Lecture #16 Mirror Descent contd. E1 260

- Algorithm
- convergence analysis

$$\min_{\underline{x} \in C} f(x)$$

- f be L- Lipschitz w.r.t $\|\cdot\|$
- $\phi : D \to \mathbb{R}$ $\quad \underline{x} \in \bar{D}$ $\quad C \cap D \neq \phi$
- $\nabla \phi : D \Rightarrow \mathbb{R}^n$ (surjective)

$$D = \partial om\ \phi$$

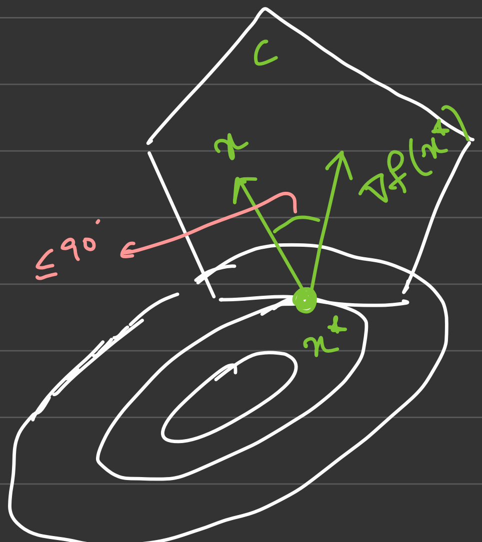
Mirror descent algorithm:

$$\underline{x}_{t+1} = \arg\min_{\underline{x} \in C} \underbrace{\eta_t \underline{g}_t^T \underline{x} + D_\phi(\underline{x}, x_t)}_{\eta_t \underline{g}_t^T \underline{x} + \phi(\underline{x}) - [\phi(x_t) + \nabla\phi^T(x_t)(\underline{x} - x_t)]}$$

Optimality condition:

$$\nabla f^T(\underline{x}^*)(\underline{x} - \underline{x}^*) \geq 0 \quad \forall x \in C$$

$$\Longleftrightarrow \quad -\nabla f(\underline{x}^*) \in \underbrace{N_C(x^*)}$$



Normal cone:

$$N_C(\underline{x}^*) = \{ \underline{x} \in \mathrm{dom}(f) \mid \underline{x}^T(\underline{x} - \underline{x}^*) \leq 0, \forall \underline{x} \in C)$$

$$\Rightarrow \quad 0 \in \nabla f(\underline{x}^*) + N_C(\underline{x}^*)$$

MD update:

$$\underline{0} \in \underbrace{\eta \underline{g}_t + \nabla\phi(x_{t+1}) - \nabla\phi(x_t)}_{y_{t+1}} + N_C(x_{t+1})$$

