Lecture #15   **Mirror Descent**   E1 260

- Algorithm :
  - proximal viewpoint of SG
  - Non. Euclidean geometry
    - Bregman divergence
    - properties

- Convergence analysis : Convex & L-Lipschitz

Recall, Projected (subgradient) method

$$\left.\begin{array}{l} \text{minimize} \quad f(x) \\ \underline{x} \in C \end{array}\right\} \quad \underline{x}_{t+1} = \text{Proj}_C(\underline{x}_t - \eta_t \underline{g}_t), \quad \underline{g}_t \in \partial f(\underline{x})$$

if f is L - Lipschitz [$\|g\| \leq L$ or

$|f(x) - f(y)| \leq L\|\underline{x} - \underline{y}\| \quad \forall\, \underline{x}, \underline{y} \in \text{dom}(f)$

and $\underline{g} \in \partial f(\underline{x})$], then $\eta_t = \eta = \dfrac{R}{L\sqrt{T}}$

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\underline{x}_t) - f(\underline{x}^*) \leq RL \frac{1}{\sqrt{T}}$$

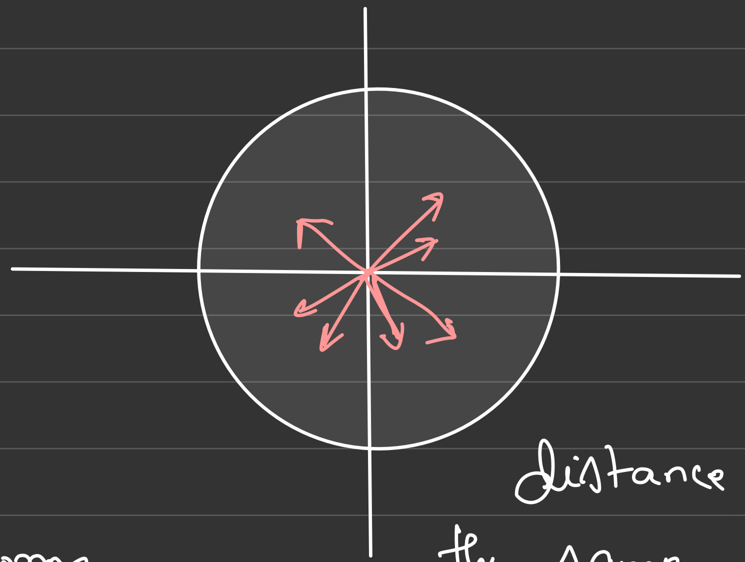
- Are these dimension independent?
  - Are dimension dependent const. in 'L'?

Recall: Proximal view of proj. subgradient method

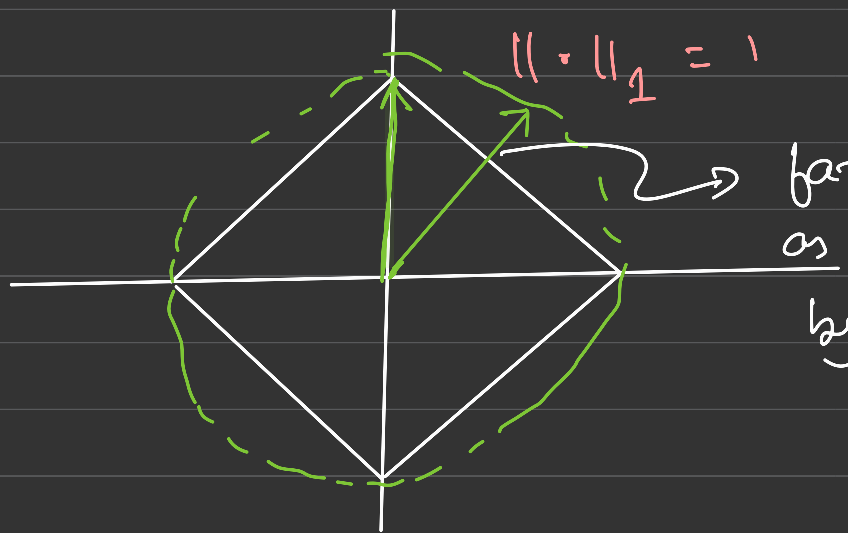
$$\underline{x}_{t+1} = \arg\min_{\underline{x}} \; f(\underline{x}_t) + \underline{g}_t^T(\underline{x} - \underline{x}_t) + \frac{1}{2\eta}\|\underline{x} - \underline{x}_t\|_2^2$$

$$= \arg\min_{\underline{x}} \; \eta\, g_t^T \underline{x} + \underbrace{\frac{1}{2}\|\underline{x} - \underline{x}_t\|_2^2}_{\text{(Proximal term)}}$$

Euclidean norm: Spherical Symmetry



Not true about other norms:



$\rightarrow$ farther away as measured by the l_1-norm

distance is the same for all these steps.

