Convex Optimization Notes

Jonathan Siegel

January 2017

1 Convex Analysis

This section is devoted to the study of convex functions $f : \mathbb{B} \to \mathbb{R} \cup \{+\infty\}$ and convex sets $U \subset \mathbb{B}$, for $\mathbb{B}$ a Banach space. The case of $\mathbb{B} = \mathbb{R}^n$ will be of particular interest. We start with the fundamental definition of a convex function and a convex set.

Definition 1.1. *A function $f : \mathbb{B} \to \mathbb{R} \cup \{+\infty\}$, with $\mathbb{B}$ a Banach space, is called convex if for all $x, y \in \mathbb{B}$, $\lambda \in [0, 1]$, we have*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \tag{1}$$

Note that we are assuming the usual arithmetic rules regarding $+\infty$ in this definition. To rule out the trivial case where $f(x) = +\infty$ for all x, we introduce the following definition.

Definition 1.2. *A convex function $f : \mathbb{B} \to \mathbb{R} \cup \{+\infty\}$ is called proper if there exists a $y \in \mathbb{B}$ such that $f(y) < \infty$.*

Of course, we also introduce the notion of a convex set.

Definition 1.3. *A set $U \subset \mathbb{B}$, with $\mathbb{B}$ a Banach space, is called convex if $x, y \in U$ and $\lambda \in [0, 1]$ implies that $\lambda x + (1 - \lambda)y \in U$.*

We are allowing our convex functions to take the value $+\infty$. Another way of affecting the same thing is to consider convex functions $f : U \subset \mathbb{B} \to \mathbb{R}$ where U is a convex set. The advantage of our approach is that the function contains all of the information about its domain instead of this being specified in addition.

Definition 1.4. *Let $f : \mathbb{B} \to \mathbb{R} \cup \{+\infty\}$. The domain of f, written* $\mathrm{dom}(f)$*, is the set*

$$\mathrm{dom}(f) = \{x \in \mathbb{B} : \ f(x) < +\infty\} \tag{2}$$

It is of course clear that the domain of a convex function is a convex set, namely we have

Proposition 1.1. *If f is convex, then* $\mathrm{dom}(f)$ *is a convex set.*

Proof. Let $x, y \in \mathrm{dom}(f)$, and let $\lambda \in [0,1]$. Since f is convex and $x, y \in \mathrm{dom}(f)$, we have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) < +\infty \tag{3}$$

Hence $\lambda x + (1-\lambda)y \in \mathrm{dom}(f)$ as desired. $\square$

One of the remarkable properties of convex functions is that they enjoy surprising regularity. This is true even though the definition makes no reference to differentiability or continuity. We begin with an elementary result in this direction.

Proposition 1.2. *Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a convex function such that* $\mathrm{dom}(f)$ *is open. Then f is continuous on* $\mathrm{dom}(f)$*.*

Proof. If $\mathrm{dom}(f)$ is empty, the assertion is trivial. Otherwise, let $x \in \mathrm{dom}(f)$. Then there exist $a, b \in \mathrm{dom}(f)$ such that $a < x < b$. Assume that $x^* \in (a, x)$. Then we have, by convexity

$$f(x^*) \leq \lambda f(a) + (1-\lambda)f(x) \tag{4}$$

$$f(x) \leq \mu f(b) + (1-\mu)f(x^*) \tag{5}$$

for some $\lambda, \mu \in (0,1)$. Moreover, as $x^* \to x$, we have $\lambda, \mu \to 0$. Rewriting the last equation above yields

$$\frac{1}{1-\mu}f(x) + \frac{\mu}{1-\mu}f(b) \leq f(x^*) \leq \lambda f(a) + (1-\lambda)f(x) \tag{6}$$

Letting $x^* \to x$, which forces $\lambda, \mu \to 0$, and noting that $f(a), f(b) < +\infty$, we see that $f(x^*) \to f(x)$. Hence f is left continuous. A completely analogous argument show that f is also right continuous and thus continuous. $\square$

We now examine the higher dimensional case, which requires more effort. However, we will derive a stronger result. In particular, we will obtain local Lipschitz continuity.

Theorem 1.1. *Let $f : \mathbb{B} \to \mathbb{R} \cup \{+\infty\}$ be a convex function on a Banach space B. Let $U \subset B$ be open such that f is bounded on U, i.e. $f(x) \leq C < \infty$ on U. Then f is locally Lipschitz continuous on U.*

Proof. Let $x \in U$ and choose $\epsilon > 0$ such that $B_\epsilon(x) = \{y \in B : \|x - y\| \leq \epsilon\} \subset U$. Let $y \in B_\epsilon(x)$, $y \neq x$ and define

$$w = x - (y - x)\frac{\epsilon}{\|x - y\|} \tag{7}$$

Then we have that

$$x = \mu w + (1-\mu)y \tag{8}$$

with $\mu = (\|x - y\|)/(\epsilon + \|x - y\|)$. Thus convexity ensures that

$$f(x) \leq \mu f(w) + (1 - \mu) f(y) \tag{9}$$

In particular, the above equation implies that if $f(x) < 0$

$$f(y) \geq \frac{1}{1-\mu} f(x) - \frac{\mu}{1-\mu} f(w) \geq 2f(x) - C \tag{10}$$

since $\mu \leq 1/2$. Additionally, if $f(x) \geq 0$ by the same logic we have

$$f(y) \geq -C \tag{11}$$

Since $y \in B_\epsilon(x)$ was arbitrary, it follows that f is bounded below on $B_\epsilon(x)$. This implies that there exists $K < \infty$ such that $|f(x)| \leq K$ on $B_\epsilon(x)$.

