CIS3333: Mathematics of Machine Learning Fall 2025

Lecture: Convergence of Gradient Descent

Date: November 12th, 2025 *Author: Surbhi Goel*

In the last lecture, we established that for convex functions, any critical point where $\nabla f(x) = 0$ is a global minimum. Today, we will show that gradient descent converges to the global minima for smooth convex functions.

1 Smoothness (Lipschitz Gradient)

While convexity prevents a function from having multiple disconnected minima, it doesn't prevent the function's curvature from being arbitrarily large. An algorithm can easily overshoot and diverge if the gradient changes too rapidly. *Smoothness* is a condition that limits how fast the gradient can change.

Definition 1 (Smoothness)**.** *A function f is L-**smooth** for some $L > 0$ if its gradient is Lipschitz continuous with constant L. Formally, for all x, y:*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

This definition means the gradient cannot change arbitrarily fast. For twice-differentiable functions, smoothness is equivalent to having an upper bound on the Hessian, which means the function's curvature cannot be infinite.

Property 2 (Second-Order Condition for Smoothness)**.** *A twice-differentiable function f is L-smooth if and only if its Hessian is bounded above:*

$$\nabla^2 f(x) \preceq LI \quad \text{for all } x$$

This means the largest eigenvalue of the Hessian is at most L.

An L-smooth function is quadratically upper-bounded. This is a crucial consequence that we will use in our proof.

Lemma 3 (Quadratic Upper Bound)**.** *If f is L-smooth, then for all x, y:*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2$$

This inequality says that the function doesn't curve *upwards* too much faster than a quadratic with curvature L. It provides a global upper bound on the function's value.

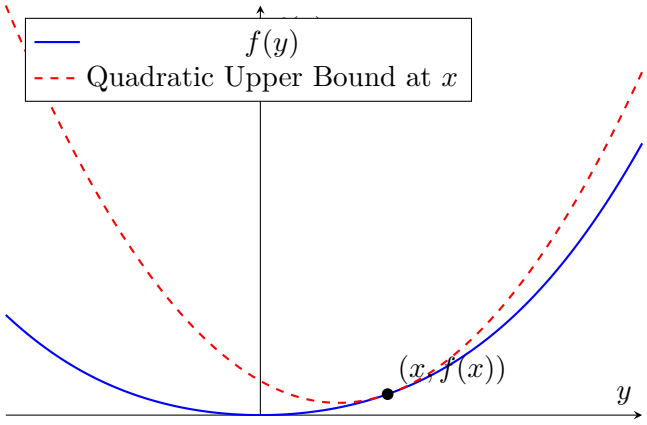

Figure 1: An L-smooth function is upper-bounded by a quadratic function at every point.

2 Convergence of Gradient Descent for Convex Functions

Recall the gradient descent algorithm:

Algorithm 1: Gradient Descent

Input: Starting point x_0, step size η, number of iterations T
Output: Final parameters x_T
Initialize x_0;
for $t = 0, \ldots, T-1$ **do**
  $x_{t+1} = x_t - \eta \nabla f(x_t)$;
end
return x_T;

We have all the pieces to prove the convergence of gradient descent for functions that are both convex and smooth. We will prove the following theorem:

Theorem 4. *Let f be a convex and L-smooth function. If we run gradient descent with a fixed step size $\eta \leq 1/L$, then for any optimal point x^*:*

$$f(x_T) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2\eta T}$$

This result shows that the "optimality gap" $f(x_T) - f(x^*)$ shrinks at a rate of $O(1/T)$. This means that to guarantee the error is less than some ϵ, we need to run for $T \geq \frac{\|x_0 - x^*\|_2^2}{2\eta\epsilon}$ iterations, so the number of iterations is $O(1/\epsilon)$.

2.1 Proof of Theorem 4

The proof rests on two key lemmas. The first, the **Descent Lemma**, guarantees that our function value decreases at each step. The second, the **Progress Lemma**, relates this decrease to the distance from the optimum. We will first prove these two lemmas and then combine them to prove the main theorem.

Lemma 5 (Descent Lemma)**.** *For an L-smooth function, a gradient descent step with $\eta \leq 1/L$ guarantees that*

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

Proof. We start with the quadratic upper bound from the smoothness property, letting $x = x_t$ and $y = x_{t+1}$:

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \quad \text{(smoothness upper bound)} \\
&= f(x_t) + \nabla f(x_t)^\top (-\eta \nabla f(x_t)) + \frac{L}{2}\| - \eta \nabla f(x_t)\|_2^2 \quad \text{(substitute } x_{t+1} = x_t - \eta \nabla f(x_t)\text{)} \\
&= f(x_t) - \eta \|\nabla f(x_t)\|_2^2 + \frac{L\eta^2}{2}\|\nabla f(x_t)\|_2^2 \quad \text{(simplify inner product and norm)} \\
&= f(x_t) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(x_t)\|_2^2 \quad \text{(factor out } \eta\|\nabla f(x_t)\|_2^2\text{)}
\end{aligned}$$

