Lecture #13   Proximal gradient methods   E1 260

- Proximal operator
- Proximal gradient descent
- Convergence analysis

Ref:
- Proximal algorithms, N. Parikh and S. Boyd
- First-order methods, A. Beck

Recall gradient descent method:

$$\underline{x}_{t+1} = \underline{x}_t - \eta_t \nabla f(\underline{x}_t)$$

$\equiv$

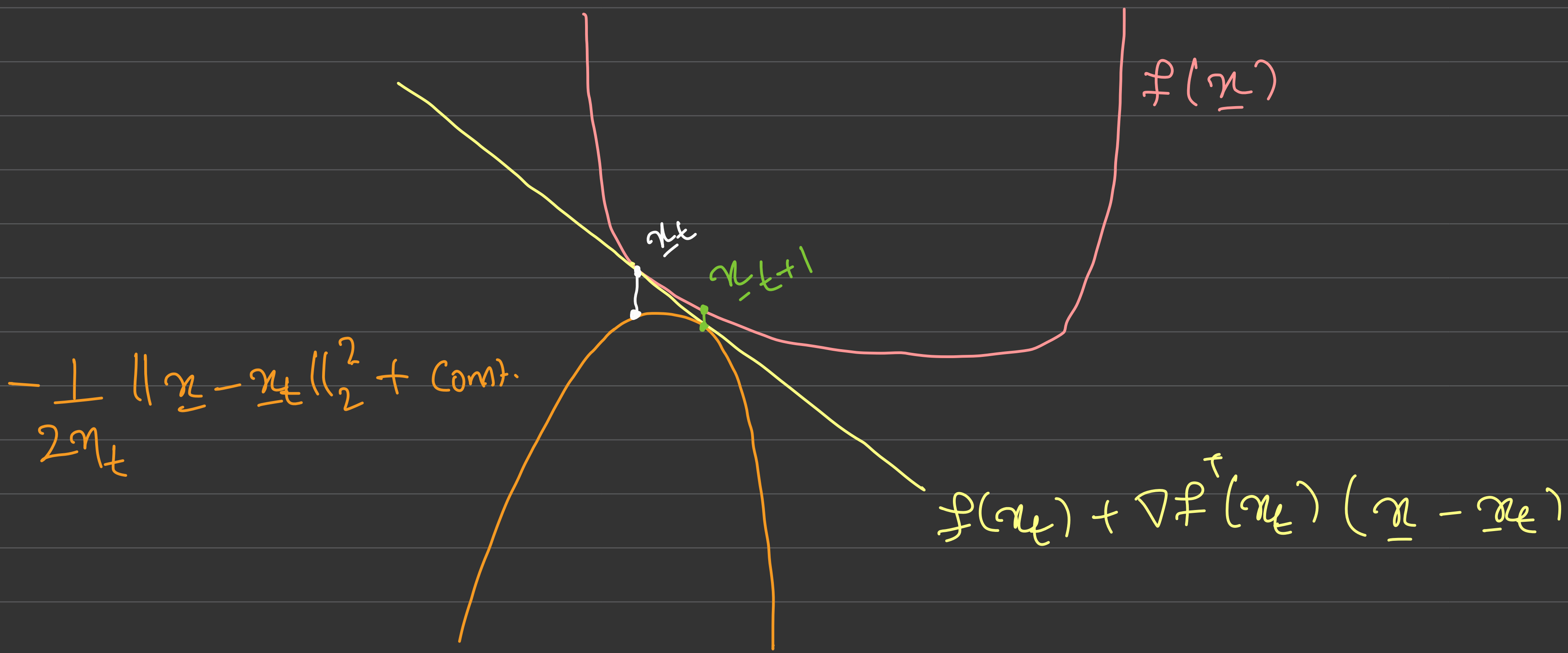
$$\underline{x}_{t+1} = \arg\min_{\underline{x}} \Big[\underbrace{f(\underline{x}_t) + \nabla f^T(\underline{x}_t)(\underline{x} - \underline{x}_t)}_{\text{affine approximation}} + \underbrace{\frac{1}{2\eta_t} \|\underline{x} - \underline{x}_t\|_2^2}_{\text{Proximal term}} \Big]$$

