Chapter 8. Gradient Descent for Constrained & non-smooth optimization.

1. Constrained Optimization

$$\min_x f(x) \quad \text{st } x \in C.$$

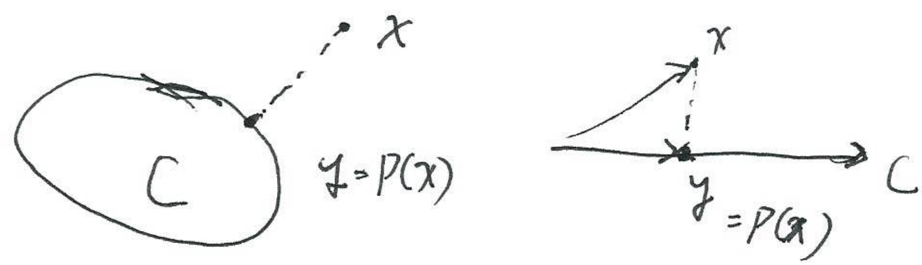
C: Convex, set

Recall gradient descent: $x^{k+1} \leftarrow x^k - \eta^k \cdot \nabla f(x^k)$

$\uparrow$ may not $\in C$

Def: Projection of x onto C

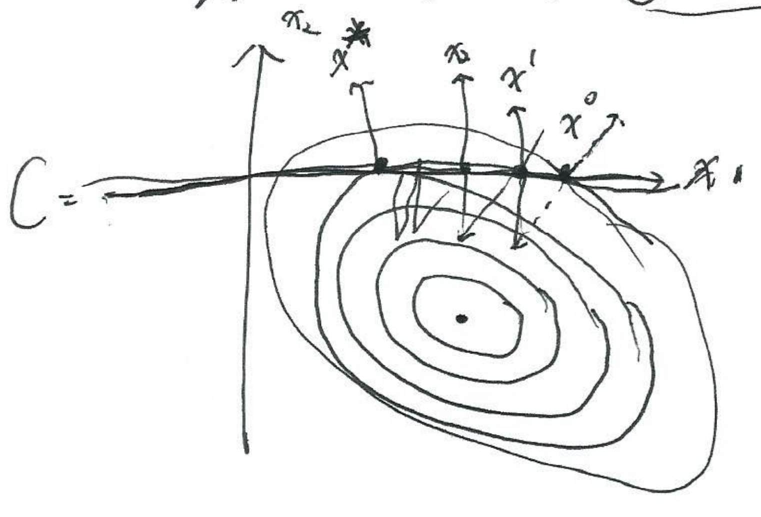
$$P_C(x) = \underset{y \in C}{\operatorname{argmin}} \|x - y\|_2$$



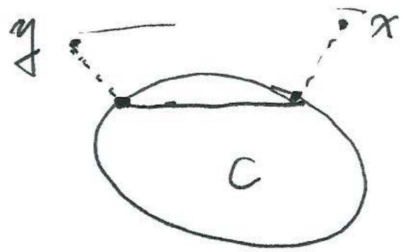
Projected Gradient Descent.

$$x^{k+1} = P_C(\underbrace{x^k - \eta^k \cdot \nabla f(x^k)}))$$

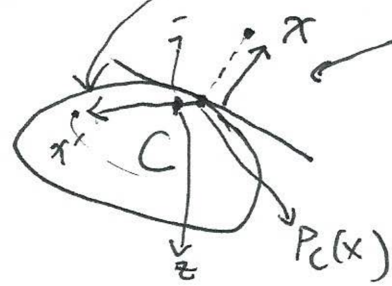
$C: \{x \mid x_2 = 0\}$



Lemma. $\| P_C(x) - P_C(y) \| \le \|x - y\| \quad \forall x, y$



pf: Step 1: $\forall x' \in C$, $\langle x' - P_C(x),\ x - P_C(x) \rangle \le 0$



(If this not true,
$\exists x'$, st $\langle \qquad \rangle > 0$

$z \neq P_C(x)$, $\|z - x\|^2 = \|z - P_C(x) + P_C(x) - x\|^2$

$= \|z - P_C(x)\|^2 + 2\langle z - P_C(x),\ P_C(x) - x \rangle$

$+ \|P_C(x) - x\|^2$

take $z = P_C(x) + \varepsilon \cdot (x' - x)$

$\|z - x\|^2 = \text{const} \cdot \varepsilon^2 - \text{const} \cdot \varepsilon$

