Recitation 14

Carles Domingo

Convergence of gradient descent

Given an initial point $x_0 \in \mathbb{R}^n$, the gradient descent algorithm follows the updates:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t), \tag{1}$$

Definition (Smoothness and strong convexity)

For $L, \mu > 0$, we say that a twice-differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$ is

- L-smooth if for all $x \in \mathbb{R}^n$, $\lambda_{\max}(Hf(x)) \leq L$.
- μ-strongly convex if for all $x \in \mathbb{R}^n$, $\lambda_{\min}(Hf(x)) \geq \mu$.

L-smooth and μ-strongly convex functions are very convenient since they can be “sandwiched” as follows (see homework 9 for a proof): for all $x, h \in \mathbb{R}^n$,

$$f(x) + \langle h, \nabla f(x) \rangle + \frac{\mu}{2}\|h\|^2 \leq f(x+h) \leq f(x) + \langle h, \nabla f(x) \rangle + \frac{L}{2}\|h\|^2, \tag{2}$$

Convergence of gradient descent

Theorem (Convex functions)

Assume that f is convex, L-smooth and that f admits a (global) minimizer $x^\star \in \mathbb{R}^n$. Then the gradient descent iterates (1) with constant step-size $\alpha_t = 1/L$ verify

$$f(x_t) - f(x^\star) \leq \frac{2L\|x_0 - x^\star\|^2}{t+4} \tag{3}$$

Theorem (Strongly convex functions)

Assume that f is L-smooth and μ-strongly convex. Then f admits a unique minimizer global $x^\star$ and the gradient descent iterates (1) with constant step-size $\alpha_t = 1/L$ verify

$$f(x_t) - f(x^\star) \leq \left(1 - \frac{\mu}{L}\right)^t (f(x_0) - f(x^\star)). \tag{4}$$

Convergence of gradient descent

Proof of the result for strongly convex functions. Let $t \geq 0$. Applying (2) for $x = x_t$ and $h = -L^{-1}\nabla f(x_t)$, we get $f(x_{t+1}) \leq f(x_t) - \frac{1}{L}\|\nabla f(x_t)\|^2 + \frac{1}{2L}\|\nabla f(x_t)\|^2 = f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2$. Now, since f is μ-strongly convex, we have (**exercise!**) for all $x \in \mathbb{R}^n$,

$$f(x) - f(x^\star) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

We get that $f(x_{t+1}) \leq f(x_t) - \frac{\mu}{L}(f(x_t) - f(x^\star))$, hence

$$f(x_{t+1}) - f(x^\star) \leq \left(1 - \frac{\mu}{L}\right)(f(x_t) - f(x^\star)),$$

from which the theorem follows.

Convergence of gradient descent

Show that if f is a μ-strongly convex with minimizer $x^\star$, then for all $x \in \mathbb{R}^n$,

$$f(x) - f(x^\star) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

Convergence of gradient descent

Show that if f is a μ-strongly convex with minimizer $x^\star$, then for all $x \in \mathbb{R}^n$,

$$f(x) - f(x^\star) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

Problem 0.14, 2019 review

Assume that we are doing standard gradient descent to minimize the least-square cost $f(x) = \|Ax - y\|^2$. Assume that the columns of A are linearly dependent, meaning that Ker$(A) \neq \{0\}$. At which speed should gradient descent converge to the minimum? If now Ker$(A) = \{0\}$, at which speed should gradient descent converge? By speed, we only ask about the dependence in t, the number of iterations, of the gap $f(x_t) - \min f$, where x_t is the position of gradient descent after t iterations.

Problem 0.14, 2019 review

Assume that we are doing standard gradient descent to minimize the least-square cost $f(x) = \|Ax - y\|^2$. Assume that the columns of A are linearly dependent, meaning that $\mathrm{Ker}(A) \neq \{0\}$. At which speed should gradient descent converge to the minimum? If now $\mathrm{Ker}(A) = \{0\}$, at which speed should gradient descent converge? By speed, we only ask about the dependence in t, the number of iterations, of the gap $f(x_t) - \min f$, where x_t is the position of gradient descent after t iterations.

Problem 0.14, 2019 review

Assume that we are doing standard gradient descent to minimize the least-square cost $f(x) = \|Ax - y\|^2$. Assume that the columns of A are linearly dependent, meaning that Ker$(A) \neq \{0\}$. At which speed should gradient descent converge to the minimum? If now Ker$(A) = \{0\}$, at which speed should gradient descent converge? By speed, we only ask about the dependence in t, the number of iterations, of the gap $f(x_t) - \min f$, where x_t is the position of gradient descent after t iterations.

