ESE415 optimization

Xiaojian Xu

02/05/2019

Lecture 7
Convergence and Step-size

Guideline

SPEED OF CONVERGENCE

SELECTING THE STEP-SIZE FOR THE GRADIENT METHOD

1. Speed of Convergence

Definition1: Convergence

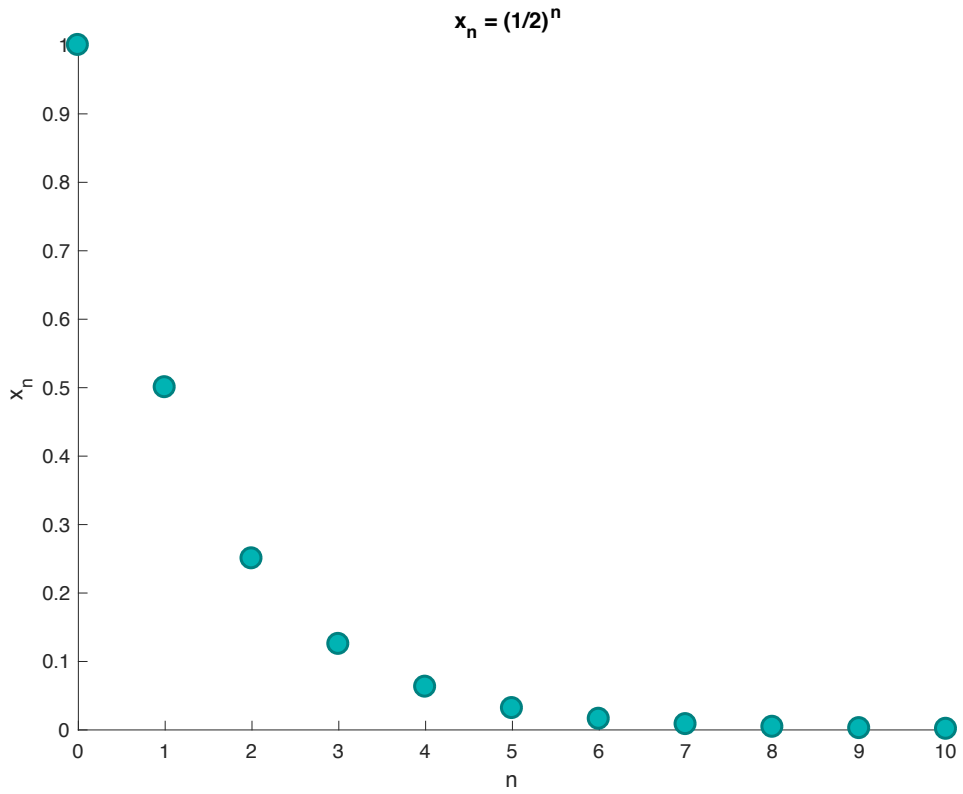
Definition 1. A sequence $\{\boldsymbol{x}_n\}$ in a metric space (X, d) is said to *converge* to $\boldsymbol{x} \in X$ if for every $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that $d(\boldsymbol{x}_n, \boldsymbol{x}) < \epsilon$ for every $n \geq N$. We denote the convergence as

$$\boldsymbol{x}_n \to \boldsymbol{x} \quad \text{or} \quad \lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{x}.$$

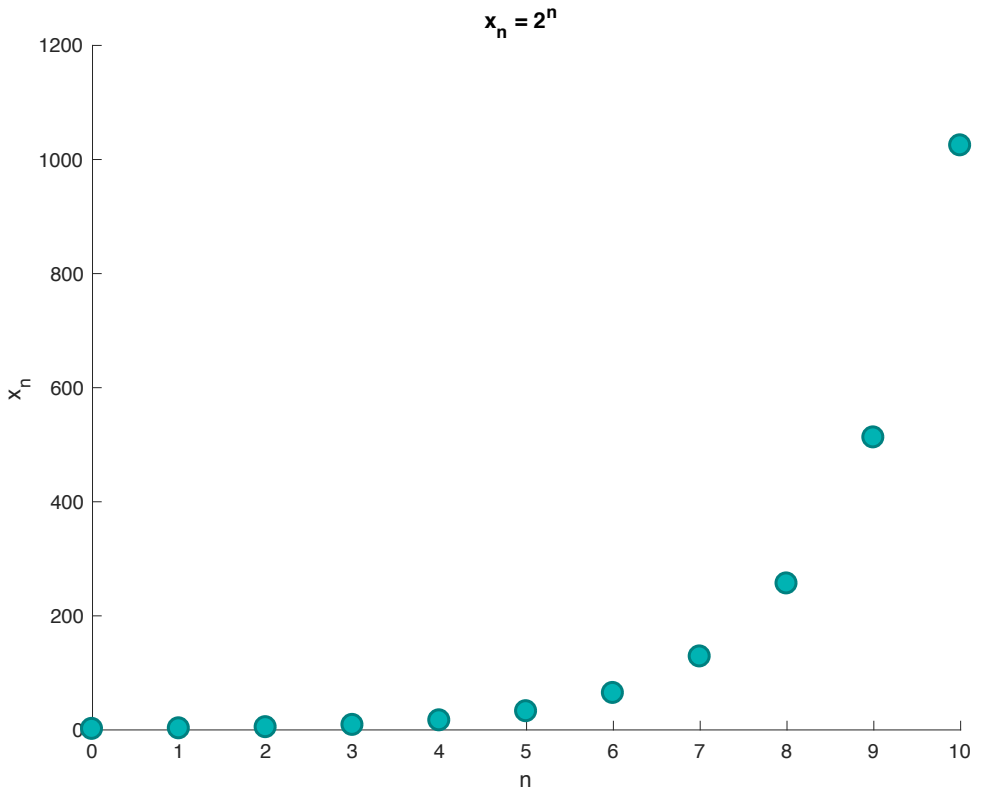
Note that a sequence which does not converge is said to be *divergent*.