$+ \|P_C(x) - x\|^2$

$\varepsilon \to 0$, $\|z - x\|^2 < \|P_C(x) - x\|^2$

contradict)

② $\langle P_C(y) - P_C(x),\ x - P_C(x) \rangle \le 0$

$+\ \langle P_C(x) - P_C(y),\ y - P_C(y) \rangle \le 0$

$\therefore\ \langle P_C(y) - P_C(x),\ x - P_C(x) - y + P_C(y) \rangle \le 0$

$\|P_C(x)-P_C(y)\|^2 = \langle P_C(x)-P_C(y),\ P_C(x)-P_C(y) \rangle \le \langle P_C(y) - P_C(x),\ y - x \rangle$

$\Downarrow$

$\|P_C(x)-P_C(y)\|^2 \le \langle P_C(y) - P_C(x),\ y - x \rangle \le \|P_C(y)-P_C(x)\|\,\|y-x\|$

$\Rightarrow \|P_C(x) - P_C(y)\| \le \|y - x\|$ #

Convergee: If f is convex, bunded below, continuosly differentiable, Then $\{x^k\}$ generated by PGD with line search (backtracking line search)

$$\lim_{k \to \infty} x^k = x^*$$

Example:

$$\min_x \frac{1}{2} x^T Q x + b^T x \quad \text{st} \quad x \geq 0$$

PGD: For $k = 0, 1, \dots$

$$g = Qx^k + b$$

$$\bar{x} = x - \eta^k g$$

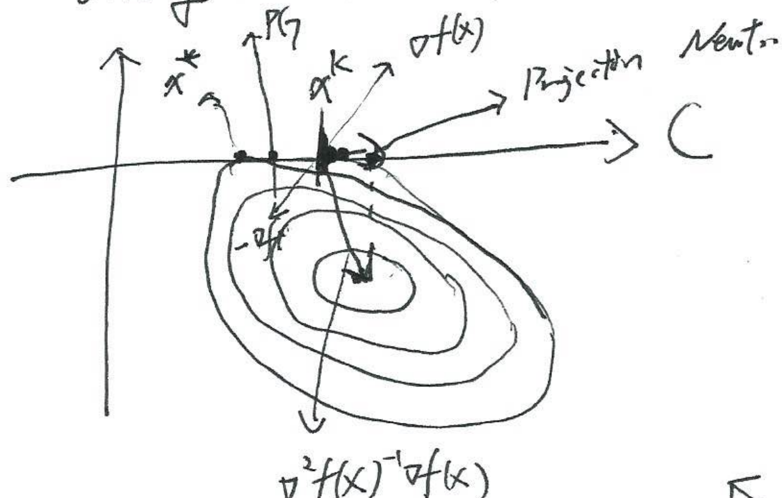
$$\boxed{x_i^{k+1} = \begin{cases} \bar{x}_i & \text{if } \bar{x}_i \geq 0 \\ 0 & \text{if } \bar{x}_i < 0 \end{cases}} = P_C(\bar{x})$$

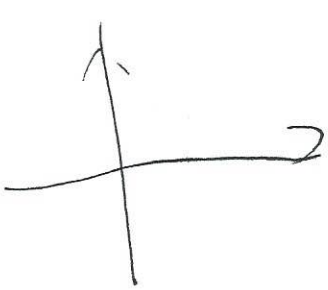
end.

Can we do projected Newton?

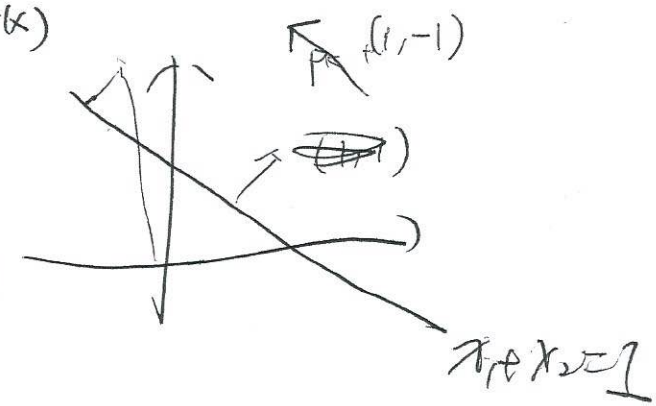
$$x^{k+1} = P_C\left(x^k - \eta^k \cdot \nabla^2 f(x^k)^{-1} \nabla f(x^k)\right) \quad X$$

Wrong....





$$p^T x \cdot \left(\frac{p}{\|p\|}\right)$$



Gradient Descent for Non-smooth Functions.

