IFT 6085 - Lecture 3
Gradients 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
Winter 2020: Thong (Bob) Vo
Winter 2019: Amir Raza, Philippe Lacaille, Jonathan Plante
Winter 2018: Philippe Brouillard, Massimo and Lucas Caccia

Instructor: Ioannis Mitliagkas

1 Summary

In the previous lecture we covered the notions of convexity as well as Lipschitsz continuity. After introducing these concepts, a bound on the convergence rate of gradient descent of a convex and L-Lipschitz function was demonstrated to scale with the $\sqrt{T}$.

Building on some of the previous lecture notions, we will introduce guarantees on the convergence rate of gradient descent for a stronger family of functions (using stronger assumptions), namely β-smooth and α-strong convex functions.

2 Gradient Descent for smooth functions

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

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

If we recall Lipschitz continuity from Lecture 2, simply speaking, an L-Lipschitz function is limited by how *quickly* its output can change. By imposing this same constraint on the gradients of a function, β-smoothness implies they cannot change abruptly and must be bounded by some value as defined above.

In other other words, β-smoothness is putting an upper bound on the curvature of the function. This is equivalent to the eigenvalues of the Hessian being less than β. Note that there can be β-smooth functions which are not twice differentiable. One key benefit of these functions is that their gradients tend to decay when x gets closer to the minimum. In contrast, non-smooth functions may have abrupt bends at the minimum, which cause significant oscillations for gradient descent. Figure 1 illustrates this point by comparing the two scenarios.

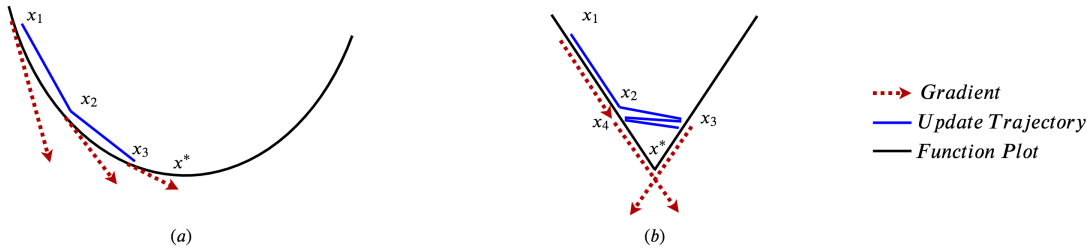

Figure 1: (a) A smooth function with decaying updates. (b) A non-smooth function with oscillating updates.

2.1 Convex and smooth functions

Here we introduce a bound on the convergence rate of a convex and β-smooth function.

Lemma 2 (Quadratic bounds)**.** *Let f be β-smooth on $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$, one has*

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

Proof. Let $g : \mathbb{R} \to \mathbb{R}$ by the rule $g(t) \triangleq f(y + t(x - y))$. The following holds true:

$$f(x) = g(1) \; and \; f(y) = g(0)$$

Then, we also observe that:

$$f(x) - f(y) = g(1) - g(0) = \int_0^1 g'(t)\, dt \tag{2}$$

Taking derivative of $g(t)$:

$$g'(t) = \nabla f(y + t(x - y))^T (x - y) \tag{3}$$

Plugging in Equation 2, and 3 to the LHS of 1:

$$\begin{aligned}
&|f(x) - f(y) - \nabla f(y)^\intercal (x - y)| \\
&= |\int_0^1 \nabla f(y + t(x - y))^\intercal (x - y)\, dt - \nabla f(y)^\intercal (x - y)| \\
&\leq \int_0^1 |\nabla f(y + t(x - y) - \nabla f(y))^\intercal (x - y)|\, dt \\
&\leq \int_0^1 \|\nabla f(y + t(x - y)) - \nabla f(y)\| \cdot \|x - y\|\, dt && \text{(applying Cauchy–Schwarz inequality)} \\
&\leq \int_0^1 \beta t |x - y\|^2\, dt && \text{(the gradient } \nabla f \text{ is } \beta \text{ -Lipschitz.)} \\
&= \frac{\beta}{2}\|x - y\|^2
\end{aligned}$$

□

Lemma 3. *Let f be such that $0 \leq f(x) - f(y) - \nabla f(y)^\intercal (x - y) \leq \frac{\beta}{2}\|x - y\|^2$. Then for any $x, y \in \mathbb{R}^n$, one has*

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

Proof. See [1] Lemma 3.5 for proof. □

Theorem 4. *Let f be convex and β-smooth on $\mathbb{R}^n$. Then the gradient descent with step size $\gamma = 1/\beta$ satisfies*

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

Proof. See [1] page 268. □

In comparison to the convergence analysis for a convex and *L*-Lipschitz function, the following points of improvement are observed in Theorem 4 over the previous result:

- No averaging of terms

- x_k is a good solution, i.e. no reference to previous iterations
- Convergence scales linearly with number of steps, convergence rate of order $O(1/T)$ compared to $O(1/\sqrt{T})$
- Ideal step size γ is constant and does not depend on T (number of steps taken).

