Convex Analysis Notes

Lecturer: Adrian Lewis, Cornell ORIE
Scribe: Kevin Kircher, Cornell MAE

These are notes from ORIE 6328, Convex Analysis, as taught by Prof. Adrian Lewis at Cornell University in the spring of 2015. They cover the basic theory of convex sets and functions, several flavors of duality, a variety of optimization algorithms (with a focus on nonsmooth problems), and an introduction to variational analysis building up to the Karush-Kuhn-Tucker conditions. Proofs are mostly omitted. If you find any errors – which are the fault of the scribe, not the lecturer – feel free to let Kevin know.

Contents

Chapter 1

Basic theory

1.1 Preliminaries

This section lays out some definitions and notation, following §1.1 of [1] and Appendix A of [2]. The setting is a Euclidean space $\mathbf{E}$, *i.e.*, a real vector space with finite-dimensional elements, inner product $\langle \mathbf{x}, \mathbf{y} \rangle$, and norm $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. Without loss of generality, we could set $\mathbf{E} = \mathbf{R}^n$ and $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$, but the coordinate-free notation allows one framework to handle a variety of optimization problems. For example, in semidefinite programming $\mathbf{E} = \mathbf{S}^n \equiv \left\{ X \in \mathbf{R}^{n \times n} \;\middle|\; X^T = X \right\}$ and $\langle X, Y \rangle = \operatorname{tr}(XY)$.

Unless otherwise noted, all vectors and sets lie in $\mathbf{E}$.

1.1.1 Sets

The sum of sets S_1 and S_2 is

$$S_1 + S_2 = \{\mathbf{x}_1 + \mathbf{x}_2 \mid \mathbf{x}_1 \in S_1, \mathbf{x}_2 \in S_2\}$$

with $S_1 - S_2$ defined similarly. The sum of a point $\mathbf{x}_0$ and set S is

$$\mathbf{x}_0 + S = \{\mathbf{x}_0 + \mathbf{x} \mid \mathbf{x} \in S\}.$$

The product of sets $\Lambda \subseteq \mathbf{R}$ and S is

$$\Lambda S = \{\lambda \mathbf{x} \mid \lambda \in \Lambda, \mathbf{x} \in S\}.$$

For a scalar $\lambda \in \mathbf{R}$ and set S,

$$\lambda S = \{\lambda \mathbf{x} \mid \mathbf{x} \in S\}.$$

Definition 1 (Unit ball) *The unit ball is the set*

$$B = \{\mathbf{x} \mid \|\mathbf{x}\| \leq 1\}.$$

Definition 2 (Interior) *The interior of a set S, denoted $\operatorname{int} S$, is the set of all points $\mathbf{x}$ for which $\mathbf{x} + \epsilon B \subset S$ for some $\epsilon > 0$.*

If $\mathbf{x} \in \operatorname{int} S$, we call S a neighborhood of $\mathbf{x}$.

Definition 3 (Limit) *A point $\mathbf{x}$ is the limit of the sequence of points $\mathbf{x}_1, \mathbf{x}_2, \ldots$, written $\mathbf{x}_i \to \mathbf{x}$, if $\|\mathbf{x}_i - \mathbf{x}\| \to 0$ as $i \to \infty$.*

Definition 4 (Closure) *The closure of S, written $\operatorname{cl} S$, is the set of limits of sequences of points in S.*

Equivalently, $\mathbf{x} \in \operatorname{cl} S$ if for all $\epsilon > 0$, there exists a $\mathbf{y} \in S$ such that $\mathbf{y} \in \mathbf{x} + \epsilon B$.

Definition 5 (Boundary) *The boundary of S is the set*

$$\operatorname{bd} S = \operatorname{cl} S \setminus \operatorname{int} S.$$

Equivalently, $\mathbf{x} \in \operatorname{bd} S$ if for all $\epsilon > 0$, there exist a $\mathbf{y} \in S$ and $\mathbf{z} \notin S$ such that $\mathbf{y}, \mathbf{z} \in \mathbf{x} + \epsilon B$.

Definition 6 (Open) *A set S is open if* $\operatorname{int} S = S$.

Equivalently, S is open if $S \cap \operatorname{bd} S = \emptyset$.

Definition 7 (Closed) *A set S is closed if* $\operatorname{cl} S = S$.

Equivalently, S is closed if $\operatorname{bd} S \subseteq S$, or if $\mathbf{E} \setminus S$ is open. The intersection of any family of closed sets is closed. The union of any family of finitely many closed sets is closed.

Definition 8 (Bounded) *A set S is bounded if $S \subset kB$ for some $k > 0$.*

A closed and bounded set is called compact.

1.1.2 Functions

A simple but very useful trick in convex analysis is to allow functions to take on values on the extended real line,

$$\overline{\mathbf{R}} = \mathbf{R} \cup \{-\infty, +\infty\}.$$

Arithmetic on the extended real line works like it does on the real line, except that we define $+\infty + (-\infty) = +\infty$ and $0(\pm\infty) = (\pm\infty)0 = 0$.

For sets other than the extended reals, the notation $f : S_1 \to S_2$ follows the standard mathematical conventions. It means that f maps each element of S_1 to some element of S_2. The domain of f is S_1, and the range of f is a subset of S_2. For example, a function $f : \mathbf{E} \to \mathbf{R}$ is defined on all of $\mathbf{E}$.

For the extended reals, the notation $f : S \to \overline{\mathbf{R}}$ means that f maps every element of S to some element of $\overline{\mathbf{R}}$ (possibly $\pm\infty$). For functions that map to the extended reals, we define the domain as follows.

Definition 9 (Domain) *The domain of $f : \mathbf{E} \to \overline{\mathbf{R}}$ is the set*

$$\operatorname{dom} f = \{\mathbf{x} \mid f(\mathbf{x}) < +\infty\}.$$

Definition 10 (Continuous) *A function $\mathbf{f} : \mathbf{E} \to \mathbf{F}$ is continuous at $\bar{\mathbf{x}} \in \mathbf{E}$ if for any $\epsilon > 0$, there exists a $\delta > 0$ such that for all $\mathbf{x}$,*

$$\|\mathbf{x} - \bar{\mathbf{x}}\| \leq \delta \qquad \Longrightarrow \qquad \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\bar{\mathbf{x}})\| \leq \epsilon.$$

A function $f : S \to \mathbf{R}$ is continuous at $\mathbf{x} \in S$ if $f(\mathbf{x}_i) \to f(\mathbf{x})$ for any sequence $\mathbf{x}_i \to \mathbf{x}$ in S. If f is continuous at every $\mathbf{x} \in S$, we say f is continuous on S.

1.1.3 Suprema and infima

For a finite set $S \subseteq \mathbf{R}$ (meaning a set containing a finite number of real numbers), $\max S$ is the largest element of S and $\min S$ is the smallest. The supremum $\sup S$ and infimum $\inf S$ generalize the maximum and minimum for sets that aren't finite. The supremum and infimum are also known as the greatest lower bound and least upper bound, respectively.

Definition 11 (Upper bound) *A scalar u is an upper bound on $S \subseteq \mathbf{R}$ if for all $x \in S$, $x \leq u$.*

Let U be the set of upper bounds on S. There are three possibilities:

1. $U = \mathbf{R}$, in which case S is empty and $\sup S = -\infty$,

2. $U = \emptyset$, in which case S is unbounded above and $\sup S = +\infty$, or

3. $U = [a, \infty)$ for some $a \in \mathbf{R}$ and $\sup S = a$.

Definition 12 (Lower bound) *A scalar l is a lower bound on $S \subseteq \mathbf{R}$ if for all $x \in S$, $x \geq l$.*

Let L be the set of lower bounds on S. As with U, there are three possibilities:

1. $L = \mathbf{R}$, in which case S is empty and $\inf S = +\infty$,

2. $L = \emptyset$, in which case S is unbounded below and $\inf S = -\infty$, or

3. $L = (-\infty, b]$ for some $b \in \mathbf{R}$ and $\inf S = b$.

If $\sup S \in S$, we say the supremum of S is attained. Similarly, if $\inf S \in S$, we say the infimum of S is attained.

The supremum and infimum of a function $f : \mathbf{E} \to \overline{\mathbf{R}}$ are

$$\begin{aligned}\sup_{\mathbf{x}\in\mathbf{E}} f(\mathbf{x}) &= \sup\left\{f(\mathbf{x}) \mid \mathbf{x} \in \operatorname{dom} f\right\} \\ \inf_{\mathbf{x}\in\mathbf{E}} f(\mathbf{x}) &= \inf\left\{f(\mathbf{x}) \mid \mathbf{x} \in \operatorname{dom} f\right\}.\end{aligned}$$

The supremum and infimum of a function can be infinite: if $\operatorname{dom} f = \emptyset$, then $\sup_{\mathbf{x}\in\mathbf{E}} f(\mathbf{x}) = \sup \emptyset = -\infty$ and $\inf_{\mathbf{x}\in\mathbf{E}} f(\mathbf{x}) = \inf \emptyset = +\infty$. Similarly, if f is unbounded above on $\operatorname{dom} f$, then $\sup_{\mathbf{x}\in\mathbf{E}} f(\mathbf{x}) = +\infty$, and if f is unbounded below on $\operatorname{dom} f$, then $\inf_{\mathbf{x}\in\mathbf{E}} f(\mathbf{x}) = -\infty$.

A few useful rules for suprema and infima of functions:

$$\begin{aligned}\sup_{\mathbf{x}\in\mathbf{E}} f(\mathbf{x}) &= -\inf_{\mathbf{x}\in\mathbf{E}} -f(\mathbf{x}) \\ \sup_{\mathbf{x}\in\mathbf{E}} f(-\mathbf{x}) &= \sup_{\mathbf{x}\in\mathbf{E}} f(\mathbf{x}) \\ \sup_{\mathbf{x}\in\mathbf{E}} \left\{ f(\mathbf{x}) + \sup_{\mathbf{y}\in\mathbf{F}} g(\mathbf{x},\mathbf{y}) \right\} &= \sup_{\mathbf{x}\in\mathbf{E},\mathbf{y}\in\mathbf{F}} \left\{f(\mathbf{x}) + g(\mathbf{x},\mathbf{y})\right\}.\end{aligned}$$

1.2 Convex sets

Definition 13 (Convex set) *A set S is convex if for all $\lambda \in [0,1]$,*

$$\mathbf{x}, \mathbf{y} \in S \quad \Longrightarrow \quad \lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in S.$$

Example: the closed halfspace

$$H = \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} \rangle \leq \beta\} = \{\mathbf{x} \mid \langle \mathbf{a}, \mathbf{x} - \mathbf{x}_0 \rangle \leq 0\} \tag{1.1}$$

is convex.

Proposition 1 *The intersection of any family of convex sets, possibly infinite in number, is convex.*

Examples:

- The intersection of finitely many halfspaces, called a polyhedron, is convex.
- The cone of positive semidefinite matrices,

$$\begin{aligned} \mathbf{S}^n_+ &\equiv \left\{ X \in \mathbf{S}^n \;\middle|\; \mathbf{y}^T X \mathbf{y} \geq 0 \text{ for all } \mathbf{y} \in \mathbf{R}^n \right\} \\ &= \left\{ X \in \mathbf{S}^n \;\middle|\; \operatorname{tr}(\mathbf{y}^T X \mathbf{y}) \geq 0 \text{ for all } \mathbf{y} \in \mathbf{R}^n \right\} \\ &= \left\{ X \in \mathbf{S}^n \;\middle|\; \operatorname{tr}(X \mathbf{y}\mathbf{y}^T) \geq 0 \text{ for all } \mathbf{y} \in \mathbf{R}^n \right\} \\ &= \bigcap_{\mathbf{y} \in \mathbf{R}^n} \left\{ X \in \mathbf{S}^n \;\middle|\; \langle X, \mathbf{y}\mathbf{y}^T \rangle \geq 0 \right\} \end{aligned}$$

  is convex.

Proposition 2 *The closure of a convex set is convex.*

Lemma 3 (Accessibility) *If a set S is convex, then for all $\lambda \in [0,1]$,*

$$\mathbf{x} \in \operatorname{int} S,\ \mathbf{y} \in \operatorname{cl} S \quad \Longrightarrow \quad \lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in \operatorname{int} S.$$

Corollary 4 *If S is nonempty and convex, then* $\operatorname{int} S$ *is convex and* $\operatorname{cl} S = \operatorname{cl}(\operatorname{int} S)$.

Theorem 5 (Best approximation) *If S is closed, nonempty and convex, then there exists a unique shortest vector $\bar{\mathbf{x}} \in S$ characterized by $\langle \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle \geq 0$ for all $\mathbf{x} \in S$.*

The proof uses the Weierstrass theorem (a continuous function attains its minimum over a compact set).

Theorem 6 (Basic separation) *If S is closed and convex and $\mathbf{y} \notin S$, then there exists a halfspace $H \supset S$ such that $\mathbf{y} \notin H$.*

The proof uses the best approximation theorem. This theorem implies that any convex set can be viewed as the intersection of the halfspaces that contain it.

Theorem 7 (Supporting hyperplane) *If S is convex and $\mathbf{y} \notin \operatorname{int} S$, then there exists a hyperplane $H \supset S$ with $\mathbf{y}$ on its boundary.*

The proof uses the accessibility lemma and the basic separation theorem.

1.3 Convex functions

Definition 14 (Epigraph) *The epigraph of a function $f : \mathbf{E} \to \overline{\mathbf{R}}$ is the set*

$$\operatorname{epi} f = \{(\mathbf{x}, t) \in \mathbf{E} \times \mathbf{R} \mid t \geq f(\mathbf{x})\}.$$

Definition 15 (Convex function) *A function $f : \mathbf{E} \to \overline{\mathbf{R}}$ is convex if* $\operatorname{epi} f$ *is convex.*

The classical definition of convexity considers functions $f : S \to \mathbf{R}$, where S is convex. Such a function is convex if for all $\mathbf{x}, \mathbf{y} \in S$ and $\lambda \in [0, 1]$,

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

and strictly convex if the inequality holds strictly whenever $\mathbf{x} \neq \mathbf{y}$ and $\lambda \in (0, 1)$.

The epigraph allows the convexity of a function to be analyzed in terms of the convexity of a set. The indicator function $\delta_S : \mathbf{E} \to \{0, +\infty\}$, defined by

$$\delta_S(\mathbf{x}) = \begin{cases} 0 & \mathbf{x} \in S \\ +\infty & \mathbf{x} \notin S, \end{cases}$$

maps in the other direction, allowing sets to be analyzed in terms of functions.

Examples: affine functions and norms are convex.

Definition 16 (Minimizer) *A point $\bar{\mathbf{x}}$ is a minimizer of $f : \mathbf{E} \to \overline{\mathbf{R}}$ if for all $\mathbf{x}$,*

$$f(\bar{\mathbf{x}}) \leq f(\mathbf{x}).$$

Proposition 8 *If a minimizer of a strictly convex function exists, it's unique.*

Definition 17 (Local minimizer) *A point $\bar{\mathbf{x}}$ is a local minimizer of f if there exists an $\epsilon > 0$ such that for all $\mathbf{x} \in \bar{\mathbf{x}} + \epsilon B$,*

$$f(\bar{\mathbf{x}}) \leq f(\mathbf{x}).$$

Lemma 9 *If $g : \mathbf{R} \to \overline{\mathbf{R}}$ is convex and $g(0) = 0$, then $t \mapsto g(t)/t$ is nondecreasing for all $t > 0$.*

The proof follows from the definition of a convex function.

Theorem 10 *If $f : \mathbf{E} \to \overline{\mathbf{R}}$ is convex and $f(\bar{\mathbf{x}}) \neq +\infty$, then $\bar{\mathbf{x}}$ is a minimizer of f if and only if $\bar{\mathbf{x}}$ is a local minimizer of f.*

This theorem makes convex optimization tractable. The proof uses lemma 9.

1.3.1 Recognizing convexity

Proposition 11 *If $h : (a, b) \to \mathbf{R}$ is differentiable, then h is convex if and only if dh/dx is nondecreasing.*

The proof uses lemma 9 and the mean value theorem.

Corollary 12 *If $h : (a, b) \to \mathbf{R}$ is twice differentiable, then*

(1) *h is convex if and only if $d^2h/dx^2 \geq 0$ on (a, b), and*

(2) *h is strictly convex if $d^2h/dx^2 > 0$ on (a, b).*

Note that the converse of the second claim is not true in general; consider $f(x) = x^4$, which is strictly convex but $\mathrm{d}^2h/\mathrm{d}x^2|_0 = 0$.

The next few results generalize this scalar convexity criterion.

