SML Recitation Notes Week 2: Convexity

Min Xu

January 21, 2011

1 Geometry Fundamentals

We can describe a hyperplane in $\mathbb{R}^d$ as a normal vector $h \in \mathbb{R}^d$ and an offset $x_0 \in \mathbb{R}^d$.

- If x is on the plane, then $h^\mathsf{T}(x - x_0) = 0$
- If x is on one side of the plane, then $h^\mathsf{T}(x - x_0) > 0$
- If x is on the other side, then $h^\mathsf{T}(x - x_0) < 0$
- we can multiply h by non-zero scalar and still describe the same plane
- Since $h^\mathsf{T} x = h^\mathsf{T} x_0$ for all x on the plane, we can change our offset to be another point x_0' on the plane and still describe the same plane

Theorem 1. *(Supporting Hyperplane Theorem)*
Let $X \subset \mathbb{R}^d$ be a convex set. Let $x_0 \in boundary(X)$, then there exist a hyperplane $H = \{x \in \mathbb{R}^d : h^\mathsf{T}(x - x_0) = 0\}$ with normal vector $h \in \mathbb{R}^d$ and offset $x_0 \in X$ such that $h^\mathsf{T}(x - x_0) \geq 0$ for all $x \in X$.

Intuitively, this means that the supporting hyperplane touches the convex set but never crosses it; the entire convex set is on the same side of the hyperplane.

Remark 1. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function, then the **epigraph** of f is the set of $n+1$-dimensional points $\{(x, z) : x \in \mathbb{R}^n, z \in \mathbb{R}, z \geq f(x)\}$. The epigraph is a **convex set**. The boundary of the epigraph is the set of $n+1$-dimensional points $\{(x, f(x))\}$; this is **graph** of f.

2 Geometry of Subgradient

Definition 2. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function.

Let $w \in \mathbb{R}^n$, we say $w \in \partial f(x_0)$ if for all $x \in \mathbb{R}^n$, $f(x) \geq w^\mathsf{T}(x - x_0) + f(x_0)$

Geometrically, the subgradient of f at x_0 forms a supporting hyperplane of the epigraph of f at x_0. We can see this with a little algebra:

$$f(x) \leq w^\mathsf{T}(x - x_0) + f(x_0) \tag{2.1}$$
$$w^\mathsf{T} x_0 - f(x_0) \leq w^\mathsf{T} x - f(x) \tag{2.2}$$
$$(w, -1)^\mathsf{T}(x_0, f(x_0)) \leq (w, -1)^\mathsf{T}(x, f(x)) \tag{2.3}$$

Since $(w,-1)^{\mathsf{T}}(x_0, f(x_0)) \le (w,-1)^{\mathsf{T}}(x, f(x))$ for all x, the $n+1$-dimensional plane specified by normal vector $(w,-1)$ and offset $(x_0, f(x_0))$ is a supporting hyperplane for the epigraph.

Similarly, any supporting hyperplane of epigraph at offset $(x_0, f(x_0))$ with normal vector $(w,-1)$ (possibly need scaling) specifies a subgradient of f at x_0. By the supporting hyperplane theorem, $\partial f(x_0)$ is non-empty for all x_0 such that $f(x_0) \neq \infty$.

Theorem 3. *A convex function f is minimized at x_0 if and only if $\bar{0} \in \partial f(x_0)$.*

A direct proof is easy, but we can also intuitively confirm this theorem by considering subgradient as supporting hyperplane.

Remark 2. **(A Quick Digression into Infinity)**
Let f be a convex function. If $f(x_0) = -\infty$ and $f(x_1)$ is finite, then $f(tx_0 + (1-t)x_1) \le tf(x_0) + (1-t)f(x_1) \le -\infty$ and so f can only be finite on a scattered set of points. Thus, we generally assume convex functions are proper, i.e., cannot be $-\infty$ anywhere and cannot be ∞ everywhere.

In contrast, a lot of important convex functions will map points to ∞; we must take special care however when f is infinity. For example, if $f(x_0) = \infty$ and f is finite somewhere else, then f does not have a subgradient at x_0.

3 Geometry of Conjugate Dual

Definition 4. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a proper convex function. The conjugate dual is a function $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ defined as

$$f^*(y) = \sup_{x \in \mathbb{R}^n} x^{\mathsf{T}}y - f(x)$$

Suppose $\sup_{x \in \mathbb{R}^n} x^{\mathsf{T}}y - f(x)$ is achieved at x_0, then

$$x_0^{\mathsf{T}}y - f(x_0) \ge x^{\mathsf{T}}y - f(x) \text{ for all } x \in \mathbb{R}^n \tag{3.1}$$

$$f(x) \ge y^{\mathsf{T}}(x - x_0) + f(x_0) \tag{3.2}$$

And so $y \in \partial f(x_0)$. We can reverse the reasoning to show that if $y \in \partial f(x_0)$, then $\sup_{x \in \mathbb{R}^n} x^{\mathsf{T}}y - f(x)$ will be achieved at x_0.

