Proximal gradient algorithm

7:19 PM

$\forall f_0(x)\{\square\}$ ← computing the projection is easy on this set.
$\neg\ x\in X\ \{\text{simple}\}$

Proximal algorithm Calafiore

Proximal gradient is a first order method

Mixed differentiable and non-differentiable objective কে handle করতে পারে.

* Proximal Mapping and projections:

$\{h\{\underbrace{[\,]}_{\text{the sublevel sets are closed}},\square,:\mathbb{R}^n\to\mathbb{R}\}\}$ $\quad \text{prox}_h(x)\{\text{proximal mapping of } h\} = \underset{z}{\text{argmin}}\left(\underbrace{h(z)}_{\{\square\}}+\underbrace{\tfrac{1}{2}\|z-x\|_2^2}_{\{\square_{\text{strongly}}\}}\right)$

$\{\square_{\text{strongly}}\}$ // $\square+\square_{\text{strongly}}=\square_{\text{strongly}}$

$\{\because f(x)\{\square_{\text{strongly}}\} \overset{\text{iff}}{\Leftrightarrow} (f(x)+\frac{m}{2}\|x\|_2^2)\{\square\}\}$ ($\exists\, m>0$)

Existence theorem of subgradient (যদি থাকে): $(f\{\square\}, x\in\text{int dom} f) \to \partial f(x)\{\neg\emptyset,\leqslant\}$

so: $\partial h(x)\{\neg\emptyset,\leqslant\} \overset{\text{iff}}{=} \forall_{z\in\text{dom}\,h}\ h(z)\geqslant h(x)+\eta_x^T(z-x)$

$x\in\text{int dom}\,h$, $\{\square\}$; η_x: subgradient of h at x

$\Leftrightarrow h(z)+\frac{1}{2}\|z-x\|_2^2 \geqslant h(x)+\eta_x^T(z-x)+\frac{1}{2}\|z-x\|_2^2$ = finite

finite; finite; $\{\leqslant\}$

$h(x)$: finite as $x\in\text{int dom} f$ // any convex function // evaluated in int dom is // finite

$\left\{\begin{array}{l}\text{unbounded below (finite lower bound)}\\ +\\ \text{strong convexity}\Rightarrow\text{strict convexity}\end{array}\right\} \Rightarrow \text{prox}_h(x)$ is $\boxed{\text{well defined}} \Leftrightarrow \boxed{\text{exists, unique}}$

• An important special case:

$h(z)=I_X(z)\ \{\text{indicator function of convex set } X\} = \begin{cases}0, & \text{if } z\in X\\ +\infty, & \text{else}\end{cases}$

$\text{prox}_{I_X}(x)=\underset{z\in X}{\text{argmin}}\left(I_X(z)+\frac{1}{2}\|z-x\|_2^2\right)$

$=\underset{z\in X}{\text{argmin}}\left(\left(\begin{cases}0, & \text{if } z\in X\\ \infty & \text{else}\end{cases}\right)+\frac{1}{2}\|z-x\|_2^2\right)=\underset{x\in X}{\text{argmin}}\left(\|z-x\|_2^2\right)$ \{ কারণ both $\frac{1}{2}\|z-x\|_2^2$ and $\|z-x\|_2^2$ are minimized by same argmin \} $=[x]_X$ \{Euclidean projection of x onto X\} $=\Pi_X(x)$

// as $z\in X$ is satisfied by the outer // argmin operator, only $\underline{0}$ will be choosen

* Any constrained minimization problem can be written in the following form using set indicator:

$\begin{pmatrix}\forall f(x)\\ \neg\ x\in X\end{pmatrix}=\left(\forall_x\ f_0(x)+\underbrace{I_X(x)}_{\text{a nondifferentiable barrier function}}\right)$

however, $I_X(x)$ is convex:

[Indicator function is convex]

$\forall_x\forall_y\forall_{\lambda\in[0,1]}$ $I_X(\lambda x+(1-\lambda)y)=\begin{cases}0, & \lambda x+(1-\lambda)y\in X\\ \infty, & \text{else}\end{cases}$

$\lambda I_X(x)=\begin{cases}0, & x\in X\\ \infty, & \text{else}\end{cases}$

$(1-\lambda)I_X(y)\ \begin{cases}0, & y\in X\\ \infty, & \text{else}\end{cases}$

Now dom f has to be convex (even for consideration)

$x\in\text{dom}\,I_X$, $y\in\text{dom}\,I_X \Rightarrow \lambda(x)+(1-\lambda)y\in\text{dom}\,I_X$

so $I_X(\lambda(x)+(1-\lambda)y)\leqslant\lambda I_X(x)+(1-\lambda)I_X(y)$

* Consider even more general problem:

$\forall_x\ \underbrace{f_0(x)}_{\{\square,\,\delta_2^x\}}+\underbrace{h(x)}_{\{\text{simple},\,\square\}}$ $\{I_X(x)\in\{h(x):\{\text{simple},\square\}\}\}$

proximal mapping is easy to compute

• Proximal gradient algorithm:

$(\forall \underbrace{f(x)}_{f_0(x)+h(x)})$: Algorithm: $\forall_{k\in\{0,1,\dots\}}\ x_{k+1}=\text{prox}_{s_k h}(x_k-s_k\nabla f_0(x_k))$

$f_0(x)$: $\{\square,\delta_2^x\}$; $h(x)$: $\square$ \{may not be differentiable\}

would-have-been update law if the problem were $\forall f_0(x)$; $s_k\in\mathbb{R}_{++}$

$=\underset{z}{\text{argmin}}\left(s_k h(z)+\frac{1}{2}\|z-(x_k-s_k\nabla f_0(x_k))\|_2^2\right)$ \{ $\text{prox}_{\square}(\square)=\underset{\square}{\text{argmin}}\left(\square(\square)+\frac{1}{2}\|\square-\square\|_2^2\right)$ \}

