CPSC 536C: Algorithms for Convex Optimization
University of British Columbia

Gradient Descent

Akshay Ram
February 2, 2025

These notes are subject to change and may contain errors.

1 Introduction

The goal of this note is to study and analyze the following simple procedure for unconstrained convex optimization:

$$x_{t+1} := x_t - \eta_t \nabla f(x_t).$$

Here and in the remainder we assume f is differentiable (so $\nabla f(x)$ is well-defined). We note that much of the analysis carries over to the non-differentiable convex functions by using any sub-gradient.

In contrast to cutting plane methods, these gradient methods usually have $\text{poly}(1/\varepsilon)$ convergence to an ε-approximate solution, but have much better dependence on the problem dimension n. At the end we will also show that such dependencies are necessary by constructing and analyzing certain hard instances for gradient descent.

1.1 Approximate Solutions

We consider more formally what it means to produce an approximate solution:

Definition 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function with optimizer $x^* := \arg\min_x f(x)$.*

- *x has ε-function gap if $f(x) \le f(x^*) + \varepsilon$;*
- *x is ε-close if $\|x - x^*\|_2 \le \varepsilon$;*
- *x is ε-critical if $\|\nabla f(x)\|_2 \le \varepsilon$.*

These three notions of approximate solutions are in general incomparable, and their importance depends on the application at hand.

Exercise 1. *Compute the function gap, distance to opt, and gradient for the following:*

1. *$f(x) := \|x\|_2^2/2$;*
2. *$f(x) := \|Ax - b\|_2^2$;*
3. *$f(Q) := -\log\det(Q) + \sum_i \lambda_i \langle x_i, Qx_i\rangle$.*

1.2 Assumptions

We need some assumptions on our initial iterate and the function in order to guarantee any reasonable approximation. The following are natural assumptions satisfied by many functions in practice.

Definition 2. *$f : \mathbb{R}^n \to \mathbb{R}$ is G-Lipschitz if*

$$\forall x, y : \quad |f(y) - f(x)| \leq G\|y - x\|_2.$$

This intuitively allows us to relate f to an affine function with slope G. We can also relate f to quadratic functions, which are the next simplest convex functions. Recall the first-order definition of convexity:

$$\forall x, y : \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Definition 3. *$f : \mathbb{R}^n \to \mathbb{R}$ is α-strongly convex if*

$$\forall x, y : \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|_2^2.$$

f is β-smooth if

$$\forall x, y : \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|_2^2.$$

Note that the first gives a stronger lower bound than standard convexity, whereas the smoothness condition gives an upper bound. In both cases this allows us to compare our function to a convex quadratic, which is much simpler to analyze.

Exercise 2.

1. *If f is α-strongly convex, then it is also α'-strongly convex for any $\alpha' \leq \alpha$. I.e. strong convexity is an increasingly strong assumption for α increasing;*
2. *If f is β-smooth, then it is also β'-strongly convex for any $\beta' \geq \beta$. I.e. smoothness is a weaker assumption for β increasing;*
3. *If f, g are α_f, α_g-strongly convex and β_f, β_g-smooth, respectively, then $f+g$ is $\alpha_f+\alpha_g$-strongly convex and $\beta_f + \beta_g$-smooth;*

Claim 4. *For quadratic $q_\gamma(x) := q_0 + \langle b, x - x_0 \rangle + \frac{\gamma}{2}\|x - x_0\|_2^2$, we can optimize*

$$\min_x q_\gamma(x) = q_0 - \frac{\|b\|_2^2}{2\gamma} \quad \textit{with} \quad x^* := \arg\min_x q_\gamma(x) = x_0 - \frac{b}{\gamma}.$$

We leave the proof as an exercise (show directly by computing gradient and setting to 0).

The point of the above assumptions are that they allow us to relate various notions of approximate solutions.

Fact 5. *Let x^* be the optimizer of f, which is (1) G-Lipschitz, (2) β-smooth, and (3) α-strongly convex, respectively. Then*

1. $f(x) - f(x^*) \leq G\|x - x^*\|_2$;
2. $f(x) - f(x^*) \leq \frac{\beta}{2}\|x - x^*\|_2^2$;
3. $f(x) - f(x^*) \leq \frac{\|\nabla f(x)\|_2^2}{2\alpha}$.

Proof:

1. Follows directly from the definition;

2. Apply the definition of smoothness between x and x^*:

$$f(x) \leq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{\beta}{2}\|x - x^*\|_2^2 = f(x^*) + \frac{\beta}{2}\|x - x^*\|_2^2,$$

   where in the last step we used $\nabla f(x^*) = 0$ since x^* is the optimizer.

