1

Ref. Parikh, Boyd, "Proximal algorithms" Foundations and trends in optimization, 2013.

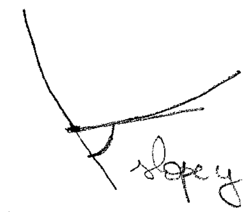
04/10/18 Proximal iterative methods (nonsmooth problems)

ex. $f(x) = \|Ax-y\|_2^2 + \lambda \|x\|_1$

ex. $f(x) = \|Ax-y\|_1$

Def. $\underbrace{\partial f(x)}_{\text{subdifferential}} = \{ y : f(z) \geq f(x) + y^T(z-x) \ \forall z \}$

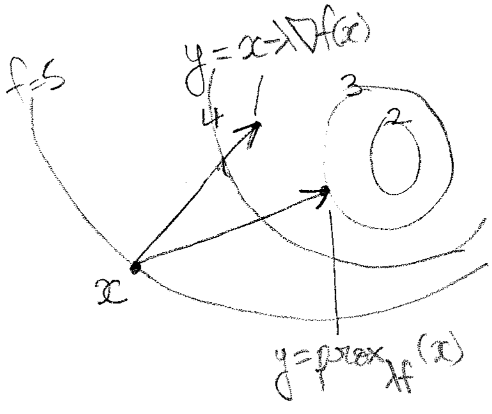
y: subgradient / subderivative



if $f \in C^1$, $\partial f(x) = \{\nabla f(x)\}$.
but in general it is a set.

Prop. f convex, C^0:
$x^* = \arg\min_x f(x) \iff 0 \in \partial f(x^*)$

Pf. $f(x) \geq f(x^*)$
$= f(x^*) + 0^T(x-x^*)$ $\square$



Def. proximal operator

$$\mathrm{prox}_{\lambda f}(x) = \arg\min_y \left(f(y) + \frac{1}{2\lambda} \|x-y\|_2^2 \right)$$

if $f \in C^1$, $\nabla f(y) + \frac{1}{\lambda}(y-x) = 0$.

$$y = x - \lambda \nabla f(y). \quad (= \mathrm{prox}_{\lambda f}(x))$$
$$y + \lambda \nabla f(y) = x$$

$\Rightarrow$ y is the point from which, if you look <u>backwards</u> along $-\lambda\nabla f(y)$, you reach x.

Gradient step is $\frac{y-x}{\lambda} = -\nabla f(x)$
$\rightarrow$ forward step (like Forward Euler)

Proximal step is $\frac{y-x}{\lambda} = -\nabla f(y)$
$\rightarrow$ backward step. (like Backward Euler)

if $f \in C^0$, $0 \in \partial f(y) + \frac{1}{\lambda}(y-x)$

$$\Rightarrow \quad x \in \underbrace{y + \lambda \partial f(y)}_{\text{operator } (I+\lambda\partial f)(y),\ \text{point-to-set}} \qquad (\$)$$

$$\Rightarrow \quad y = (I + \lambda \partial f)^{-1}(x) \qquad \text{(definition)}$$

resolvent of $I + \lambda\partial f$
(single-valued by def. of prox. for convex fn.).

<u>ex</u>. $f(x) = \mathbb{1}_{x\in C}(x)$, C convex set

$$= \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C. \end{cases}$$

then $\underset{y}{\mathrm{argmin}}\ f(y) + \frac{1}{2\lambda}\|y-x\|_2^2$

$= \mathrm{Proj}_C(x)$, orthogonal projection more generally, $\text{prox}_{\lambda f}(x)$ is an orthogonal projection on a <u>level set</u> of $f(x)$

<u>Prop.</u> f convex, C^0. For any $\lambda > 0$,
$x^* = \text{argmin}\, f(x) \Leftrightarrow x^* = \text{prox}_{\lambda f}(x^*)$
(fixed point)

<u>Pf.</u> WLOG $\lambda = 1$.