\{ s_k constant w.r.t z তাই কে ভাগ করে দিলেও resultant function এর $\forall$ same থাকবে \}

$=\underset{z}{\text{argmin}}\left(h(z)+\frac{1}{2s_k}\|(z-x_k)+s_k\nabla f_0(x_k)\|_2^2\right)$

\{ $\|(z-s_k)\|_2^2+s_k^2\|\nabla f_0(x_k)\|_2^2+2s_k\nabla f_0(x_k)^T(z-s_k)$ \}

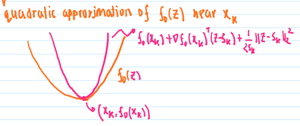
$=\underset{z}{\text{argmin}}\left(h(z)+\frac{1}{2s_k}\|z-x_k\|_2^2+\frac{1}{2}s_k\|\nabla f_0(x_k)\|_2^2+\nabla f_0(x_k)^T(z-x_k)\right)$

বাদ দেয়া যায় (constant তাই), $\forall$ same থাকবে

$=\underset{z}{\text{argmin}}\left(h(z)+\frac{1}{2s_k}\|z-x_k\|_2^2+\nabla f_0(x_k)^T(z-x_k)+f_0(x_k)\right)$

$f_0(x_k)$: constant w.r.t z, so add করলে $\forall$ same থাকবে

$=\underset{z}{\text{argmin}}\left(h(z)+\underbrace{f_0(x_k)+\nabla f_0(x_k)^T(z-x_k)+\frac{1}{2s_k}\|z-x_k\|_2^2}_{\text{local quadratic approximation of } f_0(z) \text{ near } x_k}\right)$



\{ $q_k(z)=f_0(x_k)+\nabla f_0(x_k)^T(z-x_k)+\frac{1}{2s_k}\|z-x_k\|_2^2$, $\psi_k(z)=h(z)+q_k(z)$ \}

$=\underset{z}{\text{argmin}}\left(h(z)+q_k(z)\right)=\underset{z}{\text{argmin}}\left(\psi_k(z)\right)$

$\boxed{\therefore \forall_k\ x_{k+1}=\underset{z}{\text{argmin}}\left(\psi_k(z)\right)=\underset{z}{\text{argmin}}\left(h(z)+q_k(z)\right)}$

$\nabla_z\left(f_0(x_k)+\nabla f_0(x_k)^T(z-x_k)+\frac{1}{2s_k}\|z-x_k\|_2^2\right)=\nabla f_0(x_k)+\frac{1}{s_k}(z-s_k)$ $\quad$ [$(z-s_k)^T I(z-s_k)$]

$\forall_k\ 0\in\left[\partial h(z)+\nabla q_k(z)\right]_{z=x_{k+1}}=\left[\partial h(z)+\nabla f_0(x_k)+\frac{1}{s_k}(z-x_k)\right]_{z=x_{k+1}}=\partial h(x_{k+1})+\nabla f_0(x_k)+\frac{1}{s_k}(x_{k+1}-x_k)$

$\Leftrightarrow \forall_k\ \exists_{\eta_{k+1}\in\partial h(x_{k+1})}\ 0=\eta_{k+1}+\nabla f_0(x_k)+\frac{1}{s_k}(x_{k+1}-x_k)$

Again:

$x_{k+1}=\text{prox}_{s_k h}(x_k-s_k\nabla f_0(x_k))$ থেকে motivated হয়ে আমরা following difference vector define করি:

$x_k-x_{k+1}=x_k-\underbrace{\text{prox}_{s_k h}(x_k-s_k\nabla f_0(x_k))}_{\{x_k\}}=s_k\, g_{s_k}(x_k)$ \{ defined as \}

$\Leftrightarrow \boxed{x_{k+1}=x_k-s_k g_k}\ \boxed{\{\ g_k(x_k)=\frac{1}{s_k}\left(x_k-\text{prox}_{s_k h}(x_k-s_k\nabla f_0(x_k))\right)\ \}}$ \{gradient map of f_0 on h at x_k\}

g_k is a pseudo-gradient \{ $g_s(x^*)=0$, $h=0$ \{no constraint\} $\to g_k=\nabla f_0(x_k)$ \{becomes same as gradient

$\to \forall_k\ \exists_{\eta_{k+1}\in\partial h(x_{k+1})}\ 0=\eta_{k+1}+\nabla f_0(x_k)-g_k$

$\Leftrightarrow \boxed{\forall_k\ \exists_{\eta_{k+1}\in\partial h(x_{k+1})}\ \nabla f_0(x_k)=g_k-\eta_{k+1}}$ (eq: 345)

$\boxed{\forall_{z\in\text{dom}\,h}\ h(z)\geqslant h(x_{k+1})+\eta_{k+1}^T(z-x_{k+1})}$ (eq: 346)

$\leftrightarrow \boxed{x_{k+1} = x_k - s_k g_k} \boxed{\{ g_k(x_k) = \frac{1}{s_k}\left(x_k - \text{prox}_{s_k h}(x_k - s_k \nabla f_0(x_k))\right) \}}$

g_k is a pseudo-gradient $\{ g_s(x^*) = 0$, $h=0$ {no constraint} $\rightarrow g_k = \nabla f_0(x_k)$ {becomes same as gradient in the case of unconstrained problem}

//At optimal point

//the pseudo gradient is 0

$\forall_{z \in \text{dom}\, h} \quad h(z) \geq h(x_{k+1}) + \eta_{k+1}^T (z - x_{k+1})$ (eq: 346)

