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

Lecture 6: Optimality Conditions

*Instructor:*[1] *Matt Gormley* *September 13, 2023*

# 6.1 Smooth, Strongly Convex and Strictly Convex Functions

## 6.1.1 Smoothness

**Recall from a previous lecture**

In optimization smoothness has a very particular meaning (it has a slightly different meaning in stats, and other areas of math).

**Definition 6.1** ($\beta$-Smooth)**.** *A function $f$ is $\beta$-smooth, if its gradient is Lipschitz continuous with parameter $\beta$, i.e. for any $x, y \in dom(f)$,*

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

There are several useful implications of smoothness that we will briefly discuss now:

1. Another implication of smoothness, is that it implies a quadratic upper bound on the function, i.e. if $f$ is $\beta$-smooth then,

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \|y - x\|^2.$$

   To interpret this fix a point $x$. Convex functions always lie *above* their tangent lines (i.e. $f(y) \geq f(x) + \nabla f(x)^T (y - x)$). *Smooth* convex functions always lie *below* a parabola which passes through the point $(x, f(x))$ (defined by the RHS above).

[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.

2. Suppose $x^*$ is a minimum of a $\beta$-smooth function $f$, then for all $y \in \text{dom}(f)$

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

That is, if we are at a point $y$ that is close to the minimum $x^*$, then the gradient at $y$, $\nabla f(y)$ must also be small. So any algorithm we have that follows the gradients of the functions should intuitively slow down as it approaches the minimum.

3. Finally, if $f$ is twice differentiable, then $\beta$-smoothness is equivalent to the condition that,

$$0 \preceq \nabla^2 f(x) \preceq \beta I_d.$$

where the lower bound $0 \preceq$ comes from convexity of $f$ and the upper bound $\preceq \beta I_d$ comes from $\beta$-smoothness of $f$.

4. If $f$ is $\beta$-smooth then the function $\frac{\beta}{2}\|x\|^2 - f(x)$ is convex. Typically, we would not expect $-f(x)$ to be convex (except when $f$ is affine).

**Examples:** It is worth briefly considering two examples (canonical examples of non-smooth and smooth convex functions):

1. **Absolute value:** Here we consider $f(x) = |x|$, and observe that at $x = 0$, it's impossible to seat a parabola at the origin which is always above the function. Roughly, a parabola must have close to zero derivative near its minimum, but the absolute value function has constant derivative near its minimum.

2. **Quadratic function:** Suppose we consider $f(x) = x^T Q x + a^T x + b$ where $Q \succeq 0$. Its now easy to see that this function has Hessian $2Q$, and consequently it satisfies smoothness for any $\beta \geq 2\lambda_{\max}(Q)$ (i.e. twice the largest eigenvalue of $Q$).

> **Background:** For any positive semidefinite (PSD) matrix $A \in \mathbb{R}^{n \times n}$, written $A \succeq 0$, we have that $\forall x \in \mathbb{R}, x^T A x \geq 0$ and all its eigenvalues are non-negative $\lambda_{\min}(A) \geq 0$.
>
> PSD matrices also enjoy some convenient properties regarding eigenvalues.

If we have two PSD matrices $0 \preceq A$ and $0 \preceq B$ with $A, B \in \mathbb{R}^{n \times n}$, their sum $C = A + B$ is also PSD. We can bound the eigenvalues of matrix $C$:

$$\lambda_{\min}(C) \leq \lambda_{\min}(A) + \lambda_{\min}(B)$$
$$\lambda_{\max}(C) \geq \lambda_{\max}(A) + \lambda_{\max}(B)$$

We can show this by examining the Rayleigh quotient of $C$. The *Rayleigh quotient* $R(x; A) = \frac{x^T A x}{x^T x}$ for any PSD matrix (actually for any Hermitian matrix) is bounded by the largest and smallest eigenvalues of $A$, i.e. $R(x; A) \in [\lambda_{\min}(A), \lambda_{\max}(A)]$.

### 6.1.2 Strong Convexity

The twin assumption to smoothness is strong convexity.

**Definition 6.2** ($\alpha$-Strongly Convex)**.** *A function $f$ is $\alpha$-strongly convex, if the function $g(x) = f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.*

As with smoothness there are several important implications of strong convexity that you will explore in your HW.

1. If $f$ is strongly convex then an equivalent definition is that it satisfies the following inequality for any $x, y \in \text{dom}(f)$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2}\|y - x\|^2.$$

   Again to interpret this, fix a point $x$, and observe that this expression tells us that a strongly-convex function is *above* a parabola which passes through the point $(x, f(x))$.

2. If $f$ is twice differentiable, an equivalent characterization is that,

$$\nabla^2 f(x) \succeq \alpha I_d.$$

**Examples:**

1. **Absolute value:** Consider the same function as before. It is not strongly convex. For instance, if we consider $x = 1, y = 2$, then $f(y) - (f(x) + \nabla f(x)^T (y - x))$ is 0, so the definition can only hold with $\alpha = 0$.

2. **Quadratic function:** Once again using the second-order characterization of strong convexity we see that the quadratic function satisfies the definition of strong convexity for any $\alpha \leq 2\lambda_{\min}(Q)$.

It is possible to have strongly convex functions which are not smooth and vice versa, and it is worth trying to "draw" some examples to convince yourself of this.

