CME 307: Optimization CME 307 / MS&E 311 / OIT 676

Lecture 14: Operators

Fall 2025 *Prof. Udell*

This unit develops a compact framework that unifies three important algorithms: proximal gradient (forward-backward splitting), Douglas-Rachford splitting, and ADMM. These algorithms are general enough to solve arbitrary convex optimization problems, including conic optimization problems, without the extremely high per-iteration computation complexity of interior point methods. They access the functions involved in the problem through one of two interfaces: either a *gradient* (first-order) oracle, or a *proximal* oracle, which generalized projection. With some abuse of terminology, these algorithms are generally called *first-order* algorithms, as they require no access to the Hessian.

We will understand these optimization algorithms by identifying solutions of convex optimization problems with fixed points of an operator $T : \mathbb{R}^n \to \mathbb{R}^n$. To provide a preview, we will show

$$\min_x f(x) + g(x) \quad \iff \quad 0 \in \partial f(x) + \partial g(x) \quad \iff \quad x = T(x).$$

When this operator T is either *contractive* or *averaged* (terms that we will define below), we will show convergence of the algorithm to a global optimum using this operator perspective.

1 Introduction and computational motivation

1.1 Motivation and examples

Many models decompose naturally into a smooth loss and a nonsmooth regularizer or constraint:

$$\min_x f(x) + g(x).$$

Typical instances:

- Lasso: $f(x) = \frac{1}{2}\|Ax - b\|^2$, $g(x) = \lambda\|x\|_1$.
- Box- or set-constrained least squares: $f(x) = \frac{1}{2}\|Ax - b\|^2$, $g(x) = \mathbf{1}_\Omega(x)$ for a convex set Ω.
- Total variation denoising: $f(x) = \frac{1}{2}\|x - b\|^2$, $g(x) = \lambda\|Dx\|_1$.

We will solve these by splitting the problem structure so each iteration evaluates either a gradient of f or a proximal operator (projection or shrinkage) for g.

We can even write a conic optimization problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \in K \end{array}$$

for some cone K, in a similar form:

$$\begin{array}{ll} \text{minimize} & c^T x + \mathbf{1}_K(z) \\ \text{subject to} & Ax = z. \end{array}$$

We will develop algorithms in this unit to solve all of the above problems efficiently, at least to moderate accuracy (1e-3 or 1e-4).

1.2 The proximal mapping

Definition 1.1 (Proximal operator)**.** For proper closed convex h and $\gamma > 0$,

$$\mathbf{prox}_{\gamma h}(z) = \operatorname*{\mathbf{argmin}}_x \left(h(x) + \tfrac{1}{2\gamma}\|x - z\|^2 \right).$$

Then $\mathbf{prox}_{\gamma h} = J_{\gamma \partial h}$.

Let's look at some examples.

Geometric picture (projection). For a convex set C, the proximal operator of the indicator function is the projection onto the set:

$$\mathbf{prox}_{\mathbf{1}_C}(z) = \mathbf{proj}_C(z) = \operatorname*{\mathbf{argmin}}_{x \in C} \|x - z\|^2.$$

Box constraints. The proximal operator of the indicator function of a box is clipping:

$$\mathbf{prox}_{\mathbf{1}_{[l,u]}}(z) = \min\{\max\{z, l\}, u\}$$

Soft-thresholding. The proximal operator of the ℓ_1 norm is soft-thresholding:

$$\mathbf{prox}_{\gamma|\cdot|_1}(z) = \operatorname{sign}(z) \cdot \max\{|z| - \gamma, 0\}$$

applied elementwise. Proof: from the optimality condition of the prox definition. First notice that the problem separates across coordinates, so we can consider a single coordinate z. The optimality condition is

$$0 \in \partial|x| + \tfrac{1}{\gamma}(x - z).$$

We consider three cases:

- If $x > 0$, then $\partial|x| = 1$, so $0 = 1 + \frac{1}{\gamma}(x - z)$ giving $x = z - \gamma$.
- If $x < 0$, then $\partial|x| = -1$, so $0 = -1 + \frac{1}{\gamma}(x - z)$ giving $x = z + \gamma$.
- If $x = 0$, then $\partial|x| = [-1, 1]$, so $0 \in [-1, 1] + \frac{1}{\gamma}(-z)$ giving $|z| \leq \gamma$.

Combining these cases gives the soft-thresholding formula.

1.3 Proximal gradient method

Suppose f is smooth, g is non-smooth. The proximal gradient method solves

$$\text{minimize} \quad f(x) + g(x)$$

using proximal operators together with gradient steps. The iteration is simple:

$$x^+ = \mathbf{prox}_{tg}(x - t\nabla f(x)).$$

- The proximal operator steps towards the minimum of g, and
- the gradient method steps towards minimum of f.