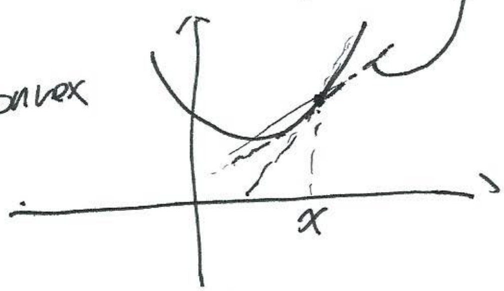
Non-(differentiable)

- Subgradient (for non-differentiable function)

recall : ~~the~~ $\nabla f(x)$ for a convex function

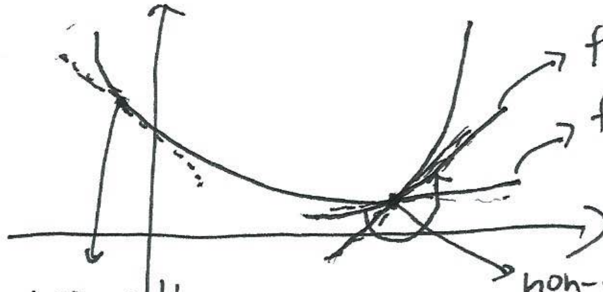
$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall y$$

convex



Def : $g \in R^n$ is a subgradient for $f(\cdot)$ on x

iff $\quad f(y) \geq f(x) + g^T (y - x) \quad \forall y$



$f(y) = f(x) + g_2^T (y - x)$

$f(y) = f(x) + g_1^T (y - x)$

non-differentiable

differentiable

$\Rightarrow$ subgradient = gradient

Example : Lasso : $f(x) = |x| \quad x \in R$

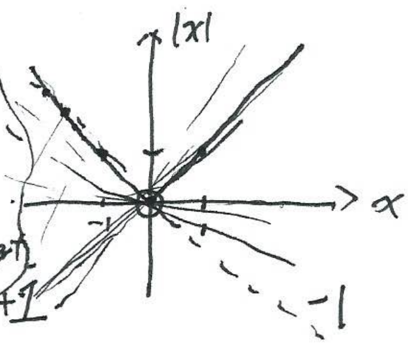
if $x \neq 0$

$\nabla f(x) = \text{sign}(x)$

if $x = 0$

$[-1, 1]$

can be subgradient



If $x > 0 \Rightarrow f(x) = x$

$\nabla f(x) = 1$

if $x < 0 \Rightarrow f(x) = -x$

$\nabla f(x) = -1$

if $x = 0$

$$f(y) \geq f(0) + g^T (y - 0)$$
$$= g^T y$$

$\forall g \in [-1, 1]$ satisfy

Def : Subdifferential

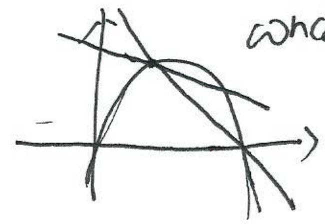
$$\partial f(x) = \{ g : g \text{ is a subgradient of } f \text{ at } x \}$$

① Existence :

$\partial f(x)$ is nonempty if f is convex

② If f is differentiable , then

$$\partial f(x) = \{ \nabla f(x) \}$$



③ If $\partial f(x) = \{\nabla f(x)\}$, then f is differentiable at x.