$y_{t+1} \longrightarrow$ Bregman projection

$$\eta_t \underline{g}_t + \nabla\phi(x) - \nabla\phi(x_t) = 0$$

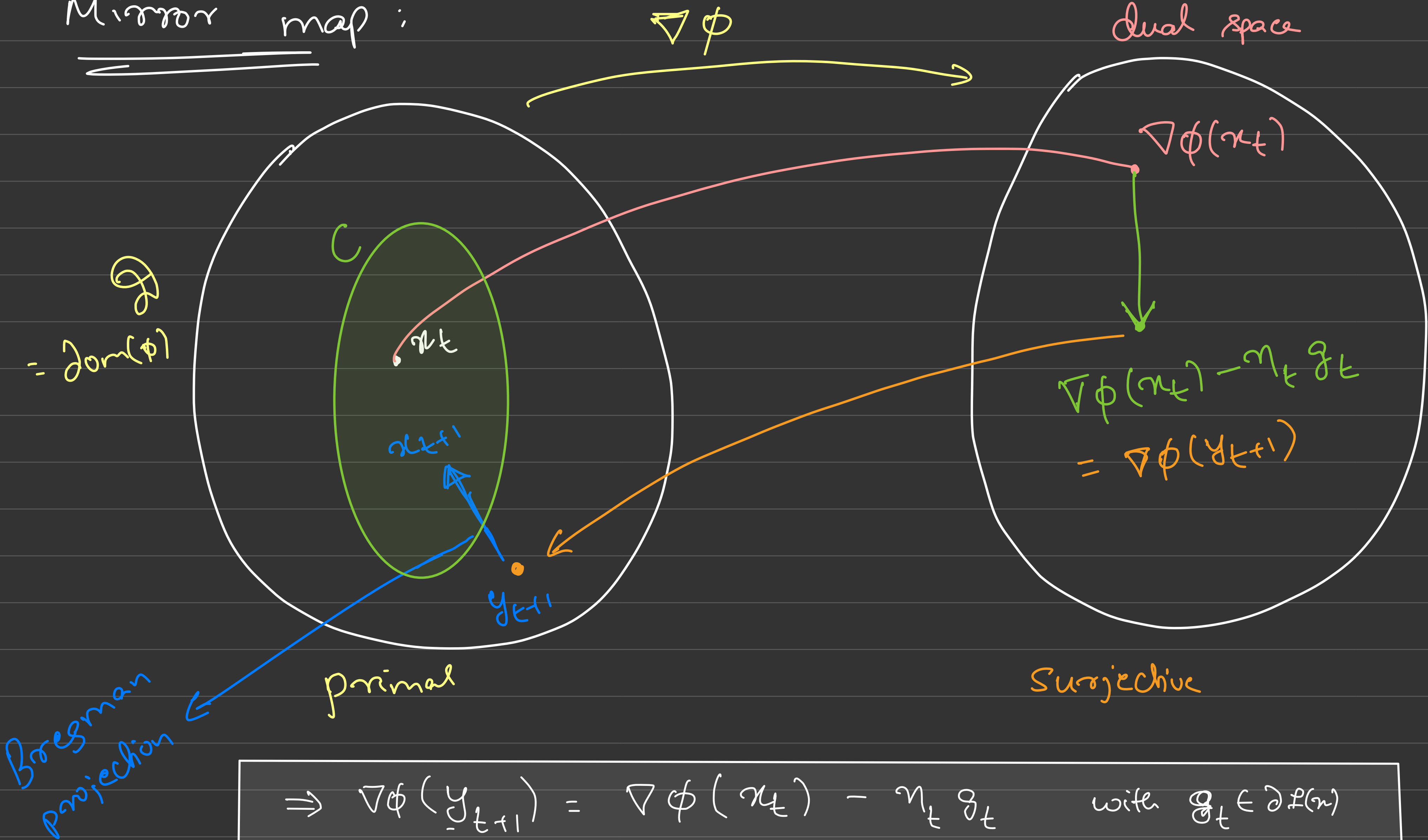
$$\boxed{\begin{aligned} &\Rightarrow \nabla\phi(\underline{y}_{t+1}) = \nabla\phi(x_t) - \eta_t g_t \quad \text{with } \underline{g}_t \in \partial f(x) \\ &\underline{x}_{t+1} \in P_{C,\phi}(\underline{y}_{t+1}) = \underset{\underline{z} \in C}{\arg\min}\ D_\phi(\underline{z}, \underline{y}_{t+1}) \end{aligned}}$$

"dual space"

ϕ : mirror map

$\nabla\phi$: mapping to a dual space

Mirror map :



$$\Rightarrow \nabla\phi(\underline{y}_{t+1}) = \nabla\phi(x_t) - \eta_t g_t \quad \text{with } g_t \in \partial f(x)$$

$$\underline{x}_{t+1} \in P_{C,\phi}(\underline{y}_{t+1}) = \arg\min_{\underline{z} \in C} D_\phi(\underline{z}, \underline{y}_{t+1})$$

Theorem: Suppose f is convex and L-Lipschitz continuous (i.e., $\|g\|_* \leq L$, $g \in \partial f(x)$) on C and ϕ is ρ-strongly convex w.r.t. $\|\cdot\|$. Then

$$f_T^{best} - f^{opt} \leq \frac{R^2 + \frac{L^2}{2\rho} \sum_{k=0}^{T-1} \eta_k^2}{\sum_{k=0}^{T-1} \eta_k}$$

with $R^2 = \sup_{x \in C} D_\phi(x, x_0)$: diameter of the set

- If $\eta_t = \frac{R}{L}\sqrt{\frac{2\rho}{L}}$ then

$$f_T^{best} - f^{opt} \leq \sqrt{\frac{2}{\rho}} \cdot \frac{RL}{\sqrt{T}}$$

Basic inequality:

$$\eta_t \left(f(\underline{x}_t) - f^{opt} \right) \leq D_\phi(\underline{x}^*, \underline{x}_t) - D_\phi(\underline{x}^*, \underline{x}_{t+1}) + \frac{\eta_t^2 L^2}{2\rho}$$

Proof:

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

(subgradient defn.)

