P14 . Lemma 1.

(i) ~~strongly~~ strong convexity $\Rightarrow$ strict convexity

f is strongly convex with $\mu > 0$

$\forall\, x, y \qquad f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \| y - x \|^2$

For $x \neq y \qquad \frac{\mu}{2} \| x - y \|^2 > 0$, implies

$$f(y) > f(x) + \nabla f(x)^T (y-x)$$

Let $z := \lambda x + (1-\lambda) y \qquad \lambda \in (0,1)$

By strong convexity of f.

$f(x) > f(z) + \nabla f(z)^T (x - z) \qquad (*)$

$\quad = f(\lambda x + (1-\lambda) y) + \nabla f(z)^T (1-\lambda)(x-y)$

and

$f(y) > f(z) + \nabla f(z)^T (y - z)$

$\quad = f(\lambda x + (1-\lambda) y) + \nabla f(z)^T \lambda (y - x) \qquad (**)$.

$\lambda (*) + (1-\lambda)(**)$, gradient terms will cancel.

$$\lambda f(x) + (1-\lambda) f(y) > f(\lambda x + (1-\lambda) y)$$

Thus, f is strictly convex.

(ii) strong convexity $\Rightarrow$ unique global minimum

$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \| y - x \|^2$

If x^* is the minimizer of f, then by first-order optimality

$\nabla f(x^*) = 0$

substituting $x = x^*$, then

$f(y) \geq f(x^*) + \frac{\mu}{2} \| y - x^* \|^2$

When $y \neq x^*$, $\frac{\mu}{2} \| y - x^* \|^2 > 0$

This implies $f(y) > f(x^*)$

x^* is the unique global minimizer of f. $\square$.

P15. GD: $x_{t+1} = x_t - \eta \nabla f(x_t)$. for f: μ-scv. L-smooth.

show $\lim_{t\to\infty} x_t = x^*$

Proof: Vanilla analysis from last lecture.

- $\nabla f(x_t)^T (x_t - x^*) = \frac{\eta}{2} \|\nabla f(x_t)\|^2 + \frac{1}{2\eta} (\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)$

note that f is μ-scv

- $f(x_t) - f(x^*) + \frac{\mu}{2} \|x_t - x^*\|^2 \leq \nabla f(x_t)^T (x_t - x^*)$.

Putting it together,

$$f(x_t) - f(x^*) + \frac{\mu}{2} \|x_t - x^*\|^2 \leq \frac{\eta}{2} \|\nabla f(x_t)\|^2 + \frac{1}{2\eta} (\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)$$

Rearranging,

$$\frac{1}{2\eta} \|x_{t+1} - x^*\|^2 \leq -f(x_t) + f(x^*) - \frac{\mu}{2} \|x_t - x^*\|^2 + \frac{\eta}{2} \|\nabla f(x_t)\|^2 + \frac{1}{2\eta} \|x_t - x^*\|^2$$

Multiplying 2η on both sides.

$$\underline{\|x_{t+1} - x^*\|^2} \leq \underbrace{2\eta (f(x^*) - f(x_t)) + \eta^2 \|\nabla f(x_t)\|^2}_{\text{"noise"}} + \underline{(1 - \mu\eta) \|x_t - x^*\|^2}$$

$0 < 1 - \mu\eta < 1$ if $\eta < \frac{1}{\mu}$

P17. Theorem 1.

① show $\|x_{t+1} - x^*\|^2 \leq (1 - \frac{\mu}{L}) \|x_t - x^*\|^2$

$\eta = \frac{1}{L}$, then

$2\eta (f(x^*) - f(x_t)) + \eta^2 \|\nabla f(x_t)\|^2$

$= \frac{2}{L} (f(x^*) - f(x_t)) + \frac{1}{L^2} \|\nabla f(x_t)\|^2$

$\leq \frac{2}{L} (f(x_{t+1}) - f(x_t)) + \frac{1}{L^2} \|\nabla f(x_t)\|^2$

Recall "sufficient decrease" (Lemma 3 from last lecture)

$f(x_{t+1}) - f(x_t) \leq - \frac{1}{2L} \|\nabla f(x_t)\|^2$

This implies.

$2\eta (f(x^*) - f(x_t)) + \eta^2 \|\nabla f(x_t)\|^2 \leq - \frac{1}{L^2} \|\nabla f(x_t)\|^2 + \frac{1}{L^2} \|\nabla f(x_t)\|^2 = 0$.

