Lecture #3 — 2/2, 2022

Lecturer: Yura Malitsky *Scribe: Robin Forsling*

1 Gradient Descent

Let $f\colon \mathbb{R}^n \to \mathbb{R}$. In this lecture an unconstrained problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x), \tag{1}$$

is considered. The *gradient descent* (GD) method iteratively solves (1) by the following recursion

$$\begin{aligned} x_0 &\in \mathbb{R}^n, \\ x_{k+1} &= x_k - \alpha_k \nabla f(x_k), \end{aligned} \tag{2}$$

where x_0 is the initial point, $\alpha_k > 0$ is the step size and $\nabla f(x_k)$ is the gradient of $f(x)$ at $x = x_k$. The focus of this lecture is on interpreting the GD method and analyze its convergence properties.

2 Interpretation

In this section several interpretations of GD are provided.

2.1 Fixed Point Operator

Let $T(x)\colon \mathbb{R}^n \to \mathbb{R}^n$ be an operator. A *fixed point* $\bar{x}$ of $T(x)$ is given by the equation

$$T(\bar{x}) = \bar{x}.$$

Let $T(x) \triangleq x - \nabla f(x)$, then *critical points* of f are found by solving for fixed points of $T(x)$, *i.e.*,

$$T(\bar{x}) = \bar{x} \implies \nabla f(\bar{x}) = 0.$$

An attractive fixed point can be to computed iteratively by considering

$$x_{k+1} = T(x_k), \tag{3}$$

which, for $\alpha_k = 1$, is identical to (2) since

$$\begin{aligned} x_0 &\in \mathbb{R}^n, \\ x_{k+1} &= T(x_k) = x_k - \nabla f(x_k). \end{aligned} \tag{4}$$

2.2 Taylor Series Expansion

If f is a complicated function but still the objective is to solve (1) it is reasonable to approximate f. A *Taylor series* expansion of $f(x)$ around x_k is given by

$$f(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \mathcal{O}(\|x_k\|^3), \tag{5}$$

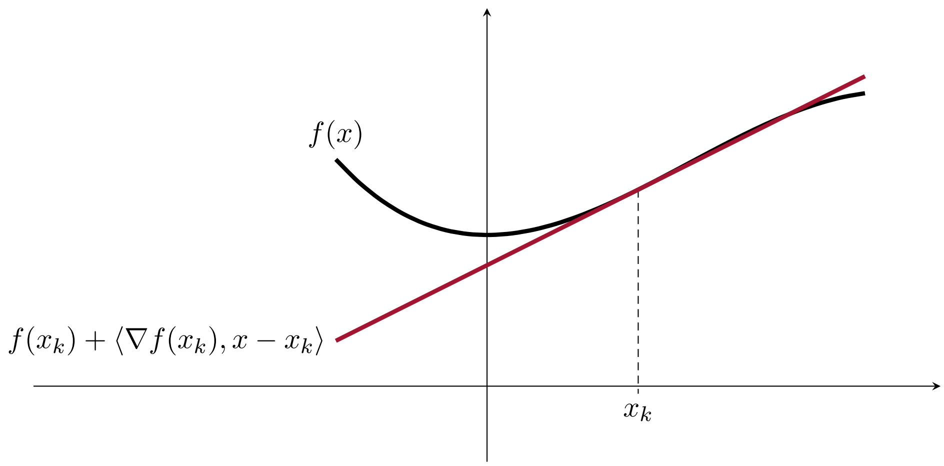

Figure 1: Approximating $f(x)$ by a linear function $f(x_k) + \langle \nabla f(x_k), x - x_k \rangle$. Since a linear function is unbounded this approximation is not suitable as the next iterate x_{k+1} would tend to $-\infty$.

where $\mathcal{O}(\|x_k\|^3)$ collects higher-order terms. For a first-order Taylor approximation $f(x) \approx f(x_k) + \langle \nabla f(x_k), x - x_k \rangle$ the minimization problem in (1) reduces to

$$x_{k+1} = \arg\min_x \quad \{f(x_k) + \langle \nabla f(x_k), x - x_k \rangle\}.$$

This approximation yields a function which is linear in x and since linear functions are unbounded this approximation is too rough, see Figure 1. Including the second-order term, see Figure 2, $\frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle$ yields

$$x_{k+1} = \arg\min_x \quad \left\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle \right\},$$

where the constant term has been removed. If the *Hessian* $\nabla^2 f(x_k)$ is positive semi-definite this problem has a solution.