Converge vs. Diverge

Convergent sequence



Divergent sequence



Definition of Supremum & Infimum

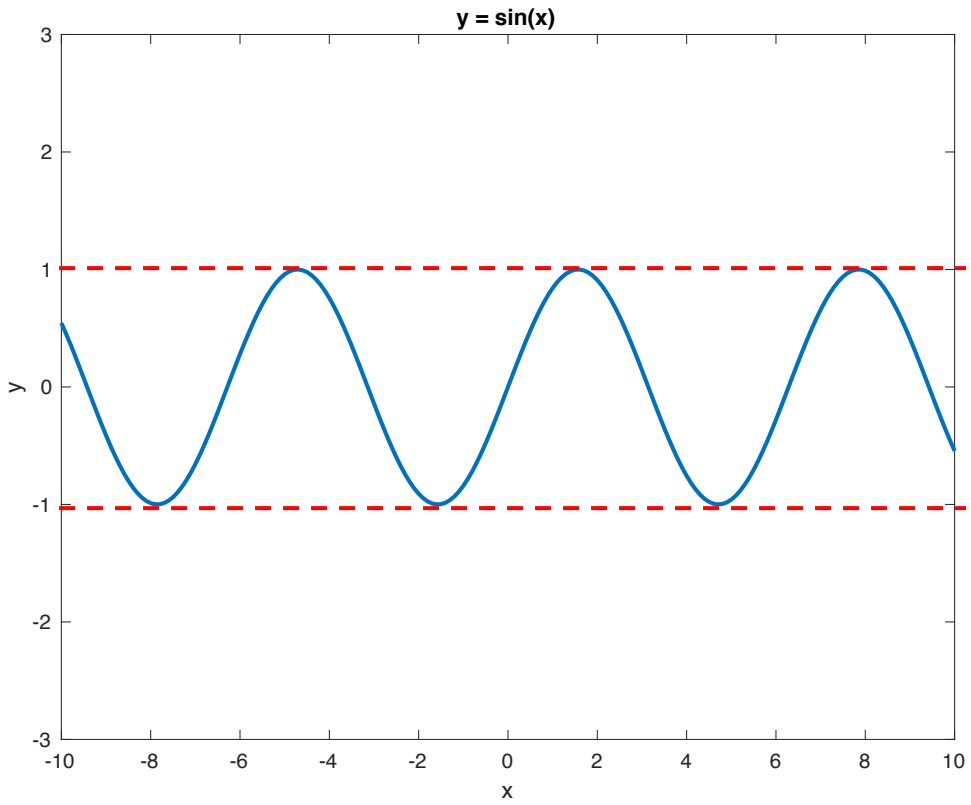
Upper bound & lower bound

Definition 2.1. A set $A \subset \mathbb{R}$ of real numbers is bounded from above if there exists a real number $M \in \mathbb{R}$, called an upper bound of A, such that $x \leq M$ for every $x \in A$. Similarly, A is bounded from below if there exists $m \in \mathbb{R}$, called a lower bound of A, such that $x \geq m$ for every $x \in A$. A set is bounded if it is bounded both from above and below.

