic Course*, Vol. 87, Springer, 2004.

Motivation

Motivation

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

Recall: convex, unconstrained, smooth minimization

Problem (Mathematical formulation)

$$\boxed{F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{F(\mathbf{x}) := f(\mathbf{x})\right\}} \tag{1}$$

where f is proper, closed, convex and twice differentiable.
Note that (1) *is unconstrained.*

How de we design efficient optimization algorithms with accuracy-computation tradeoffs for this class of functions?

Basic principles of descent methods

Iterative descent

1. Let $\mathbf{x}^0 \in \text{dom}(f)$ be a starting point.
2. Generate a sequence of vectors $\mathbf{x}^1, \mathbf{x}^2, \cdots \in \text{dom}(f)$ so that we have descent:

$$f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k), \quad \text{for all } k = 0, 1, \ldots$$

   until $\mathbf{x}_k$ is ϵ-optimal.

Such a sequence $\left\{\mathbf{x}^k\right\}_{k \geq 0}$ can be generated as:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{p}^k$$

where $\mathbf{p}^k$ is a descent direction and $\alpha_k > 0$ a step-size.

Remark

- Iterative algorithms can implicitly use various **oracle** information from the objective, such as its value, gradient, or Hessian, in different ways to obtain α_k and $\mathbf{p}^k$, which determine their overall convergence rate and complexity. The type of oracle information they use becomes their defining characteristic.

Basic principles of descent methods

A condition for local descent directions

The iterates are given as:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{p}^k$$

By Taylor's theorem, we have

$$f(\mathbf{x}^{k+1}) = f(\mathbf{x}^k) + \alpha_k \langle \nabla f(\mathbf{x}^k),\ \mathbf{p}^k \rangle + O(\alpha_k^2).$$

For α_k small enough, the term $\alpha_k \langle \nabla f(\mathbf{x}^k),\ \mathbf{p}^k \rangle$ dominates $O(\alpha_k^2)$ for a fixed $\mathbf{p}^k$. Therefore, in order to have $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$, we require:

$$\langle \nabla f(\mathbf{x}^k),\ \mathbf{p}^k \rangle < 0$$

Local steepest descent direction

Since

$$\langle \nabla f(\mathbf{x}^k),\ \mathbf{p}^k \rangle = \|\nabla f(\mathbf{x}^k)\| \|\mathbf{p}^k\| \cos\theta \ ,$$

where θ is the angle between $\nabla f(\mathbf{x}^k)$ and $\mathbf{p}^k$, we have that

$$\mathbf{p}^k := -\nabla f(\mathbf{x}^k)$$

is the local *steepest descent* direction.

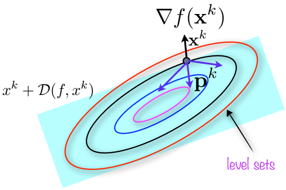


Figure: Descent directions in 2D should be an element of the cone of descent directions $\mathcal{D}(f, \cdot)$.

Gradient descent methods

Gradient descent (GD) algorithm

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

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

so that

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

Key question: How do we choose α_k so that we have descent/contraction?

Gradient descent methods

Gradient descent (GD) algorithm

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

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

so that

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

Key question: How do we choose α_k so that we have descent/contraction?

Answer: By exploiting the structures within the convex function

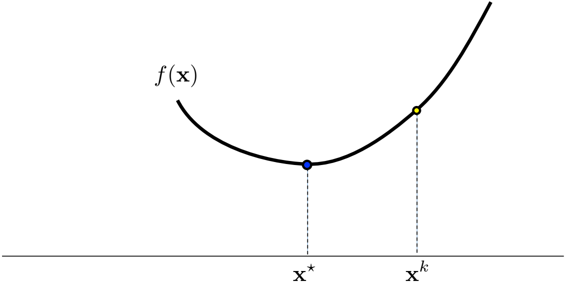
When $f \in \mathcal{F}_L^{2,1}$, we can use $\alpha_k = 1/L$ so that $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)$ is contractive.

- Note that the above GD method only uses the gradient information, and hence, it is called a **first-order method**.

