P14. Assume $x^* = 0$

$$x^* = \arg\min_{x \in \mathbb{R}^d} f(x)$$

Define shifted function

$$g(y) = f(y + x^*)$$

since $g(0) = f(0\ x^*)$

the minimizer of g is $y=0$

so we can assume W.L.O.G that the minimizer is at 0

rather than x^*.

P14. Theorem 1.

NAGD: $y_t = x_t + \beta_t (x_t - x_{t-1})$

$x_{t+1} = y_t - \eta_t \nabla f(y_t)$

$\eta_t = \frac{1}{L}$ $\beta_t = \frac{\sqrt{L/\mu} - 1}{\sqrt{L/\mu} + 1} = \frac{\sqrt{k} - 1}{\sqrt{k} + 1}$

Assume $x^* = 0$.

Set $\rho^2 := 1 - \frac{1}{\sqrt{k}}$ $u_t = \frac{1}{L} \nabla f(y_t)$ $\boxed{1}$.

$V_t := f(x_t) - f(x^*) + \frac{L}{2} \| x_t - \rho^2 x_{t-1} \|^2$

step 1: show $V_{t+1} \leq \rho^2 V_t$

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

Step 1: note that f is smooth

$$f(x_{k+1}) \leq f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{L}{2} \| x_{k+1} - y_k \|^2$$

$$= f(y_k) + \langle L u_k, x_{k+1} - y_k \rangle + \frac{L}{2} \| x_{k+1} - y_k \|^2 \qquad (*)$$

Thus. $V_{t+1} = f(x_{t+1}) - f(x^*) + \frac{L}{2} \| x_{t+1} - \rho^2 x_t \|^2$

$$\overset{t=k}{\leq} f(y_k) + \langle L u_k, x_{k+1} - y_k \rangle + \frac{L}{2} \| x_{k+1} - y_k \|^2 - f(x^*)$$

$$+ \frac{L}{2} \| x_{t+1} - \rho^2 x_t \|^2 \qquad \boxed{\text{By } (*)}$$

$$\overset{t=k}{=} f(y_k) - f(x^*) - \frac{L}{2} \| u_k \|^2 + \frac{L}{2} \| x_{k+1} - \rho^2 x_t \|^2 \qquad \boxed{\text{By } x_{k+1} - y_k = -u_k}$$

$$= \rho^2 [\underbrace{f(y_k) - f^*}_{} + \underline{L \langle u_k, x_k - y_k \rangle}] - \underline{\rho^2 L \langle u_k, x_k - y_k \rangle}$$

$$+ (1-\rho^2) [\underbrace{f(y_k) - f^*}_{} - L \langle u_k, y_k \rangle] + (1-\rho^2) L \langle u_k, y_k \rangle$$

$$- \frac{L}{2} \| u_k \|^2 + \frac{L}{2} \| x_{k+1} - \rho^2 x_k \|^2 \qquad \boxed{\text{By adding and subtract terms}}$$

By strong convexity of f,

$$f(y_k) + \langle \underbrace{\nabla f(y_k)}_{= L u_k}, x_k - y_k \rangle \leq f(x_k) - \frac{\mu}{2} \| x_k - y_k \|^2$$

$\boxed{2}$

$$f(x^*) \geq f(y_k) - \langle \nabla f(y_k), y_k - x^* \rangle + \frac{\mu}{2} \| y_k - x^* \|^2$$

$$= f(y_k) - L \langle u_k, y_k \rangle + \frac{\mu}{2} \| y_k \|^2 \qquad (\text{By } x^* = 0)$$

Combining last three displays yields

$$V^{k+1} \leq \rho^2 [f(x_k) - f(x^*) - \frac{\mu}{2} \| x_k - y_k \|^2] - \rho^2 L \langle u_k, x_k - y_k \rangle$$

$$- \frac{L}{2} \| u_k \|^2 + \frac{L}{2} \| x_{k+1} - \rho^2 x_k \|^2$$

$$= \underbrace{\rho^2 [f(x_k) - f(x^*) + \frac{L}{2} \| x_k - \rho^2 x_{k-1} \|^2]}_{V_k} + R_k$$

$$R_k := -\rho^2 \frac{\mu}{2} \|x_k - y_k\|^2 - (1-\rho^2)\frac{\mu}{2}\|y_k\|^2$$
$$+ L\langle u_k, y_k - \rho^2 x_k\rangle - \frac{L}{2}\|u_k\|^2$$
$$+ \frac{L}{2}\|x_{k+1} - \rho^2 x_k\|^2 - \frac{\rho^2 L}{2}\|x_k - \rho^2 x_{k-1}\|^2$$

