1 Overview

In the previous lecture we reviewed results from multivariate calculus in preparation for our journey into convex optimization. In this lecture we present the gradient descent algorithm for minimizing a convex function and analyze its convergence properties.

2 The Gradient Descent Algorithm

From the previous lecture, we know that in order to minimize a convex function, we need to find a stationary point. As we will see in this lecture as well as the upcoming ones, there are different methods and heuristics to find a stationary point. One possible approach is to start at an arbitrary point, and move along the gradient at that point towards the next point, and repeat until (hopefully) converging to a stationary point. We illustrate this in the figure below.

Direction and step size. In general, one can consider a search for a stationary point as having two components: the direction and the step size. The direction decides which direction we search next, and the step size determines how far we go in that particular direction. Such methods can be generally described as starting at some arbitrary point $\mathbf{x}^{(0)}$ and then at every step $k \geq 0$ iteratively moving at direction $\Delta\mathbf{x}^{(k)}$ by step size t_k to the next point $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k \cdot \Delta\mathbf{x}^{(k)}$. In gradient descent, the direction we search is the negative gradient at the point, i.e. $\Delta\mathbf{x} = -\nabla f(\mathbf{x})$. Thus, the iterative search of gradient descent can be described through the following recursive rule:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - t_k \nabla f(\mathbf{x}^{(k)})$$

Choosing a step size. Given that the search for a stationary point is currently at a certain point $\mathbf{x}^{(k)}$, how should we choose our step size t_k? Since our objective is to minimize the function, one reasonable approach is to choose the step size in manner that will minimize the value of the new point, i.e. find the step size that minimizes $f(\mathbf{x}^{(k+1)})$. Since $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - t\nabla f(\mathbf{x}^{(k)})$ the step size $t_k^\star$ of this approach is:

$$t_k^\star = \text{argmin}_{t \geq 0} f(\mathbf{x}^{(k)} - t\nabla f(\mathbf{x}^{(k)}))$$

For now we will assume that $t_k^\star$ can be computed analytically, and later revisit this assumption.

The algorithm. Formally, given a desired precision $\epsilon > 0$, we define the gradient descent as described below.

Algorithm 1 Gradient Descent

1: Guess $\mathbf{x}^{(0)}$, set $k \leftarrow 0$
2: **while** $||\nabla f(\mathbf{x}^{(k)})|| \geq \epsilon$ **do**
3: $\quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - t_k \nabla f(\mathbf{x}^{(k)})$
4: $\quad k \leftarrow k + 1$
5: **end while**
6: **return** $\mathbf{x}^{(k)}$

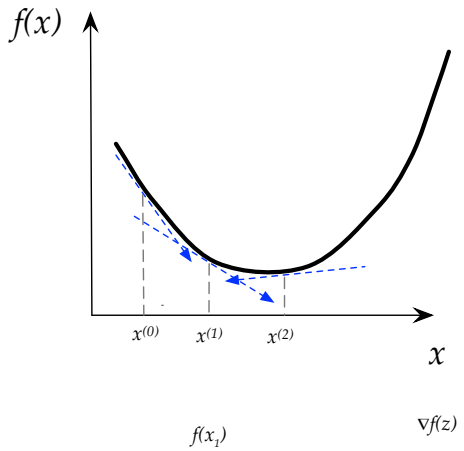

Figure 1: An example of a gradient search for a stationary point.

Some remarks. The gradient descent algorithm we present here is for *unconstrained* minimization. That is, we assume that every point we choose is feasible (inside S). In a few lectures we will see that gradient descent can be applied for constrained minimization as well. The stopping condition where we check $\|\nabla f(\mathbf{x})\| \geq \epsilon$ does not a priori guarantee us that we are ϵ close to the optimal solution, i.e. that we are at a point $\mathbf{x}^{(k)}$ for which $f(\mathbf{x}^{(k)}) - \min_{\mathbf{x} \in S} f(\mathbf{x}) \leq \epsilon$. In section however, you will show that this is implies as a consequence of the characterization of convex functions we showed in the previous lecture. Finally, computing the step size as shown here is called *exact line search*. In some cases finding $t_k^\star$ is computationally expensive and different methods are used. In your problem set this week, you will implement gradient descent and use an alternative method called *backtracking* that can be implemented efficiently.

