P9. Theorem 1.

$$y_{t+1} = x_t - \eta_t g_t \qquad g_t = \nabla f(x_t) \qquad \eta_t = \eta$$

$$x_{t+1} = \Pi_X(y_{t+1})$$

By "sufficient decrease" on P8

$$\frac{1}{2L} \|g_t\|^2 \le f(x_t) - f(x_{t+1}) + \frac{L}{2} \|y_{t+1} - x_{t+1}\|^2 \qquad (***)$$

Recall the vanilla analysis for unprojected gradient step

(P14 from lecture 2, replace x_{t+1} with y_{t+1})

$$g_t^T (x_t - x^*) = \frac{1}{2\eta} (\eta^2 \|g_t\|^2 + \|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2) \qquad (**)$$

Using Fact (ii): $\|x - \Pi_X(y)\|^2 + \|y - \Pi_X(y)\|^2 \le \|x - y\|^2$

with $x = x^*$ $y = y_{t+1}$, then $\Pi_X(y) = x_{t+1}$

$$\|x^* - x_{t+1}\|^2 + \|y - \Pi_X(y)\|^2 \le \|x^* - y_{t+1}\|^2 \qquad (*)$$

Combine $(*)$ and $(**)$

$$g_t^T (x_t - x^*) \le \frac{1}{2\eta} (\eta^2 \|g_t\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 - \|y_{t+1} - x_{t+1}\|^2)$$

By convexity of f,

$$f(x_t) - f(x^*) \le g_t^T (x_t - x^*)$$

Thus, when $\eta = \frac{1}{L}$

$$\sum_{t=0}^{T-1} (f(x_t) - f(x^*)) \le \frac{1}{2L} \sum_{t=0}^{T-1} \|g_t\|^2 + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{t=1}^{T-1} \|y_{t+1} - x_{t+1}\|^2$$

note that by $(***)$

$$\frac{1}{2L} \sum_{t=0}^{T-1} \|g_t\|^2 \le \sum_{t=0}^{T-1} \left(f(x_t) - f(x_{t+1}) + \frac{L}{2} \|y_{t+1} - x_{t+1}\|^2 \right)$$

$$= f(x_0) - f(x_T) + \frac{L}{2} \sum_{t=0}^{T-1} \|y_{t+1} - x_{t+1}\|^2$$

Putting it together,

$$\sum_{t=1}^{T} (f(x_t) - f(x^*)) \leq \frac{L}{2} \| x_0 - x^* \|^2$$

By definitions of x_{t+1} and y_{t+1}

$$\begin{aligned} \| y_{t+1} - x_{t+1} \| &\leq \| y_{t+1} - x_t \| \\ &= \eta \| \nabla f(x_t) \| \\ &= \frac{1}{L} \| \nabla f(x_t) \| \end{aligned}$$

Combining this with "sufficient decrease" (***).

$$f(x_{t+1}) \leq f(x_t)$$

make progress at every step.

$$f(x_T) - f(x^*) \leq \frac{1}{T} \sum_{t=1}^{T} (f(x_t) - f(x^*)) \leq \frac{L}{2T} \| x_0 - x^* \|^2$$

$\square$.

P15. Theorem 2.

(i) Note that from P11

$$g_t^T (x_t - x^*) \leqslant \frac{1}{2\eta} (\eta^2 \| g_t \|^2 + \| x_t - x^* \|^2 - \| x_{t+1} - x^* \|^2 - \| y_{t+1} - x_{t+1} \|^2) \quad (*)$$

By ~~strongly~~ strong convexity of f,

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

Combine $(*)$ and $(**)$

$$\| x_{t+1} - x^* \|^2 \leqslant 2\eta \, (f(x^*) - f(x_t)) + \eta^2 \| \nabla f(x_t) \|^2$$
$$- \| y_{t+1} - x_{t+1} \|^2 + (1 - \mu\eta) \| x_t - x^* \|^2 \qquad (\Delta)$$

By "sufficient decrease" on P8

$$f(x^*) - f(x_t) \leqslant f(x_{t+1}) - f(x_t)$$
$$\leqslant -\frac{1}{2L} \| \nabla f(x_t) \|^2 + \frac{L}{2} \| y_{t+1} - x_{t+1} \|^2 \qquad (\Delta\Delta)$$

Combining (Δ) and $(\Delta\Delta)$ $\quad \eta = \frac{1}{L}$

$$\| x_{t+1} - x^* \|^2 \leqslant \frac{2}{L} \left(-\frac{1}{2L} \| \nabla f(x_t) \|^2 + \frac{L}{2} \| y_{t+1} - x_{t+1} \|^2 \right)$$
$$+ \frac{1}{L^2} \| \nabla f(x_t) \|^2 - \| y_{t+1} - x_{t+1} \|^2 + (1 - \frac{\mu}{L}) \| x_t - x^* \|^2$$
$$= (1 - \frac{\mu}{L}) \| x_t - x^* \|^2$$