$\boxed{\text{Claim:}}$ Under the choice of y_t, β_t, ρ, we have

$$R_k = -\frac{1}{2}L\rho^2\left(\frac{1}{\sqrt{k}} + \frac{1}{k}\right)\|x_k - y_k\|^2 \leq 0$$

Proof: substitute the definitions of y_t, β_t, ρ, x_{k+1}, y_k into the definition of R_k (Verify it yourself!)

$$V_{k+1} \leq \rho^2 V_k \qquad \forall k$$

step 2: $f^* := f(x^*)$

$$f(x_k) - f^* \leq V_k \leq \rho^{2k} V_0$$

$\boxed{3}$

$$V_0 = f(x_0) - f^* + \frac{L}{2}\|x_0 - \rho^2 x_0\|^2 \qquad \text{By } x_{-1} = x_0$$
$$= f(x_0) - f^* + \frac{\mu}{2}\|x_0\|^2 \qquad \text{By } (1-\rho^2)^2 = \frac{1}{k} = \frac{\mu}{L}$$
$$= f(x_0) - f^* + \frac{\mu}{2}\|x_0 - x^*\|^2 \qquad \text{By } x^* = 0$$
$$\leq \frac{L}{2}\|x_0 - x^*\|^2 + \frac{\mu}{2}\|x_0 - x^*\|^2 \qquad \text{By smoothness of } f,\ \nabla f(x^*) = 0$$
$$= \frac{L+\mu}{2}\|x_0 - x^*\|^2$$

$$f(x_k) - f^* \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \frac{L+m}{2}\|x_0 - x^*\|^2 \qquad \text{By } \rho^2 = 1 - \sqrt{\frac{\mu}{L}}$$

set $R^2 := \|x_0 - x^*\|^2$

$$f(x_T) - f^* \leq \left(1 + \sqrt{\frac{\mu}{L}}\right)^T \left(\frac{L+m}{2}\right) R^2$$

$\square$

P20. Theorem 2.

Recall sufficient decrease (Lemma 3 from lecture 2).

$$f(x_{k+1}) \le f(y_k) - \frac{1}{2L} \|\nabla f(y_k)\|^2$$

Thus,

$$f(x_{k+1}) - f(x_k) = f(x_{k+1}) - f(y_k) + f(y_k) - f(x_k)$$

$$\le -\frac{1}{2L} \|\nabla f(y_k)\|^2 + \langle \nabla f(y_k), y_k - x_k \rangle$$

$$= -\frac{L}{2} \| y_k - x_{k+1} \|^2 + L \langle y_k - x_{k+1}, y_k - x_k \rangle \quad (*)$$

The last step follows from $\nabla f(y_k) = L(y_k - x_{k+1})$

Similarly,

$$f(x_{k+1}) - f(x^*) = f(x_{k+1}) - f(y_k) + f(y_k) - f(x^*)$$

$$\le -\frac{1}{2L} \|\nabla f(y_k)\|^2 + \langle \nabla f(y_k), y_k - x^* \rangle$$

$$= -\frac{L}{2} \| y_k - x_{k+1} \|^2 + L \langle y_k - x_{k+1}, y_k - x^* \rangle \quad (**)$$