MD update rule:

$$\eta_t \underline{g}_t = \nabla \phi(\underline{x}_t) - \nabla \phi(y_{t+1})$$

$$f(\underline{x}_t) - f(\underline{x}^*) \leq \frac{1}{\eta_t} \left[\nabla \phi(\underline{x}_t) - \nabla \phi(\underline{y}_{t+1}) \right]^T (\underline{x}_t - \underline{x}^*)$$

Three-point Lemma:

$$D_\phi(\overset{y_{t+1}}{\underline{z}}... $$

$$D_\phi(\underline{x},\underline{z}) = D_\phi(\overset{x^*}{\underline{x}},\overset{x_t}{\underline{y}}) + D_\phi(\underline{y},\overset{y_{t+1}}{\underline{z}})$$

$$- \left(\nabla\phi(\underset{\substack{\downarrow\\ \underline{y}_{t+1}}}{\underline{z}}) - \nabla\phi(\underset{\substack{\downarrow\\ \underline{x}_t}}{\underline{y}})\right)^T (\underset{\substack{\downarrow\\ \underline{x}^*}}{\underline{x}} - \underset{\substack{\downarrow\\ x_t}}{\underline{y}})$$

$$\Rightarrow f(x_t) - f(x^*) \le \frac{1}{\eta_t}\Big[D_\phi(\underline{x}^*, \underline{x}_t) + D_\phi(\underline{x}_t, y_{t+1})$$

$$- D_\phi(\underline{x}^*, \underline{y}_{t+1})\Big]$$

Pythagorean Lemma:

$$0 \geq D_\phi(\underline{z}, \underline{x}_{c,\phi}) + D_\phi(\underline{x}_{c,\phi}, x) - D_\phi(\underset{\substack{\downarrow\\ x^*}}{\underline{z}}, \underset{\substack{\downarrow\\ y_{t+1}}}{\underline{x}})$$

$$\Rightarrow f(x_t) - f(\underline{x}^*) \le$$

$$\frac{1}{\eta_t}\Big[D_\phi(\underline{x}^*, \underline{x}_t) + D_\phi(\underline{x}_t, \underline{y}_{t+1})\Big] - \frac{1}{\eta_t}\Big[D_\phi(\underline{x}^*, \underline{x}_{t+1}) + D_\phi(\underline{x}_{t+1}, \underline{y}_{t+1})\Big]$$

$$= \frac{1}{\eta_t}\Big[D_\phi(\underline{x}^*, \underline{x}_t) - D_\phi(\underline{x}^*, \underline{x}_{t+1})\Big] + \frac{1}{\eta_t}\Big\{D_\phi(\underline{x}_t, \underline{y}_{t+1}) - D_\phi(x_{t+1}, y_{t+1})\Big\}$$

Claim:

$$D_\phi(\underline{x}_t, \underline{y}_{t+1}) - D_\phi(\underline{x}_{t+1}, \underline{y}_{t+1}) \leq \frac{(\eta_t L)^2}{2\rho}$$

$$\Rightarrow \eta_t \left(f(\underline{x}_t) - f(\underline{x}^*) \right)$$

$$\leq \left[D_\phi(\underline{x}^*, \underline{x}_t) - D_\phi(\underline{x}^*, \underline{x}_{t+1}) \right] + \frac{(\eta_t L)^2}{2\rho}$$

$$D_\phi(\underline{x}_t, \underline{y}_{t+1}) - D_\phi(\underline{x}_{t+1}, \underline{y}_{t+1})$$

$$= \phi(\underline{x}_t) - \left[\phi(\underline{y}_{t+1}) - \nabla\phi^T(\underline{y}_{t+1}) [\underline{x}_t - \underline{y}_{t+1}] \right]$$

$$- \phi(\underline{x}_{t+1}) + \left\{ \phi(\underline{y}_{t+1}) - \nabla\phi^T(\underline{y}_{t+1}) [\underline{x}_{t+1} - \underline{y}_{t+1}] \right\}$$

$$= \phi(\underline{x}_t) - \phi(\underline{x}_{t+1}) - \nabla\phi^T(\underline{y}_{t+1}) [\underline{x}_t - \underline{x}_{t+1}]$$

