Proximity Operator

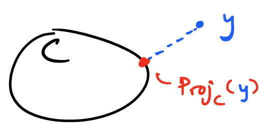
Thursday, February 11, 2021 10:43 PM

FACT: applying a monotonic transformation to the objective doesn't change location of minimizer. USEFUL TRICK!

Def The orthogonal projection of a point y onto a set C is

$$\mathrm{Proj}_C(y) := \underset{x \in C}{\mathrm{argmin}} \; \|x - y\|_2 = \underset{x \in C}{\mathrm{argmin}} \; \tfrac{1}{2}\|x - y\|_2^2$$

(e.g. exists and unique if C is a Chebyshev set, i.e., closed and convex)



Def The proximity operator or prox of a function $f \in \Gamma_0(\mathbb{R}^n)$ is

$$\mathrm{prox}_f(y) := \underset{x}{\mathrm{argmin}} \; \tfrac{1}{2}\|x - y\|_2^2 + f(x)$$

$\boxed{Q1}$ Is this argmin single valued? Why or why not?
Does it even exist?

$\boxed{A1}$ f is convex, $\frac{1}{2}\|\cdot - y\|^2$ is strongly convex $\Rightarrow$ objective is strongly convex

so $\exists!$ unique minimizer

Ex Let $f = I_C$ be an indicator function, then

$$\mathrm{prox}_f(y) = \mathrm{Proj}_C(y)$$

Note Often convenient to include a scaling, so

$$\mathrm{prox}_{tf}(y) := \underset{x}{\mathrm{argmin}} \; \tfrac{1}{2}\|x-y\|^2 + t f(x) = \mathrm{argmin} \; \frac{1}{2t}\|x-y\|^2 + f(x)$$

Ex $f(x) = \frac{1}{2}\|x\|^2$

$$\text{prox}_{tf}(y) = \underset{x}{\text{argmin}}\ \tfrac{1}{2}\|x-y\|^2 + \tfrac{t}{2}\|x\|^2$$

so solve $0 = (x-y) + t\,x$, so $\underline{\underline{x = (1+t)^{-1}y}}$.

Btw What is $\nabla f(x)$ if $f(x) = \frac{1}{2}\|Ax-b\|^2$? (i.e. for HW4)

think of as $f = g \circ h$, $g(x) = \frac{1}{2}\|x\|^2 = \frac{1}{2}\sum_{i=1}^{n} x_i^2$

so $\frac{\partial g}{\partial x_j} = x_j$, i.e. $\nabla g(x) = x$

$h(x) = A\cdot x - b$

$\text{Jac}_h(x) = A^T$

<u>Chain rule in 1D</u> $(g\circ h)'(x) = g'(h(x))\cdot h'(x)$

$= h'(x)\cdot g'(h(x))$

<u>Chain rule in n-D</u>

Be careful about orders and transposes!

$$(g\circ h)'(x) = \text{Jac}_h(x)\cdot \nabla g(h(x))$$

so in our case $\nabla\left(\frac{1}{2}\|Ax-b\|^2\right) = A^T\cdot(Ax-b)$

Ex $f(x) = \|x\|_1$

$$\text{prox}_{t\|\cdot\|_1}(y) = \underset{x}{\text{argmin}}\ \tfrac{1}{2}\|x-y\|^2 + t\|x\|_1$$

$$= \underset{x}{\text{argmin}}\ \tfrac{1}{2}\sum_{i=1}^{n}(x_i-y_i)^2 + t\sum_{i=1}^{n}|x_i|$$

SEPARABLE? yes ✓

So solve for each component x_i independently

i.e., find $\underset{x_i}{\text{argmin}}\ \underbrace{\tfrac{1}{2}(x_i-y_i)^2 + t\cdot|x_i|}_{\varphi(x)}$

NOT DIFFERENTIABLE! PANIC! :(

(drop x_i notation, let $x\in\mathbb{R}$)

GOLDEN RULE (Fermat's Principle)

subdifferential

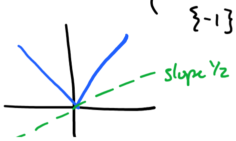
Find $0\in\partial\varphi(x) = \partial\left(\tfrac{1}{2}(x-y)^2 + t|x|\right)$

via CQ (full domain)

$= \partial\left(\tfrac{1}{2}(x-y)^2\right) + t\,\partial|x|$

$= x-y + t\cdot\partial|x|$

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



so how to find x?

① Try $x>0$

then $\partial|x| = 1$, so $0 = x-y+t$

i.e., $x = y - t$.

