15-859E: Advanced Algorithms **CMU, Spring 2015**
Lecture #16: Gradient Descent February 18, 2015
Lecturer: Anupam Gupta Scribe: Guru Guruganesh

In this lecture, we will study the gradient descent algorithm and analyze it in the context of convex optimization.

1 Preliminaries

First, recall the following definitions:

Definition 16.1. A convex set $K \subseteq \mathbb{R}^n$ is said to be convex iff

$$\big(\lambda x + (1-\lambda)y\big) \in K \qquad \forall x, y \in K,\ \forall \lambda \in [0,1]$$

Definition 16.2. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be convex iff

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \qquad \forall x, y \in \mathbb{R}^n,\ \forall \lambda \in [0,1]$$

In the context of this lecture, we will always assume that the function f is differentiable.

Fact 16.3. *A function f is convex iff $\forall x, y\ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$*

Geometrically Fact 16.3 states that the function always lies above its' tangent (see Fig 16.1). If

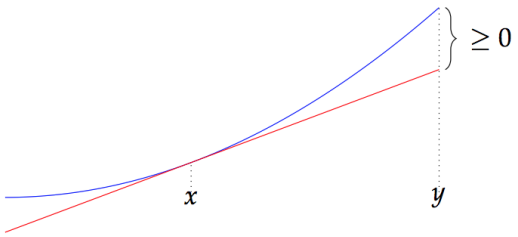

Figure 16.1: The blue line denotes the function and the red line is the tangent at x. [VT]

the function f is also twice differentiable, then we denote its' second derivative (a.k.a. Jacobian) by $H(f)_{i,j} := \frac{\partial f}{\partial x_i \partial x_j}$.

Fact 16.4. *A twice differentiable function f is convex iff $H(f) \succeq 0$* [1].

2 Convex Minimization and Gradient Descent

There are two kinds of problems that we will concern our-self with:

1. Unconstrained Convex Minimization (UCM): Given a convex function f

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

2. Constrained Convex Minimization (CCM): Given a convex function f and convex set K,

$$\min_{x \in K} f(x)$$

[1] The notation $A \succeq B$ signifies $A - B$ is positive semidefinite.

Algorithm 1: Gradient Descent

for $t \leftarrow 0$ **to** $T-1$ **do**
  $x_{t+1} \leftarrow x_t - \eta_t \cdot \nabla f(x_t)$
end
Result: $\hat{x} = \frac{1}{T}\sum_{t=0}^{T-1} x_i$

2.1 Unconstrained Convex Minimization

One useful property of convex functions is that f is convex implies that all local minima are also global minima. Hence, solving $\nabla f(x) = 0$ would enable us to compute the global minima exactly. Quite often, it is not possible to solve $\nabla f = 0$. However, we can hope to iteratively approximate the optimal solution x^*.

We do so by taking steps in the direction opposite to gradient, to get closer to a local minimum. The algorithm commonly known as Gradient Descent is described in Algorithm 1 and satisfies the following guarantee.

Theorem 16.5. *For any convex function $f : \mathbb{R}^n \to \mathbb{R}$ with $\forall x \in \mathbb{R}^n \ \|\nabla f(x)\|_2 \leq G$ and suppose $\|x^* - x_0\| \leq D$ then $f(\hat{x}) - f(x^*) \leq \epsilon$ if we set $T = \left(\frac{GD}{\epsilon}\right)^2$ and $\eta_t = \frac{D}{G\sqrt{T}}$.*

We will use the following elementary fact in the proof

$$\langle a, b\rangle = \frac{1}{2}[\|a\|^2 + \|b\|^2 - \|a-b\|^2] \tag{F1}$$

Proof.

$$\begin{aligned}
f(x_t) - f(x^*) &\leq \langle \nabla f(x_t), x_t - x^*\rangle && \text{By convexity of } f \\
&= \left\langle \frac{1}{\eta}x_t - x_{t+1}, x_t - x^* \right\rangle && \text{By update rule} \\
&= \frac{1}{2\eta}\Big(\|x_t - x_{t+1}\|^2 + \|x_t - x_t^*\|^2 - \|x_{t+1} - x^*\|^2 \Big) && \text{Using Fact F1} \\
&= \frac{1}{2\eta}\Big(\|\eta \cdot \nabla f(x_t)\|^2 + \Delta_t^2 - \Delta_{t+1}^2 \Big) && (16.1)
\end{aligned}$$

