10-425/625: Introduction to Convex Optimization (Fall 2023)

Lecture 9: Gradient Descent Convergence / Subgradients

Instructor:[1] *Matt Gormley* *September 25, 2023*

9.1 GD on Smooth Functions

9.1.1 Smooth Possibly Non-Convex Functions

For a not necessarily convex problem, we should not expect to be able to find a point which is a global optimum. Instead we'll settle for finding a point with small gradient norm, i.e. a point x for which $\|\nabla f(x)\|_2 \leq \epsilon$ (say). These points are called ϵ-substationary.

These points are called approximate saddle points (points where the gradient is 0 are called saddle points).

The main "descent" lemma:

Lemma 9.1. *For any step-size $\eta \leq 2/\beta$, the GD algorithm is a descent algorithm. For any $\eta \leq 1/\beta$ it further satisfies,*

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

Proof: (You will complete this in HW2.) ■

Worth noting that some (pretty miraculous) facts are true:

1. If $\|\nabla f(x^t)\|_2 > 0$ then we have strict descent, i.e. $f(x^{t+1}) < f(x^t)$.
2. Furthermore, if the gradient is large (in norm) then an iteration of GD decreases the function by a large amount.
3. Just by smoothness (no convexity), we already see that GD doesn't suffer from the "bouncing around" problem it encounters when applied to the (non-smooth) function $|x|$, even with a fixed step-size.

[1]These notes were originally written by Siva Balakrishnan for 10-725 Spring 2023 (original version: here) and were edited and adapted for 10-425/625.

9.1.1.1 The main theorem

Theorem 9.2. *Let x^* be any minimizer of f, then GD with step-size $\frac{1}{\beta}$ has the property that within k iterations it will reach a point x such that*

$$\|\nabla f(x)\|_2 \leq \sqrt{\frac{2\beta}{k}(f(x^0) - f(x^*))}.$$

Proof: (You will complete this in HW2.) ∎

Dimension-free: It is worth noticing an amazing fact about the above result, and more generally about many of the results about optimization algorithms you will see in this course. The result is completely dimension-free, i.e. the error goes doesn't depend at all on the ambient dimension d.

9.1.2 Gradient Descent on Smooth Convex Functions

Before we prove a result it's worth understanding how convexity might help us. In the previous section, we already showed GD will find a point with small gradient norm. We know that for convex functions we have the upper bound:

$$f(x) - f(x^*) \leq \nabla f(x)^T(x - x^*) \leq \|\nabla f(x)\|\|x - x^*\|,$$

by Cauchy-Schwarz. Suppose that we initialize in some finite neighborhood of x^*, i.e. that $\|x - x^*\| \leq R$. Intuitively, just by convexity we already know that if the gradient is small we must be close to the optimum (in function value) – this is one of the key properties of convex functions. Our subsequent proof will be a refinement of this basic intuition.

Theorem 9.3. *Let x^* be any minimizer of f, then GD with step-size $\eta = \frac{1}{\beta}$ has the property that after k iterations it will reach a point x^k such that*

$$f(x^k) - f(x^*) \leq \frac{\beta}{2k}\|x^0 - x^*\|^2.$$

1. It is worth noting that now we obtain a global guarantee (i.e. GD will find a point as good as the best point x^*). However, the guarantee is still much slower than the one we derived earlier for quadratics. To obtain ϵ-error we need to take roughly $1/\epsilon$ steps.

2. This proof – and many proofs in convex optimization will follow a few elementary steps. It might be a bit mysterious at first, but you'll get the hang of it. Usually, the steps are playing with quadratics (i.e. some form of the Pythagorean theorem) and then using the conditions (convexity, smoothness, strong convexity) in a clever way.