Example. Consider the problem of minimizing $f(x, y) = 4x^2 - 4xy + 2y^2$ using the gradient descent method. Notice that the optimal solution is $(x, y) = (0, 0)$. To apply the gradient descent algorithm let's first compute the gradient:

$$\nabla f(x, y) = \begin{pmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{pmatrix} = \begin{pmatrix} 8x - 4y \\ -4x + 4y \end{pmatrix}$$

We will start from the point $(x^{(0)}, y^{(0)}) = (2, 3)$. To find the next point $(x^{(1)}, y^{(1)})$ we compute:

$$(x^{(1)}, y^{(1)}) = (x^{(0)}, y^{(0)}) - t_0^\star \nabla f(x^{(0)}, y^{(0)}).$$

To find $t_0^\star$ we need to find the minimum of the function $\theta(t) = f\left((x^{(0)}, y^{(0)}) - t\nabla f(x^{(0)}, y^{(0)})\right)$. To do this we will look for the stationary point:

$$\begin{aligned}
\theta'(t) &= -\nabla f\Big((x^{(0)}, y^{(0)}) - t\nabla f(x^{(0)}, y^{(0)})\Big)^\intercal \nabla f(x^{(0)}, y^{(0)}) \\
&= -\nabla f(2-4t, 3-4t)^\intercal \begin{pmatrix} 4 \\ 4 \end{pmatrix} \\
&= -\Big(8(2-4t) - 4(3-4t), -4(2-4t) + 4(3-4t)\Big)^\intercal \begin{pmatrix} 4 \\ 4 \end{pmatrix} \\
&= -16(2-4t) \\
&= -32 + 64t
\end{aligned}$$

In this case $\theta'(t) = 0$ if and only $t = 1/2$. Since the function $\theta(t)$ is convex, the stationary point is a global minimum. Therefore, $t_0 = 1/2$.

The next point will be:

$$(x^{(1)}, y^{(1)}) = \begin{pmatrix} 2 \\ 3 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 4 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and the algorithm will continue finding the next point by performing similar calculations as above.

It is important to note that the directions in which the algorithm proceeds are *orthogonal*. That is:

$$\nabla f(2,3)^\intercal \nabla f(0,1) = 0$$

This is due the way in which we compute the multiplier $t_k^\star$:

$$\begin{aligned}
\theta'(t) = 0 &\iff -\nabla f\Big((x^{(k)}, y^{(k)}) - t\nabla f(x^{(k)}, y^{(k)})\Big)^\intercal \nabla f(x^{(k)}, y^{(k)}) = 0 \\
&\iff \nabla f(x^{(k+1)}, y^{(k+1)})^\intercal \nabla f(x^{(k)}, y^{(k)}) = 0
\end{aligned}$$

3 Convergence Analysis of Gradient Descent

The convergence analysis we will prove will hold for *strongly convex* functions, defined below. We will first show some important properties of strongly convex functions, and then use these properties in the proof of the convergence of gradient descent.

3.1 Strongly convex functions

Definition. *For a convex set $S \subseteq \mathbb{R}^n$, a convex function $f : S \to \mathbb{R}$ is called* ***strongly convex*** *if there exist constants $m < M \in \mathbb{R}_{\geq 0}$ s.t.:*

$$mI \leq H_f(\mathbf{x}) \leq MI$$

It is important to observe the relationship between strictly convex and strongly convex functions, as we do in the following claim.

Claim 1. *Let f be a strongly convex function, then f is strictly convex.*

Proof. For any $\mathbf{x}, \mathbf{y} \in S$, from the second-order Taylor expansion we know that there exists a $\mathbf{z} \in [\mathbf{x}, \mathbf{y}]$:

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^\intercal H_f(\mathbf{z})(\mathbf{y} - \mathbf{x})$$

strong convexity implies there exists a constant $m > 0$ s.t.:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\mathbf{y} - \mathbf{x}) + \frac{m}{2}\|y - x\|_2^2$$

and hence:

$$f(\mathbf{y}) > f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\mathbf{y} - \mathbf{x})$$

In the previous lecture we proved that a function is convex if and only if, for every $\mathbf{x}, \mathbf{y} \in S$:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\mathbf{y} - \mathbf{x})$$

The exact same proof shows that a function is *strictly* convex if and only if, for every $\mathbf{x}, \mathbf{y} \in S$ the above inequality is strict. Thus strong convexity indeed implies a strict inequality and hence the function is strictly convex. □