(ii) By smoothness of f,

$$f(x_T) - f(x^*) \leqslant \nabla f(x^*)^T (x_T - x^*) + \frac{L}{2} \| x^* - x_T \|^2$$
$$\leqslant \| \nabla f(x^*) \| \, \| x_T - x^* \| + \frac{L}{2} \| x^* - x_T \|^2 \quad \text{(Cauchy - Schwarz)}$$
$$\leqslant \underbrace{\| \nabla f(x^*) \| \, (1 - \frac{\mu}{L})^{\frac{T}{2}} \| x_0 - x^* \|}_{\text{dominate}} \qquad \text{By (i)}$$
$$+ \underbrace{\frac{L}{2} (1 - \frac{\mu}{L})^T \| x_0 - x^* \|^2}_{\text{same as the unconstrained case}}$$

P25

$$x^* = \operatorname{argmin} \{ \underbrace{g(x_t) + \nabla g(x_t)(x - x_t) + \frac{1}{2\eta} \|x - x_t\|^2}_{F(x)} \}$$

$$\Leftrightarrow \quad \nabla F_x(x^*) = 0$$

$$\Leftrightarrow \quad \nabla g(x_t) + \frac{1}{\eta}(x^* + x_t) = 0$$

$$\Leftrightarrow \quad x^* = x_t - \eta \nabla g(x_t)$$

P26

$$x_{t+1} = \underset{y}{\operatorname{argmin}} \{ \underset{\sim}{g(x_t)} + \nabla g(x_t)^T (y - x_t) + \frac{1}{2\eta} \|y - x_t\|^2 + h(y) \}$$

$$= \underset{y}{\operatorname{argmin}} \{ \nabla g(x_t)^T (y - x_t) + \frac{1}{2\eta} \|y - x_t\|^2 + \frac{\eta}{2} \|\nabla g(x_t)\|^2$$
$$+ h(y) \}$$

$$= \underset{y}{\operatorname{argmin}} \{ \frac{1}{2\eta} \| y - x_t + \eta \nabla g(x_t) \|^2 + h(y) \}$$

$$=: \operatorname{prox}_{h,\eta} (x_t - \eta \nabla g(x_t))$$

$$\operatorname{prox}_{h,\eta}(z) := \underset{y}{\operatorname{argmin}} \{ \frac{1}{2\eta} \|y - z\|^2 + h(y) \}$$

$\boxed{\text{P30 Theorem 3.}}$

Proof: $$x_{t+1} = \arg\min_y \{ \underbrace{g(x_t) + \nabla g(x_t)^T (y - x_t) + \frac{L}{2} \| y - x_t \|^2 + h(y)}_{\psi(y)} \}$$

$$= \arg\min_y \{ \psi(y) \}$$

$\psi(y)$ is strongly convex with L and $x_{t+1} = \arg\min_y \psi(y)$

Thus, $\psi(y) \geq \psi(x_{t+1}) + \frac{L}{2} \| y - x_{t+1} \|^2$

Since $\psi(y) \geq \psi(x_{t+1}) + \underbrace{\nabla \psi(x_{t+1})^T}_{=0} (y - x_{t+1}) - \frac{L}{2} \| y - x_{t+1} \|^2$

This is equivalent to

$$\nabla g(x_t)^T (y - x_t) + \frac{L}{2} \| y - x_t \|^2 + h(y)$$

$$\geq \nabla g(x_t)^T (x_{t+1} - x_t) + \frac{L}{2} \| x_{t+1} - x_t \|^2 + h(x_{t+1}) + \frac{L}{2} \| y - x_{t+1} \|^2$$

Rearranging terms and subtracting $h(x_t)$

$$\underbrace{\nabla g(x_t)^T (y - x_t)}_{\mathrm{II}} + \frac{L}{2} \| y - x_t \|^2 - \frac{L}{2} \| y - x_{t+1} \|^2 + h(y) - h(x_t)$$

$$\geq \underbrace{\nabla g(x_t)^T (x_{t+1} - x_t) + \frac{L}{2} \| x_{t+1} - x_t \|^2}_{\mathrm{I}} + h(x_{t+1}) - h(x_t)$$

By L-smoothness of g.

$$\mathrm{I} = \nabla g(x_t)^T (x_{t+1} - x_t) + \frac{L}{2} \| x_{t+1} - x_t \|^2 \geq g(x_{t+1}) - g(x_t)$$