If $\nabla^2 f(x_k)$ is unavailable but substituted with $\frac{1}{\alpha_k} I$, x_{k+1} is given by

$$x_{k+1} = \arg\min_x \quad \left\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k}\|x - x_k\|^2 \right\}.$$

Differentiating $\langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k}\|x - x_k\|^2$ w.r.t. x yields

$$\nabla f(x_k) + \frac{1}{\alpha_k}(x - x_k).$$

Equating to zero and denoting the solution by x_{k+1} yields

$$\nabla f(x_k) + \frac{1}{\alpha_k}(x_{k+1} - x_k) = 0 \iff x_{k+1} = x_k - \alpha_k \nabla f(x_k), \tag{6}$$

which is the GD method.

2.3 Steepest Descent

Consider the problem

$$\begin{aligned} &\arg\max_{\|d\|=1} \quad \lim_{t \to 0} \frac{f(x + td) - f(x)}{t} \\ &= \arg\max_{\|d\|=1} \quad \langle \nabla f(x), d \rangle. \end{aligned} \tag{7}$$

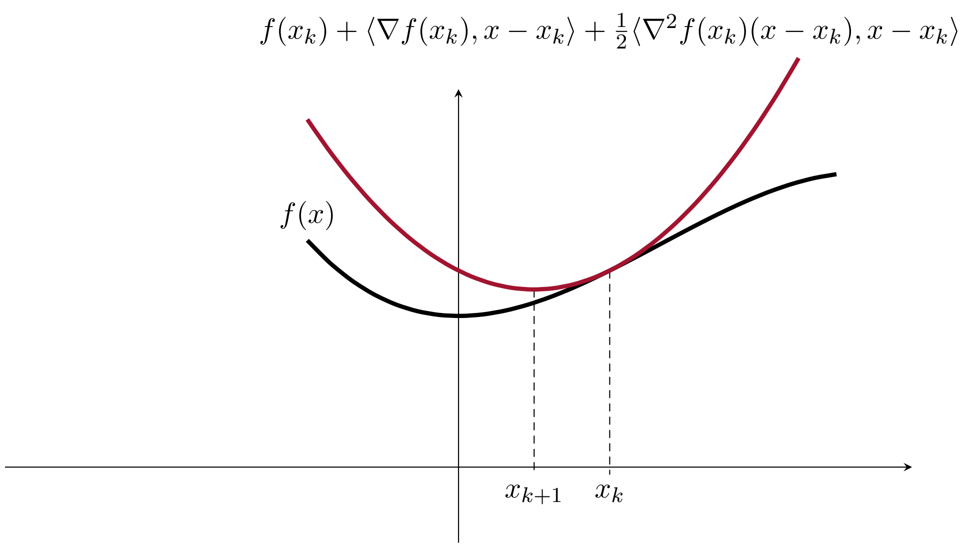

Figure 2: Approximating $f(x)$ by a quadratic function function $f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle$. The next iterate x_{k+1} is found by minimizing the quadratic approximation.

Maximizing the inner product $\langle \nabla f(x), d \rangle$ is accomplished by a vector parallel to $\nabla f(x)$, *i.e.*,

$$d^\star = \frac{\nabla f(x)}{\|\nabla f(x)\|}, \tag{8}$$

where normalization is included for $d^\star$ to satisfy $\|d^\star\| = 1$. This means that moving in the gradient direction $d^\star$ is equal to moving in the direction of steepest ascent locally at x. If instead $-d^\star$ is considered the direction of steepest descent is retrieved. By continuously pointing in the local value of $d^\star = d^\star(x)$, (non-strictly) smaller and smaller values of f are traversed. With a properly chosen step size, a local minimum will eventually be found, given that such exist.