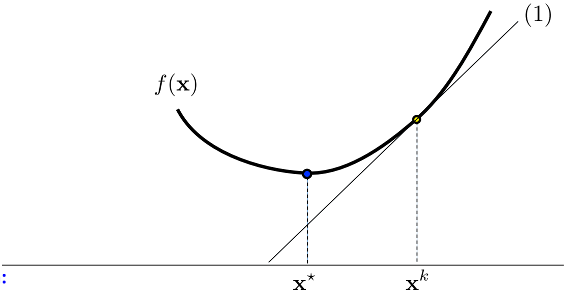
First-order methods employ only first-order oracle information about the objective, namely the value of f and ∇f at specific points.

- **Second-order methods** also use the Hessian $\nabla^2 f$.

Recall: Gradient descent methods - a geometrical intuition



Recall: Gradient descent methods - a geometrical intuition



Structure in optimization:

(1) $\qquad f(\mathbf{x}) \geq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle$

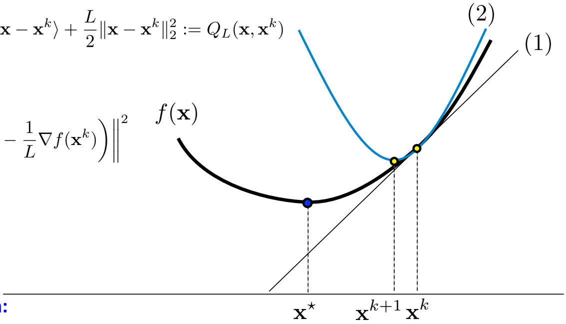
Recall: Gradient descent methods - a geometrical intuition

Majorize:

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

Minimize:

$$\begin{aligned} \mathbf{x}^{k+1} &= \arg\min_{\mathbf{x}} Q_L(\mathbf{x}, \mathbf{x}^k) \\ &= \arg\min_{\mathbf{x}} \left\| \mathbf{x} - \left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k) \right) \right\|^2 \\ &= \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k) \end{aligned}$$



Structure in optimization:

(1) $\quad f(\mathbf{x}) \geq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle$

(2) $\quad f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2$

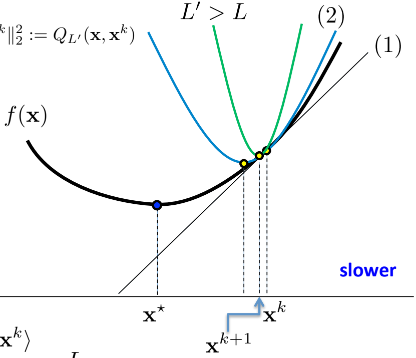
Recall: Gradient descent methods - a geometrical intuition

Majorize:

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

Minimize:

$$\begin{aligned} \mathbf{x}^{k+1} &= \arg\min_{\mathbf{x}} Q_{L'}(\mathbf{x}, \mathbf{x}^k) \\ &= \arg\min_{\mathbf{x}} \left\| \mathbf{x} - \left(\mathbf{x}^k - \frac{1}{L'}\nabla f(\mathbf{x}^k) \right) \right\|^2 \\ &= \mathbf{x}^k - \frac{1}{L'}\nabla f(\mathbf{x}^k) \end{aligned}$$



Structure in optimization:

(1) $\quad f(\mathbf{x}) \geq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle$

(2) $\quad f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2$

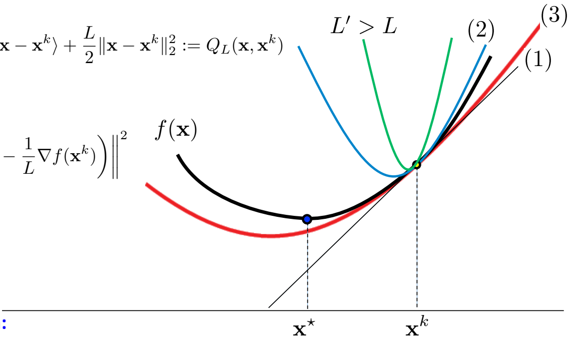
Recall: Gradient descent methods - a geometrical intuition

Majorize:

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

Minimize:

$$\begin{aligned}\mathbf{x}^{k+1} &= \arg\min_{\mathbf{x}} Q_L(\mathbf{x}, \mathbf{x}^k) \\ &= \arg\min_{\mathbf{x}} \left\| \mathbf{x} - \left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k) \right) \right\|^2 \\ &= \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\end{aligned}$$



Structure in optimization:

(1) $\quad f(\mathbf{x}) \geq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle$

(2) $\quad f(\mathbf{x}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2$

(3) $\quad f(\mathbf{x}) \geq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2$

Convergence rate of gradient descent

Theorem

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

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

$$\boxed{f(\mathbf{x}^k) - f(\mathbf{x}^\star) \leq \frac{2L}{k+4}\|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2}$$

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

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

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

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

Proof of convergence rates of gradient descent

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

Lemma

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

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

Proof.

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

$$h(\mathbf{x}) \leq h(\mathbf{y}) + \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2. \tag{3}$$

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

$$\phi(\mathbf{x}) \leq \phi\left(\mathbf{y} - \frac{1}{L}\nabla \phi(\mathbf{y})\right) \leq \phi(\mathbf{y}) - \frac{1}{2L}\|\nabla \phi(\mathbf{y})\|_2^2.$$

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

$$f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \leq f(\mathbf{y}) \tag{4}$$

By adding two copies of (4) with each other, with $\mathbf{x}$ and $\mathbf{y}$ swapped, we obtain (2). □

The proof of convergence rates - part I

Theorem

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

$$\boxed{f(\mathbf{x}^k) - f(\mathbf{x}^\star) \leq \frac{2L}{k+4}\|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2} \tag{5}$$

Proof - part I

- Consider the constant step-size iteration $\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha\nabla f(\mathbf{x}^k)$.
- Let $r_k := \|\mathbf{x}^k - \mathbf{x}^\star\|$. Show $\boxed{r_k \leq r_0}$.

$$\begin{aligned} r_{k+1}^2 &:= \|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2 = \|\mathbf{x}^k - \mathbf{x}^\star - \alpha\nabla f(\mathbf{x}^k)\|^2 \\ &= \|\mathbf{x}^k - \mathbf{x}^\star\|^2 - 2\alpha\langle\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^\star), \mathbf{x}^k - \mathbf{x}^\star\rangle + \alpha^2\|\nabla f(\mathbf{x}^k)\|^2 \\ &\leq r_k^2 - \alpha(2/L - \alpha)\|\nabla f(\mathbf{x}^k)\|^2 \quad \text{(by (2))} \\ &< r_k^2, \quad \forall\alpha < 2/L. \end{aligned}$$

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

- **An auxiliary result:** Let $\Delta_k := f(\mathbf{x}^k) - f^\star$. Show $\boxed{\Delta_k \leq r_0\|\nabla f(\mathbf{x}^k)\|}$.

$$\Delta_k \leq \langle\nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^\star\rangle \leq \|\nabla f(\mathbf{x}^k)\|\|\mathbf{x}^k - \mathbf{x}^\star\| = r_k\|\nabla f(\mathbf{x}^k)\| \leq r_0\|\nabla f(\mathbf{x}^k)\|.$$

The proof of convergence rates - part II

Proof - part II

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

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) &\le f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\
&\le f(\mathbf{x}^k) - \omega_k \|\nabla f(\mathbf{x}^k)\|^2, \quad \omega_k := \alpha(1 - L\alpha/2).
\end{aligned}
$$

  Subtract f^* from both sides and apply the last equation of the previous slide to get $\boxed{\Delta_{k+1} \le \Delta_k - (\omega_k/r_0^2)\Delta_k^2}$. Thus, dividing by $\Delta_{k+1}\Delta_k$

$$
\Delta_{k+1}^{-1} \ge \Delta_k^{-1} + (\omega_k/r_0^2)\Delta_k/\Delta_{k+1} \ge \Delta_k^{-1} + (\omega_k/r_0^2).
$$

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

$$
f(\mathbf{x}^k) - f(\mathbf{x}^\star) \le \frac{2(f(\mathbf{x}_0) - f(\mathbf{x}^\star))\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2}{2\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2 + k\alpha(2 - \alpha L)(f(\mathbf{x}_0) - f^\star)},
$$

- 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 $f \in \mathcal{F}_L^{1,1}$ is given by $\alpha = \frac{1}{L}$.
- Finally, since $f(\mathbf{x}_0) \le f^* + \nabla f(\mathbf{x}^\star)^T(\mathbf{x}_0 - \mathbf{x}^\star) + (L/2)\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2 = f^* + (L/2)r_0^2$, we obtain (5).