This is valid if $\underline{\underline{y - t > 0}}$. $\quad (y > t)$

② Try $x < 0$

then $\partial|x| = -1$, ... i.e., $x = y + t$

so valid if $y + t < 0$ $\quad (y < -t)$

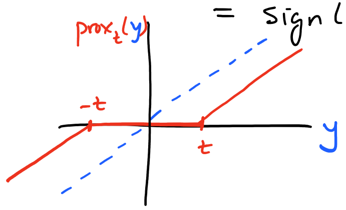
③ Try $x = 0$ then $\partial|x| = [-1, 1]$

i.e., $y - \underbrace{x}_{=0} \in [-t, t]$ $\quad (-t \le y \le t)$

... so we actually covered all cases!

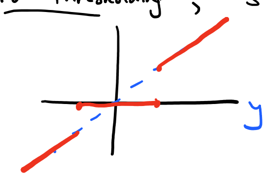
$$\text{prox}_{t|x|}(y) = \begin{cases} y - t & \text{if } y > t \\ 0 & \text{if } -t \le y \le t \\ y + t & \text{if } y < -t \end{cases}$$

$$= \text{sign}(y) \cdot \lfloor |y| - t \rfloor_+ \quad \text{where} \quad \lfloor \alpha \rfloor_+ = \max(\alpha, 0)$$



"soft-thresholding" or "shrinkage"

(vs. "hard-thresholding", $\text{sign}(y) \cdot \max(|y|, t)$



which is not the prox of a convex function)

and so $\text{prox}_{t\|\cdot\|_1}(y)$ is just component-wise soft-thresholding.

Rules If prox_f and prox_g are known, there is not a general formula for prox_{f+g}.

Even $\text{prox}_{f \circ L}$ isn't easy (in terms of prox_f) unless L is orthogonal or nearly so

Supplementary material: Moreau Envelope

Def The inf-convolution (infimal-convolution) of f and g is

$$(f \,\square\, g)(x) = \inf_y f(y) + g(x - y)$$
$$= \inf_y f(x - y) + g(y)$$

Fact f, g proper $\Rightarrow (f \,\square\, g)^* = f^* + g^*$

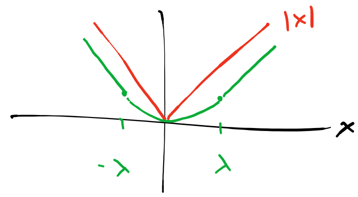
Fact $f, g \in \Gamma_0(\mathbb{R}^n) \Rightarrow f \,\square\, g = (f^* + g^*)^*$

Choosing $g(x) = \frac{1}{2\lambda}\|x\|^2$ gives the Moreau Envelope / Moreau-Yosida Regularization

$$M_\lambda(f) := f \,\square\, \tfrac{1}{2\lambda}\|\cdot\|^2$$

which preserves local min but smooths them (but doesn't smooth maxes)

Ex Huber function $H_\lambda(x) := M_\lambda(|x|)$



$$= \begin{cases} \frac{1}{2\lambda} x^2 & |x| \le \lambda \\ |x| - \lambda/2 & |x| > \lambda \end{cases}$$

Used in statistics: like $\ell_2^{\,2}$ near origin but ℓ_1 far away is more robust to outliers.

(Also used to make $\|x\|_1$ differentiable if we're lazy)

Proximal Point Algorithm PPA

Claim: $\min_x M_\lambda(f(x)) = \min_x f(x)$, i.e., preserves min

$$\min_x \min_y \underbrace{f(y) + \tfrac{1}{2\lambda}\|x-y\|^2}_{\ge f(y)} \ge \min_x \min_y f(y) = \min_y f(y) = \min_x f(x)$$

and $\le f(y) + \frac{1}{2\lambda}\|x-y\|^2$
for any y, i.e., $y = x^* \in \operatorname{argmin} f(x)$

so $\le \min_x f(x^*) + \frac{1}{2\lambda}\|x - x^*\| = f(x^*) = \min_x f(x)$

Connection to prox: the y that minimizes $f(y) + \frac{1}{2\lambda}\|x-y\|^2$ is $y = \operatorname{prox}_{\lambda f}(x)$

PPA Algorithm

Outer loop

Initialize x_0
For $k = 1, 2, \ldots$
$\quad x_k = \operatorname{prox}_{\lambda f}(x_{k-1})$] via inner loop

i.e., solve $\min f(x)$

by solving a sequence of $\min_x \underbrace{f(x) + \tfrac{1}{2\lambda}\|x - x_{k-1}\|^2}$

Why would you do this? $\longrightarrow$ Strongly convex!