$h = I_{\mathcal{X}}$ 이면 ($\{[\,],\square\}$) x_{k+1} is the euclidian projection of would_have_been_update_rule_for_unconstrained_optimization_problem onto $\mathcal{X}$

would_have_been_update_rule $= x_k - s_k \nabla f_0(x_k)$

*Proximal Gradient Algorithm: (constant stepsize) (alg: proximal_gradient_algorithm)

Require: • $f_0\{\square, \delta_2, \geq■, \exists_{L>0} \forall_{x \in \text{dom} f_0} \|\nabla^2 f_0(x)\|_F^2 = \sum_{i=1}^{n} \sigma_i^2 \leq L^2\}$

$\underbrace{\exists_{L>0} \forall_{x \in \text{dom} f_0} \|\nabla^2 f_0(x)\|_F^2 = \sum_{i=1}^{n} \sigma_i^2 \leq L^2}_{\{\text{Lipschitz continuous gradient}\} \Leftarrow}$

$\exists_{L>0} \forall_{x,y \in \text{dom} f_0} \|\nabla f_0(x) - \nabla f_0(y)\|_2 \leq L \|x-y\|_2$

• $h\{\square, [\,]\}$

• $x_0 \in \text{dom} f_0$

• $\epsilon > 0$

Alg:

$\boxed{1}$ $k=0$, $s = 1/L$

$\boxed{2}$ $\forall_{k \in \{0,1,\dots\}} \quad x_{k+1} = \text{prox}_{sh}(x_k - s \nabla f_0(x_k))$

$\boxed{3}$ if $f(x_{k+1}) - f(x^*) \leq \epsilon \Rightarrow$ exit done :)

else $k := k+1$ and go to 2

• Convergence of the proximal gradient algorithm:

$f_0\{\square_{\text{strongly}}, \exists_{L} \forall_{x \in \text{dom} f} \|\nabla^2 f_0(x)\|_F \leq L\}$

$\updownarrow \exists_{m>0}$

$(f_0(x) - \frac{m}{2}\|x\|_2^2)\{\square\}$

Note that:

(eq: derivative of $q_k(z)$)

$\nabla q_k(z) = \nabla f_0(x_k) + \frac{1}{s_k}(z - x_k) \quad \{ q_k(z) = f_0(x_k) + \nabla f_0(x_k)^T(z - x_k) + \frac{1}{2 s_k}\underbrace{\|z - x_k\|_2^2}_{(z-x_k)^T I (z-x_k)}$

$\rightarrow \nabla_z q_k(z) = 0 + \nabla f_0(x_k) + \frac{1}{2 s_k} \times I \cdot (z - s_k) = \nabla f_0(x_k) + \frac{1}{s_k}(z - s_k)\}$

$\rightarrow \nabla q_k(x_{k+1}) = \nabla f_0(x_k) - g_k \quad \{ \nabla q_k(x_{k+1}) = \nabla f_0(x_k) + \frac{1}{s_k}(x_{k+1} - x_k)$ # By defn, $x_{k+1} = x_k - s_k g_k \leftrightarrow -g_k = \frac{1}{s_k}(x_{k+1} - x_k)$

$= \nabla f_0(x_k) - g_k\}$

* $\boxed{x_{\bar{k}} = \gamma \leftrightarrow g_{\bar{k}} = 0}$ // $\bar{k}$ is a fixed // index for this proof

Proof: $(\rightarrow)$

Optimality condition for $\gamma\, f_0(x_{\bar{k}}) + h(x_{\bar{k}})$ is $0 \in \partial(f_0(x_{\bar{k}}) + h(x_{\bar{k}})) = \underbrace{\partial f_0(x_{\bar{k}})}_{\nabla f_0(x_{\bar{k}})} + \partial h(x_{\bar{k}}) = \nabla f_0(x_{\bar{k}}) + \partial h(x_{\bar{k}})\}$

$[\because f_0(x_{\bar{k}})\{\delta_2\}]$

As, $x_{\bar{k}}$ is optimal, $x_{\bar{k}+i} = x_{\bar{k}}$ ($i \geq 1$) 이고 이다 for consistency

따라서 the update rule is:

$x_{\bar{k}+1} = \underset{z}{\text{argmin}}\, \psi_{\bar{k}}(z)$

$\psi_{\bar{k}}(z) = h(z) + q_{\bar{k}}(z)$

$q_{\bar{k}}(z) = f_0(x_{\bar{k}}) + \nabla f_0(x_{\bar{k}})^T (z - x_{\bar{k}}) + \frac{1}{2 s_{\bar{k}}} \|z - x_{\bar{k}}\|_2^2$

Optimality on $\psi_{\bar{k}}(z) \Leftarrow 0 \in \partial \psi_{\bar{k}}(z) = \partial h(z) + \partial q_{\bar{k}}(z) = \partial h(z) + \nabla q_{\bar{k}}(z) = \overbrace{\partial h(z) + \nabla f_0(x_{\bar{k}})}^{0 \in} + \frac{1}{s_{\bar{k}}}(z - x_{\bar{k}})$ eq: Expression for \partial \phi_k{z}

$\partial q_{\bar{k}}(z) = \underbrace{\nabla q_{\bar{k}}(z)}_{\{\delta_2\}}$

$\nabla q_{\bar{k}}(z) = \{ \nabla f_0(x_{\bar{k}}) + \frac{1}{s_{\bar{k}}}(z - x_{\bar{k}})$ from (eq: derivative of $q_k(z)$) $\}$

$\leftrightarrow 0 \in \underbrace{\partial h(z) + \nabla f_0(x_{\bar{k}})}_{0 \in} + \frac{1}{s_{\bar{k}}}(z - x_{\bar{k}})$