□

The proof of convergence rates - part III

Theorem

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

$$\boxed{\|\mathbf{x}^k - \mathbf{x}^\star\|_2 \le \left(\frac{L-\mu}{L+\mu}\right)^k \|\mathbf{x}^0 - \mathbf{x}^\star\|_2} \tag{6}$$

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

$$\boxed{\|\mathbf{x}^k - \mathbf{x}^\star\|_2 \le \left(\frac{L-\mu}{L+\mu}\right)^{\frac{k}{2}} \|\mathbf{x}^0 - \mathbf{x}^\star\|_2} \tag{7}$$

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

Theorem

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

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ge \frac{\mu L}{\mu + L}\|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{\mu + L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2. \tag{8}$$

The proof of convergence rates - part III

Proof of (6) and (7)

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

$$
\begin{aligned}
r_{k+1}^2 &= \|\mathbf{x}_{k+1} - \mathbf{x}^\star - \alpha \nabla f(\mathbf{x}^k)\|^2 \\
&= r_k^2 - 2\alpha \langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^\star \rangle + \alpha^2 \|\nabla f(\mathbf{x}^k)\|^2 \\
&\leq \left(1 - \frac{2\alpha\mu L}{\mu + L}\right) r_k^2 + \alpha \left(\alpha - \frac{2}{\mu + L}\right) \|\nabla f(\mathbf{x}^k)\|^2
\end{aligned}
$$

- 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 less than 0, and hence

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

- 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

Convergence rate of gradient descent

$$f \in \mathcal{F}_L^{2,1}, \quad \alpha = \frac{1}{L} \qquad\qquad f(\mathbf{x}^k) - f(\mathbf{x}^\star) \leq \frac{2L}{k+4}\|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2$$

$$f \in \mathcal{F}_{L,\mu}^{2,1}, \quad \alpha = \frac{2}{L+\mu} \qquad\qquad \|\mathbf{x}^k - \mathbf{x}^\star\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|\mathbf{x}^0 - \mathbf{x}^\star\|_2$$

$$f \in \mathcal{F}_{L,\mu}^{2,1}, \quad \alpha = \frac{1}{L} \qquad\qquad \|\mathbf{x}^k - \mathbf{x}^\star\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^{\frac{k}{2}} \|\mathbf{x}^0 - \mathbf{x}^\star\|_2$$

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 $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $\mathbf{b} \in \mathbb{R}^n$ given by the model $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w} \in \mathbb{R}^n$ is some noise.
- We can try to estimate $\mathbf{x}^\natural$ by solving the Tikhonov regularized least squares

$$\min_{\mathbf{x} \in \mathbb{R}^p} f(\mathbf{x}) := \frac{1}{2} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \frac{\rho}{2} \|\mathbf{x}\|_2^2.$$

where $\rho \geq 0$ is a regularization parameter.

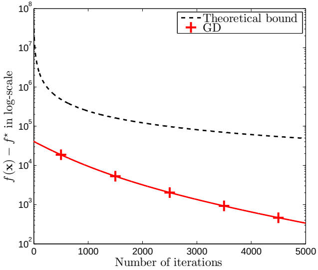
Remarks

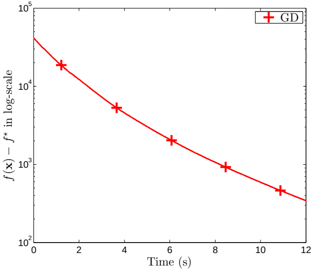
- $f \in \mathcal{F}_{L,\mu}^{2,1}$ with:
  - $L = \lambda_p(\mathbf{A}^T\mathbf{A}) + \rho$;
  - $\mu = \lambda_1(\mathbf{A}^T\mathbf{A}) + \rho$;
  - where $\lambda_1 \leq \ldots \leq \lambda_p$ are the eigenvalues of $\mathbf{A}^T\mathbf{A}$.
- The ratio $\frac{L}{\mu}$ decreases as ρ increases, leading 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:

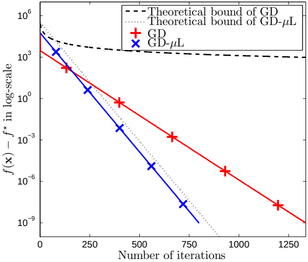
$n = 500, p = 2000, \rho = 0$

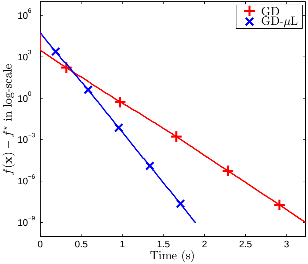




Case 2:

$n = 500, p = 2000, \rho = 0.01\lambda_p(\mathbf{A}^T\mathbf{A})$





Information theoretic lower bounds [2]

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

$f \in \mathcal{F}_L^{\infty,1}$: 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

$$f(\mathbf{x}^k) - f(\mathbf{x}^\star) \geq \frac{3L}{32(k+1)^2}\|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2 \quad \text{for all } k \leq (p-1)/2$$

$f \in \mathcal{F}_{L,\mu}^{\infty,1}$: 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

$$\|\mathbf{x}^k - \mathbf{x}^\star\|_2 \geq \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^k \|\mathbf{x}^0 - \mathbf{x}^\star\|_2$$

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!

Accelerated gradient descent algorithm

Problem

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

Global convergence of AGD [2]

Theorem (f is convex with Lipschitz gradient)

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

$$f(\mathbf{x}^k) - f^\star \leq \frac{4L}{(k+2)^2}\|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2, \ \forall k \geq 0. \tag{9}$$

Global convergence of AGD [2]

Theorem (f is convex with Lipschitz gradient)

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

$$f(\mathbf{x}^k) - f^\star \leq \frac{4L}{(k+2)^2}\|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2, \ \forall k \geq 0. \tag{9}$$

AGD-L is **optimal** *for $\mathcal{F}_L^{1,1}$ but NOT for $\mathcal{F}_{L,\mu}^{1,1}$!*

Theorem (f is strongly convex with Lipschitz gradient)

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

$$f(\mathbf{x}^k) - f^\star \leq L\left(1 - \sqrt{\frac{\mu}{L}}\right)^k \|\mathbf{x}^0 - \mathbf{x}^\star\|_2^2, \ \forall k \geq 0 \tag{10}$$

$$\|\mathbf{x}^k - \mathbf{x}^\star\|_2 \leq \sqrt{\frac{2L}{\mu}}\left(1 - \sqrt{\frac{\mu}{L}}\right)^{\frac{k}{2}} \|\mathbf{x}^0 - \mathbf{x}^\star\|_2, \ \forall k \geq 0. \tag{11}$$

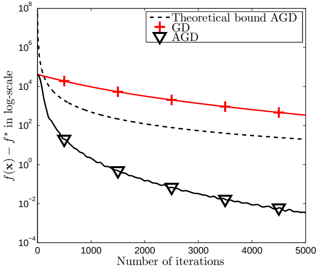
- AGD-L's iterates are not guaranteed to converge.
- AGD-L does not have a **linear** convergence rate for $\mathcal{F}_{L,\mu}^{1,1}$.
- AGD-μL does, but needs to know μ.

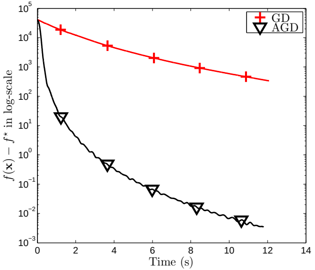
AGD achieves the iteration lowerbound within a constant!