Proof: Notice that, for any $t \in \{1, \ldots, k\}$

$$\begin{aligned}\|x^t - x^*\|_2^2 &= \|x^{t-1} - \eta \nabla f(x^{t-1}) - x^*\|_2^2 \\ &= \|x^{t-1} - x^*\|_2^2 - 2\eta \nabla f(x^{t-1})^T (x^{t-1} - x^*) + \eta^2 \|\nabla f(x^{t-1})\|_2^2.\end{aligned}$$

The first step above is simply substituting in the gradient descent update. The second step follows from the fact that for any two vectors $a, b \in \mathbb{R}^n$, we have that $\|a-b\|_2^2 = (a-b)^T(a-b) = a^Ta - 2a^Tb + b^Tb = \|a\|_2^2 - 2a^Tb + \|b\|_2^2$. Now, rearranging, we obtain:

$$\Rightarrow \nabla f(x^{t-1})^T (x^{t-1} - x^*) = \frac{1}{2\eta}\|x^{t-1} - x^*\|_2^2 - \|x^t - x^*\|_2^2 + \frac{\eta}{2}\|\nabla f(x^{t-1})\|_2^2. \tag{9.1}$$

By our main descent lemma (which holds even without convexity) we know that for our choice of step-size,

$$\|\nabla f(x^{t-1})\|_2^2 \leq \frac{2}{\eta}(f(x^{t-1}) - f(x^t)). \tag{9.2}$$

By convexity (i.e. a rearrangement of the first order convexity condition applied to x^{t-1} and x^*), we know that,

$$f(x^{t-1}) - f(x^*) \leq \nabla f(x^{t-1})^T (x^{t-1} - x^*).$$

So we obtain from Equation (9.1) and Equation (9.2),

$$\begin{aligned}f(x^{t-1}) - f(x^*) &\leq \nabla f(x^{t-1})^T (x^{t-1} - x^*) \\ &\leq \frac{1}{2\eta}\left(\|x^{t-1} - x^*\|_2^2 - \|x^t - x^*\|_2^2\right) + f(x^{t-1}) - f(x^t).\end{aligned}$$

Simply adding $(f(x^t) - f(x^{t-1}))$ to both sides gives us the fact that,

$$f(x^t) - f(x^*) \leq \frac{\beta}{2}\left(\|x^{t-1} - x^*\|_2^2 - \|x^t - x^*\|_2^2\right).$$

Summing from $t = 1, \ldots, k$ (and dividing by k), and dropping the remaining negative term we obtain that,

$$\frac{1}{k}\sum_{t=1}^{k} f(x^t) - f(x^*) \leq \frac{\beta}{2k}\|x^0 - x^*\|_2^2.$$

Now, we can conclude the proof by noticing that for our choice of step-size, $f(x^k) \leq f(x^t)$ for $t = \{1, \ldots, k\}$ (i.e. GD is a descent algorithm) and so, after k iterations we reach a point x^k s.t.:

$$f(x^k) - f(x^*) \leq \frac{\beta}{2k}\|x^0 - x^*\|_2^2. \tag{9.3}$$

■

Theorem 9.4. *We say gradient descent on a β-smooth, convex function has convergence rate $O(1/k)$. That is, it finds ϵ-suboptimal point in $O(1/\epsilon)$ iterations.*

Proof: We first define ϵ as,

$$\epsilon = \frac{\beta}{2k}\|x^0 - x^*\|_2^2$$

so that, $f(x^k) - f(x^*) \leq \epsilon$ after k steps. Next we rearrange to solve for k,

$$k = \frac{\beta}{2\epsilon}\|x^0 - x^*\|_2^2 \in O(1/\epsilon)$$

Thus, gradient descent will find an ϵ-suboptimal point in $k \in O(1/\epsilon)$ iterations. ■

9.2 GD in the Smooth and Strongly Convex Case

Recall, that in our last lecture we studied GD for (nice) quadratics, and saw that it has a very fast rate of convergence. This is more generally true of GD applied to β-smooth, α-strongly convex functions. As before we will denote the *condition number* by,

$$\kappa = \frac{\beta}{\alpha}.$$

Theorem 9.5. *Let x^* denote the minimizer of f, then after k iterations the GD iterate x^k satisfies,*

$$\|x^k - x^*\|_2^2 \leq \left(1 - \frac{1}{\kappa}\right)^k \|x^0 - x^*\|_2^2.$$