Proposition 13 *Let S be open and $f : S \to \mathbf{R}$. At any $\mathbf{x} \in S$, there exists at most one $\mathbf{g} \in \mathbf{E}$ such that for all $\mathbf{y} \in \mathbf{E}$,*

$$f(\mathbf{x} + t\mathbf{y}) = f(\mathbf{x}) + t\langle \mathbf{g}, \mathbf{y} \rangle + o(t) \text{ as } t \downarrow 0.$$

The "little o" notation $h(t) = o(t)$ means that for all $\epsilon > 0$, there exists a $t > 0$ such that $|h(t)| \leq \epsilon t$. If $h(t) = o(t)$ holds as $t \downarrow 0$, then

$$\lim_{t \downarrow 0} \frac{h(t)}{t} = 0.$$

Thus, proposition 13 implies that if $\mathbf{g}$ exists at $\mathbf{x}$, then for all $\mathbf{y}$,

$$\lim_{t \downarrow 0} \frac{f(\mathbf{x} + t\mathbf{y}) - f(\mathbf{x}) - t\langle \mathbf{g}, \mathbf{y} \rangle}{t} = 0.$$

In proposition 13, the vector $\mathbf{g}$ is called the Gateaux derivative of f at $\mathbf{x}$. If the Gateaux derivative exists at every $\mathbf{x} \in \mathbf{E}$, then we call $\mathbf{g}$ the gradient of f at $\mathbf{x}$, $\nabla f(\mathbf{x})$. If $\nabla f : \mathbf{E} \to \mathbf{E}$ is continuous (denoted $\nabla f \in C$), then we call f continuously differentiable (denoted $f \in C^1$). If $\nabla f \in C$ and for all $\mathbf{y} \in \mathbf{E}$, the function $h(\mathbf{x}) = \langle \mathbf{y}, \nabla f(\mathbf{x}) \rangle$ is continuously differentiable, then we call f twice continuously differentiable (denoted $f \in C^2$).

If $f \in C^2$, then for any $\mathbf{y} \in \mathbf{E}$, we define

$$\nabla^2 f(\mathbf{x})[\mathbf{y}, \mathbf{y}] \equiv \langle \mathbf{y}, \nabla h(\mathbf{x}) \rangle.$$

If $\mathbf{E} = \mathbf{R}^n$, this reduces to the weighted 2-norm of $\mathbf{y}$,

$$\nabla^2 f(\mathbf{x})[\mathbf{y}, \mathbf{y}] = \mathbf{y}^T \nabla^2 f(\mathbf{x}) \mathbf{y}$$

where $\nabla^2 f(\mathbf{x})$ is the Hessian of f at $\mathbf{x}$.

Theorem 14 (Taylor) *If $f : S \to \mathbf{R}$ and $f \in C^2$, then for all $\mathbf{y} \in \mathbf{E}$,*

$$f(\mathbf{x}+t\mathbf{y}) = f(\mathbf{x}) + t\langle \nabla f(\mathbf{x}), \mathbf{y}\rangle + \frac{t^2}{2}\nabla^2 f(\mathbf{x})[\mathbf{y},\mathbf{y}] + o(t^2).$$

Corollary 15 *If $f : S \to \mathbf{R}$, where S is open and convex, and $f \in C^2$, then*

(1) *f is convex if and only if $\nabla^2 f(\mathbf{x})[\mathbf{y},\mathbf{y}] \geq 0$ for all $\mathbf{x} \in S$ and $\mathbf{y} \in \mathbf{E}$, and*

(2) *f is strictly convex if $\nabla^2 f(\mathbf{x})[\mathbf{y},\mathbf{y}] > 0$ for all $\mathbf{x} \in S$ and $\mathbf{y} \in \mathbf{E}$.*

Corollary 15 generalizes corollary 12. The proof uses Taylor's theorem and corollary 12 to show that $t \mapsto f(\mathbf{x}+t\mathbf{y})$ is convex.

As an example of the usefulness of this coordinate-free convexity criterion, consider the logarithmic barrier function $f : \mathbf{S}^n_{++} \to \mathbf{R}$, defined by

$$f(X) = -\ln(\det(X)).$$

Here $\mathbf{S}^n_{++}$ is the set of $n \times n$ positive definite matrices,

$$\mathbf{S}^n_{++} \equiv \left\{ X \in \mathbf{S}^n \;\middle|\; \mathbf{y}^T X \mathbf{y} > 0 \text{ for all } \mathbf{y} \neq \mathbf{0} \right\}.$$

It can be shown that $\mathbf{S}^n_{++}$ is closed and convex (and, as an aside, that $\mathbf{S}^n_{++} = \operatorname{int}(\mathbf{S}^n_{+})$), so corollary 15 applies.

Theorem 16 *The logarithmic barrier function is strictly convex.*

Sketch of proof:

1. Show that $\nabla f(X) = -X^{-1}$.

2. Show that $(I+tZ)^{-1} = I - tZ + \mathcal{O}(t^2)$ as $t \to 0$.

3. Observe that $\nabla^2 f(X)[Y,Y] = \operatorname{tr}((X^{-1}Y)^2)$, so it suffices to show that $\operatorname{tr}((X^{-1}Y)^2) > 0$.

4. Notice that since X (and hence X^{-1}) is positive definite, it has a Cholesky factorization. Hence, there exists an upper triangular, nonsingular $A \in \mathbf{R}^n n$ such that $X^{-1} = A^T A$. Use this to prove that $\operatorname{tr}((X^{-1}Y)^2) > 0$.

1.3.2 Lipschitz continuity

Definition 18 (Lipschitz) *A function $f : \mathbf{E} \to \overline{\mathbf{R}}$ is Lipschitz around $\bar{\mathbf{x}}$ if there exists an $L \geq 0$ such that for all $\mathbf{x}, \mathbf{y}$ in a neighborhood of $\bar{\mathbf{x}}$,*

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L \left\| \mathbf{x} - \mathbf{y} \right\| .$$

If f is Lipschitz at $\bar{\mathbf{x}}$, we call L the Lipschitz constant.

Lemma 17 *If $f : \mathbf{E} \to \overline{\mathbf{R}}$ is convex and $f(\bar{\mathbf{x}})$ is finite, then f is Lipschitz around $\bar{\mathbf{x}}$ if and only if f is bounded above on $\bar{\mathbf{x}} + \epsilon B$ for some $\epsilon > 0$.*

Although lemma 17 certifies the local Lipschitz continuity of a wide class of convex functions, it's still weaker than we'd like. For example, it's tempting to think that if $f : \mathbf{E} \to \overline{\mathbf{R}}$ is convex and finite at $\bar{\mathbf{x}}$ and $\mathbf{x}_r \to \bar{\mathbf{x}}$, then $f(\mathbf{x}_r) \to f(\bar{\mathbf{x}})$. This is not true in general, however. As a counterexample, consider the function $f : \mathbf{R}^2 \to \mathbf{R}$, defined by

$$f(\mathbf{x}) = \begin{cases} x_1^2/x_2 & x_2 > 0 \\ 0 & \mathbf{x} = \mathbf{0} \\ +\infty & \text{otherwise.} \end{cases} \tag{1.2}$$

It can be shown that $f(\mathbf{x})$ is not only convex, but also has a closed epigraph and is otherwise well-behaved. Note that $f(\mathbf{x}_r) = 1$ for all $r > 0$, where $\mathbf{x}_r = \begin{bmatrix} 1/r, & 1/r^2 \end{bmatrix}^T \to \mathbf{0}$ as $r \to +\infty$, but $f(\mathbf{0}) = 0$.

Incidentally, the function defined in (1.2) illustrates many of the pathological behaviors that convex functions can exhibit. It's a useful counterexample in many proofs.

Proposition 18 *If $f : S \to \mathbf{R}$ is convex, $\mathbf{x}_1, \dots, \mathbf{x}_m \in S$, $\boldsymbol{\lambda} \geq \mathbf{0}$, and $\mathbf{1}^T \boldsymbol{\lambda} = 1$, then*

$$\sum_{i=1}^m \lambda_i \mathbf{x}_i \in S \text{ and } f\left(\sum_{i=1}^m \lambda_i \mathbf{x}_i \right) \leq \sum_{i=1}^m \lambda_i f(\mathbf{x}_i).$$

The proof is by induction, with the base case $m = 2$ given by the definition of convexity.

Definition 19 (Unit simplex) *The unit simplex in $\mathbf{R}^n$ is the set*

$$\Delta_n = \left\{ \mathbf{x} \in \mathbf{R}^n \mid \mathbf{x} \geq \mathbf{0}, \ \mathbf{1}^T \mathbf{x} \leq 1 \right\} .$$

Proposition 19 *Any convex $f : \Delta_n \to \mathbf{R}$ is bounded above, and therefore Lipschitz around any point on the interior of Δ_n.*

Theorem 20 *If S and $f : S \to \mathbf{R}$ are convex, then f is Lipschitz around any point on the interior of S.*

The proof involves fitting a tiny copy of the unit simplex around any point $\bar{\mathbf{x}} \in \operatorname{int} S$.

1.3.3 Subgradients

Definition 20 (Subgradient) *Let $f : \mathbf{E} \to \overline{\mathbf{R}}$ be convex and finite at $\bar{\mathbf{x}}$. If $\bar{\mathbf{x}}$ minimizes $f(\mathbf{x}) - \langle \mathbf{y}, \mathbf{x} \rangle$, then we call $\mathbf{y}$ a subgradient of f at $\bar{\mathbf{x}}$ and write $\mathbf{y} \in \partial f(\bar{\mathbf{x}})$.*

In other words, $\mathbf{y}$ is a subgradient of f at $\bar{\mathbf{x}}$ if for all $\mathbf{x}$,

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

The set $\partial f(\bar{\mathbf{x}})$ is called the subdifferential of f at $\bar{\mathbf{x}}$. The notation $\mathbf{y} \in \partial f(\bar{\mathbf{x}})$ is intended to suggest that $\mathbf{y}$ is some kind of derivative. In fact, if f is differentiable at $\bar{\mathbf{x}}$, then $\partial f(\bar{\mathbf{x}}) = \{\nabla f(\bar{\mathbf{x}})\}$. Subgradients generalize the gradient.

Theorem 21 *If $f : \mathbf{E} \to \overline{\mathbf{R}}$ is convex, $\bar{\mathbf{x}} \in \operatorname{int}(\operatorname{dom} f)$, and $f(\bar{\mathbf{x}})$ is finite, then f*

(1) *never takes the value $-\infty$,*

(2) *is Lipschitz around $\bar{\mathbf{x}}$, and*

(3) *has a subgradient at $\bar{\mathbf{x}}$.*

Theorem 21 is a key concept in convex duality theory, and a fundamental tool in convex optimization algorithms. The proof uses the supporting hyperplane theorem.

Chapter 2

Duality

2.1 The Fenchel framework

2.1.1 Fenchel conjugates

Definition 21 (Conjugate) *For any $f : \mathbf{E} \to \overline{\mathbf{R}}$ and $\mathbf{z} \in \mathbf{E}$, the conjugate of f at $\mathbf{z}$ is*

$$f^*(\mathbf{z}) = \sup_{\mathbf{x}\in\mathbf{E}} \left\{\langle \mathbf{z}, \mathbf{x}\rangle - f(\mathbf{x})\right\}.$$

The epigraph of the conjugate of f at $\mathbf{z}$ is

$$\operatorname{epi} f^*(\mathbf{z}) = \bigcap_{\mathbf{x}\in\mathbf{E}} \left\{(\mathbf{z}, t) \mid \langle \mathbf{z}, \mathbf{x}\rangle - t \leq f(\mathbf{x})\right\}.$$

For fixed $\mathbf{x}$, $\{(\mathbf{z}, t) \mid \langle \mathbf{z}, \mathbf{x}\rangle - t \leq f(\mathbf{x})\}$ is a closed halfspace with normal $\mathbf{a} = \begin{bmatrix}\mathbf{x}^T & -1\end{bmatrix}^T$ and offset $\beta = f(\mathbf{x})$. Thus, $\operatorname{epi} f^*$ is the intersection of a family of closed halfspaces, hence closed and, by proposition 1, convex.

Example. Consider the entropy function $f : \mathbf{R} \to \overline{\mathbf{R}}$, defined by

$$f(s) = \begin{cases} +\infty & s < 0 \\ 0 & s = 0 \\ s\ln(s) - s & s > 0. \end{cases}$$

The conjugate of the entropy function at t is

$$f^*(t) = e^t.$$

It can also be shown that

$$f^{**}(s) = f(s)$$

where f^{**} is called the **biconjugate** of f. This is a special case of a more general result.

Definition 22 (Closed) *A function $f : \mathbf{E} \to \overline{\mathbf{R}}$ is closed if* $\operatorname{epi} f$ *is closed.*

Closed functions are sometimes called lower semicontinuous.

Theorem 22 *If $f : \mathbf{E} \to \overline{\mathbf{R}}$ is closed and convex and either always or never $-\infty$, then $f^{**} = f$.*

The proof uses the basic separation theorem 6, as well as the Fenchel duality theorem 24, which is proved independently.

Definition 23 (Proper) *A convex function $f : \mathbf{E} \to \overline{\mathbf{R}}$ is proper if f is never $-\infty$, and is not identically $+\infty$.*

By theorem 22, a closed proper convex function equals its biconjugate.

Proposition 23 (Fenchel-Young inequality) *Let $f : \mathbf{E} \to \overline{\mathbf{R}}$. For all $\mathbf{x}$ and $\mathbf{y}$,*

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

The Fenchel-Young inequality follows directly from definition 21. It holds with equality if and only if $\mathbf{y} \in \partial f(\mathbf{x})$.

2.1.2 Fenchel duality

Definition 24 (Adjoint) *For Euclidean spaces* $\mathbf{E}$ *and* $\mathbf{F}$ *and a linear map* $A : \mathbf{E} \to \mathbf{F}$*, there exists a unique linear map* $A^* : \mathbf{F} \to \mathbf{E}$*, called the adjoint of* A*, that satisfies* $\langle A\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, A^*\mathbf{y}\rangle$ *for all* $\mathbf{x} \in \mathbf{E}$*,* $\mathbf{y} \in \mathbf{F}$*.*

If $\mathbf{E} = \mathbf{R}^n$, $\mathbf{F} = \mathbf{R}^m$, and A is an $m \times n$ matrix, then the adjoint of A satisfies

$$\langle A\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, A^*\mathbf{y}\rangle \quad \Longleftrightarrow \quad \mathbf{x}^T A^T \mathbf{y} = \mathbf{x}^T A^* \mathbf{y}.$$

In other words, the adjoint of a matrix is its transpose.

Theorem 24 (Fenchel duality) *Consider Euclidean spaces* $\mathbf{E}$ *and* $\mathbf{F}$*, a linear map* $A : \mathbf{E} \to \mathbf{F}$*, and functions* $f : \mathbf{E} \to \overline{\mathbf{R}}$ *and* $g : \mathbf{F} \to \overline{\mathbf{R}}$*. If*

$$\begin{aligned} p &= \inf_{\mathbf{x}\in\mathbf{E}} \{f(\mathbf{x}) + g(A\mathbf{x})\} && \textit{(primal)} \\ d &= \sup_{\mathbf{y}\in\mathbf{F}} \{-f^*(A^*\mathbf{y}) - g^*(-\mathbf{y})\} && \textit{(dual)} \end{aligned}$$

then

(1) $p \geq d$.

If, in addition, f *and* g *are convex and*

$$\mathbf{0} \in \operatorname{int}(\operatorname{dom} g - A \operatorname{dom} f), \tag{2.1}$$

then

(2) $p = d$,

(3) *if* d *is finite, then it is attained, and*

(4) $\bar{\mathbf{x}}$ *and* $\bar{\mathbf{y}}$ *are primal-dual optimal if and only if* $A^*\bar{\mathbf{y}} \in \partial f(\bar{\mathbf{x}})$ *and* $-\bar{\mathbf{y}} \in \partial g(A\bar{\mathbf{x}})$*.*

Theorem 24 is the central result in linearly-constrained convex optimization. Result 1 is analogous to weak duality in linear programming. Results 2-4 are analogous to strong duality: 2 says that the primal is infeasible ($p = +\infty$) if and only if the dual is unbounded ($d = +\infty$); and that the primal is unbounded ($p = -\infty$) if and only if the dual is infeasible ($d = -\infty$). Result 3 says that if the dual is feasible and bounded (hence by result 2, so is the primal), then the optimal value is attained. Result 4 is analogous to complementary slackness.

Proof: Weak duality ($p \geq d$) follows from definitions 21 and 24, so we'll prove the strong duality result, $p = d$.

The proof technique is "perturbational duality," where we embed the problem in a more general one, then perturb it slightly to generate the desired result. Define the value function

$$v(\mathbf{z}) = \inf_{\mathbf{x}\in\mathbf{E}} (f(\mathbf{x}) + g(A\mathbf{x} + \mathbf{z}))$$

which is a perturbation of the primal, with $p = v(\mathbf{0})$. We'd like to apply theorem 21 to v, which requires showing that v is convex and finite at some $\bar{\mathbf{z}} \in \operatorname{int}(\operatorname{dom} v)$. We'll set $\bar{\mathbf{z}} = \mathbf{0}$, and use information about p to show that $\mathbf{0} \in \operatorname{int}(\operatorname{dom} v)$ and $v(\mathbf{0})$ is finite.

To show that $\mathbf{0} \in \operatorname{int}(\operatorname{dom} v)$, observe that $v(\mathbf{z}) = +\infty$ if and only if $\{f(\mathbf{x}) + g(A\mathbf{x} + \mathbf{z})\} = \emptyset$, meaning no $\mathbf{x} \in \operatorname{dom} f$ satisfies $A\mathbf{x} + \mathbf{z} \in \operatorname{dom} g$. Thus,

$$\begin{aligned}
\operatorname{dom} v &= \{\mathbf{z} \in \mathbf{F} \mid v(\mathbf{z}) < +\infty\} \\
&= \{\mathbf{z} \in \mathbf{F} \mid \exists\ \mathbf{x} \in \operatorname{dom} f \text{ s.t. } A\mathbf{x} + \mathbf{z} \in \operatorname{dom} g\} \\
&= \{\mathbf{y} - A\mathbf{x} \mid \mathbf{y} \in \operatorname{dom} g,\ \mathbf{x} \in \operatorname{dom} f\} \\
&= \operatorname{dom} g - A \operatorname{dom} f.
\end{aligned}$$

By the regularity condition (2.1), therefore, $\mathbf{0} \in \operatorname{int}(\operatorname{dom} v)$.

