CS 726: Setting the Stage

Jelena Diakonikolas

Fall 2023

A generic optimization problem is of the form

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}). \tag{P}$$

To give problem (P) a meaning, we need to specify (i) the *vector space* in which the vector of variables $\mathbf{x}$ lives, (ii) the *feasible set* $\mathcal{X}$, which is a subset of our vector space and determines what choices of vectors $\mathbf{x}$ we are allowed to output, and (iii) the *objective function* f, which maps each vector of variables $\mathbf{x}$ to a real value. Here, we think about the objective function as some cost or loss that we want to minimize. Alternatively, in some problems, the goal may be to maximize a reward $g(\mathbf{x})$. This setup is no different than (P), as we can take $f(\mathbf{x}) = -g(\mathbf{x})$.

Another thing that we need to consider here is what it means to "solve" (P). In general, unless we make strong assumptions about the objective function and the feasible set (the set $\mathcal{X}$ of "allowed" solutions $\mathbf{x}$), we cannot hope to find the exact vector $\mathbf{x}^*$ that minimizes f over $\mathcal{X}$. Most of the time, this is also not necessary, and we can settle for points that are "good enough." What that means will become clear once we introduce the definitions for different points we call *local minima* or *local solutions* at the end of this lecture, and once we discuss different notions of approximation for such points.

1 Vector Space: Where Optimization Variables Live

Our main objects of interest are optimization problems in $\mathbb{R}^d$ whose length is measured using one of the standard ℓ_p norms for $p \geq 1$. This means that the vector $\mathbf{x}$ from (P) is a d-dimensional column vector $[x_1, x_2, \ldots x_d]^T$ and that $\mathcal{X} \subseteq \mathbb{R}^d$. We will use $\langle \cdot, \cdot \rangle$ to denote any inner product on $\mathbb{R}^d$. Most frequently, it suffices to consider the standard inner product, which, given vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, is defined as:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{d} x_i y_i. \tag{1}$$

The simplest ℓ_p norm is the Euclidean norm, obtained for $p = 2$. It is denoted as $\| \cdot \|_2$ and defined by

$$\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\sum_{i=1}^{d} {x_i}^2}. \tag{2}$$

More generally, ℓ_p norms for $p \geq 1$ are defined by

$$\|\mathbf{x}\|_p = \Big(\sum_{i=1}^{d} |x_i|^p \Big)^{1/p}. \tag{3}$$

The most commonly used ℓ_p norms are the ℓ_2 (Euclidean) norm, the ℓ_1 norm, and the ℓ_∞ norm. The latter two are just

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i| \qquad \text{and} \qquad \|\mathbf{x}\|_\infty = \max_{1 \leq i \leq d} |x_i|. \tag{4}$$

Most of the time, it will suffice to think about Euclidean norms, though we will sometimes also consider other ℓ_p norms. Once we fix the ℓ_p norm we want to use, we have defined our normed vector space $(\mathbb{R}^d, \| \cdot \|_p)$. The pair

$(\mathbb{R}^d, \|\cdot\|_p)$ tells us that the optimization variables are vectors in $\mathbb{R}^d$ and that their length is measured using the ℓ_p norm. We will also call this space the *primal space*. This is it to contrast it with something called the *dual space*. By definition, the dual space is the space of all linear functions $\mathbf{z}$ acting on $\mathbf{x}$. In our case, these are just d-dimensional vectors, while the value of the linear function $\mathbf{z}$ is simply $\langle \mathbf{z}, \mathbf{x} \rangle$. An important property of the dual space is that the vectors in the dual space are measured with respect to the *dual norm*. The definition of a norm $\|\cdot\|_*$ that is dual to $\|\cdot\|$ is

$$\|\mathbf{z}\|_* = \sup_{\mathbf{x}\in\mathbb{R}^d,\ \|\mathbf{x}\|=1} \langle \mathbf{z}, \mathbf{x} \rangle . \tag{5}$$

An immediate consequence of the definition of the dual norm is the inequality that relates the inner product of two vectors $\mathbf{x}, \mathbf{z}$ and their primal and dual norms, respectively, as stated in the following proposition. This inequality can be seen as a generalization of the classical Cauchy-Schwarz inequality (called Hölder inequality for ℓ_p norms), and it will come in handy when we analyze the convergence of different algorithms.

Proposition 1.1. *For any two vectors* $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$:

$$|\langle \mathbf{z}, \mathbf{x} \rangle| \leq \|\mathbf{z}\|_* \cdot \|\mathbf{x}\|. \tag{2.1}$$

Proof. Fix any two vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$. The inequality from the statement of the proposition holds trivially if either $\mathbf{x} = \mathbf{0}$ or $\mathbf{z} = \mathbf{0}$, where $\mathbf{0}$ denotes the vector with all coordinates equal to zero, so assume $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{z} \neq \mathbf{0}$. Let $\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$. Then $\|\hat{\mathbf{x}}\| = 1$, and using the definition of the dual norm, we have:

$$\|\mathbf{z}\|_* = \sup_{\mathbf{x}\in\mathbb{R}^d,\ \|\mathbf{x}\|=1} \langle \mathbf{z}, \mathbf{x} \rangle \geq \langle \mathbf{z}, \hat{\mathbf{x}} \rangle = \frac{\langle \mathbf{z}, \mathbf{x} \rangle}{\|\mathbf{x}\|}. \tag{6}$$

Multiplying both sides of the last inequality by $\|\mathbf{x}\|$, we get $\langle \mathbf{z}, \mathbf{x} \rangle \leq \|\mathbf{z}\|_* \cdot \|\mathbf{x}\|$. To prove $-\langle \mathbf{z}, \mathbf{x} \rangle \leq \|\mathbf{z}\|_* \cdot \|\mathbf{x}\|$, we can use the same sequence of arguments with $\hat{\mathbf{x}}$ replaced by $-\hat{\mathbf{x}}$. □

A standard fact about ℓ_p norms is that the norm dual to $\|\cdot\|_p$ is the norm $\|\cdot\|_q$ where q is such that $\frac{1}{p} + \frac{1}{q} = 1$. Thus, the space $(\mathbb{R}^d, \|\cdot\|_q)$ is dual to $(\mathbb{R}^d, \|\cdot\|_p)$ whenever $\frac{1}{p} + \frac{1}{q} = 1$ and $p \geq 1$.