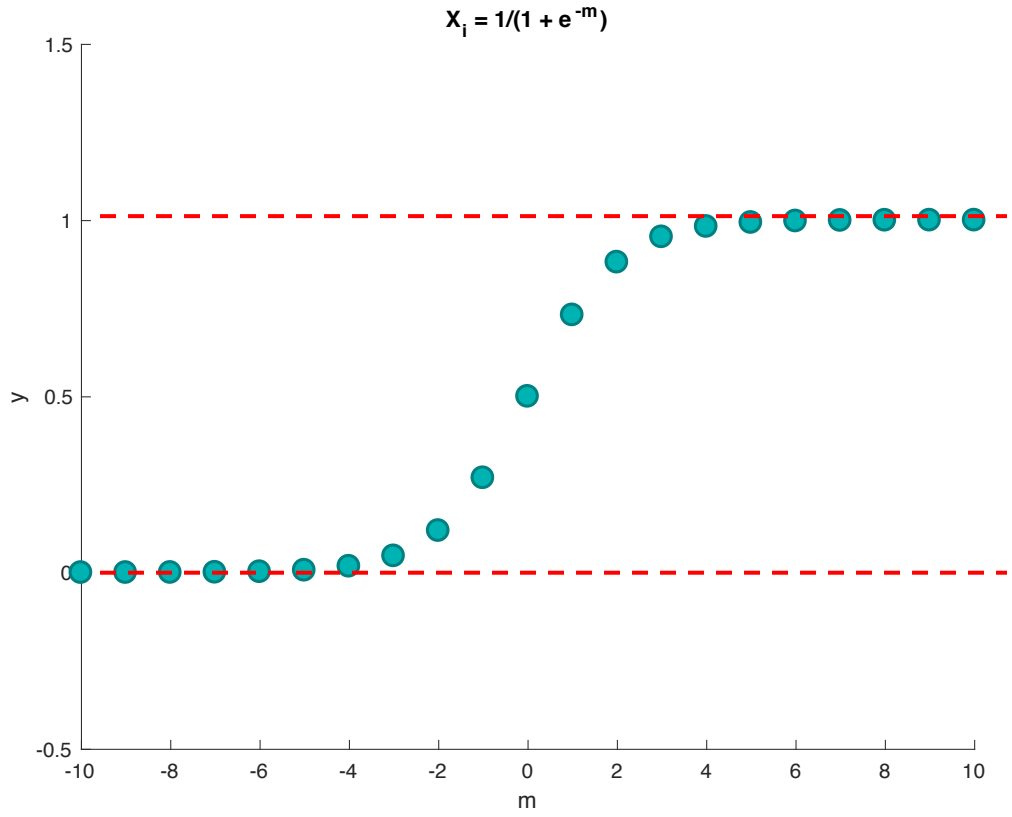
Supremum & Infimum

Definition 2.2. Suppose that $A \subset \mathbb{R}$ is a set of real numbers. If $M \in \mathbb{R}$ is an upper bound of A such that $M \leq M'$ for every upper bound M' of A, then M is called the supremum of A, denoted $M = \sup A$. If $m \in \mathbb{R}$ is a lower bound of A such that $m \geq m'$ for every lower bound m' of A, then m is called the or infimum of A, denoted $m = \inf A$.

Example for Supremum & Infimum



sup(y) = 1, inf(y) = -1



sup(y) = 1, inf(y) = 0

Definition 2: Order of Convergence

Definition 2. Let $\{f_k\}$ be a convergent sequence with $f_k \to f$. The *order of convergence* of $\{f_k\}$ is

$$p^* = \sup \left\{ p \in \mathbb{N} : \limsup_{k \to \infty} \frac{|f_k - f|}{|f_{k-1} - f|^p} < \infty \right\}.$$

The larger p^* implies faster convergence.

Remark of lim *sup* & lim *inf*

Remark. The definition above was given in terms of lim sup. Unlike the limit, the lim sup and lim inf of every bounded sequence of real numbers exists. A sequence converges if and only if its lim sup and lim inf are equal, in which case its limit is their common value. Suppose that $\{x_n\}$ is a sequence of real numbers. Then,

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} y_n \quad \text{where} \quad y_n = \sup\{x_k : k \geq n\}$$

$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty} z_n \quad \text{where} \quad z_n = \inf\{x_k : k \geq n\}.$$

Note that as n increases, the supremum and infimum are taken over smaller sets, so $\{y_n\}$ is monotone decreasing and $\{z_n\}$ is monotone increasing.