To show that $v(\mathbf{0})$ is finite, observe that $\mathbf{0} \in \operatorname{int}(\operatorname{dom} v)$ implies that $v(\mathbf{0}) < +\infty$. If $p = -\infty$, then $d \leq p$ implies that $p = d$, and we're done. So we need only consider the case where $p > -\infty$, meaning $v(\mathbf{0})$ is finite.

We can now apply theorem 21 to v, which states that v has a subgradient at $\mathbf{0}$. Let $-\mathbf{y}$ be a subgradient of v at $\mathbf{0}$. Then the Fenchel-Young inequality 23 holds with equality, *i.e.*, $v(\mathbf{0}) + v^*(-\mathbf{y}) = \langle \mathbf{0}, -\mathbf{y}\rangle$. Thus,

$$\begin{aligned}
p = v(\mathbf{0}) = -v^*(-\mathbf{y}) &= -\sup_{\mathbf{z}\in\mathbf{F}} \{\langle \mathbf{z}, -\mathbf{y}\rangle - v(\mathbf{z})\} \\
&= -\sup_{\mathbf{z}\in\mathbf{F}} \left\{ \langle \mathbf{z}, -\mathbf{y}\rangle - \inf_{\mathbf{x}\in\mathbf{E}} \{f(\mathbf{x}) + g(A\mathbf{x} + \mathbf{z})\} \right\} \\
&= -\sup_{\mathbf{z}\in\mathbf{F}} \left\{ \langle \mathbf{z}, -\mathbf{y}\rangle + \sup_{\mathbf{x}\in\mathbf{E}} \{-f(\mathbf{x}) - g(A\mathbf{x} + \mathbf{z})\} \right\} \\
&= -\sup_{\mathbf{x}\in\mathbf{E},\mathbf{z}\in\mathbf{F}} \{\langle \mathbf{z}, -\mathbf{y}\rangle - f(\mathbf{x}) - g(A\mathbf{x} + \mathbf{z})\} \\
&= -\sup_{\mathbf{x}\in\mathbf{E},\mathbf{z}\in\mathbf{F}} \{\langle A\mathbf{x}, \mathbf{y}\rangle + \langle A\mathbf{x} + \mathbf{z}, -\mathbf{y}\rangle - f(\mathbf{x}) - g(A\mathbf{x} + \mathbf{z})\} \\
&= -\sup_{\mathbf{x}\in\mathbf{E}} \left\{ \langle A\mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x}) + \sup_{\mathbf{z}\in\mathbf{F}} \{\langle -\mathbf{y}, A\mathbf{x} + \mathbf{z}\rangle - g(A\mathbf{x} + \mathbf{z})\} \right\} \\
&= -\sup_{\mathbf{x}\in\mathbf{E}} \{\langle A\mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x}) + g^*(-\mathbf{y})\} \\
&= -\sup_{\mathbf{x}\in\mathbf{E}} \{\langle A^*\mathbf{y}, \mathbf{x}\rangle - f(\mathbf{x})\} - g^*(-\mathbf{y}) \\
&= -f^*(A^*\mathbf{y}) - g^*(-\mathbf{y}) \\
&\leq d.
\end{aligned}$$

By weak duality, $p \geq d$, so the last line must hold with equality. Thus, if condition (2.1) holds, then $p = d$ and $-f^*(A^*\mathbf{y}) - g^*(\mathbf{y}) = d$, *i.e.*, $\mathbf{y}$ attains d.

□

Interpretation of $p = d$. The idea that the duality gap is zero is equivalent to saying that at $\mathbf{z} = \mathbf{0}$, the value function is equal to its biconjugate. To see this, note that in the proof of Fenchel duality we showed that for any $-\mathbf{y} \in \partial v(\mathbf{0})$,

$$-v^*(-\mathbf{y}) = -f^*(A^*\mathbf{y}) - g^*(-\mathbf{y}).$$

The right-hand side is exactly the dual objective function, so the dual can be written as

$$\begin{aligned} d &= \sup_{\mathbf{y}\in\mathbf{F}} \{-v^*(-\mathbf{y})\} \\ &= \sup_{\mathbf{y}\in\mathbf{F}} \{-v^*(\mathbf{y})\} \\ &= \sup_{\mathbf{y}\in\mathbf{F}} \{\langle \mathbf{0}, \mathbf{y}\rangle - v^*(\mathbf{y})\} \\ &= v^{**}(\mathbf{0}). \end{aligned}$$

Since $v(\mathbf{0}) = p$, therefore, $p = d$ if and only if $v(\mathbf{0}) = v^{**}(\mathbf{0})$.

Regularity condition. The regularity condition (2.1) (sometimes called a constraint qualification) may be difficult to check. It can be shown that (2.1) holds if either of the following stronger conditions are satisfied:

- there exists an $\hat{\mathbf{x}} \in \operatorname{dom} f$ such that $A\hat{\mathbf{x}} \in \operatorname{int}(\operatorname{dom} g)$, or
- A is surjective, and there exists an $\hat{\mathbf{x}} \in \operatorname{int}(\operatorname{dom} f)$ such that $A\hat{\mathbf{x}} \in \operatorname{dom} g$.

The proof that the first condition implies (2.1) is straightforward, but the second condition requires the following theorem.

Theorem 25 (Open mapping) *A surjective linear $A : \mathbf{E} \to \mathbf{F}$ maps open sets to open sets.*

The proof of the open mapping theorem uses the singular value decomposition and the fact that any matrix representation of a surjective linear map has full row rank.

2.1.3 Subdifferential calculus

Fenchel duality gives some insight into the calculus of subdifferentials. For instance, recall from vector calculus that for continuously differentiable $f, g : \mathbf{E} \to \mathbf{R}$, $\nabla(f + g)(\mathbf{x}) = \nabla f(\mathbf{x}) + \nabla g(\mathbf{x})$ and $\nabla(g \circ A)(\mathbf{x}) = A^*\nabla g(A\mathbf{x})$. (Notation: $(f + g)(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ and $(g \circ A)(\mathbf{x}) = g(A\mathbf{x})$.) Fenchel duality proves the following analog for subdifferentials:

Theorem 26 *For $f : \mathbf{E} \to \overline{\mathbf{R}}$, $g : \mathbf{F} \to \overline{\mathbf{R}}$ and affine $A : \mathbf{E} \to \mathbf{F}$,*

$$\partial(f + g \circ A)(\bar{\mathbf{x}}) \supseteq \partial f(\bar{\mathbf{x}}) + A^*\partial g(A\bar{\mathbf{x}}).$$

The above holds with equality under the Fenchel duality conditions.

2.2 Conic programming

Definition 25 (Convex hull) *The convex hull of S is*

$$\operatorname{conv}(S) = \left\{ \sum_{i=1}^{m} \lambda_i \mathbf{x}_i \;\middle|\; m \in \mathbf{N},\ \mathbf{1}^T \boldsymbol{\lambda} = 1,\ \boldsymbol{\lambda} \geq \mathbf{0},\ \mathbf{x}_i \in S \right\}.$$

Proposition 27 *The smallest convex set containing S is* $\operatorname{conv}(S)$.

Proposition 28 *If S is finite, then* $\operatorname{conv}(S)$ *is compact.*

Corollary 29 *If S is compact, then* $\operatorname{conv}(S)$ *is compact.*

Definition 26 (Cone) *A set K is a cone if $\mathbf{0} \in K$ and for all $\alpha \geq 0$ and $\mathbf{x} \in K$, $\alpha\mathbf{x} \in K$.*

A cone K is convex if and only if for all $\mathbf{x}, \mathbf{y} \in K$, $\mathbf{x} + \mathbf{y} \in K$. The only difference between a convex cone and a subspace, therefore, is that a convex cone only needs to be closed under vector addition and *nonnegative* scalar multiplication. A subspace is closed under general scalar multiplication.

Examples: the nonnegative orthant $\mathbf{R}^n_+$, the cone of symmetric positive definite matrices $\mathbf{S}^n_+$, and the second-order cone

$$\mathbf{SO}^n_+ = \{(\mathbf{x}, t) \in \mathbf{R}^n \times \mathbf{R} \mid t \geq \|\mathbf{x}\|_2\}$$

are all convex cones. Note that $\mathbf{SO}^n_+$ is the epigraph of the 2-norm on $\mathbf{R}^n$.

Definition 27 (Dual cone) *For any set K, the dual cone of K is*

$$K^+ = \{\mathbf{y} \mid \langle \mathbf{x}, \mathbf{y} \rangle \geq 0 \textit{ for all } \mathbf{x} \in K\}.$$

All dual cones are convex. The nonnegative orthant, semidefinite cone, and second-order cone are all self dual: $(\mathbf{R}^n_+)^+ = \mathbf{R}^n_+$, $(\mathbf{S}^n_+)^+ = \mathbf{S}^n_+$, and $(\mathbf{SO}^n_+)^+ = \mathbf{SO}^n_+$.

Theorem 30 (Conic duality) *Consider Euclidean spaces $\mathbf{E}$ and $\mathbf{F}$, a linear map $A : \mathbf{E} \to \mathbf{F}$, vectors $\mathbf{c} \in \mathbf{E}$ and $\mathbf{b} \in \mathbf{F}$, and cones $K \subset \mathbf{E}$ and $H \subset \mathbf{F}$. If p and d are the values of the primal-dual pair*

$$\begin{array}{ll} \text{minimize} & \langle \mathbf{c}, \mathbf{x} \rangle \\ \text{subject to} & A\mathbf{x} \in \{\mathbf{b} + \mathbf{y} \mid \mathbf{y} \in H\} \\ & \mathbf{x} \in K \end{array} \qquad \begin{array}{ll} \text{maximize} & \langle \mathbf{b}, \mathbf{y} \rangle \\ \text{subject to} & A^*\mathbf{y} \in \{\mathbf{c} - \mathbf{x} \mid \mathbf{x} \in K^+\} \\ & \mathbf{y} \in K^+ \end{array}$$

then

(1) $p \geq d$.

If, in addition, K and H are convex and either of the following hold:

(a) *there exists an* $\hat{\mathbf{x}} \in K$ *such that* $A\hat{\mathbf{x}} \in \mathbf{b} + \operatorname{int} H$*, or*

(b) A *is surjective and there exists an* $\hat{\mathbf{x}} \in \operatorname{int}(K)$ *such that* $A\hat{\mathbf{x}} \in \mathbf{b} + H$*,*

then

(2) $p = d$*,*

(3) *if* d *is finite, then it is attained, and*

(4) $\bar{\mathbf{x}}$ *and* $\bar{\mathbf{y}}$ *are primal-dual optimal if and only if* $\langle \bar{\mathbf{y}}, A\bar{\mathbf{x}} - \mathbf{b} \rangle = 0$ *and* $\langle \bar{\mathbf{x}}, \mathbf{c} - A^*\bar{\mathbf{y}} \rangle = 0$.

The proof of conic duality is by reduction to Fenchel duality using the functions

$$f(\mathbf{x}) = \begin{cases} \langle \mathbf{c}, \mathbf{x} \rangle & \mathbf{x} \in K \\ +\infty & \text{otherwise} \end{cases}$$

and

$$g(\mathbf{y}) = \begin{cases} 0 & \mathbf{y} - \mathbf{b} \in H \\ +\infty & \text{otherwise.} \end{cases}$$

2.3 The Lagrangian framework

Consider a convex set $U \subset \mathbf{E}$, convex functions $f, g_1, \ldots, g_m : U \to \mathbf{R}$, the primal

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & g_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m \\ & \mathbf{x} \in U \end{array} \tag{2.2}$$

and the Lagrangian

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}).$$

Note that the primal is equivalent to

$$p = \inf_{\mathbf{x} \in U} \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$$

since the supremum is $+\infty$ if $g_i(\mathbf{x}) > 0$ for any i. The Lagrangian dual is

$$\begin{aligned} d &= \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \inf_{\mathbf{x} \in U} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \\ &= -\inf_{\boldsymbol{\lambda} \geq \mathbf{0}} \left\{ -\inf_{\mathbf{x} \in U} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \right\} \\ &= -\inf_{\boldsymbol{\lambda}} \psi(\boldsymbol{\lambda}) \end{aligned}$$

where

$$\psi(\boldsymbol{\lambda}) = \begin{cases} -\inf_{\mathbf{x} \in U} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) & \boldsymbol{\lambda} \geq \mathbf{0} \\ +\infty & \text{otherwise} \end{cases}$$

is called the **dual function**.

It's easy to show that weak duality holds, *i.e.*, that $p \geq d$. The duality gap can be nonzero, however. To see this, consider

$$p = \inf_{\mathbf{x} \in \mathbf{R}^2} \left\{ e^{x_2} \;\middle|\; \sqrt{x_1^2 + x_2^2} - x_1 \leq 0 \right\}.$$

Here $p = 1$ and $d = 0$.

Definition 28 (Slater point) *A Slater point $\hat{\mathbf{x}} \in U$ satisfies $g_i(\hat{\mathbf{x}}) < 0$ for all i.*

Theorem 31 (Lagrangian duality) *If a Slater point exists, then $p = d$ and d is attained if finite.*

Proof: Like Fenchel duality, Lagrangian duality is proved by a perturbation argument. In fact, Lagrangian duality can be viewed as a special case of Fenchel duality, but it's more natural to prove it directly.

We start by defining the value function

$$v(\mathbf{z}) = \inf_{\mathbf{x}\in U} \{f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{z}\}.$$

It's easy to check that v is convex and that $v(\mathbf{0}) = p$. The Fenchel conjugate of v is

$$\begin{aligned}
v^*(-\boldsymbol{\lambda}) &= \sup_{\mathbf{z}} \left\{-\boldsymbol{\lambda}^T\mathbf{z} - v(\mathbf{z})\right\} \\
&= \sup_{\mathbf{z}} \left\{-\boldsymbol{\lambda}^T\mathbf{z} + \sup_{\mathbf{x}\in U} \{-f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{z}\}\right\} \\
&= \sup_{\mathbf{x}\in U,\mathbf{z}} \left\{-\boldsymbol{\lambda}^T\mathbf{z} - f(\mathbf{x}) \;\middle|\; \mathbf{g}(\mathbf{x}) \leq \mathbf{z}\right\} \\
&= \sup_{\mathbf{w}\geq\mathbf{0},\mathbf{x}\in U,\mathbf{z}} \left\{-\boldsymbol{\lambda}^T\mathbf{z} - f(\mathbf{x}) \;\middle|\; \mathbf{g}(\mathbf{x}) + \mathbf{w} = \mathbf{z}\right\} \\
&= \sup_{\mathbf{w}\geq\mathbf{0},\mathbf{x}\in U} \left\{-\boldsymbol{\lambda}^T(\mathbf{g}(\mathbf{x}) + \mathbf{w}) - f(\mathbf{x})\right\} \\
&= \begin{cases} \sup_{\mathbf{x}\in U} \left\{-\boldsymbol{\lambda}^T\mathbf{g}(\mathbf{x}) - f(\mathbf{x})\right\} & \boldsymbol{\lambda} \geq \mathbf{0} \\ +\infty & \text{otherwise} \end{cases} \\
&= \begin{cases} -\inf_{\mathbf{x}\in U} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) & \boldsymbol{\lambda} \geq \mathbf{0} \\ +\infty & \text{otherwise} \end{cases} \\
&= \psi(\boldsymbol{\lambda}).
\end{aligned}$$

From the definition of the dual,

$$d = -\inf_{\boldsymbol{\lambda}} \psi(\boldsymbol{\lambda}) = -\inf_{\boldsymbol{\lambda}} v^*(-\boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda}} -v^*(\boldsymbol{\lambda}) = v^{**}(\mathbf{0}),$$

so we have $p = v(\mathbf{0})$ and $d = v^{**}(\mathbf{0})$. To prove $p = d$, therefore, we only need to show that $v(\mathbf{0}) = v^{**}(\mathbf{0})$.

To see this, let $\hat{\mathbf{x}}$ be a Slater point, meaning $\hat{\mathbf{x}} \in U$ and $\mathbf{g}(\hat{\mathbf{x}}) < \mathbf{0}$. Then

$$\mathbf{0} \in \operatorname{int}(\operatorname{dom} v) = \operatorname{int} \{\mathbf{z} \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{z} \text{ for some } \mathbf{x} \in U\}.$$

By Theorem 21, therefore, v is never $-\infty$ and has a subgradient at $\mathbf{0}$. It can also be shown that v is closed, so Theorem 22 applies and $v = v^{**}$. Hence $p = d$.

To see that d is attained when finite, recall that $d = -\inf_{\boldsymbol{\lambda}} v^*(-\boldsymbol{\lambda})$. Let $-\bar{\boldsymbol{\lambda}}$ be a subgradient of v at $\mathbf{0}$. Then by the Fenchel-Young inequality,

$$v(\mathbf{0}) + v^*(-\bar{\boldsymbol{\lambda}}) = 0. \tag{2.3}$$

Since $v(\mathbf{0}) = p = d$, equation (2.3) implies that $d = -v^*(-\bar{\boldsymbol{\lambda}})$. Thus, $\bar{\boldsymbol{\lambda}}$ attains the infimum and is dual optimal.