When we are solving an optimization problem, we are typically not told which norm to use. The performance of any algorithm that we will see in this class depends on certain properties of the feasible set $\mathcal{X}$ and the objective function f that are measured with respect to the norm that we select. If, for one choice of the norm, we get that these quantities are finite, then they will be finite with respect to every ℓ_p norm. This is because ℓ_p norms are related by the following (tight) inequalities

$$(\forall \mathbf{x} \in \mathbb{R}^d)(\forall p \geq 1)(\forall r > p): \qquad \|\mathbf{x}\|_r \leq \|\mathbf{x}\|_p \leq d^{\frac{1}{p}-\frac{1}{r}} \|\mathbf{x}\|_r. \tag{2.2}$$

However, the choice of the norm will make a big difference in terms of how fast the algorithm that we apply to our optimization problem will be. This is because the speed with which our algorithm converges will scale with the quantities that we measure with respect to the selected ℓ_p norm, and, thus, selecting a wrong norm may mean we end up with an algorithm that is slower by a factor polynomial in d. For large-scale problems that we are interested in here, this difference determines whether we would be able to use our algorithm in practice or not.

2 Feasible Set

The feasible set determines what choices of vectors $\mathbf{x}$ are acceptable as possible solutions to (P). If $\mathcal{X} \equiv \mathbb{R}^d$, we say that the problem (P) is *unconstrained*. Otherwise, we are dealing with a *constrained* optimization problem. This distinction between constrained and unconstrained problems is not always completely clear or even necessary, as we will see in the next section.

We typically use one of the following two ways of thinking about and expressing feasible sets: (i) as *abstract* sets $\mathcal{X}$ that we view as some geometrical bodies, such as, e.g., a ball, a box, or a polyhedron; and (ii) as *constraint-based* sets described by functional constraints of the form $f_i(\mathbf{x}) \leq 0$, for $i \in \{1, 2, \dots, m_1\}$ and $h_j(\mathbf{x}) = 0$, for $j \in \{1, 2, \dots, m_2\}$. Of course, the main difference here is in how we express or think about the feasible set; for example, if our feasible set is a unit Euclidean ball centered at the origin, then we could describe it by the constraint $\|\mathbf{x}\|_2^2 - 1 \leq 0$ or just think about it as an abstract geometric body. Observe here that the two types of constraints we wrote in (ii) are without loss of generality, as, for example, constraints of the form $\bar{f}_i(\mathbf{x}) \geq C$ can equivalently be written as $f_i \leq 0$ for $f_i(\mathbf{x}) = -\bar{f}_i(\mathbf{x}) + C$.

One of the most important properties of sets is convexity. Except for some special cases, we often need to make an assumption that the feasible set is convex to be able to guarantee that any algorithm we select converges to an acceptable solution (defined later), and that it does so at some reasonable speed.

Definition 2.1. We say that a set $\mathcal{X}$ is convex if for any pair of points from the set $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, the line between $\mathbf{x}$ and $\mathbf{y}$ is fully contained in $\mathcal{X}$. In other words, $\mathcal{X}$ is convex if for any pair of points $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and any $\alpha \in (0, 1)$:

$$(1-\alpha)\mathbf{x} + \alpha\mathbf{y} = \mathbf{x} + \alpha(\mathbf{y} - \mathbf{x}) \in \mathcal{X}. \tag{7}$$

If a set is not convex, we say that it is nonconvex.

One of the reasons why we do not consider completely arbitrary nonconvex sets is that we would very quickly run into computational bottlenecks. For example, a simple nonconvex constraint of the form $x_i(1 - x_i) = 0$ would force the i^{th} coordinate of $\mathbf{x}$ to only take one of the two possible values: 0 or 1. This would enable us to write optimization problems that are not too large (e.g., have size polynomial in d) but are not solvable with any known polynomial-time algorithms. In particular, we would be able to express problems that are from the computational class known as NP-complete, for which existing provably convergent algorithms can only handle very small problem instances, with d at the order of 10-100.

We will always assume that the feasible set is *closed.* Of course, to talk about closed (or open) sets, we need to define a topology. Since we will only be working in $\mathbb{R}^d$, whenever we talk about open or closed sets, we will be assuming the *standard topology on* $\mathbb{R}^d$.

Since we are only concerned with finite-dimensional optimization problems, Heine-Borel theorem implies that every closed and bounded set is compact. Compactness, in turn, guarantees that every continuous function defined and continuous on this set attains its extremal values on it, by the well-known Weierstrass's theorem. Thus, compact sets are of particular interest, as we do not need to worry about whether a minimizer of the function exists – as already noted, this is guaranteed by the theorem of Weierstrass. On the other hand, whenever we work with unconstrained optimization problems, to obtain any meaningful convergence guarantees, we will always assume that the objective function f is bounded below on its domain. Further, to avoid dealing with degenerate problems, we will always be assuming that the objective function f is well defined on $\mathcal{X}$ and its effective domain (the set of points on which it is smaller than $+\infty$) has a non-empty intersection with $\mathcal{X}$.

3 Basic Properties of the Objective Function

We will never work with completely arbitrary objective functions. The reason is that, similarly as for feasible sets, we would quickly run into computationally intractable problems.

Recall that the domain of a function is the set of points on which the function is well-defined. We will always assume that the domain $\mathcal{D}$ of the objective function f is a subset of $\mathbb{R}^d$. It will also be convenient to assume that f maps $\mathcal{D}$ to the extended real line $\bar{\mathbb{R}} \stackrel{\text{def}}{=} \mathbb{R} \cup \{-\infty, +\infty\}$, i.e., that f is *extended real valued.* Working with extended real valued functions is particularly useful, as we can define each function $f : \mathcal{D} \to \bar{\mathbb{R}}$ on the entire $\mathbb{R}^d$ by assigning it value $+\infty$ on all points that lie outside the domain $\mathcal{D}$. Note that we do not lose anything by doing so, as our goal is to minimize f. Once we have defined f on the entire $\mathbb{R}^d$, we only need to consider the *effective domain* of f, defined as the set of points on which f takes values lower than $+\infty$, i.e., $\mathrm{dom}(f) = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) < \infty\}$. In the sequel, whenever we talk about the domain of f, we will be referring to the effective domain of f, as this simplifies the discussion and is without loss of generality. We say that a function f is *proper* if it is not equal to $\pm\infty$ everywhere, i.e., if it takes at least some real values. Note that a necessary condition for f to be proper is that its effective domain is non-empty.

The field of linear and nonlinear optimization is often referred to as the "continuous" optimization, typically to contrast it with discrete, or combinatorial, optimization. Thus, it seems natural that the minimum assumption we would make about the objective function is that it be *continuous*. In most settings, this will indeed be the minimum assumption we make about the objective function. However, continuity can be relaxed slightly, if additional structure such as convexity (defined later in this section) and certain "simplicity" structure can be assumed. In those cases we may allow the objective function to be *lower semicontinuous*, defined as follows.

