Lecture 12: Subgradient methods

E1 260

- Subgradients
- Subgradient methods
- Convergence analysis
  - Lipschitz convex functions
  - Strong convexity

Gradient descent method:

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

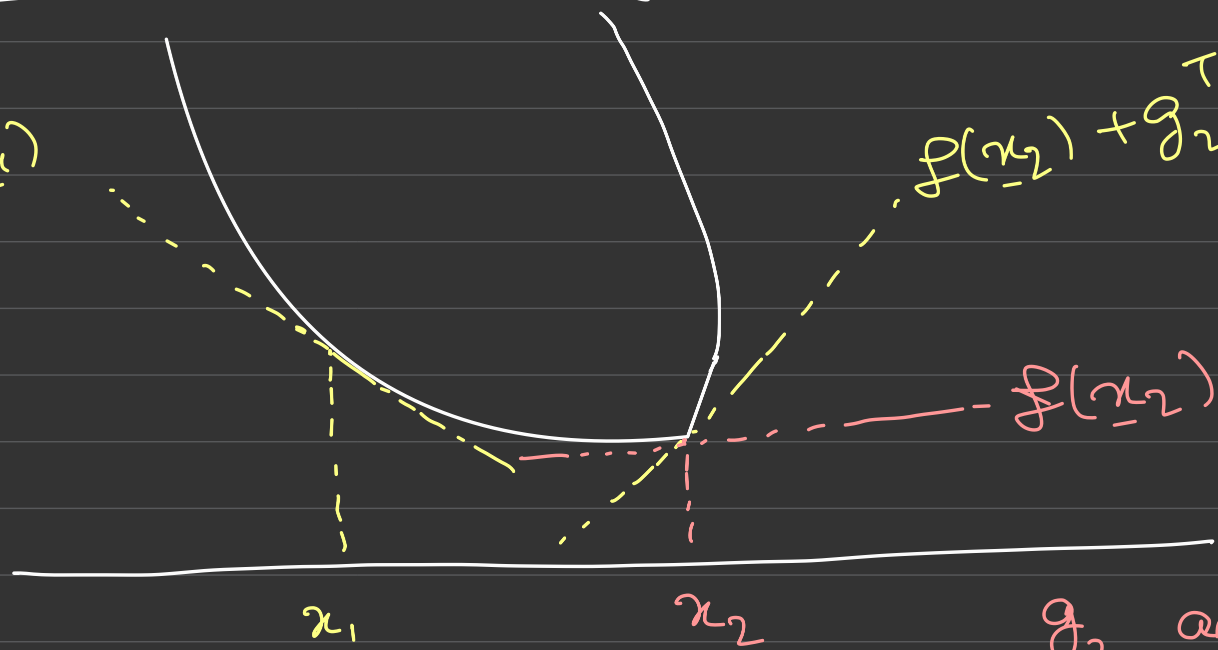
Differentiability of the objective function f is essential

How about

$$\begin{aligned} &\text{minimize} && \|\underline{x}\|_1 \\ &\text{s.t.} && \|A\underline{x} - \underline{b}\| \leq \varepsilon \end{aligned}$$

$$\begin{aligned} &\text{minimize} && \|X\|_* \\ &\text{s. to} && \|P_\Omega(X - Y)\|_F^2 \leq \varepsilon \end{aligned}$$

Subgradient :



$f(\underline{x}_1) + g_1^T(\underline{x} - \underline{x}_1)$

$f(\underline{x}_2) + g_2^T(\underline{x} - \underline{x}_2)$

$f(\underline{x}_2) + g_3^T(\underline{x} - \underline{x}_3)$

g_1 is a subgradient at $\underline{x}_1$

g_2 and g_3 are subgradients at $\underline{x}_2$

- $\underline{g}$ is a **subgradient** of f at $\underline{x}$ if

$$\boxed{f(\underline{y}) \geq \underbrace{f(\underline{x}) + \underline{g}^T(\underline{y} - x)}_{\text{a global linear underestimate of } f}, \quad \forall\ y}$$