Example of Quadratic minimization:

$$f(x_1, x_2) = x_1^2 \cdot \frac{1}{100} + x_2^2 \, 100$$

Suppose we are at $x_t = \begin{pmatrix} -10 \\ -0.1 \end{pmatrix}$

$$\nabla f(x_t) = \begin{pmatrix} 2x_1/100 \\ 2x_2 \cdot 100 \end{pmatrix} \Bigg|_{\begin{pmatrix} -10 \\ -0.1 \end{pmatrix}} = \begin{pmatrix} -1/5 \\ -20 \end{pmatrix}$$

$$\underline{x}_{t+1} = \arg\min_{\underline{x}} \; \eta \, \nabla f(x_t)^T \underline{x} + \frac{1}{2} (\underline{x} - \underline{x}_t)^T I (\underline{x} - x_t)$$

Suppose :

$$\underline{x}_{t+1} = \arg\min_{\underline{x}} \; \eta \, \nabla f(x_t)^T \underline{x} + \frac{1}{2} \overbrace{(\underline{x} - \underline{x}_t)^T Q (\underline{x} - x_t)}^{\|\underline{x} - x_t\|_Q^2}$$

Skew gradients to fit geometry better

$$\eta \nabla f(x_t) + Q(\underline{x} - x_t) = 0$$

$$\underline{x}_{t+1} = x_t - \eta Q^{-1} \nabla f(\underline{x}_t)$$

$$Q := \begin{pmatrix} 1/50 & 0 \\ 0 & 200 \end{pmatrix} \Rightarrow \quad x_{t+1} = x_t - \eta \begin{pmatrix} 50 & 0 \\ 0 & \frac{1}{200} \end{pmatrix} \begin{pmatrix} -1/5 \\ -20 \end{pmatrix}$$

$$= \begin{pmatrix} -10 \\ -0.1 \end{pmatrix} - \eta \begin{pmatrix} -10 \\ -0.1 \end{pmatrix}$$

Mirror descent:

- measure distance using a different norm ?
- adjust gradient updates to fit the geometry of the problem.

This changes Lipschitzness of the function.

$$| f(x) - f(y) | \leq L \; \| \underline{x} - \underline{y} \|_2$$

- Example: $f(\underline{x}) = \|\underline{x}\|_1$ with $\underline{x} = \underline{1}$ and $\underline{y} = (1+\epsilon)\,\underline{1}$

$$\big| \|\underline{x}\|_1 - \|\underline{y}\|_1 \big| \leq L \; \|\underline{x} - \underline{y}\|_2$$

$$| n - (1+\epsilon) n | = \epsilon n \leq L \; \|\underline{x} - \underline{y}\|_2 = L \; \|\underline{1} - (1+\epsilon)\underline{1}\|_2$$

$$\epsilon n \leq L \, \epsilon \sqrt{n}$$

For the upper bound to be valid: $\underbrace{L = \sqrt{n}}_{\text{depends on the dimension}}$