Now let $y, z \in B_{\epsilon/4}(x)$ and note that $z \in B_{\epsilon/2}(y) \subset B_\epsilon(x)$. For y and z, define

$$v = y + (z - y)\frac{\epsilon}{2\|z - y\|} \tag{12}$$

$$w = y - (z - y)\frac{\epsilon}{2\|z - y\|} \tag{13}$$

Thus v and w are the points on the boundary of $B_{\epsilon/2}(y)$ which are collinear with y and z. The important thing to note is that

$$z = \lambda v + (1 - \lambda) y \tag{14}$$

with $\lambda = (2\|z - y\|)/\epsilon$ and

$$y = \mu w + (1 - \mu) z \tag{15}$$

with $\mu = (2\|z - y\|)/(\epsilon + 2\|z - y\|)$. Thus convexity ensures that

$$f(z) \leq \lambda f(v) + (1 - \lambda) f(y) \tag{16}$$

$$f(y) \leq \mu f(w) + (1 - \mu) f(z) \tag{17}$$

Rearranging, we obtain

$$f(z) - f(y) \leq \lambda(f(v) - f(y)) \tag{18}$$

$$f(y) - f(z) \leq \mu(f(w) - f(z)) \tag{19}$$

which implies that

$$-\lambda(f(v) - f(y)) \leq f(y) - f(z) \leq \mu(f(w) - f(z)) \tag{20}$$

and thus

$$|f(z) - f(y)| \leq \max\{\lambda|f(v) - f(y)|, \mu|f(w) - f(z)|\} \tag{21}$$

So by the triangle inequality we have

$$|f(z) - f(y)| \leq 2\max\{\lambda, \mu\}K \tag{22}$$

Recalling the definition of λ and μ we finally obtain

$$|f(z) - f(y)| \leq \frac{4K}{\epsilon}\|z - y\| \tag{23}$$

which completes the proof of local Lipschitz continuity. □

This immediately implies Lipschitz continuity on open subsets of $\mathbb{R}^d$

Proposition 1.3. *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be convex and let $U \subset \operatorname{dom}(f)$ be an open set. Then f is locally Lipschitz continuous on U.*

Proof. Let $x \in U$. In light of the preceding theorem, we only need to show that $x \in V \subset U$ with V open and $f \leq K < \infty$ on V. To do so define the subset

$$P_\epsilon(x) = \{y \in \mathbb{R}^d : \ |x - y|_1 \leq \epsilon\} \tag{24}$$

and note that $P_\epsilon(x)$ has non-empty interior, in particular $x \in P_\epsilon(x)^{\mathrm{o}}$.

More importantly, note that $P_\epsilon(x)$ is the convex hull of the finitely many points $(x \pm \epsilon e_i)_{i=1}^d$.

Now choose ϵ so small that $P_\epsilon(x) \subset U \subset \operatorname{dom}(f)$. Thus, in particular,

$$f(x \pm \epsilon e_i) < +\infty \tag{25}$$

for all i. Consequently, for all $y \in P_\epsilon(x)^{\mathrm{o}} \subset P_\epsilon(x)$, we have

$$f(y) \leq K \tag{26}$$

where $K = \max\{f(x \pm \epsilon e_i) : \ i = 1, ..., d\} < \infty$. Thus, the Lipschitz continuity follows from the previous theorem. □

We now continue our study of the regularity of convex functions by introducing a generalization of the derivative, known as the sub-differential.

Definition 1.5. *Let $f : \mathbb{B} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The sub-differential of f, $D_f \subset \mathbb{B} \times \mathbb{B}^*$ (here $\mathbb{B}^*$ denotes the dual space of $\mathbb{B}$) is defined as*

$$D_f = \{(x, p) \in \mathbb{B} \times \mathbb{B}^* : \ f(y) \geq f(x) + \langle p, y - x\rangle \textit{ for all } y \in \mathbb{B}\} \tag{27}$$

The sub-differential of f at x is

$$\partial f(x) = \{p \in \mathbb{B}^* : \ (x, p) \in D_f\} \tag{28}$$

First we wish to characterize which subsets $A \subset \mathbb{B} \times \mathbb{B}^*$ can be sub-differentials of proper convex functions. To this end, we introduce the notion of a cyclically monotone subset of $\mathbb{B} \times \mathbb{B}^*$.

Definition 1.6. *A subset $A \subset \mathbb{B} \times \mathbb{B}^*$ is called cyclically monotone if for any $n \geq 2$ and $(x_1, p_1), ..., (x_n, p_n) \in A$, we have*

$$\sum_{i=1}^{n} \langle p_i, x_{i+1} - x_i \rangle \leq 0 \tag{29}$$

where x_{n+1} is set equal to x_1.

It is relatively easy to prove that the sub-differential of a proper convex function is cyclically monotone.

Proposition 1.4. *Let $f : \mathbb{B} \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. Then D_f is cyclically monotone.*

Proof. Let $(x_1, p_1), ..., (x_n, p_n) \in D_f$. Note first that since f is proper, $f(x_i) < \infty$ for all i. This follows since for $f(y) < \infty$ (there must exist such a y as f is proper), we have

$$f(x_i) + \langle p_i, y - x_i \rangle \leq f(y) < \infty \tag{30}$$

which is clearly impossible if $f(x_i) = \infty$.

Now consider the following inequalities which follow from the definition of the sub-differential D_f.

$$f(x_{i+1}) \geq f(x_i) + \langle p_i, x_{i+1} - x_i \rangle \tag{31}$$

where $x_{n+1} = x_1$. Adding these inequalities and noting that all of the $f(x_i)$ cancel (here we use that $f(x_i) < \infty$), we see that

$$\sum_{i=1}^{n} \langle p_i, x_{i+1} - x_i \rangle \leq 0 \tag{32}$$

As $(x_1, p_1), ..., (x_n, p_n) \in D_f$ were chosen arbitrarily, we see that D_f is cyclically monotone as desired. □

A literal converse of the above statement cannot hold since the empty set is cyclically monotone, for instance, yet is not the sub-differential of any proper convex function. However, every cyclically monotone set is contained in the sub-differential of a lower-semicontinuous proper convex function. Of course, we must first introduce the concept of lower-semicontinuity.

Definition 1.7. *A function $f : \mathbb{B} \to \mathbb{R} \cup \{+\infty\}$ is lower-semicontinuous if for all sequences $x_n \in \mathbb{B}$ with $x_n \to x$, we have*

$$f(x) \leq \liminf_{n} f(x_n) \tag{33}$$

The following simple and useful result allows us to construct many different lower-semicontinuous convex functions.