If we choose the step size $\eta \leq 1/L$, then $(1 - L\eta/2) \geq 1/2$, which gives the desired result. □

Lemma 6 (Progress towards Optimum)**.** *Let f be convex and L-smooth. For a gradient descent step with $\eta \leq 1/L$, we have:*

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

Proof. First, we analyze the change in distance to the optimum x^*:

$$\begin{aligned}
\|x_{t+1} - x^*\|_2^2 &= \|x_t - \eta \nabla f(x_t) - x^*\|_2^2 \quad \text{(substitute GD update)} \\
&= \|x_t - x^*\|_2^2 - 2\eta \nabla f(x_t)^\top (x_t - x^*) + \eta^2 \|\nabla f(x_t)\|_2^2 \quad \text{(expand squared norm)}
\end{aligned}$$

Using the first-order condition for convexity, $\nabla f(x_t)^\top (x_t - x^*) \geq f(x_t) - f(x^*)$, we can bound the middle term:

$$\|x_{t+1} - x^*\|_2^2 \leq \|x_t - x^*\|_2^2 - 2\eta(f(x_t) - f(x^*)) + \eta^2 \|\nabla f(x_t)\|_2^2 \quad \text{(apply convexity)}$$

The Descent Lemma implies $\|\nabla f(x_t)\|_2^2 \leq \frac{2}{\eta}(f(x_t) - f(x_{t+1}))$. Substituting this in:

$$\|x_{t+1} - x^*\|_2^2 \leq \|x_t - x^*\|_2^2 - 2\eta(f(x_t) - f(x^*)) + 2\eta(f(x_t) - f(x_{t+1})) \quad \text{(apply Descent Lemma)}$$

The $2\eta f(x_t)$ terms cancel. Rearranging the remaining terms gives the desired result:

$$2\eta(f(x_{t+1}) - f(x^*)) \leq \|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \quad \text{(simplify and rearrange)}$$

□

With these lemmas in hand, the proof of the main theorem is straightforward.

Proof of Theorem 4. We sum the inequality from Lemma 6 over all iterations $t = 0, \ldots, T-1$:

$$\sum_{t=0}^{T-1}(f(x_{t+1}) - f(x^*)) \leq \frac{1}{2\eta}\sum_{t=0}^{T-1}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right)$$

The right hand side is a **telescoping sum**. To see how it collapses, let's write out the first few and last terms of the sum:

$$\begin{aligned}\sum_{t=0}^{T-1}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) = &(\|x_0 - x^*\|_2^2 - \|x_1 - x^*\|_2^2) \quad \text{(t=0)} \\ &+ (\|x_1 - x^*\|_2^2 - \|x_2 - x^*\|_2^2) \quad \text{(t=1)} \\ &+ \dots \\ &+ (\|x_{T-1} - x^*\|_2^2 - \|x_T - x^*\|_2^2) \quad \text{(t=T-1)}\end{aligned}$$

The negative part of each term cancels the positive part of the next, leaving only the first part of the first term and the last part of the last term. The sum simplifies to $\|x_0 - x^*\|_2^2 - \|x_T - x^*\|_2^2$. Since norms are non-negative, we can bound this by simply dropping the negative term: $\|x_0 - x^*\|_2^2$.

Plugging this back into our main inequality, we get:

$$\sum_{t=0}^{T-1}(f(x_{t+1}) - f(x^*)) \leq \frac{1}{2\eta}\|x_0 - x^*\|_2^2$$

Since $f(x_t)$ is non-increasing (from the Descent Lemma), the final value $f(x_T)$ is the smallest value in the sequence $f(x_1), \dots, f(x_T)$. Therefore, we can bound the sum on the left:

$$T(f(x_T) - f(x^*)) \leq \sum_{t=0}^{T-1}(f(x_{t+1}) - f(x^*)) \leq \frac{\|x_0 - x^*\|_2^2}{2\eta} \quad \text{(bound sum by its smallest term)}$$

Finally, dividing by T gives the desired result.

$$f(x_T) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2\eta T} \quad \text{(divide by } T\text{)}$$

□

3 Faster Convergence under Strong Convexity

If we additionally assume the function is μ-strongly convex, we get a much faster rate.

Theorem 7. *Let f be μ-strongly convex and L-smooth. With $\eta = 1/L$, GD converges at a linear rate:*

$$\|x_T - x^*\|_2 \leq \left(1 - \frac{\mu}{L}\right)^T \|x_0 - x^*\|_2$$

This is called a *linear* convergence rate because the error is multiplied by a constant factor < 1 at each step, analogous to how a geometric series decreases. The error gets reduced by a constant factor every iteration, so the number of iterations to get ϵ error is $O(\log(1/\epsilon))$. This is much faster than the $O(1/\epsilon)$ rate for general convex functions.

While we will not see the general proof here, in the homework, we will show the result for the least squares problem.