26 Foundations of Convex Analysis (†)

Our coverage of convexity is necessarily extremely brief. We introduce only what is necessary and refer the reader to standard texts for the proofs.

26.1 Convex sets and functions

A set $A \subseteq \mathbb{R}^d$ is convex if for any $x, y \in A$ it holds that $\alpha x + (1-\alpha)y \in A$ for all $\alpha \in (0,1)$. The convex hull of a collection of points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ is the smallest convex set containing the points, which also happens to satisfy

$$\mathrm{co}(x_1, x_2, \ldots, x_n) = \left\{ x \in \mathbb{R}^d : x = \sum_{i=1}^n p_i x_i \text{ for some } p \in \mathcal{P}_{d-1} \right\} .$$

The convex hull is also defined for an arbitrary set $A \subset \mathbb{R}^d$: $\mathrm{co}(A)$, the convex hull of A, is defined to be the smallest convex set containing A (see (c) in Figure 26.1). Let $A \subset \mathbb{R}^d$. The **polar** of A is denoted by A° and is defined as

$$A^\circ = \left\{ u \in \mathbb{R}^d : \sup_{x \in A} |\langle u, x \rangle| \leq 1 \right\} .$$

Of course, if $\sup_{x \in A} |\langle x, u \rangle| \leq 1$ and $\sup_{x \in A} |\langle x, v \rangle| \leq 1$, then $\sup_{x \in A} |\langle x, \alpha u + (1-\alpha)v \rangle| \leq 1$, which ensures that the polar is convex (even for non-convex A). For the rest of the section we let $A \subseteq \mathbb{R}^d$ be convex. Let $\bar{R} = \mathbb{R} \cup \{-\infty, \infty\}$ be the extended real number system and define operations involving infinities in the usual way (see notes).

DEFINITION 26.1 An extended real-valued function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is **convex** if its **epigraph** $E_f = \{(x, y) : x \in \mathbb{R}^d, y \geq f(x)\} \subset \mathbb{R}^{d+1}$ is a convex set.

The **domain** of an extended real-valued function on $\mathbb{R}^d$ is $\mathrm{dom}(f) = \{x \in \mathbb{R}^d : f(x) < \infty\}$. For $S \subset \mathbb{R}^d$, a function $f : S \to \bar{\mathbb{R}}$ is identified with the function $\bar{f} : \mathbb{R}^d \to \bar{\mathbb{R}}$ which coincides with f on S and is defined to take the value $+\infty$ outside of S. It follows that if $f : S \to \mathbb{R}$ then $\mathrm{dom}(f) = S$. A convex function is **proper** if its domain is nonempty and its range does not include $-\infty$.

For the rest of the chapter we will write "let f be a convex" to mean that $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is a proper convex function.

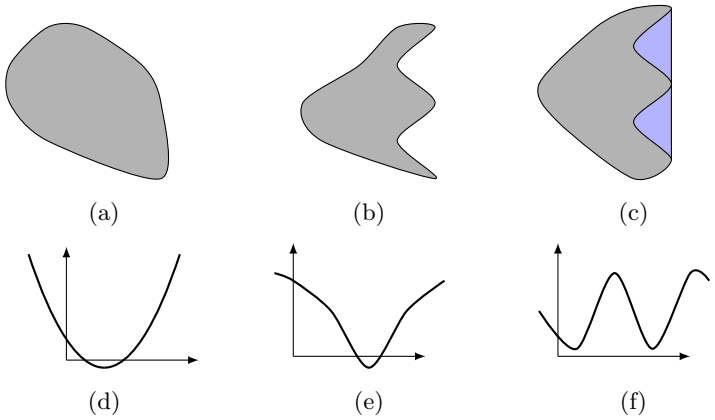


Figure 26.1 (a) is a convex set. (b) is a nonconvex set. (c) is the convex hull of a nonconvex set. (d) is a convex function. (e) is nonconvex, but all local minimums are global. (f) is not convex.

Permitting convex functions to take values of $-\infty$ is a convenient standard because certain operations on proper convex functions result in improper ones (infimal convolution, for example). These technicalities will never bother us in this book, however.