$\rightarrow z = x_{\bar{k}}$ {otherwise $0 \notin$ RHS 이다}

$\left.\begin{array}{l} x_{\bar{k}+1} = \underset{z}{\text{argmin}}\, \psi_{\bar{k}}(z) = x_{\bar{k}} \\ x_{\bar{k}+1} = x_{\bar{k}} - \underset{>0}{s_{\bar{k}}}\, g_{\bar{k}} \end{array}\right\} \Rightarrow g_{\bar{k}} = 0$

$(\leftarrow)$ $g_{\bar{k}} = 0$

now

$\forall_i \quad x_{i+1} = x_i - s_i g_i$

$i = \bar{k}: \quad x_{\bar{k}+1} = x_{\bar{k}} - s_{\bar{k}} g_{\bar{k}} = x_{\bar{k}}$

By defn:

$\underbrace{x_{\bar{k}+1} = \underset{z}{\text{argmin}}\, \psi_{\bar{k}}(z) = x_{\bar{k}}}$ // $\because z^* = x_{\bar{k}}$ 이고 $x_{\bar{k}}$ is not variable, it is known // so, $x_{\bar{k}}$ is what minimizes $\psi_{\bar{k}}(z)$

By definition:

$\forall_k \quad x_{k+1} = \underset{z}{\text{argmin}}\, \overbrace{\psi_k(z)}^{h(z) + q_k(z)}$

$\Rightarrow 0 \in [\partial \psi_{\bar{k}}(z)]_{z = x_{\bar{k}}}$ {By optimality}

$= [\partial h(z) + \nabla f_0(x_{\bar{k}}) + \frac{1}{s_{\bar{k}}}(z - x_{\bar{k}})]_{z = x_{\bar{k}}} = \partial h(x_{\bar{k}}) + \nabla f_0(x_{\bar{k}})$

(eq: Expression for \partial \phi_k{z} (참조))

$\Rightarrow 0 \in \partial h(x_{\bar{k}}) + \nabla f_0(x_{\bar{k}})$

$\therefore x_{\bar{k}} = \underset{x}{\text{argmin}}\, [f_0(x) + h(x)]$ ■

$\therefore x_{\bar{k}} = \underset{x}{\operatorname{argmin}}\left[f_0(x)+h(x)\right]$

$\Rightarrow \exists\, \eta_{\bar{k}} \in \partial h(x_{\bar{k}})$ $\quad g_{\bar{k}} = 0 = \eta_{\bar{k}} + \nabla f_0(x_{\bar{k}})$

Need to remove

$\partial h(x_{\bar{k}}) = \partial h(x_{\bar{k}+1})$ হলে আমরা $\eta_{\bar{k}}$ কে $\eta_{\bar{k}+1}$ দিয়ে replace করতে পারি.

By def

$\forall_{z\in \operatorname{dom} h} \quad h(z) \geq h(x_{\bar{k}}) + \eta_{\bar{k}}^T (z - x_{\bar{k}})$

By assumption $f_0 \{\mathcal{D}_{\text{strongly}}, \mathcal{L}_L^{\nabla}$ {Lipschitz continuous gradient}$\}$

$\leftrightarrow \exists_{L,m:\, 0<L\leq m}\ \forall_{z\in \operatorname{dom} f_0}\quad f_0(x_k) + \nabla f_0(x_k)^T(z-x_k) + \frac{m}{2}\|z-x_k\|_2^2 \leq f_0(z) \leq f_0(x_k) + \nabla f_0(x^k)^T(z-x_k) + \frac{L}{2}\|z-x_k\|_2^2$

{Lemma 12.1.2.

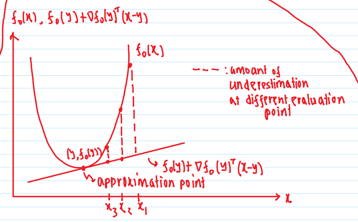
amount of underestimation

[lemma: Underestimation Lemma]

$f_0 \{\mathcal{D}, \delta_2^{\text{continuous}}, \mathcal{L}_L^{\nabla}\} \Rightarrow \forall_x \forall_y \left(0 \leq \underbrace{f_0(x) - f_0(y) - \nabla f_0(y)^T(x-y)}_{\text{amount of underestimation}} \leq \frac{L}{2}\|x-y\|_2^2\right)$

// figure

For a convex function which is continuously differentiable and lipschitz continuous gradient, the amount of underestimation between two function evaluations (the original function and the 1st order estimation around an **approximation point**) at some **evaluation point** is upper bounded by L/2 times the norm squared of the difference between the evaluation point and approximation point



আবার

$f_0(x)\{\mathcal{D}_{\text{strongly}}, \delta_2\} \overset{\text{def}}{\equiv} \exists_m \left(f_0(x) - \frac{m}{2}\|x\|_2^2\right)\{\mathcal{D}, \delta_2\}$

$\overset{\text{def}}{\equiv} \forall_{x,y\in \operatorname{dom} f}\quad f_0(y) - \frac{m}{2}\|y\|_2^2 \geq f_0(x) - \frac{m}{2}\|x\|_2^2 + \underbrace{\nabla\left(f_0(x) - \frac{m}{2}\|x\|_2^2\right)^T}_{\left(\nabla f_0(x) - \frac{m}{2}2x\right)^T}(y-x)$

$\leftrightarrow f_0(y) \geq f_0(x) + \nabla f_0(x)^T(y-x) \quad \nabla f_0(x)^T(y-x) - m x^T y + m\|x\|_2^2$

$+ \underbrace{\frac{m}{2}\|x\|_2^2 + \frac{m}{2}\|y\|_2^2 - 2\cdot\frac{m}{2}x^Ty}_{= \frac{m}{2}\|y-x\|_2^2}$

$\leftrightarrow f_0(y) - f_0(x) - \nabla f_0(x)^T(y-x) \geq \frac{m}{2}\|y-x\|_2^2$