Example : If C is a convex set

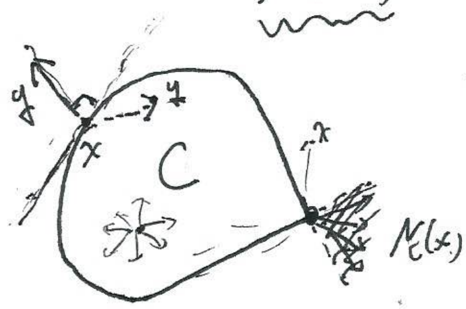
we can define $I_C(x)$ (indicator function)

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

$$\left(\begin{array}{l} \min\limits_{x} f(x) \text{ st } x \in C \\ \underset{\text{equivalent}}{\Leftrightarrow} \quad \min\limits_{x} f(x) + I_C(x) \end{array} \right)$$

Property : $\partial I_C(x) = N_C(x)$ (normal vectors), $x \in C$

$$N_C(x) = \{ g : g^T(y-x) \leq 0 , \forall y \in C \}$$



if $x \in C$

$I_C(x) = 0$

Why? By definition, we want.

$$I_C(y) \geq I_C(x) + g^T(y-x) \quad \forall y$$

If $y \notin C \Rightarrow I_C(y) = \infty$ OK

If $y \in C \Rightarrow I_C(y) = 0$

$\hookrightarrow 0 \geq 0 + g^T(y-x) \Rightarrow g^T(y-x) \leq 0$

want $\nearrow$

$\Leftrightarrow N_C(x)$

~~Rec~~ Optimality condition

recall if f is differentiable

$$x^* \text{ is optimal iff } \nabla f(x^*) = 0$$

For non-differentiable function (convex function)

$$x^* = \underset{x}{\operatorname{argmin}}\, f(x) \text{ iff } 0 \in \partial f(x^*)$$

Why? $f(y) \geq f(x^*) + \underline{0}^T (y - x^*) = f(x^*) \quad \forall y$

For constrained optimization:

$$\min_x f(x) \text{ st } x \in C, \quad (f, C \text{ are convex}) \quad f \text{ differentiable}$$

equivalent

$$\boxed{x^* \text{ is optimal solution iff } \nabla f(x^*)^T (y - x) \geq 0 \quad \forall y \in C}$$

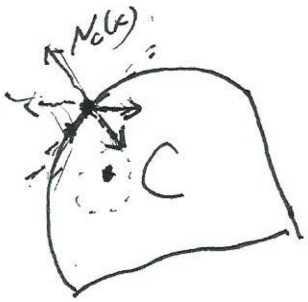
Why? $\min_x \{ f(x) + I_C(x) \} = F(x)$

x^* is optimal $\iff$ $0 \in \partial F(x^*)$

$$0 \in \partial F(x^*) = \{\nabla f(x^*)\} + \partial I_C(x^*)$$

$$= \{\nabla f(x^*)\} + N_C(x^*)$$

$$N_C(x) = \{ g : g^T (y - x) \leq 0, \ \forall y \in C \}$$



$$-\nabla f(x^*)^T (y - x) \leq 0 \quad \forall y \in C$$

$$\iff \nabla f(x^*)^T (y - x) \geq 0 \quad \forall y \in C$$

\#

Example:

$$\min_x \{ g(x) + \lambda \|x\|_1 \} := f(x)$$

Lasso $g(x) = \frac{1}{2}\|Ax - b\|_2^2$

$$\|x\|_1 := \sum_i |x_i|$$

$\{x_1\} + \{y_1, y_2, \cdots, y_n\} = \{x_1 + y_1, x_1 + y_2 \cdots\}$

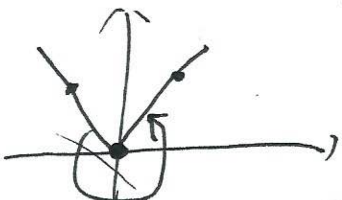
x^* Optimal solution $\Leftrightarrow$ $0 \in \partial f(x^*)$

$$= \{\nabla g(x^*)\} + \lambda \cdot \partial(\|x^*\|_1)$$

$$= \nabla g(x^*) + \lambda V$$

if $a \in \mathbb{R}$

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



$$\left(V_i = \begin{cases} 1 & \text{if } x_i^* > 0 \\ -1 & \text{if } x_i^* < 0 \\ [-1, 1] & \text{if } x_i^* = 0 \end{cases} \right)$$

$\Rightarrow x^*$ is optimal iff

$$\begin{cases} \nabla_i g(x^*) = -\lambda & \text{if } x_i^* > 0 \\ \nabla_i g(x^*) = \lambda & \text{if } x_i^* < 0 \\ \nabla_i g(x^*) \in [-\lambda, \lambda] & \text{if } x_i^* = 0 \end{cases}$$