□

Chapter 3

Algorithms

3.1 The alternating directions method of multipliers

The alternating direction method of multipliers (ADMM) is a popular algorithm for large-scale optimization of Fenchel-type problems. It's a splitting method, meaning it operates on f and g separately. There are many variants of ADMM; we'll discuss one.

Consider linear $A : \mathbf{E} \to \mathbf{F}$, convex $f : \mathbf{E} \to \overline{\mathbf{R}}$ and $g : \mathbf{F} \to \overline{\mathbf{R}}$, and the primal-dual pair

$$\begin{aligned} p &= \inf_{\mathbf{x}} \{f(\mathbf{x}) + g(A\mathbf{x})\} && \text{(primal)} \\ d &= \sup_{\mathbf{y}} \{-f^*(A^*\mathbf{y}) - g^*(-\mathbf{y})\} && \text{(dual)} \end{aligned}$$

Assume that p is finite, $p = d$ (a stronger condition than the results given by Fenchel duality), and that the primal and dual have optimal solutions $\mathbf{x}^*$ and $\mathbf{y}^*$, respectively, not necessarily unique. From Fenchel duality, a necessary and sufficient optimality condition is that $-A^*\mathbf{y}^* \in \partial f(\mathbf{x}^*)$ and $\mathbf{y}^* \in \partial g(A\mathbf{x}^*)$.

If we think of the primal as minimizing $f(\mathbf{x})+g(\mathbf{z})$ subject to $A\mathbf{x} = \mathbf{z}$, then it makes sense to define Lagrange multipliers (*i.e.*, dual variables) $\mathbf{y} \in \mathbf{F}$. We can also form the augmented Lagrangian $\mathcal{L}_\rho : \mathbf{E} \times \mathbf{F} \times \mathbf{F} \to \overline{\mathbf{R}}$, parameterized by $\rho > 0$, such that

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) + \langle \mathbf{y}, A\mathbf{x} - \mathbf{z} \rangle + \frac{\rho}{2} \|A\mathbf{x} - \mathbf{z}\|^2 .$$

Notice that

$$p = \inf_{\mathbf{x},\mathbf{z}} \sup_{\mathbf{y}} \mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y})$$

since the augmented Lagrangian is $+\infty$ whenever $A\mathbf{x} \neq \mathbf{z}$.

The ADMM algorithm is

```
given $\mathbf{x}_k$, $\mathbf{z}_k$, $\mathbf{y}_k$, and parameter $\rho > 0$
repeat
    1. $\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x}} \mathcal{L}_\rho(\mathbf{x}, \mathbf{z}_k, \mathbf{y}_k)$
    2. $\mathbf{z}_{k+1} = \operatorname{argmin}_{\mathbf{z}} \mathcal{L}_\rho(\mathbf{x}_{k+1}, \mathbf{z}, \mathbf{y}_k)$
    3. $\mathbf{y}_{k+1} = \mathbf{y}_k + \rho\mathbf{r}_{k+1}$
stop if $\mathbf{r}_{k+1}$ and $\mathbf{s}_{k+1}$ are small.
```

In the above,

$$\begin{aligned} \mathbf{r}_{k+1} &= \mathbf{z}_{k+1} - A\mathbf{x}_{k+1} \\ \mathbf{s}_{k+1} &= \rho A^*(\mathbf{z}_{k+1} - \mathbf{z}_k) \end{aligned}$$

are the primal and dual residuals, respectively. The norm of $\mathbf{r}_{k+1}$ measures the distance from constraint satisfaction, and the norm of $\mathbf{s}_{k+1}$ measures the distance to optimality. To see the latter, one can show that $\mathbf{y}_{k+1} \in \partial g(\mathbf{z}_{k+1})$ and $-A^*\mathbf{y}_{k+1} - \mathbf{s}_{k+1} \in \partial f(\mathbf{x}_{k+1})$, so $\mathbf{x}_{k+1}$

and $\mathbf{y}_{k+1}$ are optimal if and only if $\mathbf{r}_{k+1} = \mathbf{0}$ and $\mathbf{s}_{k+1} = \mathbf{0}$. From step 3, the parameter ρ can be interpreted as a step size. Large ρ will result in stricter enforcement of the primal constraints.

ADMM is *slow*. Rather than stopping when $\|\mathbf{r}_{k+1}\|$ and $\|\mathbf{s}_{k+1}\|$ are less than 10^{-6} like we might in a Newton-type algorithm, in ADMM we typically stop when the norms are less than 10^{-2} or so. The power of ADMM comes from the fact that step 1 only involves f, and step 2 only involves g. For many interesting problems, the $\mathbf{x}$ and $\mathbf{z}$ updates can be performed quickly.

Theorem 32 *Let $p_k = f(\mathbf{x}_k) + g(\mathbf{z}_k)$. If the ADMM assumptions hold, then*

$$\mathbf{r}_k \to \mathbf{0}, \quad \mathbf{s}_k \to \mathbf{0}, \quad p_k \to p.$$

The proof involves the monotonicity property $\langle \mathbf{y}_{k+1} - \mathbf{y}_k, \mathbf{z}_{k+1} - \mathbf{z}_k \rangle \geq 0$, which follows from the definition of subgradient. It involves showing that

$$V_k = \frac{1}{\rho}\|\mathbf{y}_k - \mathbf{y}^*\| + \rho\|\mathbf{z}_k - \mathbf{z}_*\|$$

is a Lyapunov function.

3.1.1 Proximity operators

The ADMM algorithm is only useful if the $\mathbf{x}$ and $\mathbf{z}$ updates in steps 1 and 2 are easy. Thankfully, for many interesting problems they are. This is because the updates often take the form

$$\operatorname{argmin}\left\{f + \frac{\rho}{2}\|\cdot - \mathbf{v}\|^2\right\} \tag{3.1}$$

This problem of minimizing a function plus the 2-norm of a translation is called a **proximity operator**. Proximity operators are easy to compute for many convex functions. The operator 3.1 is a smooth function of $\mathbf{v}$, so even if f and g are nonsmooth or extended valued, the proximity operator can be analyzed using smooth tools.

To see that steps 1 and 2 can be proximity operators, note that the augmented Lagrangian can be written as

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) + \frac{\rho}{2}\|A\mathbf{x} - \mathbf{z} + \mathbf{u}\|^2 - \frac{\rho}{2}\|\mathbf{u}\|^2,$$

where we have changed dual variables to $\mathbf{u} = \mathbf{y}/\rho$. When $A = I$, ADMM reduces to

```
given $\mathbf{x}_k$, $\mathbf{z}_k$, $\mathbf{u}_k$, and parameter $\rho > 0$
repeat
    1. $\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x}}\left\{f(\mathbf{x}) + (\rho/2)\|\mathbf{x} - (\mathbf{z}_k - \mathbf{u}_k)\|^2\right\}$
    2. $\mathbf{z}_{k+1} = \operatorname{argmin}_{\mathbf{z}}\left\{g(\mathbf{z}) + (\rho/2)\|\mathbf{z} - (\mathbf{x}_{k+1} + \mathbf{u}_k)\|^2\right\}$
```

3. $\mathbf{u}_{k+1} = \mathbf{u}_k + \mathbf{x}_{k+1} - \mathbf{z}_{k+1}$

stop if $\mathbf{r}_{k+1}$ and $\mathbf{s}_{k+1}$ are small.

Some common proximity operators follow.

Projection. If $f = \delta_C$, then the proximity operator is

$$\underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \delta_C(\mathbf{x}) + \frac{\rho}{2} \|\mathbf{x} - \mathbf{v}\|^2 \right\} = \underset{\mathbf{x} \in C}{\operatorname{argmin}} \frac{\rho}{2} \|\mathbf{x} - \mathbf{v}\|^2 = P_C(\mathbf{v})$$

where $P_C(\mathbf{v})$ is the projection of $\mathbf{v}$ onto C. It's easy to compute $P_C(\mathbf{v})$ when C is the nonnegative orthant, the semidefinite cone, any affine subspace, the second-order cone, or $\bar{\mathbf{x}} + kB$.

Quadratic forms on subspaces. The proximity operator is easy to compute if

$$f(\mathbf{x}) = \begin{cases} \frac{1}{2}\mathbf{x}^T P \mathbf{x} & \mathbf{x} \in L \\ +\infty & \text{else} \end{cases}$$

where L is a subspace.

The 1-norm. If

$$f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

then

$$\underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \|\mathbf{x}\|_1 + \frac{\rho}{2} \|\mathbf{x} - \mathbf{v}\|^2 \right\} = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} |x_i| + \frac{\rho}{2}(x_i - v_i)^2 \right\}$$

which can be computed separately for each x_i. The solution is the so-called soft-thresholding operator,

$$x_i = S_{1/\rho}(v_i) = \begin{cases} v_i + 1/\rho & v_i < -1/\rho \\ 0 & |v_i| \leq 1/\rho \\ v_i - 1/\rho & v_i > 1/\rho. \end{cases}$$

Soft thresholding can be done in linear time.

3.1.2 Applications

Set intersection. If

$$p = \inf_{\mathbf{x}} \{\delta_C(\mathbf{x}) + \delta_D(\mathbf{x})\}$$

then steps 1 and 2 are projection onto C and D, and ADMM reduces to Dykstra's alternating projection algorithm.

Linear programming. A linear program can be written

$$p = \inf_{\mathbf{x}} \left\{ \langle \mathbf{q}, \mathbf{x} \rangle + \delta_{\mathbf{b}+L}(\mathbf{x}) + \delta_{\mathbf{R}^n_+}(\mathbf{x}) \right\}$$

where L is a subspace. This can be solved by setting $f(\mathbf{x}) = \langle \mathbf{q}, \mathbf{x} \rangle + \delta_{\mathbf{b}+L}(\mathbf{x})$ and $g(\mathbf{x}) = \delta_{\mathbf{R}^n_+}(\mathbf{x})$ and applying ADMM. Similarly, linearly constrained quadratic programs can be solved by replacing $\langle \mathbf{q}, \mathbf{x} \rangle$ by $\mathbf{x}^T P \mathbf{x}$, where $P \in \mathbf{S}^n_+$.

Basis pursuit. The basis pursuit problem

$$p = \inf_{\mathbf{x}} \{ \|\mathbf{x}\|_1 + \delta_{\mathbf{b}+L}(\mathbf{x}) \}$$

can be solved for billions of variables, since soft thresholding can be done in linear time. On this problem, ADMM typically proceeds in two phases. It starts by discovering the correct basis (*i.e.*, which components of $\mathbf{x}$ should be zero), then it solves the linear system $\mathbf{x} \in \mathbf{b} + L$ in linear time. The first phase may take many iterations. ADMM basis pursuit resembles an active set method, where the work of the algorithm is finding the set of constraints that are binding at the solution.

LASSO regression. A popular statistics problem is

$$p = \inf_{\mathbf{x}} \left\{ \frac{\lambda}{2} \|A\mathbf{x} - \mathbf{b}\|^2 + \|\mathbf{x}\|_1 \right\}.$$

The 1-norm term encourages sparsity, so this problem is often used instead of ordinary least squares regression when there are many more predictors than responses ($n \gg m$, where $A \in \mathbf{R}^{m \times n}$) and a parsimonious model is sought.

Least absolute deviations. The problem

$$p = \inf_{\mathbf{x}} \{ \|A\mathbf{x} - \mathbf{b}\|_1 \}$$

can be solved by setting $f \equiv 0$ and $g(\mathbf{z}) = \|\mathbf{z} - \mathbf{b}\|_1$.

3.1.3 Example: consensus optimization

An example that gives nice intuition for ADMM, and particularly for its usefulness in distributed problems, is the consensus optimization problem

$$\inf_{\mathbf{z} \in \mathbf{F}} \sum_{i=1}^{N} f_i(\mathbf{z})$$

where N may be very large. It models agents $1, \ldots, N$, each with their local objective f_i, trying to agree on an acceptable value of $\mathbf{z}$.

To apply ADMM to consensus optimization, we introduce the slightly more general primal

$$p = \inf_{\mathbf{x}} \{f(\mathbf{x}) + g(\mathbf{z}) \mid A\mathbf{x} = B\mathbf{z}\}$$

where $A : \mathbf{E} \to \mathbf{G}$ and $B : \mathbf{F} \to \mathbf{G}$ are linear and $f : \mathbf{E} \to \overline{\mathbf{R}}$ and $g : \mathbf{F} \to \overline{\mathbf{R}}$ are convex. The convergence proof and algorithm follow through for this formulation, with two minor changes: the augmented Lagrangian is

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) + \langle \mathbf{y}, A\mathbf{x} - B\mathbf{z}\rangle + \frac{\rho}{2}\|A\mathbf{x} - B\mathbf{z}\|^2$$

and the $\mathbf{y}$ update is

$$\mathbf{y}_{k+1} = \mathbf{y}_k + \rho(A\mathbf{x}_{k+1} - B\mathbf{z}_{k+1}).$$

The residuals and stopping criterion generalize naturally.

The consensus optimization problem can be put into this form by setting $\mathbf{E} = \mathbf{G} = \mathbf{F}^N$ (the space of N-tuples $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, with $\mathbf{x}_i \in \mathbf{F}$), $A = I$, $g = 0$, $B\mathbf{z} = (\mathbf{z}, \dots, \mathbf{z})$, and $f(\mathbf{x}) = \sum_{i=1}^N f_i(\mathbf{x}_i)$. This gives the ADMM primal

$$p = \inf_{\mathbf{x} \in \mathbf{F}^N} \{f(\mathbf{x}) \mid \mathbf{x}_i = \mathbf{z} \text{ for all } i = 1, \dots, N\}.$$

Defining dual variables $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$, the augmented Lagrangian is

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) &= f(\mathbf{x}) + \sum_{i=1}^N \langle \mathbf{y}_i, \mathbf{x}_i - \mathbf{z}\rangle + \frac{\rho}{2}\|\mathbf{x} - (\mathbf{z}, \dots, \mathbf{z})\|^2 \\ &= \sum_{i=1}^N \left[f_i(\mathbf{x}_i) + \langle \mathbf{y}_i, \mathbf{x}_i - \mathbf{z}\rangle + \frac{\rho}{2}\|\mathbf{x}_i - \mathbf{z}\|^2\right]\end{aligned}$$

where we've used the fact that the inner product on $\mathbf{E} \times \mathbf{F}$ is $\langle(\mathbf{x}, \mathbf{y}), (\mathbf{u}, \mathbf{v})\rangle = \langle \mathbf{x}, \mathbf{u}\rangle + \langle \mathbf{y}, \mathbf{v}\rangle$, hence $\|(\mathbf{x}, \mathbf{y})\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$.

The ADMM algorithm is

```
given $\mathbf{x}_k$, $\mathbf{z}_k$, $\mathbf{y}_k$, and parameter $\rho > 0$
repeat
    1. $\mathbf{x}_i^{k+1} = \text{argmin}_{\mathbf{x}_i}\{f_i(\mathbf{x}_i)) + \langle \mathbf{y}_i^k, \mathbf{x}_i - \mathbf{z}_k\rangle + \rho\|\mathbf{x}_i - \mathbf{z}_k\|^2/2\ \}$ for all $i$
    2. solve $\nabla_{\mathbf{z}_{k+1}}\mathcal{L}_\rho(\mathbf{x}_{k+1}, \mathbf{z}_{k+1}, \mathbf{y}_k) = \mathbf{0}$ for $\mathbf{z}_{k+1}$
    3. $\mathbf{y}_i^{k+1} = \mathbf{y}_i^k + \rho(\mathbf{x}_i^{k+1} - \mathbf{z}_{k+1})$ for all $i$
```

with the appropriate stopping criterion. It's easy to show that step 2 is equivalent to

$$\mathbf{z}_{k+1} = \frac{1}{N}\sum_{i=1}^N \left(\mathbf{x}_i^{k+1} + \frac{1}{\rho}\mathbf{y}_i\right) = \bar{\mathbf{x}}_{k+1} + \frac{1}{\rho}\bar{\mathbf{y}}_k.$$

Step 3 implies that

$$\bar{\mathbf{y}}_{k+1} = \bar{\mathbf{y}}_k + \rho\left(\bar{\mathbf{x}}_{k+1} - \mathbf{z}_{k+1}\right) = \bar{\mathbf{y}}_k + \rho\left(-\frac{1}{\rho}\bar{\mathbf{y}}_k\right) = \mathbf{0}$$

so for all $k > 1$, $\mathbf{z}_{k+1} = \bar{\mathbf{x}}_{k+1}$. Changing variables to $\mathbf{u} = \mathbf{y}/\rho$, ADMM simplifies to

```
given $\mathbf{x}_k$, $\mathbf{u}_k$, and parameter $\rho > 0$
repeat
    1. $\mathbf{x}_i^{k+1} = \mathrm{argmin}_{\mathbf{x}_i}\{f_i(\mathbf{x}_i)) + (\rho/2)\left\|\mathbf{x}_i - \bar{\mathbf{x}}_k + \mathbf{u}_i^k\right\|^2\}$ for all $i$
    2. $\mathbf{u}_i^{k+1} = \mathbf{u}_i^k + (\mathbf{x}_i^{k+1} - \bar{\mathbf{x}}_{k+1})$ for all $i$
```

At each iteration of the consensus optimization solution, ADMM first evaluates N proximity operators to update $\mathbf{x}_1, \ldots, \mathbf{x}_N$. Each proximity operator evaluation is distributed, *i.e.*, involves only local information. ADMM then performs the central step of averaging the $\mathbf{x}_i$ and updating the (scaled) dual variables $\mathbf{u}_1, \ldots, \mathbf{u}_N$ according to the discrepancy from the mean.