Thus.

$$\|x_{t+1} - x^*\|^2 \leq \underbrace{2\eta (f(x^*) - f(x_t)) + \eta^2 \|\nabla f(x_t)\|^2}_{\leq 0} + (1 - \mu\eta) \|x_t - x^*\|^2$$

$$\leq (1 - \mu\eta) \|x_t - x^*\|^2$$

② show $f(x_T) - f(x^*) \leq \frac{L}{2}(1-\frac{\mu}{L})^T \|x_0 - x^*\|^2$

It is shown that

$$\|x_T - x^*\|^2 \leq (1-\frac{\mu}{L})^T \|x_0 - x^*\|^2 \qquad (**).$$

Smoothness together with $\nabla f(x^*) = 0$

$$f(x_T) - f(x^*) \leq \nabla f(x^*)^T (x_T - x^*) + \frac{L}{2}\|x_T - x^*\|^2$$

$$= \frac{L}{2}\|x_T - x^*\|^2 \qquad (*)$$

Putting $(*)$ $(**)$ together,

$$f(x_T) - f(x^*) \leq \frac{L}{2}\|x_T - x^*\|^2$$

$$\leq \frac{L}{2}(1-\frac{\mu}{L})^T \|x_0 - x^*\|^2$$

$\square$.

$R := \|x_0 - x^*\|$.

Then. $f(x_T) - f(x^*) \leq \frac{L}{2}(1-\frac{\mu}{L})^T R^2$

$$\frac{L}{2}(1-\frac{\mu}{L})^T R^2 \leq \varepsilon \iff (1-\frac{\mu}{L})^T \leq \frac{2\varepsilon}{LR^2}$$

$$\iff T \ln(1-\frac{\mu}{L}) \leq \ln(\frac{2\varepsilon}{LR^2})$$

Note that $\ln(1-\frac{\mu}{L}) < 0$ as $1-\frac{\mu}{L} < 1$

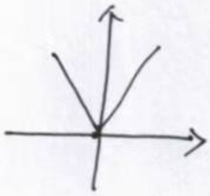
$$T \geq \frac{\ln(\frac{2\varepsilon}{LR^2})}{\ln(1-\frac{\mu}{L})}.$$

$\ln(1-x) \approx -x$ for small x

$$\ln(1-\frac{\mu}{L}) \approx -\frac{\mu}{L}$$

$$T \geq \frac{\ln(\frac{2\varepsilon}{LR^2})}{-\frac{\mu}{L}} = \frac{L}{\mu}\ln(\frac{R^2 L}{2\varepsilon}).$$

$f(x) = |x|$.



P29. For convex and nondifferentiable f, bounded subgradients not equivalent to Lipschitz.

① Lipschitz $\Rightarrow$ bounded subgradient.

$\forall\ g_x \in \partial f(x)$.

If $g_x = 0$, done.

Otherwise, set $y = x + g_x$.

$$L \| g_x \| = L \| y - x \| \geq |f(y) - f(x)| \quad \text{Lipschitz.}$$

$$\geq |\langle g_x, g_x \rangle| \quad \text{convexity.}$$

$$= \| g_x \|^2$$

Thus $\| g_x \| \leq L$

② bounded subgradient $\not\Rightarrow$ Lipschitz.

$f(x) = -\sqrt{x} \qquad 0 \leq x \leq 1$

- $f(x)$ is convex
- $\partial f(x) = \begin{cases} -\frac{1}{2\sqrt{x}} & x > 0. \\ [-1, 1] & x = 0 \end{cases}$
- $\dfrac{|f(x) - f(0)|}{|x - 0|} = \dfrac{1}{\sqrt{x}} \to \infty \quad (x \to 0).$

Bounded subgradient doesnot guarantee Lipschitzness.

$\square$

P29. Strong convexity and Lipschitzness contradict.

Assume $f: \mathbb{R}^d \to \mathbb{R}$ is non-differentiable.