Example: Ridge regression

Case 1:

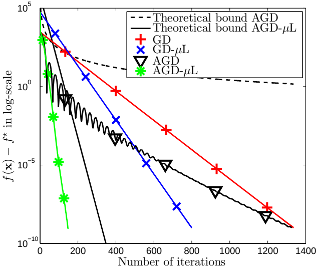
$$n = 500, p = 2000, \rho = 0$$

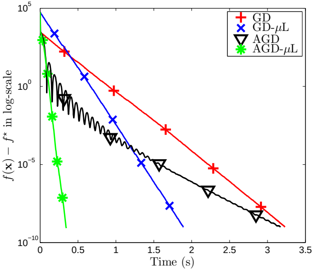




Case 2:

$$n = 500, p = 2000, \rho = 0.01\lambda_p(\mathbf{A}^T\mathbf{A})$$





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**;

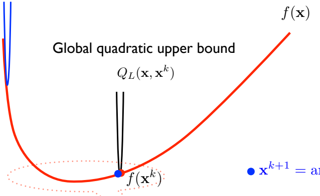
Line-search

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

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

Here: $L_0 > 0$ is given (e.g., $L_0 := c\frac{\|\nabla f(\mathbf{x}^1) - \nabla f(\mathbf{x}^0)\|_2}{\|\mathbf{x}^1 - \mathbf{x}^0\|_2}$) for $c \in (0, 1]$.

How can we better adapt to the local geometry?

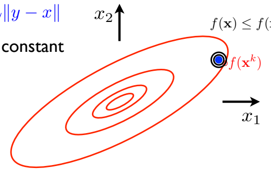


- $\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2 \right\}$

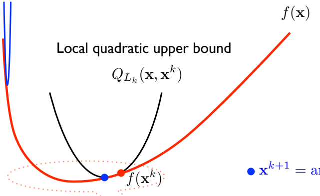
$$\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|$$

L is a global worst-case constant

$$f(\mathbf{x}) \leq f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|_2^2$$



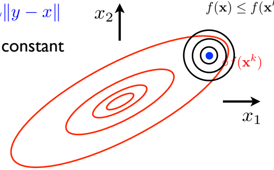
How can we better adapt to the local geometry?



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

$\|\nabla f(x) - \nabla f(y)\| \leq L\|y - x\|$

L is a global worst-case constant



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

applies only locally

Enhancements

Why do we need a restart strategy?

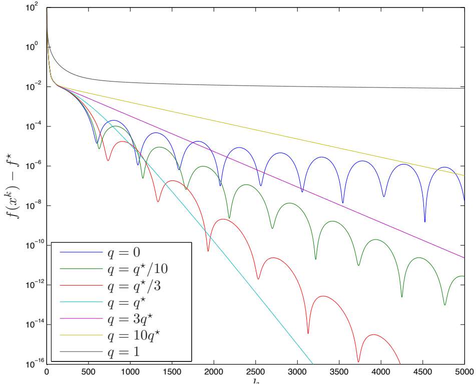
- AG-μL requires knowledge of μ and AG-L does not have optimal convergence for strongly convex f.
- AG is non-monotonic (i.e., $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k)$ is not always satisfied).
- AG has a periodic behavior, where the momentum depends on the local condition number $\kappa = L/\mu$.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

Two restart strategies

1. **O'Donoghue - Candes's strategy [3]:** 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 [1]:** Do not require $t_k = 1$ and do not necessary require function evaluations.

Oscillatory behavior of AGD

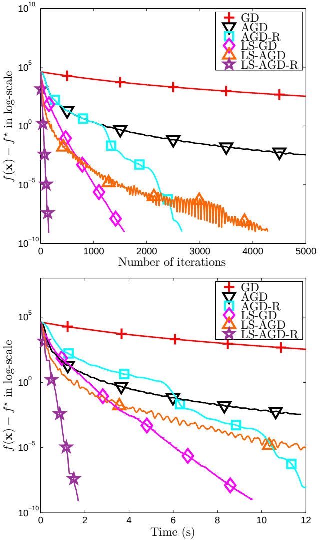
- Minimize a quadratic function $f(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Phi} \mathbf{x}$, with $p = 200$ and $\kappa(\boldsymbol{\Phi}) = L/\mu = 2.4 \times 10^4$
- Use stepsize $\alpha = 1/L$ and update $\mathbf{x}^{k+1} + \gamma_{k+1}(\mathbf{x}^{k+1} - \mathbf{x}^k)$ where
  - $\gamma_{k+1} = \theta_k(1-\theta_k)/(\theta_k^2 + \theta_{k+1})$
  - θ_{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.



