proximal Algorithms. Neal Parikh 2013.

setting: $f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ closed proper convex. $\Rightarrow$ epi f nonempty closed convex.

proximal operator:

$$\boxed{\mathrm{prox}_f(v) = \underset{x}{\mathrm{argmin}}\left(f(x) + \frac{1}{2}\|x-v\|_2^2\right)}$$

$$\mathrm{prox}_{\lambda f}(v) = \underset{x}{\mathrm{argmin}}\left(f(x) + \frac{1}{2\lambda}\|x-v\|_2^2\right) = \underset{x}{\mathrm{argmin}}\left(\lambda f(x) + \frac{1}{2}\|x-v\|_2^2\right)$$

$\downarrow$ 在 v 附近最小化 f(x) 的点

v 的偏移：距离增加和 f(x) 增加的 tradeoff.

Gradient descent perspective

$$\mathrm{prox}_{\lambda f}(v) \approx v - \underset{\substack{\uparrow \\ \text{stepsize}}}{\lambda} \nabla f(v).$$

important property: link with fixed point theory: $\mathrm{prox}_{\lambda f}(x^*) = x^*$ iff x^* minimize $f(x)$.

properties & definitions of proximal operators.:

① f separable $f(x,y) = \varphi(x) + \psi(y) \Rightarrow \mathrm{prox}_f(v,w) = (\mathrm{prox}_\varphi(v), \mathrm{prox}_\psi(w))$

sum. joint.

$\downarrow$

$v, x \in \mathbb{R}^n$. $f(x) = \sum_{i=1}^n f_i(x_i) \Rightarrow (\mathrm{prox}_f(v))_i = \cancel{\mathrm{prox}_f}\ \mathrm{prox}_{f_i}(x_i)$

② $f(x) = \alpha\varphi(x) + b$. $\alpha > 0 \Rightarrow \mathrm{prox}_{\lambda f}(v) = \mathrm{prox}_{\alpha\lambda\varphi}(v)$

$f(x) = \varphi(\alpha x + b)$ $\alpha \neq 0 \Rightarrow \mathrm{prox}_{\lambda f}(v) = \frac{1}{\alpha}(\mathrm{prox}_{\alpha^2\lambda\varphi}(\alpha v + b) - b)$

③ orthogonal Q, $f(x) = \varphi(Qx) \Rightarrow \mathrm{prox}_{\lambda f}(v) = Q^T \mathrm{prox}_{\lambda\varphi}(Qv)$

④ $f(x) = \varphi(x) + a^T x + b$. $\Rightarrow \mathrm{prox}_{\lambda f}(v) = \mathrm{prox}_{\lambda\varphi}(v - \lambda a)$.

⑤ $f(x) = \varphi(x) + \frac{\rho}{2}\|x-a\|_2^2$. $\Rightarrow \mathrm{prox}_{\lambda f}(v) = \mathrm{prox}_{\tilde{\lambda}\varphi}\left(\left(\frac{\tilde{\lambda}}{\lambda}\right)v + (\rho\tilde{\lambda})a\right)$ $\tilde{\lambda} = \frac{\lambda}{1+\lambda\rho}$

Fixed point algorithms:

defs: ① contraction: f Lipschitz continuous with $K<1$ $d(f(x), f(y)) \le K\, d(x,y)$. $K<1$

② non-expansive: f Lipschitz continuous with $K=1$ $d(f(x), f(y)) \le d(x,y)$

③ firmly nonexpansive: f ~~Lipschitz~~ $\mathrm{dis}(f(x), f(y))^2 \le (x-y)^T(f(x)-f(y))$ [vector space]

can prove to be Lipschitz-1

④ averaged operator: N is nonexpansive: $T = (1-\alpha)I + \alpha N$.

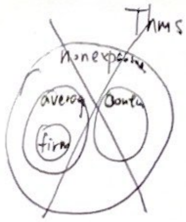
Thms: ① contraction can find fixed point $f(x^*) = x^*$ by $x^{(k+1)} = f(x^{(k)})$

② averaged operator can also find fixed point using ①

③ Firmly nonexpansive operators are indeed. $\frac{1}{2}$ averaged operator. $\forall\ T$ firm $\Rightarrow 2T - I$ nonexpan

④ averaged operators are closed under composition, but firmly nonexpansives are not.