3 Convergence

In this section convergence of the GD is analyzed, but first a couple of useful concepts are introduced.

3.1 L-Smooth Functions and the Descent Lemma

Definition 1 (L-smooth function)**.** *The function $f\colon \mathbb{R}^n \to \mathbb{R}$ is L-smooth if and only if*

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|, \quad \forall x, y \in \mathbb{R}^n, \tag{9}$$

where $L \geq 0$ is the Lipschitz constant.

Example: L-Smooth Function

Let $f(x) = x^2$. The gradient is $\nabla f(x) = 2x$. Since

$$\|\nabla f(y) - \nabla f(x)\| = |2y - 2x| \leq L|y - x| = L\|y - x\|, \quad \forall x, y \in \mathbb{R},$$

is satisfied for $L = 2$, $f(x)$ is L-smooth.

Example: Non-L-Smooth Function

Let $f(x) = x^3$. The gradient is $\nabla f(x) = 3x^2$. With $x = 0$, the l.h.s. of

$$|3y^2| \leq L|y|,$$

will grow faster than the r.h.s. and hence $|\nabla f(y) - \nabla f(x)|$ cannot be bounded by $L|y - x|$, $\forall x, y \in \mathbb{R}$.

Proposition 2 (L-smoothness of twice-differentiable functions)**.** *If f is twice-differentiable, then the condition in (9) is equivalent to*

$$\lambda_{\max}\left(\nabla^2 f(x)\right) \leq L, \quad \forall x, \tag{10}$$

where $\lambda_{\max}\left(\nabla^2 f(x)\right)$ is the maximum eigenvalue of the Hessian $\nabla^2 f(x)$ of f.

Proof. The mean-value theorem is given in the notes of Lecture 2. It states that for any $x, y \in \mathbb{R}^n$ there exists a $z \in \mathbb{R}^n$ in between x and y such that

$$\nabla f(y) = \nabla f(x) + \nabla^2 f(z)(y - x).$$

Using $y = x + td$, where $d \in \mathbb{R}^n$ and $t \geq 0$ is a scalar, (9) can be written as

$$\begin{aligned} &\|\nabla^2 f(z) td\| \leq L\|td\| \\ \iff &\|\nabla^2 f(z) d\| \leq L\|d\|. \end{aligned}$$

Let $t \to 0$, by continuity $x = y = z$ and hence

$$\|\nabla^2 f(x) d\| \leq L\|d\|. \tag{11}$$

Now, since the maximum eigenvalue $\lambda_{\max}(A)$ of a matrix A is given by

$$\underset{\|u\| \neq 0}{\text{maximize}} \quad \frac{\|Au\|}{\|u\|},$$

and since (11) holds for all d and in particular for d associated with $\lambda_{\max}$, we have that

$$\lambda_{\max}\left(\nabla^2 f(x)\right) \leq L.$$

□

Lemma 3 (Descent lemma)**.** *If $f\colon \mathbb{R}^n \to \mathbb{R}$ is L-smooth with Lipschitz constant L, then*

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|y - x\|^2. \tag{12}$$

Proof. Let $z = x + t(y - x)$ where $t \in [0, 1]$. By the fundamental theorem of calculus

$$f(y) - f(x) = \int_0^1 \frac{d}{dt} f(z) dt = \int_0^1 \langle \nabla f(z), y - x \rangle dt, \tag{13}$$

where in the second equality the chain rule

$$\frac{d}{dt} f(z(t)) = \left\langle \nabla f, \frac{d}{dt} z(t) \right\rangle,$$

was used. Subtracting $\langle \nabla f(x), y - x \rangle$ from both sides of (13) yields

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \int_0^1 \langle \nabla f(z) - \nabla f(x), y - x \rangle dt. \tag{14}$$

