Monotone Operators

Stephen Boyd (with help from Neal Parikh)

EE364b, Stanford University

Outline

Relations

- a *relation* R on a set $\mathbf{R}^n$ is a subset of $\mathbf{R}^n \times \mathbf{R}^n$
- $\mathbf{dom}\, R = \{x \mid \exists y\ (x, y) \in R\}$
- overload $R(x)$ to mean the *set* $R(x) = \{y \mid (x, y) \in R\}$
- can think of R as 'set-valued mapping', *i.e.*, from $\mathbf{dom}\, R$ into $2^{\mathbf{R}^n}$
- when $R(x)$ is always empty or a singleton, we say R is a function
- any function (or operator) $f : C \to \mathbf{R}^n$ with $C \subseteq \mathbf{R}^n$ is a relation ($f(x)$ is then ambiguous: it can mean $f(x)$ or $\{f(x)\}$)

Examples

- empty relation: $\emptyset$
- full relation: $\mathbf{R}^n \times \mathbf{R}^n$
- identity: $I = \{(x, x) \mid x \in \mathbf{R}^n\}$
- zero: $0 = \{(x, 0) \mid x \in \mathbf{R}^n\}$
- $\{x \in \mathbf{R}^2 \mid x_1^2 + x_2^2 = 1\}$
- $\{x \in \mathbf{R}^2 \mid x_1 \leq x_2\}$
- *subdifferential relation*: $\partial f = \{(x, \partial f(x)) \mid x \in \mathbf{R}^n\}$

Operations on relations

- *inverse* (relation): $R^{-1} = \{(y, x) \mid (x, y) \in R\}$
  - inverse exists for any relation
  - coincides with inverse function, when inverse function exists
- *composition*: $RS = \{(x, y) \mid \exists z \ (x, z) \in S, \ (z, y) \in R\}$
- *scalar multiplication*: $\alpha R = \{(x, \alpha y) \mid (x, y) \in R\}$
- *addition*: $R + S = \{(x, y + z) \mid (x, y) \in R, \ (x, z) \in S\}$

Example: Resolvent of operator

for relation R and $\lambda \in \mathbf{R}$, *resolvent* (much more on this later) is relation

$$S = (I + \lambda R)^{-1}$$

- $I + \lambda R = \{(x, x + \lambda y) \mid (x, y) \in R\}$
- $S = (I + \lambda R)^{-1} = \{(x + \lambda y, x) \mid (x, y) \in R\}$
- for $\lambda \neq 0$, $S = \{(u, v) \mid (u - v)/\lambda \in R(v)\}$

Generalized equations

- goal: solve *generalized equation* $0 \in R(x)$
- *i.e.*, find $x \in \mathbf{R}^n$ with $(x, 0) \in R$
- *solution set* or *zero set* is $X = \{x \in \mathbf{dom}\, R \mid 0 \in R(x)\}$
- if $R = \partial f$ and $f : \mathbf{R}^n \to \mathbf{R}^n$, then $0 \in R(x)$ means x minimizes f

Outline

Monotone operators

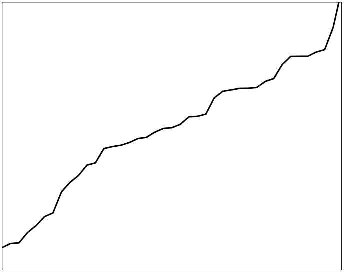
- relation F on $\mathbf{R}^n$ is *monotone* if

$$(u-v)^T(x-y) \geq 0 \quad \text{for all } (x,u),\ (y,v) \in F$$

- F is *maximal monotone* if there is no monotone operator that properly contains it
- we'll be informal (*i.e.*, sloppy) about maximality, other analysis issues
- solving generalized equations with maximal monotone operators subsumes many useful problems

Maximal monotone operators on R

F is maximal monotone iff it is a connected curve with no endpoints, with nonnegative (or infinite) slope