⑤ proximal operators are firmly expansive operators



⑥ if N is only nonexpansive, update by $x^{(k+1)} = (1-\alpha)x^{(k)} + \alpha N(x^{(k)})$ [寻讨 average].

def: proximal average of $f_1 \cdots f_m$. closed proper convex g s.t. $\frac{1}{m}\sum_{i=1}^{m} \mathrm{prox}_{f_i} = \mathrm{prox}_g$.

Moreau decomposition: $v = \mathrm{prox}_f(v) + \mathrm{prox}_{f^*}(v)$ $\quad f^*(y)$: conjuage $= \sup_x (y^T x - f(x))$

e.g. $v = \Pi_L(v) + \Pi_{L^\perp}(v)$. $\quad f(x) = I_L(x)$, $(I_L(x))^* = I_{L^\perp}(x)$ $\quad \mathrm{prox}_{I_L}(v) = \Pi_L(v)$

e.g. $v = \Pi_K(v) + \Pi_{K^\circ}(v)$ $\quad K^\circ = \{y : y^T x \le 0, \forall x \in K\}$ polar cone, negtive ~~cone~~ of dual cone.

can be used as. $\mathrm{prox}_f(v) = v - \mathrm{prox}_{f^*}(v)$.

"smooth" approximation perspective of proximal operator.

infimal convolution: $f \,\square\, g(v) = \inf_x [f(x) + g(v-x)]$ $\quad v \in \mathrm{dom} f + \mathrm{dom}\, g$.

Moreau ~~ev~~envelope: $M_{\lambda f} = \frac{1}{\lambda}(\lambda f \,\square\, \frac{1}{2}\|\cdot\|_2^2) = \inf_x [f(x) + \frac{1}{2\lambda}\|v-x\|_2^2]$

(Moreau Tosida regularization).

$M_{\lambda f}(v)$ & $f(x)$ have the same minimizers.

Always continuously differentiable

Always have domain R^n

if define $h(x, v) = f(x) + \frac{1}{2\lambda}\|v-x\|_2^2$

$M_{\lambda f}(v) = h(\mathrm{prox}_{\lambda f}(v), v)$

Establish a approximation relationship between $M_{\lambda f}(x)$ and $f(x)$:

M_f is a smoothed version of f.

$\frac{1}{2}\|\cdot\|_2^2$ is self dual

property $(f \,\square\, g)^* = f^* \,\square\, g^* \Longrightarrow M_f^* = f^* + (\frac{1}{2}\|\cdot\|_2^2)^* \Longrightarrow M_f^* = f^* + \frac{1}{2}\|\cdot\|_2^2$

$\xRightarrow{M_f^{**} = M_f} M_f = (f^* + \frac{1}{2}\|\cdot\|_2^2)^*$ $\quad$ dual $\to$ regulorize $\to$ dual. (it's smooth)

(get strong convex)

$M_{\lambda f}$ differentiable, so we can take derivative.

~~$M_{\lambda f}(\mathrm{prox}_{\lambda f}(x))$~~ $M_{\lambda f}$ ~~$(x) = h(\mathrm{prox}_{\lambda f}(x), x)$~~ $=$

~~$M_{\lambda f}(v) = h(\mathrm{prox}_{\lambda f}(v), v) = f(\mathrm{prox}_{\lambda f}(v)) + \frac{1}{2\lambda}\|v - \mathrm{prox}_{\lambda f}(v)\|_2^2$~~

$\Rightarrow \mathrm{prox}_{\lambda f}(x) = x - \lambda \nabla M_{\lambda f}(x)$. $\quad$ approximate $\nabla f(x)$. [so it's a gradient descent step]

$\mathrm{prox}_f(x) = \nabla M_{f^*}(x)$.

proximal operator is the resolvent of subdifferential operator.

$\mathrm{prox}_{\lambda f} = (I + \lambda \partial f)^{-1}$ $\quad$ what's nontrivial here: $(I + \lambda \partial f)^{-1}$ becomes single-valued mapping

subgradient of f

More perspective from gradient descent:

① $\mathrm{prox}_{\lambda f}(x) = x - \lambda \nabla M_{\lambda f}(x)$.