Proposition 1.5. *Let f_α, $\alpha \in A$ be a family of functions. Then if each f_α is lower-semicontinuous, so is $f = \sup_{\alpha\in A} f_\alpha$. Also, if each f_α is convex, then so is $f = \sup_{\alpha\in A} f_\alpha$.*

Proof. The first statement is a consequence of the well-known inequality

$$\sup_\alpha \liminf_\beta g(\alpha,\beta) \le \liminf_\beta \sup_\alpha g(\alpha,\beta) \tag{34}$$

Namely, if $x_n \in \mathbb{B}$ with $x_n \to x$, then

$$f(x) = \sup_{\alpha\in A} f_\alpha(x) \le \sup_{\alpha\in A} \liminf_n f_\alpha(x_n) \le \liminf_n \sup_{\alpha\in A} f_\alpha(x_n) = \liminf_n f(x_n) \tag{35}$$

where the first inequality follows since each f_α is lower-semicontinuous.

To prove the second statement, let $x, y \in \mathbb{B}$, $\lambda \in \mathbb{R}$, and let $\epsilon > 0$. Then for some $\alpha \in A$ we have

$$f(\lambda x + (1-\lambda)y) < f_\alpha(\lambda x + (1-\lambda)y) + \epsilon \tag{36}$$

Thus since f_α is convex, we have

$$f(\lambda x + (1-\lambda)y) < \lambda f_\alpha(x) + (1-\lambda) f_\alpha(y) + \epsilon \le \lambda f(x) + (1-\lambda) f(y) + \epsilon \tag{37}$$

Since $\epsilon > 0$ was arbitrary, we see that

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \tag{38}$$

and f is convex as desired. □

We can now prove the following.

Proposition 1.6. *Let $A \subset \mathbb{B} \times \mathbb{B}^*$ be cyclically monotone. Then there exists a function $f : \mathbb{B} \to \mathbb{R}$, which is lower-semicontinuous, proper and convex, such that*

$$A \subseteq D_f \tag{39}$$

The clever proof, due to Rockafellar, proceeds by constructing the required function f.

Proof. If A is empty the statement is obvious, so assume that $(z,q) \in A$ for some $(z,q) \in \mathbb{B} \times \mathbb{B}^*$. Define the function f as follows

$$f(x) = \sup_{(y_1,r_1),\ldots,(y_m,r_m)\in A} \langle q, y_1 - z\rangle + \sum_{i=1}^{m-1} \langle r_i, y_{i+1} - y_i\rangle + \langle r_m, x - y_m\rangle \tag{40}$$

I.e. we form the sum from the definition of cyclical monotonicity with the pairs $(z,q), (y_1,r_1), \ldots, (y_m,r_m)$, but we set $y_{m+1} = x$ instead of z. Then we take the supremum over all choices of $(y_1,r_1), \ldots, (y_m,r_m) \in A$.

We proceed to show that f satisfies the conditions of the theorem. Note first that f is lower-semicontinuous and convex since it is the supremum of a collection of linear functions. Now, we check that f is proper. In particular, we check that $f(z) = 0$. This is a result of the cyclical monotonicity of A. Note that for any choice of $(y_1, r_1), ..., (y_m, r_m) \in A$, we have

$$\langle q, y_1 - z\rangle + \sum_{i=1}^{m-1} \langle r_i, y_{i+1} - y_i\rangle + \langle r_m, z - y_m\rangle \leq 0 \tag{41}$$

exactly by the cyclical monotonicity of A applied to $(z, q), (y_1, r_1), ..., (y_m, r_m)$. Now note that the choice $(y_1, r_1) = (z, q)$ gives $f(z) = 0$.

Finally, we check that $A \subset D_f$. To this end, assume that $(y, r) \in A$ and $x \in \mathbb{R}^n$. Now let $(y_1, r_1), ..., (y_m, r_m) \in A$ be arbitrary. Then by applying the definition of $f(x)$ to the tuple $(y_1, r_1), ..., (y_m, r_m), (y, r)$, we see that

$$f(x) \geq \langle q, y_1 - z\rangle + \sum_{i=1}^{m-1} \langle r_i, y_{i+1} - y_i\rangle + \langle r_m, y - y_m\rangle + \langle r, x - y\rangle \tag{42}$$

Taking the supremum over tuples $(y_1, r_1), ..., (y_m, r_m) \in A$ and noting the definition of $f(y)$, we see that

$$f(x) \geq f(y) + \langle r, x - y\rangle \tag{43}$$

and thus $(y, r) \in D_f$. Since $(y, r) \in A$ was arbitrary, we see that $A \subseteq D_f$ as desired. $\square$

This doesn't completely answer the question of which sets can occur as sub-differentials of proper convex functions. We know that every sub-differential is cyclically monotone and every cyclically monotone subset is contained in a sub-differential. This seems to suggest that maximal cyclically monotone sets (with respect to inclusion) would correspond to sub-differentials of lower-semicontinuous proper convex functions. However, we must first show that the sub-differential of a lower-semicontinuous proper convex function is indeed maximal with respect to inclusion (what could fail is that the subdifferential of a lower-semicontinuous proper convex function could be strictly contained in the subdifferential of a different function). This will turn out to be the case.

Another, related question which we would like to ask is whether the sub-differential characterizes a convex function f (up to a constant, of course), i.e. whether it is possible for two different convex functions (which differ by more than a constant) to have the same sub-differential. Unfortunately, this is possible, so the sub-differential of a convex function does not uniquely characterize the function.

A example is given by the functions $f_a \to \mathbb{R} \cup \{+\infty\}$ defined by

$$f_a(x) = \begin{cases} 0 & -1 < x < 1 \\ a & x = \pm 1 \\ +\infty & |x| > 1 \end{cases}$$

Clearly, f_a is convex if $a \geq 0$. Moreover, a simple computation shows that

$$D_{f_a} = \begin{cases} (-1,1) \times \{0\} & a > 0 \\ ((-1,1) \times \{0\}) \cup (\{-1\} \times (-\infty, 0]) \cup (\{1\} \times [0, \infty)) & a = 0 \end{cases}$$

Thus the sub-differential is the same for all functions f_a with $a > 0$. The issue is with the values on the boundary of the domain. To remedy this, we must require that our functions to be lower semi-continuous. This ensures that the domain of our function is closed and that it takes the "correct" values on the boundary. We will then see that the sub-differential does characterize lower-semicontinuous functions.

Both of these considerations demonstrate the role that the assumption of lower-semicontinuity plays.

The following proposition solves both of the above problems.