As an observation of $\|x_1 - x^*\|^2$, the bound will be tighter if x_1 will be closer to x^* (minima), and looser if x_1 will be farther. Bounds that don't depend on the initial value of x will be discussed later in this lecture.

3 Strong convexity

Definition 5 (Strong convexity)**.** *A function $f(x)$ is α-strongly convex, if for $\alpha > 0$, $\forall x \in \mathrm{dom}(f)$,*

$$f(x) - \frac{\alpha}{2}\|x\|^2 \textit{is convex.}$$

Strong convexity provides a lower bound for the function's curvature. The function must have strictly positive curvature. In other words, all eigenvalues of the Hessian of a α-strongly convex function are lower bounded by α. We can write this in terms of positive-semi definiteness as

$$\nabla^2 f(x) \succeq \alpha I \iff \nabla^2 f(x) - \alpha I \succeq 0$$

For example, $f : \mathbb{R} \to \mathbb{R}, f(x) = \frac{h}{2}x^2$ is h-strongly convex, but not $(h+\epsilon)$-strongly convex for $\epsilon > 0$. Figure 2 illustrates examples of two convex functions, of which only one is strongly convex.

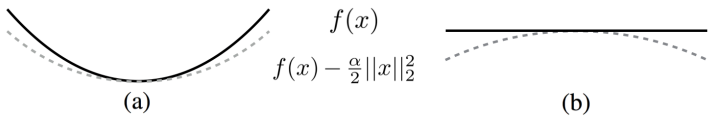

Figure 2: (a) A convex function which is also strongly convex. (b) A convex function which is not strongly convex.

3.1 Strongly convex and Lipschitz functions

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

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

Proof. See [1] page 277 □

With Theorem 6, we can notice how moving from convexity to strong convexity may affect the convergence rate of gradient descent. We previously tackled convexity paired with Lipschitz continuity in Lecture 2 and the similarity with the new convergence guarantees is pretty noticeable. By moving to strong-convexity, we can notice a few things:

- Still have averaging of terms, where $\sum_{k=1}^{T} \frac{2k}{T(T+1)} x_k$ is a non uniform averaging scheme with more recent iterates having more weight
- Current x_k solution is not appropriate, i.e. ideally only evaluate at x_k
- Convergence now scales linearly with number of steps
- Although not constant, the step size γ_k is diminishing at every step

Unlike previous results, the right hand side does not have any terms dependent on distance from x^* here. Intuitively this is a consequence of two conditions which have to be met for a Strong Convex function:

- Norm of Gradient increases as we go further away from minima.
- Gradient for a L-Lipschitz function is bounded, and can not keep increasing.

Thus to have a finite bound, no distance term on the right hand side of the bound.

One could hope that by combining strong convexity along with smoothness, gradient descent may present stronger convergence guarantees. We will see how smoothness removes dependency from the averaging scheme.

Note: α Strongly convex and L-Lipschitz condition is a special case because the upper bound L-Lipschitz condition will ultimately conflict with the lower bound α Strongly convex grow rate. Therefore, such functions are typically defined in a range, e.g. $x \in [-1, 1]$.

3.2 Strongly convex and smooth functions

Recalling Lemma 2 (Quadratic bounds), it tells us that a β-smooth function f is sandwiched between 2 quadratics because of the following inequality:

$$f(y) + \nabla f(y)^\intercal (x - y) - \frac{\beta}{2}\|x - y\|^2 \leq f(x) \leq f(y) + \nabla f(y)^\intercal (x - y) + \frac{\beta}{2}\|x - y\|^2 \tag{4}$$

Now we introduce another lemma which allows us to lower bound an α-strong convex function.

Lemma 7. *Let f be λ-strongly convex. Then $\forall x, y$, we have:*

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

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

By rearranging terms, this tells us that a λ-strong convex function can be lower bounded by the following inequality:

$$f(x) \geq f(y) - \nabla f(y)^T (y - x) + \frac{\lambda}{2}\|x - y\|^2 \tag{5}$$

Figure 3 showcases the resulting bounds from both the smoothness and the strong convexity constraints. The shaded area in each sub-figure is showing the area validated by the respective bound(s).