② if $\nabla f(x)$ exists, first-order approximation $\hat{f}_v^{(1)}(x) = f(v) + \nabla f(v)^T (x - v)$.

then $\mathrm{prox}_{\hat{f}_v^{(1)}}(v) = v - \lambda \nabla f(v)$. $\quad$ explain: minimize first-order approximaton of $f(x)$ at point v (near v) result in a gradient descent step from v.

③ if $\nabla^2 f(x)$ exists, second-order approximation $\hat{f}_v^{(2)}(x) = f(v) + \nabla f(v)^T (x-v) + \frac{1}{2}(x-v)^T \nabla^2 f(v)(x-v)$

then $\mathrm{prox}_{\hat{f}_v^{(2)}}(v) = v - (\nabla^2 f(v) + \frac{1}{\lambda} I)^{-1} \nabla f(v)$. explain: similar, but 2-order, result in a Levenberg-Marquardt step.

Trust Region problem perspective.

trust region problem:

$$\min \ f(x) \quad \text{s.t.} \quad \|x-v\|_2 \le \rho.$$

proximal problem:

$$\min \ f(x) + \frac{1}{2\lambda}\|x-v\|_2^2$$

relationship: solution for some ρ $\xleftarrow{\text{is a}}$ solution

solution $\longrightarrow$ unconstrained minimizer of f / solution for some λ

Above is the interpretation of proximal operator.

Algorithms.

[1] Direct proximal ~~min~~ minimization

$x^{k+1} = \text{prox}_{\lambda f}(x^k)$. guarantee convergence with $\lambda_k > 0$, $\sum_{k=1}^{\infty} \lambda^k = \infty$.

application: ill-conditioned f [we add a quadratic term to be strong convex]

perspective: $x^{k+1} = \underset{x}{\text{argmin}}\ f(x) + \frac{1}{2\lambda_k}\|x - x^k\|_2^2$ regularization gets smaller as $x \to x^*$. the impact of the term disappears with iterations.

e.g. of application: iterative refinement. $f(x) = \frac{1}{2}x^T A x - b^T x$. ill-conditioned A.

$$\text{prox}_{\lambda f}(x^k) = (A + \tfrac{1}{\lambda}I)^{-1}(b + \tfrac{1}{\lambda}x^k)$$
$$= x^k + \underbrace{(A + \tfrac{1}{\lambda}I)^{-1}}_{\hat{A}}(b - Ax^k)$$

iteratively compensating for $\hat{A} - A$. difference.

[2] Proximal gradient Method.

$\min f(x) + g(x)$. $f(x)$ differentiable. both closed proper convex., $g(x)$ can be non-smooth.

$x^{k+1} = \text{prox}_{\lambda^k g}(x^k - \lambda^k \nabla f(x^k)) \iff x^k - \lambda^k \nabla f(x^k) - \lambda^k \nabla M_{\lambda^k g}(x^k). \approx x^k - \lambda^k \nabla (f+g)(x^k)$.

Convergence: ∇f Lipschitz $-L \Rightarrow \lambda^k = \lambda \in (0, \frac{1}{L}]$, Converge with $O(\frac{1}{k})$.

L not known: back-tracking "line" search. Beck & Teboulle.

parameter $\beta \in (0,1)$. $\lambda = \lambda^{k-1}$
$\Rightarrow z = \text{prox}_{\lambda g}(x^k - \lambda \nabla f(x^k))$
if $f(z) > \hat{f}_\lambda(z, x^k)$, $\lambda := \beta\lambda$
$\hat{f}_\lambda(x,y) = f(y) + \nabla f(y)^T(x-y) + \frac{1}{2\lambda}\|x-y\|_2^2$

perspective ① Majorization – minimization. (like EM algorithm)

Consider minimizing $\varphi(x)$.

Step 1: majorization: find convex upper (surrogate) bound $\hat{\varphi}$ of φ tight at x^k: $\begin{cases} \hat{\varphi}(x, x^k) \ge \varphi(x) \\ \hat{\varphi}(x,x) = \varphi(x) \end{cases}$

step 2: minimization: $x^{k+1} = \underset{x}{\text{argmin}}\ \hat{\varphi}(x, x^k)$.