Proposition 1.7. *Let $f : \mathbb{B} \to \mathbb{R} \cup \{+\infty\}$ be lower-semicontinuous, proper and convex. Then for any lower-semicontinuous proper convex function g, if $D_f \subseteq D_g$, then $g = f + C$ for some $C \in \mathbb{R}$. Thus, $D_f = D_g$.*

We won't prove this proposition until we've established more refined regularity results, but let us mention some of the consequences.

Corollary 1.1. *The sub-differential of a lower-semicontinuous proper convex is a maximal cyclically monotone set, i.e. it is a cyclically monotone set which is not strictly contained in any other cyclically monotone set.*

Proof. Let f be a lower-semicontinuous proper convex function. If $D_f \subseteq A$ with A cyclically monotone, then by proposition (1.6) we know that there exists a lower-semicontinuous proper convex function g such that $A \subseteq D_g$. Thus $D_f \subseteq D_g$, which implies that $D_f = D_g$ by the previous proposition. Hence $D_f = A$ and D_f is a maximal cyclically monotone set. $\square$

Corollary 1.2. *Maximal cyclically monotone sets are in one-to-one correspondence with lower-semicontinuous proper convex functions up to addition of a constant.*

Proof. Given a lower-semicontinuous proper convex function f, its sub-differential is a maximal cyclically monotone set. Given a maximal cyclically monotone set A, proposition (1.6) produces a lower-semicontinuous proper convex function g, with $A \subset D_g$. Now we need to show that these are inverses of each other (where we consider two convex functions equivalent if they differ by a constant).

But it is clear that if $A = D_f$, then proposition (1.6) produces a lower-semicontinuous proper convex function g with $D_f \subset D_g$ and hence $g = f + C$. Also, if A is a maximal cyclically monotone set, then proposition (1.6) produces a lower-semicontinuous proper convex function g with $A \subset D_g$. Then the maximality of A implies that $A = D_g$.

Thus these two maps are inverses of each other as desired. $\square$

This correspondence between lower-semicontinuous proper convex functions and maximal cyclically monotone sets is very useful in understanding the splitting methods of non-smooth convex optimization. This is because minimizing a convex function is the same as finding an x such that $0 \in \partial f(x)$, i.e. is the same as finding a zero of a maximal cyclically monotone set.

2 Dual of a Convex Function

In this section, we introduce the Fenchel dual of a lower-semicontinuous proper convex function and derive some of its properties.

To begin with, recall that lower-semicontinuous proper convex functions (up to adding a constant) are in bijection with maximal cyclically monotone sets. We now note the following simple fact about maximal cyclically monotone sets.

Lemma 2.1. *let $A \subset \mathbb{B} \times \mathbb{B}^*$ be a maximal monotone set. Then the set B obtained by swapping each element of A, i.e.*

$$B = \{(x, y) \in \mathbb{B}^* \times \mathbb{B} \ s.t. \ (y, x) \in A\} \tag{44}$$

is also maximal cyclically monotone.

Proof. First, we shall show that B is cyclically monotone. To this end we let $(p_1, x_1), ..., (p_n, x_n) \in B$ and note that

$$\sum_{i=1}^{n} \langle x_i, p_i - p_{i+1} \rangle = \sum_{i=1}^{n} \langle x_i - x_{i-1}, p_i \rangle \leq 0 \tag{45}$$

here $x_0 = x_n$, $p_{n+1} = p_1$ and the last inequality follows since by the definition of B, $(x_n, p_n), ..., (x_1, p_1) \in A$ and A is cyclically monotone.

The maximality of B follows since if $B \subseteq B'$ with B' cyclically monotone, then by swapping each element of B', we obtain a cyclically monotone set containing A, which must then be equal to A, implying that $B = B'$. $\square$

In summary, maximal cyclically monotone sets correspond to lower-semicontinuous proper convex functions (up to adding a constant) and flipping the entries of a maximal cyclically monotone preserves maximality and cyclical monotonicity.

So given a (lower-semicontinuous proper convex) function $f : \mathbb{B} \to \mathbb{R}$, there exists a (lower-semicontinuous proper convex) function $g : \mathbb{B}^* \to \mathbb{R}$ (unique up to adding a constant) whose sub-differential consists of the flipped entries of D_f. It turns out that there is a simple formula for g in terms of f.

Proposition 2.1. *Let $f : \mathbb{B} \to \mathbb{R}$ be a lower-semicontinuous proper convex function and define $g : \mathbb{B}^* \to \mathbb{R}$ as follows*

$$g(z) = \sup_{x \in \mathbb{B}} \{\langle z, x \rangle - f(x)\} \tag{46}$$

Then g is lower-semicontinuous, proper and convex and

$$D_g = \{(x, y) \ s.t. \ (y, x) \in D_f\} \tag{47}$$

Proof. It is clear that g is lower-semicontinuous and convex, since it is the supremum of a collection of linear functions.

Now, assume that $(x, p) \in D_f$. Recall that this means that for all $y \in \mathbb{B}$ we have

$$f(y) \geq f(x) + \langle p, y - x \rangle \tag{48}$$

This implies that $\langle p, x \rangle - f(x) \geq \langle p, y \rangle - f(y)$ for all $y \in \mathbb{B}$ and so

$$g(p) = \sup_{z \in \mathbb{B}} \{\langle p, z \rangle - f(z)\} = \langle p, x \rangle - f(x) \tag{49}$$

Note first that this implies that g is proper ($g(p) < \infty$). Also, if we let $q \in \mathbb{B}^*$ be arbitrary, then

$$\begin{aligned} g(q) = \sup_{z \in \mathbb{B}} \{\langle q, z \rangle - f(z)\} \geq \langle q, x \rangle - f(x) &= \langle p, x \rangle - f(x) + \langle q - p, x \rangle \\ &= g(p) + \langle q - p, x \rangle \end{aligned} \tag{50}$$

Thus, $(p, x) \in D_g$. So we have that

$$\{(x, y) \ s.t. \ (y, x) \in D_f\} \subseteq D_g \tag{51}$$

Now g is a proper convex function and so D_g is cyclically monotone. Moreover, D_f is maximal cyclically monotone since f is lower-semicontinuous. Also, by the previous proposition, maximal cyclical monotonicity is preserved when the entries of our set are swapped. Thus we have

$$D_g = \{(x, y) \ s.t. \ (y, x) \in D_f\} \tag{52}$$

as desired. $\square$

3 Refined Regularity Results

Recall that we previously proved that a convex function on $\mathbb{R}^n$ is Lipschitz continuous on any open subset of its domain. In this section, we prove more refined regularity results concerning the differentiability of convex functions. This will allow us to give a proof of proposition (1.7).