Define $\Delta_k = f(x_k) - f(x^*)$

Taking $(*) \times \lambda_k(\lambda_k - 1) + (**) \times \lambda_k$

$\boxed{4}$

$$\lambda_k(\lambda_k - 1)(\Delta_{k+1} - \Delta_k) + \lambda_k \Delta_{k+1}$$

$$\le \frac{L}{2} \left[2 \langle \lambda_k (y_k - x_{k+1}), \lambda_k y_k - (\lambda_k - 1) x_k - x^* \rangle - \| \lambda_k (y_k - x_{k+1}) \|^2 \right]$$

note that $\lambda_k^2 - \lambda_k = \lambda_{k-1}^2$

$$2\langle a, b \rangle - \|a\|_2^2 = \|b\|_2^2 - \|b - a\|_2^2$$

$$\lambda_k^2 \Delta_{k+1} - \lambda_{k-1}^2 \Delta_k \le \frac{L}{2} \left[\| \lambda_k y_k - (\lambda_k - 1) x_k - x^* \|_2^2 - \| \lambda_k x_{k+1} - (\lambda_k - 1) x_k - x^* \|_2^2 \right] \quad (\Delta\Delta).$$

By definitions of β_{k+1} , y_{k+1}

$$y_{k+1} = x_{k+1} + \beta_{k+1} (x_{k+1} - x_k)$$

$$= x_{k+1} + \frac{\lambda_k - 1}{\lambda_{k+1}} (x_{k+1} - x_k)$$

This implies

$$\lambda_{k+1} y_{k+1} - (\lambda_{k+1} - 1) x_{k+1} = \lambda_k x_{k+1} - (\lambda_k - 1) x_k \qquad (\Delta)$$

Combining (Δ) $(\Delta\Delta)$

$$\lambda_k^2 \Delta_{k+1} - \lambda_{k-1}^2 \Delta_k \leqslant \frac{L}{2} [\| \lambda_k y_k - (\lambda_k - 1) x_k - x^* \|_2^2$$

$$- \| \lambda_{k+1} y_{k+1} - (\lambda_{k+1} - 1) x_{k+1} - x^* \|_2^2]$$

Summing over k , note that $\lambda_0 = 0$ $\lambda_1 = 1$ $\beta_1 = -1$ $y_1 = x_0$

$$\lambda_k^2 \Delta_{k+1} - \lambda_0^2 \Delta_1 \leqslant \frac{L}{2} \| \lambda_1 y_1 - (\lambda_1 - 1) x_1 - x^* \|^2$$

$$\Rightarrow \qquad \lambda_k^2 \Delta_{k+1} \leqslant \frac{L}{2} \| x_0 - x^* \|^2 \qquad \boxed{\Sigma}$$

Finally, note that

$$\lambda_k \geqslant \frac{1 + \sqrt{4 \lambda_{k-1}^2}}{2} = \lambda_{k-1} + \frac{1}{2}$$

Together with $\lambda_1 = 1$, $\lambda_k \geqslant \frac{k+1}{2}$ $\forall k$

Thus, $f(x_{k+1}) - f(x^*) = \Delta_{k+1} \leqslant \dfrac{2L \| x_0 - x^* \|_2^2}{(k+1)^2}$

When $R^2 = \| x_0 - x^* \|^2$

$$f(x_{T+1}) - f(x^*) \leqslant \frac{2LR^2}{T^2} \qquad \square.$$

P41. Theorem 3.

$$x_{t+1} = x_t - \underbrace{\frac{\alpha}{\sqrt{G_t}}} \nabla f(x_t) \qquad G_t = \sum_{j=1}^{t} \|\nabla f(x_j)\|^2$$

$$= x_t - \eta_t \nabla f(x_t)$$

From lecture 4.

$$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

is equivalent to

$$x^t_{t+1} = \arg\min_x \underbrace{\{ \nabla f(x_t)^T (x - x_t) + \frac{1}{2\eta_t} \|x - x_t\|^2 \}}_{\phi(x)}$$

The first-order optimality condition gives

$$\langle \nabla \phi(x^\circ), x^\circ - x \rangle \le 0 \qquad \forall x$$

if $x^\circ = \arg\min_x \phi(x)$

Thus $\langle \nabla \phi(x_{t+1}), x_{t+1} - x^* \rangle \le 0$ $\boxed{6}$

$$\Rightarrow \langle \nabla f(x_t), x_{t+1} - x^* \rangle \le \frac{1}{\eta_t} \langle x_t - x_{t+1}, x_{t+1} - x^* \rangle$$

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

The last step is based on $ab = \frac{1}{2}(a+b)^2 - \frac{a^2}{2} - \frac{b^2}{2}$ $\forall a, b$

Note that

$$\langle \nabla f(x_t), x_t - x^* \rangle = \langle \nabla f(x_t), x_{t+1} - x^* \rangle - \langle \nabla f(x_t), x_{t+1} - x_t \rangle$$

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

$$- \langle \nabla f(x_t), x_{t+1} - x_t \rangle - \frac{1}{2\eta_t} \|x_{t+1} - x_t\|^2$$

(**).