3. Rearranging, we see that this is equivalent to a lower bound on $f(x^*)$ in terms of the gradient norm. For this we use our strong convexity assumption

$$f(x^*) \geq q_\alpha(x^*) := f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\alpha}{2}\|x^* - x\|_2^2.$$

   Then by theorem 4 we have

$$q_\alpha(x^*) \geq \min_y q_\alpha(y) = f(x) - \frac{\|\nabla f(x)\|_2^2}{2\alpha}.$$

   The result follows by the lower bound $f(x^*) \geq q_\alpha(x^*)$ and rearranging.

□

2 Analysis of Gradient Descent

Intuitively, if we have a simple quadratic function, theorem 4 shows how to find the optimum directly. In general, convexity (and strong convexity and smoothness) allows us to approximate our function f by an affine or quadratic function. The algorithm and analysis of gradient involves using these proxies to compute updates and reason about the progress made.

Theorem 6. *Let f be α-strongly convex and β-smooth, and consider update*

$$x_{t+1} := x_t - \frac{1}{\beta}\nabla f(x_t).$$

Then the function gap of iteration T can be bounded as

$$f(x_T) - f(x^*) \leq \left(1 - \frac{\alpha}{\beta}\right)^T (f(x_0) - f(x^*)),$$

*i.e. we achieve ε-*relative *function gap in $O(\kappa \log(1/\varepsilon))$ iterations, where $\kappa := \beta/\alpha$.*

Proof: The update step can be derived as the optimizer of the proxy given by smoothness:

$$x_{t+1} = \arg\min_x q_\beta(x) := \arg\min_x f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\beta}{2}\|x - x_t\|_2^2,$$

where the minimum calculation follows by theorem 4. By smoothness we have $f \leq q_\beta$, so we can show significant progress in function value:

$$f(x_{t+1}) \leq q_\beta(x_{t+1}) = f(x_t) - \frac{\|\nabla f(x_t)\|_2^2}{2\beta},$$

where again the last step was by theorem 4. Next, we use strong convexity to show that this gives a significant improvement:

$$\frac{\|\nabla f(x_t)\|_2^2}{2\alpha} \geq f(x_t) - f(x^*),$$

where we applied theorem 5(3) for α-strongly convex f. Putting this together gives

$$\begin{aligned} f(x_{t+1}) - f(x^*) &= (f(x_{t+1}) - f(x_t)) + (f(x_t) - f(x^*)) \\ &\leq -\frac{\|\nabla f(x_t)\|_2^2}{2\beta} + (f(x_t) - f(x^*)) \leq (1 - \alpha/\beta)(f(x_t) - f(x^*)), \end{aligned}$$

where in the second step we used the upper bound for the update, and in the final step we used that theorem 5(3) to compare the gradient norm and $f(x_t) - f(x^*)$. □

Note that the above gives exponential convergence in terms of function gap. Also by theorem 5 we can show convergence in terms of the solution $\|x_t - x^*\|_2$ and gradient $\|\nabla f(x_t)\|_2$. The above setting is in some sense the most structured, since we have upper and lower bounds on f via smoothness and strong convexity. The next case is the least structured, and will have much worse guarantees.

Theorem 7. *Let f be G-Lipschitz and assume $\|x_0 - x^*\|_2 \leq R$. Then for update*

$$x_{t+1} := x_t - \eta \nabla f(x_t) \quad \textit{where} \quad \eta := \sqrt{\frac{R^2}{G^2 T}},$$

the function gap can be bounded as

$$\min_{t \in [T]} f(x_t) - f(x^*) \lesssim \sqrt{\frac{R^2 G^2}{T}},$$

*i.e. we achieve ε-*relative *function gap in $O(\kappa \log(1/\varepsilon))$ iterations, where $\kappa := \beta/\alpha$.*

Proof: By the same theorem 4 we argue that

$$x_{t+1} = \arg\min_x q_\eta(x) := \arg\min_x f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta}\|x - x_t\|_2^2.$$