4 Monotone Operators

In this subsection we introduce the notion of a monotone set. This generalizes the concept of a cyclically monotone set and thus it generalizes the notion of a sub-differential of a convex function. This notion will prove very useful is studying non-smooth convex optimization.

Definition 4.1. *A set $A \subset \mathbb{B} \times \mathbb{B}^*$ is called monotone if for any $(x_1, p_1), (x_2, p_2) \in A$ we have*

$$\langle p_1 - p_2, x_2 - x_1 \rangle \leq 0 \tag{53}$$

Note that applying the definition of a cyclically monotone set to the pair $(x_1, p_1), (x_2, p_2)$ produces exactly this condition. This definition is more general, though, since we only require the cyclical monotonicity condition for pairs of elements in A, as opposed to arbitrary sequences of elements of A. This gives us the following lemma.

Lemma 4.1. *If $A \subset \mathbb{B} \times \mathbb{B}^*$ is cyclically monotone, then it is monotone.*

Monotone sets are indeed more general than cyclically monotone sets, as the following example shows.

Example 4.1. *Let $\mathbb{B} = \mathbb{R}^2$, and consider the set $A \subset \mathbb{R}^2 \times \mathbb{R}^2$ defined as follows.*

$$A = \{((x, y), (-y, x)) \text{ with } (x, y) \in \mathbb{R}^2\} \tag{54}$$

Then A is monotone, but not cyclically monotone. To see this, let $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$ and note that for the corresponding points in A, we have

$$\langle p_1 - p_2, x_2 - x_1 \rangle = (-y_1 + y_2, x_1 - x_2) \cdot (x_2 - x_1, y_2 - y_1) = 0 \tag{55}$$

and thus A is monotone. However, the definition of cyclical monotonicity fails for the sequence of points of A corresponding to $(1, 0), (0, 1), (-1, 0), (0, -1) \in \mathbb{R}^2$.

Recall that by corollary (1.2) sub-differentials of lower semi-continuous convex functions and maximal cyclically monotone sets are the same thing. By the above, these are monotone sets. An important question is whether they are maximal monotone sets, i.e. even though they aren't strictly contained in any cyclically monotone set, can they be strictly contained in a monotone set?

Proposition 4.1. *Let $f : \mathbb{B} \to \mathbb{R}$ be a lower-semicontinuous proper convex function. Then D_f is a maximal monotone set.*

Proof. $\square$

The utility of these results is that it allows us to generalize the problem of minimizing a convex function to the problem of finding a zero of a maximal monotone set. This is defined as follows.

Definition 4.2. *Let $A \subset \mathbb{B} \times \mathbb{B}^*$ be a monotone set. A zero of A is a point $x \in \mathbb{B}$ such that $(x, 0) \in A$.*

Note that not every monotone set is the sub-differential of a convex function, and so this is indeed a strict generalization. In the next section we will introduce the proximal point method for solving this problem (in the case where $\mathbb{B}$ is a Hilbert space) and investigate how various algorithms in convex optimization are special cases of it.

First, we define some operations on monotone sets and investigate their properties.

Definition 4.3. *Let $A, B \subset \mathbb{B} \times \mathbb{B}^*$ be monotone sets, and $0 \leq \lambda \in \mathbb{R}$. Then we define*

$$\lambda A = \{(x, \lambda y) \text{ s.t. } (x, y) \in A\} \tag{56}$$

$$A + B = \{(x, y + z) \text{ s.t. } (x, y) \in A \text{ and } (x, z) \in B\} \tag{57}$$

Proposition 4.2. *Let $A, B \subset \mathbb{B} \times \mathbb{B}^*$ be monotone sets, and $0 \leq \lambda \in \mathbb{R}$. Then λA and $A + B$ are both monotone.*

Proof. We merely need to check directly the definition. Let $(x, p), (y, q) \in \lambda A$. This means that there are $(x, p^*), (y, q^*) \in A$ such that $\lambda p^* = p$ and $\lambda q^* = q$. Thus we have

$$\langle p - q, y - x\rangle = \lambda\langle p^* - q^*, y - x\rangle \leq 0 \tag{58}$$

as $\lambda \geq 0$ and $\langle p^* - q^*, y - x\rangle \leq 0$ since A is monotone.

Now assume that $(x, p), (y, q) \in A + B$. Thus there exist $(x, p_A), (y, q_A) \in A$ and $(x, p_B), (y, q_B) \in B$ such that $p = p_A + p_B$ and $q = q_A + q_B$. Then we have

$$\langle p - q, y - x\rangle = \langle p_A - q_A, y - x\rangle + \langle p_B - q_B, y - x\rangle \leq 0 \tag{59}$$

since both A and B are monotone. □

Note that if A and B are sub-differentials of convex functions, then these operations correspond to scaling and addition of the corresponding convex functions.

The first question we will ask is how these operations relate to maximal monotonicity.

Proposition 4.3. *If A is a maximal monotone set and $0 < \lambda \in \mathbb{R}$, then λA is a maximal monotone set.*

Proof. The key to the proof is to note that by expanding the definition $\lambda^{-1}(\lambda A) = A$, as

$$\lambda^{-1}(\lambda A) = \{(x, \lambda^{-1}y) \text{ s.t. } (x, y) \in \lambda A\} = \{(x, \lambda^{-1}\lambda y) \text{ s.t. } (x, y) \in A\} \tag{60}$$

Also, we definition we clearly have that if $A \subseteq B$, then $\lambda A \subseteq \lambda B$.

So now we assume that $\lambda A \subseteq B$. Multiplying by λ^{-1}, we have

$$A \subseteq \lambda^{-1}B \tag{61}$$

which implies by the maximality of A that $A = \lambda^{-1}B$. Multiplying this by λ we get

$$\lambda A = B \tag{62}$$

which proves that λA is maximal monotone as desired. □

Note that the above theorem fails if $\lambda = 0$. For instance, consider the maximal monotone set $A = \{(0, p)\ p \in \mathbb{B}^*\}$. Then $\lambda A = \{(0, 0)\}$ which is not maximal monotone. In terms of convex functions, what is happening is that multiplying a convex function by 0 doesn't make sense since the function takes on infinite values outside of its domain.