The lasso demo in class, also found at `https://github.com/stanford-cme-307/demos/`, demonstrates the power of this method. The remainder of the notes is devoted to analyzing this method and related methods, and showing their convergence.

2 Review

We first recall a few definitions from earlier units, and define a few more. In much of what follows, we'll need to assume functions are

Definition 2.1 (CCP function)**.**
- closed: $\mathbf{epi}(f)$ is a closed set
- convex: f is convex
- proper: $\mathbf{dom}\, f$ is non-empty

which we abbreviate as *CCP* (closed, convex, proper).

Definition 2.2 (Subdifferential)**.** For a proper closed convex function $h : \mathbb{R}^n \to \mathbb{R} \cup +\infty$, the subdifferential at x is

$$\partial h(x) = \{v \in \mathbb{R}^n : \ h(y) \geq h(x) + \langle v, y - x \rangle \text{ for all } y\}.$$

If h is differentiable at x, then $\partial h(x) = \nabla h(x)$.

We will use subdifferentials to express optimality conditions for convex optimization problems, such as the problem of minimizing a sum of CCP functions.

Proposition 2.3 (First-order optimality)**.** *For CCP functions $f, g : \mathbb{R}^n \to \mathbb{R}$, a point $x^\star$ minimizes $f + g$ if and only if*

$$0 \in \partial f(x^\star) + \partial g(x^\star).$$

3 Relations

Definition 3.1 (Relation)**.** A *relation* R on $\mathbb{R}^n$ is a subset of $\mathbb{R}^n \times \mathbb{R}^n$.

- We write $\mathbf{dom}\, R = \{x : (x, y) \in R\}$
- We let $R(x) = \{y : (x, y) \in R\}$
- If $R(x)$ is always empty or a singleton, we say R is a function. We often call such a relation a *map* or an *operator* to distinguish it from a real-valued function.

Any function $F : \mathbb{R}^n \to \mathbb{R}^n$ defines a relation $\{(x, F(x)) : x \in \mathbf{dom}\, F\}$.

Example 3.2 (Important relations)**.** The following relations will be important in what follows:

- empty relation: $\emptyset$
- full relation: $\mathbb{R}^n \times \mathbb{R}^n$
- identity: $\{(x, x) : x \in \mathbb{R}^n\}$
- zero: $\{(x, 0) : x \in \mathbb{R}^n\}$
- subdifferential: $\partial f = \{(x, g : x \in \mathbf{dom}\, f, g \in \partial f(x)\}$

More concretely, consider the subdifferential of the absolute value function:

$$\partial|x| = \begin{cases} \{-1\}, & x < 0 \\ [-1, 1], & x = 0 \\ \{1\}, & x > 0 \end{cases}$$

This is a relation that is not a function, since $\partial|0|$ contains more than one point. We write the associated relation concretely as

$$\partial|x| = \{(x, u) : u \in \partial|x|\} = \{(x, -1) : x < 0\} \cup \{(0, u) : u \in [-1, 1]\} \cup \{(x, 1) : x > 0\}.$$

Exercise 3.3. Plot the graph of the relation $\partial|x|$.

Definition 3.4 (Operations on relations)**.** If R and S are relations, define

- composition: $RS = \{(x, z) : (x, y) \in R, (y, z) \in S\}$
- addition: $R + S = \{(x, y + z) : (x, y) \in R, (x, z) \in S\}$
- inverses: $R^{-1} = \{(y, x) : (x, y) \in R\}$

We use inequality on sets to mean the inequality holds for any element in the set, *e.g.*,

$$f(y) \geq f(x) + \partial f^T(y - x).$$

Example 3.5 (Fenchel duality via relations)**.** Recall the definition of the Fenchel dual: $f^*(y) := \max_x(y^T x - f(x))$. We will prove $\partial f^* = \partial f^{-1}$.

To show this, recall that if f is CPP, $(f^*)^* = f^{**} = f$, so

$$\begin{aligned}
(u,v) \in (\partial f)^{-1} &\iff (v,u) \in \partial f \\
&\iff u \in \partial f(v) \\
&\iff 0 \in \partial f(v) - u \\
&\iff v \in \operatorname*{\mathbf{argmin}}_x (f(x) - u^T x) \\
&\iff v \in \operatorname*{\mathbf{argmax}}_x (u^T x - f(x)) \\
&\iff f(v) + f^*(u) = u^T v \\
&\iff u \in \operatorname*{\mathbf{argmax}}_y (y^T v - f^*(y)) \\
&\iff 0 \in v - \partial f^*(u) \\
&\iff (u,v) \in \partial f^*
\end{aligned}$$

Exercise 3.6. As a concrete application, use the result $\partial f^* = \partial f^{-1}$ to compute the Fenchel dual of the absolute value function $f(x) = |x|$.