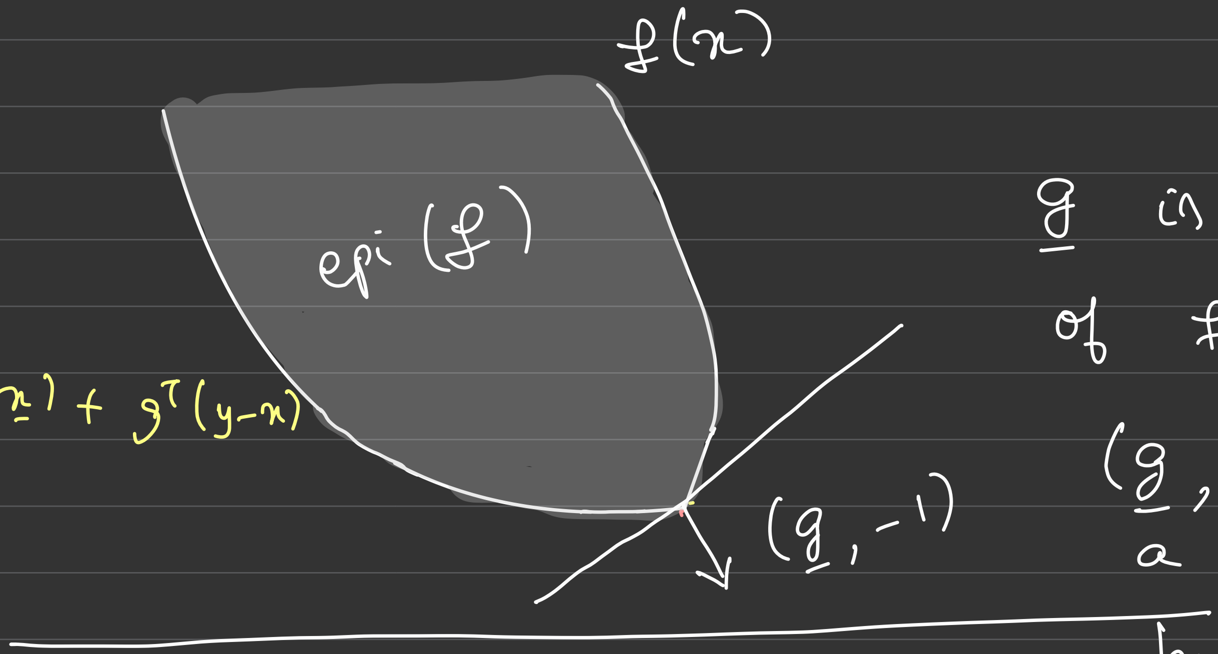
- Convexity is equivalent to the existence of subgradients everywhere

- if a function is convex and differentiable, $\nabla f(x)$ is a subgradient of a f at x

- the set of subgradients of f at x is called the subdifferential of f at $\underline{x}$

$$\partial f(\underline{x})$$

$t \geq f(y) \geq f(x) + g^T(y-x)$



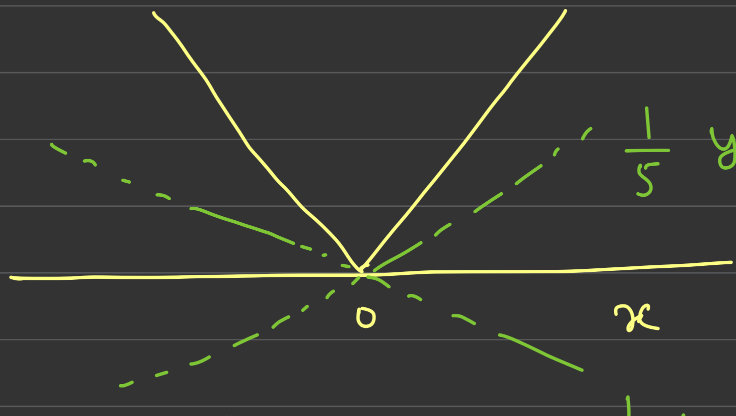
g is a subgradient of f at x iff $(g, -1)$ defines a supporting hyperplane of epi(f) at $(x, f(x))$

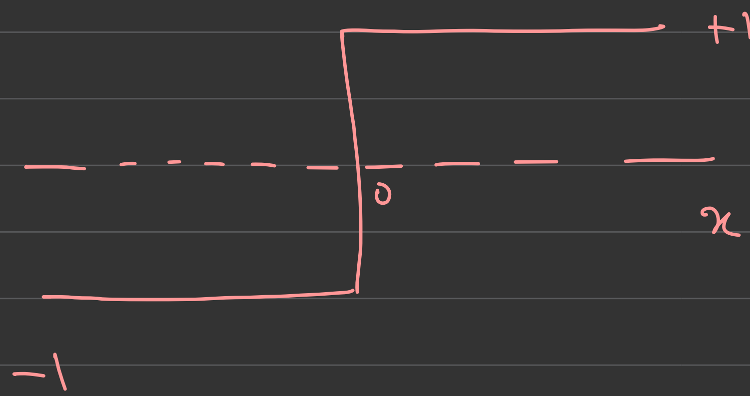
$$(y,t) \in \mathrm{epi}(f) \Rightarrow \begin{bmatrix} g \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0$$