ρ - Strong Convexity of $\phi(\underline{x}_t)$

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

$$- \nabla \phi^T(\underline{y}_{t+1})(\underline{x}_t - \underline{x}_{t+1})$$

$$= [\nabla \phi(x_t) - \nabla \phi(\underline{y}_{t+1})]^T(\underline{x}_t - \underline{x}_{t+1})$$

$$- \frac{\rho}{2} \| \underline{x}_t - \underline{x}_{t+1} \|^2$$

MD update rule:

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

$$= \eta_t \underline{g}_t^T (\underline{x}_t - \underline{x}_{t+1}) - \frac{\rho}{2} \| \underline{x}_t - \underline{x}_{t+1} \|^2$$

$$\leq \| \underline{g}_t \|_* \| \underline{x}_t - \underline{x}_{t+1} \|$$

$$\leq \eta_t L \| \underline{x}_t - \underline{x}_{t+1} \| - \frac{\rho}{2} \| \underline{x}_t - \underline{x}_{t+1} \|^2$$

(*) (Cauchy-Schwarz)

min. the upper bound with variable $\|\underline{x}_t - \underline{x}_{t+1}\|$

$$\eta_t L - \frac{\rho}{2} \cdot 2 \|\underline{x}_t - \underline{x}_{t+1}\| = 0$$

$$\Rightarrow \|\underline{x}_t - \underline{x}_{t+1}\| = \frac{\eta_t L}{\rho}$$

$$\Rightarrow^{*} \quad \leq \frac{(\eta_t L)^2}{\rho} - \frac{\rho}{2} \cdot \frac{(\eta_t L)^2}{\rho^2}$$

$$= \frac{(\eta_t L)^2}{2\rho}$$

claim:

$$D_\phi(\underline{x}_t, \underline{y}_{t+1}) - D_\phi(\underline{x}_{t+1}, \underline{y}_{t+1}) \leq \frac{(\eta_t L)^2}{2\rho}$$

$$\Rightarrow \eta_t \left(f(\underline{x}_t) - f(\underline{x}^*) \right)$$

$$\leq \left[D_\phi(\underline{x}^*, \underline{x}_t) - D_\phi(\underline{x}^*, \underline{x}_{t+1}) \right] + \frac{(\eta_t L)^2}{2\rho}$$

Now, the theorem:

Sum $t = 0, \dots, T-1$ (telescope D_ϕ)

$$\sum_{t=0}^{T-1} \eta_t \left(f(\underline{x}_t) - f(\underline{x}^*) \right)$$

$$\leq D_\phi(\underline{x}^*, \underline{x}_0) - D_\phi(\underline{x}^*, \underline{x}_T)$$

$$+ \frac{L^2}{2\rho} \sum_{t=0}^{T-1} \eta_t^2$$

$$\leq R^2 + \frac{L^2}{2\rho} \sum_{t=0}^{T-1} \eta_t^2 \qquad \text{C1}$$

Since

$$\frac{\sum_{t=0}^{T-1} \eta_t f(\underline{x}_t) - \left(\sum_{t=0}^{T-1} \eta_t \right) f(\underline{x}^*)}{\sum_{t=0}^{T-1} \eta_t} \geq f_T^{best} - f^{opt}$$

But $$\sum_{t=0}^{T-1} \eta_t \left(f_t^{best} - f^{opt}\right) \leq \sum_{t=0}^{T-1} \eta_t \left(f(x_t) - f^{opt}\right)$$

$$\left(\sum_{t=0}^{T-1} \eta_t f_t^{best}\right) - f^{opt} \sum_{t=0}^{T-1} \eta_t \leq \sum_{t=0}^{T-1} \eta_t \left(f(x_t) - f^{opt}\right)$$

Or

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

$$\Rightarrow \quad f_T^{best} - f^{opt} \leq \frac{R^2 + \frac{L^2}{2\rho} \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t}$$