$$= \arg\min_{\underline{x}} \Big[\frac{1}{2\eta_t} \| x - (x_t - \eta_t \nabla f(x_t))\|_2^2 \Big]$$

optimality condition:

$$\nabla f(\underline{x}_t) + \frac{1}{2\eta_t} 2(x - x_t) = 0$$

$\Rightarrow$ x_{t+1} is a point for which the slope of $f(\underline{x}_t) + \nabla f^T(\underline{x}_t)(\underline{x} - \underline{x}_t)$ and $-\frac{1}{2\eta_t}\|\underline{x} - \underline{x}_t\|_2^2$ are the <u>same</u>.



Recall projected gradient descent:

minimize $f(\underline{x})$ subject to $\underline{x} \in C$

$$\underline{x}_{t+1} = P_C\left(\underline{x}_t - \eta_t \nabla f(\underline{x}_t)\right)$$

Define

$$\mathbb{1}_C(\underline{x}) = \begin{cases} 0 & ; \text{ if } \underline{x} \in C \\ \infty & , \text{ otherwise} \end{cases}$$

Then,

$$\underline{x}_{t+1} = \arg\min_{\underline{x}} \Big[f(\underline{x}_t) + \nabla f^T(\underline{x}_t)(\underline{x} - \underline{x}_t) + \frac{1}{2\eta_t} \|\underline{x} - \underline{x}_t\|_2^2 + \eta_t \mathbb{1}_C(\underline{x}) \Big]$$

$$= \arg\min_{\underline{x}} \Big[\frac{1}{2} \|\underline{x} - (\underline{x}_t - \eta_t \nabla f(\underline{x}_t))\|_2^2 + \eta_t \mathbb{1}_C(\underline{x}) \Big]$$

Proximal operator :

$$\text{Prox}_h(\underline{x}) = \arg\min_{\underline{z}} \left[\frac{1}{2} \| \underline{z} - \underline{x} \|^2 + h(\underline{z}) \right]$$

With this, the projected gradient descent becomes

$$\underline{x}_{t+1} = \text{Prox}_{\eta \mathbb{1}_C} \left(\underline{x}_t - \eta_t \nabla f(\underline{x}_t) \right)$$

$$\text{Prox}_{\eta h}(\underline{x}) = \arg\min_{\underline{z}} \left[\frac{1}{2\eta} \| \underline{z} - \underline{x} \|^2 + h(\underline{z}) \right]$$

- We can generalize $\underline{h}(\cdot)$ and handle an important class of functions, namely, composite models

$$F(\underline{x}) = f(\underline{x}) + h(\underline{x})$$

$f(x)$: (nice) Convex and smooth

$h(\underline{x})$: (simple) Convex (may not be differentiable)

Proximal gradient method :

gradient descent: $y_{t+1} = \underline{x}_t - \eta_t \nabla f(\underline{x}_t)$

Proximal minimization : $\underline{x}_{t+1} = \text{Prox}_{\eta_t h}(\underline{y}_{t+1})$

$$\Rightarrow \underline{x}_{t+1} = \underline{x}_t - \eta_t G_h(\underline{x}_t)$$

$$\min_{x \in \mathbb{R}^d} \quad F(\underline{x}) = f(\underline{x}) + h(\underline{x})$$

$$G_h(\underline{x}) = \frac{1}{\eta_t}\left[\underline{x} - \text{Prox}_{\eta_t h}\left(\underline{x} - \eta_t \nabla f(\underline{x}_t)\right)\right]$$

is the so-called generalized gradient of f.