3.2 The proximal point method

Consider a closed proper convex $f : \mathbf{E} \to \overline{\mathbf{R}}$ and the problem

$$p = \inf_{\mathbf{x}} f(\mathbf{x}). \tag{3.2}$$

This is a very general convex optimization problem, since f maps to $\overline{\mathbf{R}}$ and hence models constraints. It's not clear how to solve problem (3.2), since gradient descent-type methods may fail if f is nonsmooth.

A naive approach is to apply ADMM to (3.2) with $g = 0$ and $A = I$. The augmented Lagrangian is

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + \langle \mathbf{y}, \mathbf{x} - \mathbf{z} \rangle + \frac{\rho}{2} \|\mathbf{x} - \mathbf{z}\|^2$$

and the dual is

$$d = \sup_{\mathbf{y}} -f^*(\mathbf{y}) = \sup_{\mathbf{y}} - \sup_{\mathbf{x}} \{\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x})\} = \sup_{\mathbf{y}} -\delta_{\{\mathbf{0}\}}(\mathbf{y}) = 0$$

with optimal dual variable $\mathbf{y}^* = \mathbf{0}$. In ADMM, therefore, $\mathbf{y}_{k+1} = \mathbf{0}$ and $\mathbf{z}_{k+1} = \mathbf{x}_{k+1}$, so the algorithm reduces to

```
given $\mathbf{x}_k$ and parameter $\rho > 0$
repeat $\mathbf{x}_{k+1} = \text{argmin}_{\mathbf{x}}\{f(\mathbf{x}) + (\rho/2)\|\mathbf{x} - \mathbf{x}_k\|^2\}$
```

which is called the proximal point method (PPM).

Corollary 33 *If f has a minimizer, then $f(\mathbf{x}_k) \to \min f$ and $\mathbf{x}_{k+1} - \mathbf{x}_k \to \mathbf{0}$.*

The proof follows from theorem 32 with $\mathbf{y}^* = \mathbf{0}$.

3.2.1 Convergence

Corollary 33 establishes that PPM will eventually converge to the solution of problem (3.2). We could consider our work with PPM done, but we will explore its convergence further in order to understand the convergence of ADMM from a more abstract perspective.

We start by showing that the subproblem at each PPM iteration has a solution. This is guaranteed if and only if

$$\tilde{f}(\mathbf{u}) = \inf_{\mathbf{x}} \left\{ f(\mathbf{x}) + \frac{\rho}{2} \|\mathbf{x} - \mathbf{u}\|^2 \right\}$$

is finite for all $\mathbf{u}$. The function $\tilde{f}$ is called the Moreau envelope of f.

Proposition 34 *If a proper $f : \mathbf{E} \to \overline{\mathbf{R}}$ has a subgradient, then the Moreau envelope of f is always finite.*

All we need to know, therefore, is when a closed proper convex function has a subgradient. It turns out the answer is "always."

Theorem 35 *A convex function has a subgradient if and only if it's proper.*

The proof uses the following lemma.

Lemma 36 *If $\mathbf{0} \in S \subset \mathbf{E}$ and S is open, then $\operatorname{int} S \neq \emptyset$ if and only if S spans $\mathbf{E}$.*

We have established that the PPM subproblems are well-defined. Next, we characterize their solutions.

Theorem 37 *If $f : \mathbf{E} \to \overline{\mathbf{R}}$ is closed convex proper and $\mathbf{z} \in \mathbf{E}$, then*

$$\inf_{\mathbf{x}} \left\{ f(\mathbf{x}) + \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|^2 \right\} + \inf_{\mathbf{y}} \left\{ f^*(\mathbf{y}) + \frac{1}{2} \|\mathbf{y} - \mathbf{z}\|^2 \right\} = \frac{1}{2} \|\mathbf{z}\|^2$$

uniquely attained by $\mathbf{x}^*$ and $\mathbf{y}^*$. Furthermore, $(\mathbf{x}^*, \mathbf{y}^*)$ is the unique solution of the system

$$\begin{cases} \mathbf{x} + \mathbf{y} = \mathbf{z} \\ \mathbf{y} \in \partial f(\mathbf{x}). \end{cases}$$

This theorem is symmetric in $\mathbf{x}$ and $\mathbf{y}$, since $f = f^{**}$ by theorem 22. The proof applies Fenchel duality to the problem in $\mathbf{x}$, observing that f has a subgradient. Uniqueness follows from the fact that $\|\mathbf{x} - \mathbf{z}\|^2$ is strictly convex, so $f(\mathbf{x}) + (1/2) \|\mathbf{x} - \mathbf{z}\|^2$ is too.

Theorem 37 gives us a unique characterization of $\mathbf{x}_{k+1}$, the solution of the k^{th} PPM subproblem. Replacing $\mathbf{x}$ by $\mathbf{x}_{k+1}$ and $\mathbf{z}$ by $\mathbf{x}_k$ gives

$$\mathbf{x}_k - \mathbf{x}_{k+1} \in \partial f(\mathbf{x}_{k+1}).$$

Since the solution is unique, we write

$$\mathbf{x}_{k+1} = (I + \partial f)^{-1}(\mathbf{x}_k)$$

where $(I + \partial f)^{-1}$ is called the resolvent of the subdifferential operator ∂f.

3.2.2 ADMM as a special case of PPM

We now explore what properties of ∂f, which maps points in $\mathbf{E}$ to sets in $\mathbf{E}$, we need in order to guarantee PPM convergence.

Definition 29 (Monotone) *A set-valued mapping $T : \mathbf{E} \rightrightarrows \mathbf{E}$ is monotone if*

$$\mathbf{y} \in T(\mathbf{x}) \text{ and } \mathbf{y}' \in T(\mathbf{x}') \quad \Longrightarrow \quad \langle \mathbf{y} - \mathbf{y}', \mathbf{x} - \mathbf{x}' \rangle \geq 0.$$

Examples: if f is convex, then ∂f is monotone. If $T = f : \mathbf{R} \to \mathbf{R}$ with f nondecreasing, then T is monotone.

Definition 30 (Maximal) *A monotone $T : \mathbf{E} \rightrightarrows \mathbf{E}$ is maximal if*

$$T' : \mathbf{E} \rightrightarrows \mathbf{E} \textit{ monotone and } T(\mathbf{x}) \subset T'(\mathbf{x}) \textit{ for all } \mathbf{x} \in \mathbf{E} \quad \Longrightarrow \quad T \equiv T'.$$

It turns out that PPM works with ∂f replaced by any maximal monotone operator.

Theorem 38 (Minty) *A monotone $T : \mathbf{E} \rightrightarrows \mathbf{E}$ is maximal if and only if $\mathbf{z} \in \mathbf{x} + T(\mathbf{x})$ has a unique solution $\mathbf{x}$ for all $\mathbf{z} \in \mathbf{E}$.*

The "only if" direction of Minty's theorem is an easy exercise, but the "if" direction is hard to prove in general. We do know one special case where the "if" direction is easy, however: if $T = \partial f$ and f is closed, proper, and convex, then theorem 37 does the job. In this case, Minty's theorem also tells us that ∂f is maximal monotone only if $f : \mathbf{E} \to \overline{\mathbf{R}}$ is closed, proper, and convex.

Theorem 39 *Suppose T is maximal monotone and $\{\lambda_k\}$ is a bounded sequence with $\lambda_k > 0$ for all k. Consider the PPM iteration*

$$\mathbf{x}_{k+1} = (I + (1/\lambda_k)F)^{-1}(\mathbf{x}_k).$$

If $\mathbf{0} \in F(\mathbf{x})$ is solvable, then the sequence $\{\mathbf{x}_k\}$ converges to a solution.

The sequence λ_k is included to allow online tuning of the step size ρ.

This theorem is important because, by a clever choice of maximal monotone T, it can be used to show that ADMM is a special case of PPM.

3.3 The augmented Lagrangian method

It's not clear how to actually implement PPM for a general problem. A special case where PPM is easy to implement is the Lagrangian framework discussed in §2.3.

Observe that the Lagrangian primal (2.2) is equivalent to

$$p = \inf_{\mathbf{x} \in U, \mathbf{z} \geq \mathbf{0}} \left\{ f(\mathbf{x}) \mid \mathbf{g}(\mathbf{x}) + \mathbf{z} = \mathbf{0} \right\}.$$

We can also add a quadratic penalty, giving the equivalent primal

$$p = \inf_{\mathbf{x} \in U, \mathbf{z} \geq \mathbf{0}} \left\{ f(\mathbf{x}) + \frac{1}{2} \left\| \mathbf{g}(\mathbf{x}) + \mathbf{z} \right\|^2 \,\middle|\, \mathbf{g}(\mathbf{x}) + \mathbf{z} = \mathbf{0} \right\}.$$

Taking the Lagrangian dual gives

$$d = \sup_{\boldsymbol{\lambda}} \inf_{\mathbf{x} \in U, \mathbf{z} \geq \mathbf{0}} \left\{ f(\mathbf{x}) + \frac{1}{2} \left\| \mathbf{g}(\mathbf{x}) + \mathbf{z} \right\|^2 + \boldsymbol{\lambda}^T (\mathbf{g}(\mathbf{x}) + \mathbf{z}) \right\}.$$

The minimization over $\mathbf{z} \geq \mathbf{0}$ can be done in closed form using the following lemma.

Lemma 40 *Given $\lambda, g \in \mathbf{R}$,*

$$\inf_{z \geq 0} \left\{ \frac{1}{2}(g+z)^2 + \lambda(g+z) \right\} = \frac{1}{2}(\lambda + g)_+^2 - \frac{1}{2}\lambda^2,$$

where $(\cdot)_+ = \max(0, \cdot)$. The minimizing z is $z = (\lambda + g)_+$.

Applying Lemma 40 componentwise to the Lagrangian dual gives

$$d = \sup_{\boldsymbol{\lambda}} \inf_{\mathbf{x} \in U} \left\{ f(\mathbf{x}) + \frac{1}{2} \left\| (\boldsymbol{\lambda} + \mathbf{g}(\mathbf{x}))_+ \right\|^2 - \frac{1}{2} \left\| \boldsymbol{\lambda} \right\|^2 \right\},$$

where the positive part is taken componentwise. This motivates defining the augmented Lagrangian

$$\bar{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \frac{1}{2} \left\| (\boldsymbol{\lambda} + \mathbf{g}(\mathbf{x}))_+ \right\|^2 - \frac{1}{2} \left\| \boldsymbol{\lambda} \right\|^2.$$

It also motivates the augmented Lagrangian method (ALM),

```
given $\mathbf{x}_k$ and $\boldsymbol{\lambda}_k$
repeat
    1. $\mathbf{x}_k = \mathrm{argmin}_{\mathbf{x}}\, \bar{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}_k)$
    2. $\boldsymbol{\lambda}_{k+1} = (\boldsymbol{\lambda}_k + \mathbf{g}(\mathbf{x}_k))_+$
```

ALM is one of a handful of standard methods used by commercial codes for nonlinear optimization. It can be shown (see Lemma 41) that

$$\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1} \in \partial\bar{\psi}(\boldsymbol{\lambda}_{k+1}),$$

which means that an ALM iteration is exactly a PPM iteration on the dual. Therefore, all the properties of PPM also apply to ALM. Note that the inner minimization can be done using a Newton-style method, although the Hessian may not be well-defined since the augmented Lagrangian is nonsmooth.

Lemma 41 *Let $\bar{\mathbf{x}} \in U$ minimize $\bar{\mathcal{L}}(\cdot, \bar{\boldsymbol{\lambda}})$ for some $\bar{\boldsymbol{\lambda}} \geq \mathbf{0}$, and define $\bar{\mathbf{z}} = (\bar{\boldsymbol{\lambda}} + \mathbf{g}(\bar{\mathbf{x}}))_+$. Then*

$$\bar{\boldsymbol{\lambda}} - \bar{\mathbf{z}} \in \partial\bar{\psi}(\bar{\mathbf{z}}),$$

where

$$\bar{\psi}(\bar{\mathbf{z}}) = \begin{cases} -\inf_{\mathbf{x}\in U} \bar{\mathcal{L}}(\mathbf{x}, \bar{\mathbf{z}}) & \bar{\mathbf{z}} \geq \mathbf{0} \\ +\infty & \textit{otherwise.} \end{cases}$$

The proof involves showing that for convex $h : \mathbf{E} \to \overline{\mathbf{R}}$, the function $\mathbf{x} \mapsto (1/2)h(\mathbf{x})_+^2$ is convex with subgradients $h(\mathbf{x})_+\partial h(\mathbf{x})$ at $\mathbf{x}$. Subdifferential calculus then shows that $\bar{\mathbf{x}}$ minimizes $\bar{\mathcal{L}}(\cdot, \bar{\boldsymbol{\lambda}})$ if and only if $\bar{\mathbf{x}}$ minimizes $\mathcal{L}(\cdot, \bar{\mathbf{z}})$.

3.4 The subgradient method

We now consider the problem of minimizing a convex, proper $f : \mathbf{E} \to \overline{\mathbf{R}}$ over a closed, convex, nonempty $Q \subset \operatorname{dom} f$. We assume that

- f has at least one minimizer $\mathbf{x}^* \in Q$,
- it's easy to project onto Q, and
- it's easy to produce one element of $\partial f(\mathbf{x})$ for any $\mathbf{x} \in Q$.

These assumptions hold in a variety of interesting situations. One is the Lagrangian framework, where we're interested in the primal

$$p = \inf_{\mathbf{z}} \{h(\mathbf{z}) \mid \mathbf{g}(\mathbf{z}) \leq \mathbf{0}\}$$

and dual

$$d = \inf_{\boldsymbol{\lambda} \geq \mathbf{0}} \psi(\boldsymbol{\lambda}).$$

The dual function

$$\psi(\boldsymbol{\lambda}) = -\inf_{\mathbf{z}} \left\{h(\mathbf{z}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{z})\right\}$$

is convex, and projecting onto $\mathbf{R}^m_+$ is easy (just take the componentwise positive part). It's also easy to produce a subgradient of $\psi(\boldsymbol{\lambda})$, as the following proposition shows.

Proposition 42 *If $\bar{\mathbf{z}}$ minimizes $h + \bar{\boldsymbol{\lambda}}^T \mathbf{g}$, then $-\mathbf{g}(\bar{\mathbf{z}}) \in \partial\psi(\bar{\boldsymbol{\lambda}})$.*

Proof: Observe that $\psi(\boldsymbol{\lambda}) \geq -h(\bar{\mathbf{z}}) - \boldsymbol{\lambda}^T \mathbf{g}(\bar{\mathbf{z}})$ and $\psi(\bar{\boldsymbol{\lambda}}) = -h(\bar{\mathbf{z}}) - \bar{\boldsymbol{\lambda}}^T \mathbf{g}(\bar{\mathbf{z}})$. Hence

$$\psi(\boldsymbol{\lambda}) - \psi(\bar{\boldsymbol{\lambda}}) \geq -(\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}})^T \mathbf{g}(\bar{\mathbf{z}}),$$

so $-\mathbf{g}(\bar{\mathbf{z}})$ is a subgradient of ψ at $\bar{\boldsymbol{\lambda}}$. □

3.4.1 Algorithm

The subgradient method uses two key ideas. The first is moving in a direction opposite to a subgradient. The second is projecting onto Q, an operation which we denote by P_Q. These ideas are motivated by the following proposition.

Proposition 43 *For nonminimizing $\mathbf{x} \in Q$ and nonzero $\mathbf{g} \in \partial f(\mathbf{x})$, $\|P_Q(\mathbf{x} - t\mathbf{g}) - \mathbf{x}^*\|$ is strictly decreasing for small $t \geq 0$.*

The proof uses the fact that for $\mathbf{x}^* \in Q$ and $\mathbf{x} \in \mathbf{E}$,

$$\|\mathbf{x} - P_Q(\mathbf{x})\|^2 + \|P_Q(\mathbf{x}) - \mathbf{x}^*\|^2 \leq \|\mathbf{x} - \mathbf{x}^*\|^2 .$$

Proposition 43 guarantees that by taking a small step in a direction opposite a subgradient and projecting the result onto Q, we reduce the distance to a minimizing $\mathbf{x}^* \in Q$. Repeating this process will eventually drive the iterates to $\mathbf{x}^*$. This motivates the subgradient method,

```
given $\mathbf{x}_k$, $\mathbf{g}_k \in \partial f(\mathbf{x}_k)$, and step size $t_k > 0$
repeat $\mathbf{x}_{k+1} = P_Q(\mathbf{x}_k - t_k \mathbf{g}_k / \|\mathbf{g}_k\|)$
```

It's important to note that proposition 43 makes no mention of improving the objective value. To show that the subgradient method converges to an optimal value, we need a Lipschitz assumption on f in the neighborhood of $\mathbf{x}^*$. We also need to understand how to choose the step size t_k.

3.4.2 Convergence and step size

To prove that the subgradient method converges, we need to add the following assumption:

- the objective function f is M-Lipschitz on $\mathbf{x}^* + RB$, for some $R > 0$.

This gives the following lemma.