$\frac{m}{2}\|x-y\|_2^2 \leq f_0(x) - f_0(y) - \nabla f_0(x)^T(y-x) \leq \frac{L}{2}\|y-x\|_2^2$

$\leftrightarrow f_0(y) + \nabla f_0(x)^T(y-x) + \frac{m}{2}\|x-y\|_2^2 \leq f_0(x) \leq f_0(y) + \nabla f_0(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2$

• আলাদা আলাদাভাবে লিখলে:

$\forall_k\ \forall_{z\in \operatorname{dom} f_0}\quad f_0(z) \geq f_0(x_k) + \nabla f_0(x_k)^T(z-x_k) + \frac{m}{2}\|z-x_k\|_2^2$

$\forall_k\ \forall_{z\in \operatorname{dom} f_0}\quad f_0(z) \leq f_0(x_k) + \nabla f_0(x_k)^T(z-x_k) + \frac{L}{2}\|z-x_k\|_2^2$

$z := x_{k+1}: \quad f_0(\underbrace{x_{k+1}}_{=x_k = \forall}) \leq f_0(x_k) + \nabla f_0(x_k)^T$

$z = x_{k+1},\ 0 < s_k \leq \frac{1}{L}:$

$f_0(x_{k+1}) \leq f_0(x_k) + \nabla f_0(x_k)^T(x_{k+1}-x_k) + \frac{L}{2}\|x_{k+1}-x_k\|_2^2$

$\{g_k(z) = f_0(x^k) + \nabla f_0(x_k)^T(z-x_k) + \frac{1}{2s_k}\|z-x_k\|_2^2,\quad 0 < s_k \leq \frac{1}{L} \to L \leq \frac{1}{s_k}\}$ $\quad \leq f_0(x_k) + \nabla f_0(x_k)^T(x_{k+1}-x_k) +$

$\frac{1}{2s_k}\|x_{k+1}-x_k\|_2^2 = g_k(x_{k+1})$

$\therefore f_0(x_{k+1}) \leq g_k(x_{k+1})$ (eq: 432)

{adding $h(z)$ on bothsides}

$\leftrightarrow \forall_k\ \underbrace{\forall_{z\in \operatorname{dom} f}}\ f(z) - \frac{m}{2}\|z-x_k\|_2^2 \geq h(z) + f_0(x_k) + \nabla f_0(x_k)^T\underbrace{(z-x_k)}_{(x_{k+1}-x_k+z-x_{k+1})}$

{extended value extension ✓}

$\underbrace{\nabla f_0(x_k)^T(x_{k+1}-x_k) + \nabla f_0(x_k)^T(z-x_{k+1})}$

$= h(z) + f_0(x_k) + \nabla f_0(x_k)^T(x_{k+1}-x_k) + \nabla f_0(x_k)^T(z-x_{k+1})$

{from: (eq: 345) থেকে: $\forall_k \exists_{\eta_{k+1}\in\partial h(x_{k+1})}\ \nabla f_0(x_k) = g_k - \eta_{k+1}$}

$= h(z) + f_0(x_k) + \nabla f_0(x_k)^T(x_{k+1}-x_k) + (g_k - \eta_{k+1})^T(z-x_{k+1})$

$= h(z) - \eta_{k+1}^T(z-x_{k+1}) + f_0(x_k) + \nabla f_0(x_k)^T(x_{k+1}-x_k) + g_k^T(z-x_{k+1})$

{from (eq: 346): $\forall_{z\in\operatorname{dom} h}\ h(z) \geq h(x_{k+1}) + \eta_{k+1}^T(z-x_{k+1})$, using extended-value extension $z\in\operatorname{dom} f_0 \leftrightarrow z\in\operatorname{dom} h$}

$\geq h(x_{k+1}) + f_0(x_k) + \nabla f_0(x_k)^T(x_{k+1}-x_k) + g_k^T(z-x_{k+1})$

{$g_k(z) = f_0(x_k) + \nabla f_0(x_k)^T(z-x_k) + \frac{1}{2s_k}\|z-x_k\|_2^2$

$z = x_{k+1}: \quad g_k(x_{k+1}) = f_0(x_k) + \nabla f_0(x_k)^T(x_{k+1}-x_k) + \frac{1}{2s_k}\|x_{k+1}-x_k\|^2$}

$= h(x_{k+1}) + g_k(x_{k+1}) - \frac{1}{2s_k}\|x_{k+1}-x_k\|_2^2 + g_k^T(z-x_{k+1})$

{By defn:

$g_k = \frac{1}{s_k}(x_k - x_{k+1}) = -\frac{1}{s_k}(x_{k+1}-x_k)$

$\to \|g_k\|_2^2 = g_k^T g_k = \frac{1}{s_k^2}\|x_{k+1}-x_k\|_2^2$

$\leftrightarrow \|x_{k+1}-x_k\|_2^2 = s_k^2\|g_k\|_2^2$}

$= h(x_{k+1}) + g_k(x_{k+1}) - \frac{1}{2s_k}s_k^2\|g_k\|_2^2 + g_k^T(z-x_{k+1})$

$= h(x_{k+1}) + g_k(x_{k+1}) - \frac{s_k}{2}\|g_k\|_2^2 + g_k^T(z-x_k) + g_k^T(x_k-x_{k+1})$

{$g_k^T(x_k-x_{k+1}) = \frac{1}{s_k}\|x_k-x_{k+1}\|_2^2 = \frac{1}{s_k}\cdot s_k^2\cdot\|g_k\|_2^2 = s_k\|g_k\|_2^2$}