Examples:

$$f(x) = |x|$$

$$\partial f(x) = \begin{cases} -1 & , \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$





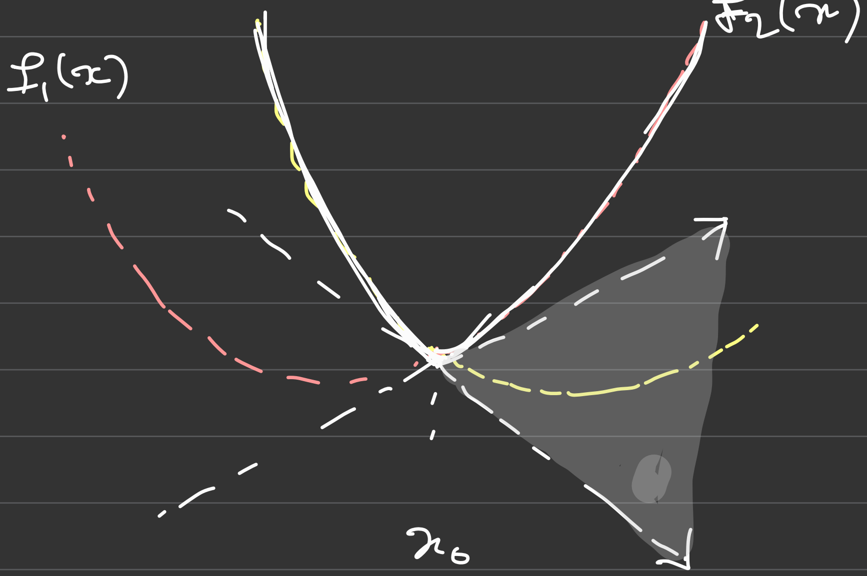
$f(y) \geq g y$

$g \in [-1, 1]$ at 0

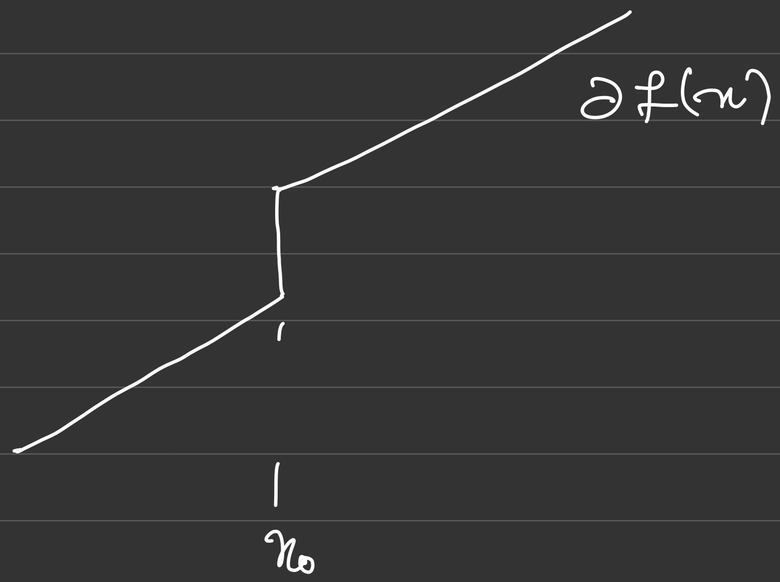
$$f(\underline{x}) = \|\underline{x}\|_1 = \sum_{i=1}^{d} |x_i| = \sum_{i=1}^{d} f_i(x)$$

$$\partial f(\underline{x}) = \begin{cases} \operatorname{sgn}(x_i)\, \underline{e}_i & \text{if } x_i \neq 0 \\ [-1, 1]\, \underline{e}_i & \text{if } x_i = 0 \end{cases}$$

Example: $f(x) = \max\{f_1(x), f_2(x)\}$



$$\partial f(x) = \begin{cases} \nabla f_1(x_0), & \text{if } f_1(x_0) > f_2(x_0) \\ [\nabla f_1(x_0), \nabla f_2(x_0)], & \text{if } f_1(x_0) = f_2(x_0) \\ \nabla f_2(x_0), & \text{if } f_2(x_0) > f_1(x_0) \end{cases}$$



Optimality condition.