A consequence of the definition is that for convex f we have

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \quad \text{for all } \alpha \in (0,1) \text{ and } x, y \in \operatorname{dom}(f). \tag{26.1}$$

In fact, the inequality holds for all $x, y \in \mathbb{R}^d$.

Some authors use Eq. (26.1) as the definition of a convex function along with a specification that the domain is convex. If $A \subseteq \mathbb{R}^d$ is convex, then $f : A \to \mathbb{R}$ is convex if it satisfies Eq. (26.1) with $f(x) = \infty$ assumed for $x \notin A$.

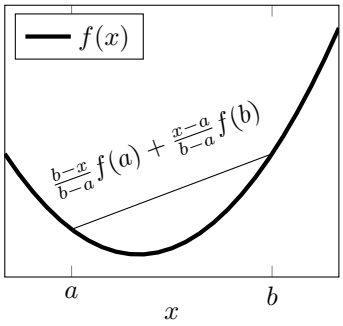
A function is **strictly convex** if the inequality in Eq. (26.1) is always strict. The **Fenchel dual** of a function f is $f^*(u) = \sup_{x \in \operatorname{dom}(f)} \langle x, u \rangle - f(x)$, which is convex because the maximum of convex functions is convex. The Fenchel dual is also called the convex conjugate. If $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is twice differentiable on its domain, then convexity of f is equivalent to its Hessian having nonnegative eigenvalues for all $x \in \operatorname{dom}(f)$. Strict convexity is equivalent to having strictly positive eigenvalues. The field of optimization is obsessed with convex functions because all local minimums are global (see figure). This means that minimizing a convex function is usually possible (efficiently) using some variation of gradient descent. A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is **concave** if $-f$ is convex.

26.2 Jensen's inequality

One of the most important results for convex functions is Jensen's inequality.

THEOREM 26.1 (Jensen's inequality) *Let $f : \mathbb{R}^d \to \bar{R}$ be a convex function and X be an $\mathbb{R}^d$-valued random element on some probability space such that $\mathbb{E}[X]$ exists and $X \in \mathrm{dom}(f)$ holds almost surely. Then $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.*

If we allowed Lebesgue integrals to take on the value of $+\infty$, the condition that X is almost surely an element of the domain of f could be removed and the result would still be true. Indeed, in this case we would immediately conclude that $\mathbb{E}[f(X)] = +\infty$ and Jensen's inequality would trivially hold. The basic inequality of (26.1) is trivially a special case of Jensen's inequality. Jensen's inequality is so central to convexity that it can actually be used as the definition (a function is convex if and only if it satisfies Jensen's inequality). The proof of Jensen's using Definition 26.1 is left to the reader (Exercise 26.1), but we include a picture to convince the reader. The direction of Jensen's inequality is reversed if 'convex' is replaced by 'concave'.

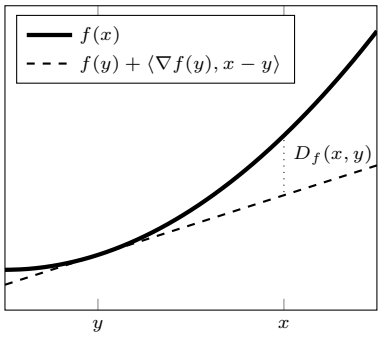


26.3 Bregman divergence

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, $x, y \in \mathbb{R}^d$, f differentiable at y. Then the **Bregman divergence** at y induced by f is defined by

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle \,,$$

where $\nabla f(y) \in \mathbb{R}^d$ is the gradient of f at y. To get a sense of the divergence function D_f, note that $D_f(x, y)$ is the difference between $f(x)$ and its first order Taylor expansion about the point y. Since f is convex, the linear approximation of f is a lower bound on f (why?) and so $D_f(x, y)$ is nonnegative over its domain with $D_f(x, x) = 0$.