$= h(x_{k+1}) + g_k(x_{k+1}) - \frac{s_k}{2}\|g_k\|_2^2 + g_k^T(z-x_k) + s_k\|g_k\|_2^2$

$$= h(x_{k+1}) + g_k(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 + g_k^T(z-x_k)$$

$$\geq \underbrace{h(x_{k+1}) + f_0(x_{k+1})}_{f(x_{k+1})\ \{\text{by defn}\}} + \frac{s_k}{2}\|g_k\|_2^2 + g_k^T(z-x_k) \quad \{\text{From (eq: 432)}\}$$

$$= f(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 + g_k^T(z-x_k)$$

$$\rightarrow \forall_k\ \forall_{z\in \text{dom} f} \quad f(z) - \frac{m}{2}\|z-x_k\|_2^2 \geq f(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 + g_k^T(z-x_k) \quad (12.55)$$

$$\xrightarrow{z=x_k} \forall_k \quad f(x_k) - \underbrace{\frac{m}{2}\|x_k-x_k\|_2^2}_{0} \geq f(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 + \underbrace{g_k^T(x_k-x_k)}_{0}$$

$$\therefore \forall_k \quad f(x_{k+1}) - f(x_k) \leq -\frac{s_k}{2}\|g_k\|_2^2 \quad (\text{eq: decreasing\_of\_f})$$

// note that this is 0 only at optimum, else $f(x_{k+1}), f(x_k)$ এর মাঝে ব্যবধান থাকবে, অতএব ফাংশন কমতে থাকবে.

so proximal gradient method is a descent method!

• From (12.55):

$$\forall_k\ \forall_{z\in \text{dom} f} \quad f(z) \geq f(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 + g_k^T(z-x_k) + \frac{m}{2}\|z-x_k\|_2^2 \quad (\text{eq: 12.57})$$

$\{(\forall_z\ f(z) \geq g(z)) \rightarrow \min_z f(z) \geq \min_z g(z) \equiv f(z_f) \geq g(z_g)$, $\gamma = z_f$, $\gamma = z_g$

Proof:

Per absurdum $(\cdot) \wedge f(z_f) < g(z_g)$

$z := z_f$ $\Downarrow$ $f(z_f) \geq g(z_f)$.

$\Rightarrow g(z_f) \leq f(z_f) < g(z_g)$

$\Rightarrow g(z_f) < g(z_g)$

$\therefore z_g \neq \gamma$ (contradiction! ∎ $\}$

$$\forall_k \quad \underbrace{\min_z f(z)}_{f(x^*)\ \{x^* = \gamma f(z)\}} \geq \min_z \left(f(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 + g_k^T(z-x_k) + \frac{m}{2}\|z-x_k\|_2^2 \right)$$

$\{\nabla$

$$g_k + \frac{m}{2}2(z-x_k) = 0 \equiv m(z-x_k) = -g_k \equiv z = x_k - \frac{g_k}{m}$$

$$f(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 \underbrace{- g_k^T\frac{g_k}{m} + \frac{m}{2}\left\|-\frac{g_k}{m}\right\|_2^2}$$

$$-\frac{1}{m}\|g_k\|_2^2 + \frac{m}{2}\cdot\frac{1}{m^2}\|g_k\|_2^2 = -\frac{1}{2}\frac{1}{m}\|g_k\|_2^2$$

$$= f(x_{k+1}) + \|g_k\|_2^2 \cdot \frac{1}{2}\cdot\left(s_k - \frac{1}{m}\right)$$

$$\rightarrow \forall_k \quad f(x^*) \geq f(x_{k+1}) + \frac{1}{2}\|g_k\|_2^2\left(s_k - \frac{1}{m}\right)$$

$$\leftrightarrow \forall_k \quad f(x_{k+1}) - f(x^*) \leq -\frac{1}{2}\|g_k\|_2^2\left(s_k - \frac{1}{m}\right) = \frac{1}{2}\|g_k\|_2^2\underbrace{\left(\frac{1}{m} - s_k\right)}$$

$\{ \geq 0$ as $L \geq m$, $s_k \leq \frac{1}{L}$ $\rightarrow$ $s_k \leq \frac{1}{L} \leq \frac{1}{m} \Rightarrow \left(\frac{1}{m} - s_k\right) \geq 0 \}$

$$\therefore \forall_k \quad f(x_{k+1}) - f(x^*) \leq \frac{1}{2}\|g_k\|_2^2\left(\frac{1}{m} - s_k\right) \quad (\text{eq: 12.58}) \quad \frac{1}{m} \geq \frac{1}{L}$$

(eq: 12.57) থেকে:

$$\forall_k\ \forall_{z\in \text{dom} f} \quad f(z) \geq f(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 + g_k^T(z-x_k) + \frac{m}{2}\|z-x_k\|_2^2$$

$\therefore z = x^*$:

$$\forall_k \quad f(x^*) \geq f(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 + g_k^T(x^*-x_k) + \underbrace{\frac{m}{2}\|x^*-x_k\|_2^2}$$

$\{\mathbb{R}_+$ termটাকে ফেলে বাদ দেয়া হবে$\}$

$$\geq f(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 + g_k^T(x^*-x_k)$$

আবার $\rightarrow$

$$\forall_k \quad \Big(0 \leq \overbrace{f(x_{k+1}) - f(x^*)}^{\text{by defn}} \leq -\frac{s_k}{2}\|g_k\|_2^2 - g_k^T(x^*-x_k) - \frac{m}{2}\|x^*-x_k\|_2^2\Big)$$

$$\Rightarrow \forall_k \quad 0 \leq g_k^T(x_k-x^*) - \frac{s_k}{2}\|g_k\|_2^2 - \frac{m}{2}\|x_k-x^*\|_2^2$$

$$\rightarrow \forall_k \quad g_k^T(x_k-x^*) \geq \frac{s_k}{2}\|g_k\|_2^2 + \frac{m}{2}\|x_k-x^*\|_2^2 \quad (\text{eq: 12.56})$$

$$\rightarrow \forall_k \quad f(x_{k+1}) - f(x^*) \leq g_k^T(x_k-x^*) - \frac{s_k}{2}\|g_k\|_2^2 = \frac{1}{2s_k}\underbrace{\left(\|x_k-x^*\|_2^2 - \|x_k-x^*-s_kg_k\|_2^2\right)} = \frac{1}{2s_k}\Big(\|x_k-x^*\|_2^2 - \|(\overbrace{x_k-s_kg_k}^{x_{k+1}})-x^*\|\Big) = \frac{1}{2s_k}\left(\|x_k-x^*\|_2^2 - \|x_{k+1}-x^*\|_2^2\right)$$