4 Fixed points and zeros

The fixed-point viewpoint unifies many first-order algorithms. Instead of searching directly for the optimizer of a function, we search for a point that is left unchanged by an appropriately chosen operator. Throughout these notes we let $F : \mathbb{R}^n \to \mathbb{R}^n$ denote a (possibly set-valued) operator.

Definition 4.1 (Zero of a relation)**.** Let R be a relation. A point x is a *zero* of R if $0 \in R(x)$, i.e., the graph of R contains $(x, 0)$. The *zero set* of R is $R^{-1}(0) = \{x : (x, 0) \in R\}$.

Every convex optimization problem can be written as the search for a zero of a subdifferential. Indeed, x minimizes f precisely when $0 \in \partial f(x)$. This equivalence gives us a dictionary between optimality conditions and zero problems.

Example 4.2 (Zeros arising in optimization)**.** Consider $f(x) = \frac{1}{2}\|Ax - b\|^2$ with full-column-rank A. The optimality condition is $0 \in \nabla f(x) = A^T(Ax - b)$, so the zero set is the affine subspace $\{x : A^T Ax = A^T b\}$. For nonsmooth functions such as $f(x) = \lambda\|x\|_1$, the zero condition becomes $0 \in \lambda\partial\|x\|_1$.

4.1 Lipschitz, nonexpansive, and contractive operators

Definition 4.3 (Lipschitz operator)**.** A relation F is *Lipschitz* with constant $L \geq 0$ if $\|u - v\| \leq L\|x - y\|$ for all $(x, u), (y, v) \in F$. We call F *nonexpansive* when $L \leq 1$ and *contractive* when $L < 1$.

When F is Lipschitz, the inequality with $x = y$ implies $u = v$, so F must in fact be a function rather than a multi-valued relation. Nonexpansive maps preserve distances while contractive maps actively shrink them.

Example 4.4 (A small catalog of Lipschitz maps)**.**

- Rotations and translations in $\mathbb{R}^n$ are nonexpansive: they change angles or positions but never stretch vectors.
- The map $x \mapsto cx$ with $|c| < 1$ is contractive since it scales every distance by $|c|$.
- Projections onto a closed convex set C satisfy $\|\mathbf{proj}_C(x) - \mathbf{proj}_C(y)\| \le \|x - y\|$, so they are nonexpansive.

Proposition 4.5 (Gradient update Lipschitz constant)**.** *Assume f is α-strongly convex and β-smooth. For any step size $t > 0$ the gradient update*

$$G_t(x) = x - t\nabla f(x)$$

is Lipschitz with constant $L = \max\{|1 - t\alpha|, |1 - t\beta|\}$, hence it is contractive whenever $t \in (0, 2/(\alpha + \beta))$.

Proof. By the fundamental theorem of calculus, $G_t(x) - G_t(y) = \int_0^1 (I - t\nabla^2 f(\theta x + (1-\theta)y))(x - y)\, d\theta$. Taking norms and using Jensen's inequality gives

$$\|G_t(x) - G_t(y)\| \le \int_0^1 \|I - t\nabla^2 f(\theta x + (1-\theta)y)\|\, \|x - y\| d\theta.$$

The Hessian satisfies $\alpha I \preceq \nabla^2 f \preceq \beta I$, so the integrand is bounded by $\max\{|1 - t\alpha|, |1 - t\beta|\}\|x - y\|$, yielding the stated Lipschitz constant. □

Corollary 4.6 (Best constant step)**.** *Choosing $t = 2/(\alpha + \beta)$ minimizes the Lipschitz constant and gives $L = (\kappa - 1)/(\kappa + 1)$ where $\kappa = \beta/\alpha$ is the condition number.*

4.2 Proximal maps

Proposition 4.7 (Proximal maps are (firmly) nonexpansive)**.** *For any closed convex function f the proximal operator $\mathbf{prox}_f$ satisfies $\|\mathbf{prox}_f(x) - \mathbf{prox}_f(y)\| \le \|x - y\|$.*

Proof. Let $u = \mathbf{prox}_f(x)$ and $v = \mathbf{prox}_f(y)$. Then $x - u \in \partial f(u)$ and $y - v \in \partial f(v)$. The subgradient inequality applied twice gives $f(v) \ge f(u) + \langle x - u, v - u\rangle$ and $f(u) \ge f(v) + \langle y - v, u - v\rangle$. Summing yields $\langle x - y, u - v\rangle \ge \|u - v\|^2$, so Cauchy-Schwarz implies $\|u - v\| \le \|x - y\|$. □

The intermediate inequality shows that $\mathbf{prox}_f$ is *firmly* nonexpansive: $\langle x - y, u - v\rangle \ge \|u - v\|^2$. This property is stronger than nonexpansiveness and will be useful later.