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} y_n \quad \text{where} \quad y_n = \sup\{x_k : k \geq n\}$$

$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty} z_n \quad \text{where} \quad z_n = \inf\{x_k : k \geq n\}.$$

Example. Consider the sequence

$$1, -1, 1, -1, \ldots \quad \Rightarrow \quad x_n = (-1)^n \quad \text{where} \quad n = 0, 1, 2, \ldots$$

Then

$$y_n = \sup\{(-1)^k : k \geq n\} = 1 \quad \text{and} \quad z_n = \inf\{(-1)^k : k \geq n\} = -1.$$

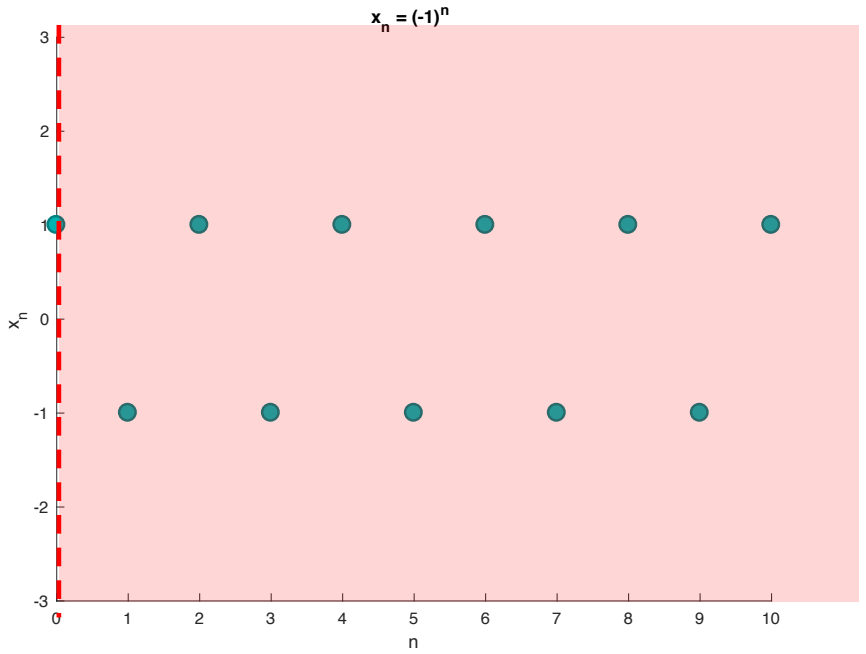
We then have

$$\limsup_{n\to\infty} x_n = 1 \quad \text{and} \quad \liminf_{n\to\infty} x_n = -1.$$

Note that the original sequence does not converge and $\lim_{n\to\infty} x_n$ is undefined.

$Y_0 = 1$

$Y_1 = 1$



Example of lim *sup* & lim *inf*

Definition 3: Convergence Description

Definition 3. Let p denote the order of convergence of $\{f_k\}$ to f and let

$$\beta = \limsup_{k \to \infty} \frac{|f_k - f|}{|f_{k-1} - f|^p}.$$

We say that the convergence is

- *linear* or *geometric* if $p = 1$ and $\beta \in (0, 1)$.
- *superlinear* if $p = 1$ and $\beta = 0$.
- *sublinear* if $p = 1$ and $\beta = 1$.
- *quadratic* if $p = 2$ and $\beta < \infty$.

An Example of Linear convergence

Consider $f_k = (0.1)^k$, which implies that

$$f_1 = \frac{1}{10}, f_2 = \frac{1}{100}, \dots \quad \text{so that} \quad f_k \to f = 0.$$

Then, we have that

$$\frac{|f_{k+1} - f|}{|f_k - f|^p} = \frac{\left|\frac{1}{10}\right|^{k+1}}{\left|\frac{1}{10}\right|^{kp}} = \left(\frac{1}{10}\right)^{k(1-p)+1}.$$

Hence, we have $p = 1$ and

$$\lim_{k\to\infty} \frac{|f_{k+1} - f|}{|f_k - f|} = \frac{1}{10},$$

which implies the linear convergence.