But in this setting this does not necessarily give us a direct guarantee for our function. We instead use convexity to argue that each iteration guarantees either function improvement or distance improvement:

$$\begin{aligned} f(x^*) + \frac{1}{2\eta}\|x^* - x_t\|_2^2 &\geq f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle + \frac{1}{2\eta}\|x^* - x_t\|_2^2 = q_\eta(x^*) \\ &\geq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{1}{2\eta}\|x_{t+1} - x_t\|_2^2 + \frac{1}{2\eta}\|x^* - x_{t+1}\|_2^2 \\ &= q_\eta(x_{t+1}) + \frac{1}{2\eta}\|x^* - x_t\|_2^2 \\ &\geq f(x_t) - \eta\frac{\|\nabla f(x_t)\|_2^2}{2} + \frac{1}{2\eta}\|x^* - x_{t+1}\|_2^2, \end{aligned}$$

where the first step was by convexity of f, the third step follows by $1/\eta$-strong convexity of q_η (which can be proven directly or using that q_η is the sum of convex f and $1/\eta$-strongly convex

$\frac{1}{2\eta}\|x^* - x_t\|_2^2$), and in the final step we used theorem 4 for the proxy q_η. Rearranging, we have a bound on the function gap in each iteration:

$$f(x_t) - f(x^*) \leq \eta \frac{\|\nabla f(x_t)\|_2^2}{2} + \frac{1}{2\eta}(\|x^* - x_t\|_2^2 - \|x^* - x_{t+1}\|_2^2).$$

We can bound the first term using the Lipschitz condition. For the second term, note that this is large iff x_{t+1} gets much closer to the optimum x^* than x_t; since we have an initial bound $\|x_0 - x^*\|_2 \leq R$, this term is bounded across all iterations. And finally, by the Lipschitz condition, if $\|x_t - x^*\|_2$ is small, then we can argue that the function gap is small by theorem 5(1).

We will first argue about the average function gap:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*) &\leq \frac{1}{T}\sum_{t=0}^{T-1} \frac{\eta\|\nabla f(x_t)\|_2^2}{2} + \frac{1}{T}\sum_{t=0}^{T-1} \frac{\|x^* - x_t\|_2^2 - \|x^* - x_{t+1}\|_2^2}{2\eta} \\ &\leq \frac{\eta G^2}{2} + \frac{\|x^* - x_0\|_2^2 - \|x^* - x_T\|_2^2}{2\eta T} \leq \frac{\eta G^2}{2} + \frac{R^2}{2\eta T},\end{aligned}$$

where the first step was by the calculation above for each iterate, in the second step we bounded each gradient $\|\nabla f(x_t)\|_2 \leq G$ by the Lipschitz condition, and we bounded the sum by noting it is a telescoping series, and in the final step we used that $\|x^* - x_T\|_2 \geq 0$ and our initial bound $\|x_0 - x^*\|_2 \leq R$. Finally we can choose $\eta = \sqrt{\frac{R^2}{G^2 T}}$ to balance terms. Therefore we can bound the minimum function gap over all iterations

$$\min_t f(x_t) - f(x^*) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*) \leq \sqrt{\frac{R^2 G^2}{T}}.$$

We can also use convexity to bound the function gap for the average iterate $\overline{x} := \frac{1}{T}\sum_{t=0}^{T-1} x_t$:

$$f(\overline{x}) - f(x^*) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*) \leq \sqrt{\frac{R^2 G^2}{T}},$$

where the first step is by Jensen's inequality. □

In the next section we show that the guarantee for the Lipschitz setting is optimal! I.e. there exists worst-case instances such that *any* algorithm with access to function evaluation and gradients must have iteration dependence of the form $1/\varepsilon^2$. On the other hand, the guarantee for the first case, with smoothness and strong convexity, is surprisingly not optimal.

3 Lower Bounds

Theorem 8 (Theorem 3.13 in Bubeck)**.** *For any $T \leq n$ and any $G > 0$, there exists a convex function $f : \mathbb{R}^n \to \mathbb{R}$ that is G-Lipschitz such that any black-box algorithm with access to function evaluation and gradients must satisfy*

$$\min_{t\in[T]} f(x_t) \geq \min_{x \in B_2^n} f(x) + \Omega\left(\frac{RG}{\sqrt{T}}\right).$$

Theorem 9 (Theorem 3.15 in Bubeck)**.** *For any $\kappa = \beta/\alpha > 1$, there exists a convex function f that is α-strongly convex and β-smooth, such that any black-box algorithm with access to function evaluation and gradients must satisfy*

$$f(x_t) - f(x^*) \geq \frac{\alpha}{2}\left(1 - \frac{1}{\Omega(\sqrt{\kappa})}\right)^t \|x_0 - x^*\|_2^2.$$