Simple Example: (Soft-thresholding)

$$\min_x \frac{1}{2}\|x - b\|_2^2 + \lambda \|x\|_1$$

$(Ax-b)$

$\nabla g(x^*) = x^* - b$

Optimal solution x^* iff

$$\begin{cases} x_i^* - b_i = -\lambda & \text{if } x_i^* > 0 \\ x_i^* - b_i = \lambda & \text{if } x_i^* < 0 \\ x_i^* - b_i \in [-\lambda, \lambda] & \text{if } x_i^* = 0 \end{cases}$$

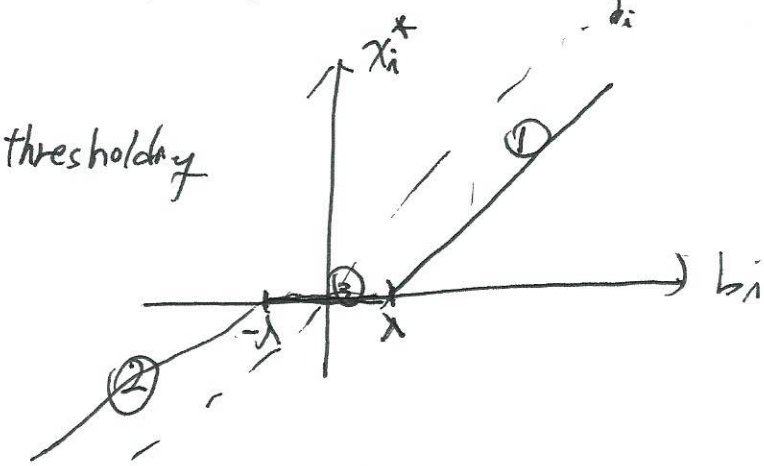
$$\Rightarrow \begin{cases} \text{if } b_i - \lambda > 0, & x_i^* = b_i - \lambda \\ \text{if } b_i + \lambda < 0, & x_i^* = b_i + \lambda \\ \text{if } |b_i| < \lambda, & x_i^* = 0 \end{cases}$$

$$\Rightarrow \left.\begin{cases} x_i^* = b_i - \lambda & \text{if } b_i > \lambda \quad ① \\ x_i^* = b_i + \lambda & \text{if } b_i < -\lambda \quad ② \\ x_i^* = 0 & \text{if } |b_i| \le \lambda \quad ③ \end{cases}\right\} := S_\lambda(b)$$

↓

we call this soft-thresholding

$x^* = S_\lambda(b)$



How to minimize non-differentiable functions?

$$\min_x f(x)$$

Subgradient Descent:

$$x^{k+1} \leftarrow x^k - \eta^k \cdot g^k$$

$$g^k \in \partial f(x^k)$$

This will converge to x^*, but slow.

Proximal Gradient Descent for Decomposible Functions:

- Decomposible function:

$$f(x) = g(x) + h(x).$$

g: is convex & differentiable

h: is convex, but may be non-differentiable.

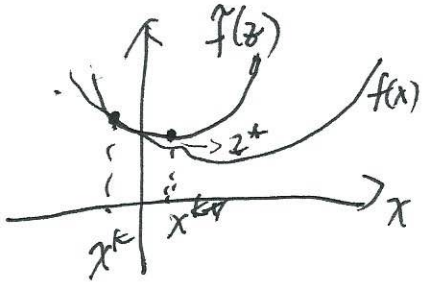
Example: $\underbrace{\frac{1}{2}\|Ax-b\|^2}_{g(x)} + \underbrace{\lambda\|x\|_1}_{h(x)}$

For ML problems: $g(\cdot)$: loss function

$h(\cdot)$ = regularization.

Recall gradient descent. $x^{k+1} \leftarrow x^k - \nabla f(x^k)$

Form approximate function:



$$\tilde{f}(z) = f(x^k) + \nabla f(x^k)^T(z - x^k) + \frac{1}{2\eta}\|z - x^k\|^2$$

$$z^* = \arg\min_z \tilde{f}(z)$$

$$x^{k+1} = z^*$$

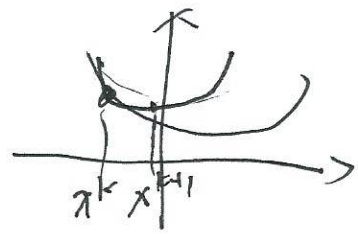
$$\nabla \tilde{f}(z^*) = 0$$

$$\nabla f(x^k) + \frac{1}{\eta}(z^* - x^k) = 0$$

$$\Rightarrow z^* = x^k - \eta \cdot \nabla f(x^k)$$

proximal gradient descent.

