EE C227C Convex Optimization and Approximation Lecture 12 - 02/24/2022

Lecture 12: Acceleration of Gradient Methods

Lecturer: Jiantao Jiao *Scribe: Yigit Efe Erginbas, Simon Xu*

In this lecture, we present Nesterov's accelerated gradient descent algorithm. This algorithm can be viewed as a hybrid of the previously introduced gradient descent and mirror descent methods. For further reference, see [1] (Chapter 8).

1 The Problem Setup

First, let us recall the conditions on acceleration. Given the optimization problem,

$$\min_{x\in\mathbb{R}^n} f(x) \tag{1}$$

where $f(x)$ is convex and its gradients are L-Lipschitz continuous. Here, we work with a general pair of dual norms $\|\cdot\|$ and $\|\cdot\|_*$. Thus, we use the notion of L-Lipschitz continuous gradient with the following definiton.

Definition 1. *A function $f:\mathbb{R}^n\to\mathbb{R}$ is said to have L-Lipschitz continuous gradients with respect to a norm $\|\cdot\|$ if for all $x,y\in\mathbb{R}^n$,*

$$\|\nabla f(x)-\nabla f(y)\|_*\le L\|x-y\| \tag{2}$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$.

For convex functions, this condition is the same as L-smoothness:

$$f(y)\le f(x)+\langle y-x,\nabla f(x)\rangle+\frac{L}{2}\|x-y\|^2,\qquad \forall x,y\in\mathbb{R}^n \tag{3}$$

Furthermore, we let $R:\mathbb{R}^n\longrightarrow\mathbb{R}$ to be σ-strongly convex regularizer with respect to a norm $\|.\|$. In other words, the Bregman divergence $D_R(x,y)$ of R satisfies

$$D_R(x,y)=R(x)-R(y)-\langle\nabla R(y),x-y\rangle\ge\frac{\sigma}{2}\|x-y\|^2 \tag{4}$$

2 Main result on accelerated gradient descent

Theorem 2. ***(Existence of an accelerated algorithm)*** *Given the conditions*

1. *a first-order oracle access to a convex function $f(x)$*
2. *a number L such that $\nabla f(x)$ is L-Lipschitz continuous with respect to a norm $\|\cdot\|$*
3. *a σ-strongly convex regularizer R*
4. *an oracle access to the gradient map (∇R) and its inverse $(\nabla R)^{-1}$ (given that they are bijections)*
5. *an initial point x_0 such that $D_R(x^*,x_0)\le D^2$*
6. *an $\epsilon>0$,*

there exists an algorithm that produces a point $x\in\mathbb{R}^n$ such that $f(x)\le f(x^*)+\epsilon$. In addition, the algorithm makes $T=O(\sqrt{\frac{LD^2}{\sigma\epsilon}})$ queries to the oracle and performs $O(nT)$ arithmetic operations.

2.1 Intuition

From Theorem 2, we start with making several observations:

- The problem dimension n does not appear in the convergence speed (dimension-free)
- The standard gradient descent algorithm requires $T = O\left(\frac{LD^2}{\sigma\epsilon}\right)$ iterations – exactly the square of what the above theorem achieves.

Now, we will further explore an intuitive way to reason about such an algorithm. In previous lectures we have observed that the gradient descent behaves better when we have large gradients, while mirror descent is better when we operate with small gradients. Therefore, in order to have an accelerated descent algorithm, we would prefer an algorithm that behaves like gradient descent for large gradients, and behaves like mirror descent for small gradients.

To further emphasise our point, let us consider the following example. For the sake of illustration, define a value ϵ such that 2ϵ is equal to the $f(x_0) - f(x^*)$. Our goal is to achieve ϵ accuracy $f(x_t) - f(x^*) \leq \epsilon$. Now, consider these two cases separately: either $\|\nabla f(x_t)\|_* \geq K$ or $\|\nabla f(x_t)\|_* \leq K$ for all t.