Lemma 2. *Let $f : S \to \mathbb{R}$ be a* strongly *convex function with parameters m, M as in the definition above, and let $\alpha^\star = \min_{\mathbf{x} \in S} f(\mathbf{x})$. Then:*

$$f(\mathbf{x}) - \frac{1}{2m}\|\nabla f(\mathbf{x})\|_2^2 \leq \alpha^\star \leq f(\mathbf{x}) - \frac{1}{2M}\|\nabla f(\mathbf{x})\|_2^2$$

Proof. For any $\mathbf{x}, \mathbf{y} \in S$, from the second-order Taylor expansion we know that there exists a $\mathbf{z} \in [\mathbf{x}, \mathbf{y}]$:

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^\intercal H_f(\mathbf{z})(\mathbf{y} - \mathbf{x})$$

strong convexity implies there exists a constant $m > 0$ s.t.:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\mathbf{y} - \mathbf{x}) + \frac{m}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$$

The function:

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\mathbf{y} - \mathbf{x}) + \frac{m}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$$

is convex quadratic in $\mathbf{y}$ and minimized at $\tilde{\mathbf{y}} = \mathbf{x} - \frac{1}{m}\nabla f(\mathbf{x})$. Therefore we can apply the above inequality to show that for any $\mathbf{y} \in S$ we have that $f(\mathbf{y})$ is lower bounded by the convex quadratic function at $\mathbf{y} = \tilde{\mathbf{y}}$:

$$\begin{aligned}
f(\mathbf{y}) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\tilde{\mathbf{y}} - \mathbf{x}) + \frac{m}{2}\|\tilde{\mathbf{y}} - \mathbf{x}\|_2^2 \\
&= f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal (\mathbf{x} - \frac{1}{m}\nabla f(\mathbf{x}) - \mathbf{x}) + \frac{m}{2}\|\mathbf{x} - \frac{1}{m}\nabla f(\mathbf{x}) - \mathbf{x}\|_2^2 \\
&= f(\mathbf{x}) - \frac{1}{m}\nabla f(\mathbf{x})^\intercal \nabla f(\mathbf{x}) + \frac{m}{2} \cdot \frac{1}{m^2}\|\nabla f(\mathbf{x})\|_2^2 \\
&= f(\mathbf{x}) - \frac{1}{m}\|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{2m}\|\nabla f(\mathbf{x})\|_2^2 \\
&= f(\mathbf{x}) - \frac{1}{2m}\|\nabla f(\mathbf{x})\|_2^2
\end{aligned}$$

Since this holds for any $\mathbf{y} \in S$, it holds for $\mathbf{y}^\star = \text{argmin}_{\mathbf{y}\in S} f(\mathbf{y})$, which implies the first side of our desired inequality. In a similar manner we can show the other side of the inequality by relying on the second-order Taylor expansion, upper bound of the Hessian H_f by MI and choosing $\tilde{\mathbf{y}} = \mathbf{x} - \frac{1}{M}\nabla f(\mathbf{x})$. □

3.2 Convergence of gradient descent

Theorem 3. *Let $f : S \to \mathbb{R}$ be a strongly convex function with parameters m, M as in the definition above. For any $\epsilon > 0$ we have that $f(\mathbf{x}^{(k)}) - \min_{\mathbf{x}\in S} f(\mathbf{x}) \leq \epsilon$ after $k^\star$ iterations for any $k^\star$ that respects:*

$$k^\star \geq \frac{\log\left(\frac{f(\mathbf{x}^{(0)})-\alpha^\star}{\epsilon}\right)}{\log\left(\frac{1}{1-m/M}\right)}$$

Proof. For a given step k define the optimal step size $t_k^\star = \text{argmin}_{t\geq 0} f(\mathbf{x}^{(k)} - t\nabla f(\mathbf{x}^{(k)}))$. From the second-order Taylor expansion we have that:

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^\intercal H_f(\mathbf{z})(\mathbf{y}-\mathbf{x})$$

Together with strong convexity we have that:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y}-\mathbf{x}) + \frac{M}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$$