Example: Ridge regression

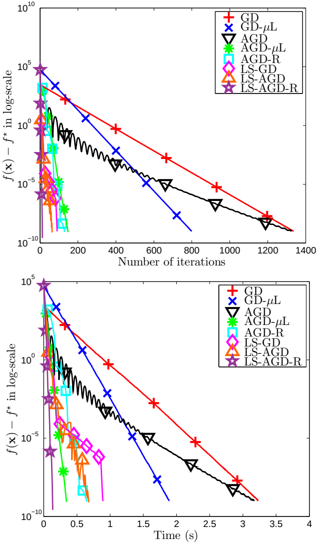
Case 1:

$n = 500, p = 2000, \rho = 0$



Case 2:

$n = 500, p = 2000, \rho = 0.01\lambda_p(\mathbf{A}^T\mathbf{A})$



The (special) quadratic case – Step-size

Consider the minimization of a quadratic function

$$\min_{\mathbf{x}} f(\mathbf{x}) := \frac{1}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle$$

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

Gradient Descent

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

Steepest descent

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

Barzilai-Borwein

$$\alpha_k = \frac{\|\nabla f(\mathbf{x}^{k-1})\|^2}{\langle \nabla f(\mathbf{x}^{k-1}), \mathbf{A}\nabla f(\mathbf{x}^{k-1})\rangle} \tag{13}$$

The (special) quadratic case – convergence rates

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

Theorem (Gradient Descent)

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

Theorem (Steepest Descent)

$$\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|_{\mathbf{A}} \leq \left(\frac{\lambda_p - \lambda_1}{\lambda_p + \lambda_1}\right)^k \|\mathbf{x}^0 - \mathbf{x}^\star\|_{\mathbf{A}}$$

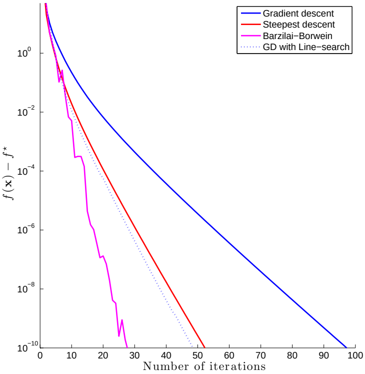
Theorem (Barzilai-Borwein)

Under the condition $\lambda_p < 2\lambda_1$

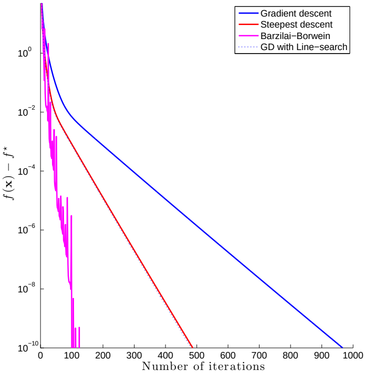
$$\|\mathbf{x}^{k+1} - \mathbf{x}^\star\|_2 \leq \left(\frac{\lambda_p - \lambda_1}{\lambda_1}\right)^k \|\mathbf{x}^0 - \mathbf{x}^\star\|_2$$

Example: Quadratic function

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



Case 1: $n = p = 100, \kappa(\mathbf{A}) = 100$



References I

[1] Pontus Giselsson and Stephen Boyd.
Monotonicity and restart in fast gradient methods.
In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 5058–5063. IEEE, 2014.

[2] Y. Nesterov.
Introductory lectures on convex optimization: A basic course, volume 87.
Springer, 2004.

[3] Brendan O'Donoghue and Emmanuel Candes.
Adaptive restart for accelerated gradient schemes.
Foundations of computational mathematics, 15(3):715–732, 2013.