$\{ \|x_k-x^*\|_2^2 - \left(\|x_k-x^*\|_2^2 + s_k^2\|g_k\|_2^2 - 2s_kg_k^T(x_k-x^*)\right)$

$= \|x_k-x^*\|_2^2 - \|x_k-x^*\|_2^2 - s_k^2\|g_k\|_2^2 + 2s_kg_k^T(x_k-x^*)$

$= -s_k^2\|g_k\|_2^2 + 2s_kg_k^T(x_k-x^*)$

$= 2s_k\left(-\frac{s_k}{2}\|g_k\|_2^2 + g_k^T(x_k-x^*)\right)$

$\therefore \left(-\frac{s_k}{2}\|g_k\|_2^2 + g_k^T(x_k-x^*)\right) = \frac{1}{2s_k}\left(\|x_k-x^*\|_2^2 - \|x_k-x^*-s_kg_k\|_2^2\right) \}$

$$\therefore \forall_k \quad f(x_{k+1}) - f(x^*) \leq \frac{1}{2s_k}\left(\|x_k-x^*\|_2^2 - \|x_{k+1}-x^*\|_2^2\right) \quad (\text{eq: 12.59})$$

Now we are going to show convergence of proximal gradient algorithm for $s_k = s = \frac{1}{L}$

(এটা stepsize এর জন্য backtracking line search দিয়ে convergence এ arrive করা যাবে)

First note that:

$$\|x_{k+1}-x^*\|_2^2 = \|\overbrace{x_k - s_kg_k}^{x_{k+1}} - x^*\|_2^2 = \|(x_k-x^*) - s_kg_k\|_2^2 = \|x_k-x^*\|_2^2 + s_k^2\|g_k\|_2^2 - 2s_kg_k^T(x_k-x^*)$$

$$\leq \|x_k-x^*\|_2^2 - ms_k\|x_k-x^*\|_2^2$$

$$= (1-ms_k)\|x_k-x^*\|_2^2$$

$$= \underbrace{\left(1-\frac{m}{L}\right)}\|x_k-x^*\|_2^2 \quad [\because s_k = s = L]$$

$\in (0,1]$ $\{\because 0 < m \leq L$, because of strongly convex and lipschitz continuous gradient$\}$

$\{$From (eq: 12.56) $g_k^T(x_k-x^*) \geq \frac{s_k}{2}\|g_k\|_2^2 + \frac{m}{2}\|x_k-x^*\|_2^2$

$\rightarrow -2\underbrace{s_k}_{>0} g_k^T(x_k-x^*) \leq -s_k^2\|g_k\|_2^2 - ms_k\|x_k-x^*\|_2^2$

$\rightarrow s_k^2\|g_k\|_2^2 - 2s_kg_k^T(x_k-x^*) \leq -ms_k\|x_k-x^*\|_2^2 \}$

$$\therefore \forall_k \quad \|x_{k+1}-x^*\|_2^2 \leq \overbrace{\left(1-\frac{m}{L}\right)}^{\in(0,1]}\|x_k-x^*\|_2^2$$

$$\text{so} \quad \|x_k-x^*\|_2^2 \leq \left(1-\frac{m}{L}\right)\|x_{k-1}-x^*\|_2^2$$

$\in [0,1]$

$\therefore \forall_k \quad \|x_{k+1}-x^*\|_2^2 \leqslant \underbrace{\left(1-\frac{m}{L}\right)}_{\in[0,1]} \|x_k - x^*\|_2^2$ $\quad \{\because 0<m\leqslant L$, because of strongly convex and lipschitz continuous gradient$\}$

so $\quad \|x_k - x^*\|_2^2 \leqslant \left(1-\frac{m}{L}\right)\|x_{k-1}-x^*\|_2^2$

$\vdots$

$\|x_1 - x^*\|_2^2 \leqslant \left(1-\frac{m}{L}\right)\|x_{\underset{k-(k)}{0}} - x^*\|_2^2$

$$\|x_{k+1}-x^*\|_2^2 \leqslant \left(1-\frac{m}{L}\right)\|x_k - x^*\|_2^2 \leqslant \left(1-\frac{m}{L}\right)^2\|x_{k-1}-x^*\|_2^2 \ldots \leqslant \left(1-\frac{m}{L}\right)^{k+1}\|x_{\underset{k-(k)}{0}}-x^*\|_2^2$$

$$\therefore \quad 0 \leqslant \|x_{k+1}-x^*\|_2^2 \leqslant \left(1-\frac{m}{L}\right)^{k+1}\|x_0 - x^*\|_2^2$$

$$\lim_{k\to\infty} \quad 0 \leqslant \|x_{k+1}-x^*\| \leqslant \underbrace{\left(1-\frac{m}{L}\right)^{k+1}}_{\lim_{k\to\infty} \to 0}\|x_0-x^*\|_2^2$$

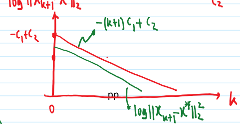
$$\therefore \lim_{k\to\infty} \|x_{k+1}-x^*\|_2^2 = 0 \Leftrightarrow \lim_{k\to\infty} x_{k+1} = x^*$$