- Recall for convex and differentiable f

$$f(\underline{x}^*) = \inf_x f(x) \iff 0 = \nabla f(\underline{x}^*)$$

- For differentiable f, $\nabla f(x) = 0$ we can only say $\underline{x}$ is a critical point.

- Suppose $f : \mathrm{dom}(f) \to \mathbb{R}$ and $\underline{x} \in \mathrm{dom}(f)$.

  if $\underline{0} \in \partial f(\underline{x})$, then $\underline{x}$ is "a global minimum"

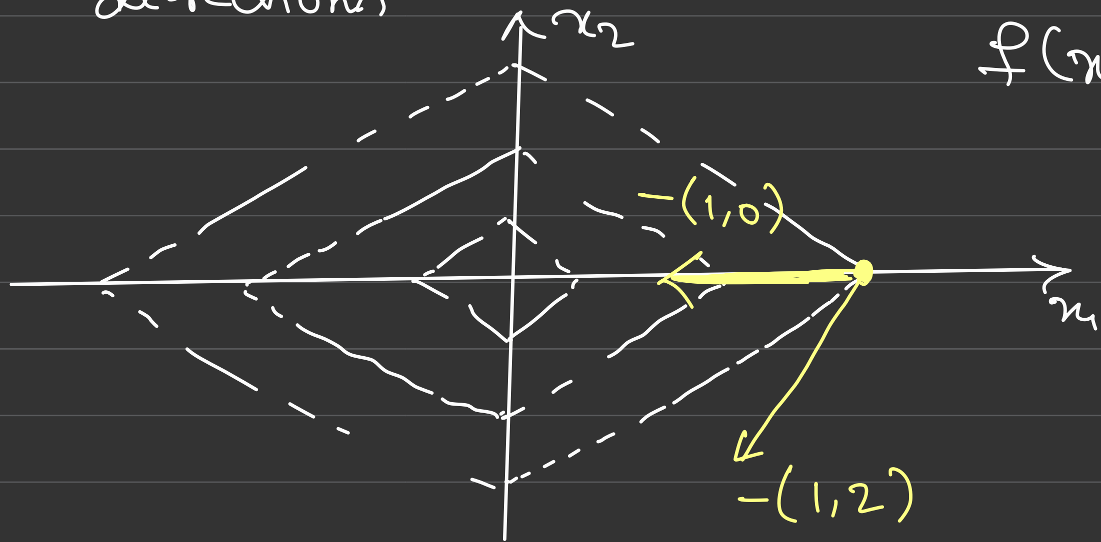
If $\underline{g} = \underline{0} \in \partial f(x)$, $f(y) \geq f(x) + g^T(y - x) = f(\underline{x})$

$$\forall y \in \mathrm{dom}(f)$$

Descent direction:

- Negative subgradients are not necessarily descent directions

$$f(x) = |x_1| + 2|x_2|$$



at $\underline{x} = (1,0)$

- $\underline{g}_1 = (1,0) \in \partial f(x)$ and $-g_1$ is a descent direction
- $g_2 = (1,2) \in \partial f(x)$, but $-g_2$ is not

The algorithm :

$$\Rightarrow \quad \underline{x}_{t+1} = \underline{x}_t - \eta_t \underline{g}_t \quad ; \quad g_t \in \partial f(x)$$

$$\Rightarrow \quad \underline{x}_{t+1} = P_C \left(\underline{x}_t - \eta_t g_t \right) \quad ; \quad \underline{g}_t \in \partial f(x_t)$$

- Since $f(x_{t+1})$ is not necessarily monotone, we also keep track of

$$f_t^{best} := \min_{1 \le i \le t} f(\underline{x}_i)$$

- Define $$f^{opt} = \min_{\underline{x}} f(\underline{x})$$

Fundamental inequality for Projected Subgradient methods:

$$\|\underline{x}_{t+1} - \underline{x}^*\|_2^2 \leq \underbrace{\|\underline{x}_t - \underline{x}^*\|_2^2 - 2\eta_t\left(f(x_t) - f^{opt}\right) + \eta_t^2 \|g_t\|_2^2}_{\text{majorizing function}}$$

We wish to optimize $\|\underline{x}_{t+1} - \underline{x}^*\|_2^2$, but without access to $\underline{x}^*$ we optimize by finding another function that majorizes it.

$$\| \underline{x}_{t+1} - \underline{x}^* \|_2^2 = \| P_C (\underline{x}_t - \eta_t g_t) - P_C (x^*) \|_2^2$$