THEOREM 26.2 *The following hold:*

(a) $D_f(x, y) \geq 0$ *for all* $x, y \in A$*;*
(b) $D_f(x, x) = 0$ *for all* $x \in A$*;*
(c) $D_f(x, y)$ *is convex as a function of* x*.*

The square root of the Bregman divergence shares many properties with a metric and for some choices of f it actually is a metric. In general, however, it is not symmetric and does not satisfy the triangle inequality.

EXAMPLE 26.1 Let $f(x) = \frac{1}{2}\|x\|_2^2$. Then $\nabla f(x) = x$ and

$$D_f(x,y) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x-y\rangle = \frac{1}{2}\|x-y\|_2^2 .$$

EXAMPLE 26.2 Let $A = [0,\infty)^d$, $\mathrm{dom}(f) = A$ and $f(x) = \sum_{i=1}^d (x_i \log(x_i) - x_i)$. Then $\nabla f(x) = \log(x)$ and

$$\begin{aligned} D_f(x,y) &= \sum_{i=1}^d (x_i \log(x_i) - x_i) - \sum_{i=1}^d (y_i \log y_i - y_i) - \sum_{i=1}^d \log(y_i)(x_i - y_i) \\ &= \sum_{i=1}^d x_i \log\left(\frac{x_i}{y_i}\right) + \sum_{i=1}^d (y_i - x_i) . \end{aligned}$$

Notice that if $x, y \in \mathcal{P}_{d-1}$ are in the unit simplex, then $D_f(x,y)$ is the relative entropy between probability vectors x and y. The function f is called the **unnormalized negentropy**, which will feature heavily in many of the chapters that follow.

26.4 Legendre functions

In this section we use various topological notions such as the interior, closed set and boundary. The definitions of these terms are given in the notes. Let f be a convex function and $A = \mathrm{dom}(f)$ and $C = \mathrm{int}(A)$. Then f is **essentially smooth** if:

(a) C is nonempty;
(b) f is differentiable on C;
(c) $\lim_{n\to\infty} \|\nabla f(x_n)\|_2 = \infty$ for any sequence $(x_n)_n$ with $x_n \in C$ for all n and $\lim_{n\to\infty} x_n = x$ and some $x \in \partial C$.

It is **essentially strictly convex** if f is strictly convex on every convex subset of $\mathrm{dom}(\nabla f)$. A **Legendre function** is a convex function f that is both essentially smooth and essentially strictly convex. The intuition is that the set $\{(x, f(x)) : x \in \mathrm{dom}(A)\}$ is a 'dish' with ever-steepening edges towards the boundary of the domain.

THEOREM 26.3 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Legendre function. Then:*

(a) ∇f is a bijection between $\mathrm{int}(\mathrm{dom}(f))$ and $\mathrm{int}(\mathrm{dom}(f^*))$ with the inverse $(\nabla f)^{-1} = \nabla f^*$.
(b) $D_f(x,y) = D_{f^*}(\nabla f(y), \nabla f(x))$ for all $x, y \in \mathrm{int}(\mathrm{dom}(f))$.

The next corollary formalizes the 'dish' intuition by showing the directional derivative along any straight path from a point in the interior to the boundary blows up.

COROLLARY 26.1 *Let f be Legendre and $x \in \operatorname{int}(\operatorname{dom}(f))$ and $y \in \partial\operatorname{int}(\operatorname{dom}(f))$, then $\lim_{\alpha\to 1}\langle \nabla f((1-\alpha)x + \alpha y), y - x\rangle = \infty$.*

EXAMPLE 26.3 Let f be the Legendre function given by $f(x) = \frac{1}{2}\|x\|_2^2$, which has domain $\operatorname{dom}(f) = \mathbb{R}^d$. Then $f^*(x) = f(x)$ and ∇f and ∇f^* are the identity functions.

EXAMPLE 26.4 Let $f(x) = -2\sum_{i=1}^d \sqrt{x_i}$ when $x_i \geq 0$ for all i and ∞ otherwise, which has $\operatorname{dom}(f) = [0,\infty)^d$ and $\operatorname{int}(\operatorname{dom}(f)) = (0,\infty)^d$. The gradient is $\nabla f(x) = -1/\sqrt{x}$, which blows up on sequence (x_n) approaching $\partial\operatorname{int}(\operatorname{dom}(f))$. Strict convexity is also obvious so f is Legendre. In Exercise 26.6 we ask you to calculate the Bregman divergences with respect to f and f^* and verify the results of Theorem 26.3.

