Chapter 7: Optimization for Data Science
Gradient Descent

TANN Chantara

Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia

October 8, 2022

Table of Contents

Unconstrained Optimization Problem

Unconstrained Minimization Problem

$$\min_{x\in\mathbb{R}^n} f(x) = f(x^*)$$

- $f : \mathbb{R}^n \to \mathbb{R}$ convex and differentiable.
- Necessary and sufficient conditions $\nabla f(x^*) = 0 \implies$ Set of n (nonlinear) equations with n variables

Goal: Find an approximate solution $\tilde{x} \in \mathbb{R}^n$ such that

$$f(\tilde{x}) - f(x^*) < \varepsilon$$

- compute $\tilde{x}$ iteratively via an algorithm (e.g., gradient descent)

Gradient Descent

- Iterative algorithm $x_{t+1} = x_t + v_t$
- Choose step v_t such that $f(x_{t+1}) < f(x_t)$
- Taylor series

$$f(x_t + v_t) = f(x_t) + \nabla f(x_t)^\top v_t + \underbrace{r(v_t)}_{o(\|v_t\|)} \approx f(x_t) + \nabla f(x_t)^\top v_t$$

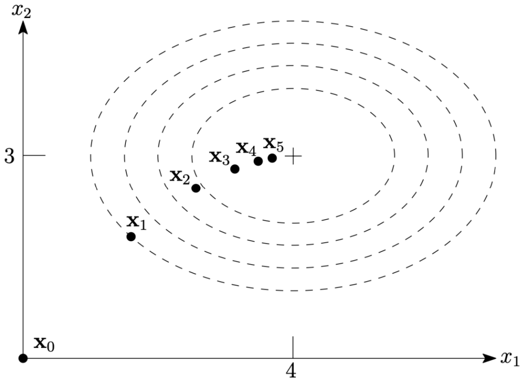
  - " $\approx$ " requires step size $||v_t||$to be small.
  - want $\nabla f(x_t)^T v_t <) \implies v_t = -\nabla f(x_t)$ not small

Gradient descent $x_{t+1} = x_t - \gamma \nabla f(x_t)$

- Step size $\gamma > 0$, how to choose it?
  - γ "too small" $\implies$ gradient descent takes long to converge.
  - γ "too large" $\implies$ gradient descent might overshoot.

Gradient Descent Cont'd

Example: Run of gradient descent on the quadratic function $f(x_1, x_2) = 2(x_1 - 4)^2 + 3(x_2 - 3)^2$ with global minimum $x^* = (4, 3)$; we have choose $x_0 = (0, 0), \gamma = 0.1$; dashed lines represent level sets of f (points of constant f-value).



Theoretical Analysis

Gradient descent: $x_{t+1} = x_t - \gamma \nabla f(x_t)$

Lemma 1 (Property of Gradient Descent):

$$\sum_{t=0}^{T-1} (f(x_t) - f(x^*)) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} ||\nabla f(x_t)||^2 + \frac{1}{2\gamma} ||x_0 - x^*||^2$$

- Clearly, $\min_{t-0,\ldots,T-1} f(x_t) - f(x^*) \leq \frac{1}{T} \sum_{t=0}^{T-1} (f(x_t) - f(x^*))$
- Dependence on $||x_0 - x^*||$ to be expected (if we start far away, we need more steps).
- Need to control the gradient $||\nabla f(x_t)||^2$

Proof of Lemma 1

The gradient descent can be equivalently written as

$$x_{t+1} = x_t - \gamma \nabla f(x_t) \iff \nabla f(x_t) = \frac{1}{\gamma}(x_t - x_{t+1})$$

Recall for any $v, w \in \mathbb{R}^n, 2v^T w = ||v||^2 + ||w||^2 - ||v - w||^2$. Hence

$$\begin{aligned}
\nabla f(x_t)^T (x_t - x^*) &= \frac{1}{\gamma}(x_t - x_{t+1})^T (x_t - x^*) \\
&= \frac{1}{2\gamma}(||x_t - x_{t+1}||^2 + ||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) \\
&= \frac{\gamma}{2}||\nabla f(x_t)||^2 + \frac{1}{2\gamma}(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2)
\end{aligned}$$

Proof of Lemma 1 Cont'D

Hence, applying a telescopic sum

$$\sum_{t=0}^{T-1} \nabla f(x_t)^T (x_t - x^*) = \frac{\gamma}{2} \sum_{t=0}^{T-1} ||\nabla f(x_t)||^2 + \frac{1}{2\gamma}(||x_0 - x^*||^2 - ||x^T - x^*||^2)$$

$$= \frac{\gamma}{2} \sum_{t=0}^{T-1} ||\nabla f(x_t)||^2 + \frac{1}{2\gamma}||x_0 - x^*||^2 \qquad (\star)$$

Since f is convex the first-order convexity conditions state

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^T (x_t - x^*)$$

which combined with $(\star)$ leads to

$$\sum_{t=0}^{T-1} (f(x_t) - f(x^*)) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} ||\nabla f(x_t)||^2 + \frac{1}{2\gamma}||x_0 - x^*||^2$$

This proves **Lemma 1**.