- This covers many commonly used regularizers in ML and "$\text{Prox}_h(\underline{x})$" is well-defined for nonsmooth convex functions

Examples:

- L_1 regularized minimization:

$$\underset{\underline{x}}{\text{minimize}} \quad f(\underline{x}) + \|\underline{x}\|_1$$

- nuclear norm regularized minimization

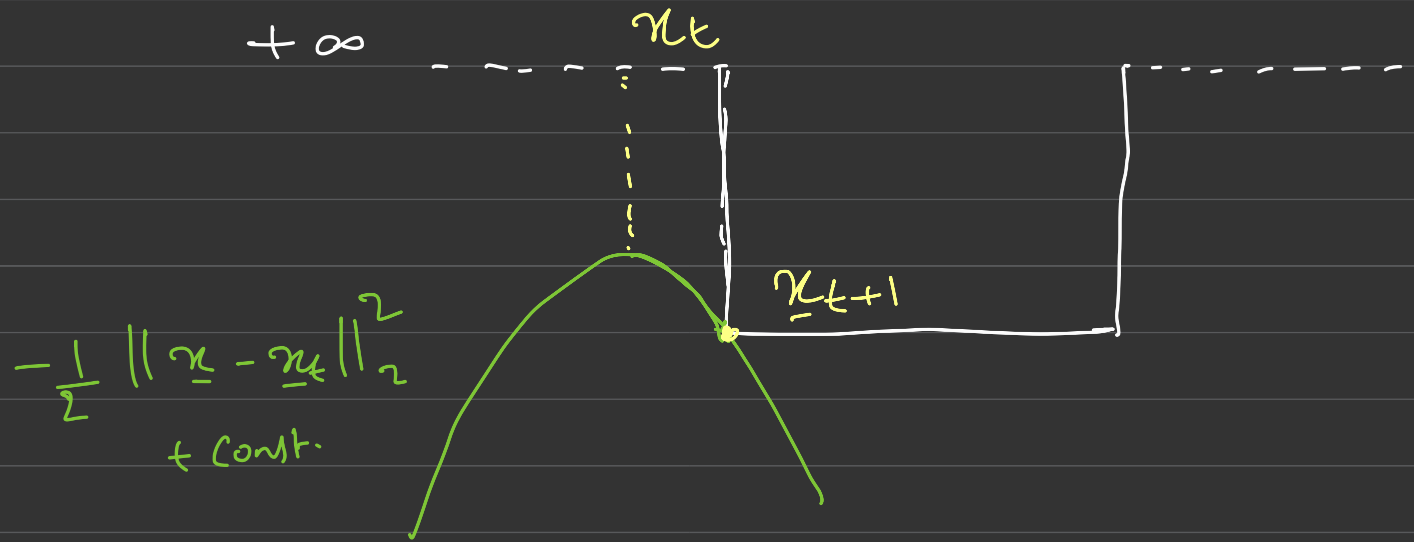
$$\underset{X}{\text{minimize}} \quad f(X) + \|X\|_*$$

Examples

1. Indicator function $\quad h = \mathbb{1}_C = \begin{cases} 0, & \text{if } \underline{x} \in C \\ \infty, & \text{o.w.} \end{cases}$

$$\text{Prox}_h(\underline{x}) = \arg\min_{z \in C} \|\underline{x} - z\|_2^2 = P_C(x)$$

(Euclidean projection)



② ℓ_1 - norm $\quad h(\underline{x}) = \lambda \|\underline{x}\|_1 = \sum_{i=1}^{d} |x_i|$

$$\operatorname{Prox}_{\lambda h}(\underline{x}) = \arg\min_{\underline{z}} \left[\frac{1}{2} \|\underline{z} - \underline{x}\|^2 + \lambda \|\underline{z}\|_1 \right]$$

$$= \arg\min_{\underline{z}} \left[\frac{1}{2} \sum_{i=1}^{d} \left((z_i - x_i)^2 + \lambda |z_i| \right) \right]$$

$$\arg\min_{z_i} \frac{1}{2} (z_i - x_i)^2 + \lambda |z_i| \; ; \text{ for } i = 1, \dots, d$$

① when z_i is positive :

