IFT 6085 - Lecture 2
Gradient descent for smooth and for strongly convex functions

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribes: Philippe Brouillard, Massimo and Lucas Caccia **Instructor:** Ioannis Mitliagkas

1 Summary

In the previous lecture we covered notions of convexity and gradient descent for Lipschitz functions.

In this lecture we will cover gradient descent for smooth and for strongly convex functions.

2 Gradient Descent for Lipschitz functions

(see [1] Section 3.1 for more details)

Before we start with new material, let us go back to the end of last class, where we proved Theorem 3.2. Let's clarify how to go from equations (9) to (10) (in the previous set of scribe notes).

You have

$$f(x_k) - f(x^*) \leq \langle \nabla f(x_k), x_k - x^* \rangle$$

Using the gradient update rule, we have $x_{k+1} = x_k - \gamma \nabla f(x_k)$, or equivalently $\nabla f(x_k) = \frac{1}{\gamma}(x_k - x_{k+1})$

Giving us

$$= \langle \frac{1}{\gamma}(x_k - x_{k+1}), x_k - x^* \rangle$$

From the identity $||a-b||_2^2 = ||a||_2^2 + ||b||_2^2 - 2a^\intercal b$, or $a^\intercal b = \frac{1}{2}(-||a-b||_2^2 + ||a||_2^2 + ||b||_2^2)$
Setting $a = (x_k - x_{k+1})$ and $b = x_k - x^*$, we get

$$= \frac{1}{2\gamma}(-||x_k - x_{k+1} - (x_k - x^*)||_2^2 + ||x_k - x_{k+1}||_2^2 + ||x_k - x^*||_2^2)$$

Replacing the second term by the gradient and cleaning up the first term we get

$$= \frac{1}{2\gamma}(-||x^* - x_{k+1}||_2^2 + ||\gamma \nabla f(x_k)||_2^2 + ||x_k - x^*||_2^2)$$

And we recover equation (10) from the last set of notes

$$f(x_k) - f(x^*) \leq \frac{1}{2\gamma}(||x_k - x^*||_2^2 - ||x_{k+1} - x^*||_2^2) + \frac{\gamma}{2}||\nabla f(x_k)||_2^2$$

3 Gradient descent for smooth functions

(see [1] Section 3.2 for more details)

Definition 1 (β-smoothness)**.** *We say that a continously differentiable function f is β-smooth if the gradient ∇f is β-Lipschitz, that is*

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

Note that if f is twice differentiable then this is equivalent to the eigenvalues of the Hessians being smaller than β.

The next theorem shows that gradient descent, which iterates $x_{t+1} = x_t \gamma \nabla f(x_t)$, attains a much faster rate in this situation than in the non-smooth case of the previous lecture.

3.1 Theorem 3.3

Theorem 2. *Let f be convex and β-smooth on $\mathbb{R}^n$. Then the gradient descent with $\gamma = 1/\beta$ satisfies*

$$f(x_k) - f(x^*) \leq \frac{2\beta \|x_1 - x^*\|^2}{k - 1}$$

Before embarking on the proof we state a few properties of smooth convex functions.

Lemma 3. *Let f be a β-smooth on $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$, one has*

$$|f(x) - f(y) - \nabla f(y)^\intercal (x - y)| \leq \frac{\beta}{2} \|x - y\|^2$$

See [1] Lemma 3.4 for proof.

Lemma 4. *Let f be such that $0 \leq f(x) - f(y) - \nabla f(y)^\intercal (x - y) \leq \frac{\beta}{2} \|x - y\|$. Then for any $x, y \in \mathbb{R}^n$, one has*

$$f(x) - f(y) \leq \nabla f(x)^\intercal (x - y) - \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|^2$$

See [1] Lemma 3.5 for proof.

We can now prove Theorem 2

Proof. See [1] page 268 □

4 Strong convexity

Definition 5 (α-strong convexity)**.** *We say that f is α-strongly convex if it satisfies the following subgradient inequality:*

$$f(x) - f(y) \leq \nabla f(x)^\intercal (x - y) - \frac{\alpha}{2} \|x - y\|^2$$

The strong convexity parameter α is a measure of the curvature of f.