Proposition 4.8 (Strongly convex prox is contractive)**.** *If f is α-strongly convex, then* $\mathbf{prox}_f$ *is $(1+2\alpha)^{-1}$-contractive.*

Proof. With the same notation as above, strong convexity produces the modified inequalities

$$f(v) \geq f(u) + \langle x-u, v-u\rangle + \alpha\|v-u\|^2 \qquad f(u) \geq f(v) + \langle y-v, u-v\rangle + \alpha\|u-v\|^2.$$

Adding and simplifying gives

$$\langle x-y, u-v\rangle \geq (1+2\alpha)\|u-v\|^2,$$

hence $\|u-v\| \leq (1+2\alpha)^{-1}\|x-y\|$. □

4.3 Fixed points

Definition 4.9 (Fixed point)**.** A point x is a *fixed point* of F if $F(x) = x$. The set $\mathrm{Fix}(F) = \{x : F(x) = x\}$ plays the role of the solution set.

Example 4.10 (Fixed points of simple maps)**.**

- For the identity map $F(x) = x$ every point is a fixed point.
- For the constant map $F(x) = 0$ the unique fixed point is 0.
- A translation $F(x) = x + a$ with $a \neq 0$ is nonexpansive but has no fixed point, illustrating that nonexpansive maps need not have fixed points.

Proposition 4.11 (Fixed point of proximal gradient map)**.** *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and $g : \mathbb{R}^n \to \mathbb{R}$ is proxable, and let $t > 0$. Any fixed point of the proximal gradient iteration minimizes $f + g$.*

Proof. Use the definition of the proximal operator to check

$$x = \mathbf{prox}_{tg}(x) \qquad \Longleftrightarrow \qquad x' = (I + t\partial g)^{-1}(x)$$

The point x minimizes $f + g$ iff

$$\begin{aligned} 0 &\in \nabla f(x) + \partial g(x) \\ x &\in t\nabla f(x) + x + t\partial g(x) \\ (I - t\nabla f)(x) &\in (I + t\partial g)(x) \\ x &\in (I + t\partial g)^{-1}(I - t\nabla f)(x) \end{aligned}$$

□

Proposition 4.12 (Uniqueness for contractions)**.** *If F is contractive, then it has at most one fixed point.*

Proof. Suppose both x and y are fixed points. Then $\|x - y\| = \|F(x) - F(y)\| \leq L\|x - y\|$ with $L < 1$, forcing $x = y$. □

Theorem 4.13 (Banach fixed point)**.** *Let F be a contraction with constant $L < 1$. The iteration $x^{k+1} = F(x^k)$ converges from any starting point x^0 to the unique fixed point $x^\star$, and the errors satisfy $\|x^k - x^\star\| \leq L^k\|x^0 - x^\star\|$.*

Proof. Because $\|x^{k+1} - x^k\| \leq L\|x^k - x^{k-1}\|$, a telescoping sum shows the sequence is Cauchy. Let $x^\star$ be the limit. Continuity of F gives $x^\star = \lim_{k\to\infty} F(x^{k-1}) = F(x^\star)$, so $x^\star$ is the desired fixed point. Finally, $\|x^k - x^\star\| = \|F(x^{k-1}) - F(x^\star)\| \leq L\|x^{k-1} - x^\star\|$, which recursively yields the linear rate. □

The iterates of a contraction are Fejér monotone: the distance to $x^\star$ never increases. This property will continue to hold for averaged operators, giving intuition for why damped iterations are stable.

5 Averaged Operators

We have seen that many important maps are nonexpansive, but not contractive. To develop algorithms that make use of these operators, but guarantee convergence, we introduce averaged operators. Averaged operators interpolate between a nonexpansive map and the identity, retaining enough damping to guarantee convergence while permitting more general behavior.

Definition 5.1 (Averaged operator)**.** An operator F is *averaged* if there exists $\theta \in (0, 1)$ and a nonexpansive map G such that $F = (1 - \theta)I + \theta G$. When $\theta = 1/2$ the map is *firmly* nonexpansive.