The Taylor series of the Bregman divergence is often a useful approximation. Let $g(y) = D_f(x,y)$, which for $y = x$ has $\nabla g(y) = 0$ and $\nabla^2 g(y) = \nabla^2 f(x)$. A second order Taylor expansion suggests that

$$D_f(x,y) = g(y) \approx g(x) + \langle \nabla g(x), y - x\rangle + \frac{1}{2}\|y-x\|^2_{\nabla^2 f(x)} = \frac{1}{2}\|y-x\|^2_{\nabla^2 f(x)}\,.$$

This approximation can be very poor if x and y are far apart. Even when x and y are close the lower order terms are occasionally problematic, but nevertheless the approximation can guide intuition. The next theorem, which is based on Taylor's theorem, gives an exact result.

THEOREM 26.4 *If f is convex and twice differentiable in $A = \operatorname{int}(\operatorname{dom}(f))$ and $x, y \in A$, then there exists an $\alpha \in [0,1]$ and $z = \alpha x + (1-\alpha)y$ such that*

$$D_f(x,y) = \frac{1}{2}\|x-y\|^2_{\nabla^2 f(z)}\,.$$

The next result will be useful.

THEOREM 26.5 *Let $\eta > 0$ and f be Legendre and twice differentiable in $A = \operatorname{int}(\operatorname{dom}(f))$. Let $z \in [x,y]$ be the point such that $D_f(x,y) = \frac{1}{2}\|x-y\|^2_{\nabla^2 g(z)}$. Then for all $u \in \mathbb{R}^d$,*

$$\langle x - y, u\rangle - \frac{D_f(x,y)}{\eta} \leq \frac{\eta}{2}\|u\|^2_{(\nabla^2 f(z))^{-1}}\,.$$

Proof Strict convexity of f ensures that $H = \nabla^2 f(z)$ is invertible. Applying Cauchy-Schwartz we have

$$\langle x - y, u\rangle \leq \|x-y\|_H \|u\|_{H^{-1}} = \|u\|_{H^{-1}}\sqrt{2D_f(x,y)}\,.$$

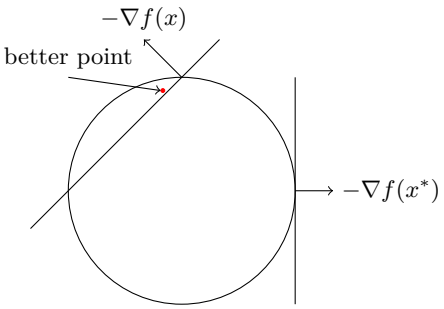


Figure 26.2 Illustration of first-order optimality conditions. The point at the top is not a minimizer because the hyperplane with normal as gradient does not support the convex set. The point at the right is a minimizer.

Therefore

$$\langle x - y, u \rangle - \frac{D_f(x,y)}{\eta} \leq \|u\|_{H^{-1}} \sqrt{2D_f(x,y)} - \frac{D_f(x,y)}{\eta} \leq \frac{\eta}{2} \|u\|_{H^{-1}}^2 ,$$

where the last step follows from the useful trick that $ax - bx^2 \leq a^2/(4b)$ for all $x \in \mathbb{R}$ and $b \geq 0$. □

26.5 Optimization

The **first-order optimality condition** states that if $x \in \mathbb{R}^d$ is the minimizer of differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, then $\nabla f(x) = 0$. One of the things we like about convex functions is that when f is convex the first-order optimality condition is both necessary and sufficient. The first-order optimality condition can also be generalized to constrained minima. In particular, if $A \subseteq \mathbb{R}^d$ is a nonempty convex set and $f : A \to \mathbb{R}$ is convex, then

$$x^* \in \operatorname{argmin}_{x \in A} f(x) \Leftrightarrow \forall x \in A : \langle \nabla f(x^*), x - x^* \rangle \geq 0 \,. \tag{26.2}$$

The necessity of this condition is easy to understand by a geometric reasoning as shown in Fig. 26.2. Since x^* is a minimizer of f over A, $-\nabla f(x^*)$ must be the outer normal of a **supporting hyperplane** H_{x^*} of A at x^* otherwise x^* could be moved by a small amount while staying inside A and improving the value of f. Since A is convex, it thus lies entirely on the side of H_{x^*} that $\nabla f(x^*)$ points into. This is clearly equivalent to (26.2). The sufficiency of the condition also follows from this geometric viewpoint as the reader may verify from the figure.