Some basic properties

suppose F and G are monotone

- *sum*: $F + G$ is monotone
- *nonnegative scaling*: if $\alpha \geq 0$, then αF is monotone
- *inverse*: F^{-1} is monotone
- *congruence*: for $T \in \mathbf{R}^{n \times m}$, $T^T F(Tz)$ is monotone (on $\mathbf{R}^m$)
- *zero set*: $\{x \in \mathbf{R}^n \mid 0 \in F(x)\}$ is convex if F is maximal monotone

affine function $F(x) = Ax + b$ is monotone iff $A + A^T \succeq 0$

Subdifferential

$F(x) = \partial f(x)$ is monotone

- suppose $u \in \partial f(x)$ and $v \in \partial f(y)$
- then

$$f(y) \geq f(x) + u^T(y - x), \qquad f(x) \geq f(y) + v^T(x - y)$$

- add these and cancel $f(y) + f(x)$ to get

$$0 \leq (u - v)^T(x - y)$$

if f is convex closed proper (CCP) then $F(x) = \partial f(x)$ is maximal monotone

KKT operator

- equality-constrained convex problem (with $A \in \mathbf{R}^{m \times n}$)

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

  with Lagrangian $L(x, y) = f(x) + y^T(Ax - b)$
- associated *KKT operator* on $\mathbf{R}^n \times \mathbf{R}^m$:

$$F(x, y) = \begin{bmatrix} \partial_x L(x, y) \\ -\partial_y L(x, y) \end{bmatrix} = \begin{bmatrix} \partial f(x) + A^T y \\ b - Ax \end{bmatrix} = \begin{bmatrix} r^{\text{dual}} \\ -r^{\text{pri}} \end{bmatrix}$$

- zero set of F is set of primal-dual optimal points (saddle points of L)
- KKT operator is monotone: write as sum of monotone operators

$$F(x, y) = \begin{bmatrix} \partial f(x) \\ b \end{bmatrix} + \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Multiplier to residual mapping

- same equality-constrained convex problem
- define $F(y) = b - Ax$ with $x \in \operatorname{argmin}_z L(z, y)$ (can be set-valued)
- $-F(y)$ is primal residual obtained from dual variable y
- interpretation: $F(y) = \partial(-g)(y)$, where g is dual function
- zero set is set of dual optimal points
- multiplier to residual mapping F is monotone
- quick proof: $F(y) = b - A(\partial f)^{-1}(-A^T y)$ (or use $F(y) = \partial(-g)(y)$)

Outline

Nonexpansive and contractive operators

- F has *Lipschitz constant* L if

$$\|F(x) - F(y)\|_2 \leq L\|x - y\|_2 \quad \text{for all } x,\ y \in \mathbf{dom}\, F$$

- for $L = 1$, we say F is *nonexpansive*
- for $L < 1$, we say F is a *contraction* (with contraction factor L)

Properties

- if F and G have Lipschitz constant L, so does

$$\theta F + (1-\theta)G, \qquad \theta \in [0,1]$$

- composition of nonexpansive operators is nonexpansive
- composition of nonexpansive operator and contraction is contraction
- *fixed point set* of nonexpansive F (with $\mathbf{dom}\, f = \mathbf{R}^n$)

$$\{x \mid F(x) = x\}$$

  is convex (but can be empty)
- a contraction has a single fixed point (more later)

Outline

Resolvent and Cayley operator

- for $\lambda \in \mathbf{R}$, *resolvent* of relation F is

$$R = (I + \lambda F)^{-1}$$

- when $\lambda \geq 0$ and F monotone, R is nonexpansive (thus a function)
- when $\lambda \geq 0$ and F maximal monotone, $\mathbf{dom}\, R = \mathbf{R}^n$
- *Cayley operator* of F is

$$C = 2R - I = 2(I + \lambda F)^{-1} - I$$

- when $\lambda \geq 0$ and F monotone, C is nonexpansive
- we write R_F and C_F to explicitly show dependence on F

Proof that C is nonexpansive

assume $\lambda > 0$ and F monotone

- suppose $(x, u) \in R$ and $(y, v) \in R$, *i.e.*,

$$u + \lambda F(u) \ni x, \qquad v + \lambda F(v) \ni y$$

- subtract to get $u - v + \lambda(F(u) - F(v)) \ni x - y$
- multiply by $(u - v)^T$ and use monotonicity of F to get

$$\|u - v\|_2^2 \leq (x - y)^T (u - v)$$

- so when $x = y$, we must have $u = v$ (*i.e.*, R is a function)