For $\mathbf{y} = \mathbf{x}^{(k)} - t\nabla f(\mathbf{x}^{(k)})$ and $\mathbf{x} = \mathbf{x}^{(k)}$we get:

$$\begin{aligned} f(\mathbf{x}^{(k)} - t\nabla f(\mathbf{x}^{(k)})) &\leq f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)})^\intercal(-t\nabla f(\mathbf{x}^{(k)})) + \frac{M}{2}\|-t\nabla f(\mathbf{x}^{(k)})\|_2^2 \\ &= f(\mathbf{x}^{(k)}) - t\cdot\|\nabla f(\mathbf{x}^{(k)})\|_2^2 + \frac{M}{2}\cdot t^2\|\nabla f(\mathbf{x}^{(k)})\|_2^2 \end{aligned}$$

In particular, using $t = t_M = 1/M$ we get:

$$\begin{aligned} f(\mathbf{x}^{(k)} - t_M\nabla f(\mathbf{x}^{(k)})) &\leq f(\mathbf{x}^{(k)}) - \frac{1}{M}\cdot\|\nabla f(\mathbf{x}^{(k)})\|_2^2 + \frac{M}{2}\cdot\frac{1}{M^2}\|\nabla f(\mathbf{x}^{(k)})\|_2^2 \\ &= f(\mathbf{x}^{(k)}) - \frac{1}{M}\cdot\|\nabla f(\mathbf{x}^{(k)})\|_2^2 + \frac{1}{2M}\|\nabla f(\mathbf{x}^{(k)})\|_2^2 \\ &= f(\mathbf{x}^{(k)}) - \frac{1}{2M}\cdot\|\nabla f(\mathbf{x}^{(k)})\|_2^2 \end{aligned}$$

By the minimality of $t_k^\star$ we know that $f(\mathbf{x}^{(k)} - t_k^\star\nabla f(\mathbf{x}^{(k)})) \leq f(\mathbf{x}^{(k)} - t_M\nabla f(\mathbf{x}^{(k)}))$ and thus:

$$f(\mathbf{x}^{(k)} - t_k^\star\nabla f(\mathbf{x}^{(k)})) \leq f(\mathbf{x}^{(k)}) - \frac{1}{2M}\cdot\|\nabla f(\mathbf{x}^{(k)})\|_2^2$$

Notice that $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - t_k^\star\nabla f(\mathbf{x}^{(k)})$ and thus:

$$f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)}) - \frac{1}{2M}\cdot\|\nabla f(\mathbf{x}^{(k)})\|_2^2$$

subtracting $\alpha^\star = \min_{\mathbf{x}\in S} f(\mathbf{x})$ from both sides we get:

$$f(\mathbf{x}^{(k+1)}) - \alpha^\star \leq f(\mathbf{x}^{(k)}) - \alpha^\star - \frac{1}{2M} \cdot \|\nabla f(\mathbf{x}^{(k)})\|_2^2$$

Applying Lemma 2 we know that $\|\nabla f(\mathbf{x}^{(k)})\|_2^2 \geq 2m(f(\mathbf{x}^{(k)}) - \alpha^\star)$, hence:

$$\begin{aligned} f(\mathbf{x}^{(k+1)}) - \alpha^\star &\leq f(\mathbf{x}^{(k)}) - \alpha^\star - \frac{1}{2M} \cdot \|\nabla f(\mathbf{x}^{(k)})\|_2^2 \\ &\leq f(\mathbf{x}^{(k+1)}) - \alpha^\star - \frac{m}{M}\left(f(\mathbf{x}^{(k+1)}) - \alpha^\star\right) \\ &= \left(1 - \frac{m}{M}\right)\left(f(\mathbf{x}^{(k+1)}) - \alpha^\star\right) \end{aligned}$$

Applying this rule recursively on our initial point $\mathbf{x}^{(0)}$ we get:

$$f(\mathbf{x}^{(k+1)}) - \alpha^\star \leq \left(1 - \frac{m}{M}\right)^k \left(f(\mathbf{x}^{(0)}) - \alpha^\star\right)$$

Thus, $f(\mathbf{x}^{(k)}) - \alpha^\star \leq \epsilon$ when

$$k \geq \frac{\log\left(\frac{f(\mathbf{x}^{(0)}) - \alpha^\star}{\epsilon}\right)}{\log\left(\frac{1}{1-m/M}\right)}.$$

□

Notice that the rate converges to ϵ both as a function of how far our initial point was from the optimal solution, as well as the ratio between m and M. As m and M get closer, we have tigher bounds on the strong convexity property of the function, and the algorithm converges faster as a result.

4 Further Reading

For further reading on gradient descent and general descent methods please see Chapter 9 of the Convex Optimization book by Boyd and Vandenberghe.