In the first case ($\|\nabla f(x)\|_* \geq K$), we can imagine applying a gradient descent algorithm, such that

$$f(x_{t+1}) - f(x_t) \leq \frac{1}{2L}\|\nabla f(x_t)\|_*^2 \tag{5}$$

for the choice of step size $\eta = 1/L$. Recall that in this case the number of required steps by gradient descent algorithm is

$$T_1 := \frac{L\epsilon}{K^2} \tag{6}$$

Similarly, in the second case ($\|\nabla f(x)\|_* \leq K$), an application of the mirror descent algorithm (for a 1-strongly convex regularizer) requires

$$T_2 := \frac{K^2 D^2}{\epsilon^2} \tag{7}$$

for a K-Lipschitz convex function f.

As we can see from previous exposition, T_1 is decreasing in K while T_2 is increasing in K. If we were to determine a threshold between two regimes, we would solve $T_1 = T_2$ for K. Using the definitions from Eq. (6) and (7), we obtain the threshold

$$K_{\text{th}} = \left(\frac{L\epsilon^3}{D^2}\right)^{1/4} \tag{8}$$

Therefore, if $K \geq K_{\text{th}}$, we would prefer to have $\|\nabla f(x)\|_* \geq K$ and use gradient descent. On the other hand, if $K \leq K_{\text{th}}$, we would prefer to have $\|\nabla f(x)\|_* \leq K$ and use mirror descent. Hence, given an ability to choose the regime in which we will operate, the number of required steps would be at most

$$T_1(K_{\text{th}}) = T_2(K_{\text{th}}) = \sqrt{\frac{LD^2}{\epsilon}} \tag{9}$$

As we can see, this iteration complexity exactly matches with the iteration complexity claimed for accelerated gradient descent (by setting $\sigma = 1$). While this thought experiment may help motivate why using a mixture of gradient descent and mirror descent can help accelerate convergence, the reality is that this thought experiment does not work in practice: one iteration of gradient descent can break down another iteration of mirror descent, and vice versa, causing any progress to be lost. Hence, we require a more rigorous analysis and understanding of accelerated gradient descent.

2.2 Linear Coupling

Linear coupling is the solution to this problem in a high level. It is a method that will allow us to accelerate gradient descent by combining it with mirror descent. With the idea of linear coupling, we will be able to construct an optimal balance between gradient descent and mirror descent algorithms.

Starting from an initial point x_0, a duplicate of x_0 is made, along with a point z_0. Next, y_0 is produced as a linear combination x_0 and z_0. On the next iteration, x_1 is produced from an iteration of gradient descent, while z_1 is produced from an iteration of mirror descent. y_1 is then produced from x_1 and z_1, and the process continues with x_2, z_2, y_2, and so on. Below is a figure that illustrates this process.

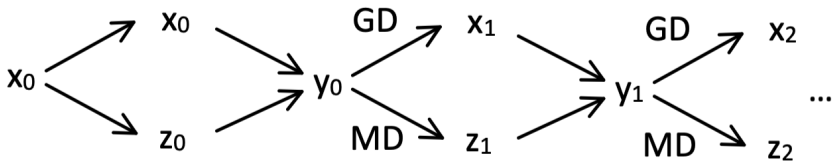

Figure 1: A diagram illustrating linear coupling.

The linear coupling equation between the iterations are given by:

$$y_{t-1} = (1-\gamma_t)x_{t-1} + \gamma_t z_{t-1} \tag{10}$$

$$x_t \leftarrow \text{apply a GD step to } y_{t-1} \tag{11}$$

$$z_t \leftarrow \text{apply a MD step to } y_{t-1} \tag{12}$$

3 Proof strategy: estimate sequences

In our proof of Theorem 2, we first start by introducing the concept of estimate sequence and show how we can use an estimate sequence to prove our result on accelerated gradient descent. Then, in the process of proving the existence of estimate sequences, we derive an accelerated gradient descent algorithm, which then turns out to imply Theorem 2.