As a consequence of smoothness (and the fact that $\nabla f(x^*) = 0$ we know that,

$$f(x^k) - f(x^*) \leq \frac{\beta}{2}\|x^k - x^*\|^2 \leq \frac{\beta}{2}\left(1 - \frac{1}{\kappa}\right)^k \|x^0 - x^*\|_2^2.$$

As with quadratics, to reach a point with $f(x^k) - f(x^*) \leq \epsilon$, ignoring β, κ dependent constants roughly $\log(1/\epsilon)$ iterations suffice. This (linear) convergence is much faster than GD under just smoothness and convexity (i.e. without strong convexity).

Proof: See Recitation for HW2. ∎

We have by now developed some understanding of GD, and how well it solves optimization problems where the function is β-smooth over an unconstrained domain. Our next goal will be to try to understand (unconstrained) optimization in the non-smooth setting, i.e. we'll no longer assume our function is differentiable and won't be able to rely on gradients any longer.

9.3 Subgradient Method

Exactly like gradient descent, but we replace gradients by subgradients, i.e. we initialize at x^0, and iterate:

$$x^{t+1} = x^t - \eta_t g_{x^t},$$

where $g_{x^t} \in \partial f(x^t)$, is any subgradient of f at x^t.

Since it often will not be a descent method, we'll usually keep track of the best iterate found so far, and output:

$$x^{\text{best}} = \arg\min_{t \in \{0,\ldots,k\}} f(x^t).$$

It is common to use the term subgradient method (instead of subgradient descent) since often the method is not a descent method (i.e. in most cases where we apply the method, and for reasonable choices of the step-size, function values can go up between iterations).

We'll have lots more to say about this method, but for now, let's develop some intuitions for subgradients.

9.4 Subgradients

We've defined subgradients before. We'll stick to a convex function f (although one can define subgradients more generally). For any $x \in \text{dom}(f)$ we'll say $g_x \in \partial f(x)$ if for all $y \in \text{dom}(f)$,

$$f(y) \geq f(x) + g_x^T(y - x).$$

1. For a convex f subgradients exist everywhere except in some pathological examples, on the boundary of the domain of f.
2. When unique the subgradient is equal to the gradient (and the function is differentiable).
3. The collection of vectors g_x which satisfy the above inequality form the subdifferential $\partial f(x)$.

9.4.1 Examples

Here are a couple of useful examples:

1. **Absolute Value:** We have discussed this example before: $f(x) = |x|$. Here if $x \neq 0$, then the function is differentiable and $g_x = \text{sign}(x)$. At 0 it is not differentiable, but it is easy to check that any $g \in [-1, 1]$ satisfies the above inequality, so the subdifferential $\partial f(0) = [-1, 1]$.

   To denote this more conveniently, we can define the sign function:

$$\text{sign}(x) = \begin{cases} +1, & x > 0 \\ [-1, +1], & x = 0 \\ -1, & x < 0. \end{cases}$$

Then we have that, $\partial f(x) = \text{sign}(x)$.

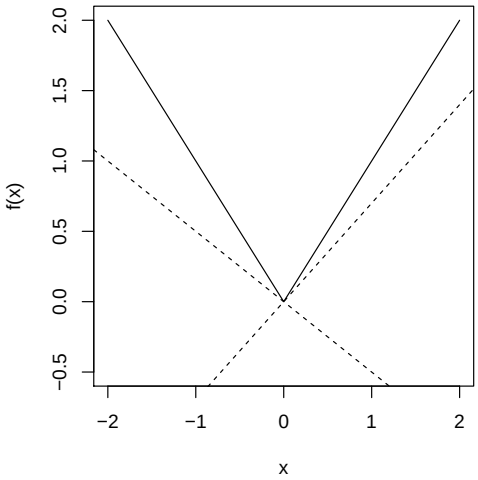

Figure 9.1: Absolute Value

2. ℓ_1 **Norm:** A slight generalization of this is: $f(x) = \|x\|_1$ where $x \in \mathbb{R}^d$. In this case, we just obtain (applying the same logic as above, elementwise) that $\partial f(x) = \text{sign}(x)$, where now we apply the sign function elementwise.