Lipschitz Continuous Functions

Definition: A function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $l > 0$ if $|f(x) - f(y)| \leq l||x - y||, \forall x, y \in \mathbb{R}^n$

- f is l Lipschitz $\iff ||\nabla f(x)|| \leq l, \forall x \in \mathbb{R}^n$
- Ex: $f(x) = x^2$ is not Lipschitz cont. as $\nabla f(x) = 2x$ is unbounded.

Lipschitz Continuous Functions Cont'd

Theorem (Gradient Descent for Lipschitz Function): Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable with global minimum x^*. Suppose that $||x_0 - x^*|| \leq R$ and $||\nabla f(x)|| \leq B, \forall x \in \mathbb{R}^n$. Choosing the step size $\gamma = \frac{R}{B\sqrt{T}}$, the gradient descent yields.

$$\frac{1}{T}\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \frac{RB}{\sqrt{T}}$$

- For achieving $\min_{t=0,\dots,T-1} f(x_t) - f(x^*) \leq \varepsilon$, we need $T \geq \frac{R^2B^2}{\varepsilon^2}$ $\implies$ # of iterations scale as $\mathcal{O}(1/\varepsilon^2)$.
- No specific dependence on n.

Proof of Theorem

Using Lemma 1 directly leads to

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^\star)) \leq \underbrace{\frac{\gamma}{2}TB^2 + \frac{1}{2\gamma}R^2}_{=:q(\gamma)},$$

which holds for any $\gamma > 0$. Solving $\min_{\gamma>0} q(\gamma)$ gives the optimal step size as $\gamma^* = \dfrac{R}{L\sqrt{T}}$ and $q(\gamma^*) = RB\sqrt{T}$. Hence,

$$\frac{1}{T}\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \frac{RB}{\sqrt{T}}$$

which completes the proof.

- What happen if we do not know R and/or B?
- Can we improve the $\mathcal{O}(1/\varepsilon^2)$ complexity?

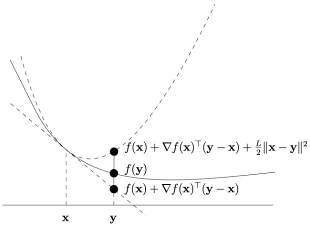
Smooth Convex Functions

Recall the first order convexity conditions

$$f(y) \geq f(x) + \nabla f(x)^T (y - x), \forall x, y \in \mathbb{R}^n$$

Definition: Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $\mathbb{X} \subset \mathbb{R}^n$ be convex and $L > 0$. Then function f is called L-smooth over $\mathbb{X}$, if

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} ||y - x||^2, \forall x, y \in \mathbb{X}$$



- No convexity of f required
- Ex: $f(x) = x^2$ is smooth with $L = 2$
- How can we easily check smoothness of a function?

Properties of Smooth Functions

Lemma: Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. The following are equivalent

(i) f is L-smooth.

(ii) $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||, \forall x, y \in \mathbb{R}^n$

(iii) (ii) $\implies$ (i) holds without convexity

Lemma (Smoothness Preserving Operation):

(i) Let $f_1, ..., f_m$ be smooth with parameters $L_1, ..., L_m$, let $\lambda_1, ..., \lambda_m > 0$. Then the function $f = \sum_{i=1}^{m} \lambda_i f_i$ is L-smooth with $L = \sum_{i=1} \lambda_i L_i$ over $\text{dom}(f) = \cap_{i=1}^{m} \text{dom}(f_i)$.

(ii) Consider $f : \mathbb{R}^n \to \mathbb{R}$ L-smooth and $g : \mathbb{R}^m \to \mathbb{R}^n$ affine, i.e., $g(x) = Ax + b$. Then $f(g(x)) = f(Ax + b)$ is smooth with parameter $L||A||^2$, where $||A||$ = the spectral norm.

Convergence analysis for smooth functions

Lemma 2 (Sufficient Decrease): Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, differentiable and L-smooth. For $\gamma = 1/L$, gradient descent yields

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||\nabla f(x_t)||^2, t \geq 0$$

- GD (with suitable stepsize γ) makes progress in function value on smooth functions in every step.