The above statement continues to hold with a small modification even when f is not everywhere differentiable. In particular, in this case the equivalence (26.2) holds for any $x^* \in \operatorname{dom}(\nabla f)$ with the modification that on both sides of the equivalence, A should be replaced by $A \cap \operatorname{dom}(f)$:

PROPOSITION 26.1 *Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be a convex function, $A \neq \emptyset$, $A \subset \mathbb{R}^d$ convex. Then for any $x^* \in \mathrm{dom}(\nabla f)$, it holds that:*

$$x^* \in \mathrm{argmin}_{x \in A \cap \mathrm{dom}(f)} f(a) \Leftrightarrow \forall x \in A \cap \mathrm{dom}(f) : \langle \nabla f(x^*), x - x^* \rangle \geq 0\,. \tag{26.3}$$

26.6 Projections

If $A \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$, then the Euclidean projection of x on A is $\Pi_A(x) = \mathrm{argmin}_{y \in A} \|x - y\|_2^2$. One can also project with respect to a Bregman divergence induced by convex function f. Let $\Pi_{A,f}$ by

$$\Pi_{A,f}(x) = \mathrm{argmin}_{y \in A} D_f(y, x)\,.$$

A property of the projection that will be exploited heavily in subsequent chapters is that minimizing a Legendre function f on a convex constrained set A is (usually) equivalent to finding the unconstrained minimum on the domain of f and then projecting that point onto A.

THEOREM 26.6 *Let f be Legendre, $A \subset \mathbb{R}^d$ a closed convex set and assume that $\tilde{y} = \mathrm{argmin}_{z \in \mathrm{dom}(f)} f(y)$ exists. Then the following hold:*

(a) $y = \mathrm{argmin}_{z \in A \cap \mathrm{dom}(f)} f(z)$ *exists and is unique;*
(b) $y = \mathrm{argmin}_{z \in A \cap \mathrm{dom}(f)} D_f(z, \tilde{y})$.

The assumption that $\tilde{y}$ exists is necessary. For example $f(x) = -\sqrt{x}$ for $x \geq 0$ and $f(x) = \infty$ for $x < 0$ is Legendre with domain $\mathrm{dom}(f) = [0, \infty)$, but f does not have a minimum on its domain.

26.7 Notes

1 The 'infinity arithmetic' on the extended real line is as follows:

$$\begin{aligned}
&\alpha + \infty = \infty \qquad \text{for } \alpha \in (-\infty, \infty] \\
&\alpha - \infty = -\infty \qquad \text{for } \alpha \in [-\infty, \infty) \\
&\alpha \cdot \infty = \infty \text{ and } \alpha \cdot (-\infty) = -\infty \qquad \text{for } \alpha > 0 \\
&\alpha \cdot \infty = -\infty \text{ and } \alpha \cdot (-\infty) = \infty \qquad \text{for } \alpha < 0 \\
&0 \cdot \infty = 0 \cdot (-\infty) = 0\,.
\end{aligned}$$

Like $\alpha/0$ the value of $\infty - \infty$ is not defined. We also have $\alpha \leq \infty$ for all α and $\alpha \geq -\infty$ for all α.