By strong convexity: $g_x \in \partial f(x)$ $\quad x \neq x^*$

$$f(x^*) \geq f(x) + g_x^T (x^* - x) + \frac{\mu}{2} \|x - x^*\|$$

Note that $f(x^*) \leq f(x)$

$$\frac{\mu}{2} \|x - x^*\|^2 \leq g_x^T (x - x^*)$$

$$\leq \|g_x\| \|x - x^*\| \qquad \text{Cauchy-Schwarz.}$$

Thus $\|g_x\| \geq \frac{\mu}{2} \|x - x^*\|$

Note that Lipschitzness implies bounded subgradient

$\exists L > 0$ s.t.

$$\sup_{g_x \in \partial f(x)} \|g_x\| \leq L, \quad \forall x$$

If a function f is both μ-strongly convex and L-Lipschitz.

$$\frac{\mu}{2} \|x - x^*\| \leq \|g_x\| \leq L.$$

as $\|x - x^*\| \to \infty$ $\qquad \frac{\mu}{2} \|x - x^*\| \to \infty$

Thus, non-differentiable function cannot be both Lipschitz and strongly convex.

□

P30 Theorem 3.

By vanilla analysis from previous lecture.

$$g_t^T (x_t - x^*) = \frac{\eta_t}{2} \|g_t\|^2 + \frac{1}{2\eta_t} (\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2).$$

$$g_t = \partial f(x_t)$$

By strong convexity.

$$g_t^T (x_t - x^*) \geq f(x_t) - f(x^*) + \frac{\mu}{2} \|x_t - x^*\|^2$$

Putting it together with $\|g_t\|^2 \leq B^2$

$$f(x_t) - f(x^*) \leq \frac{\eta_t}{2} B^2 + \left(\frac{1}{2\eta_t} - \frac{\mu}{2}\right) \|x_t - x^*\|^2$$

$$- \frac{1}{2\eta_t} \|x_{t+1} - x^*\|^2$$

set $\eta_t = \frac{2}{\mu(1+t)}$.

$$t(f(x_t) - f(x^*)) \leq \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4} [t(t-1) \|x_t - x^*\|^2 - t(t+1) \|x_{t+1} - x^*\|^2]$$

$$\leq \frac{B^2}{\mu} + \frac{\mu}{4} [t(t-1) \|x_t - x^*\|^2 - t(t+1) \|x_{t+1} - x^*\|^2]$$

$$\sum_{t=1}^{T} t(f(x_t) - f(x^*)) \leq \frac{TB^2}{\mu} + \frac{\mu}{4} [0 - T(T+1) \|x_{T+1} + x^*\|^2]$$

$$\leq \frac{TB^2}{\mu}. \quad (**).$$

note that $\sum_{t=1}^{T} t = \frac{(1+T)T}{2} \Rightarrow \frac{2}{T(T+1)} \sum_{t=1}^{T} t = 1.$

Recall Jensen's inequality:

f is convex. $\lambda_1 \cdots \lambda_m$ $\sum_{i=1}^{m} \lambda_i = 1$. $f(\sum_{i=1}^{m} \lambda_i x_i) \leq \sum_{i=1}^{m} \lambda_i f(x_i).$

$$f\left(\frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot x_t\right) - f(x^*) \leq \frac{2}{T(T+1)} \sum_{t=1}^{T} t f(x_t) - f(x^*)$$

$$= \frac{2}{T(T+1)} \sum_{t=1}^{T} t [f(x_t) - f(x^*)] \quad (*)$$

Putting (*) (**) together, we have:

$$f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t\, x_t\right) - f(x^*) \leq \frac{2B^2}{\mu(T+1)} \qquad \square$$

$$B := \max_{t=1}^{T} \|g_t\|$$

$$f(x) = x^2 \qquad \nabla f(x) = 2x$$

$$\frac{2B^2}{\mu(T+1)} \leq \frac{2B^2}{\mu T} \leq \varepsilon \iff T \geq \frac{2B^2}{\mu\varepsilon}$$

P43. Fact 1.