By convexity of g,

$$\mathrm{II} = \nabla g(x_t)^T (y - x_t) \leq g(y) - g(x_t)$$

Putting this together

$$g(x_{t+1}) - g(x_t) + h(x_{t+1}) - h(x_t)$$

$$\leq g(y) - g(x_t) + \frac{L}{2} \| y - x_t \|^2 - \frac{L}{2} \| y - x_{t+1} \|^2 + h(y) - h(x_t)$$

that is

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

set $y = x^*$, sum up over from $t=0$ to $t = T-1$

$$\sum_{t=0}^{T-1}(f(x_{t+1}) - f(x_t)) \leqslant \sum_{t=0}^{T-1}(f(x^*) - f(x_t)) + \frac{L}{2}\|x^* - x_0\|^2 - \frac{L}{2}\|x^* - x_T\|^2$$

$$f(x_T) - f(x_0) \leqslant \sum_{t=0}^{T-1}(f(x^*) - f(x_t)) + \frac{L}{2}\|x^* - x_0\|^2 - \frac{L}{2}\|x^* - x_T\|^2$$

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leqslant f(x_0) - f(x_T) + \frac{L}{2}\|x^* - x_0\|^2$$

Recall the definition of x_{t+1} and ψ

$$\psi(x_{t+1}) \leqslant \psi(x_t) \qquad 0 \leqslant t \leqslant T$$

thus $f(x_{t+1}) \leqslant f(x_t)$

$$f(x_T) - f(x^*) \leqslant \frac{1}{T}\sum_{t=1}^{T}(f(x_t) - f(x^*)) \leqslant \frac{L}{2T}\|x^* - x_0\|^2$$

$\square$

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

$$f(x_T) - f(x^*) \leqslant \frac{LR^2}{2T} \leqslant \varepsilon \quad \Longleftrightarrow \quad T \geqslant \frac{LR^2}{2\varepsilon}$$

converge rate for proximal GD $\quad O(\frac{1}{\varepsilon})$

P37. $\text{Prox}_{h,\eta}(z) = \underset{\beta}{\arg\min} \left\{ \frac{1}{2\eta} \|\beta - z\|_2^2 + \lambda \sum_{i=1}^{d} |\beta_i| \right\} =: S_{\lambda,\eta}(z)$

Define $F(\beta) := \frac{1}{2\eta} \sum_{i=1}^{d} (\beta_i - z_i)^2 + \lambda \sum_{i=1}^{d} |\beta_i|$

$F_i(\beta) := \frac{1}{2\eta} (\beta_i - z_i)^2 + \lambda |\beta_i|$

Task: find minimizer β^* of $F(\beta)$

Case 1: $\beta_i > 0$

$F_i(\beta) = \frac{1}{2\eta} (\beta_i - z_i)^2 + \lambda \beta_i$

$\frac{dF_i(\beta)}{d\beta_i} = \frac{1}{\eta} (\beta_i - z_i) + \lambda$

setting it to 0 for optimality

$\frac{1}{\eta} (\beta_i - z_i) + \lambda = 0 \Rightarrow \beta_i = z_i - \lambda\eta$

If $z_i > \lambda\eta$, then $\beta_i^* = z_i - \lambda\eta$

Case 2: $\beta_i < 0$

$F_i(\beta) = \frac{1}{2\eta} (\beta_i - z_i)^2 - \lambda \beta_i$

$\frac{1}{\eta} (\beta_i - z_i) - \lambda = 0 \Rightarrow \beta_i = z_i + \lambda\eta$

If $z_i < -\lambda\eta$. then $\beta_i^* = z_i + \lambda\eta$

Case 3: $\beta_i = 0$

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

From first-order optimality

$0 \in \frac{1}{\eta} (\beta_i - z_i) + \lambda \cdot \partial |\beta_i|$

At $\beta_i = 0$

$0 \in \frac{1}{\eta} (-z_i) + \lambda \cdot [-1, 1]$

$$\Leftrightarrow \quad -\lambda \le \frac{z_i}{\eta} \le \lambda$$

$$\Leftrightarrow \quad -\eta\lambda \le z_i \le \eta\lambda$$

Thus. if $|z_i| \le \eta\lambda$, then $\beta_i^* = 0$

Combining all three cases

$$\beta_i^* = \begin{cases} z_i - \eta\lambda & z_i > \eta\lambda \\ 0 & |z_i| \le \eta\lambda \\ z_i + \eta\lambda & z_i < -\eta\lambda \end{cases} \qquad i = 1 \dots d$$

$\square$

P48.