By taking the absolute value of the r.h.s. and using the Cauchy-Schwarz inequality, which states that

$$|\langle x, y \rangle| \leq \|x\| \|y\|,$$

we arrive at

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &\leq \left| \int_0^1 \langle \nabla f(z) - \nabla f(x), y - x \rangle dt \right| \\ &\leq \int_0^1 |\langle \nabla f(z) - \nabla f(x), y - x \rangle| \, dt \\ &\leq \int_0^1 \|\nabla f(z) - \nabla f(x)\| \|y - x\| dt. \end{aligned} \tag{15}$$

By assumption f is L-smooth, hence

$$\|\nabla f(z) - \nabla f(x)\| \leq L\|z - x\| = L\|t(y - x)\| = Lt\|y - x\|.$$

Plugging this expression into (15) finally gives us

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq L\|y - x\|^2 \int_0^1 t \, dt = \frac{L}{2} \|y - x\|^2.$$

□

3.2 Convergence Analysis

Assumption 4 (Convergence of the gradient descent method)**.** *The following assumptions are made for the convergence analysis of the GD method:*

1. $f(x) \colon \mathbb{R}^n \to \mathbb{R}$ *is L-smooth.*
2. $f(x)$ *is bounded from below,* i.e., $f(x) \geq f_{low}, \forall x$.
3. *The Lipschitz constant L is known.*

Since f is L-smooth we can use Lemma 3 with y, x replaced by x_{k+1}, x_k to get

$$f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle \leq \frac{L}{2} \|x_{k+1} - x_k\|^2.$$

Using the GD recursion $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ this can be written as

$$\begin{aligned} & f(x_{k+1}) - f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 \leq \frac{L}{2} \alpha_k^2 \|\nabla f(x_k)\|^2 \\ \iff & f(x_{k+1}) - f(x_k) \leq -\alpha_k \left(1 - \frac{\alpha_k L}{2}\right) \|\nabla f(x_k)\|^2. \end{aligned}$$

What we want is

$$-\alpha_k \left(1 - \frac{\alpha_k L}{2}\right) < 0,$$

while its absolute value is as large as possible, since in this case $f(x_{k+1}) - f(x_k)$ becomes as small as possible which is desirable w.r.t. convergence. The optimal value for the step size is $\alpha_k = \frac{1}{L}$ which gives

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2,$$

where availability of L is guaranteed by the third assumption. Now we want to make $\|\nabla f(x_k)\|$ as small as possible since a local minimum is characterized by $\nabla f(x) = 0$. To get rid of x we first use

$$\|\nabla f(x_k)\|^2 \leq 2L\left(f(x_k) - f(x_{k+1})\right),$$

and then construct the sum

$$\sum_{i=0}^{K} \|\nabla f(x_i)\|^2 \leq 2L\left(f(x_0) - f(x_{K+1})\right) \leq 2L\left(f(x_0) - f_{\text{low}}\right), \tag{16}$$

where the second assumption was used together with the fact that

$$\begin{aligned}\|\nabla f(x_{k-1})\|^2 + \|\nabla f(x_k)\|^2 &\leq 2L\left(f(x_{k-1}) - f(x_k)\right) + 2L\left(f(x_k) - f(x_{k+1})\right)\\ &= 2L\left(f(x_{k-1}) - f(x_{k+1})\right).\end{aligned}$$

Since (16) holds we have that

$$\min_i \quad \|\nabla f(x_i)\|^2 \leq \frac{2L\left(f(x_0) - f_{\text{low}}\right)}{K},$$

i.e., the smallest value of a sequence cannot be larger than the mean of the same sequence. If a tolerance of $\|\nabla f(x_k)\| \leq \varepsilon$ is required, then

$$\frac{2L\left(f(x_0) - f_{\text{low}}\right)}{K} \leq \varepsilon^2,$$

and hence

$$K = \frac{2L\left(f(x_0) - f_{\text{low}}\right)}{\varepsilon^2}, \tag{17}$$

iterations are required to guarantee that $\|\nabla f(x_k)\| \leq \varepsilon$ is reached. In other words the complexity is $\frac{1}{\varepsilon^2}$. The main drawback is the restriction imposed by the third assumption, *i.e.*, the Lipschitz constant L must be available. If L is unavailable, then these convergence results do not hold.