We can now think about conjugate dual geometrically. Given a plane passing through origin with normal vector $(y,-1)$, $f^*(y)$ is the distance we have to shift this plane up and down until this plane becomes a supporting hyperplane of f. $f^*(y)$ is negative if we have to shift the plane up, positive if we shift the plane down.

Remark 3. There are cases where $\sup_{x \in \mathbb{R}^n} x^{\mathsf{T}}y - f(x)$ is not achieved at any $x_0 \in \mathbb{R}^n$. For example, it might be that $x^{\mathsf{T}}y - f(x)$ increases as x gets farther and farther away and $f^*(y) = \infty$. As another less important example, it might be that the gradient of f levels off as x goes to infinity.

Theorem 5. ***(Dual at Zero Theorem)***
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a proper convex function. Then

$$\inf_{x \in \mathbb{R}^n} f(x) = -f^*(\bar{0})$$

$(\bar{0}, -1)$ specifies precisely the hyperplane parallel to the n-dimensional domain of f so the theorem intuitively makes sense. The proof is also easy:

Proof.

$$-f^*(\bar{0}) = -\left(\sup_{x \in \mathbb{R}^n} x^\top \bar{0} - f(x)\right) \tag{3.3}$$

$$= \inf_{x \in \mathbb{R}^n} f(x) \tag{3.4}$$

□

Taking the conjugate dual of convex functions is a sensitive operation. For instance, $(f+g)^* \neq f^* + g^*$.

Theorem 6. *Let $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be closed-proper-convex functions. Then*

$$(f+g)^*(y) = \inf_{u,w \in \mathbb{R}^n : u+w=y} f^*(u) + g^*(w)$$

Hand-wavily, we can see that this theorem is natural because $\partial(f+g) = \partial f + \partial g$. So we can form one subgradient of $f+g$ by taking many different sums of a subgradient of f and a subgradient at g.

Proof. First, we recall that if f is closed-proper-convex, then $f^{**} = f$. Hence:

$$(f+g)^*(y) = \sup_{x \in \mathbb{R}^n} x^\top y - f(x) - g(x) \tag{3.5}$$

$$= \sup_{x \in \mathbb{R}^n} x^\top y - f^{**}(x) - g^{**}(x) \tag{3.6}$$

$$= \sup_{x \in \mathbb{R}^n} x^\top y - \left(\sup_{u \in \mathbb{R}^n} u^\top x - f^*(u)\right) - \left(\sup_{w \in \mathbb{R}^n} w^\top x - g^*(w)\right) \tag{3.7}$$

In general, we know that $-\sup_x f(x) = \inf_x -f(x)$. Hence we can continue:

$$(f+g)^*(y) = \sup_{x \in \mathbb{R}^n} x^\top y + \left(\inf_{u \in \mathbb{R}^n} -u^\top x + f^*(u)\right) + \left(\inf_{w \in \mathbb{R}^n} -w^\top x + g^*(w)\right) \tag{3.8}$$

$$= \sup_{x \in \mathbb{R}^n} \inf_{u,w \in \mathbb{R}^n} x^\top y - u^\top x - w^\top x + f^*(u) + g^*(w) \tag{3.9}$$

$$= \sup_{x \in \mathbb{R}^n} \inf_{u,w \in \mathbb{R}^n} x^\top (y - (u+w)) + f^*(u) + g^*(w) \tag{3.10}$$

Let u', w' minimize $\inf_{u,w \in \mathbb{R}^n} x^\top (y - (u+w)) + f^*(u) + g^*(w)$ (to be fully rigorous, we cannot assume such u', w' exist). Suppose also that $(y - (u' + w')) \neq \bar{0}$. Then $\sup_{x \in \mathbb{R}^n} x^\top (y - (u' + w')) + f^*(u) + g^*(w)$ is infinity because as we increase magnitude of x, we can make $x^\top (y - (u' + w'))$ arbitrarily large while $f^*(u)$ and $g^*(w)$ do not change. Thus, it must be that $(y - (u' + w') = \bar{0}$ and we can argue:

$$(f+g)^*(y) = \sup_{x \in \mathbb{R}^n} \inf_{u,w \in \mathbb{R}^n : u+w=y} x^\top \bar{0} + f^*(u) + g^*(w) \tag{3.11}$$

$$= \inf_{u,w \in \mathbb{R}^n : u+w=y} f^*(u) + g^*(w) \tag{3.12}$$

□

We can now derive Fenchel Duality Theorem as a corollary:

Corollary 7. ***(Fenchel Duality Theorem)***
Let $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be closed-proper-convex functions. Then:

$$\begin{aligned}
\inf_{x \in \mathbb{R}^n} (f+g)(x) &= -(f+g)^*(\bar{0}) && (3.13)\\
&= -(\inf_{\lambda \in \mathbb{R}^n} f^*(\lambda) + g^*(-\lambda)) && (3.14)\\
&= \sup_{\lambda \in \mathbb{R}^n} -f^*(\lambda) - f^*(-\lambda) && (3.15)
\end{aligned}$$

4 Connecting Subgradient and Conjugate Dual

There is an interesting connection between subgradient of the conjugate dual function and the subgradient of the original function.