$$\Rightarrow \; \|\nabla f(x)\|_2 \leq \sqrt{n}$$

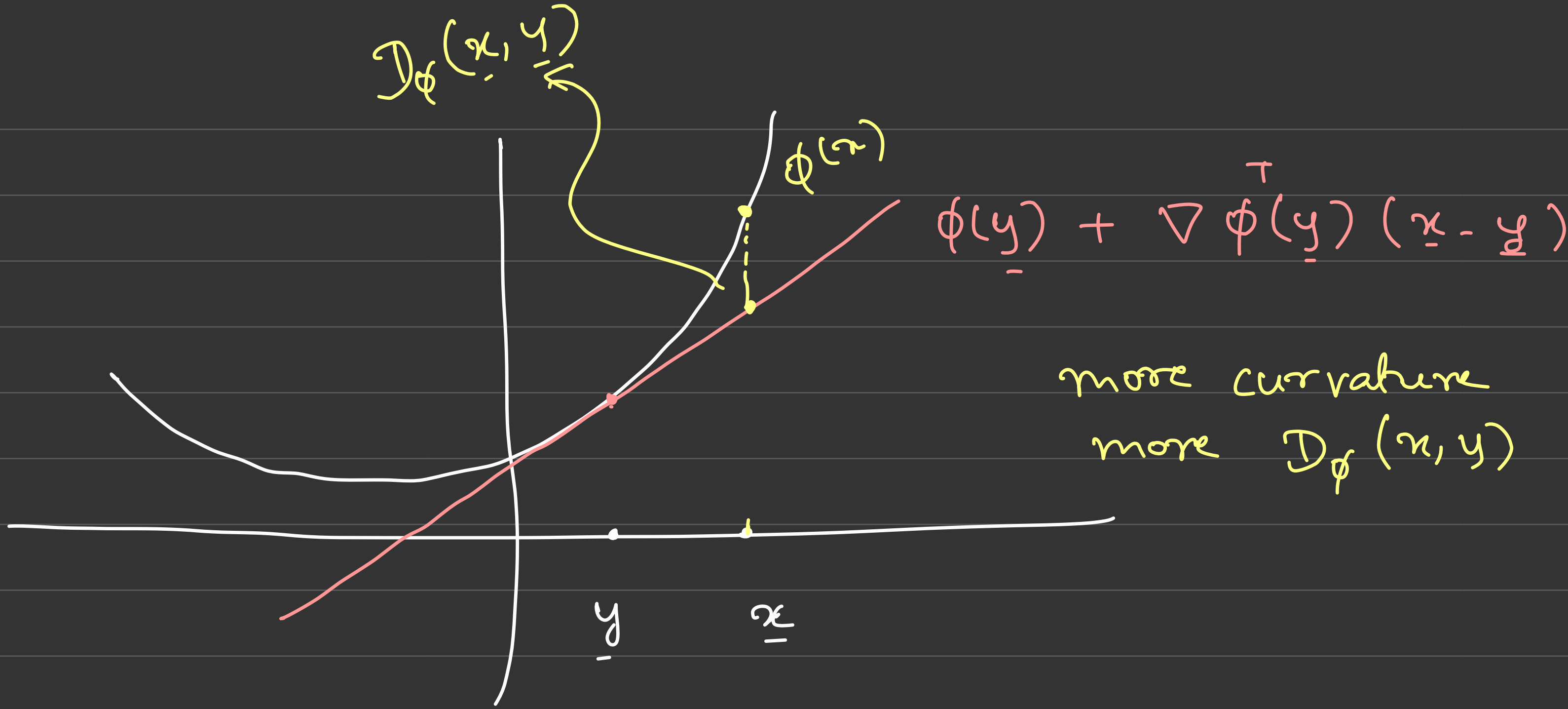
- Suppose f is 1 - Lipschitz w.r.t. a different norm say $\|\nabla f(x)\|_\infty \leq 1$ (we had this for $\|\cdot\|_1$) $\Rightarrow \|\nabla f(x)\|_2 < \sqrt{n}$

Mirror descent fixes this by:

replacing the $\|\underline{x} - \underline{x}_t\|_2^2$ with distance - like metric

$$\underbrace{D_\phi(\underline{x}, \underline{y})}_{\text{Bregman divergence}} = \phi(\underline{x}) - \left[\phi(\underline{y}) + \nabla\phi^T(\underline{y})(\underline{x} - \underline{y})\right]$$

Bregman divergence for convex and differentiable $\phi(\underline{x})$



more curvature
more $D_\phi(x, y)$

Example:

$$\phi(x) = \frac{1}{2}\|\underline{x}\|_2^2$$

$$D_\phi(\underline{x}, \underline{y}) = \frac{1}{2}\|\underline{x}\|_2^2 - \left[\frac{1}{2}\|\underline{y}\|_2^2 + \underline{y}^T(\underline{x} - \underline{y})\right]$$

$$= \frac{1}{2}\left[\|\underline{x}\|_2^2 - 2\underline{y}^T\underline{x} - \|\underline{y}\|_2^2 + 2\|y\|^2\right]$$

$$= \frac{1}{2}\|\underline{x} - \underline{y}\|_2^2$$

(2) $\phi(\underline{x}) = \frac{1}{2}\underline{x}^T Q \underline{x} \qquad Q \succ 0$

$$Q = \begin{bmatrix} 1/50 & \\ & 200 \end{bmatrix}$$

$D_\phi(\underline{x}, \underline{y}) = \frac{1}{2}(\underline{x} - \underline{y})^T Q (\underline{x} - \underline{y})$

$D_\phi(x, y)$ v.s $\|x - y\|_2^2$?