Analyzing the objective,

$$
\begin{aligned}
f(\hat{x}) - f(x^*) &= f\left(\frac{\sum_{t=0}^{T-1} x_t}{T}\right) - f(x^*) \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}[f(x_t) - f(x^*)] && \text{By convexity of } f \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{2\eta}\Big(\|\eta\cdot\nabla f(x_t)\|^2 + \Delta_t^2 - \Delta_{t+1}^2\Big) && \text{Substituting in 16.1} \\
&= \frac{1}{2\eta}\Big(\|\eta\cdot\nabla f(x_t)\|^2 + \frac{\Delta_0^2 - \Delta_T^2}{T}\Big) \\
&\leq \frac{1}{2\eta}\Big(\|\eta\cdot\nabla f(x_t)\|^2 + \frac{\Delta_0^2}{T}\Big) \\
&\leq \frac{1}{T\cdot 2\eta}[\eta^2 G^2 + D^2\frac{1}{T}] \\
&\leq \frac{DG}{\sqrt{T}} && \text{Plugging in } \eta_t = \frac{D}{G\sqrt{T}} \\
&\leq \epsilon && \text{and } T = \big(\frac{GD}{\varepsilon}\big)^2
\end{aligned}
$$

□

Remark 16.6. The algorithm's path resembles that of a random walk towards the optimum goal (see Fig 16.2). In practice, one may choose to vary the step length as time progresses. The above

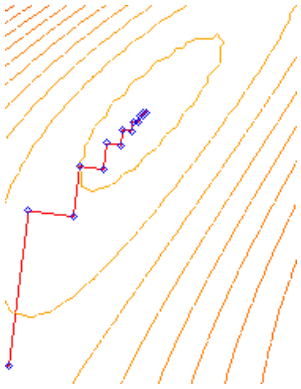

Figure 16.2: The yellow lines denote the level sets of the function f and the red walk denotes the steps of gradient descent. [Com06]

analysis is tight assuming lipschitzness.

2.2 Constrained Convex Minimization

Unlike the case of UCM, the derivative may not be 0 at the optimum. Nonetheless, the main idea of gradient descent still yields a good algorithm. We would like to take steps opposite to the gradient but the update rule needs to ensure that the new point x_{t+1} lies within K. To ensure this, we simply project each step back onto K. Let $\prod_K : \mathbb{R}^n \to K$ be defined as

$$\prod_K(y) = \text{argmin}_{x\in K}\,\|x - y\|$$

The modified algorithm is given in Algorithm 2 with changes highlighted in blue. We will show

Algorithm 2: Gradient Descent For CCM

for $t \leftarrow 0$ **to** $T-1$ **do**
  $y_{t+1} \leftarrow x_t - \eta_t \cdot \nabla f(x_t)$;
  $x_{t+1} \leftarrow \prod_K (y_{t+1})$;
end
Result: $\hat{x} = \frac{1}{T}\sum_{t=0}^{T-1} x_i$

below that a theorem (and analysis) similar to that of Theorem 16.5 holds.

Theorem 16.7. *Given a convex set K with diameter D and any convex function $f : \mathbb{R}^n \to \mathbb{R}$ with $\forall x \in \mathbb{R}^n \ \|\nabla f(x)\|_2 \leq G$ and suppose $x_0 \in K$ then $f(\hat{x}) - f(x^*) \leq \epsilon$ if we set $T = \left(\frac{GD}{\epsilon}\right)^2$ and $\eta_t = \frac{D}{G\sqrt{T}}$.*

Proof.

$$\begin{aligned}
f(x_t) - f(x^*) &\leq \langle \nabla f(x_t), x_t - x^* \rangle && \text{By convexity of } f \\
&= \left\langle \frac{1}{\eta} x_t - y_{t+1}, x_t - x^* \right\rangle && \text{By update rule} \\
&= \frac{1}{2\eta}\Big(\|x_t - y_{t+1}\|^2 + \|x_t - x_t^*\|^2 - \|y_{t+1} - x^*\|^2 \Big) && \text{Using Fact F1} \\
&\leq \frac{1}{2\eta}\Big(\|x_t - y_{t+1}\|^2 + \|x_t - x_t^*\|^2 - \|x_{t+1} - x^*\|^2 \Big) && \text{By definition of } \prod_K \\
&= \frac{1}{2\eta}\Big(\|\eta \cdot \nabla f(x_t)\|^2 + \Delta_t^2 - \Delta_{t+1}^2 \Big)
\end{aligned}$$

The rest of the argument is identical to the proof of Theorem 16.5. □

3 Relation with Multiplicative Weights

Consider the following problem: At each time step, you propose an x_t and an adversary exhibits a function f_t with $\|\nabla f_t\| \leq G$. The cost of each time step is $f_t(x_t)$ and your objective is to minimize regret.