From non expansiveness

$$\leq \| \underline{x}_t - \eta_t g_t - \underline{x}^* \|_2^2$$

$$= \| \underline{x}_t - \underline{x}^* \|_2^2 - 2\eta_t \underline{g}_t^T (\underline{x}_t - \underline{x}^*) + \eta_t^2 \| \underline{g}_t \|_2^2$$

$$\leq \| \underline{x}_t - \underline{x}^* \|_2^2 - 2\eta_t [f(\underline{x}_t) - f(\underline{x}^*)] + \eta_t^2 \| g_t \|_2^2$$

as

$$f(\underline{x}^*) - f(\underline{x}_t) \geq \underline{g}_t (\underline{x}^* - \underline{x}_t)$$

Polyak step size :

$$\eta_t = \frac{f(\underline{x}_t) - f^{opt}}{\|\underline{g}_t\|^2}$$

- we get an error reduction

$$\|\underline{x}_{t+1} - \underline{x}^*\|_2^2 \leq \|\underline{x}_t - x^*\|_2^2 - \frac{\left(f(\underline{x}_t) - f(x^*)\right)^2}{\|g_t\|_2^2}$$

- But needs f^{opt} to be known

Suppose f is convex and B Lipschitz continuous

Then

① $\| g \| \leq B \quad \forall \; \underline{g} \in \partial f(\underline{x})$

② $| f(x) - f(y) | \leq B \, \| \underline{x} - \underline{y} \| \quad \forall x, y \in \operatorname{dom} f$

<u>Claim:</u> The projected subgradient descent with Polyak's step size rule satisfies

$$f_T^{best} - f^{opt} \leq \frac{B}{\sqrt{T}} \| \underline{x}_0 - \underline{x}^* \|_2$$

- Sublinear convergence rate of $O\left(\frac{1}{\sqrt{T}}\right)$

We had for Polyak's step size rule:

$$\|\underline{x}_{t+1} - \underline{x}_t^*\|_2^2 \leq \|\underline{x}_t - x^*\|_2^2 - \frac{(f(\underline{x}_t) - f(x^*))^2}{\|g_t\|_2^2}$$

$$\Rightarrow \left(f(\underline{x}_t) - f(\underline{x}^*)\right)^2 \leq \left[\|\underline{x}_t - \underline{x}^*\|_2^2 - \|\underline{x}_{t+1} - \underline{x}^*\|_2^2\right] B^2$$

Apply recursively and sum over iterations

$t = 0$ to $T-1$

$$\sum_{t=0}^{T-1} (f(x_t) - f(\underline{x}^*))^2 \leq B^2\left[\|\underline{x}_0 - x^*\|_2^2 - \|\underline{x}_T - \underline{x}^*\|_2^2\right]$$

$$\Rightarrow T\left(f_{T-1}^{best} - f(\underline{x}^*)\right)^2 \leq B^2 \|\underline{x}_0 - \underline{x}^*\|_2^2$$

$$\Rightarrow f_t^{best} - f^{opt} \leq \frac{B}{\sqrt{T}} \|\underline{x}_0 - \underline{x}^*\|$$

How about other step sizes (diminishing ?)

Claim:

Suppose f is convex and B Lipschitz continuous.

The projected subgradient method

$$f_T^{best} - f^{opt} \le \frac{\| \underline{x}_0 - \underline{x}^* \| + B^2 \sum_{t=0}^{T-1} \eta_t^2}{2 \sum_{t=0}^{T-1} \eta_t}$$

$$\|\underline{x}_{t+1} - \underline{x}^*\|_2^2 \le \|\underline{x}_t - \underline{x}^*\|_2^2 - 2\eta_t [f(\underline{x}_t) - f(\underline{x}^*)] + \eta_t^2 \|g_t\|_2^2$$

$$\|\underline{x}_T - \underline{x}^*\|_2^2 \le \|\underline{x}_0 - \underline{x}^*\|_2^2 - 2\sum_{t=0}^{T-1} \eta_t (f(\underline{x}_t) - f^{opt}) + \sum_{t=0}^{T-1} \eta_t^2 \|\underline{g}_t\|_2^2$$

$\Rightarrow$

$$2\sum_{t=0}^{T-1} \eta_t (f(\underline{x}_t) - f^{opt}) \le \|\underline{x}_0 - \underline{x}^*\|_2^2 - \|\underline{x}_T - \underline{x}^*\|_2^2 + B^2 \sum_{t=0}^{T-1} \eta_t^2$$