Proof (continued)

- now let's show C is nonexpansive:

$$\begin{aligned}\|C(x)-C(y)\|_2^2 &= \|(2u-x)-(2v-y)\|_2^2 \\ &= \|2(u-v)-(x-y)\|_2^2 \\ &= 4\|u-v\|_2^2 - 4(u-v)^T(x-y) + \|x-y\|_2^2 \\ &\leq \|x-y\|_2^2\end{aligned}$$

  using inequality above

- R is nonexpansive since it is the average of I and C:

$$R = (1/2)I + (1/2)(2R-I)$$

Example: Linear operators

- linear mapping $F(x) = Ax$ is
  - monotone iff $A + A^T \succeq 0$
  - nonexpansive iff $\|A\|_2 \leq 1$
- $\lambda \geq 0$ and $A + A^T \succeq 0 \Longrightarrow$
  - $I + \lambda A$ nonsingular
  - $\|R_A\|_2 = \|(I + \lambda A)^{-1}\|_2 \leq 1$
  - $\|C_A\|_2 = \|2(I + \lambda A)^{-1} - I\|_2 \leq 1$
- for matrix case, we have alternative formula for Cayley operator:

$$2(I + \lambda A)^{-1} - I = (I + \lambda A)^{-1}(I - \lambda A)$$

cf. bilinear function $\dfrac{1 - \lambda a}{1 + \lambda a}$, which maps

$$\{s \in \mathbf{C} \mid \Re s \geq 0\} \quad \text{into} \quad \{s \in \mathbf{C} \mid |s| \leq 1\}$$

Resolvent of subdifferential: Proximal mapping

- suppose $z = (I + \lambda \partial f)^{-1}(x)$, with $\lambda > 0$, f convex
- this means $z + \lambda \partial f(z) \ni x$
- rewrite as
$$0 \in \partial_z \left(f(z) + (1/2\lambda)\|z - x\|_2^2 \right)$$
which is the same as
$$z = \operatorname*{argmin}_u \left(f(u) + (1/2\lambda)\|u - x\|_2^2 \right)$$
- RHS called *proximal mapping* of f, denoted $\mathbf{prox}_{\lambda f}(x)$

Example: Indicator function

- take $f = I_C$, indicator function of convex set C
- ∂f is the *normal cone operator*

$$N_C(x) = \begin{cases} \emptyset & x \notin C \\ \{w \mid w^T(z-x) \leq 0 \;\; \forall z \in C\} & x \in C \end{cases}$$

- proximal operator of f (*i.e.*, resolvent of N_C) is

$$(I + \lambda \partial I_C)^{-1}(x) = \operatorname*{argmin}_u \left(I_C(u) + (1/2\lambda)\|u - x\|_2^2 \right) = \Pi_C(x)$$

where Π_C is (Euclidean) projection onto C

Resolvent of multiplier to residual map

- take F to be multiplier to residual mapping for convex problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

- $F(y) = b - Ax$ where $x \in \operatorname{argmin}_w L(w, y)$
- $z = (I + \lambda F)^{-1}(y)$ means $z + \lambda F(z) \ni y$
- $z + \lambda(b - Ax) = y$ for some $x \in \operatorname{argmin}_w L(w, z)$
- write as

$$z = y + \lambda(Ax - b), \qquad \partial f(x) + A^T z \ni 0$$

Resolvent of multiplier to residual map

- write second term as $\partial f(x) + A^T y + \lambda A^T(Ax - b) \ni 0$, so

$$x \in \operatorname*{argmin}_w \left(f(w) + y^T(Aw - b) + (\lambda/2)\|Aw - b\|_2^2 \right)$$

- function on right side is *augmented Lagrangian* for the problem
- so $z = R(y)$ can be found as

$$\begin{aligned} x &:= \operatorname*{argmin}_w \left(f(w) + y^T(Aw - b) + (\lambda/2)\|Aw - b\|_2^2 \right) \\ z &:= y + \lambda(Ax - b) \end{aligned}$$