$\square$

(C1) with fixed step size η

$$\eta\left(\sum_{t=1}^{T} f(\underline{x}_t) - f(\underline{x}^*)\right) \leq R^2 + \frac{L^2\eta^2 T}{2\rho}$$

$$\Rightarrow \frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*) \leq \frac{R^2}{\eta T} + \frac{\eta L^2}{2\rho} = q(\eta)$$

Now, optimize for the best choice of η

$$q(\eta) = \frac{R^2}{\eta T} + \frac{\eta L^2}{2T}$$

$$\frac{dq(\eta)}{d\eta} = -\frac{R^2}{T\eta^2} + \frac{L^2}{2\ell} = 0$$

$$\frac{R^2}{\eta^2 T} = \frac{L^2}{2e}$$

$$\eta = \sqrt{\frac{2e\,R^2}{L^2 T}} = \frac{R}{L}\sqrt{\frac{2e}{T}}$$

$$\Rightarrow \frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(\underline{x}^*) \leq \frac{R^2}{T}\cdot\frac{L}{R}\cdot\sqrt{\frac{T}{2e}} + \frac{R}{L}\sqrt{\frac{2e}{T}}\cdot\frac{L^2}{2e}$$

$$= \frac{RL}{\sqrt{T}}\cdot\frac{1}{\sqrt{2e}} + \frac{RL}{\sqrt{T}}\cdot\frac{1}{\sqrt{2e}}$$

$$\Rightarrow f_T^{best} - f(\underline{x}^*) \leq \frac{RL}{\sqrt{T}}\cdot\sqrt{\frac{2}{e}} \qquad \square$$

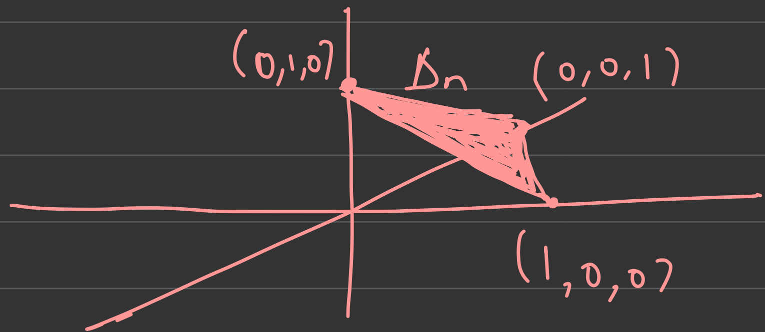
Example:

minimize $\quad f(x) = \|A\underline{x} - b\|_1$

s.t. $\quad \underline{x} \in \Delta_n \quad , \quad \Delta_n \in \{ \underline{x} \in \mathbb{R}^n \mid x \geq 0 , \sum_{i=1}^{n} x_i = 1 \}$

f : 1- Lipschitz w.r.t. $\|\cdot\|_1$

equivalently $\quad \|\nabla f(x)\|_\infty \leq 1$



Recall w.r.t $\|\cdot\|_2$, we had $\|\nabla f(x)\|_2 \leq \sqrt{n}$

Projected subgradient method : $\|\underline{x}_0 - x^*\| \cdot \sqrt{\frac{n}{T}}$

$\underline{\text{MD}}$: $\quad \underline{x}_{t+1} = \underset{\underline{x} \in C}{\arg\min} \quad \eta g_t + D_\phi(\underline{x}, \underline{x}_t)$

What ϕ should we choose?

Suppose $\quad \phi = \sum_{i=1}^{n} x_i \log x_i$.

ϕ is 1 - strongly convex on Δ_n w.r.t. $\|\cdot\|_1$

so $\rho = 1$ (check)

$$R^2 = D_\phi(\underline{x}_0, \underline{x}^*) \quad . \quad \text{Let} \quad \underline{x}_0 = \begin{bmatrix} 1/n \\ \vdots \\ 1/n \end{bmatrix} : \text{Center of } \Delta_n$$

$$\leq \log n$$

$\Rightarrow$ MD : $\quad RL\sqrt{\frac{2}{\rho}} = \sqrt{2} \cdot \sqrt{\frac{\log n}{T}} \quad \left[\text{compared to } \sqrt{\frac{n}{T}}\right]$