Squared Mahalanobis distance

(3) $\phi(\underline{x}) = \sum_i x_i \log x_i$ (negative entropy)

$D_\phi(\underline{x}, \underline{y}) = KL(\underline{x} \| \underline{y}) = \sum_i x_i \log\left(\frac{x_i}{y_i}\right)$

(Check as an excercise)

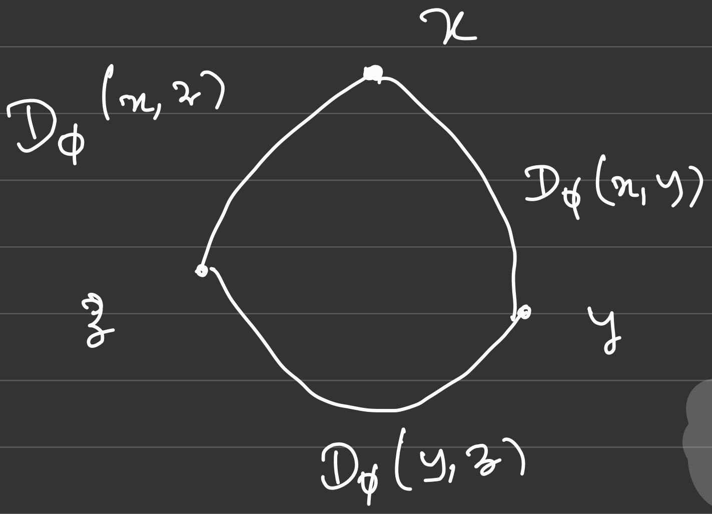
Homework: With $\Delta = \{ \underline{x} \in \mathbb{R}_+^n \mid \underline{1}^T \underline{x} = 1 \}$ probability simplex

MD results in "entropic descent".

Properties of Bregman divergence:

Three-point Lemma:

$$D_\phi(\underline{x},\underline{z}) = D_\phi(\underline{x},\underline{y}) + D_\phi(\underline{y},\underline{z}) - \left(\nabla\phi(\underline{z}) - \nabla\phi(\underline{y})\right)^T(\underline{x}-\underline{y})$$



Cosine Law for the Euclidean case.

- $$\|\underline{x}-\underline{z}\|_2^2 = \|\underline{x}-\underline{y}\|_2^2 + \|y-z\|_2^2 - 2(z-y)^T(\underline{x}-y)$$

Proof:

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

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

$$= \cancel{\phi(x)} - \cancel{\phi(y)} - \nabla\phi^T(y)(\underline{x}-y) + \cancel{\phi(y)} - \cancel{\phi(z)} - \nabla\phi^T(z)(y-\underline{z}) - \left[\cancel{\phi(\underline{x})} - \cancel{\phi(\underline{z})} - \nabla\phi^T(z)(x-z)\right]$$

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

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

$$= [\nabla\phi(\underline{z}) - \nabla\phi(\underline{y})]^T(\underline{x}-\underline{y}) \quad \blacksquare$$

② Convexity of $D_\phi(\underline{x},\underline{y})$ in $\underline{x}$

$$D_\phi(\underline{x},y) = \underbrace{\phi(x)} - \phi(y) - \nabla\phi^T(y)(\underline{x}-\underline{y})$$

follows from convexity of $\phi(\cdot)$

L- Lipschitz definition for arbitrary norm:

$$|f(x) - f(y)| \leq L \, \|\underline{x} - y\|$$

Defn: of subgradients : $f(y) \geq f(x) + \underline{g}^T (\underline{y} - \underline{x})$

$$f(\underline{x}) - f(\underline{y}) \leq \underline{g}^T (\underline{y} - \underline{x})$$

$$|f(\underline{x}) - f(\underline{y})| \leq \|\underline{g}\|_* \, \|\underline{x} - \underline{y}\|$$

(Generalize Cauchy-Schwarz or Holder's inequality)

This gives us another definition of L- Lipschitz w.r.t. $\|\cdot\|$

$$\|\underline{g}\|_* \leq L$$

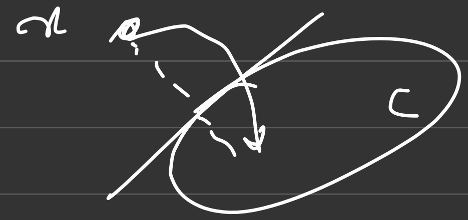
Convex and Lipschitz problems:

$$\text{minimize} \quad f(x)$$
$$\text{s. to} \quad x \in C$$

- f is convex and Lipschitz continuous
  - $\|\underline{g}\|_* \leq L$ for any $\underline{g} \in \partial f(\underline{x})$
- $\phi(\cdot)$ is ρ - strongly convex wrt. $\|\cdot\|$

$$\Rightarrow \quad D_\phi(\underline{x}, \underline{y}) = \phi(x) - \left[\phi(\underline{y}) + \nabla\phi^T(\underline{y})(\underline{x} - \underline{y})\right]$$
$$\geq \frac{\rho}{2} \|\underline{x} - \underline{y}\|^2$$