Definition 3.1. We say that a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is lower semicontinuous at $\bar{\mathbf{x}} \in \mathbb{R}^d$ if $\liminf_{\mathbf{x} \to \bar{\mathbf{x}}} f(\mathbf{x}) = f(\bar{\mathbf{x}})$ and lower semicontinuous on $\mathbb{R}^d$ if it is lower semicontinuous at every point $\bar{\mathbf{x}} \in \mathbb{R}^d$.

The main usefulness of lower semicontinuous functions (and likely the main reason they are even considered in nonlinear optimization) is that the indicator function of a closed set $\mathcal{X}$, defined as

$$I_{\mathcal{X}}(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in \mathcal{X} \\ +\infty, & \text{otherwise,} \end{cases} \tag{8}$$

is a lower semicontinuous function (verify this!). Thus, we can view constrained and unconstrained optimization problems in a unified way: minimizing f over a closed set $\mathcal{X}$ is the same as minimizing $f + I_{\mathcal{X}}$ over the entire $\mathbb{R}^d$!

Of course, while lower semicontinuous functions are useful for abstracting away constraints, we cannot work with arbitrary such functions. To see this, consider a function $f(\cdot) = I_{\mathcal{X}}(\cdot)$, where $\mathcal{X}$ is some closed set that is *not* known to us. If we access f by only querying its local information – e.g., for any candidate point $\mathbf{x} \in \mathbb{R}^d$, we get back the value of f at $\mathbf{x}$ and all of f's derivatives that exist, if any, at $\mathbf{x}$ (most optimization algorithms will work in this way) – then unless we are lucky and query a point inside $\mathcal{X}$, we get no clue in which direction to move. Thus, whenever we make an assumption that some function $f(\mathbf{x})$ is lower semicontinuous, we will also assume that it is convex (defined below) and that problems of the form $\min_{\mathbf{x} \in \mathbb{R}^d} g(\mathbf{x}) + f(\mathbf{x})$ are easily solvable for any function g from a given class of "simple" functions, such as linear or quadratic functions.

Most often, the minimum assumption we will make about a function in terms of its continuity is that it is Lipschitz-continuous.

Definition 3.2. We say that a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is Lipschitz-continuous on a set $\mathcal{X} \subseteq \mathbb{R}^d$ if there exists a constant $M < \infty$ such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ we have:

$$|f(\mathbf{x}) - f(\mathbf{y})| \le M \|\mathbf{x} - \mathbf{y}\|.$$

Observe that, as we are only considering finite-dimensional normed vector spaces, if a function is Lipschitz-continuous w.r.t. some norm $\|\cdot\|$ then it is Lipschitz-continuous w.r.t. every other norm. However, the Lipschitz constant M depends on the selected norm, and this is crucial to the performance of the algorithms in different setups.

Lipschitz-continuity of a function on its own is not sufficient for obtaining efficient optimization algorithms even if the function is bounded below; we always need to assume more structure such as, e.g., convexity. Thus, without convexity, we make additional assumptions about the objective function, such as that it is differentiable and its gradients are Lipschitz-continuous. Such functions are referred to as *smooth* functions.

Definition 3.3. We say that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is smooth w.r.t. a norm $\|\cdot\|$ on a set $\mathcal{X} \subseteq \mathbb{R}^d$ if there exists a constant $L < \infty$ such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \le L \|\mathbf{x} - \mathbf{y}\|,$$

where $\|\cdot\|, \|\cdot\|_*$ is a pair of dual norms.

For brevity, we will refer to functions that satisfy the conditions of Definition 3.3 as L-smooth functions.

There are also generalizations of smoothness that are sometimes used in optimization. One specific class of such functions comprises functions with Hölder-continuous gradients, sometimes referred to as weakly smooth functions.

Definition 3.4. We say that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is (κ, L)-weakly smooth for $\kappa \in [1, 2]$ w.r.t. a norm $\|\cdot\|$ on a set $\mathcal{X} \subseteq \mathbb{R}^d$ if there exists a constant $L < \infty$ such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \le L \|\mathbf{x} - \mathbf{y}\|^{\kappa - 1},$$

where $\|\cdot\|, \|\cdot\|_*$ is a pair of dual norms.

Note that when $\kappa = 2$, we recover the definition of smooth functions. When $\kappa = 1$, the weak smoothness condition reduces to bounded variation of the gradient, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \le L \|\mathbf{x} - \mathbf{y}\|^0 = L$, $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$. Such a condition implies that the function is Lipschitz-continuous (see Exercise 5). Some examples of (weakly) smooth functions are provided below.

Example 3.5. The following functions are weakly smooth:

1. $\frac{1}{p}\|\mathbf{x}\|_p^p$ for $p \in [1, 2]$ is $(p, 1)$-weakly smooth w.r.t. $\|\cdot\|_p$;
2. $\frac{1}{2}\|\mathbf{x}\|_p^2$ for $p \geq 2$ is $(p-1)$-smooth w.r.t. $\|\cdot\|_p$;
3. log-sum-exp (or soft max) function $\log\left(\sum_{i=1}^d \exp(x_i)\right)$ is 1-smooth w.r.t. $\|\cdot\|_\infty$.

Functions that are convex are of particular interest in nonlinear optimization, due to their rich structure that enables the design of efficient optimization algorithms.

Definition 3.6. We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex, if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and any $\alpha \in (0, 1)$:

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$

If a function is not convex, we say that it is nonconvex.

It is also possible to equivalently define convex functions via the definition of convex sets, using the notion of an epigraph, as follows.

Lemma 3.7. *Let $f : \mathbb{R}^d \to \mathbb{R}$. Then f is convex if and only if its epigraph, defined as*

$$\mathrm{epi}(f) = \{(\mathbf{x}, a) : \mathbf{x} \in \mathbb{R}^d, a \in \mathbb{R}, f(\mathbf{x}) \leq a\},$$

is convex.

The proof is left as an exercise.

Finally, we will frequently use an equivalent definition of convex functions that are further differentiable, as summarized in the following lemma.

Lemma 3.8. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then, f is convex if and only if $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle. \tag{9}$$

The proof is a special case of the proof of Lemma 3.14, provided later in this lecture.

3.1 Continuous Differentiability and Taylor Theorem

Very often, we will work with functions that are continuously differentiable. When we say that a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is continuously differentiable, what we mean is that its gradient exists at every point in $\mathbb{R}^d$ and is continuous. In particular,

$$(\forall \epsilon > 0)(\exists \delta > 0)(\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d) : \|\mathbf{x} - \mathbf{y}\| \leq \delta \quad \Rightarrow \quad \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq \epsilon. \tag{10}$$

As a side note, we can also talk about functions that are continuously differentiable on $\mathcal{X} \subset \mathbb{R}^d$ by restricting $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ in (10). Note that a continuously differentiable function must be continuous (otherwise, it would not be differentiable). Observe further from (10) and Definition 3.3 that smooth functions must be continuously differentiable (smoothness is a stronger property).