$$\min_x \{ g(x) + h(x) \} := f(x)$$



approximate function (approximate)

$$\tilde{f}(z) = \tilde{g}(z) + h(z)$$

$$= g(x^k) + \nabla g(x^k)^T (z - x^k) + \frac{1}{2\eta} \|z - x^k\|^2 + h(z)$$

$$\frac{1}{2\eta}\|z - x^k\|^2 + (z - x^k)^T \nabla g(x^k) + \eta^2 \|\nabla g(x^k)\|^2 \cdot \frac{1}{2\eta}$$

$$= \frac{1}{2\eta} \|z - x^k + \eta \cdot \nabla g(x^k)\|^2 + h(z) + \text{constants}.$$

$$= \frac{1}{2\eta} \|z - (x^k - \eta \nabla g(x^k))\|^2 + h(z)_{\text{const.}}$$

$$z^* = \arg\min_z \tilde{f}(z)$$

$$x^{k+1} = z^*$$

"proximal operator"

close to x — to have small $h(z)$

$$\text{prox}_\eta(x) = \arg\min_z \frac{1}{2\eta} \cdot \|z - x\|^2 + h(z)$$

proximal gradient descent:

$$\boxed{x^{k+1} = \text{prox}_{\eta^k}(x^k - \eta^k \cdot \nabla g(x^k))}$$

$$\left(= P_C(x^k - \eta^k \cdot \nabla g(x^k)) \right)$$

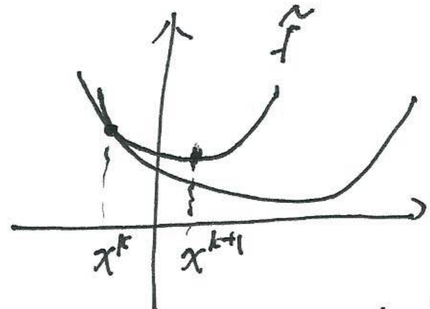
Decomposable functions

$$\min_x \{ g(x) + h(x) \} := f(x)$$

smooth loss function ($\|Ax-b\|_2^2$)

non-differentiable ($\lambda\|x\|_1$) $= \lambda \sum_i |x_i|$

Proximal Gradient Descent



At each iteration

$$\tilde{f}(y) = \tilde{g}(y) + h(y)$$
$$= g(x^k) + \nabla g(x^k)^T (y - x^k) + \frac{1}{2\eta}\|y - x^k\|^2 + h(y)$$

$$x^{k+1} = \operatorname*{argmin}_y \tilde{f}(y)$$
$$= \operatorname*{argmin}_y \boxed{\frac{1}{2\eta}\| y - x^k + \eta \cdot \nabla g(x^k)\|^2} + h(y) + \text{const}$$

$$\frac{1}{2\eta}\|y - x^k\|^2 + \nabla g(x^k)^T (y - x^k) + \frac{1}{2\eta} \cdot \|\eta \nabla g(x^k)\|^2$$

$$= \operatorname*{argmin}_y \frac{1}{2}\|y - (x^k - \eta \nabla g(x^k))\|^2 + \eta h(y)$$

$$= \text{prox}_\eta (x^k - \eta \nabla g(x^k))$$

$$\left(\text{prox}_\eta (x) = \operatorname*{argmin}_y \frac{1}{2}\|y - x\|^2 + \eta h(y) \right)$$

When $h(x) = \lambda \|x\|_1$

$$\text{prox}_\eta(x) = \arg\min_y \frac{1}{2}\|y - x\|^2 + \lambda\eta \cdot \|y\|_1$$

$$= S_{\eta\lambda}(x) = \begin{cases} x - \eta\lambda & \text{if } x > \eta\lambda \\ x + \eta\lambda & \text{if } x < -\eta\lambda \\ 0 & \text{if } |x| < \eta\lambda \end{cases}$$

proximal gradient ~~for~~ when $h(x) = \lambda \|x\|_1$

For $k = 0, 1, \cdots$

$$z = x^k - \eta^k \nabla g(x^k)$$

$$x^{k+1} = S_{\eta^k \lambda}(z)$$

end.