Bregman projection:



Given a point $\underline{x}$,

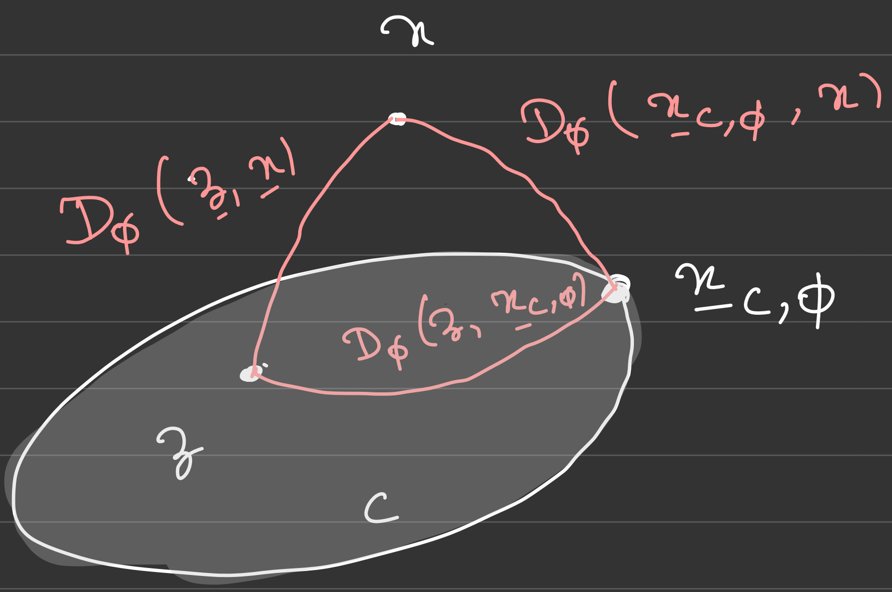
$$P_{C,\phi}(\underline{x}) = \underset{\underline{z} \in C}{\arg\min}\; D_\phi(\underline{x}, \underline{z})$$

if $\phi = \frac{1}{2}\|\underline{x}\|_2^2$, we have

$$D_\phi(\underline{x}, \underline{z}) = \frac{1}{2}\|\underline{x} - \underline{z}\|_2^2.$$

Then, $P_{C,\phi}(x) = P_C(\underline{x})$ [Orthogonal Euclidean projection]

Generalized Pythagorean theorem:



if $\underline{x}_{c,\phi} = P_{c,\phi}(\underline{x})$

Then,

$$D_\phi(\underline{z}, \underline{x}) \geq D_\phi(\underline{z}, \underline{x}_{c,\phi}) + D_\phi(\underline{x}_{c,\phi}, x)$$

$$\forall\ \underline{z} \in C$$

$$\underline{x}_{c,\phi} = \underset{\underline{z} \in C}{\arg\min}\ D_\phi(\underline{z}, \underline{x})$$

Recall optimality condn:

$$\underline{g}^T(\underline{z} - \underline{x}_{c,\phi}) \geq 0 \quad \forall\ \underline{z} \in C$$

Let $\underline{g} = \nabla_{\underline{z}} D_\phi(\underline{z}, \underline{x})\Big|_{\underline{z} = \underline{x}_{c,\phi}}$

$$D_\phi(\underline{z}, \underline{x}) = \phi(\underline{z}) - \left[\phi(x) + \nabla\phi^T(x)(\underline{z} - x)\right]$$

$$\Rightarrow \quad \underline{g} = \nabla\phi(\underline{x}_{c,\phi}) - \nabla\phi(\underline{x})$$

$$g^T(\underline{z} - \underline{x}_{c,\phi}) \geq 0$$

$$\Rightarrow \left[\nabla\phi(\underline{x}_{c,\phi}) - \nabla\phi(x)\right]^T \left[\underline{z} - \underline{x}_{c,\phi}\right] \geq 0$$

Three-point Lemma : $D_\phi(\underline{x}, \underline{z}) = D_\phi(\underline{x}, \underline{y}) + D_\phi(\underline{y}, \underline{z})$
$- \left(\nabla\phi(\underline{z}) - \nabla\phi(\underline{y})\right)^T(\underline{x} - \underline{y})$

$$0 \geq D_\phi(\underline{z}, \underline{x}_{c,\phi}) + D_\phi(\underline{x}_{c,\phi}, x) - D_\phi(\underline{z}, \underline{x})$$

$\blacksquare$