Fixed points of Cayley and resolvent operators

- assume F is maximal monotone, $\lambda > 0$
- solutions of $0 \in F(x)$ are fixed points of R:

$$F(x) \ni 0 \iff x + \lambda F(x) \ni x \iff x = (I + \lambda F)^{-1}(x) = R(x)$$

- solutions of $0 \in F(x)$ are fixed points of C:

$$x = R(x) \iff x = 2R(x) - x = C(x)$$

- key result: we can solve $0 \in F(x)$ by finding fixed points of C or R
- next: how to actually find these fixed points

Outline

Contraction mapping theorem

- also known as *Banach fixed point theorem*
- assume F is contraction, with Lipschitz constant $L < 1$, $\mathbf{dom}\, F = \mathbf{R}^n$
- the iteration

  $$x^{k+1} := F(x^k)$$

  converges to the unique fixed point of F

- proof (sketch):
  - sequence x^k is Cauchy: $\|x^{k+m} - x^k\|_2 \leq \|x^{k+1} - x^k\|_2/(1-L)$
  - hence converges to a point $x^\star$
  - $x^\star$ must be (the) fixed point

Example: Gradient method

- assume f is convex, $mI \preceq \nabla^2 f(x) \preceq LI$
  (*i.e.*, f strongly convex, ∇f Lipschitz)
- gradient method is

$$x^{k+1} := x^k - \alpha \nabla f(x^k) = F(x^k)$$

  (fixed points are exactly solutions of $F(x) = x$)
- $DF(x) = I - \alpha \nabla^2 f(x)$
- F is a Lipschitz with parameter $\max\{|1 - \alpha m|, |1 - \alpha L|\}$
- F is a contraction when $0 < \alpha < 2/L$
- so gradient method converges (geometrically) when $0 < \alpha < 2/L$

Damped iteration of a nonexpansive operator

- suppose F is nonexpansive, $\mathbf{dom}\, F = \mathbf{R}^n$, with fixed point set $X = \{x \mid F(x) = x\}$
- can have $X = \emptyset$ (*e.g.*, translation)
- simple iteration of F need not converge, even when $X \neq \emptyset$ (*e.g.*, rotation)
- *damped* iteration:

$$x^{k+1} := (1 - \theta^k)x^k + \theta^k F(x^k)$$

  $\theta^k \in (0, 1)$
- important special case: $\theta^k = 1/2$ (more later)
- another special case: $\theta^k = 1/(k+1)$, which gives simple averaging

$$x^k = \frac{1}{k+1}\left(x^0 + \cdots + F(x^{k-2}) + F(x^{k-1})\right)$$

Convergence results

- assume F is nonexpansive, $\mathbf{dom}\, F = \mathbf{R}^n$, $X \neq \emptyset$, and

$$\sum_{k=0}^{\infty} \theta^k (1 - \theta^k) = \infty$$

  (which holds for special cases above)

- then we have

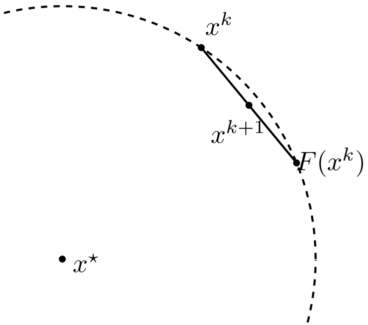
$$\min_{j=0,\ldots,k} \mathbf{dist}(x^j, X) \to 0$$

  i.e., (some) iterates get arbitrarily close to fixed point set, and

$$\min_{j=0,\ldots,k} \|F(x^j) - x^j\|_2 \to 0$$

  i.e., (some) iterates yield arbitrarily good 'almost fixed points'

Idea of proof



- $F(x^k)$ is no farther from $x^\star$ than x^k is (by nonexpansivity)
- so x^{k+1} is *closer* to $x^\star$ than x^k is

Proof

- start with identity

$$\|\theta a + (1-\theta)b\|_2^2 = \theta\|a\|_2^2 + (1-\theta)\|b\|_2^2 - \theta(1-\theta)\|b-a\|_2^2$$