(Iterative Soft Thesholding Algorithm)

ISTA

Convergence:

① will converge to stationary points if

$0 < \underline{\eta} \le \eta^k \le 1/L \quad \forall k$

$(\nabla^2 g(x) \preceq LI \quad \forall x)$

if $\eta^k = 1/k$

$\not\Rightarrow$ converge to solution

② If $g(\cdot)$ is strongly convex

$(mI \preceq \nabla^2 g(x) \preceq LI \quad \forall x)$

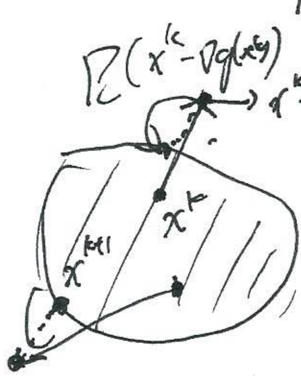
$\Rightarrow$ Linear convergence.

Recall : projected gradient for

Constrained minimization

$$\min_x f(x) \text{ st } x \in C, \quad C \text{ is convex set.}$$

Projected Gradient (PG) :



$$x^{k+1} = P_C(x^k - \eta \cdot \nabla f(x^k))$$

$P_C(z)$: projection of z on C.

$$P_C(z) := \operatorname*{argmin}_y \|y - z\|_2 \text{ st } y \in C$$

PG is a special case for proximal gradient.

Why? equivalent to

$$\min_x f(x) + I_C(x) \quad \text{---- } (*)$$

$$I_C(x) := \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

solve (*) by proximal gradient

$$x^{k+1} = \text{prox}_\eta(x^k - \eta \nabla f(x^k))$$

$$\text{prox}_\eta(z) := \operatorname*{argmin}_y \|y - z\|^2 + \eta I_C(y)$$

$$= \operatorname*{argmin}_y \|y - z\|^2 \text{ st } y \in C$$

$$= P_C(z)$$

Proximal Newton:

Newton: $x^{k+1} \leftarrow \underbrace{x^k - \eta^k \cdot \nabla^2 f(x^k)^{-1} \nabla f(x^k)}$

Hessian

Another: At each iteration. form approximate function

$$\tilde{f}(y) = f(x^k) + \nabla f(x^k)^T (y - x^k) + \frac{1}{2} (y - x^k)^T \frac{\nabla^2 f(x^k)}{\eta} (y - x^k)$$

$$\underset{y}{\operatorname{argmin}} \ \tilde{f}(y)$$

$$\nabla f(x^k) + \frac{1}{\eta} \nabla^2 f(x^k)(y - x^k) = 0$$

$\Rightarrow$ ~~y~~

$$\eta \nabla f(x^k) + \nabla^2 f(x^k)(y - x^k) = 0$$

$$\Rightarrow y = x^k - \eta \nabla^2 f(x^k)^{-1} \nabla f(x^k)$$

proximal Newton: for decomposable function.

$$\min_x \{ \underset{\text{differentiable}}{g(x)} + \underset{\text{non-differentiable.}}{h(x)} \} := f(x)$$

At each iteration

$$\tilde{f}(y) = \tilde{g}(y) + h(y)$$

$$= \underbrace{g(x^k) + \nabla g(x^k)^T (y - x^k) + \frac{1}{2} (y - x^k)^T \nabla^2 g(x^k) (y - x^k)} + h(y)$$

$$x^{k+1} = \underset{y}{\operatorname{argmin}} \ \tilde{f}(y)$$

$(H = \nabla^2 g(x^k))$

$$\arg\min_y \tilde{f}(y)$$

$$= \arg\min_y \frac{1}{2} \| y - (x^k - H^{-1} \nabla g(x^k)) \|_H^2 + h(y)$$

$$= \operatorname{prox}^H (x^k - H^{-1} \nabla g(x^k))$$

$$\|x\|_H^2 = x^T H x$$

$$(y - (x^k - H^{-1} \nabla g(x^k)))^T H \, (y - (x^k) - H^{-1} \nabla g(x^k))$$

$$= (y - x^k)^T H (y - x^k) + 2 (y - x^k)^T \cancel{H \cdot H^{-1}} \nabla g(x^k)$$