For $f(x)$, a surrogate is $\hat{f}_\lambda(x,y)$, EM alg ~~with~~ will give precisely gradient descent.

For $f(x)+g(x)$. $q_\lambda(x,y) = \hat{f}_\lambda(x,y) + g(x) \iff$ proximal gradient.

perspective ② solution is a fixed point for $\mathrm{prox}_{\lambda g}((I-\lambda\nabla f)(x)) = (I+\lambda\partial g)^{-1}(I-\lambda f)$.

[3] Accelerated proximal gradient.

$$y^{k+1} = x^k + \omega^k (x^k - x^{k-1})$$

$$x^{k+1} = \mathrm{prox}_{\lambda^k g}(y^{k+1} - \lambda^k \nabla f(y^{k+1}))$$

recommend. $\omega^k = \frac{k}{k+3}$

convergence: $\lambda^k = \lambda = O(0, \frac{1}{L}]$, rate $O(\frac{1}{k^2})$.

If L not know. again Beck & Teboulle search but use $\cancel{y^{k+1}}$ y^k

[4] ADMM version. of proximal gradient.

$\min f(x) + g(x)$. $\iff$ $\min f(x) + g(z)$ s.t. $x = z$

both f, g can be non-smooth.

recall ADMM (scaled)

$$x^{k+1} := \underset{x}{\mathrm{argmin}}\ f(x) + \frac{\rho}{2}\|x - z^k + u^k\|_2^2 \quad \cancel{\mathrm{prox}_{\lambda f}(z^k - u^k)}$$

$$z^{k+1} := \underset{z}{\mathrm{argmin}}\ g(z) + \frac{\rho}{2}\|x^{k+1} - z + u^k\|_2^2 \quad \cancel{\mathrm{prox}_{\lambda g}(x^{k+1} + u^k)}$$

$$u^{k+1} := u^k + x^{k+1} - z^{k+1}$$

$\iff$

$$x^{k+1} = \mathrm{prox}_{\lambda f}(z^k - u^k)$$

$$z^{k+1} = \mathrm{prox}_{\lambda g}(x^{k+1} + u^k).$$

$$u^{k+1} = u^k + x^{k+1} - z^{k+1}.$$

$\lambda = \frac{1}{\rho}$.

an interesting insight: prove the convergence of ADMM $\iff$ fixedpoint algorithm of a firmly nonexpansive operator.

[5]. Linearized ADMM.

$\min f(x) + g(Ax)$ $\iff$ $\min f(x) + g(z)$ s.t. $Ax - z = 0$.

Original ADMM.

$$x^{k+1} = \underset{x}{\mathrm{argmin}}\ f(x) + \cancel{\frac{\rho}{2}\|Ax - z^k + u^k\|_2^2}\ \rho u^{k^T}(Ax - z^k) + \frac{\rho}{2}\|Ax - z^k\|_2^2$$

$$z^{k+1} = \underset{z}{\mathrm{argmin}}\ g(z) + \rho u^{k^T}(Ax^{k+1} - z) + \frac{\rho}{2}\|Ax^{k+1} - z\|_2^2$$

$$u^{k+1} = u^k + Ax^{k+1} - z^{k+1}.$$

$\longrightarrow$

only modify x step:

replace $\frac{\rho}{2}\|Ax - z^k\|_2^2 \propto \cancel{\frac{\rho}{2} x^T A^T A x}$ $\frac{\rho}{2}(A^TAx - A^Tz^k)^T x$ with.

$\rho(A^TAx^k - A^Tz^k)^T x + \frac{\mu}{2}\|x - x^k\|_2^2$. $0 < \mu \leq \frac{\lambda}{\|A\|_2^2}$

then new algorithm:

$$x^{k+1} = \mathrm{prox}_{\mu f}(x^k - \frac{\mu}{\lambda}A^T(Ax^k - z^k + u^k))$$

$$z^{k+1} = \mathrm{prox}_{\lambda g}(Ax^{k+1} + u^k).$$

$$u^{k+1} = u^k + Ax^{k+1} - z^{k+1}.$$