P11. $\Pi_X^{\phi}(y) = \underset{x \in X}{\arg\min}\, D_\phi(x, y)$ (*).

$\Delta_d = \{ x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i = 1, \ x_i \geq 0, \ \forall i \}$

$\phi(x) = \sum_{i=1}^{d} x_i \log x_i \qquad x_i \in \mathbb{R}_{++}^d = \{ x \in \mathbb{R}^d : x_i > 0, \ \forall i \}$

$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle$

$\nabla \phi(y) = (\log y_1 + 1, \log y_2 + 1, \cdots, \log y_d + 1)$

The Lagrangian for (*) is

$$L(x, \lambda) = D_\phi(x, y) + \lambda \left(\sum_{i=1}^{d} x_i - 1 \right)$$

λ: Lagrange multiplier associated with $\sum_{i=1}^{d} x_i = 1$

$$\frac{\partial}{\partial x_i} D_\phi(x, y) = \log\left(\frac{x_i}{y_i}\right)$$

Thus. $\nabla_x D_\phi(x, y) = \left(\log\left(\frac{x_1}{y_1}\right), \cdots, \log\left(\frac{x_d}{y_d}\right) \right)$

$$\frac{\partial}{\partial x_i} L(x, \lambda) = \log \frac{x_i}{y_i} + \lambda = 0 \Rightarrow \frac{x_i}{y_i} = e^{-\lambda}$$

then $e^{-\lambda} \sum_{i=1}^{d} y_i = 1$ since $\sum_{i=1}^{d} x_i = 1$

Thus. $e^{-\lambda} = \frac{1}{\sum_{i=1}^{d} y_i}$

so, $x_i = e^{-\lambda} y_i = \frac{y_i}{\sum_{i=1}^{d} y_i}$

$$x^* = \frac{y}{\sum_{i=1}^{d} y_i} = \frac{y}{\|y\|_1} \qquad (y_i > 0).$$

$\square$.

P11 $\quad \phi(x) = \sum_{i=1}^{d} x_i \log x_i$

Sketch:

Extreme cases: $\quad x_1 = \begin{pmatrix} \frac{1}{d} \\ \vdots \\ \frac{1}{d} \end{pmatrix} \qquad x_2 = \begin{pmatrix} \varepsilon \\ \varepsilon \\ \vdots \\ \varepsilon \\ 1-(d-1)\varepsilon \end{pmatrix}$ $\quad \varepsilon \to$ small

$$\phi(x_1) = -\log d. \qquad \phi(x_2) \to 0 \quad (\varepsilon \to 0).$$

$$-\log d \leq \phi(x) < 0$$

$$x_0 \in \underset{x \in X}{\arg\min}\, \phi(x) \qquad x_0 = x_1$$

$$R^2 = \sup_x \phi(x) - \phi(x_0)$$

$$R^2 \leq \log d.$$

P18.

$f(x) = \frac{1}{2}(f_1(x) + f_2(x))$

$f_1(x) = 2x^2 \qquad f_2(x) = -x^2$

$f(x) = \frac{1}{2}x^2$

SGD : $\quad x_{k+1} = x_k - \eta \nabla f_{i_k}(x_k).$

When $\quad i_k = 2 \qquad f_2(x) = -x^2 \qquad \nabla f_2(x) = -2x$

SGD update : $\quad x_{k+1} = x_k - \eta(-2x_k)$

$= (1+2\eta)x_k$

Thus, $\quad x_{k+1} = (1+2\eta)x_k > x_k$

Recall $\quad f(x) = \frac{1}{2}x^2$

$f(x_{k+1}) = \frac{1}{2}(1+2\eta)^2 x_k^2$

since $\quad (1+2\eta)^2 > 1 \quad$ for any $\quad \eta > 0$

then $\quad f(x_{k+1}) > f(x_k).$

P21. $\mathbb{E}[X] = \sum_i \mathbb{E}[X \mid A_i] \Pr[A_i]$

X is partitioned into disjoint events $A_1, A_2 \ldots$ (countable).