Since our focus is on finite dimensions, the choice of the norms in (10) is not important, as all norms in $\mathbb{R}^d$ are within a constant factor of each other, where the constant may depend on the dimension (this is known as the equivalence of norms; see (2.2) for inequalities relating ℓ_p norms). For concreteness, one may choose the ℓ_2 norm, $\|\cdot\| = \|\cdot\|_2$.

We will say that a function f is twice continuously differentiable, if its second order derivative $\nabla^2 f$, known as the Hessian matrix (or just Hessian) exists at every point in $\mathbb{R}^d$ and is continuous. In the (ϵ, δ) notation, this can be expressed as

$$(\forall \epsilon > 0)(\exists \delta > 0)(\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d) : \|\mathbf{x} - \mathbf{y}\|_2 \leq \delta \quad \Rightarrow \quad \|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq \epsilon. \tag{11}$$

Observe that in (11) we are applying the ℓ_2 norm to a matrix. When applied to a symmetric matrix $\mathbf{A}$, the Euclidean norm is also known as the operator norm and is defined by $\|\mathbf{A}\|_2 = \sup_{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2$. In this case, $\|\mathbf{A}\|_2$ is equal to the maximum absolute eigenvalue of $\mathbf{A}$. It is also possible to use other norms for measuring the distance between the points $\mathbf{x}, \mathbf{y}$ and the corresponding Hessian matrices $\nabla^2 f(\mathbf{x}), \nabla^2 f(\mathbf{y})$, though this is not important for the definition of continuity due to the equivalence of norms, same as for the gradients.

It is also possible to talk about functions that are k-times continuously differentiable for k larger than 2. In this case, the k^{th} order derivative of f at $\mathbf{x} \in \mathbb{R}^d$ is a k^{th} order tensor and we need an appropriate definition of a norm to measure distances between tensors, generalizing further from matrix norms. However, for our purposes, it will suffice to only consider functions that are up to two times continuously differentiable.

A central result that leads to different characterizations of functions that are once or twice continuously differentiable is a multivariate version of Taylor theorem stated below.

Theorem 3.9. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function. Then, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have:*

(i) $f(\mathbf{y}) = f(\mathbf{x}) + \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle \, \mathrm{d}t$;

(ii) There exists $t \in (0, 1)$ such that $f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle$.

If f is twice continuously differentiable, then we further have:

(iii) $\nabla f(\mathbf{y}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) \mathrm{d}t$;

(iv) There exists $t \in (0, 1)$ such that $f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \left\langle \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle$.

Part (*ii*) of Theorem 3.9 is often referred to as the Mean Value Theorem. You may notice that we did not state such an identity for vectors but instead had an appropriate identity for f in part (*iv*) of the theorem. You may be wondering whether we could state the Mean Value Theorem for the gradients (as we did for the function value). The answer is **no** and you should remember this, as it is a common mistake.

Taylor theorem allows us to obtain different characterizations of smooth functions that turn out to be useful for the analysis of optimization algorithms. A particularly useful result is the quadratic upper approximation of a smooth function, as summarized in the following lemma.

Lemma 3.10. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth function w.r.t. a norm $\|\cdot\|$. Then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, both of the following two inequalities hold*

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

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{13}$$

Proof. Applying Theorem 3.9(*i*), we have that

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| = \Big| \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \, \mathrm{d}t \Big|.$$

Applying Jensen's inequality, as the absolute value is a convex function, we have

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \int_0^1 |\langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \mathrm{d}t. \tag{14}$$

By Proposition 1.1, we have that $|\langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|_* \|\mathbf{y} - \mathbf{x}\|$. Further, as f is L-smooth, we also have $\|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|_* \leq Lt\|\mathbf{y} - \mathbf{x}\|$. Combining with (14), it follows that

$$\begin{aligned}
|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| &\leq \int_0^1 tL\|\mathbf{y} - \mathbf{x}\|^2 \mathrm{d}t \\
&= L\|\mathbf{y} - \mathbf{x}\|^2 \int_0^1 t\mathrm{d}t \\
&= \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2.
\end{aligned}$$

To complete the proof of (12), it remains to rearrange the last inequality and use the definition of the absolute value. □

It is further possible to characterize smooth functions by looking at their Hessian, as shown in the following two lemmas.

Lemma 3.11. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice continuously differentiable function. If for all $\mathbf{x} \in \mathbb{R}^d$ it holds $\|\nabla^2 f(\mathbf{x})\|_{p,q} = \sup_{\mathbf{y}:\|\mathbf{y}\|_p=1} \|\nabla^2 f(\mathbf{x})\mathbf{y}\|_q \leq L < \infty$, where $\frac{1}{p} + \frac{1}{q} = 1, p \geq 1$, then f is L-smooth w.r.t. $\|\cdot\|_p$.*

Proof. Using Theorem 3.9(*iii*), we have

$$\begin{aligned} \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_q &= \left\| \int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) \mathrm{d}t \right\|_q \\ &\leq \int_0^1 \left\| \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) \right\|_q \mathrm{d}t, \end{aligned} \tag{15}$$

where we have used Jensen's inequality, which holds as $\|\cdot\|_q$ is a convex function.

Using the same argument as in the proof of Proposition 1.1, we have that

$$\begin{aligned} \left\| \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) \right\|_q &\leq \|\nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))\|_{p,q} \|\mathbf{y} - \mathbf{x}\|_p \\ &\leq L\|\mathbf{y} - \mathbf{x}\|_p, \end{aligned}$$

where the second inequality is by the Lemma assumption. Combining with (15), we have

$$\begin{aligned} \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_q &\leq \int_0^1 L \|\mathbf{y} - \mathbf{x}\|_p \mathrm{d}t \\ &= L\|\mathbf{y} - \mathbf{x}\|_p, \end{aligned}$$

and thus f is L-smooth w.r.t. $\|\cdot\|_p$, by definition. □

Lemma 3.12. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice continuously differentiable function. Then f is L-smooth w.r.t. the Euclidean norm $\|\cdot\|_2$ if and only if $-L\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.*

Proof. The "if" part of the statement follows by Lemma 3.11, as the condition $-L\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$ is equivalent to $\|\nabla^2 f(\mathbf{x})\|_2 = \|\nabla^2 f(\mathbf{x})\|_{2,2} \leq L$.