Additionally, even if A and B are both maximal monotone, the sum $A + B$ need not be maximal monotone. Determining conditions on A and B which ensure that $A + B$ is maximally monotone is an intricate problem.

5 The Proximal Point Method

Here we introduce the proximal point method for finding a zero of a maximal monotone set. The key idea here are the notions of the resolvent and proximal operator. Throughout, we will assume that $\mathbb{B}$ is a Hilbert Space and so $\mathbb{B}^*$ is canonically identified with $\mathbb{B}$.

Definition 5.1. *The identity set is* $I = \{(x, x) \ s.t. \ x \in \mathbb{B}\} \subset \mathbb{B} \times \mathbb{B}$.

Note that this is a maximal monotone set since it is the subdifferential of the function $\frac{1}{2}\|x\|_{\mathbb{B}}^2$.

Definition 5.2. *Let* $A \subset \mathbb{B} \times \mathbb{B}$ *be a monotone set, then we define* A^{-1} *as*

$$A^{-1} = \{(y, x) \ s.t. \ (x, y) \in A\} \tag{63}$$

Recall from the previous section that if A *is the subdifferential of a convex function, then this corresponds to taking the Fenchel dual of the function.*

Definition 5.3. *Let* $A \subset \mathbb{B} \times \mathbb{B}$ *be a monotone set. We define the resolvent of* A *(which is a new monotone set) as*

$$R_A = (I + A)^{-1} \tag{64}$$

The following important definition and lemma collects some properties of the resolvent operator.

Definition 5.4. *Let* $R \subset \mathbb{B} \times \mathbb{B}$ *be a monotone set. Then we say that* R *is said to be non-expansive if*

$$\|y_1 - y_2\|_{\mathbb{H}} \leq \|x_1 - x_2\|_{\mathbb{H}} \tag{65}$$

whenever $(x_1, y_1), (x_2, y_2) \in RA$.

R is said to be firmly non-expansive if

$$\|y_1 - y_2\|_{\mathbb{H}}^2 \leq \langle y_1 - y_2, x_1 - x_2 \rangle_H \tag{66}$$

whenever $(x_1, y_1), (x_2, y_2) \in R$.

Note that a contraction is obviously non-expansive, and a firmly non-expansive set is also non-expansive since

$$\langle f(x) - f(y), x - y \rangle_H \leq \|f(x) - f(y)\|_{\mathbb{H}} \|x - y\|_{\mathbb{H}} \tag{67}$$

Also, a firmly non-expansive map need not be a contraction (take for instance f to be the identity) and a contraction need not be firmly non-expansive (take for instance $H = \mathbb{R}^2$ and R to be $((x, y), (-\lambda y, \lambda x)))$. Finally, firmly non-expansive sets are also monotone since $\|y_1 - y_2\|^2 \geq 0$.

Lemma 5.1. *Assume that* A *is a monotone set. Then* R_A *is a maximal monotone operator.*

Proof. The proof of this is simply a matter of expanding the definition of R_A. Note that

$$R_A = \{(x + y, x) \text{ s.t. } (x, y) \in A\} \tag{68}$$

So we must show that if $(x_1, y_1), (x_2, y_2) \in A$, then

$$\langle x_1 - x_2, x_1 - x_2 \rangle \leq \langle x_1 - x_2, (x_1 - x_2) + (y_1 - y_2) \rangle \tag{69}$$

but this is equivalent to

$$0 \leq \langle x_1 - x_2, y_1 - y_2 \rangle \tag{70}$$

which is true iff A is monotone. $\square$

Note that we can reverse these implications and this gives us a bijection between monotone sets and firmly non-expansive sets (note that $A = R_A^{-1} - I$).

Lemma 5.2. *A non-expansive set (in particular a firmly non-expansive set) A is single valued, i.e. $(x, y_1), (x, y_2) \in A$ implies that $y_1 = y_2$.*

Proof. This is obvious since the condition of non-expansiveness implies that $\|y_1 - y_2\| \leq \|x - x\| = 0$. $\square$

The previous lemma implies that the resolvent of a maximal monotone set defines a function $R_A : D(R_A) \to \mathbb{B}$ where $D(S) = \{x \text{ s.t. } (x, y) \in S \text{ for some } y\}$ (the domain of the set S). We can characterize the monotone sets A for which $D(R_A) = \mathbb{B}$.

Theorem 5.1 (Minty)**.** *The domain of R_A is all of $\mathbb{B}$ iff A is a maximal monotone set.*

The previous theorem tells us that if A is maximal monotone, then R_A defines a map $R_A : \mathbb{B} \to \mathbb{B}$. From now on, we will abuse notation and use R_A to denote this map as well as its graph (the subset of $\mathbb{B} \times \mathbb{B}$).

From the definitions, we can describe explicitly how to calculate $R_A(z)$ for $z \in \mathbb{B}$. First, solve for $(x, y) \in A$ satisfying $x + y = z$ (that this can always be done if A is maximal monotone is essentially the content of Minty's theorem). Then return x.

For a given maximal monotone set A, it may be difficult to solve the equation $x + y = z$ with $(x, y) \in A$. However, if $A = D_f$ with f convex and lower-semicontinuous, then we easily verify that

$$R_A(z) = \arg\min_{x \in \mathbb{B}} \frac{1}{2}\|x - z\|_{\mathbb{H}}^2 + f(x) \tag{71}$$

since x^* being the optimizer of this problem implies that $z - x^* \in Df(x^*)$. From now on we abuse notation and write R_f for R_{D_f} and call this the resolvent of the convex function f (in the literature it may also be referred to as the proximal map of f).

For a variety of convex function f, there are efficient algorithms for solving this optimization problem. One may argue that in each of these cases there is also a simple algorithm for minimizing f itself. This is correct, however, the usefulness of the resolvent map is in the development of splitting methods, which allow us to bootstrap algorithms for solving the optimization problem above to algorithms for minimizing sums of convex functions.