So, the proximal gradient method convergences at a linear rate

On a log scale:

$$2\log\|x_{k+1}-x^*\|_2 \leqslant (k+1)\underbrace{\log\left(1-\frac{m}{L}\right)}_{-2c_1 \ [\because \log[(0,1)]<0]} + 2\log\|x_0 - x^*\|_2$$

$$\rightarrow \log\|x_{k+1}-x^*\|_2 \leqslant -(k+1)c_1 + \underbrace{\log\|x_0-x^*\|_2}_{c_2}$$



that is why the convergence rate is called linear #

* m, L জানা থাকলে (eq: 12.58) থেকে আমরা stopping criterion বের করতে পারি:

By checking norm of g_k:

$$\boxed{\|g_k\|_2^2 \leqslant 2\epsilon \frac{mL}{L-m}} \Rightarrow f(x_{k+1}) - f(x^*) \leqslant \frac{1}{2}\|g_k\|_2^2\left(\frac{1}{m} - \underset{1/L}{s_k}\right)$$

$\epsilon \geqslant 0$

Stopping criterion

$$= \frac{1}{2}\cdot 2\epsilon\cdot\frac{mL}{L-m}\cdot\frac{L-m}{mL} = \epsilon$$

$$\therefore f(x_{k+1}) - f(x^*) \leqslant \epsilon \quad \text{(eq: 12.60)}$$

from (eq: 12.59):

$$\forall_k \quad f(x_{k+1}) - f(x^*) \leqslant \frac{1}{2s_k}\left(\|x_k - x^*\|_2^2 - \|x_{k+1}-x^*\|_2^2\right)$$

$$\underline{k} \quad \forall_i \quad \left(f(x_i) - f(x^*)\right) \leqslant \frac{1}{2s_i}\left(\|x_{i-1}-x^*\|_2^2 - \|x_i - x^*\|_2^2\right)$$

$$\rightarrow \sum_{i=1}^{k}\left(f(x_i) - f(x^*)\right) \leqslant \sum_{i=1}^{k}\frac{1}{2\underset{1/L}{s_i}}\left(\|x_{i-1}-x^*\|_2^2 - \|x_i - x^*\|_2^2\right) = \frac{L}{2}\left(\sum_{i=1}^{k}\left(\|x_{i-1}-x^*\|_2^2 - \|x_i - x^*\|_2^2\right)\right)$$

$$= \frac{L}{2}\Big(\|x_{k-1}-x^*\|_2^2 - \|x_k - x^*\|_2^2$$
$$+ \|x_{k-2}-x^*\|_2^2 - \|x_{k-1}-x^*\|_2^2$$
$$\vdots$$
$$+ \|x_1 - x^*\|_2^2 - \|x_2 - x^*\|_2^2$$
$$+ \|x_0 - x^*\|_2^2 - \|x_1 - x^*\|_2^2 \Big)$$

$$= \frac{L}{2}\left(\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2\right)$$

$$\rightarrow \sum_{i=1}^{k}\left(f(x_i) - f(x^*)\right) \leqslant \frac{L}{2}\|x_0 - x^*\|_2^2 \underbrace{- \frac{L}{2}\|x_k - x^*\|_2^2}_{\text{nonpositive term}} \leqslant \frac{L}{2}\|x_0 - x^*\|_2^2$$

nonpositive term টা বাদ দিলে আমরা বড় মান পাবো।

আবার, proximal gradient method is a descent method বলে:

(eq: decreasing of f) থেকে: $\quad f(x_{k+1}) - f(x_k) \leqslant -\frac{s_k}{2}\|g_k\|_2^2 \leqslant 0$

$\rightarrow f(x_{k+1}) \leqslant f(x_k)$

$\rightarrow f(x_{k+1}) \leqslant f(x_k) \leqslant \ldots \leqslant f(x_1)$

$\rightarrow f(x_{k+1}) - f(x^*) \leqslant f(x_k) - f(x^*) \leqslant \ldots \leqslant f(x_1) - f(x^*)$

$\rightarrow f(x_{k+1}) - f(x^*) \leqslant \frac{1}{k}\left(\sum_{i=1}^{k}\left(f(x_i) - f(x^*)\right)\right)$ $\quad \{\because (x \leqslant x_n \leqslant \ldots \leqslant x_1) \rightarrow x \leqslant x_n$

$x \leqslant x_{n-1}$

$\vdots$

$x \leqslant x_1$

$nx \leqslant (x_1 + \ldots + x_n)$

$\rightarrow x \leqslant \frac{1}{n}(x_1 + \ldots + x_n)\}$

$$\boxed{\forall_k \ 0 \leqslant f(x_{k+1}) - f(x^*) \leqslant \frac{L}{2k}\|x_0 - x^*\|_2^2}$$

$$\therefore \lim_{k\to\infty} f(x_{k+1}) = f(x^*) \text{ with a rate } \frac{1}{k}.$$

Everything we have done so far is summarized in the following result:

Theorem 12.1:

(alg: proximal gradient algorithm) satisfies:

$$\forall_k \quad 0 \leqslant f(x_k) - f(x^*) \leqslant \frac{L}{2k}\|x_0 - x^*\|_2^2$$

if f_0 is strongly convex then: $\{f = f_0 + h\}$

$$\forall_k \left(0 \leqslant \|x_k - x^*\|_2^2 \leqslant \left(1-\frac{m}{L}\right)^k\|x_0 - x^*\|_2^2\right)$$

$$\forall_k \left(0 \leqslant f(x_{k+1}) - f(x^*) \leqslant \frac{1}{2}\|g_k\|_2^2\left(\frac{1}{m} - \frac{1}{L}\right)\right)$$