2. Selecting the Step-size for the Gradient Method

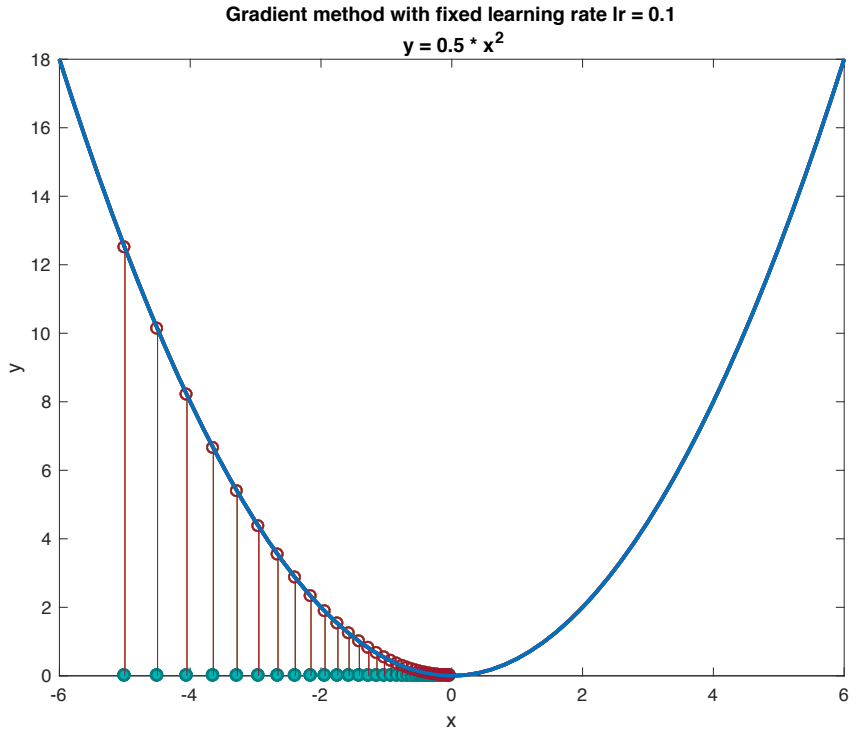
Selecting the step-size for the gradient method

1. Fixed
2. Decreasing
3. Exact line-search
4 .Backtracking line-search

1.Fixed step size

$$\boldsymbol{x}^t = \boldsymbol{x}^{t-1} - \gamma_t \nabla f(\boldsymbol{x}^{t-1}), \quad (t = 1, 2, 3, \dots)$$

- *Fixed*: The step-size is fixed to a constant $\gamma_t = \gamma > 0$. In Lecture 5, we saw that for a Lipschitz continuous f with a constant L, the gradient method converges for any $\gamma \in (0, 2/L)$.

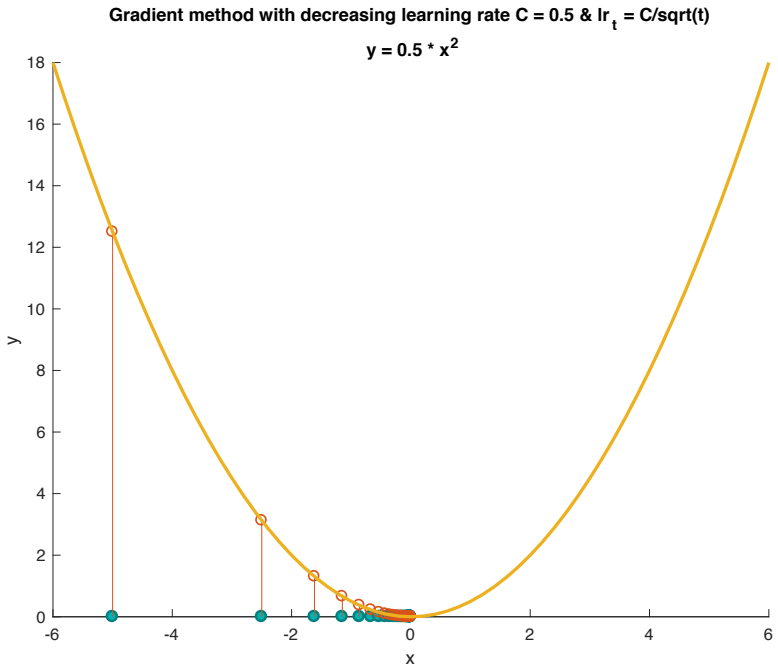


2.Decreasing step size

- *Decreasing*: In some applications, it is common to set the step-size to a fixed decreasing sequence. For example

$$\gamma_t = \frac{C}{t} \quad \text{or} \quad \gamma_t = \frac{C}{\sqrt{t}},$$

where $t = 1, 2, \ldots$ and $C > 0$ is some constant.



3.Exact line-search

- *Exact line-search*: In every iteration, one can choose the step-size to minimize