- apply to $x^{k+1} - x^\star = (1-\theta^k)(x^k - x^\star) + \theta^k(F(x^k) - x^\star)$:

$$\begin{aligned} &\|x^{k+1} - x^\star\|_2^2 \\ &\quad = \quad (1-\theta^k)\|x^k - x^\star\|_2^2 + \theta^k\|F(x^k) - x^\star\|_2^2 - \theta^k(1-\theta^k)\|F(x^k) - x^k\|_2^2 \\ &\quad \leq \quad \|x^k - x^\star\|_2^2 - \theta^k(1-\theta^k)\|F(x^k) - x^k\|_2^2 \end{aligned}$$

using $\|F(x^k) - x^\star\|_2 \leq \|x^k - x^\star\|_2$

Proof (continued)

- iterate inequality to get

$$\sum_{j=0}^{k} \theta^j(1-\theta^j)\|F(x^j)-x^j\|_2^2 \leq \|x^0-x^\star\|_2^2 - \|x^{k+1}-x^\star\|_2^2$$

- if $\|F(x^j)-x^j\|_2 \geq \epsilon$ for $j=0,\ldots,k$, then

$$\epsilon^2 \leq \frac{\|x^0-x^\star\|_2^2}{\sum_{j=0}^{k}\theta^j(1-\theta^j)}$$

- RHS goes to zero as $k \to \infty$

Outline

Damped Cayley iteration

- want to solve $0 \in F(x)$ with F maximal monotone
- damped Cayley iteration:

$$\begin{aligned} x^{k+1} &:= (1-\theta^k)x^k + \theta^k C(x^k) \\ &= (1-\theta^k)x^k + \theta^k(2R(x^k) - I(x^k)) \\ &= (1-2\theta^k)x^k + 2\theta^k R(x^k) \end{aligned}$$

  with $\theta^k \in (0,1)$ and $\sum_k \theta^k(1-\theta^k) = \infty$
- converges (assuming $X \neq \emptyset$) in sense given above
- important: requires ability to evaluate resolvent map of F

Proximal point algorithm

- take $\theta^k = 1/2$ in damped Cayley iteration
- gives *resolvent iteration* or *proximal point algorithm*:

$$x^{k+1} := R(x^k) = (I + \lambda F)^{-1}(x^k)$$

- if $F = \partial f$ with f convex, yields *proximal minimization algorithm*

$$x^{k+1} := \mathbf{prox}_{f,1/\lambda}(x^k) = \underset{x}{\operatorname{argmin}} \left(f(x) + (1/2\lambda)\|x - x^k\|_2^2 \right)$$

  can interpret as quadratic regularization that goes away in limit
- many classical algorithms are just proximal point method applied to appropriate maximal monotone operator

Method of multipliers

- take F to be multiplier to residual mapping for

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

- $F(y) = b - Ax$ with $x \in \operatorname{argmin}_z L(z, y)$
- proximal point algorithm becomes *method of multipliers*:

$$\begin{aligned} x^{k+1} &:= \operatorname*{argmin}_w \left(f(w) + (y^k)^T (Aw - b) + (\lambda/2)\|Aw - b\|_2^2 \right) \\ y^{k+1} &:= y^k + \lambda(Ax^{k+1} - b) \end{aligned}$$

Method of multipliers

- first step is augmented Lagrangian minimization
- second step is dual variable update
- y^k converges to an optimal dual variable
- primal residual $Ax^k - b$ converges to zero

Method of multipliers dual update

- optimality conditions (primal and dual feasibility):

$$Ax - b = 0, \qquad \partial f(x) + A^T y \ni 0$$

- from definition of x^{k+1} we have

$$\begin{aligned} 0 &\in \partial f(x^{k+1}) + A^T y^k + \lambda A^T (Ax^{k+1} - b) \\ &= \partial f(x^{k+1}) + A^T y^{k+1} \end{aligned}$$

- so dual update makes (x^{k+1}, y^{k+1}) dual feasible
- primal feasibility occurs in limit as $k \to \infty$

Comparison with dual (sub)gradient method

method of multipliers

- like dual method, but with augmented Lagrangian, specific step size
- converges under *far more general* conditions than dual subgradient
- f need not be strictly convex, or differentiable
- f can take on value $+\infty$
- but not amenable to decomposition (more later ...)