To solve this problem, we can use the same algorithm (with one slight modification). The update rule is now taken with respect to gradient of the current function f_t.

$$x_{t+1} \leftarrow x_t - \eta_t \cdot \nabla f_{\mathbf{t}}(x_t)$$

The same analysis shows that if $T \geq \left(\frac{DG}{\varepsilon}\right)^2$

$$\sum_{t=0}^{T} \big(f_t(x_t) - f_t(x^*)\big) \leq \frac{DG}{\sqrt{T}} \leq \varepsilon$$

One advantage of this algorithm (and analysis) is that it holds for all convex bodies, as opposed to MW algorithm which holds just for the simplex. However, it now depends D and G (instead of $\log(\#\text{ of experts})$), while the $(\frac{1}{\varepsilon^2})$ dependence is the same.

In the special case of linear functions (i.e. $f_i(x) = \langle l_i, x \rangle$) and K is the simplex, the previous theorem will give us $\frac{2N}{\varepsilon^2}$ as $\text{diam}(K) = \sqrt{2}$ and $\|\nabla l_i\| \leq \sqrt{N}$. This is substantially worse than the guarantees that multiplicative weights provides, but we will see how this can be improved.

4 Further Assumptions and Modifications

4.1 Sub-Gradients

Definition 16.8. A function z is called a sub-gradient if

$$\forall y \in \mathbb{R}^n \, f(y) \geq f(x) + \langle z(x), y - x \rangle$$

If a function is not differentiable, we can use sub-gradient instead and the entire proof goes through again. Even an approximate sub-gradient suffices.

4.2 β-smooth function

Definition 16.9. A function f is a β-smooth convex function if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|x - y\|^2 \text{ for all } x, y \in K$$

Fact 16.10. *f is l-smooth convex function $\Leftrightarrow H(f) \succeq \beta I$.*

In this case, the gradient descent algorithm with step size $\eta_t = O\left(\frac{1}{lt}\right)$ yields an $\hat{x}$ which satisfies $f(\hat{x}) - f(x^*) \leq \varepsilon$ when $T = O(\frac{G^2 \log D}{\varepsilon})$.

4.3 l-strongly convex functions

Definition 16.11. A function is l-strongly convex if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{l}{2} \|x - y\|^2 \text{ for all } x, y \in K$$

Fact 16.12. *f is t-smooth convex function $\Leftrightarrow H(f) \preceq lI$.*

In this case, the gradient descent algorithm converges to a solution with error ε in $T = O(\frac{1}{\varepsilon})$

4.4 Well-conditioned Functions

Functions that are β-smooth and l-strongly convex are known as "well-conditioned" functions with condition number $\frac{\beta}{l}$. In this case, we get the much stronger convergence in time $T = O(\log(\frac{1}{\varepsilon}))$.

Theorem 16.13. *Given a function f which is β-smooth and l-strongly convex and let x^* be the solution to the unconstrained convex minimization problem $argmin_{x \in \mathbb{R}^n} f(x)$ then*

$$\|x_t - x^*\|^2 \leq \exp(-\delta t) \|x_t - x^*\|^2 \qquad \text{where } \delta = \delta(\beta, l)$$

$$f(x_t) - f(x^*) \leq \frac{\beta}{2} \exp\left(\frac{-4t}{\beta/l + 1}\right)$$

The proof of the above theorem and many others is included in the upcoming book by Hazan (see Chapter 3 in [Haz]).

References

[Com06] Wikimedia Commons. Gradient ascent(countour), 2006. Available at http://en.wikipedia.org/wiki/Gradient_descent#/media/File:Gradient_ascent_(contour).png. 16.2

[Haz] Elad Hazan. Introduction to Online Convex Optimization. Graduate text in machine learning and optimization. Available at http://ocobook.cs.princeton.edu/. 4.4

[VT] Nisheeth Vishnoi and Jakub Tarnawski. Fundamentals of convex optimization. lecture 1 basics, gradient descent and its variants. Available at http://tcs.epfl.ch/files/content/sites/tcs/files/Lec1-Fall14-Ver1.pdf. 16.1