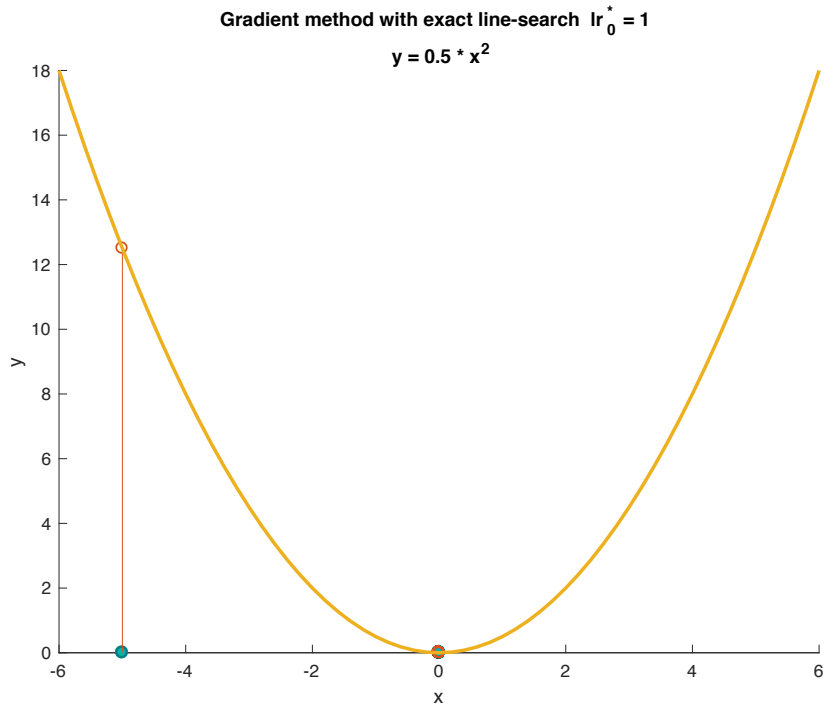
$$\gamma_t = \underset{\gamma \in \mathbb{R}_+}{\arg\min} \{\phi(\gamma)\} \quad \text{where} \quad \phi(\gamma) = f\left(\boldsymbol{x}^{t-1} - \gamma \nabla f(\boldsymbol{x}^{t-1})\right).$$

$f(x) = \frac{1}{2}x^2,\ \nabla f(x_0) = x,\ x_0 = -5;$

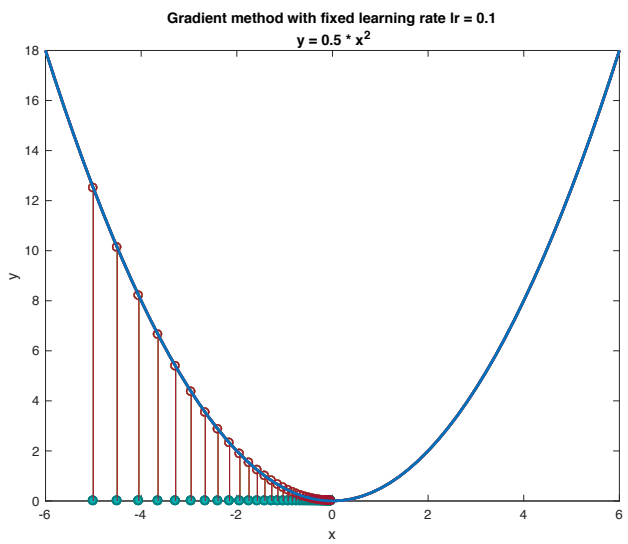
$\emptyset(\gamma_1) = f(x_0 - \gamma_1 \nabla f(x_0)) = \frac{25}{2}(1-\gamma_1)^2;$

$\gamma_1^* = argmin_{\gamma_1}(\emptyset(\gamma_1)) = argmin_{\gamma_1} \frac{25}{2}(1-\gamma_1)^2 = 1;$

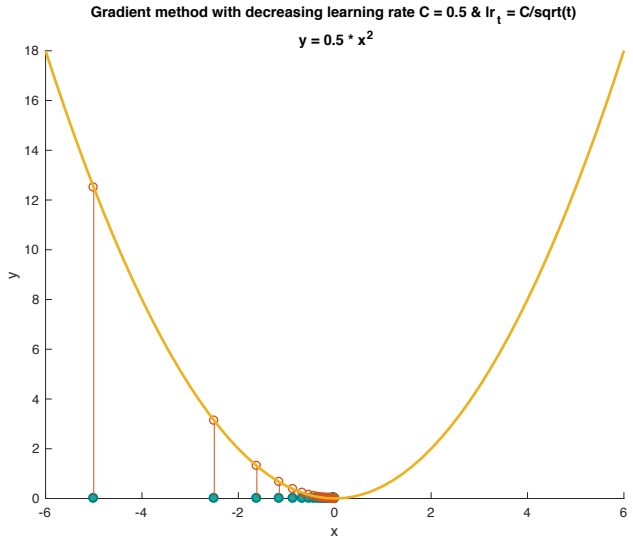
$x_1 = x_0 - \gamma_1^* * x_0 = x_0 - 1 * x_0 = 0$