For the "only if" part, assume that f is L-smooth w.r.t. $\|\cdot\|_2$. Fix any $\mathbf{x} \in \mathbb{R}^d$ and let $\mathbf{y} = \mathbf{x} + \alpha\mathbf{p}$, where $\alpha > 0$ and $\mathbf{p} \in \mathbb{R}^d$, $\|\mathbf{p}\|_2 = 1$. Applying Theorem 3.9(*iv*), there exists $t \in (0, 1)$ such that

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = \frac{\alpha^2}{2} \left\langle \nabla^2 f(\mathbf{x} + t\alpha\mathbf{p})\mathbf{p}, \mathbf{p} \right\rangle. \tag{16}$$

Further, by Lemma 3.10, we have that

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \frac{L\alpha^2}{2} \|\mathbf{p}\|_2^2. \tag{17}$$

Combining (16) and (17), we have

$$|\left\langle \nabla^2 f(\mathbf{x} + t\alpha\mathbf{p})\mathbf{p}, \mathbf{p} \right\rangle| \leq L\|\mathbf{p}\|_2^2.$$

Letting $\alpha \downarrow 0$ and using continuity of $\nabla^2 f$, we get $|\left\langle \nabla^2 f(\mathbf{x})\mathbf{p}, \mathbf{p} \right\rangle| \leq L\|\mathbf{p}\|_2^2$. Since $\mathbf{p}$ was an arbitrary unit vector, we conclude that the maximum absolute eigenvalue of $\nabla^2 f(\mathbf{x})$ is bounded by L, or, equivalently, $-L\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$. □

3.2 Strong and Uniform Convexity

So far we have focused on what we can think of as regularity conditions for objective functions f, which intuitively bound above how quickly a function can change. Without any additional assumptions, such properties are necessary in ensuring optimization problems we deal with remain tractable. One could plausibly ask whether bounding the function growth below could be useful in any way. On its own, of course, given what we have discussed so far, we cannot expect that bounding the growth below would be useful. However, if we are able to bound the growth both below and above, this gives us a good sense of how fast the function could be changing and how far we can explore. This, in turn, becomes extremely useful for arguing about fast convergence of algorithms.

One of the most commonly used such properties in the design and analysis of gradient-based algorithms is known as strong convexity.

Definition 3.13. A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is said to be strongly convex with modulus $\mu > 0$ w.r.t. the norm $\|\cdot\|$ if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and any $\alpha \in (0, 1)$

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \le (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

For brevity, we will often refer to functions that are strongly convex with modulus μ as just being μ-strongly convex. When the context is clear, we will omit referring to the specific norm w.r.t. which a function is strongly convex. Observe that, comparing to Definition 3.6, we can think of convex functions as being strongly convex with modulus zero (although $\mu = 0$ is not allowed in the definition of strongly convex functions; we can think of this as the limiting case as $\mu \downarrow 0$).

Strong convexity ensures that, informally speaking, a function is at least as curved as a quadratic. For functions that are continuously differentiable, one such characterization is provided in the following lemma.

Lemma 3.14. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be a continuously differentiable function. Then, f is strongly convex with modulus $\mu > 0$ w.r.t. a norm $\|\cdot\|$ if and only if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ it holds*

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{18}$$

Proof. Suppose first that f is μ-strongly convex. Given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have that for any $\alpha \in (0, 1)$,

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \le (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Rearranging,

$$f(\mathbf{y}) - f(\mathbf{x}) \ge \frac{f(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\alpha} + (1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

To complete the proof of this part, it remains to take $\alpha \downarrow 0$.

Now suppose f satisfies (18). Fix any $\mathbf{x}, \mathbf{y}$ and any $\alpha \in (0, 1)$ and let $\mathbf{z} = (1-\alpha)\mathbf{x} + \alpha\mathbf{y}$. Then, applying (18) for $\mathbf{x}, \mathbf{z}$ and $\mathbf{y}, \mathbf{z}$, we get:

$$f(\mathbf{x}) \ge f(\mathbf{z}) + \alpha\langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{y}\rangle + \frac{\mu}{2}\alpha^2\|\mathbf{y} - \mathbf{x}\|^2, \tag{19}$$

$$f(\mathbf{y}) \ge f(\mathbf{z}) + (1-\alpha)\langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}(1-\alpha)^2\|\mathbf{y} - \mathbf{x}\|^2. \tag{20}$$

Multiplying (19) by $1 - \alpha$, multiplying (20) by α, summing them up and simplifying, we get

$$(1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) \ge f(\mathbf{z}) + \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2,$$

which is precisely the definition of a μ-strongly convex function. □

Using Taylor theorem (Theorem 3.9), it is further possible to show that twice continuously differentiable functions that are μ-strongly convex w.r.t. the Euclidean norm also satisfy $\nabla^2 f(\mathbf{x}) \succeq \mu\mathbf{I}, \forall\mathbf{x}$. The converse is also true: functions that are twice continuously differentiable and satisfy $\nabla^2 f(\mathbf{x}) \succeq \mu\mathbf{I}, \forall\mathbf{x}$ are μ-strongly convex w.r.t. the Euclidean norm.

It is natural to ask whether bounding the growth of a function by a quadratic is the only useful example for optimization. It turns out that there is a generalization of strong convexity, known as uniform convexity, where a function is bounded below by a different, higher-order polynomial. This concept is introduced in the following definition.

Definition 3.15. Given $p \ge 2$, a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is said to be p-uniformly convex with modulus $\mu > 0$ w.r.t. the norm $\|\cdot\|$ if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and any $\alpha \in (0, 1)$

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \le (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \alpha(1-\alpha)\frac{\mu}{p}\|\mathbf{y} - \mathbf{x}\|^p.$$

For brevity, we will often refer to such functions as being (p, μ)-uniformly convex. Observe that, according to Definition 3.13, 2-uniformly convex functions are strongly convex.

Similar to Lemma 3.14, it is not hard to argue that p-uniformly convex differentiable functions satisfy, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{p} \|\mathbf{y} - \mathbf{x}\|^p. \tag{21}$$

Note further that it is possible to define functions that are strongly or uniformly convex on a bounded set $\mathcal{X} \subset \mathbb{R}^d$ by restricting $\mathbf{x}$ and $\mathbf{y}$ from the definitions to belong to $\mathcal{X}$.

Some common examples of functions that are strongly or uniformly convex are provided in the following.

Example 3.16. The following functions are either strongly or uniformly convex.

1. $\frac{1}{2}\|\mathbf{x}\|_p^2$ for $p \in (1, 2]$ is $(p-1)$-strongly convex w.r.t. $\|\cdot\|_p$ on $\mathbb{R}^d$;
2. $\frac{1}{p}\|\mathbf{x}\|_p^p$ for $p \geq 2$ is p-uniformly convex with modulus 1 and w.r.t. $\|\cdot\|_p$ on $\mathbb{R}^d$;
3. Negative entropy $\sum_{i=1}^d x_i \log(x_i)$ is 1-strongly convex w.r.t. $\|\cdot\|_1$ on the probability simplex $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : x_i \geq 0, 1 \leq i \leq d, \sum_{i=1}^d x_i = 1\}$.