($\Rightarrow$) $f(x) \geq f(x^*) \quad \forall x$

$$\Rightarrow f(x) + \frac{1}{2}\|x - x^*\|_2^2 \geq f(x^*) + \frac{1}{2}\|x^* - x^*\|_2^2$$

$$\Rightarrow x^* = \underset{x}{\text{argmin}}\, f(x) + \frac{1}{2}\|x - x^*\|_2^2$$

$$\Rightarrow x^* = \text{prox}_f(x^*)$$

($\Leftarrow$) $x^* = \text{argmin}\, f(x) + \frac{1}{2}\|x - x^*\|_2^2$

$$\partial\left(f(x) + \frac{1}{2}\|x - x^*\|_2^2\right) = \partial f(x) + x - x^*$$

$$\Rightarrow 0 \in \partial f(x^*) + x^* - x^*$$
$$\Rightarrow 0 \in \partial f(x^*)$$
$$\Rightarrow x^* \text{ minimizes } f(x) \qquad \square$$

<u>Def.</u> Proximal minimization / Proximal point algorithm

$$x_{k+1} = \text{prox}_{\lambda f}(x_k).$$

Rmk: prox is not in general a contraction, but it is nonexpansive

$(\|\text{prox}(x) - \text{prox}(y)\| \leq \|x-y\|)$,

and slightly more (firm nonexpansiveness), which guarantees convergence.

Splitting (the point of proximal operators)

$$\min \underbrace{f(x)}_{\substack{C^1 \\ \text{differentiable}}} + \underbrace{g(x)}_{\substack{C^0 \\ \text{"proximable"}}}$$

$$0 \in \nabla f(x) + \partial g(x)$$

$$0 \in (\lambda \nabla f(x) - x) + (x + \lambda \partial g(x))$$

$$(I - \lambda \nabla f)(x) \in (I + \lambda \partial g)(x)$$

$$x = (I + \lambda \partial g)^{-1} (I - \lambda f)(x)$$

Suggests:

Def (forward-backward splitting, proximal gradient method)

$$x_{k+1} = \text{prox}_{\lambda f}(x_k - \lambda \nabla f(x_k))$$

Thm. Convergence with rate $O(1/k)$ ($)
when $\lambda \in (0, 1/L]$ fixed,
where ∇f is L-Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$$

⚠ Backward-backward is in general unstable

Ref. Beck-Teboulle 2010
(for a line search)

ex. $g(x) = \|x\|_1$ $\quad$ $(f(x) = \|Ax - y\|_2^2)$

$$\min_y \underbrace{\|y\|_1 + \frac{1}{2\lambda}\|y - x\|_2^2}_{\sum_i |y_i| + \frac{1}{2\lambda}(y_i - x_i)^2}$$

(separable).

$$\rightarrow \min_y \underbrace{|y| + \frac{1}{2\lambda}(y-x)^2}_{h(y)} \qquad (*)$$

$$\partial h(y) = \begin{cases} -1 + \frac{1}{\lambda}(y-x) & y < 0 \\ [-1,1] + \frac{1}{\lambda}(0-x) & y = 0 \\ 1 + \frac{1}{\lambda}(y-x) & y > 0. \end{cases}$$

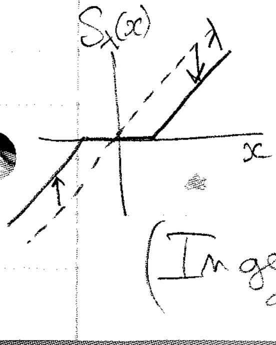
Case $y<0$: $h'(y) = 0 \iff y = x + \lambda$
works when $x < -\lambda$

Case $y=0$: $0 \in \partial h(0)$ $\quad$ $y = 0$
works when $|x| < \lambda$

Case $y>0$: $h'(y) = 0 \iff y = x - \lambda$
works when $x > \lambda$.

Concl: y solution of $(*)$ $(y = \mathrm{prox}_{\lambda g}(x))$ obeys

$$y = \begin{cases} x + \lambda & x < -\lambda \\ 0 & |x| < \lambda \\ x - \lambda & x > \lambda \end{cases}$$



$= S_\lambda(x)$, soft-thresholding at level λ.
or shrinkage by λ

(In general, apply componentwise)