Geometrically, an averaged step moves partway towards the action of G and partway towards the previous iterate, preventing the oscillations that plague undamped nonexpansive maps such as rotations.

Example 5.2 (Why nonexpansive needs damping)**.** Let G rotate points in $\mathbb{R}^2$ by ninety degrees. Then $\|Gx\| = \|x\|$, so G is nonexpansive, but unless x^0 starts at the origin the iteration $x^{k+1} = Gx^k$ simply spins around the origin. The averaged map $F = \frac{1}{2}(I + G)$ shrinks the radius each step, and x^k spirals toward the unique fixed point at the origin.

Proposition 5.3 (Fixed points of averaged maps)**.** *If $F = (1 - \theta)I + \theta G$ with $0 < \theta < 1$, then x is a fixed point of F if and only if it is a fixed point of G.*

Proof. The equality $x = Fx$ expands to $x = (1 - \theta)x + \theta Gx$, so $\theta(x - Gx) = 0$. Since $\theta > 0$, we conclude $x = Gx$. The reverse implication is immediate. □

Theorem 5.4 (Krasnosel'skii-Mann iteration)**.** *Let $F = (1-\theta)I + \theta G$ be averaged with a fixed point. Then the fixed-point iteration $x^{k+1} = F(x^k)$ converges to a fixed point of F for every starting point. Moreover, the residual obeys*

$$\|Gx^{(k)} - x^{(k)}\|^2 \leq \frac{\|x^{(0)} - x^\star\|^2}{(k+1)\theta(1-\theta)}.$$

Although the bound is sublinear, the Fejér property ensures the iterates remain stable. In practice we often combine averaging with momentum or relaxation to accelerate convergence.

5.1 Averaged gradient maps

Proposition 5.5 (Gradient descent is averaged)**.** *If f is β-smooth, then $I - \frac{2}{\beta}\nabla f$ is nonexpansive. Hence $I - t\nabla f$ is averaged for every $t \in (0, 2/\beta)$.*

Proof. Smoothness implies $\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\|$. We compute

$$\begin{aligned}\|(I - \frac{2}{\beta}\nabla f)(x) - (I - \frac{2}{\beta}\nabla f)(y)\|^2 &= \|x-y\|^2 - \frac{4}{\beta}\langle x - y, \nabla f(x) - \nabla f(y)\rangle \\ &\quad + \frac{4}{\beta^2}\|\nabla f(x) - \nabla f(y)\|^2 \\ &\leq \|x-y\|^2\end{aligned}$$

□

6 Proximal Gradient Method

We are now ready to analyze the proximal gradient method for minimizing $f+g$ where f is smooth and g is CCP. We have already established that fixed points of the proximal gradient map

$$T(x) = \mathbf{prox}_{tg}(x - t\nabla f(x))$$

are minimizers of $f + g$. To show convergence of the iteration $x^{k+1} = T(x^k)$, we need to show that T is contractive (for linear convergence) or averaged (for sublinear convergence).

Proposition 6.1 (Proximal-gradient convergence)**.** *Let f be β-smooth and g be CCP. For any $t \in (0, 2/\beta)$ the proximal-gradient map*

$$T(x) = \mathbf{prox}_{tg}(x - t\nabla f(x))$$

is averaged, and the iteration $x^{k+1} = T(x^k)$ converges to a minimizer of $f + g$. If either f or g is strongly convex, the convergence is linear.

Proof. The operator $I - t\nabla f$ is averaged by the previous proposition, while $\mathbf{prox}_{tg}$ is firmly nonexpansive. The composition of a firmly nonexpansive map with an averaged map is averaged, establishing convergence by the Krasnosel'skii-Mann theorem. Strong convexity of f makes $I - t\nabla f$ contractive; strong convexity of g makes $\mathbf{prox}_{tg}$ contractive, and either case transfers the linear rate to T. □

Example 6.2 (Interpreting the theory)**.**

- **LASSO.** Here $f(x) = \frac{1}{2}\|Ax - b\|^2$ is smooth with $\beta = \|A\|_2^2$, $g(x) = \lambda\|x\|_1$ admits the soft-thresholding prox, and the theory predicts $O(1/k)$ convergence with step sizes below $2/\|A\|_2^2$. Adding an ℓ_2 term renders g strongly convex and yields linear convergence.
- **Box-constrained least squares.** Taking g to be the indicator of $[l, u]$ shows proximal gradient alternates between a gradient step and a simple projection, again converging globally.
- **Logistic regression with ℓ_1 penalty.** The gradient of the logistic loss is Lipschitz with constant given by the largest singular value of the design matrix, so the theory prescribes a safe step size and guarantees convergence of iterates often used in practice.