Proof:

$$\begin{aligned} f(x_{t+1}) &\overset{L\text{-smooth}}{\leq} f(x_t) + \nabla f(x_t)^\top \underbrace{(x_{t+1} - x_t)}_{-1/L\nabla f(x_t)} + \frac{L}{2}\underbrace{\|x_t - x_{t+1}\|^2}_{1/L^2\|\nabla f(x_t)\|^2} \\ &= f(x_t) - \frac{1}{L}\|\nabla f(x_t)\|^2 + \frac{1}{2L}\|\nabla f(x_t)\|^2 \\ &= f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2 \end{aligned}$$

Convergence analysis for smooth functions Cont'd

Theorem (Gradient Descent for Smooth Functions): Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex with global minimum x^*, differentiable and L-smooth. For $\gamma = 1/L$, gradient descent yields

$$f(x_T) - f(x^*) \leq \frac{1}{2L}||x_0 - x^*||^2, \quad T > 0$$

- For $R = ||x_0 - x^*||$, to get $f(x_T) - f(x^*) \leq \varepsilon$, we need $T \geq \dfrac{R^2 L}{2\varepsilon}$
  $\implies$ complexity $\mathcal{O}(1/\varepsilon)$

Lipschitz functions:	Smooth functions:
$T = \mathcal{O}(1/\varepsilon^2)$	$T = \mathcal{O}(1/\varepsilon)$

- Could we do even better? E.g., we could achieve $\mathcal{O}(1/\sqrt{\varepsilon})$ or $\mathcal{O}(1/\log \varepsilon)$?

Proof of Theorem

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^\star)) \overset{\text{Lemma 1}}{\leq} \frac{\gamma}{2}\sum_{t=0}^{T-1} \underbrace{\|\nabla f(x_t)\|^2}_{\overset{\text{Lemma 2}}{\leq} 2L(f(x_t)-f(x_{t+1}))} + \frac{1}{2\gamma}\|x_0 - x^\star\|^2$$

$$\leq \quad f(x_0) - f(x_T) + \frac{L}{2}\|x_0 - x^\star\|^2$$

This is equivalent to

$$\sum_{t=1}^{T}(f(x_t) - f(x^*)) \leq \frac{L}{2}||x_0 - x^*||^2 \qquad (1)$$

Recall that from Lemma 2, we know that $f(x_{t+1}) \leq f(x_t)$. Hence, take averages in (1) gives

$$f(x_T) - f(x^*) \leq \sum_{t=1}^{T}(f(x_t) - f(x^*) \leq \frac{L}{2T}||x_0 - x^*||^2$$

Acceleration for Smooth Convex Functions

Accelerated gradient descent[1]:
Choose $z_0 = y_0 = x_0$ arbitrarily.
For $t \geq 0$ set

$$
\begin{aligned}
y_{t+1} &= x_t - \frac{1}{L}\nabla f(x_t) \\
z_{t+1} &= z_t - \frac{t+1}{2L}\nabla f(x_t) \\
x_{t+1} &= \frac{t+1}{t+3} y_{t+1} + \frac{2}{t+3} z_{t-1}
\end{aligned}
$$

[1] Yurii Nesterov 1983

Acceleration for Smooth Convex Functions Cont'd

Theorem (Accelerated Gradient Descent): Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex with global minimum x^*, differentiable and L-smooth. Accelerated gradient descent yields

$$f(y_T) - f(x^*) \leq \frac{2L||z_0 - x^*||^2}{T(T+1)}, \quad T > 0$$

- To reach error ε, we need $\mathcal{O}(1/\sqrt{\varepsilon})$ steps.

An Observation

- Consider the smooth function $f(x) = x^2$. Gradient descent according to our theorem ensures.

$$f(x_T) \leq \frac{1}{T} x_0^2$$

- For $\gamma = 1/L = 1/2$, gradient descent yields

$$x_{t+1} = x_t - 1/2 \nabla f(x_t) = x_t - x_t = 0$$

$\implies$ we converge in only one step

- For a suboptimal (valid) step size $\gamma = 1/4$, gradient descent yields

$$x_{t+1} = x_t - 1/4 \nabla f(x_t) = \frac{x_t}{2} \implies f(x_T) = f(x_0/2^T) = \frac{x_0^2}{2^{2^T}}$$

To achieve $f(x_T) \leq \varepsilon$, we require $T \approx \frac{1}{2} \log(x_0^2/\varepsilon)$

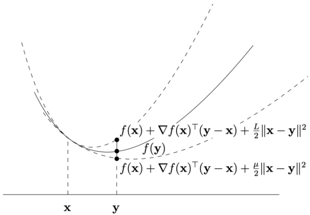
Strong Convexity

Recall the first order convexity conditions

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

Definition: Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $\mathbb{X} \subset \mathbb{R}^n$ be convex and $L > 0$. The function f is called μ-strongly convex over $\mathbb{X}$, if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}||x - y||^2, \forall x, y \in \mathbb{X}$$