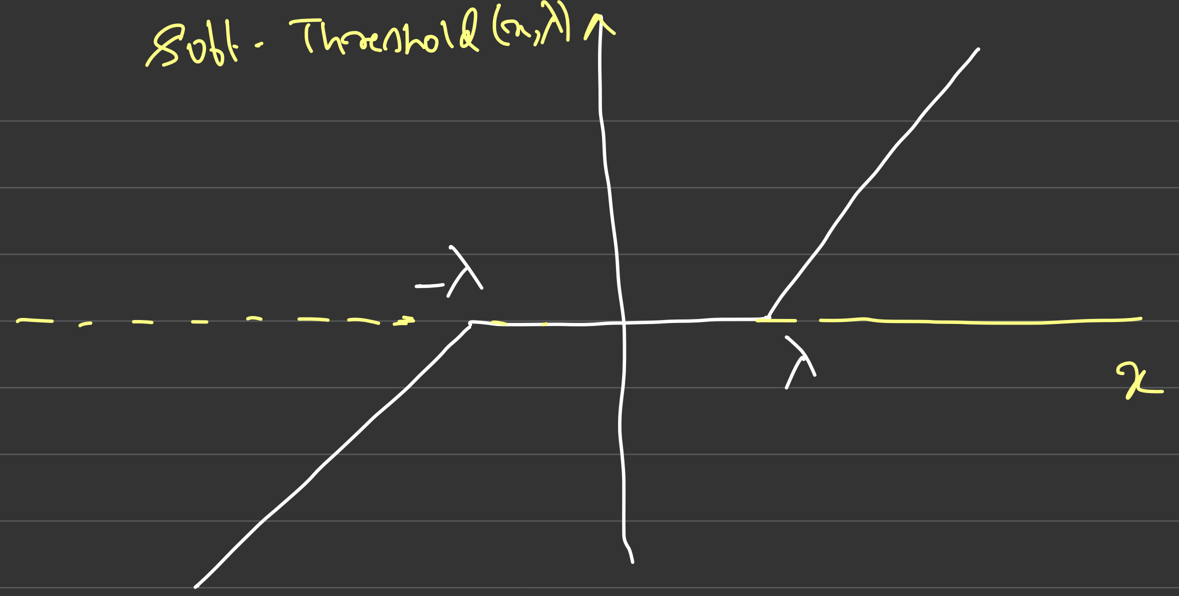
$$\frac{1}{2} (z_i - x_i)^2 + \lambda z_i$$

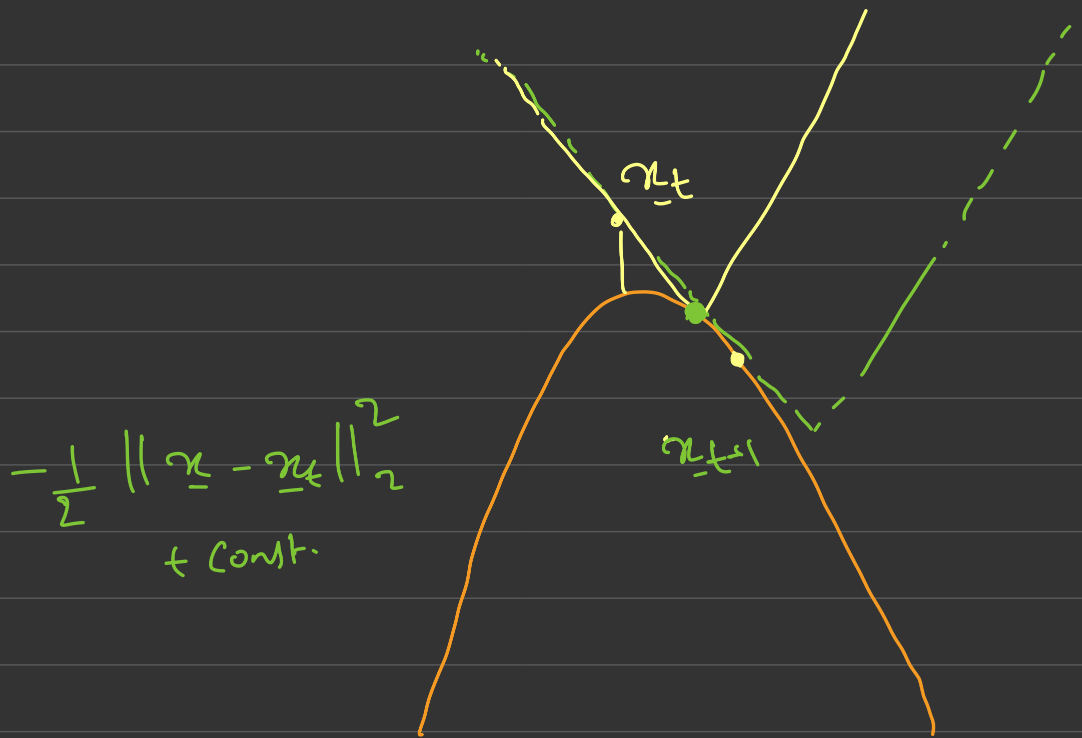
1st order condition : $\quad z_i = x_i - \lambda \quad$ so $\quad x_i > \lambda$