(i) By definition of $D_\phi(x,y) = \phi(x) - \phi(y) - \nabla\phi(y)^T(x-y)$

$$D_\phi(x,y) + D_\phi(z,x) - D_\phi(z,y)$$

$$= \cancel{\phi(x)} - \cancel{\phi(y)} - \nabla\phi(y)^T(x-y)$$

$$+ \cancel{\phi(z)} - \cancel{\phi(x)} - \nabla\phi(x)^T(z-x)$$

$$- \cancel{\phi(z)} + \cancel{\phi(y)} + \nabla\phi(y)^T(z-y)$$

$$= -\nabla\phi(x)^T(z-x) - \nabla\phi(y)^T(x-y-z+y)$$

$$= (\nabla\phi(x) - \nabla\phi(y))^T(x-z)$$

(ii) By first-order optimality condition

f is convex and X: closed convex set on which

f is differentiable.

$$x^* \in \underset{x\in X}{\operatorname{argmin}}\, f(x) \quad \text{iff} \quad \nabla f(x^*)^T(x^*-y) \le 0 \quad \forall y \in X.$$

Recall $\Pi_X^\phi(y) = \underset{x\in X}{\operatorname{argmin}}\, D_\phi(x,y)$

$$\left[\nabla_x D_\phi(x,y)\Big|_{x=\Pi_X^\phi(y)}\right]^T(\Pi_X^\phi(y) - y) \le 0$$

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

$$(*) := \left(\nabla\phi(\Pi_X^\phi(y)) - \nabla\phi(y)\right)^T(\Pi_X^\phi(y) - y) \le 0$$

Moreover, By (i).

$$(\nabla\phi(x) - \nabla\phi(y))^T(x-z) = D_\phi(x,y) + D_\phi(z,x) - D_\phi(z,y).$$

set $x = \Pi_X^\phi(y)$ $\quad z = x$

$$(*) = D_\phi(\Pi_X^\phi(y), y) + D_\phi(x, \Pi_X^\phi(y)) - D_\phi(x,y) \le 0 \qquad \square$$

P47. Theorem 4.

By convexity of f

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

Recall that $\nabla\phi(y_{t+1}) = \nabla\phi(x_t) - \eta g_t$

$$g_t^T(x_t - x) = \frac{1}{\eta}(\nabla\phi(x_t) - \nabla\phi(y_{t+1}))$$

$$\Rightarrow f(x_t) - f(x) \leq \frac{1}{\eta}(\nabla\phi(x_t) - \nabla\phi(y_{t+1}))(x_t - x)$$

$$= \frac{1}{\eta}(D_\phi(x, x_t) + D_\phi(x_t, y_{t+1}) - D_\phi(x, y_{t+1})) \qquad \text{By (i)}$$

$$\leq \frac{1}{\eta}(D_\phi(x, x_t) + D_\phi(x_t, y_{t+1}) - D_\phi(x, x_{t+1}) - D_\phi(x_{t+1}, y_{t+1})) \qquad \text{By (ii)}$$

$$D_\phi(x_t, y_{t+1}) - D_\phi(x_{t+1}, y_{t+1}) \qquad (**)$$

$$= \phi(x_t) - \phi(x_{t+1}) - \nabla\phi(y_{t+1})^T(x_t - x_{t+1})$$

Recall that ϕ is ρ-strongly convex.

$$\phi(x_t) - \phi(x_{t+1}) - \nabla\phi(y_{t+1})^T(x_t - x_{t+1})$$

$$\leq (\nabla\phi(x_t) - \nabla\phi(x_{t+1}))^T(x_t - x_{t+1}) - \frac{\rho}{2}\|x_t - x_{t+1}\|^2$$

$$= \eta g_t^T(x_t - x_{t+1}) - \frac{\rho}{2}\|x_t - x_{t+1}\|^2$$

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

$$\left\{ az - bz^2 \leq \frac{a^2}{4b} \quad \forall z. \qquad F(z) = az - bz^2,\quad z^* = -\frac{a}{2(-b)} = \frac{a}{2b},\quad F(z^*) = \frac{a^2}{4b}. \right\}$$

$$\leq \frac{(\eta L)^2}{2\rho} \qquad (*)$$

Putting $(*)$ and $(**)$ together

$$\sum_{t=1}^{T} \left(f(x_t) - f(x) \right) \leq \frac{D_\phi(x, x_1)}{\eta} + \eta \frac{L^2 T}{2\rho}$$

Recall $x_1 \in \underset{x \in X}{\arg\min}\ \phi(x)$ $\quad \eta = \frac{2R}{L}\sqrt{\frac{\rho}{T}}$.

$$x := x^*$$

Thus,

$$\frac{1}{T} \sum_{t=1}^{T} \left(f(x_t) - f(x^*) \right) \leq \frac{2R}{L}\sqrt{\frac{\rho}{T}} \frac{L^2}{2\rho} = RL\sqrt{\frac{1}{\rho T}}$$

By Jensen's inequality

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

$$\leq RL\sqrt{\frac{1}{\rho T}} \qquad \square.$$