2 There are many ways to define the topological notions used in this chapter. The most elegant is also the most abstract, but there is no space for that here. Instead we give the classical definitions that are specific to $\mathbb{R}^d$ and subsets. Let A be a subset of $\mathbb{R}^d$. A point $x \in A$ is an **interior point** if there exists

an $\varepsilon > 0$ such that $B_\varepsilon(x) = \{y : \|x - y\|_2 \leq \varepsilon\} \subset A$. The **interior** of A is $\text{int}(A) = \{x \in A : x \text{ is an interior point}\}$. The set A is **open** if $\text{int}(A) = A$ and **closed** if its complement $A^c = \mathbb{R}^d \setminus A$ is open. The boundary of A is denoted by ∂A and is the set of points in $x \in \mathbb{R}^d$ such that for all $\varepsilon > 0$ the set $B_\varepsilon(x)$ contains points from A and A^c. Note that points in the boundary need not be in A. Some examples: $\partial \mathbb{R}^n = \emptyset$ and $\partial[0, \infty) = \{0\}$.

26.8 Bibliographic remarks

The main source for these notes is the excellent book by Rockafellar [2015]. The basic definitions are in Part I. The Fenchel dual is analyzed in Part III while Legendre functions are are found in Part V. Convex optimization is a huge topic. The standard text is by Boyd and Vandenberghe [2004].

26.9 Exercises

26.1 Prove Jensen's inequality (Theorem 26.1). Precisely, let $X \in \mathbb{R}^d$ be a random variable for which $\mathbb{E}[X]$ exists and $f : \mathbb{R}^d \to \bar{R}$ a convex function. Prove that $\mathbb{E}[f(x)] \geq f(\mathbb{E}[X])$.

Let $x_0 = \mathbb{E}[X] \in \mathbb{R}^d$ and define a linear function $g : \mathbb{R}^d \to \mathbb{R}$ such that $g(x_0) = f(x_0)$ and $g(x) \leq f(x)$ for all $x \in \mathbb{R}^d$.

26.2 For each of the real-valued functions below decide whether or not it is Legendre on the given domain.

(a) $f(x) = x^2$ on $[-1, 1]$.
(b) $f(x) = -\sqrt{x}$ on $[0, \infty)$.
(c) $f(x) = \log(1/x)$ on $[0, \infty)$ with $f(0) = \infty$.
(d) $f(x) = x \log(x)$ on $[0, \infty)$ with $f(0) = 0$.
(e) $f(x) = |x|$ on $\mathbb{R}$.
(f) $f(x) = \max\{|x|, x^2\}$ on $\mathbb{R}$.

26.3 Prove Theorem 26.2.

26.4 Prove Corollary 26.1.

26.5 Prove Proposition 26.1.

26.6 Let f be the convex function given in Example 26.4.

(a) For $x, y \in \text{dom}(f)$ find $D_f(x, y)$.
(b) Compute $f^*(u)$ and $\nabla f^*(u)$.

(c) Find $\text{dom}(\nabla f^*)$.
(d) Show that for $u, v \in (-\infty, 0]^d$,

$$D_{f^*}(u, v) = -\sum_{i=1}^{d} \frac{(u_i - v_i)^2}{u_i v_i^2}\,.$$

(e) Verify the claims in Theorem 26.3.

26.7 Let f be Legendre. Show that $\tilde{f}$ given by $\tilde{f}(x) = f(x) + \langle x, u \rangle$ is also Legendre for any $u \in \mathbb{R}^d$.

26.8 Let f be the unnormalized negentropy function from Example 26.2.

(a) Prove that f is Legendre.
(b) Given $y \in [0, \infty)^d$, prove that $\text{argmin}_{x \in \mathcal{P}_{d-1}} D_f(x, y) = y/\|y\|_1$.

26.9 Let $\alpha \in [0, 1/d]$ and $\mathcal{A} = \mathcal{P}_{d-1} \cap [\alpha, 1]^d$ and f be the unnormalized negentropy function. Let $y \in [0, \infty)^d$ and $x = \text{argmin}_{x \in \mathcal{A}} D_f(x, y)$ and assume that $y_1 \leq y_2 \leq \cdots \leq y_d$. Let m be the smallest value such that $y_m / \sum_{i=m}^{d} y_i \geq \alpha$. Show that $x_i = \alpha$ if $i < m$ and $x_i = y_i / \sum_{j=m}^{d} y_j$ otherwise.