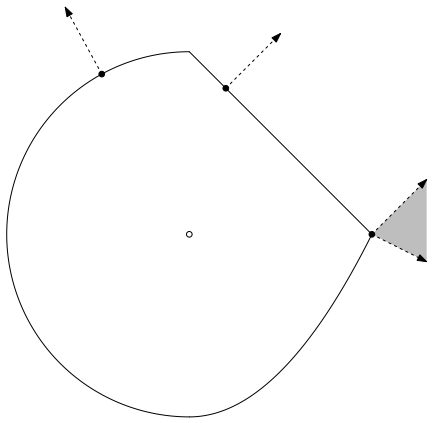
## 6.2 The Tangent Cone and Normal Cone

There is a fundamental reason why cones will be important to us. We will use them to characterize optimality. Two cones are important in this context: the normal cone and its polar cone (which has its own name, the tangent cone).

1. **Normal Cone:** Given a set $C$, and a point $x \in C$ the normal cone of $C$ at $x$ is defined as:

$$N_C(x) = \{g : g^T(y - x) \leq 0, \text{for all } y \in C\}.$$

It is important to make sense of the following figure (for clarity in the figure, the normal cone $N_C(x)$ has been translated to $x$).



There are three different types of points for which we should understand what the normal cone looks like: (1) Interior points (the normal cone is empty), (2) Boundary points where the boundary is smooth (the

normal cone is a single ray) (3) Boundary points where the boundary is not smooth (the normal cone is "fat").

Even if $C$ is not convex this cone is a convex cone (think about how you might show this).

2. **Tangent Cone:** For *convex sets* the polar of the normal cone is the tangent cone, i.e. $T_C(x) = N_C(x)^\circ$. In this case, the tangent cone is a convex cone.

   More generally (i.e. for non-convex sets) the tangent cone is defined to be the set of feasible (limiting) directions, i.e. roughly directions along which you can move and stay in the set $C$. This is possibly the more intuitive way of thinking about the tangent cone at a point (it is simply the set of feasible directions we can move and stay in the set). For general sets $C$, the tangent cone need not be convex.

**Segue...** Next we will consider optimality conditions, but for now we'll just summarize the punchline: in a convex optimization problem, a point $x$ will be optimal if the negative gradient belongs to $N_C(x)$, i.e. roughly if the direction we'd like to move makes at least a 90-degree angle with every direction that we *can* move in.

## 6.3 Optimality Conditions

Here we will revisit some things we discussed briefly in the previous lecture. Here is the basic question. We are interested in solving a problem:

$$\min_{x \in C} f(x),$$

where $f$ is a convex function, and $C$ is a convex set. What can I say about a solution $x^*$ to this problem?

1. **Unconstrained Case:** Suppose first that $C = \mathbb{R}^d$, and that $\mathrm{dom}(f) = \mathbb{R}^d$ then our characterization should be familiar to us from usual calculus classes.

   **Theorem 6.3.** *$x^*$ is optimal, if (and only if) $0 \in \partial f(x^*)$.*

**Proof:** If $0 \in \partial f(x^*)$, then from the first-order condition we know that,

$$f(y) \geq f(x^*) + 0^T(y - x^*) = f(x^*).$$

Conversely, if $x^*$ is optimal, then $f(y) \geq f(x^*)$ and we know that, $f(y) \geq f(x^*) + g_{x^*}^T(y - x^*)$ for all $y$, when $g_{x^*} = 0$ and so we know that 0 is valid subgradient at $x^*$.

■

2. **Constrained, Differentiable Case:** Now suppose that $C \subset \mathbb{R}^d$ and we wish to solve the constrained optimization problem: $\min_{x \in C} f(x)$. Recall that the minimum of the function within $C$ might be *different* than the minimum of the unconstrained function. So the gradient might not be zero. We need a new notion of optimality to cover this case.

   A feasible point $x^*$ is optimal, if and only if $\nabla f(x^*)^T(y - x^*) \geq 0$ for all $y \in C$.

   We will only verify one direction of this (the other direction requires a bit of analysis to check). Suppose that, $\nabla f(x^*)^T(y - x^*) \geq 0$ for all $y \in C$, then from the first-order condition we have that,

   $$f(y) \geq f(x^*) + \nabla f(x^*)^T(y - x^*) \geq f(x^*),$$

   so $f(y) \geq f(x^*)$ and $x^*$ is optimal.

   If you recall the definition of the normal cone, then you will see that this condition says that,

   $$-\nabla f(x^*) \in N_C(x^*).$$

   Consider three cases of optimality corresponding to the three cases we discussed with normal cones:

   (a) $x_1^*$ inside $C$: the gradient vanishes.

   (b) $x_2^*$ at a point where the boundary of $C$ is smooth: the gradient is orthogonal to the supporting hyperplane, i.e. a single ray.

   (c) $x_3^*$ at a non-smooth boundary point of $C$: the gradient is within the corresponding normal cone, which is wide.

3. **General, Constrained Case:** Now we consider the case where the function is nondifferentiable.

   A feasible point $x^*$ is optimal, if and only if $0 \in \partial f(x^*) + N_C(x^*)$. Here we are adding two sets, i.e. $C + D = \{y : y = u + v, u \in C, v \in D\}$.

   We leave one direction of the proof here in the lecture notes, but might not cover this in detail. Again it's only easy to verify one direction of this, i.e. suppose that $0 \in \partial f(x^*) + N_C(x^*)$, this means that there are two vectors $u \in \partial f(x^*)$ and $v \in N_C(x^*)$ such that,

   $$u + v = 0.$$

   Now, we know that for any $y$ which is feasible,

   $$\begin{aligned} f(y) &\geq f(x^*) + u^T(y - x^*) \\ &= f(x^*) - v^T(y - x^*). \end{aligned}$$

   Since $v \in N_C(x^*)$ we know that $v^T(y - x^*) \leq 0$ for every feasible $y$, and so we conclude that $f(y) \geq f(x^*)$.