3.3 Local Error Bounds

Properties such as strong and uniform convexity rarely hold for functions we commonly encounter in optimization, especially since they are even stronger properties than convexity. There are relaxations of these properties that hold even for some nonconvex functions while enabling the development of fast optimization algorithms, as we will see in later lectures. These properties are known as local error bounds. They hold for almost all interesting convex optimization problems and even some nonconvex problems (either locally or globally). The first set of such conditions are known as "sharpness" conditions, Łojasiewicz inequality or Hölderian error bounds.

Definition 3.17. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy the sharpness property or satisfy Łojasiewicz inequality on a set $\mathcal{X} \in \mathbb{R}^d$ if the set of minimizers of f, $\mathcal{X}^* = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$, is non-empty and there exist constants $\mu > 0, r > 0$ such that $\forall \mathbf{x} \in \mathcal{X}$

$$f(\mathbf{x}) - \min_{\mathbf{y}} f(\mathbf{y}) \geq \frac{\mu}{r} \operatorname{dist}(\mathbf{x}, \mathcal{X}^*)^r, \tag{22}$$

where $\operatorname{dist}(\mathbf{x}, \mathcal{X}^*) = \min_{\mathbf{x}^* \in \mathcal{X}^*} \|\mathbf{x} - \mathbf{x}^*\|$.

This property is usually stated using the Euclidean norm $\|\cdot\| = \|\cdot\|_2$, however it is possible to define it using any norm. We will refer to functions that satisfy the conditions of Definition 3.17 as being (r, μ)-sharp. It is not hard to verify that μ-strongly convex functions are $(2, \mu)$-sharp w.r.t. the same norm, while (p, μ)-uniformly convex functions are (p, μ)-sharp w.r.t. the same norm.

A related set of local error bound conditions known as the gradient dominated property, Łojasiewicz gradient inequality, or the Polyak-Łojasiewicz (PŁ) condition are defined below. Such conditions can be stated even for functions that are not differentiable but satisfy a weaker property of being subdifferentiable (we will encounter and properly define such functions in later lectures). For simplicity however, below we define the gradient dominated property for differentiable functions.

Definition 3.18. A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy the gradient dominated property or Łojasiewicz gradient inequality on a set $\mathcal{X} \in \mathbb{R}^d$ if the set of minimizers of f, $\mathcal{X}^* = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ is non-empty and there exist constants $\mu > 0, r > 0$ such that $\forall \mathbf{x} \in \mathcal{X}$

$$f(\mathbf{x}) - \min_{\mathbf{y}} f(\mathbf{y}) \leq \frac{\mu}{r} \|\nabla f(\mathbf{x})\|^r. \tag{23}$$

4 Local and Global Solutions

We finish this lecture by formally defining what it means to "solve" an optimization problem (P). Throughout this section, we assume that $\mathcal{X}$ has a nonempty interior, so that the statements we make are nontrivial.

The weakest notion of a solution that we use in this class is what we will refer to as the stationary solution.

Definition 4.1. Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be a differentiable function and let $\mathcal{X}$ be closed and convex. We say that $\mathbf{x} \in \mathcal{X}$ is a stationary point for (P) if for all $\mathbf{y} \in \mathcal{X}$,

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0. \tag{24}$$

It should be immediately clear that if there exists $\mathbf{x} \in \mathcal{X}$ such that $\nabla f(\mathbf{x}) = \mathbf{0}$, then $\mathbf{x}$ is a stationary point for (P). For unconstrained problems with objective functions that are bounded below, we will often seek points $\mathbf{x}$ that have small gradient norm $\|\nabla f(\mathbf{x})\|_*$.

Stationarity alone does not guarantee that a point minimizes the objective function on the set $\mathcal{X}$, even locally. As simple examples, consider univariate functions $f_1(x) = x^2$, $f_2(x) = -x^2$, and $f_3(x) = x^3$, for which the point $x = 0$ is stationary (verify this!). For f_1, $x = 0$ is a local minimum, for f_2 it is a local maximum, while for f_3 it is neither a local minimum nor a local maximum.

We now formally define what it means for a point to be a local or a global solution.

Definition 4.2. Given a problem (P) and a point $\mathbf{x}^* \in \mathcal{X}$, we say that

- $\mathbf{x}^*$ is a local solution to (P) or a local optimum for (P) if there exists a neighborhood $\mathcal{N}_{\mathbf{x}^*}$ of $\mathbf{x}^*$ such that for all $\mathbf{x} \in \mathcal{N}_{\mathbf{x}^*} \cap \mathcal{X}$, $f(\mathbf{x}) \geq f(\mathbf{x}^*)$.
- $\mathbf{x}^*$ is a global solution to (P) or a global optimum for (P) if for all $\mathbf{x} \in \mathcal{X}$, $f(\mathbf{x}) \geq f(\mathbf{x}^*)$.

When (P) is unconstrained (that is, when $\mathcal{X} \equiv \mathbb{R}^d$), we will refer to local/global solutions to (P) as local/global minimizers of f. A local/global minimizer $\mathbf{x}^*$ of f will be called strict if the inequality $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ from its definition is strict for any $\mathbf{x} \neq \mathbf{x}^*$.

In the rest of the section, we provide some necessary and sufficient conditions for a point to be a local or global solution.

4.1 Necessary and Sufficient Conditions for Possibly Nonconvex Objective Functions

The first optimality condition shows that every local solution to (P) must be a stationary point for (P). These statements will apply to both constrained and unconstrained problems.

Theorem 4.3 (First-order Necessary Optimality Conditions)**.** *Given* (P) *where f is continuously differentiable and $\mathcal{X}$ is closed and convex: if $\mathbf{x}^* \in \mathcal{X}$ is a global solution to* (P)*, then $\mathbf{x}$ must be a stationary point for* (P)*;*

If $\mathbf{x}^* \in \mathcal{X}$ is a local solution to (P)*, then there exists a neighborhood $\mathcal{N}$ of $\mathbf{x}^*$ such that for all $\mathbf{y} \in \mathcal{N} \cap \mathcal{X}$, $\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geq 0$.*

Proof. For the first claim, suppose, for the purpose of contradiction (f.p.o.c.) that $\mathbf{x}^* \in \mathcal{X}$ is a global solution to (P), but there exists some point $\mathbf{y} \in \mathcal{X}$ such that $\langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle < 0$. Since f is continuously differentiable, there must exists a sufficiently small $\alpha \in (0, 1)$ such that $\langle \nabla f(\mathbf{x}^* + \alpha'(\mathbf{y} - \mathbf{x}^*)), \mathbf{y} - \mathbf{x}^* \rangle < 0$ for all $\alpha' \in (0, \alpha]$. Let $\mathbf{z} = (1 - \alpha)\mathbf{x}^* + \alpha\mathbf{y}$. Then, applying Theorem 3.9(*ii*), there exists $t \in (0, 1)$ such that

$$f(\mathbf{z}) - f(\mathbf{x}^*) = \langle \nabla f(\mathbf{x}^* + t(\mathbf{z} - \mathbf{x}^*)), \mathbf{z} - \mathbf{x}^* \rangle = \alpha \langle \nabla f(\mathbf{x}^* + \alpha t(\mathbf{y} - \mathbf{x}^*)), \mathbf{y} - \mathbf{x}^* \rangle < 0, \tag{25}$$

as $t\alpha < \alpha$ and $t > 0$, $\alpha > 0$. Thus, $\mathbf{x}^*$ is not a global solution and we reach a contradiction.