The proximal point algorithm consists of the iteration $x_{n+1} = R_A(x_n)$. Note that if $x^* = R_A(x^*)$, then by definition $(x, 0) \in A$. Hence fixed points of the proximal point iteration are zeros of the maximal monotone set A. In the next section investigate the convergence properties of the proximal point algorithm.

6 Peaceman-Rachford and Douglas-Rachford Splitting

We begin with a simple lemma about firmly non-expansive operators.

Lemma 6.1. *Let $R : \mathbb{B} \to \mathbb{B}$. Then the following are equivalent:*
1. *R is a firmly non-expansive operator.*
2. *$2R - I$ is a non-expansive operator.*
3. *$R = \frac{1}{2}S + \frac{1}{2}I$ for some non-expansive operator S.*

Proof. $(1 \to 2)$ We compute

$$\begin{aligned}&\|(2R(x) - x) - (2R(y) - y)\|^2 = \\ &\langle 2(R(x) - R(y)) - (x - y), 2(R(x) - R(y)) - (x - y)\rangle = \\ &4(\|R(x) - R(y), R(x) - R(y)\|^2 - \langle R(x) - R(y), x - y\rangle) + \|x - y\|^2\end{aligned} \tag{72}$$

Now, since R is firmly non-expansive, we have that

$$\|R(x) - R(y), R(x) - R(y)\|^2 - \langle R(x) - R(y), x - y\rangle \leq 0 \tag{73}$$

and thus

$$\|(2R(x) - x) - (2R(y) - y)\|^2 \leq \|x - y\|^2 \tag{74}$$

so that $2R - I$ is non-expansive.

$(2 \to 3)$ This is obvious since $R = \frac{1}{2}(2R - I) + \frac{1}{2}I$.

$(3 \to 1)$ Note that $S = 2R - I$ and reverse the argument for $(1 \to 2)$. □

The operator $2R - I$ is called the reflection of R about the identity. Note that it may be the case that $2R - I$ is also firmly non-expansive (for instance, if R is a translation, then $2R - I$, $4R - 3I$, ... are all firmly non-expansive).

The problem that we want a splitting method for is solving $0 \in (A + B)(x)$ for maximal monotone sets A and B. Note that this is the same as finding an x such that $(x, a) \in A$, $(x, b) \in B$ and $a + b = 0$. Recall that iterating R_A calculates x s.t. $(x, 0) \in A$ and iterating R_B calculates x s.t. $(x, 0) \in B(x)$.

A reasonable approach seems to be to iterate $R_A \circ R_B$ or $\frac{1}{2}(R_A \circ R_B) + \frac{1}{2}I)$. $R_A \circ R_B$ is certainly non-expansive (as R_A and R_B are) which means that averaging it with the identity produces a firmly non-expansive operator by the above lemma. However, it has the wrong fixed points, as we shall now compute.

Recall that to compute $R_B(z)$, we first find $(x, y) \in B$ such that $x + y = z$ and return x. Then to apply R_A we calculate $(a, b) \in A$ such that $a + b = x$ and return a. We can write this as

$$R_A \circ R_B = \{(z, a) \text{ s.t. } (x, y) \in B,\ (a, b) \in A,\ x + y = z,\ a + b = x\} \tag{75}$$

Thus z is a fixed point of $R_A \circ R_B$ i.e. $(z, z) \in R_A \circ R_B$ iff $(z, b) \in A$, $(z+b, y) \in B$ with $z+y+b = z$ or $y+b = 0$. What we really want is a z such that $(z, y) \in B$, $(z, b) \in A$ and $y + b = 0$. But here we have B evaluated at the wrong point $z + b$.

This was really a warm up for introducing the Peaceman-Rachford and Douglas-Rachford iterations. These splitting methods were motivated by some numerical methods developed in the 50s for solving the heat equation. We begin first with the Peaceman-Rachford iteration.

Definition 6.1 (Peaceman-Rachford)**.** *Let A and B be maximal monotone sets. Then the Peaceman-Rachford iteration is*

$$x_{n+1} = (2R_A - I) \circ (2R_B - I)(x_n) \tag{76}$$

Thus, instead of composing the resolvent operators, we compose their reflections.

Since the resolvents are firmly non-expansive, the previous lemma implies that their reflections are non-expansive. Thus their composition is non-expansive and the Peaceman-Rachford iteration is equivalent to iterating a non-expansive map.

Now we calculate the fixed points of the Peaceman-Rachford iteration.

Lemma 6.2. *If z is a fixed point of $(2R_A - I) \circ (2R_B - I)$, then $(R_B(z), 0) \in (A + B)$, i.e. $R_B(z)$ is a zero of $A + B$.*

Proof. We first characterize the calculating of $(2R_A - I)(z)$. Calculating this is equivalent to finding $(a, b) \in A$ such that $a + b = z$ and returning $2a - (a + b) = a - b$. We apply $2R_B - I$ in a similar fashion. This yields the following characterization of the Peaceman-Rachford map.

$$\begin{aligned} &(2R_A - I) \circ (2R_B - I) = \\ &\{(z, a - b) \text{ s.t. } (x, y) \in B,\ (a, b) \in A,\ x + y = z,\ a + b = x - y\} \end{aligned} \tag{77}$$

Thus, z is a fixed point of the iteration iff $z = a - b$ with

$$(x, y) \in B,\ (a, b) \in A,\ x + y = z = a - b,\ a + b = x - y$$

But the linear system $x + y = a - b$ and $a + b = x - y$ implies that $x = a$ and $y = -b$.

But now $(x, y) \in A$, $(a = x, b) \in B$ and $y+b = 0$, so x is a zero of $A+B$ (what we wanted to find!). Moreover, $(x, y) \in B$ and $x+y = z$ so that $x = R_B(z)$. $\square$

The Peaceman-Rachford algorithm finds a 0 of $A+B$ by iterating the map $(2R_A - I) \circ (2R_B - I)$ and then applying R_B to the fixed point.

One issue is that $(2R_A - I) \circ (2R_B - I)$ is only a non-expansive map, so its iterates are not guaranteed to converge (even weakly) to a fixed point (take for instance a rotation in $\mathbb{R}^2$).

However, we have seen that we can obtain a firmly non-expansive map with the same fixed points by averaging with the identity! This produces the Douglas-Rachford iteration.