4.1 Theorem 3.9

Theorem 6. *Let f be α-strongly convex and L-Lipschitz. Then the projected subgradient descent with $\gamma_k = \frac{2}{\alpha(k+1)}$ satisfies*

$$f\left(\sum_{k=1}^{T} \frac{2k}{T(T-1)} x_k\right) - f(x^*) \leq \frac{2L^2}{\alpha(k+1)}$$

Proof. See [1] page 277 □

4.2 Theorem 3.12

Theorem 7. *For f a λ-strongly convex and β-smooth function, the gradient descent with $\gamma = \frac{2}{\lambda+\beta}$ satisfies:*

$$f(x_{k+1}) - f(x^*) \leq \frac{\beta}{2} \exp\left(-\frac{4k}{\kappa+1}\right) \|x_1 - x^*\|^2$$

where κ is the condition number.

In order to prove the theorem, we will use the following lemma:

Lemma 8 (Coercivity of the gradient)**.** *Let f be β-smooth and a-strongly. The for all x and y, we have:*

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

See [1] Lemma 3.11 for proof.

Proof. (Theorem 7) First, let's define the distance between the coordinate at the iteration k and the optimal point as:

$$D_k \triangleq \|x_k - x^*\|$$

Then,

$$D_{k+1}^2 = \|x_{k+1} - x^*\|^2$$

By replacing x_{k+1} by its definition $x_k - \gamma\nabla f(x_k)$, we get:

$$= \|x_k - \gamma\nabla f(x_k) - x^*\|^2$$

By expanding the square, we get:

$$= \|x_k - x^*\|^2 - 2\gamma\langle \nabla f(x_k), x_k - x^* \rangle + \gamma^2\|\nabla f(x_k)\|^2$$

By doing a slight modification to the second term, we can apply the lemma of coercivity of the gradient. Since $\nabla f(x^*) = 0$ this equality holds:

$$\langle \nabla f(x_k), x_k - x^* \rangle = \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle$$

And by applying the lemma of coercivity of the gradient, we get an upper bound:

$$\geq \frac{\lambda\beta}{\lambda+\beta} D_k^2 + \frac{1}{\lambda+\beta}\|\nabla f(x_k) - \nabla f(x^*)\|^2$$

By replacing the second term by the upper bound we just found, we get:

$$D_{k+1}^2 \leq D_k^2 - 2\gamma\left(\frac{\lambda\beta}{\lambda+\beta} D_k^2 + \frac{1}{\lambda+\beta}\|\nabla f(x_k)\|^2\right) + \gamma^2\|\nabla f(x_k)\|^2$$

We can rearrange the terms and add $-\nabla f(x^*)$ again inside the norm terms:

$$= \left(1 - \frac{2\gamma\lambda\beta}{\lambda+\beta}\right) D_k^2 + \left(\frac{-2\gamma}{\lambda+\beta} + \gamma^2\right) \|\nabla f(x_k) - \nabla f(x^*)\|^2$$

We can change the norm term by using the fact that f is β-smooth:

$$\leq \left(1 - \frac{2\gamma\lambda\beta}{\lambda+\beta}\right) D_k^2 + \left(\frac{-2\gamma}{\lambda+\beta} + \gamma^2\right) \beta^2 D_k^2$$

Let $\gamma = \frac{2}{\lambda+\beta}$, then:

$$D_{k+1}^2 \leq \left(1 - \frac{4\lambda\beta}{(\lambda+\beta)^2}\right) D_k^2$$

By unrolling the recursion and since $\left(\frac{\kappa-1}{\kappa+1}\right)^2 = \left(1 - \frac{4\lambda\beta}{(\lambda+\beta)^2}\right)$ we get:

$$\leq \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} D_1^2$$

Since $\exp(-x) \geq 1 - x$ for every x, we get:

$$\leq \exp\left(-\frac{4k}{\kappa+1}\right) D_1^2$$

By β-smoothness we finally have:

$$f(x_{k+1}) - f(x^*) \leq \frac{\beta}{2} \exp\left(-\frac{4k}{\kappa+1}\right) \|x_1 - x^*\|^2$$

□

References

[1] S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.