Lemma 44 *If* $\|\mathbf{x} - \mathbf{x}^*\| \leq R$ *and* $\mathbf{g} \in \partial f(\mathbf{x})$, *then*

$$\frac{1}{M}(f(\mathbf{x}) - f(\mathbf{x}^*)) \leq \left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|}, \mathbf{x} - \mathbf{x}^* \right\rangle .$$

Suppose that we start the subgradient method from some $\mathbf{x}_0$ satisfying $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq R$, and that $f(\mathbf{x}_i) > f(\mathbf{x}^*) + \epsilon$ for all $i = 0, \ldots, k$. In other words, we never get within $\epsilon > 0$ of optimality. In the worst case, how big could the optimality gap ϵ be after k iterations?

To answer this question, recall the subgradient update equation

$$\mathbf{x}_{i+1} = P_Q\left(\mathbf{x}_i - t_i \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|}\right) .$$

Subtracting $\mathbf{x}^*$ from both sides and taking norms gives

$$\begin{aligned}
\|\mathbf{x}_{i+1} - \mathbf{x}^*\|^2 &= \left\| P_Q\left(\mathbf{x}_i - t_i \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|}\right) - \mathbf{x}^* \right\|^2 \\
&\leq \left\| \mathbf{x}_i - t_i \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|} - \mathbf{x}^* \right\|^2 \\
&= \|\mathbf{x}_i - \mathbf{x}^*\|^2 + t_i^2 - 2t_i \left\langle \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|}, \mathbf{x}_i - \mathbf{x}^* \right\rangle .
\end{aligned}$$

This gives the $k+1$ inequalities

$$\begin{aligned} R^2 - \|\mathbf{x}_0 - \mathbf{x}^*\|^2 &\geq 0 \\ \|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_1 - \mathbf{x}^*\|^2 + (t_1)^2 - 2t_1 \left\langle \frac{\mathbf{g}_1}{\|\mathbf{g}_1\|}, \mathbf{x}_1 - \mathbf{x}^* \right\rangle &\geq 0 \\ &\vdots \\ \|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_k - \mathbf{x}^*\|^2 + (t_k)^2 - 2t_k \left\langle \frac{\mathbf{g}_k}{\|\mathbf{g}_k\|}, \mathbf{x}_k - \mathbf{x}^* \right\rangle &\geq 0. \end{aligned}$$

Adding these inequalities gives

$$R^2 + \sum_{i=0}^{k} t_i^2 \geq 2 \sum_{i=0}^{k} t_i \left\langle \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|}, \mathbf{x}_i - \mathbf{x}^* \right\rangle. \tag{3.3}$$

Claim 45 *For $i = 0, \ldots, k$,*

$$\left\langle \frac{\mathbf{g}_i}{\|\mathbf{g}_i\|}, \mathbf{x}_i - \mathbf{x}^* \right\rangle > \frac{\epsilon}{M}.$$

The proof is by contradiction. It uses lemma 44 and the fact that

$$\epsilon < f(\mathbf{x}_0) - f(\mathbf{x}^*) \leq RM.$$

This holds because $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq R$, $f(\mathbf{x}_i) > f(\mathbf{x}^*) + \epsilon$, and f is M-Lipschitz on $\mathbf{x}^* + RB$.

Combining this claim with inequality 3.3 gives

$$\epsilon < M \left(\frac{R^2 + \sum_{i=0}^k t_i^2}{2 \sum_{i=0}^k t_i} \right).$$

This is the desired upper bound on the optimality gap. It shows that after k iterations of the subgradient method, started from any $\mathbf{x}_0 \in \mathbf{x}^* + RB$, we get at least as close to optimality as the right-hand side. This is stated formally in the next theorem.

Theorem 46 *If f is M-Lipschitz on $\mathbf{x}^* + RB$, where $\mathbf{x}^*$ is a minimizer of f, then*

$$\min_{i=1,\ldots,k} \{f(\mathbf{x}_i) - f(\mathbf{x}^*)\} \leq M \left(\frac{R^2 + \sum_{i=0}^k t_i^2}{2 \sum_{i=0}^k t_i} \right).$$

Step size. Theorem 46 establishes that the subgradient method converges, provided the step sizes t_i are chosen such that the denominator goes to $+\infty$ faster than the numerator. One such choice is $t_i = 1/(i+1)$. A better choice, it turns out, is

$$t_i = \frac{r}{\sqrt{i+1}},$$

which gives the convergence bound

$$\min_{i=1,\ldots,k} \{f(\mathbf{x}_i) - f(\mathbf{x}^*)\} \leq M \left(\frac{R^2 + r^2(1 + \ln(k+1))}{4r\left(\sqrt{k+2} - 1\right)} \right).$$

For large k, the right-hand side is $\mathcal{O}(1/\sqrt{k})$. This is *very slow*: to get within $\epsilon = 10^{-3}$ of optimality, the subgradient method requires $k \approx 1/\epsilon^2 = 10^6$ iterations. An interesting fact, however, is that the bound is independent of the dimension of $\mathbf{x}$, making the method potentially attractive for very large problems.

Although the subgradient method is slow, it's actually optimal for the problem in this section, where we assume only that a subgradient is available, projection is easy, and the objective function is Lipschitz on a neighborhood of a minimizer in Q. There are counterexamples showing that (roughly speaking) no algorithm can do better than $\mathcal{O}(1/\sqrt{k})$ on this problem; to improve the convergence bound requires further assumptions about the problem structure. The next section narrows the scope to smooth objective functions, for which $\mathcal{O}(1/k)$ and even $\mathcal{O}(1/k^2)$ algorithms have been developed.

3.5 The gradient method

In §3.4, we considered the problem of minimizing a convex, proper (but possibly nonsmooth) objective function over a closed, convex set. We now narrow our focus to objective functions that are *smooth*, *i.e.*, that are differentiable and whose gradients satisfy a Lipschitz condition.

For any unit direction $\mathbf{d}$, the rate of decrease in the function f at $\mathbf{x}$ is

$$-\lim_{t\downarrow 0}\frac{f(\mathbf{x}+t\mathbf{d})-f(\mathbf{x})}{t}=-\langle\nabla f(\mathbf{x}),\mathbf{d}\rangle.$$

Assuming $\nabla f(\mathbf{x})\neq \mathbf{0}$, the rate of decrease is optimized by the negative gradient, $\mathbf{d}=-\nabla f(\mathbf{x})$. This is the basis for the gradient method,

```
given $\mathbf{x}_k$ and step size $t_k > 0$
repeat $\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)$
```

Theorem 47 *If $f:\mathbf{E}\to\mathbf{R}$ and ∇f is L-Lipschitz, then for all $\mathbf{x}$ and $\mathbf{y}$,*

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

Theorem 47 bounds the error between $f(\mathbf{y})$ and $f(\mathbf{x})+\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle$, the linear approximation to $f(\mathbf{y})$. The proof uses the fundamental theorem of calculus and the chain rule.

Assuming that the linear approximation underestimates $f(\mathbf{y})$, theorem 47 gives the quadratic upper bound

$$f(\mathbf{y})\leq f(\mathbf{x})+\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle+\frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2. \tag{3.4}$$

The $\mathbf{y}$ that minimizes this upper bound is

$$\mathbf{y}=\mathbf{x}-\frac{1}{L}\nabla f(\mathbf{x}),$$

which looks like a gradient method iteration with $t_k=1/L$. This is the step size that we'll use to analyze the algorithm. Note that this $\mathbf{y}$ gives the upper bound

$$f\left(\mathbf{x}-\frac{1}{L}\nabla f(\mathbf{x})\right)\leq f(\mathbf{x})-\frac{1}{2L}\|\nabla f(\mathbf{x})\|^2. \tag{3.5}$$

The results so far hold for convex or nonconvex f. If we assume that f is convex, we get the following result.

Theorem 48 (Co-coercivity) *For convex $f:\mathbf{E}\to\mathbf{R}$ with ∇f L-Lipschitz,*

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

This theorem strengthens the monotonicity that we already know the gradient, a monotone operator, satisfies. With a little algebra, co-coercivity implies that with step size $t_k = 1/L$, the gradient method iterates satisfy

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

In other words, the distance to any minimizer $\mathbf{x}^*$ is nonincreasing.

Theorem 49 *Let f be convex and have L-Lipschitz gradient. If $\mathbf{x}_{k+1} = \mathbf{x}_k - (1/L)\nabla f(\mathbf{x}_k)$ and $\mathbf{x}^*$ minimizes f, then*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{2L\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|^2}{k+4}.$$

The proof uses the quadratic upper bound (3.5) and the subgradient inequality.

Theorem 49 says that the gradient method is a descent method with $\mathcal{O}(1/k)$ convergence. This is a big improvement over the $\mathcal{O}(1/\sqrt{k})$ convergence of the subgradient method: to get within $\epsilon = 10^{-3}$ of optimality, the gradient method requires only $k \approx 1/\epsilon = 10^3$ iterations, rather than the 10^6 subgradient iterations. Of course, the gradient method applies only to smooth objective functions.

Although we proved theorem 49 using the step size $t_k = 1/L$, the $\mathcal{O}(1/k)$ convergence holds for other step sizes, provided that they're not too much larger than $1/L$. In practice, t_k is often chosen by backtracking.

3.6 Fast proximal gradient methods

In 1983, Yurii Nesterov published a method with $\mathcal{O}(1/k^2)$ convergence for the problem in §3.5. Although this was a huge improvement over the gradient method's $\mathcal{O}(1/k)$ convergence, Nesterov's first method was fairly opaque and gained little attention. Nesterov published a more transparent $\mathcal{O}(1/k^2)$ method in 1988, and another in 2005. In 2008, Amir Beck and Marc Teboulle developed the fast iterative shrinkage-thresholding algorithm (FISTA), an $\mathcal{O}(1/k^2)$ splitting method that has since gained widespread popularity. [3] Later in 2008, Paul Tseng gave a clean derivation of the convergence properties of a family of splitting methods that includes FISTA. [4] The development in this section follows Tseng's.

3.6.1 General framework

The setting is the problem

$$\inf \{ f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}) \},$$

where $g : \mathbf{E} \to \mathbf{R}$ is convex, ∇g is L-Lipschitz, and $h : \mathbf{E} \to \overline{\mathbf{R}}$ is closed, proper, and convex. We also assume

- the infimum is attained at $\mathbf{x}^*$ (not necessarily uniquely), and
- the proximity operator of h is easy to evaluate.

We consider the following family of methods, parameterized by θ.

```
given $\mathbf{x}$, $\mathbf{v}$ (initialized by $\mathbf{v}_0 = \mathbf{x}_0$), parameter $\theta$, and Lipschitz constant $L > 0$
repeat
    1. $\theta \leftarrow \theta_k$
    2. $\mathbf{y} \leftarrow (1-\theta)\mathbf{x} + \theta\mathbf{v}$
    3. $\mathbf{y}^+ \leftarrow \mathbf{y} - (1/L)\nabla g(\mathbf{y})$
    4. $\mathbf{x}^+$ minimizes $h + (L/2)\left\|\cdot - \mathbf{y}^+\right\|^2$
    5. $\mathbf{v}^+ \leftarrow \mathbf{x}^+ + (1/\theta - 1)(\mathbf{x}^+ - \mathbf{x})$
    6. $\mathbf{x} \leftarrow \mathbf{x}^+$
    7. $\mathbf{v} \leftarrow \mathbf{v}^+$
```

If $\theta_k = 1$ for all k, then at every iteration $\mathbf{v} = \mathbf{x} = \mathbf{y}$, and the algorithm reduces to a gradient step followed by a proximity operation. The $\theta_k = 1$ case, called the proximal gradient method (PGM), further reduces to PPM when $g \equiv 0$.

We now search for a choice of θ that results in $\mathcal{O}(1/k^2)$ convergence. Since g is L-Lipschitz, the quadratic upper bound (3.4) gives

$$g(\mathbf{x}^+) \leq g(\mathbf{y}) + \langle \nabla g(\mathbf{y}), \mathbf{x}^+ - \mathbf{y} \rangle + \frac{L}{2} \left\| \mathbf{x}^+ - \mathbf{y} \right\|^2 . \tag{3.6}$$

From the definition of $\mathbf{x}^+$,

$$\begin{aligned} \mathbf{0} &\in \partial \left(h + \frac{L}{2} \left\| \cdot - \mathbf{y}^+ \right\|^2 \right) (\mathbf{x}^+) \\ &= \partial h(\mathbf{x}^+) + L(\mathbf{x}^+ - \mathbf{y}^+), \end{aligned}$$

where the second line follows from the subdifferential calculus theorem 26 and the differentiability of $\|\cdot\|^2$. But $L(\mathbf{y}^+ - \mathbf{x}^+) \in \partial h(\mathbf{x}^+)$ means that for all $\mathbf{z}$,

$$h(\mathbf{x}^+) \leq h(\mathbf{z}) + L\langle \mathbf{x}^+ - \mathbf{z}, \mathbf{y}^+ - \mathbf{x}^+ \rangle.$$

Plugging in the definition of $\mathbf{y}^+$ and simplifying,

$$h(\mathbf{x}^+) \leq h(\mathbf{z}) + L\langle \mathbf{x}^+ - \mathbf{z}, \mathbf{y} - \mathbf{x}^+ \rangle - \langle \mathbf{x}^+ - \mathbf{z}, \nabla g(\mathbf{y}) \rangle. \tag{3.7}$$

Adding inequalities (3.6) and (3.7) gives

$$f(\mathbf{x}^+) \leq g(\mathbf{y}) + h(\mathbf{z}) + \langle \nabla g(\mathbf{y}), \mathbf{z} - \mathbf{y} \rangle + L\langle \mathbf{x}^+ - \mathbf{z}, \mathbf{y} - \mathbf{x}^+ \rangle + \frac{L}{2} \left\| \mathbf{x}^+ - \mathbf{y} \right\|^2 .$$

Since g is convex,

$$g(\mathbf{y}) + \langle \nabla g(\mathbf{y}), \mathbf{z} - \mathbf{y} \rangle \leq g(\mathbf{z}).$$

For all $\mathbf{z}$, therefore,

$$f(\mathbf{x}^+) \leq f(\mathbf{z}) + L\langle \mathbf{x}^+ - \mathbf{z}, \mathbf{y} - \mathbf{x}^+ \rangle + \frac{L}{2} \left\| \mathbf{x}^+ - \mathbf{y} \right\|^2 . \tag{3.8}$$

Evaluating inequality (3.8) at $\mathbf{z} = \mathbf{x}$ and $\mathbf{z} = \mathbf{x}^*$ and multiplying through by $1 - \theta$ and θ, respectively, gives

$$\begin{aligned} (1-\theta) f(\mathbf{x}^+) &\leq (1-\theta) \left(f(\mathbf{x}) + L\langle \mathbf{x}^+ - \mathbf{x}, \mathbf{y} - \mathbf{x}^+ \rangle + \frac{L}{2} \left\| \mathbf{x}^+ - \mathbf{y} \right\|^2 \right) \\ \theta f(\mathbf{x}^+) &\leq \theta \left(f(\mathbf{x}^*) + L\langle \mathbf{x}^+ - \mathbf{x}^*, \mathbf{y} - \mathbf{x}^+ \rangle + \frac{L}{2} \left\| \mathbf{x}^+ - \mathbf{y} \right\|^2 \right) . \end{aligned}$$

Adding these inequalities, writing the terms involving $\mathbf{y}$ as a difference of squares, and simplifying gives

$$\frac{2}{L\theta^2} \left(f(\mathbf{x}^+) - f(\mathbf{x}^*) \right) + \left\| \mathbf{v}^+ - \mathbf{x}^* \right\|^2 \leq \frac{2(1-\theta)}{L\theta^2} \left(f(\mathbf{x}) - f(\mathbf{x}^*) \right) + \left\| \mathbf{v} - \mathbf{x}^* \right\|^2 . \tag{3.9}$$

We've also used the definitions of $\mathbf{y}^+$ and $\mathbf{v}^+$.

Recall that $\theta \in [0, 1]$ is a free parameter that can be independently chosen at each iteration. A clever choice of θ gives the following theorem.

Theorem 50 *For $\theta_k = 2/(k+1)$,*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \frac{2L\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|^2}{(k+1)^2}.$$

The proof is by induction, using the fact that

$$\frac{1-\theta_k}{\theta_k} = \frac{k^2-1}{4} < \frac{k^2}{4} = \frac{1}{\theta_{k-1}^2}.$$

3.6.2 The fast iterative shrinkage-thresholding algorithm

The FISTA algorithm is the following.

```
given $\mathbf{x}_{k-2}$, $\mathbf{x}_{k-1}$, and Lipschitz constant $L > 0$
repeat
    1. $\mathbf{y} = \mathbf{x}_{k-1} + \frac{k-2}{k+1}(\mathbf{x}_{k-1} - \mathbf{x}_{k-2})$
    2. $\hat{\mathbf{y}} = \mathbf{y} - (1/L)\nabla g(\mathbf{y})$
    3. $\mathbf{x}_k = \operatorname{argmin}_{\mathbf{x}} \left\{ h(\mathbf{x}) + (L/2)\left\|\mathbf{x} - \hat{\mathbf{y}}\right\|^2 \right\}$
```

FISTA converges according to

$$\min_{i=1,\ldots,k} \left\{ f(\mathbf{x}_i) - f(\mathbf{x}^*) \right\} \le \frac{2L\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|^2}{(k+1)^2}.$$

Note that FISTA is not a descent method, and that when $h \equiv 0$, FISTA reduces to Nesterov's second method. The method combines a gradient step (step 2) with a proximity operator (step 3). There are counterexamples to show that FISTA's $\mathcal{O}(1/k^2)$ convergence is optimal within the class of algorithms using only first-order information.