The second part follows from the first, by restricting (P) to the set $\mathcal{N}_{\mathbf{x}^*} \cap \mathcal{X}$. □

In addition to first-order necessary and sufficient conditions for local solutions, there exist second-order conditions for problem with twice continuously differentiable objective functions. Such conditions are complicated to state for constrained problems without introducing more background, however they are simple to state for unconstrained problems.

Theorem 4.4 (Second-Order Necessary and Sufficient Conditions for Unconstrained Problems)**.** *Given* (P)*, where f is twice continuously differentiable and $\mathcal{X} \equiv \mathbb{R}^d$:*

- *If $\mathbf{x}^* \in \mathbb{R}^d$ is a local solution to* (P) *(local minimizer of f), then $\|\nabla f(\mathbf{x}^*)\|_2 = 0$ and $\nabla^2 f(\mathbf{x}^*) \succeq 0$;*
- *If there exists $\mathbf{x}^* \in \mathbb{R}^d$ such that $\|\nabla f(\mathbf{x}^*)\|_2 = 0$ and $\nabla^2 f(\mathbf{x}^*) \succ 0$, then $\mathbf{x}^*$ is a strict local minimizer of f.*

The proof uses similar ideas to the proof of Theorem 4.3 and is left as an exercise.

4.2 Optimality Conditions for Convex Problems

Convex problems are special in the sense that all local minima/solutions are global and they are easy to characterize using necessary conditions from Theorems 4.3 and 4.4, which turn out to be also sufficient. This is formally shown in the following theorem.

Theorem 4.5 (Optimality Conditions for Convex Problems)**.** *Given a problem* (P) *where f is convex and $\mathcal{X}$ is closed and convex, the following statements all hold:*

(i) *Every local solution to* (P) *is also global;*

(ii) *The set of solutions to* (P) *is convex;*

(iii) *If f is differentiable then $\mathbf{x}^*$ is a global solution to* (P) *if and only if $\mathbf{x}^*$ satisfies the stationarity condition* (24) *for* (P)*.*

Proof. For Part (*i*), suppose that $\mathbf{x}^*$ is a local but not a global solution to (P). Let $\mathcal{N}_{\mathbf{x}^*}$ be any neighborhood of $\mathbf{x}^*$ such that for $\mathbf{x} \in \mathcal{N}_{\mathbf{x}^*} \cap \mathcal{X}$, $f(\mathbf{x}) \geq f(\mathbf{x}^*)$. (Such a neighborhood must exist as $\mathbf{x}^*$ is a local solution.) Since $\mathbf{x}^*$ is not a global solution, there must exist $\mathbf{y} \in \mathcal{X}$ such that $f(\mathbf{y}) < f(\mathbf{x}^*)$. By convexity of f, for any $\alpha \in (0, 1)$,

$$f((1-\alpha)\mathbf{x}^* + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}^*) + \alpha f(\mathbf{y}) < f(\mathbf{x}^*).$$

By making $\alpha > 0$ sufficiently small, we obtain a point $\hat{\mathbf{x}} = (1-\alpha)\mathbf{x}^* + \alpha\mathbf{y} \in \mathcal{N}_{\mathbf{x}^*} \cap \mathcal{X}$ such that $f(\hat{\mathbf{x}}) < f(\mathbf{x}^*)$. This contradicts the assumption that $\mathbf{x}^*$ was a local solution, thus (*i*) holds.

To prove Part (*ii*), we need to show that if $\mathbf{x}^*, \hat{\mathbf{x}}^*$ are any two solutions to (P), then for any $\alpha \in (0, 1)$, $(1-\alpha)\mathbf{x}^* + \alpha\hat{\mathbf{x}}^*$ is also a solution to (P). This follows simply by convexity, as for any $\alpha \in (0, 1)$, we have

$$f((1-\alpha)\mathbf{x}^* + \alpha\hat{\mathbf{x}}^*) \leq (1-\alpha)f(\mathbf{x}^*) + \alpha f(\hat{\mathbf{x}}^*) = f(\mathbf{x}^*) = f(\hat{\mathbf{x}}^*).$$

For Part (*iii*), the "only if" direction follows from the first part of Theorem 4.3. For the "if" part of the statement, assume that $\mathbf{x}^*$ is a stationary point for (P). Then, by convexity and Lemma 3.8, we have that $\forall \mathbf{y} \in \mathcal{X}$,

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geq f(\mathbf{x}^*).$$

Thus, $\mathbf{x}^*$ is a global solution, by definition. □

We can say even more about convex problems with strongly convex objective functions, as shown in the following theorem.

Theorem 4.6 (Optimality Conditions for Convex Problems with Strongly Convex Objectives)**.** *Given a problem* (P)*, if f is strongly convex with modulus $\mu > 0$ and continuous on its domain, and $\mathcal{X}$ is closed, convex, and has a non-empty intersection with the effective domain of f, then a solution to* (P) *is attained and unique.*

Proof. Let $\mathbf{x}_0 \in \mathcal{X} \cap \mathrm{dom}(f)$ be an arbitrary point. Then $f(\mathbf{x}_0) < \infty$. If there is only one such point, then it must be the unique solution to (P) and we would be done, so assume that this is not the case.

Consider the set $S = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$, which is closed and non-empty. Further, by strong convexity of f, we have that if $\mathbf{x}, \mathbf{y} \in S$, then for any $\alpha \in (0, 1)$,

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq f(\mathbf{x}_0) - \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{26}$$

Thus, $(1-\alpha)\mathbf{x} + \alpha\mathbf{y} \in S$, which implies that S is convex.