A crucial notion used in deriving the accelerated gradient descent algorithm is that of an estimate sequence.

Definition 3 (**Estimate Sequence**). *A sequence $(\phi_t, \lambda_t, x_t)_{t\geq 0}$ with function $\phi_t : \mathbb{R}^n \to \mathbb{R}$, value $\lambda_t \in [0,1]$, and vector $x_t \in \mathbb{R}^n$ (for all $t \geq 0$) is said to be an estimate sequence for a function $f : \mathbb{R}^n \to \mathbb{R}$ if it satisfies the following properties:*

1. ***Lower bound:*** *For all $t \geq 0$ and for all $x \in \mathbb{R}^n$,*

$$\phi_t(x) \leq (1-\lambda_t)f(x) + \lambda_t \phi_0(x)$$

2. ***Upper bound:*** *For all $t \geq 0$ and for all $x \in \mathbb{R}^n$,*

$$f(x_t) \leq \phi_t(x)$$

Intuitively, we can think of the sequence $(x_t)_{t\geq 0}$ as converging to a minimizer of f. The functions $(\phi_t)_{t\geq 0}$ serve as approximations to f, which provide tighter and tighter (as t increases) bounds on the gap $f(x_t) - f(x^*)$. More precisely, condition (1) says that $\phi_t(x)$ is an approximate lower bound to $f(x)$ and condition (2) says that the minimum value of ϕ_t is above $f(x_t)$.

To illustrate this definition, suppose that $\lambda_t = 0$ for some t. Then, the condition (1) implies that $\phi_t(x) \leq f(x)$ for all $x \in \mathbb{R}^n$. Choosing $x = x^*$ and combining with the condition (2) we obtain

$$f(x_t) \leq \phi_t(x^*) \leq f(x^*) \tag{13}$$

Since x^* is the minimizer of f by definition, this inequality implies that x_t is an optimal solution. Thus, even though achieving $\lambda_t = 0$ may be too ambitious, we aim for obtaining an estimate sequence that will guarantee $\lambda_t \to 0$ as t increases. In fact, as we will show later, the accelerated gradient method constructs a sequence λ_t which goes to zero as $1/t^2$. Formally, we state the following theorem and defer its proof to our next lecture.

Theorem 4 (**Existence of optimal estimate sequences**). *For every convex, L-smooth (with respect to norm $\|\cdot\|$) function $f : \mathbb{R}^n \to \mathbb{R}$, for every σ-strongly convex regularizer R (with respect to the same norm $\|\cdot\|$), and for every $x_0 \in \mathbb{R}^n$, there exists an estimate sequence $(\phi_t, \lambda_t, x_t)_{t\geq 0}$ with*

$$\phi_0(x) := f(x_0) + \frac{L}{\sigma} D_R(x, x_0)$$

and

$$\lambda_t \leq \frac{c}{t^2}$$

for some absolute constant $c > 0$.

We leave the proof of Theorem 4 for our next lecture, and we continue with a lemma that will enable us to conclude our main result Theorem 2.

Lemma 5. *Under the problem setting described in Section 1, an estimate sequence $(\phi_t, \lambda_t, x_t)_{t\geq 0}$ with*

$$\phi_0(x) := f(x_0) + \frac{L}{\sigma} D_R(x, x_0)$$

and

$$\lambda_t \leq \frac{c}{t^2}$$

for some absolute constant $c > 0$, satisfies

$$f(x_t) \leq f(x^*) + \frac{2cLD^2}{\sigma t^2}$$

Proof For an estimate sequence $(\phi_t, \lambda_t, x_t)_{t\geq 0}$ that satisfies the given properties, we obtain