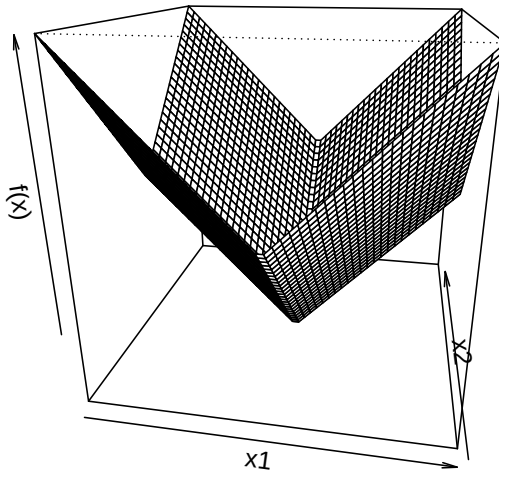

Figure 9.2: ℓ_1 norm in 3D

3. ℓ_2 **Norm** A more interesting example is when we consider $f(x) = \|x\|_2$. When $x \neq 0$ we can find the gradient directly and see that $\nabla f(x) = x/\|x\|_2$.

   The function is not differentiable at $x = 0$, so we need to check which vectors g_0 satisfy the condition that,

   $$\|y\| \geq g_0^T y,$$

   for every $y \in \mathbb{R}^d$. As a consequence of the Cauchy-Schwarz inequality, any g_0 with $\|g_0\|_2 \leq 1$ satisfies this condition, and therefore is in the subdifferential at 0.

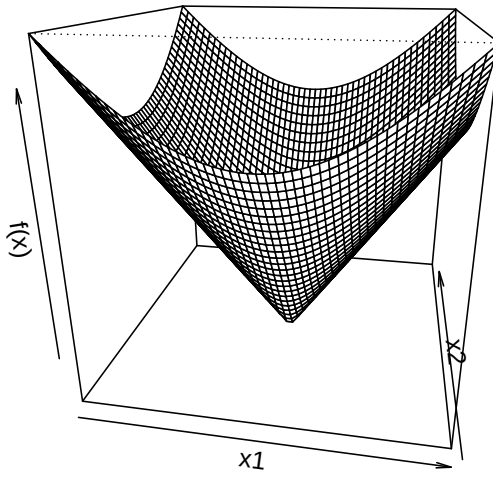

Figure 9.3: ℓ_2 norm in 3D

4. **Indicator Function of a Set:** An even more interesting example is to consider the function $f(x) = \mathbb{I}_C(x)$ the indicator function for a convex set. Now it turns out that for any $x \in C$,

$$\partial f(x) = N_C(x) = \{g \in C : g^T x \geq g^T y, \text{ for some } y \in C\},$$

i.e. the subdifferential of the indicator function is the same as the normal cone.

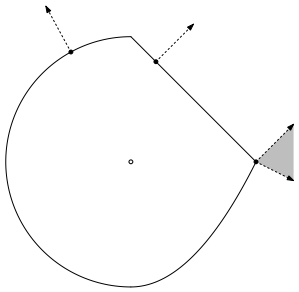

Figure 9.4: Three normal cone examples

To see this, fix a point $x \in C$ and observe that if $g_x \in \partial f(x)$ then we

must have that,

$$f(y) \geq f(x) + g_x^T(y - x) = g_x^T(y - x).$$

Now, there are two possibilities: if $y \notin C$ then the above condition is trivially satisfied (since the LHS is ∞), so the only interesting possibility is when $y \in C$. The vector g_x must thus satisfy,

$$g_x^T(y - x) \leq 0, \quad \text{for all } y \in C,$$

which is the same as requiring that $g_x \in N_C(x)$. Conversely, any vector in the normal cone is a valid subgradient via similar reasoning.