In the form above, FISTA requires knowing the Lipschitz constant L. In practice, however, the Lipschitz constant is often unknown.A simple work-around is to replace $1/L$ by a step size t, which is determined by backtracking. With this modification, the algorithm is

```
given $\mathbf{x}_{k-2}$, $\mathbf{x}_{k-1}$, and Lipschitz constant $L > 0$
repeat
    1. $\mathbf{y} = \mathbf{x}_{k-1} + \frac{k-2}{k+1}(\mathbf{x}_{k-1} - \mathbf{x}_{k-2})$
    2. $\hat{\mathbf{y}} = \mathbf{y} - t\nabla g(\mathbf{y})$
    3. while $g(\mathbf{x}^+) > g(\mathbf{y}) + \langle \nabla g(\mathbf{y}), \mathbf{x}^+ - \mathbf{y} \rangle + \frac{1}{2t}\left\|\mathbf{x}^+ - \mathbf{y}\right\|^2$, set $t \leftarrow t/2$ and return to step 2
    4. $\mathbf{x}_k = \operatorname{argmin}_{\mathbf{x}} \left\{ h(\mathbf{x}) + (L/2)\left\|\mathbf{x} - \hat{\mathbf{y}}\right\|^2 \right\}$
```

The $\mathcal{O}(1/k^2)$ convergence result carries through for FISTA with backtracking.

Chapter 4

Variational analysis and nonconvex optimization

4.1 Generalized derivatives

4.1.1 Regular subgradients

Recall that for convex $f : \mathbf{E} \to \overline{\mathbf{R}}$, $\mathbf{y}$ is a subgradient of f at $\mathbf{x}$ if for all $\mathbf{z}$,

$$f(\mathbf{x}+\mathbf{z}) \geq f(\mathbf{x}) + \langle \mathbf{y}, \mathbf{z} \rangle.$$

This subgradient inequality is a one-sided, global estimate of $f(\mathbf{x}+\mathbf{z})$. We used subgradients extensively in both duality theory and algorithms for convex optimization.

For continuously differentiable $f : \mathbf{E} \to \mathbf{R}$, we have the analogous result that $\mathbf{y}$ is the gradient of f at $\mathbf{x}$ if in the limit $\mathbf{z} \to \mathbf{0}$,

$$f(\mathbf{x}+\mathbf{z}) = f(\mathbf{x}) + \langle \mathbf{y}, \mathbf{z} \rangle + o(\mathbf{z}).$$

This two-sided, local estimate of $f(\mathbf{x}+\mathbf{z})$ is somewhat less useful for optimization. We'd like to develop a one-sided estimate for nonconvex functions. To do so, we extend the idea of subgradients to nonconvex functions with the following definition.

Definition 31 (Regular subgradient) *For $f : \mathbf{E} \to \overline{\mathbf{R}}$ finite at $\mathbf{x}$, $\mathbf{y} \in \hat{\partial} f(\mathbf{x})$ if as $\mathbf{z} \to \mathbf{0}$,*

$$f(\mathbf{x}+\mathbf{z}) \geq f(\mathbf{x}) + \langle \mathbf{y}, \mathbf{z} \rangle + o(\mathbf{z}).$$

We call $\mathbf{y}$ a regular subgradient and $\hat{\partial} f$ the regular subdifferential of f at $\mathbf{x}$.

The regular subdifferential provides a one-sided, local estimate of $f(\mathbf{x}+\mathbf{z})$. It has a few familiar-looking properties:

1. If $\mathbf{x}$ is a local minimizer of f, then $\mathbf{0} \in \hat{\partial} f(\mathbf{x})$.
2. If $\mathbf{0} \in \hat{\partial} f(\mathbf{x})$, then $\mathbf{x}$ is a strict local minimizer of $f + \delta \left\| \cdot - \mathbf{x} \right\|$ for all $\delta > 0$.
3. If $g : \mathbf{E} \to \mathbf{R}$ is continuously differentiable, then

$$\hat{\partial}(f+g)(\mathbf{x}) = \hat{\partial} f(\mathbf{x}) + \nabla g(\mathbf{x}).$$

These properties are as close as we can get to the powerful result that for convex f, $\mathbf{x}^*$ is a global minimizer of f if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

We've also used subdifferential calculus (see theorem 26) in several proofs. This calculus extends to nonconvex functions in the following sense.

Theorem 51 (Fuzzy sum rule) *Let $f_1, f_2 : \mathbf{E} \to \overline{\mathbf{R}}$ be finite at $\mathbf{x}$ and closed. If $\mathbf{y} \in \hat{\partial}(f_1 + f_2)(\mathbf{x})$, then for $i = 1, 2$ there exists an $\mathbf{x}_i$ close to $\mathbf{x}$ and a $\mathbf{y}_i \in \hat{\partial} f_i(\mathbf{x}_i)$ such that $f_i(\mathbf{x}_i)$ is close to $f_i(\mathbf{x})$ and $\mathbf{y}_1 + \mathbf{y}_2$ is close to $\mathbf{y}$.*

The proof uses local arguments and quadratic penalties. This is quite different from our convex analysis proofs, which relied on global arguments and separating hyperplanes. This is a recurring theme that distinguishes nonconvex and convex analysis.

4.1.2 Limiting subgradients

The graph of the regular subdifferential is generally not closed. This is unsatisfying, since we'd like to take limits, *e.g.*, to prove the convergence of algorithms. An object that allows this sort of analysis is the limiting subdifferential.

Definition 32 (Limiting subgradient) *Let $f : \mathbf{E} \to \overline{\mathbf{R}}$ be finite at $\mathbf{x}$. A vector $\mathbf{y}$ is a limiting subgradient of f at $\mathbf{x}$, denoted $\mathbf{y} \in \partial f(\mathbf{x})$, if there exists a sequence $\mathbf{x}_r \to \mathbf{x}$ with $\mathbf{y}_r \in \hat{\partial} f(\mathbf{x}_r)$ and $f(\mathbf{x}_r) \to f(\mathbf{x})$, so $\mathbf{y}_r \to \mathbf{y}$. We call $\partial f(\mathbf{x})$ the limiting subdifferential of f at $\mathbf{x}$.*

Like the regular subdifferential, the limiting subdifferential has some familiar properties:

1. If f is continuously differentiable, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

2. If $f : \mathbf{E} \to \overline{\mathbf{R}}$ is closed and convex, then $\partial f(\mathbf{x})$ is exactly the subdifferential defined in §1.3.3. In other words, there's no clash of notation.

3. If f is Lipschitz, then ∂f has a closed graph with nonempty compact (set) values.

4. If $\mathbf{x}$ is a local minimizer of f, then $\mathbf{0} \in \partial f(\mathbf{x})$.

The limiting subdifferential also satisfies a cleaner version of the fuzzy sum rule:

Theorem 52 (Sum rule) *If $f_1, f_2 : \mathbf{E} \to \overline{\mathbf{R}}$ are closed and f_1 or f_2 is Lipschitz, then*

$$\partial(f_1 + f_2)(\mathbf{x}) \subset \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}).$$

The proof uses the fuzzy sum rule.

4.2 Metric regularity

4.2.1 The Ekeland variational principle

Theorem 53 (Ekeland 1974) *Consider $f : \mathbf{E} \to \overline{\mathbf{R}}$ closed and $\epsilon, \lambda > 0$. If $f(\hat{\mathbf{x}}) \leq \inf f + \epsilon$, then there exists an $\bar{\mathbf{x}}$ satisfying*

1. $\|\bar{\mathbf{x}} - \hat{\mathbf{x}}\| \leq \lambda$,
2. $f(\bar{\mathbf{x}}) \leq f(\hat{\mathbf{x}})$, *and*
3. $\bar{\mathbf{x}}$ *strictly minimizes* $f + (\epsilon/\lambda) \|\cdot - \bar{\mathbf{x}}\|$.

The Ekeland variational principle was published by the French mathematician Ivar Ekeland in 1974. It holds much more generally than the form stated here, and is used in many fields of applied math. It's unique in that it gives an existence criterion that requires no compactness assumptions (unlike the usual Weierstrass arguments).

The proof involves showing that $\hat{f} = f + (\epsilon/\lambda) \|\cdot - \hat{\mathbf{x}}\|$ is closed, that $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\hat{\mathbf{x}})\}$ is nonempty and compact, and hence that the set of minimizers of $\hat{f}$ is nonempty and compact.

When we apply the Ekeland variational principle, we usually think of λ as being some function of ϵ (typically $\sqrt{\epsilon}$), and argue that near an approximate minimizer $\hat{\mathbf{x}}$ lies an exact minimizer $\bar{\mathbf{x}}$ of a "nearby function." This approach gives the following result.

Corollary 54 *For closed $f : \mathbf{E} \to \overline{\mathbf{R}}$, if $f(\hat{\mathbf{x}}) \leq \inf f + \epsilon$, then*

- *there exists an $\bar{\mathbf{x}} \in \hat{\mathbf{x}} + \sqrt{\epsilon}B$ with $f(\bar{\mathbf{x}}) \leq f(\hat{\mathbf{x}})$, and*
- *there exists a $\bar{\mathbf{y}} \in \sqrt{\epsilon}B$ with $\bar{\mathbf{y}} \in \partial f(\bar{\mathbf{x}})$.*

4.2.2 Definitions

Metric regularity is a fundamental topic in many fields of numerical analysis, particularly inverse problems. It provides a general framework from which to view nonlinear root finding, matrix rank, linear programming sensitivity, and other familiar ideas. We will introduce metric regularity in general terms, then discuss its application to nonconvex optimization.

Consider Euclidean spaces $\mathbf{E}, \mathbf{Y}$ and a set valued operator $\Phi : \mathbf{E} \rightrightarrows \mathbf{Y}$. Given the data $\mathbf{y} \in \mathbf{Y}$, we're interested in the generalized equation $\mathbf{y} \in \Phi(\mathbf{x})$. In particular, we'd like to understand whether an approximate solution $\mathbf{x}$ (meaning an $\mathbf{x}$ for which $\mathbf{y}$ is close to $\Phi(\mathbf{x})$) is necessarily close to a true solution. Loosely, this depends on the smoothness of the inverse operator Φ^{-1}.

Definition 33 (Distance) *The distance from $\mathbf{x}$ to a set Z is*

$$d_Z(\mathbf{x}) = \inf_{\mathbf{z} \in Z} \{\|\mathbf{x} - \mathbf{z}\|\}.$$

Note that $d_\emptyset(\mathbf{x}) = +\infty$ for all $\mathbf{x}$.

Example. For $\Phi : \mathbf{R}^n \rightrightarrows \mathbf{R}^m$ defined by

$$\Phi(\mathbf{x}) = \begin{cases} A\mathbf{x} + \mathbf{R}^m_+ & \mathbf{x} \in \mathbf{R}^n_+ \\ \emptyset & \text{otherwise,} \end{cases}$$

the generalized equation $\mathbf{y} \in \Phi(\mathbf{x})$ is a linear programming feasibility problem with solutions satisfying $\mathbf{x} \geq \mathbf{0}$ and $A\mathbf{x} \leq \mathbf{y}$. Suppose $\hat{\mathbf{x}} \geq \mathbf{0}$ and $\|(A\hat{\mathbf{x}} - \mathbf{y})_+\|$ is small. Is $\hat{\mathbf{x}}$ necessarily close to the feasible region $\{\mathbf{x} \geq \mathbf{0} \mid A\mathbf{x} \leq \mathbf{y}\}$? It turns out that the answer is yes.

Theorem 55 (Hoffman 1952) *For any $\hat{\mathbf{x}}$, there exists a k (depending on the data A and $\mathbf{y}$) such that*

$$d_{\{\mathbf{x}\geq\mathbf{0} \ \mid \ A\mathbf{x}\leq\mathbf{y}\}}(\hat{\mathbf{x}}) \leq k \left\|(A\hat{\mathbf{x}} - \mathbf{y})_+\right\|.$$

Metric regularity generalizes this idea to other types of inverse problems.

Definition 34 (Metrically regular) *Suppose $\bar{\mathbf{y}} \in \Phi(\bar{\mathbf{x}})$. A set-valued operator $\Phi : \mathbf{E} \rightrightarrows \mathbf{Y}$ is metrically regular at $\bar{\mathbf{x}}$ for $\bar{\mathbf{y}}$ if there exists a $k > 0$ such that for all $(\mathbf{x}, \mathbf{y})$ close to $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$,*

$$d_{\Phi^{-1}(\mathbf{y})}(\mathbf{x}) \leq k d_{\Phi(\mathbf{x})}(\mathbf{y}).$$

A particular consequence of metric regularity is that $\Phi^{-1}(\mathbf{y})$, the inverse image of $\mathbf{y}$, is nonempty for all $\mathbf{y}$ near $\bar{\mathbf{y}}$.

Examples.

1. If the single-valued map $\mathbf{g} : \mathbf{Y} \to \mathbf{E}$ is Lipschitz and $\Phi(\mathbf{x}) = \mathbf{g}^{-1}(\mathbf{x})$, then Φ is metrically regular (to see this, set k equal to the Lipschitz constant of $\mathbf{g}$). This suggests that metric regularity is a sort of inverse of the idea of Lipschitz continuity.

2. If $\Phi : \mathbf{R} \to \mathbf{R}$ is defined by $\Phi(x) = x^3$, then $\Phi^{-1}(x) = y^{1/3}$. This Φ is *not* metrically regular at 0, since the ratio

$$\frac{d_{\Phi^{-1}(y)}(x)}{d_{\Phi(x)}(y)} = \frac{\left|x - y^{1/3}\right|}{\left|x^3 - y\right|}$$

is not bounded for small (x, y). Note that the inverse map $\Phi^{-1}(y) = y^{1/3}$ is not Lipschitz.

3. The linear map $A : \mathbf{E} \to \mathbf{Y}$ is metrically regular at $\mathbf{0}$ (and in fact, anywhere) if and only if A is surjective or, equivalently, if the matrix representation of A has full row rank. This can be proved by the singular value decomposition, with $k = 1/\sigma_{\min}(A)$.

4.2.3 Conditions for metric regularity

We've seen that an answer to the question "When is an approximate solution to $\mathbf{y} \in \Phi(\mathbf{x})$ necessarily close to a true solution?" is "When Φ is metrically regular." We now develop conditions for metric regularity.

To begin, consider the simple case of a continuous, single-valued mapping $\mathbf{f} : \mathbf{E} \to \mathbf{Y}$. Suppose $\bar{\mathbf{x}} \in S \subset \mathbf{E}$ with S closed, and define $F_S : \mathbf{E} \rightrightarrows \mathbf{Y}$ by

$$F_S(\mathbf{x}) = \begin{cases} \{\mathbf{f}(\mathbf{x})\} & \mathbf{x} \in S \\ \emptyset & \text{otherwise.} \end{cases}$$

If F_S is metrically regular at $\bar{\mathbf{x}}$, then $F_S^{-1}(\mathbf{y}) = \{\mathbf{x} \in S \mid \mathbf{f}(\mathbf{x}) = \mathbf{y}\}$ is nonempty for all $\mathbf{y}$ near $F(\bar{\mathbf{x}})$. In other words, the problem of finding an $\mathbf{x}$ with $\mathbf{f}(\mathbf{x}) = \mathbf{y}$ is solvable for $\mathbf{y}$ near $\mathbf{f}(\bar{\mathbf{x}})$. Hence $\min_S \|\mathbf{f}(\cdot) - \mathbf{y}\| = 0$, and if $\mathbf{x} \in \operatorname{argmin}_S \|\mathbf{f}(\cdot) - \mathbf{y}\|$, then $\mathbf{f}(\mathbf{x}) = \mathbf{y}$. The following theorem establishes an approximate converse of this result.

Lemma 56 (Ioffe 1979) *Suppose $\|\mathbf{x} - \bar{\mathbf{x}}\|$, $\|\mathbf{y} - \mathbf{f}(\mathbf{x})\|$, and $\delta > 0$ are small. If the implication*

$$\mathbf{x} \in \operatorname*{argmin}_S \{\|\mathbf{f}(\cdot) - \mathbf{y}\| + \delta \|\cdot - \mathbf{x}\|\} \quad \Longrightarrow \quad \mathbf{f}(\mathbf{x}) = \mathbf{y}$$

holds, then F_S is metrically regular at $\bar{\mathbf{x}}$.

The proof is by contradiction, and uses the Ekeland variational principle (53).

It's not clear how to check the conditions in Ioffe's lemma. To establish more constructive conditions, we add the assumption that $\mathbf{f}$ is Lipschitz and use the following definition.