Similarly , when z_i is negative , gradient vanishes at $\quad z_i = x_i + \lambda \quad$ and $\quad x_i < -\lambda$;

$$\Rightarrow \left[\operatorname{Prox}_{\lambda h}(\underline{x}) \right]_i = \text{Soft-Threshold}(x_i, \lambda)$$

$$\text{Soft-Threshold}(x, \lambda) = \begin{cases} x - \lambda & ; \text{ if } x > \lambda \\ x + \lambda & , \text{ if } x < \lambda \\ 0 & ; \text{ if } -\lambda \le x \le \lambda \end{cases}$$


Soft - Threshold (x, λ)
$-\lambda$
λ
x


x_t
x_{t+1}
$-\frac{1}{2} \| \underline{x} - \underline{x}_t \|_2^2$
$+$ const.

Monotonicity of the cost:

Claim: Suppose f is convex and L-smooth and $\eta_t = \frac{1}{L}$. Then

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

2. $\| \underline{x}_{t+1} - \underline{x}^* \|_2^2 \leq \| \underline{x}_t - \underline{x}^* \|_2^2$

- For subgradient methods, objective value might not be monotonic.

If the following is true:

$$\boxed{\begin{aligned} F(\underline{x}_{t+1}) - F(\underline{x}) &\le \frac{L}{2}\left[\|\underline{x} - \underline{x}_t\|_2^2 - \|\underline{x} - \underline{x}_{t+1}\|_2^2\right] \\ &\quad - \psi(\underline{x}, \underline{x}_t) \\ \text{where} \quad \psi(\underline{x}, \underline{x}_t) &= f(\underline{x}) - f(\underline{x}_t) - \nabla f^T(\underline{x}_t)(\underline{x} - \underline{x}_t) \end{aligned}}$$

≥ 0 [by convexity of f]

Then, taking $x = x_t$ we have. ① as
$F(x_{t+1}) - F(x_t) \le 0$

and taking $x = x^*$ we have ②

as $\underbrace{F(x_{t+1}) - F(x^*) + \psi(\underline{x}^*, x_t)}_{\geq 0} \le \text{RHS}$

Proof:

Let $\phi(\underline{z}) = f(\underline{x}_t) + \nabla f^T(\underline{x}_t)(\underline{z} - \underline{x}_t) + \frac{L}{2}\|\underline{z} - \underline{x}_t\|_2^2$
$+ h(\underline{z})$

See that

$$\underline{x}_{t+1} = \arg\min_{\underline{z}} \phi(\underline{z})$$

- Since $\phi(z)$ is L-strongly convex

$$\phi(\underline{x}) \geq \phi(x_{t+1}) + \frac{L}{2}\|\underline{x} - \underline{x}_{t+1}\|_2^2$$

- $$\phi(\underline{x}_{t+1}) = f(x_t) + \nabla f^T(\underline{x}_t)(\underline{x}_{t+1} - \underline{x}_t) + \frac{L}{2}\|\underline{x}_{t+1} - x_t\|_2^2 + h(\underline{x}_{t+1})$$

upper bound on $f(\underline{x}_{t+1})$

$$\geq f(\underline{x}_{t+1}) + h(\underline{x}_{t+1})$$ [From smoothness]

$$\phi(\underline{x}_{t+1}) \geq F(\underline{x}_{t+1})$$

$$\Rightarrow \quad \phi(\underline{x}) \geq F(\underline{x}_{t+1}) + \frac{L}{2}\|\underline{x} - \underline{x}_{t+1}\|_2^2$$