$A_i = \{Y = y\}$.

$\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y=y] \Pr(Y=y)$

P22. Show that for convex f,

$$\mathbb{E}[g_t^T(x_t - x^*)] \geq \mathbb{E}[f(x_t) - f(x^*)]$$

$\mathbb{E}[g_t^T(x - x^*) \mid x_t = x]$

$= \mathbb{E}[g_t \mid x_t = x]^T (x - x^*)$

$= \nabla f(x)^T (x - x^*)$

By Partition theorem,

$\mathbb{E}[g_t^T(x_t - x^*)] = \sum_x \mathbb{E}[g_t^T(x - x^*) \mid x_t = x] \Pr(x_t = x)$

$= \sum_x \nabla f(x)^T (x - x^*) \Pr(x_t = x)$

$= \mathbb{E}[\nabla f(x)^T (x_t - x^*)]$

Thus. $\mathbb{E}[g_t^T(x_t - x^*)] = \mathbb{E}[\nabla f(x)^T(x_t - x^*)]$

$\geq \mathbb{E}[f(x_t) - f(x^*)]$

X, Y are random varibles

$X \geq Y \Rightarrow \mathbb{E}X \geq \mathbb{E}Y$

$\square$

P23. Theorem 2.

By vanilla analysis from lecture 2.

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

g_t : stochastic gradient.

Taking expectation and using convexity in expectation.

$$\sum_{t=0}^{T-1} \mathbb{E}[f(x_t) - f(x^*)] \leq \sum_{t=0}^{T-1} \mathbb{E}[g_t^T (x_t - x^*)] \qquad P22$$

$$\leq \mathbb{E}\left[\frac{\gamma}{2} \sum_{t=0}^{T-1} \|g_t\|^2 + \frac{1}{2\gamma} \|x_0 - x^*\|^2\right]$$

$$= \frac{\gamma}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|g_t\|^2] + \frac{1}{2\gamma} \|x_0 - x^*\|^2$$

$$\leq \frac{\gamma}{2} T B^2 + \frac{1}{2\gamma} R^2$$

solving $h(\gamma) = \frac{\gamma}{2} B^2 T + \frac{1}{2\gamma} R^2$ to find optimal γ

similar to P17 from lecture 2.

$$\gamma^* = \frac{R}{B\sqrt{T}} \Rightarrow h\left(\frac{R}{B\sqrt{T}}\right) = RB\sqrt{T}$$

Thus $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[f(x_t)] - f(x^*) \leq \frac{RB}{\sqrt{T}}$

$T \geq \frac{R^2 B^2}{\varepsilon} \Rightarrow$ expected error $\leq \frac{RB}{\sqrt{T}} \leq \varepsilon$

same order as gradient descent.

but in expectation !

$\boxed{5}$.

$\square$

P25

GD: $\|\nabla f(x)\|^2 \leq B_{GD}^2$

$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$

$\nabla f(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x)$

Thus $\|\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x)\|^2 \leq B_{GD}^2$

SGD: $\mathbb{E}[\|g_t\|^2] \leq B_{SGD}^2$

$$\mathbb{E}[\|g_t\|^2] = \mathbb{E}[\|\nabla f_{i_t}(x)\|^2] = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x)\|^2 \leq B_{SGD}^2$$

Take $B_{GD}^2 \approx \|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x)\|^2 \leq \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x)\|^2 \approx B_{SGD}^2$

Jensen's inequality: $f(\sum_{i=1}^{n}\frac{1}{n}a_i) \leq \sum_{i=1}^{n}\frac{1}{n}f(a_i)$.

$f(x) = \|x\|^2$

B_{GD}^2 can be smaller than B_{SGD}^2

but often comparable.

P26. Theorem 3.

By vanilla analysis from lecture 2.

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

g_t: stochastic gradient.