By ~~the~~ Cauchy-Schwarz inequality

$$-\langle \nabla f(x_t), x_{t+1} - x_t \rangle \leq \|\nabla f(x_t)\| \, \|x_{t+1} - x_t\|$$

Thus $$-\langle \nabla f(x_t), x_{t+1} - x_t \rangle - \frac{1}{2\eta_t} \|x_{t+1} - x_t\|^2$$

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

$$\leq \frac{\eta_t}{2} \|\nabla f(x_t)\|^2 \qquad (*)$$

$f(x) = ax - bx^2$

$f'(x) = a - 2bx = 0 \Rightarrow x = \frac{a}{2b}$

Combining $(*)$ $(**)$

$$\langle \nabla f(x_t), x_t - x^* \rangle \leq \frac{1}{2\eta_t} \|x_t - x^*\|^2 - \frac{1}{2\eta_t} \|x_{t+1} - x^*\|^2 + \frac{\eta_t}{2} \|\nabla f(x_t)\|^2$$

summing up and collecting terms

$$\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle \leq \sum_{t=2}^{T} \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \underbrace{\|x_t - x^*\|^2}_{R^2}$$

$$+ \frac{1}{2\eta_1} \underbrace{\|x_2 - x^*\|^2}_{R^2}$$

$$+ \sum_{t=1}^{T} \frac{\eta_t}{2} \|\nabla f(x_t)\|^2 \qquad (1)$$

$$\leq \frac{R^2}{2\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \|\nabla f(x_t)\|^2 \qquad (2)$$

Recall $\eta_t = \dfrac{R}{\sqrt{\sum_{i=1}^{t} \|\nabla f(x_i)\|^2}}$

(3) $$\frac{R^2}{\eta_T} = R\sqrt{\sum_{t=1}^{T} \|\nabla f(x_t)\|^2}$$

$$\sum_{t=1}^{T} \eta_t \|\nabla f(x_t)\|^2 = R \sum_{t=1}^{T} \frac{\|\nabla f(x_t)\|^2}{\sqrt{\sum_{i=1}^{t} \|\nabla f(x_i)\|^2}}$$

Claim: For $\forall\, a_1 \ldots a_T > 0$

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{\sum_{s=1}^{t} a_s}} \leq 2\sqrt{\sum_{t=1}^{T} a_t}$$

think of $\frac{a_t}{\sqrt{\sum_{s=1}^{t} a_s}}$ as $\frac{dx}{\sqrt{x}}$ and recall that

$$\int \frac{1}{\sqrt{x}}\, dx = \sqrt{x} + C$$

Using this claim,

$$\sum_{t=1}^{T} \frac{\|\nabla f(x_t)\|^2}{\sqrt{\sum_{i=1}^{t} \|\nabla f(x_i)\|^2}} \leq 2\sqrt{\sum_{t=1}^{T} \|\nabla f(x_t)\|^2} \qquad (1)$$

Combining (1) (2) (3)

$$\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle \leq \frac{3}{2} R \sqrt{\sum_{t=1}^{T} \|\nabla f(x_t)\|^2} \leq \frac{3}{2} RL\sqrt{T}$$

By convexity of f

$$\begin{aligned}
f(\underbrace{\tfrac{1}{T}\textstyle\sum_{t=1}^{T} x_i}_{\bar{x}_T}) - f(x^*) &\leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \\
&= \frac{1}{T}\sum_{t=1}^{T} [f(x_t) - f(x^*)] \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle
\end{aligned}$$

Thus, $f(\bar{x}_T) - f(x^*) \leq \frac{3}{2}\frac{RL}{\sqrt{T}}$ $\square$