But

$$2\sum_{t=0}^{T-1} \eta_t (f_t^{best} - f^{opt}) \le 2\sum_{t=0}^{T-1} \eta_t (f(\underline{x}_t) - f^{opt})$$

or

$$f_T^{best} - f^{opt} \le \sum_{t=0}^{T-1} \eta_t (f(x_t) - f^*) \Big/ \sum_{t=0}^{T-1} \eta_t$$

$$f_T^{best} - f^{opt} \leq \frac{\|\underline{x}_0 - \underline{x}^*\|_2^2 + B^2 \sum_{t=0}^{T-1} \eta_t^2}{2 \sum_{t=0}^{T-1} \eta_t}$$

- **Constant step-size:** $\eta_t = \eta$

$$\lim_{T \to \infty} f_T^{best} \leq \frac{B^2 \eta}{2}$$

may not converge to optimal points

- **Diminishing step size:** $\sum_{t=0}^{T-1} \eta_t^2 < \infty$ and $\sum_{t=0}^{T-1} \eta_t \to \infty$

$$\lim_{T \to \infty} f_T^{best} = 0$$

e.g. $\frac{a}{b+t}$

$a > 0$; $b \geq 0$

Converges to optimal points.

Strongly convex : $O(1/\varepsilon)$

better than $O(1/\varepsilon^2)$, worse than $O(\log(1/\varepsilon))$

Claim:

$f: \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and $\underline{x}^*$ be the unique minimizer of f. With $\eta_t = \frac{2}{\mu(t+1)}$.

Then Subgradient method, yields

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

where $B = \max_{1 \leq t \leq T} \|g_t\|$.

Recall :

$$\underline{g}_t^T(\underline{x}_t - \underline{x}^*) = \frac{\eta_t}{2}\underbrace{\|g_t\|^2}_{\le B^2} + \frac{1}{2\eta_t}\Big[\|\underline{x}_t - \underline{x}^*\|^2 - \|\underline{x}_{t+1} - \underline{x}^*\|^2\Big]$$

use the quadratic lower bound :

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

$$\Rightarrow f(\underline{x}_t) - f(\underline{x}^*) \le \frac{\eta_t}{2}B^2 + \frac{(\eta_t^{-1} - \mu)}{2}\|\underline{x}_t - \underline{x}^*\|^2 - \frac{\eta_t^{-1}}{2}\|\underline{x}_{t+1} - \underline{x}^*\|^2$$

- Unlike Gradient descent with fixed step size we cannot telescope anymore when we sum over iterations

- To get a telescopic sum

$$\eta_t^{-1} = \eta_{t+1}^{-1} - \mu$$

One choice of η_t that satisfies this is:

$$\eta_t^{-1} = \mu(t+1)$$

Actually our choice

$$\eta_t^{-1} = \mu(t+1)/2$$ does not

Homework 2: Check what happens if we proceed with $\eta_t^{-1} = \mu(t+1)$

$$t\left(f(\underline{x}_t)-f(\underline{x}^*)\right) \leq \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\Big[t(t-1)\,\|\underline{x}_t-\underline{x}^*\|^2$$

$$-(t+1)t\,\|\underline{x}_{t+1}-\underline{x}^*\|^2\Big]$$

$$\leq \frac{B^2}{\mu} + \frac{\mu}{4}\Big[t(t-1)\,\|\underline{x}_t-\underline{x}^*\|^2$$

$$-(t+1)t\,\|\underline{x}_{t+1}-\underline{x}^*\|^2\Big]$$

Summing over $t = 1$ to T :

$$\sum_{t=1}^{T} t\left(f(\underline{x}_t)-f(\underline{x}^*)\right) \leq T\frac{B^2}{\mu} + \frac{\mu}{4}\Big[0 -$$

$$T(T+1)\,\|\underline{x}_T-\underline{x}^*\|^2\Big] \leq \frac{TB^2}{\mu}$$

Since $\frac{2}{T(T+1)} \sum_{t=1}^{T} t = 1$

and $f(\cdot)$ is convex (Jensen's inequality)

$$f\left(\frac{2}{T(T+1)} \sum_{t=1}^{T} t\, \underline{x}_t\right) - f(\underline{x}^*) \leq \frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot \left[f(\underline{x}_t) - f(\underline{x}^*)\right]$$

$$\Rightarrow \quad f\left(\frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot \underline{x}_t\right) - f(\underline{x}^*) \leq \frac{2B^2}{\mu(T+1)}$$