(i) $\Pi_X(y) := \underset{x \in X}{\arg\min} \|x - y\|$

$\Pi_X(y)$ is the minimizer of $d_y = \|x-y\|^2$ over $x \in X$

By first-order characterization of optimality

(Lemma 4 from lecture 2)

$$0 \leq \nabla d_y(\Pi_X(y))^T (x - \Pi_X(y))$$

$$= 2(\Pi_X(y) - y)(x - \Pi_X(y))$$

This implies $\quad 0 \geq (x - \Pi_X(y))^T (y - \Pi_X(y))$

(ii) Note that

$$2v^T w = \|v\|^2 + \|w\|^2 - \|v-w\|^2 \quad \forall\, v, w$$

$$v := x - \Pi_X(y) \qquad w := y - \Pi_X(y).$$

By (i), $0 \geq 2v^T w = \|v\|^2 + \|w\|^2 - \|v-w\|^2$

$$= \|x - \Pi_X(y)\|^2 + \|y - \Pi_X(y)\|^2 - \|x-y\|^2$$

$$\|x - \Pi_X(y)\|^2 + \|y - \Pi_X(y)\|^2 \leq \|x-y\|^2 \qquad \square$$

P46 . Theorem 4.

GD: $x_{t+1} = x_t - \gamma \nabla f(x_t)$ $\qquad g_t = \nabla f(x_t)$.

Vanilla analysis

$$g_t^T(x_t - x^*) = \frac{1}{2\gamma}\left(\gamma^2\|g_t\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right)$$

PGD $\quad y_{t+1} = x_t - \gamma \nabla f(x_t) \qquad x_{t+1} = \Pi_X(y_{t+1})$.

$$g_t^T(x_t - x^*) = \frac{1}{2\gamma}\left(\gamma^2\|g_t\|^2 + \|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right) \quad (*).$$

By Fact 1 (ii):

$$\|x - \Pi_X(y)\|^2 + \|y - \Pi_X(y)\|^2 \leq \|x - y\|^2$$

set $x = x^*$ $\qquad y = y_{t+1}$

note that $\Pi_X(y) = x_{t+1}$

$$\|x^* - x_{t+1}\|^2 \leq \|x^* - y_{t+1}\|^2 \qquad (**).$$

Thus, $(*)$ $(**)$ gives

$$g_t^T(x_t - x^*) \leq \frac{1}{2\gamma}\left(\gamma^2\|g_t\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right).$$

$$= \frac{\gamma}{2}\|g_t\|^2 + \frac{1}{2\gamma}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right)$$

$$\sum_{t=0}^{T-1} g_t^T(x_t - x^*) = \frac{\gamma}{2}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{1}{2\gamma}\left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2\right).$$

note that $f(x_t) - f(x^*) \leq g_t^T(x_t - x^*)$

$$\sum_{t=0}^{T-1}\left(f(x_t) - f(x^*)\right) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{1}{2\gamma}\left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2\right)$$

$\|x_0 - x^*\| \leq R \qquad \|g_t\| \leq B$

Then, $\displaystyle\sum_{t=0}^{T-1}\left(f(x_t) - f(x^*)\right) \leq \frac{\gamma}{2}B^2T + \frac{1}{2\gamma}R^2$

$$\gamma := \frac{R}{B\sqrt{T}}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(f(x_t) - f(x^*)\right) \leq \frac{RB}{\sqrt{T}} \qquad \square$$

P49. Sufficient Decrease.

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L} \|\nabla f(x_t)\|^2 + \underbrace{\frac{L}{2} \|y_{t+1} - x_{t+1}\|^2}$$

PGD: $\quad y_{t+1} = x_t - \gamma \nabla f(x_t) \qquad x_{t+1} = \Pi_X(y_{t+1})$.

$$f(x_{t+1}) \leq f(x_t) + \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{L}{2} \|x_t - x_{t+1}\|^2 \qquad L\text{-smooth.}$$

$$= f(x_t) - L(y_{t+1} - x_t)^T (x_{t+1} - x_t) + \frac{L}{2} \|x_t - x_{t+1}\|^2$$

Using fact $2v^T w = \|v\|^2 + \|w\|^2 - \|v - w\|^2$

$$f(x_{t+1}) \leq f(x_t) - \frac{L}{2}\left(\|y_{t+1} - x_t\|^2 + \|x_{t+1} - x_t\|^2 - \|y_{t+1} - x_{t+1}\|^2\right)$$

$$+ \frac{L}{2} \|x_t - x_{t+1}\|^2$$

$$= f(x_t) - \frac{L}{2} \|y_{t+1} - x_t\|^2 + \frac{L}{2} \|y_{t+1} - x_{t+1}\|^2$$

$$= f(x_t) - \frac{1}{2L} \|\nabla f(x_t)\|^2 + \frac{L}{2} \|y_{t+1} - x_{t+1}\|^2$$

□