Theorem 8. ***(Dual Subgradient Theorem)***
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a proper convex function. Let f^* be its conjugate dual. Then for all $x_0, y_0 \in \mathbb{R}^n$:

$$y_0 \in \partial f(x_0) \textit{ if and only if } x_0 \in \partial f^*(y_0)$$

This theorem says that "gradient" and "position" switches role for the conjugate dual. If f has gradient y_0 at position x_0, then f^* has gradient x_0 at position y_0.

To get an intuition for this theorem, let us consider the case where $x_0 = \bar{0}$ and let us suppose $f(\bar{0}) = 0$ and f has subgradient y_0 at $x_0 = \bar{0}$. In this case, the Dual Subgradient Theorem says that $\bar{0} \in \partial f^*(y_0)$ which means that f^* achieves its minimum at y_0.

Why is this so? If $f(\bar{0}) = 0$, then the origin in the $n+1$-dimensional space lies at the boundary of the epigraph of f. Hence, the hyperplane passing through origin with normal vector $(y_0, -1)$ is the the supporting hyperplane of the epigraph and it touches the epigraph at the origin. If we change y_0 by a bit, the hyperplane will "cut" into the epigraph and hence increase the value of f^*.

We can see this intuition appear in the proof as well:

Proof. **(Dual Subgradient Theorem)**
Step 1: We will first prove this for case where $x_0 = \bar{0}$ and $f(\bar{0}) = 0$. Let $y_0 \in \partial f(\bar{0})$, then for all $x \in \mathbb{R}^n$:

$$\begin{aligned}
f(x) &\geq y_0^\top (x - x_0) + f(x_0) && (4.1)\\
0 &\geq y_0^\top x - f(x) && (4.2)\\
0 &\geq \sup_{x \in \mathbb{R}^n} y_0^\top x - f(x) && (4.3)\\
0 &\geq f^*(y_0) && (4.4)
\end{aligned}$$

On the other hand, $f^*(y) = \sup_{x \in \mathbb{R}^n} y^\top x - f(x) \geq y^\top x_0 - f(x_0) \geq 0$. Hence, f^* is minimized at y_0 and $\bar{0} \in \partial f^*(y_0)$. We can reverse the reasoning to prove the converse.

Step 2: now we prove the theorem for the general case. Suppose $y_0 \in \partial f(x_0)$. Define function $g(x) = f(x + x_0) - f(x_0)$ (note that $f(x_0) < \infty$ since f has subgradient at x_0). Then we see that $g(\bar{0}) = 0$ and $\partial g(\bar{0}) = \partial f(x_0)$. Hence, $y_0 \in \partial f(x_0)$ if and only if $y_0 \in \partial g(\bar{0})$. By step 1, we know that $\bar{0} \in \partial g^*(y_0)$ and so we need to relate ∂g^* to ∂f^*

$$
\begin{aligned}
g^*(y_0) &= \sup_{x \in \mathbb{R}^n} x^\mathsf{T} y_0 - g(x) && (4.5)\\
&= \sup_{x \in \mathbb{R}^n} x^\mathsf{T} y_0 - f(x + x_0) + f(x_0) && (4.6)\\
&= \sup_{x \in \mathbb{R}^n} (x + x_0)^\mathsf{T} y_0 - f(x + x_0) - x_0^\mathsf{T} y_0 + f(x_0) && (4.7)\\
&= f^*(y_0) - x_0^\mathsf{T} y_0 + f(x_0) && (4.8)
\end{aligned}
$$

Hence, $\partial g^*(y_0) = \partial f^*(y_0) - x_0$ and we see that $\bar{0} \in \partial g^*(y_0)$ is equivalent to $x_0 \in \partial f^*(y_0)$. $\square$

We see then that the original function, the conjugate dual, and the subgradient have the following interesting connection:

- Finding the minimum value of f is equivalent to evaluating f^* at $\bar{0}$.

$$\inf_x f(x) = -f^*(\bar{0})$$

- Finding x that achieves minimum value of f is equivalent to finding the subgradient of f^* at $\bar{0}$.

$$\bar{0} \in \partial f(x) \text{ if and only if } x \in \partial f^*(\bar{0})$$