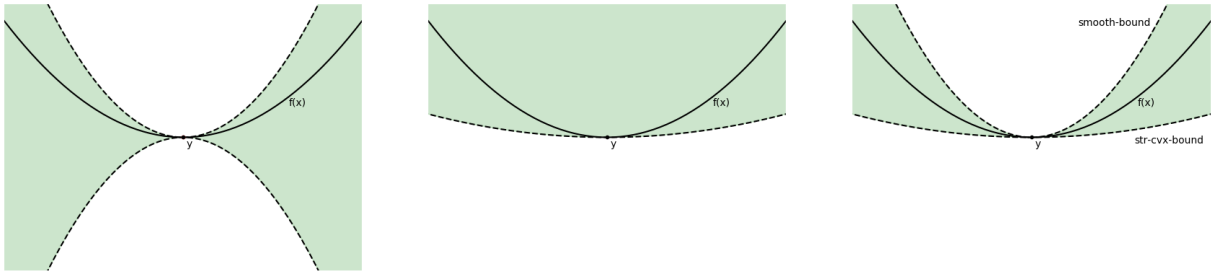

Figure 3: (Left) Upper and lower bounds from equation 4 for smoothness constraint (Middle) Lower bound from equation 5 for strong convexity constraint (Right) Combination of upper bound from smoothness and lower bound from strong convexity

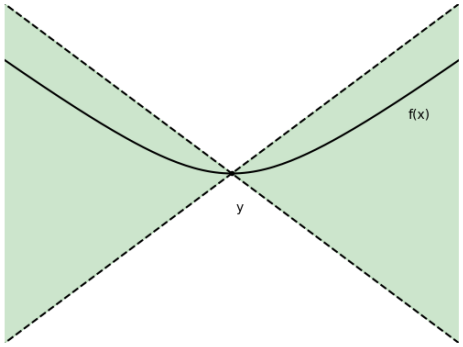

Figure 4: For an L_f-Lipschitz continuous function, the green region shows where the function would exist . We can imagine that without smoothness and only L-Lipschitz in equation 4, the accepted region would be having linear boundaries

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

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

Proof. See [1] Lemma 3.11 for proof. □

Theorem 9. *For f a λ-strongly convex and β-smooth function, 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 $\kappa \triangleq \frac{\beta}{\lambda}$ is the condition number.

Proof. (Theorem 9) 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:

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

By expanding the square, we get:

$$D_{k+1}^2 = \|x_k - x^*\|^2 - 2\gamma\langle \nabla f(x_k), x_k - x^*\rangle + \gamma^2\|\nabla f(x_k)\|^2$$

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:

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

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:

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

The term $\left(1 - \frac{2\gamma\lambda\beta}{\lambda+\beta}\right)$ is useful, because we can show that it is less than 1 and has geometric convergence. Further steps will try to simplify the remaining terms.
We can change the norm term by using the fact that f is β-smooth:

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

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:

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

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

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

By β-smoothness we finally have:

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

□

Theorem 9 observations

- x_k solution is a good solution, i.e. no reference to previous iterations
- Convergence rate of order $O(\exp(-T))$
- κ measures how *far apart* the upper and lower bounds are (see Figure 3). It can be interpreted as the ratio of largest to smallest curvature of the function.
- The smaller the condition number κ is, the less iterations are required to converge. Intuitively, the accepted region between the bounds will be smaller.
- Consequently, the greater β is the more iterations will be required to converge. This is logical since a constant step size on a function with a steep gradient will cause the a greater change in the function value.

4 Comparison of optimization properties for different function classes

The table below summarizes various convergence properties of discussed functions classes. From left to right, the assumptions on properties of these classes increase, from Convex and L-Lipschitz to λ strongly convex and β smooth. In the same direction, the convergence rates are also increase, aligned with stronger assumptions on those functions.

Table 1: Comparison of different function classes

Function Class	cvx, L-Lipschitz	cvx, β smooth	α str-cvx, L-Lipschitz	λ str-cvx, β smooth
Optimal Step size	$\gamma = \frac{\|x_1 - x^*\|_2}{L\sqrt{T}}$	$\gamma = \frac{1}{\beta}$	$\gamma = \frac{2}{\alpha(k+1)}$	$\gamma = \frac{2}{\lambda+\beta}$
Convergence Rate	$\mathcal{O}(1/\sqrt{T})$	$\mathcal{O}(1/T)$	$\mathcal{O}(1/T)$	$\mathcal{O}(exp(-T))$
Sub-optimal gap	$f\left(\frac{1}{T}\sum_{k=1}^{T} x_k\right) - f(x^*)$	$(x_k) - f(x^*)$	$f\left(\sum_{k=1}^{T} \frac{2k}{T(T+1)} x_k\right) - f(x^*)$	$f(x_{k+1}) - f(x^*)$
Bounds of the gap	$\leq \frac{\|x_1 - x^*\| L}{\sqrt{T}}$	$\leq \frac{2\beta\|x_1 - x^*\|^2}{k-1}$	$\leq \frac{2L^2}{\alpha(T+1)}$	$\leq \frac{\beta}{2} \exp\left(-\frac{4k}{\kappa+1}\right) D_1^2$

References

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