- Smooth and strongly convex function
- Smooth $\rightarrow$ “not too curved"
- Strongly convex $\rightarrow$ “not too flat"
- Ex: $f(x) = x^2$ is smooth and strongly convex

Smooth and Strongly Convex Case

Theorem (Smooth and Strongly Convex Function): Let $f : \mathbb{R}^n \to \mathbb{R}$ be μ-strongly convex with global minimum x^*, differentiable and L-smooth. For a step size $\gamma = 1/L$, gradient descent yields

(i) $||x_{t+1} - x^*||^2 \leq (1 - \frac{\mu}{L})||x_t - x^*||^2, t \geq 0$

(ii) $f(x_T) - f(x^*) \leq \frac{L}{2}(1 - \frac{\mu}{L})^T ||x_0 - x^*||^2, T > 0$

- For $R = ||x_0 - x^*||$, to achieve $f(x_T) - f(x^*) \leq \varepsilon$, using (i) one can chooseT $\log(1 - \frac{\mu}{L}) \leq \log(\frac{2\varepsilon}{LR^2}) \qquad (2)$
  Using the bound $\log(1 - \frac{\mu}{L}) \leq -\frac{\mu}{L}$, the condition (2) is implied by

$$T \geq \frac{L}{\mu} \log\left(\frac{R^2 L}{2\varepsilon}\right)$$

- Hence, the over iteration complexity is $\mathcal{O}(\log(1/\varepsilon))$.

Proof of Theorem

We first show (i):

$$
\begin{aligned}
& f(x_t) - f(x^\star) + \frac{\mu}{2}\|x_t - x^\star\|^2 \\
& \overset{\mu \text{ strongly convex}}{\leq} \quad \nabla f(x_t)^\top (x_t - x^\star) \\
& \overset{\text{Proof of Lemma 1}}{\leq} \quad \frac{\gamma}{2}\|\nabla f(x_t)\|^2 + \frac{1}{2\gamma}(\|x_t - x^\star\|^2 - \|x_{t+1} - x^\star\|^2)
\end{aligned}
$$

Therefore,

$$
f(x_t) - f(x^\star) \leq \frac{1}{2\gamma}(\gamma^2\|\nabla f(x_t)\|^2 + \|x_t - x^\star\|^2 - \|x_{t+1} - x^\star\|^2) - \frac{\mu}{2}\|x_t - x^\star\|^2,
$$

which is equivalent to

$$
\|x_{t+1} - x^\star\|^2 \leq 2\gamma(f(x_t) - f(x^\star)) + \gamma^2\|\nabla f(x_t)\|^2 + (1 - \mu\gamma)\|x_t - x^\star\|^2 \quad (3)
$$

Proof of Theorem

$$f(x^\star) - f(x_t) \le f(x_{t+1}) - f(x_t) \overset{\text{Lemma 2}}{\le} -\frac{\gamma}{2}\|\nabla f(x_t)\|^2$$

$$\begin{aligned}
&\Longrightarrow 2\gamma(f(x^\star) - f(x_t)) + \gamma^2\|\nabla f(x_t)\|^2 \le 0 \\
&\overset{(3)}{\Longrightarrow} \|x_{t+1} - x^\star\|^2 \le (1-\mu\gamma)\|x_t - x^\star\|^2 = (1 - \tfrac{\mu}{L})\|x_t - x^\star\|^2 \\
&\Longrightarrow \|x_T - x^\star\|^2 \le (1 - \tfrac{\mu}{L})^T\|x_0 - x^\star\|^2,
\end{aligned}$$

which shows (i). To show (ii), note that

$$\begin{aligned}
f(x_T) - f(x^\star) &\overset{\text{smooth}}{\le} \nabla f(x^\star)(x_T - x^\star) + \tfrac{L}{2}\|x_T - x^\star\|^2 \\
&\overset{\nabla f(x^\star)=0}{=} \tfrac{L}{2}\|x_T - x^\star\|^2 \\
&\overset{(i)}{\le} \tfrac{L}{2}(1 - \tfrac{\mu}{L})^T\|x_0 - x^\star\|^2,
\end{aligned}$$

Main Take-Away Points

- **Definitions:** smooth convex functions, strongly convex functions
- **Iteration complexity analysis for gradient descent:** For a convex function f.

	Lipschitz	smooth	smooth & strongly con.
gradient descent	$\mathcal{O}(1/\varepsilon^2)$	$\mathcal{O}(1/\varepsilon)$	$\mathcal{O}(\log(1/\varepsilon))$
acc. gradient desc.		$\mathcal{O}(1/\sqrt{\varepsilon})$	