By assumption, f is continuous on its domain, thus it is continuous at $\mathbf{x}_0$. We now argue that f must be bounded below. Fix an arbitrary $\mathbf{y} \in S$. By strong convexity, for any $\alpha \in (0, 1)$,

$$f(\mathbf{y}) \geq f(\mathbf{x}_0) + \frac{f((1-\alpha)\mathbf{x}_0 + \alpha\mathbf{y}) - f(\mathbf{x}_0)}{\alpha} + (1-\alpha)\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}_0\|^2.$$

Because f is continuous at $\mathbf{x}_0$, there exists a sufficiently small $\alpha \in (0, 1/2]$ and a constant $M < \infty$ such that

$$f(\mathbf{y}) \geq f(\mathbf{x}_0) - M\|\mathbf{y} - \mathbf{x}_0\| + \frac{\mu}{4}\|\mathbf{y} - \mathbf{x}_0\|^2.$$

(Verify this using the (ϵ, δ) definition of continuity.) This implies that $f(\mathbf{y})$ is bounded below, as the function on the right-hand side of the last inequality is bounded below (by $f(\mathbf{x}_0) - \frac{M^2}{\mu}$).

Further, rearranging (26) and taking $\alpha = \frac{1}{2}$, we have

$$\|\mathbf{y} - \mathbf{x}\|^2 \leq \frac{8}{\mu}\Big(f(\mathbf{x}_0) - f(\frac{1}{2}\mathbf{x} + \frac{1}{2}\mathbf{y})\Big) \leq \Big(\frac{8}{\mu} - 1\Big)f(\mathbf{x}_0) + \frac{M^2}{\mu} < \infty,$$

which implies that S is bounded. By Heine-Borel theorem, S is compact. By the Extreme Value Theorem of Weierstrass, f attains its minimum value on S. Thus a solution to (P) is attained. To complete the proof, it remains to argue that it is unique. Suppose f.p.o.c. that there were two solutions $\mathbf{x}^*, \hat{\mathbf{x}}^*$, $\mathbf{x}^* \neq \hat{\mathbf{x}}^*$ to (P). Then for any $\alpha > 0$,

$$f((1-\alpha)\mathbf{x}^* + \alpha\hat{\mathbf{x}}^*) \leq (1-\alpha)f(\mathbf{x}^*) + \alpha f(\hat{\mathbf{x}}^*) - \alpha(1-\alpha)\frac{\mu}{2}\|\mathbf{x}^* - \hat{\mathbf{x}}^*\|^2 < f(\mathbf{x}^*).$$

Since $(1-\alpha)\mathbf{x}^* + \alpha\hat{\mathbf{x}}^* \in \mathcal{X}$ (by convexity of $\mathcal{X}$), this contradicts the assumption that $\mathbf{x}^*, \hat{\mathbf{x}}^*$ were (optimal) solutions to (P), completing the proof. □

Exercises

1. For what kinds of vectors are the inequalities in (2.2) tight?

2. Can a function that is weakly smooth on $\mathbb{R}^d$ take infinite values? Argue why or why not.

3. Prove that if a function $f : \mathbb{R}^d \to \mathbb{R}$ is (κ, L)-weakly smooth w.r.t. a norm $\|\cdot\|$, then it satisfies $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{1}{\kappa}\|\mathbf{y} - \mathbf{x}\|^\kappa.$$

4. Prove that if a function $f : \mathbb{R}^d \to \mathbb{R}$ is $(1, L)$-weakly smooth and there exists a point $\mathbf{x}^*$ such that $\nabla f(\mathbf{x}^*) = 0$, then f is also Lipschitz continuous with constant L.

5. Prove that if a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous with constant L, then it is also $(1, 2L)$-weakly smooth.

6. Prove that log-sum-exp function $\log(\sum_{i=1}^d \exp(x_i))$ is 1-smooth w.r.t. $\|\cdot\|_\infty$.

7. Prove Lemma 3.7.

8. Consider (P) where f is L-smooth w.r.t. a norm $\|\cdot\|$ (but not necessarily convex), $\mathcal{X} \equiv \mathbb{R}^d$, and $\mathbf{x}^* \in \mathrm{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$. Use Lemma 3.10 to prove that $\forall \mathbf{x} \in \mathbb{R}^d$,

$$\frac{1}{2L}\|\nabla f(\mathbf{x})\|_*^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*).$$

9. Prove that a function is convex and L-smooth w.r.t a norm $\|\cdot\|$ if and only if $\forall \mathbf{x}, \mathbf{y}$,

$$\frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle.$$

Hint: For the "only if" direction, consider a fixed $\mathbf{x}$ and the function $h_{\mathbf{x}}(\mathbf{y}) = f(\mathbf{y}) - \langle \nabla f(\mathbf{x}), \mathbf{y}\rangle$. Argue that this function is convex, L-smooth, and minimized by $\mathbf{x}$.

10. Prove that the sum of a convex and a μ-strongly convex function is μ-strongly convex.

11. Prove that a function f is μ-strongly convex w.r.t. the ℓ_2-norm if and only if $f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|_2^2$ is convex.

12. Prove that a twice continuously differentiable function is convex if and only if $\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$.

13. Prove that a function f is μ-strongly convex w.r.t. $\|\cdot\|$ if and only if $\forall \mathbf{x}, \mathbf{y}$,

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq \mu\|\mathbf{x} - \mathbf{y}\|^2.$$

14. Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable and μ-strongly convex w.r.t the Euclidean norm $\|\cdot\|_2$. Prove that f must satisfy $\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}$, $\forall \mathbf{x}$.

15. Prove that if $f_1 : \mathbb{R}^d \to \bar{R}$ is μ_1-strongly convex and $f_2 : \mathbb{R}^d \to \bar{R}$ is μ_2-strongly convex, then $f_1 + f_2$ is $(\mu_1 + \mu_2)$-strongly convex.

16. Prove that if a function is convex and (r, μ)-sharp on a set $\mathcal{X}$, then it also satisfies

$$(\forall \mathbf{x} \in \mathcal{X})(\forall \mathbf{x}^* \in \mathcal{X}^*) \quad \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \geq \frac{\mu}{r} \operatorname{dist}(\mathbf{x}, \mathcal{X}^*)^r.$$

17. Prove that if a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is μ-strongly convex w.r.t. the Euclidean norm $\|\cdot\|_2$ and minimized by some $\mathbf{x}^* \in \mathbb{R}^d$, then f satisfies the gradient dominated property with constants $1/\mu$ and $r = 2$.

18. Prove Theorem 4.4.

19. Prove that if $\mathbf{x}^*$ solves (P), where $\mathcal{X}$ is closed and convex and $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex w.r.t. a norm $\|\cdot\|$, then it holds $\forall \mathbf{x} \in \mathcal{X}$,

$$\|\mathbf{x} - \mathbf{x}^*\| \leq \sqrt{\frac{2}{\mu}(f(\mathbf{x}) - f(\mathbf{x}^*))}.$$