Taking expectation on both sides

$$\mathbb{E}[g_t^T (x_t - x^*)] = \frac{\eta_t}{2} \mathbb{E}[\|g_t\|^2] + \frac{1}{2\eta_t} \left(\mathbb{E}[\|x_t - x^*\|^2] - \mathbb{E}[\|x_{t+1} - x^*\|^2] \right)$$

By "strong convexity in expectation"

$$\mathbb{E}[g_t^T (x_t - x^*)] = \mathbb{E}[\nabla f(x_t)^T (x_t - x^*)]$$

$$\geq \mathbb{E}[f(x_t) - f(x^*)] + \frac{\mu}{2} \mathbb{E}[\|x_t - x^*\|^2]$$

Putting together with $\mathbb{E}[\|g_t\|^2] \leq B^2$

$$\mathbb{E}[f(x_t)] - f(x^*) \leq \frac{\eta_t}{2} B^2 + \frac{\eta_t^{-1} - \mu}{2} \mathbb{E}[\|x_t - x^*\|^2]$$

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

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

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

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

summing over $t = 1$ to $t = T$

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

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

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

By Jensen's inequality

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

Taking expectation

$$\mathbb{E}\left[f\left(\sum_{t=1}^{T} \frac{2t}{T(T+1)} x_t \right) \right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{2t}{T(T+1)} f(x_t) \right]$$

Combining ~~Combing~~ this with $(*) \cdot \frac{2}{T(T+1)}$

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

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

same rate as subgradient method

but in expectation ! $\square$

P31.

$$\mathbb{E}[\|\tilde{g}_t - \nabla f(x_t)\|^2]$$

$$= \mathbb{E}[\|\frac{1}{m}\sum_{i\in S_t} \nabla f_i(x_t) - \nabla f(x_t)\|^2]$$

$$= \mathbb{E}[\|\frac{1}{m}\sum_{i\in S_t} (\nabla f_i(x_t) - \nabla f(x_t))\|^2]$$

Note that $\mathbb{E}[\|X_1 + X_2\|^2] = \mathbb{E}[\|X_1\|^2] + \mathbb{E}[\|X_2\|^2]$

if X_1 independent of X_2.

Since individual gradient $\nabla f_i(x_t)$ are independent

and from the same distribution

$$\mathbb{E}[\|\frac{1}{m}\sum_{i\in S_t} (\nabla f_i(x_t) - \nabla f(x_t))\|^2]$$

$$= \frac{1}{m^2}\sum_{i\in S_t} \mathbb{E}[\|\nabla f_i(x_t) - \nabla f(x_t)\|^2]$$

$$= \frac{1}{m^2} m \, \mathbb{E}[\|\nabla f_1(x_t) - \nabla f(x_t)\|^2]$$

$$= \frac{1}{m}\mathbb{E}[\|\nabla f_1(x_t)\|^2 + \|\nabla f(x_t)\|^2 - 2\nabla f(x_t)^T \nabla f_1(x_t)]$$

$$= \frac{1}{m}\mathbb{E}[\|\nabla f_1(x_t)\|^2] + \frac{1}{m}\|\nabla f(x_t)\|^2 - \frac{2}{m}\underbrace{\nabla f(x_t)^T \mathbb{E}[\nabla f_1(x_t)]}_{\|\nabla f(x_t)\|^2}$$

$$= \frac{1}{m}\mathbb{E}[\|\nabla f_1(x_t)\|^2] - \frac{1}{m}\|\nabla f(x_t)\|^2$$

$$\leq \frac{B^2}{m} \to 0 \quad (m \to \infty)$$

P41.

SGD:

$$\|x_k - x^*\| \le \Big(\underbrace{1 - \frac{2\mu}{L+\mu}}_{\rho_{GD} = 1 - \frac{2}{1+\kappa}} \Big)^{k} \|x_0 - x^*\|$$

$$\rho_{GD} \to 1 \quad (\kappa \to \infty).$$

$\kappa = 100$ $\qquad \rho_{GD} \approx 0.98$ $\qquad \rho_{HB} = 0.9.$