So

$$f(\underline{x}_t) + \nabla f^T(\underline{x}_t)(\underline{x} - \underline{x}_t) + h(\underline{x}) + \frac{L}{2}\|\underline{x} - \underline{x}_t\|_2^2$$

$$\geq F(\underline{x}_{t+1}) + \frac{L}{2}\|\underline{x} - \underline{x}_{t+1}\|_2^2$$

$$f(\underline{x}) - \psi(\underline{x}, \underline{x}_t) + h(\underline{x}) + \frac{L}{2}\|\underline{x} - \underline{x}_t\|_2^2$$

$$\geq F(\underline{x}_{t+1}) + \frac{L}{2}\|\underline{x} - \underline{x}_{t+1}\|_2^2$$

$$\Rightarrow \quad F(\underline{x}_{t+1}) - F(\underline{x}) \leq \frac{L}{2}\left[\|\underline{x} - \underline{x}_t\|_2^2 - \|\underline{x} - \underline{x}_{t+1}\|^2\right] - \psi(\underline{x}, \underline{x}_t) \quad \blacksquare$$

Convergence: $F(x) = f(x) + h(x)$

Suppose f is convex and L-smooth and $\eta_t = \frac{1}{L}$

Then proximal gradient descent satisfies

$$F(x_t) - F^* \leq \frac{L}{2T} \| x_0 - x^* \|_2^2$$

projected subgradient : $O(1/\varepsilon^2)$, whereas we now have $O(1/\varepsilon)$

Proof: $F(x_{t+1}) - F(x^*) \leq \frac{L}{2} \left[\| x^* - x_t \|_2^2 - \| x^* - x_{t+1} \|^2 \right] - \underbrace{\varphi(x, x_t)}_{\geq 0}$

$$\leq \frac{L}{2}\left[\|\underline{x}^* - \underline{x}_t\|_2^2 - \|\underline{x}^* - \underline{x}_{t+1}\|_2^2\right]$$

Sum over $t = 0$ to $T-1$,

$$\sum_{t=0}^{T-1}\left(F(x_{t+1}) - F(\underline{x}^*)\right) \leq \frac{L}{2}\|\underline{x}_0 - \underline{x}^*\|_2^2 - \frac{L}{2}\|\underline{x}_T - x^*\|_2^2$$

Since last iterate is the best,

$$\Rightarrow F(x_T) - F(\underline{x}^*) \leq \frac{L}{2T}\|\underline{x}_0 - \underline{x}^*\|_2^2 \qquad \blacksquare$$

μ- strongly convex and L-smooth functions

With $\eta_t = \frac{1}{L}$, we have

$$\| \underline{x}_T - \underline{x}^* \|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^T \| \underline{x}_0 - \underline{x}^* \|_2^2$$

- Iteration complexity : $O(\log(1/\varepsilon))$

Proof:

$$F(\underline{x}_{t+1}) - F(\underline{x}^*) \leq \frac{L}{2}\left[\| \underline{x}^* - \underline{x}_t \|_2^2 - \| \underline{x}^* - \underline{x}_{t+1} \|^2\right]$$
$$- \varphi(\underline{x}^*, \underline{x}_t)$$

Since f is μ - strongly convex

$$\varphi(\underline{x}^*, \underline{x}_t) = f(\underline{x}^*) - f(\underline{x}_t) - \nabla f^T(\underline{x}_t)(\underline{x}^* - \underline{x}_t)$$
$$\geq \frac{\mu}{2} \| \underline{x}^* - \underline{x}_t \|_2^2$$

$$\Rightarrow \quad \underbrace{F(\underline{x}_{t+1}) - F(\underline{x}^*)}_{\geq 0} \leq \frac{L-\mu}{2} \|\underline{x}_t - \underline{x}^*\|_2^2 - \frac{L}{2} \|\underline{x}_{t+1} - \underline{x}^*\|_2^2$$

$$\Rightarrow \quad \|\underline{x}_{t+1} - \underline{x}^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right) \|\underline{x}_t - \underline{x}^*\|_2^2$$

$\blacksquare$