$$
\begin{aligned}
f(x_t) &\overset{(a)}{\leq} \phi_t(x^*) && (14)\\
&\overset{(b)}{\leq} (1-\lambda_t)f(x^*) + \lambda_t\phi_0(x^*) && (15)\\
&\overset{(c)}{=} (1-\lambda_t)f(x^*) + \lambda_t f(x_0) + \lambda_t\frac{L}{\sigma}D_R(x^*,x_0) && (16)\\
&= f(x^*) + \lambda_t(f(x_0)-f(x^*)) + \lambda_t\frac{L}{\sigma}D_R(x^*,x_0) && (17)\\
&\overset{(d)}{\leq} f(x^*) + \lambda_t\left(\langle x_o - x^*, \nabla f(x^*)\rangle + \frac{L}{2}\|x_o-x^*\|^2\right) + \lambda_t\frac{L}{\sigma}D_R(x^*,x_0) && (18)\\
&= f(x^*) + \lambda_t L\left(\frac{1}{2}\|x_o-x^*\|^2 + \frac{1}{\sigma}D_R(x^*,x_0)\right) && (19)\\
&\overset{(e)}{\leq} f(x^*) + \lambda_t L\left(\frac{1}{\sigma}D_R(x^*,x_0) + \frac{1}{\sigma}D_R(x^*,x_0)\right) && (20)\\
&= f(x^*) + \lambda_t\frac{2L}{\sigma}D_R(x^*,x_0) && (21)\\
&\overset{(f)}{\leq} f(x^*) + \lambda_t\frac{2LD^2}{\sigma} && (22)\\
&\overset{(g)}{\leq} f(x^*) + \frac{2cLD^2}{\sigma t^2} && (23)
\end{aligned}
$$

Here, (a) uses the upper bound property of estimate sequences, (b) uses the lower bound property of estimate sequences, (c) uses the definition of $\phi_0(x^*)$, (d) uses L-smoothness of f, (e) uses σ-strong convexity of R, (f) uses the condition $D_R(x^*,x_0)\leq D^2$, and (g) uses the property $\lambda_t \leq c/t^2$. □

Thus, given an estimate sequence that satisfies the conditions in Lemma 5, it is enough to take $t = O\left(\sqrt{\frac{LD^2}{\sigma\epsilon}}\right)$ to make sure that $f(x_t) - f(x^*) \leq \epsilon$. Next, we need to prove that such an estimate sequence $(\phi_t, \lambda_t, x_t)_{t\geq 0}$ for f exists and can be efficiently computed using a first order oracle to f and R only. As we will see, the proof of Theorem 4 will also provide us an efficient algorithm to compute estimate sequences.

4 Construction of an estimate sequence

To start, we make a simplifying assumption that $L = 1$ and $\sigma = 1$ without loss of generality. The construction of the estimate sequence is iterative. Let $x_0 \in \mathbb{R}^n$ be an arbitrary initial point. We set

$$\phi_0(x) := f(x_0) + D_R(x,x_0) \quad \text{and} \quad \lambda_0 = 1 \tag{24}$$

Thus, the lower bound condition in Definition 3 is trivially satisfied. The upper bound condition follows from noting that

$$\phi_0^* = \min_x \phi_0(x) = f(x_0) \tag{25}$$

Thus,

$$\phi_0(x) = \phi_0^* + D_R(x,x_0)$$

The construction of subsequent elements of the estimate sequence is inductive. Suppose we are given $(\phi_{t-1}, \lambda_{t-1}, x_{t-1})$. Then ϕ_t will be a convex combination of ϕ_{t-1} and a linear lower bound L_{t-1} to f at a carefully chosen point $y_{t-1} \in \mathbb{R}^n$. More precisely, we set

$$L_{t-1}(x) := f(y_{t-1}) + \langle x - y_{t-1}, \nabla f(y_{t-1})\rangle \tag{26}$$

and we set the new estimate to be

$$\phi_t(x) := (1 - \gamma_t)\phi_{t-1}(x) + \gamma_t L_{t-1}(x) \tag{27}$$

for some $\gamma_t \in [0, 1]$ determined later.

References

[1] N. K. Vishnoi, "Algorithms for convex optimization." [Online]. Available: https://convex-optimization.github.io/ACO-v1.pdf