Definition 35 (Normal cone) *For $\bar{\mathbf{x}} \in S \subset \mathbf{E}$ with S closed, the set $N_S(\bar{\mathbf{x}}) = \partial\delta_S(\bar{\mathbf{x}})$ is called the normal cone to S at $\bar{\mathbf{x}}$.*

Note that $\partial\cdot$ is the limiting subdifferential. By definition, $N_S(\bar{\mathbf{x}})$ is always a closed convex cone. In the special case of convex S, the normal cone to S at $\bar{\mathbf{x}}$ is the set of all vectors $\mathbf{y}$ making an obtuse angle with vectors $\mathbf{x} - \bar{\mathbf{x}}$, for any $\mathbf{x} \in S$:

$$N_S(\bar{\mathbf{x}}) = \{\mathbf{y} \mid \langle \mathbf{y}, \mathbf{x} - \bar{\mathbf{x}} \rangle \leq 0 \text{ for all } \mathbf{x} \in S\}.$$

In the convex case, the supporting hyperplane theorem implies that $\bar{\mathbf{x}} \in \operatorname{int} S$ if and only if $N_S(\bar{\mathbf{x}}) = \{\mathbf{0}\}$. The following theorem allows us to generalize this result to the nonconvex case (see corollary 58).

Theorem 57 *Suppose $\mathbf{f} : \mathbf{E} \to \mathbf{Y}$ is Lipschitz and $\bar{\mathbf{x}} \in S \subset \mathbf{E}$ with S closed. If the implication*

$$\mathbf{0} \in \partial\langle \mathbf{w}, \mathbf{f}(\cdot)\rangle(\bar{\mathbf{x}}) + N_S(\bar{\mathbf{x}}) \quad \Longrightarrow \quad \mathbf{w} = \mathbf{0}$$

holds, then the mapping

$$\mathbf{x} \mapsto \begin{cases} \{\mathbf{f}(\mathbf{x})\} & \mathbf{x} \in S \\ \emptyset & \mathbf{x} \notin S \end{cases}$$

is metrically regular at $\bar{\mathbf{x}}$.

The proof of theorem 57 has a similar structure to Ioffe's lemma: it's by contradiction, using the Ekeland variational principle. It also involves the sum rule (theorem 52) and the following properties of the limiting subdifferentiall:

- For Lipschitz $\mathbf{g} : \mathbf{E} \to \mathbf{Y}$ and continuously differentiable $g : \mathbf{Y} \to \mathbf{R}$,

  $$\partial(g \circ \mathbf{g})(\mathbf{x}) = \partial\langle \nabla g(\mathbf{g}(\mathbf{x})), \mathbf{g}(\cdot)\rangle(\mathbf{x}).$$

  In other words, $g \circ \mathbf{g}$ can be replaced by its linear approximation.

- If $g : \mathbf{E} \to \mathbf{R}$ is L-Lipschitz, then the graph of ∂g is closed and for all $\mathbf{x}$,

  $$\partial g(\mathbf{x}) \subset LB.$$

- The graph of $\partial \delta_S$ is closed.

Example. Consider the special case where $S = \mathbf{E}$ and $\mathbf{f}$ is smooth. In this case, $\delta_S \equiv 0$, $\partial\delta_S \equiv \{\mathbf{0}\}$, and $\partial\langle \mathbf{w}, \mathbf{f}(\cdot)\rangle(\mathbf{x}) = \langle \mathbf{w}, \nabla \mathbf{f}(\mathbf{x})\rangle$. The metric regularity implication therefore reduces to

$$\mathbf{0} = \langle \mathbf{w}, \nabla \mathbf{f}(\mathbf{x})\rangle \quad \Longrightarrow \quad \mathbf{w} = \mathbf{0},$$

which holds if and only if the Jacobian $\nabla \mathbf{f}$ is surjective. Hence a smooth Lipschitz map defined on all of $\mathbf{E}$ with surjective Jacobian is metrically regular. This result is also known as the inverse function theorem.

Corollary 58 *For any closed set S, $\bar{\mathbf{x}} \in \operatorname{int} S$ if and only if $N_S(\bar{\mathbf{x}}) = \{\mathbf{0}\}$.*

Instead of using the supporting hyperplane theorem, which does not apply to the nonconvex case, the proof follows from theorem 57 with $\mathbf{f}$ the identity map.

4.2.4 Error bounds

We now consider the slightly more general system

$$\begin{cases} \mathbf{f}(\mathbf{x}) \in \mathbf{z} + P \\ \mathbf{x} \in S. \end{cases}$$

Theorem 59 *Suppose $\mathbf{f} : \mathbf{E} \to \mathbf{Y}$ is Lipschitz, $S \subset \mathbf{E}$ and $P \subset \mathbf{Y}$ are closed, $\bar{\mathbf{x}} \in S$, and $\mathbf{f}(\bar{\mathbf{x}}) \in P$. If there is no $k > 0$ such that for all $\mathbf{x} \in S$ near $\bar{\mathbf{x}}$ and small $\mathbf{z}$,*

$$d_{S \cap \mathbf{f}^{-1}(\mathbf{z}+P)}(\mathbf{x}) \leq k d_P(\mathbf{f}(\mathbf{x}) - \mathbf{z}),$$

then there exists a nonzero $\mathbf{w} \in N_P(\mathbf{f}(\bar{\mathbf{x}}))$ with $\mathbf{0} \in \partial\langle \mathbf{w}, \mathbf{f}(\cdot)\rangle(\bar{\mathbf{x}}) + N_S(\bar{\mathbf{x}})$.

The proof uses theorem 57 and the fact that $N_{S\times P}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = N_S(\bar{\mathbf{x}}) \times N_P(\bar{\mathbf{y}})$.

Example. Suppose $\mathbf{Y} = \mathbf{E}$, and that the *transversality* condition $N_S(\bar{\mathbf{x}}) \cap -N_P(\bar{\mathbf{x}}) = \{\mathbf{0}\}$ holds. Then for all $\delta > 0$, the set

$$S \cap (\mathbf{z} + P) \cap (\bar{\mathbf{x}} + \delta B)$$

is nonempty. The analog of this result for convex S and P follows from the separating hyperplane theorem. Theorem 59 therefore gives a kind of local, nonconvex separation. An interesting application of this result is that transversality implies that the method of alternating projections, an algorithm for finding a point in the intersection of two sets, converges locally.

We'd like to extend theorem 59 from its current form, which requires that $\bar{\mathbf{x}} \in S$, to the general case of $\mathbf{x} \in \mathbf{E}$. To do this, we introduce the idea of exact penalty functions.

Proposition 60 (Exact penalty function) *If $f : \mathbf{E} \to \mathbf{R}$ is L-Lipschitz around $\bar{\mathbf{x}}$, a local minimizer of f over $S \subset \mathbf{E}$, then $\bar{\mathbf{x}}$ locally minimizes $f + kd_S$ for all $k \geq L$.*

Proposition 60 allows us to exchange the constrained problem of minimizing f over S for the unconstrained problem of minimizing $f + kd_S$ over $\mathbf{E}$. Note that the modified objective in the unconstrained problem, which involves the (nonsmooth, 1-Lipschitz) distance function, is Lipschitz but nonsmooth.

4.3 Tangent cones and optimality

Definition 36 (Tangent cone) *The tangent cone to a set S at $\bar{\mathbf{x}}$ is*

$$T_S(\bar{\mathbf{x}}) = \left\{ \lim_r \mathbf{d}_r \;\middle|\; \textit{there exist } t_r \downarrow 0 \textit{ with } \bar{\mathbf{x}} + t_r \mathbf{d}_r \in S \right\}.$$

In other words, the tangent cone is the set of limit points of sequences of directions in which you can take a small step from $\bar{\mathbf{x}}$ and remain in S. For any S, $T_S(\bar{\mathbf{x}})$ is a closed cone.

Proposition 61 *The dual cone of $T_S(\bar{\mathbf{x}})$ is $-\hat{\partial}\delta_S(\bar{\mathbf{x}})$.*

Corollary 62 *For closed, convex S, $T_S(\bar{\mathbf{x}}) = \mathrm{cl}(\mathbf{R}_+(S - \bar{\mathbf{x}})) = -N_S(\bar{\mathbf{x}})^+$.*

Corollary 63 *For any $S \subset \mathbf{E}$, S^{++} is the smallest closed, convex cone containing S.*

Equivalently, S^{++} is the intersection of all closed, convex cones containing S.

Proposition 64 (Basic optimality condition) *If $\bar{\mathbf{x}}$ is a local minimizer of $f : \mathbf{E} \to \mathbf{R}$ over $S \subset \mathbf{E}$ and f is differentiable at $\bar{\mathbf{x}}$, then $\nabla f(\bar{\mathbf{x}}) \in T_S(\bar{\mathbf{x}})^+$.*

Proposition 64 is the basic first-order necessary optimality condition for nonlinear optimization. The proof uses properties of the regular subdifferential. To apply the result, it's useful to know how to compute tangent cones to sets with structure, such as $S \cap \mathbf{f}^{-1}(P) = \{\mathbf{x} \in S \mid \mathbf{f}(\mathbf{x}) \in P\}$, the feasible region of an optimization problem.

Proposition 65 *For $S \subset \mathbf{E}$, $P \subset \mathbf{Y}$, and $\mathbf{f} : \mathbf{E} \to \mathbf{Y}$ continuously differentiable at $\bar{\mathbf{x}} \in S$ with $\mathbf{f}(\bar{\mathbf{x}}) \in P$,*

$$T_{S \cap \mathbf{f}^{-1}(P)}(\bar{\mathbf{x}}) \subset T_S(\bar{\mathbf{x}}) \cap \nabla \mathbf{f}(\bar{\mathbf{x}})^{-1} T_P(\mathbf{f}(\bar{\mathbf{x}})).$$

Note that the set $\nabla \mathbf{f}(\bar{\mathbf{x}})^{-1} T_P(\mathbf{f}(\bar{\mathbf{x}}))$ is the inverse image of the tangent cone to P at $\mathbf{f}(\bar{\mathbf{x}})$, under the inverse image of $\nabla \mathbf{f}(\bar{\mathbf{x}})$.

Proposition 65 allows us to compute the tangent cone to $S \cap \mathbf{f}^{-1}(P)$ by computing a Jacobian and the tangent cones to S and P. However, the basic optimality condition involves the *dual* of the tangent cone of the the feasible region, rather than the tangent cone itself. To construct duals to tangent cones, it helps to view the tangent cone in terms of the distance function, as follows.

Proposition 66 *For $\bar{\mathbf{x}} \in S$, a direction $\mathbf{d}$ lies in $T_S(\bar{\mathbf{x}})$ if and only if there exists a sequence $t_r \downarrow 0$ with*

$$\lim (1/t_r) d_S(\bar{\mathbf{x}} + t_r \mathbf{d}) = 0.$$

When S is convex, the condition reduces to $\lim_{t \downarrow 0} (1/t) d_S(\bar{\mathbf{x}} + t\mathbf{d}) = 0$. Proposition 66 allows us to introduce the machinery of metric regularity, which bounds the distance to the feasible region.

Theorem 67 *Assume $S \subset \mathbf{E}$ and $P \subset \mathbf{Y}$ are closed and convex, $\mathbf{f} : \mathbf{E} \to \mathbf{Y}$ is continuously differentiable, $\bar{\mathbf{x}} \in S$, $\mathbf{f}(\bar{\mathbf{x}}) \in P$, and the constraint qualification*

$$\begin{cases} \mathbf{w} \in N_P(\mathbf{f}(\bar{\mathbf{x}})) \\ -\nabla \mathbf{f}(\bar{\mathbf{x}})^* \mathbf{w} \in N_S(\bar{\mathbf{x}}) \end{cases} \qquad \Longrightarrow \qquad \mathbf{w} = \mathbf{0}$$

holds. Then

$$T_{S \cap \mathbf{f}^{-1}(P)}(\bar{\mathbf{x}}) = T_S(\bar{\mathbf{x}}) \cap \nabla \mathbf{f}(\bar{\mathbf{x}})^{-1} T_P(\mathbf{f}(\bar{\mathbf{x}})).$$

The proof uses proposition 66, as well as our metric regularity theorem 59, extended by proposition 60. Note that although S and P are convex, the feasible region $S \cap \mathbf{f}^{-1}(P)$ is nonconvex in general.

Corollary 68 *For closed, convex $S, P \subset \mathbf{E}$ with $\bar{\mathbf{x}} \in S \cap P$, if $N_P(\bar{\mathbf{x}}) \cap -N_S(\bar{\mathbf{x}}) = \{\mathbf{0}\}$, then $T_{P \cap S}(\bar{\mathbf{x}}) = T_P(\bar{\mathbf{x}}) + T_S(\bar{\mathbf{x}})$.*

The result follows from theorem 67 with $\mathbf{f}$ set to identity.

4.4 The Karush-Kuhn-Tucker conditions

We can now derive optimality conditions for the problem

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{f}(\mathbf{x}) \in P \\ & \mathbf{x} \in S, \end{array} \tag{4.1}$$

where $f : \mathbf{E} \to \mathbf{Y}$ is continuously differentiable at a local minimizer $\bar{\mathbf{x}}$, $\mathbf{f}$ is continuously differentiable, and the sets $S \subset \mathbf{E}$ and $P \subset \mathbf{Y}$ are closed and convex.

The constraint qualification in theorem 67 can be written in terms of tangent cones as

$$T_P(\mathbf{f}(\bar{\mathbf{x}})) - \nabla \mathbf{f}(\bar{\mathbf{x}}) T_S(\bar{\mathbf{x}}) = \mathbf{Y}, \tag{4.2}$$

which is known as the Robinson constraint qualification. Showing this requires the following facts:

- For any cones K, Q and linear A, $(K + A^* Q)^+ = K^+ \cap A^{-1} Q^+$.
- For any convex cone L, $L^+ = \{\mathbf{0}\}$ if and only if $L = \mathbf{Y}$.

Assuming $\bar{\mathbf{x}}$ is a local minimizer for problem 4.1, the basic optimality condition 64 is $\nabla f(\bar{\mathbf{x}}) \in T_{S \cap \mathbf{f}^{-1}(P)}(\bar{\mathbf{x}})^+$. If the Robinson condition also holds, this is equivalent to

$$\nabla f(\bar{\mathbf{x}}) \in \left(T_S(\bar{\mathbf{x}}) \cap \nabla \mathbf{f}(\bar{\mathbf{x}})^{-1} T_P(\mathbf{f}(\bar{\mathbf{x}})) \right)^+ .$$

Applying the Krein-Rutman polar cone calculus (a consequence of Fenchel duality) to the right-hand side, the optimality condition reduces to

$$-\nabla f(\bar{\mathbf{x}}) \in N_S(\bar{\mathbf{x}}) + \nabla \mathbf{f}(\bar{\mathbf{x}})^* N_P(\mathbf{f}(\bar{\mathbf{x}})).$$

This gives the following result.

Corollary 69 *Suppose $\bar{\mathbf{x}}$ is a local minimizer for problem (4.1). If the Robinson condition (4.2) holds, then*

$$-\nabla f(\bar{\mathbf{x}}) \in N_S(\bar{\mathbf{x}}) + \nabla \mathbf{f}(\bar{\mathbf{x}})^* N_P(\mathbf{f}(\bar{\mathbf{x}})).$$

These are the first-order necessary optimality conditions for the abstract nonlinear program (4.1). If P is described by a system of inequality and equality constraints, and if $S = \mathbf{E}$, we get the famous Karush-Kuhn-Tucker conditions.

Theorem 70 (Karush-Kuhn-Tucker) *Suppose $\bar{\mathbf{x}}$ is a local minimizer for*

$$\begin{array}{lll} \text{minimize} & f(\mathbf{x}) & \\ \text{subject to} & g_i(\mathbf{x}) \geq 0 & i = 1, \ldots, q \\ & \mathbf{h}(\mathbf{x}) = \mathbf{0}, & \end{array} \tag{4.3}$$

where f is differentiable at $\bar{\mathbf{x}}$ and $\mathbf{h}$ and the g_i are continuously differentiable. Define the active set $I = \{i \mid g_i(\bar{\mathbf{x}}) = 0\}$ and assume the Mangasarian-Fromowitz condition: $\nabla \mathbf{h}(\bar{\mathbf{x}})$ is surjective and there exists a $\mathbf{d} \in \operatorname{null}(\nabla \mathbf{h}(\bar{\mathbf{x}}))$ such that $\langle \nabla g_i(\bar{\mathbf{x}}), \mathbf{d}\rangle > 0$ for all $i \in I$.

Then there exist Lagrange multipliers $\boldsymbol{\lambda} \in \mathbf{R}^q$ and $\boldsymbol{\mu}$ such that

$$\nabla f(\bar{\mathbf{x}}) = \sum_{i \in I} \lambda_i \nabla g_i(\bar{\mathbf{x}}) + \nabla \mathbf{h}(\bar{\mathbf{x}})^* \boldsymbol{\mu}$$

and $\boldsymbol{\lambda} \geq \mathbf{0}$.

The Karush-Kuhn-Tucker theorem was originally proved by William Karush in his 1939 master's thesis at the University of Chicago. The result was largely unappreciated until Harold Kuhn and Albert Tucker rediscovered it in 1951 at Princeton University. The proof involves applying corollary 69 with $\mathbf{f}(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) & \dots & g_q(\mathbf{x}) & \mathbf{h}(\mathbf{x})^T \end{bmatrix}^T$ and $\mathbf{w} = \begin{bmatrix} \boldsymbol{\lambda}^T & \boldsymbol{\mu}^T \end{bmatrix}^T$.

Bibliography

[1] J.M. Borwein and A.S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples.* Second edition, Springer Science and Business Media (2006).

[2] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[3] A. Beck and M. Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." *SIAM Journal on Imaging Sciences* **2.1** (2009): 183-202.

[4] P. Tseng. "On accelerated proximal gradient methods for convex-concave optimization." Submitted to *SIAM Journal on Optimization* (2008).