Definition 6.2 (Douglas-Rachford)**.** *Let A and B be maximal monotone sets. Then the Peaceman-Rachford iteration is*

$$x_{n+1} = \frac{1}{2}(2R_A - I) \circ (2R_B - I)(x_n) + \frac{1}{2}x_n \tag{78}$$

which can be rewritten in the form

$$x_{n+1} = R_A \circ (2R_B - I)(x_n) - (R_B - I)(x_n) \tag{79}$$

The Douglas-Rachford algorithm iterates this map (which is firmly non-expansive and so is guaranteed to converge (at least weakly) to a fixed point) and then applies R_B to the result.

6.1 Linear Convergence of Douglas-Rachford under Strong Convexity and Smoothness

We examine the behavior of the Douglas-Rachford and Peaceman-Rachford iterations in the case where at least one of the maximal monotone sets involved is the subdifferential of a strongly convex and smooth function.

We begin by recalling the following definitions.

Definition 6.3. *A function $f : \mathbb{H} \to \mathbb{R}$ is called α-strongly convex if $f(x) - \frac{1}{2}\alpha\|x\|_H^2$ is convex.*

Definition 6.4. *A function $f : \mathbb{H} \to \mathbb{R}$ is called β-smooth if it is everywhere differentiable and $\|\nabla f(x_1) - \nabla f(x_2)\|_H \leq \beta\|x_1 - x_2\|_H$ (recall that H is a Hilbert space and so it is self-dual).*

The goal of this section is to prove the following

Theorem 6.1. *Suppose that $f : \mathbb{H} \to \mathbb{R}$ is β-smooth and α-strongly convex. Then $2R_f - I$ is a contraction with $\lambda = \sqrt{\left(1 - \frac{4\alpha}{(1+\beta)^2}\right)}$, i.e.*

$$\|(2R_f - I)(x_1) - (2R_f - I)(x_2)\|_H \leq \sqrt{\left(1 - \frac{4\alpha}{(1+\beta)^2}\right)}\|x_1 - x_2\|_H \tag{80}$$

Note that this theorem implies that the Douglas-Rachford iteration converges linearly if at least one of the maximal monotone sets is the subdifferential of a strongly convex and smooth function. This is because the composition of a non-expansive map and a contraction is a contraction.

Proof. Note that since $f(y) - \frac{1}{2}\alpha\|y\|_H^2$ is convex and f is differentiable, we have that $(y, \nabla f(y) - \alpha y) \subset \mathbb{H} \times \mathbb{H}$ is a monotone set. This means that

$$\langle (\nabla f(y_1) - \alpha y_1) - (\nabla f(y_2) - \alpha y_2), y_1 - y_2 \rangle \geq 0 \tag{81}$$

so that

$$\langle \nabla f(y_1) - \nabla f(y_2), y_1 - y_2 \rangle \geq \alpha \|y_1 - y_2\|_H^2 \tag{82}$$

Using the Cauchy-Schwartz inequality and the assumption that f is β-smooth yields

$$\langle y_1 - y_2, \nabla f(y_1) - \nabla f(y_2) \rangle \leq \|y_1 - y_2\|_H \|\nabla f(y_1) - \nabla f(y_2)\|_H \leq \beta \|y_1 - y_2\|_H^2 \tag{83}$$

Combining these inequalities we obtain

$$\alpha \|y_1 - y_2\|_H^2 \leq \langle y_1 - y_2, \nabla f(y_1) - \nabla f(y_2) \rangle \leq \beta \|y_1 - y_2\|_H^2 \tag{84}$$

Note that this immediately implies that $\alpha \leq \beta$.

Now recall that $R_f(x)$ is the solution to the equation $y + \nabla f(y) = x$ and thus $(2R_f - I)(x) = 2R_f(x) - x = y - \nabla f(y)$ with $y + \nabla f(y) = x$. We now compute

$$\|(2R_f - I)(x_1) - (2R_f - I)(x_2)\|_H^2 = \|(y_1 - \nabla f(y_1)) - (y_2 - \nabla f(y_2))\|_H^2 =$$

$$\|y_1 - y_2\|_H^2 + \|\nabla f(y_1) - \nabla f(y_2)\|_H^2 - 2\langle y_1 - y_2, \nabla f(y_1) - \nabla f(y_2) \rangle$$

On the other hand, we have

$$\|x_1 - x_2\|_H^2 = \|(y_1 + \nabla f(y_1)) - (y_2 + \nabla f(y_2))\|_H^2 =$$

$$\|y_1 - y_2\|_H^2 + \|\nabla f(y_1) - \nabla f(y_2)\|_H^2 + 2\langle y_1 - y_2, \nabla f(y_1) - \nabla f(y_2) \rangle$$

Combining these two equations we get that

$$\|(2R_f - I)(x_1) - (2R_f - I)(x_2)\|_H^2 = \|x_1 - x_2\|_H^2 - 4\langle y_1 - y_2, \nabla f(y_1) - \nabla f(y_2) \rangle \tag{85}$$

so that

$$\|(2R_f - I)(x_1) - (2R_f - I)(x_2)\|_H^2 \leq \|x_1 - x_2\|_H^2 - 4\alpha \|y_1 - y_2\|_H^2 \tag{86}$$

But the equation for $\|x_1 - x_2\|_H^2$ also implies that

$$\|x_1 - x_2\|_H^2 \leq \|y_1 - y_2\|_H^2 + (\beta \|y_1 - y_2\|_H)^2 + 2\beta \|y_1 - y_2\|_H^2 \tag{87}$$

and we have

$$\|y_1 - y_2\|_H^2 \geq \frac{1}{(1+\beta)^2} \|x_1 - x_2\|_H^2 \tag{88}$$

so that we finally get

$$\|(2R_f - I)(x_1) - (2R_f - I)(x_2)\|_H^2 \leq \left(1 - \frac{4\alpha}{(1+\beta)^2}\right) \|x_1 - x_2\|_H^2 \tag{89}$$

as desired. $\square$

This theorem tells us that the Douglas-Rachford iteration is most useful when minimizing the sum of two convex functions, one of which is strongly convex and smooth. One of the main applications of this is regularizing a smooth, strongly convex optimization problem using an L1 penalty term (which is often done in statistics). This is indeed where ADMM (which we shall see is Douglas-Rachford in disguise, but on the dual problem) has proved incredibly well-behaved and useful.