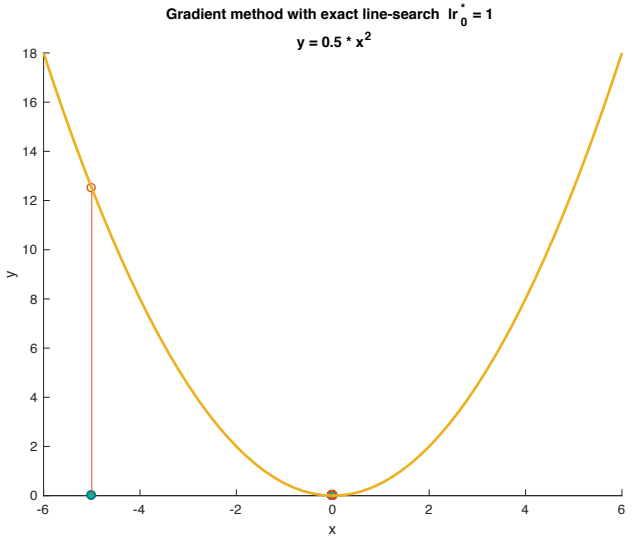
Fixed



Decreasing



Exact line-search



Comparison

4. Backtracking line-search

- *Backtracking line-search*: Majority of line-searches used in practice are *inexact*, where the step is chosen to approximately minimize f along the ray given by the gradient.

$$f(\boldsymbol{x}^+) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\mathsf{T}(\boldsymbol{x}^+ - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2,$$

$$f(\boldsymbol{x}^+) \leq q_\gamma(\boldsymbol{x}^+, \boldsymbol{x}) \triangleq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\mathsf{T}(\boldsymbol{x}^+ - \boldsymbol{x}) + \frac{1}{2\gamma}\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2$$

- *If γ above is well selected (small enough) , the inequation should always hold; If it doesn't hold, it means the γ we select is bad (too big).*
- *So correspondingly, with this well selected (small enough) γ , what can we guarantee?*

4. Backtracking line-search

- *For a well selected (small enough) γ, the following holds:*

- *For a well selected (small enough) γ, the following holds:*
- *Which implies:*

- *So, for any bad (too big) step size γ,* it breaks the bound above:
- We shrink it by a constant parameter:

- *Which implies:*

$$
\begin{aligned}
f(\boldsymbol{x}^t) &\leq f(\boldsymbol{x}^{t-1}) + \nabla f(\boldsymbol{x}^{t-1})^\mathsf{T}(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}) + \frac{1}{2\gamma_t}\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|^2 \\
&= f(\boldsymbol{x}^{t-1}) + \nabla f(\boldsymbol{x}^{t-1})^\mathsf{T}[-\gamma_t \nabla f(\boldsymbol{x}^{t-1})] + \frac{1}{2\gamma_t}\|\gamma_t \nabla f(\boldsymbol{x}^{t-1})\|^2 \\
&= f(\boldsymbol{x}^{t-1}) - \frac{\gamma_t}{2}\|\nabla f(\boldsymbol{x}^{t-1})\|^2.
\end{aligned}
$$

- *So, for any bad (too big) step size γ,* it breaks the bound above:

$$
f(\boldsymbol{x}^{t-1} - \gamma_t \nabla f(\boldsymbol{x}^{t-1})) > f(\boldsymbol{x}^{t-1}) - \frac{\gamma_t}{2}\|\nabla f(\boldsymbol{x}^{t-1})\|^2
$$

- We shrink it by a constant parameter:

$$
\gamma_t = \beta \gamma_t
$$

Remark

Remark. The discussion above leads to an elegant interpretation of the gradient method as a *majorize-minimize scheme*. To understand the scheme, consider the quadratic function on the right-hand side of eq. (2)

$$q_\gamma(\boldsymbol{x},\boldsymbol{y}) = f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\mathsf{T}(\boldsymbol{x}-\boldsymbol{y}) + \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{y}\|^2.$$

For a well chosen and fixed step-size $\gamma > 0$, we have that

$$\begin{aligned} f(\boldsymbol{x}) &\le q_\gamma(\boldsymbol{x},\boldsymbol{y}) \quad \text{for all} \quad \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n \\ f(\boldsymbol{x}) &= q_\gamma(\boldsymbol{x},\boldsymbol{x}) \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{R}^n. \end{aligned}$$