Accelerated gradient methods

Theorem (Gradient descent with momentum)

We add a momentum term to gradient descent:

$$x_{t+1} = x_t + v_t \text{ where } v_t = -\alpha_t \nabla f(x_t) + \beta_t v_{t-1}, \tag{5}$$

for some α_t, β_t. If α_t, β_t are chosen appropriately, for f L-smooth and μ-strongly convex,

$$\|x_t - x^\star\| \leq \left(\frac{\sqrt{L} - \sqrt{l}}{\sqrt{L} + \sqrt{l}}\right)^t \|x_0 - x^\star\| \tag{6}$$

Accelerated gradient methods

Theorem (Nesterov's accelerated gradient descent)

The updates are of the form

$$x_{t+1} = x_t + v_t \text{ where } v_t = \alpha_t v_{t-1} - \beta_t \nabla f(x_t + \alpha_t v_{t-1}) \tag{7}$$

If f is L-smooth and μ-strongly convex, and if its minimum is attained at some $x^\star$, then for $\alpha_t = \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}$ and $\beta_t = 1/L$ we have

$$f(x_t) - f(x^\star) \leq L\|x_0 - x^\star\|^2(1 - \mu/L)^t. \tag{8}$$

Newton's method

Theorem

Newton's method performs updates according to

$$x_{t+1} = x_t - Hf(x_t)^{-1} \nabla f(x_t). \tag{9}$$

When f is μ-strongly convex and L-smooth, for t large enough

$$\|x_t - x^\star\|_2 \leq Ce^{-\rho 2^t}, \tag{10}$$

where $C, \rho > 0$ are constants depending on f and x_0.

Convergence of Newton's method

1. Show that if $(x_t)_{t\geq 0}$ are iterates of Newton's method for f μ-strongly convex and L-smooth, then

$$\frac{L}{2\mu^2}\|\nabla f(x_{t+1})\| \leq \left(\frac{L}{2\mu^2}\|\nabla f(x_t)\|\right)^2$$

2. Show that for a μ-strongly convex differentiable function and any $x, y \in \mathbb{R}^n$,

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2$$

   "The gradient of a strongly convex function is a strongly monotone operator."

3. Show that if the gradient at initialization is small enough, then

$$\|x_t - x^\star\|_2 \leq Ce^{-\rho 2^t},$$

   where $C, \rho > 0$ are constants depending on f and x_0. Identify these constants.

Convergence of Newton's method

1. Show that if $(x_t)_{t \geq 0}$ are iterates of Newton's method for f μ-strongly convex and L-smooth, then

$$\frac{L}{2\mu^2} \|\nabla f(x_{t+1})\| \leq \left(\frac{L}{2\mu^2} \|\nabla f(x_t)\| \right)^2$$

Convergence of Newton's method

2. Show that for a μ-strongly convex differentiable function and any $x, y \in \mathbb{R}^n$,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$$

"The gradient of a strongly convex function is a strongly monotone operator."

Convergence of Newton's method

3. Show that if the gradient at initialization is small enough, then

$$\|x_t - x^\star\|_2 \le Ce^{-\rho 2^t},$$

where $C, \rho > 0$ are constants depending on f and x_0. Identify these constants.

Problem 0.15, 2019 review

Let $A \in \mathbb{R}^{n \times d}$. Assume that the columns of A are linearly independent. How many steps of Newton's method do you need to minimize $\|Ax - y\|^2$? ($y \in \mathbb{R}^n$ is a fixed vector). Justify your answer.

Problem 0.15, 2019 review

Let $A \in \mathbb{R}^{n \times d}$. Assume that the columns of A are linearly independent. How many steps of Newton's method do you need to minimize $\|Ax - y\|^2$? ($y \in \mathbb{R}^n$ is a fixed vector). Justify your answer.

Stochastic gradient descent

Instead of the full-gradient $\nabla R_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\theta)$, updates are of the following form:

1. Pick i uniformly at random in $\{1, \dots, N\}$,
2. Update $\theta_{t+1} = \theta_t - \alpha_t \nabla f_i(\theta_t)$.

Classical results on stochastic gradient descent show that

- If the f_i are μ-strongly convex and smooth, then SGD with step sizes $\alpha_t = 1/(\mu t)$ achieves after t steps an error of $O(1/t)$.
- If the f_i are convex and smooth, then SGD with step sizes $\alpha_t = 1/\sqrt{t}$ achieves after t steps an error of $O(1/\sqrt{t})$.

Convergence of SGD

In the appendix of the lecture notes (beginning of page 8), it is shown that

$$\mathbb{E}R(\theta_{t+1}) \leq (1 - \mu\alpha_t)\mathbb{E}R(\theta_t) + L\alpha_t^2\sigma^2.$$

The last steps of the proof are sketched but not detailed. Starting from this equation, show that if $\alpha_t = \frac{2}{\mu t}$, then

$$\mathbb{E}R(\theta_t) \leq \frac{2L\sigma^2}{\mu^2 t}$$

Convergence of SGD

In the appendix of the lecture notes (beginning of page 8), it is shown that

$$\mathbb{E}R(\theta_{t+1}) \leq (1 - \mu\alpha_t)\mathbb{E}R(\theta_t) + L\alpha_t^2\sigma^2.$$

The last steps of the proof are sketched but not detailed. Starting from this equation, show that if $\alpha_t = \frac{2}{\mu t}$, then

$$\mathbb{E}R(\theta_t) \leq \frac{2L\sigma^2}{\mu^2 t}$$

Problem 0.16, 2019 review

When running stochastic gradient descent, what are upsides and downsides of having a rapidly decaying learning rate?

Problem 0.16, 2019 review

When running stochastic gradient descent, what are upsides and downsides of having a rapidly decaying learning rate?