Interpretation:

use Fact $\nabla M_\lambda f(x) = \lambda^{-1}\left(x - \text{prox}_{\lambda f}(x)\right)$ (if f convex)

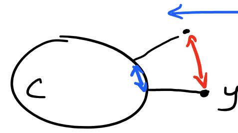
and is Lipschitz cts. w/ constant $1/\lambda$

so PPA is gradient descent (w/ stepsize $\lambda = 1/L$) on $M_\lambda f$

$$\begin{aligned} x_k &= x_{k-1} - \lambda \cdot \nabla M_\lambda f(x_{k-1}) \\ &= x_{k-1} - \left(x_{k-1} - \text{prox}_{\lambda f}(x_{k-1})\right) \\ &= \text{prox}_{\lambda f}(x_{k-1}) \end{aligned}$$

More Facts

Orthogonal Projections are non-expansive, $\|\text{Proj}_C(x) - \text{Proj}_C(y)\| \le \|x-y\|$
(if C is convex)



but recall Fixed Point Iteration / Picard / Banach (APPM 4650, 5440)

we like contractions : $\|Tx - Ty\| \le L \cdot \|x-y\|$, $L < 1$

$\Rightarrow$ $x_{k+1} = T x_k$ converges.

Non-expansive means $L = 1$. NOT GOOD FOR CONVERGENCE

Ex: $Tx := -x$, $x_{k+1} = -x_k$. Unless $x_0 = 0$, doesn't converge.

But it's non-expansive

Projections are not contractive (eg., if $x \in C$, $\text{Proj}_C(x) = x$ so it acts like identity operator)

It turns out they satisfy an in-between property : firm nonexpansivity (and will prevent $T = -I$ counterexample)

Def $T: \mathbb{R}^n \to \mathbb{R}^n$ is firmly non-expansive if $\forall x, y \in \mathbb{R}^n$

1. $\|Tx - Ty\|^2 \le \langle Tx - Ty, x-y \rangle$

or equivalently

identity operator, $Ix = x$, not indicator fcn

2. $\|Tx - Ty\|^2 + \|(I-T)x - (I-T)y\|^2 \le \|x-y\|^2$

or equiv.

3. T firmly non-expansive iff $(I-T)$ firmly non-expansive

or equiv.

4. T firmly nonexpansive iff $2T - I$ is non-expansive

$\boxed{Q2}$ Show "firmly non-expansive" (definition ①) $\Rightarrow$ "non-expansive"
$\cancel{\Leftarrow}$

$\boxed{A2}$ $\Rightarrow$ $\|Tx - Ty\|^2 \le \langle Tx - Ty, x - y\rangle$ via definition

$\le \|Tx - Ty\| \cdot \|x - y\|$ via Cauchy–Schwarz

$\div \|Tx - Ty\|$ to get $\|Tx - Ty\| \le \|x - y\|$, i.e., non-expansive

$\Leftarrow$ $T = -I$ is non-expansive (since $\|-x + y\| \le \|x - y\|$)

but not firmly non-expansive:

$\|-x + y\|^2 \overset{?}{\le} \langle -x + y, x - y\rangle = -\|x - y\|^2$

$\|x - y\|^2 \overset{?}{\le} -\|x - y\|^2$ No, False

Thm Proj_C is firmly non-expansive ($C \ne \emptyset$, closed, convex)

Thm If $f \in \Gamma_0(\mathbb{R}^n)$ then prox_f is firmly non-expansive

Take-away For a convergent algo. like gradient descent (w/ appropriate stepsize), we can add prox/projection steps and they won't mess up convergence (details later)

Moreau's Decomposition

Recall, if $V \subseteq \mathcal{H}$ is a linear subspace (either a closed subspace or $\mathcal{H} = \mathbb{R}^n$) and $V^\perp$ its orthogonal complement, then $\forall x \in \mathcal{H}$, $x = x^{\parallel} + x^{\perp}$

$x^{\parallel} = \mathrm{Proj}_V(x)$, $x^\perp = \mathrm{Proj}_{V^\perp}(x)$

This extends to prox!

Thm $f \in \Gamma_0(\mathbb{R}^n)$ then $\forall x \in \mathbb{R}^n$,

① $$\boxed{x = \underbrace{\mathrm{prox}_f(x)}_{x_1} + \underbrace{\mathrm{prox}_{f^*}(x)}_{x_2}}$$

and furthermore

② $f(x_1) + f^*(x_2) = \langle x_1, x_2\rangle$ (tight Young's ineq.)

and

(3) $\frac{1}{2}\|x\|^2 = (M_1 f)(x) + (M_2 f^*)(x)$

Moreau envelope

Take-away Using (1), if I can compute $\text{prox}_f(x)$ numerically, then I can also compute $\text{prox}_{f^*}(x) = x - \text{prox}_f(x)$