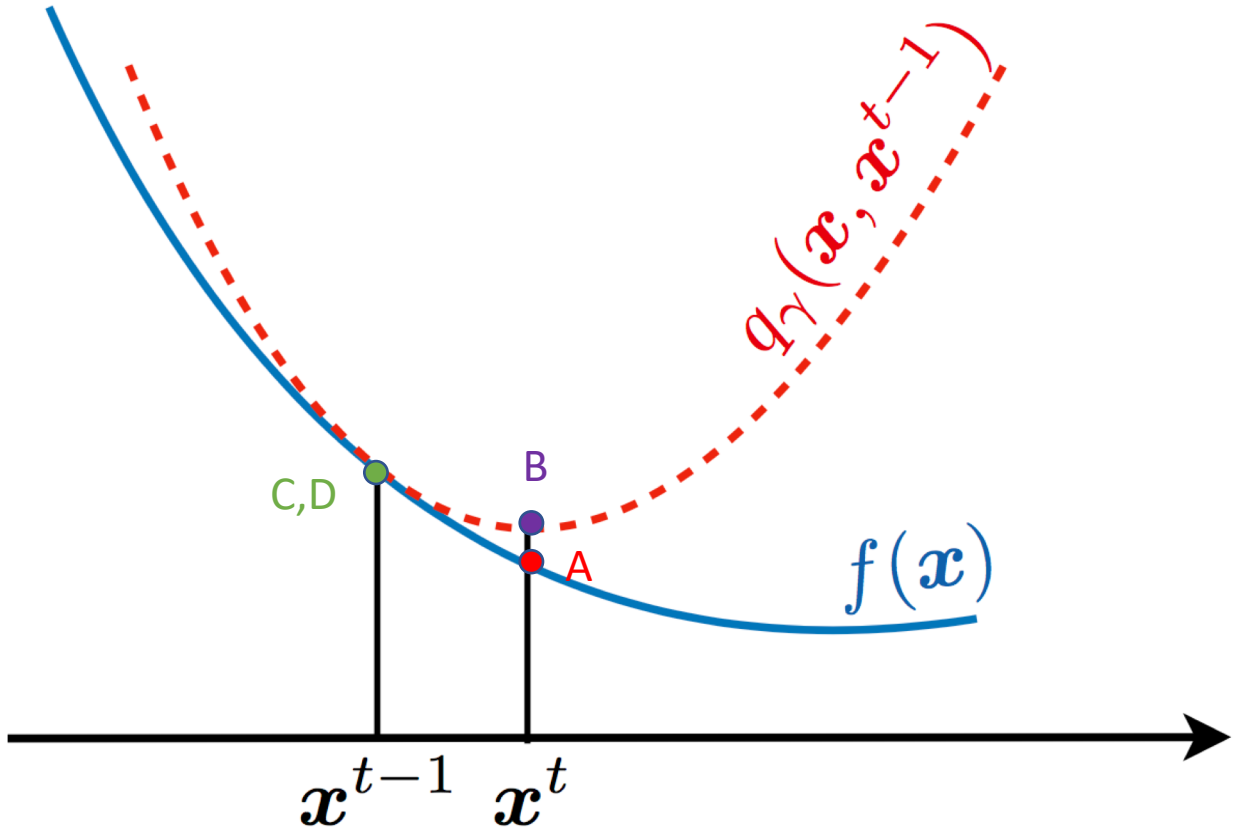
Geometrically, this means that $q_\gamma(\boldsymbol{x},\boldsymbol{y})$ lies above $f(\boldsymbol{x})$ and is tangent at $\boldsymbol{x} = \boldsymbol{y}$ (see Fig. 1). The gradient method updates the iterates by minimizing the quadratic upper-bound as

$$\boldsymbol{x}^t = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ q_\gamma(\boldsymbol{x},\boldsymbol{x}^{t-1}) \right\}.$$

This means that $q_\gamma(\boldsymbol{x}^t,\boldsymbol{x}^{t-1}) \le q_\gamma(\boldsymbol{x},\boldsymbol{x}^{t-1})$ for all $\boldsymbol{x}$, which directly implies that

$$f(\boldsymbol{x}^t) \le q_\gamma(\boldsymbol{x}^t,\boldsymbol{x}^{t-1}) \le q(\boldsymbol{x}^{t-1},\boldsymbol{x}^{t-1}) = f(\boldsymbol{x}^{t-1})$$

for every $t \in \mathbb{N}$, thus guaranteeing a monotonic reduction in the objective at every iteration.



$f(x^t) \le q_r(x^t,x^t) \le q_r(x^{t-1},x^{t-1}) = f(x^{t-1})$;
$A \qquad B \qquad C \qquad D$

$A = f(x^t)$; $\quad B = q_r(x^t,x^t)$; $\quad C = q_r(x^{t-1},x^{t-1})$; $\quad D = f(x^{t-1})$;

$A \le B$, $B \le C = D$, so $A \le C = D$, which guarantees a reduction in each iteration.

Thanks