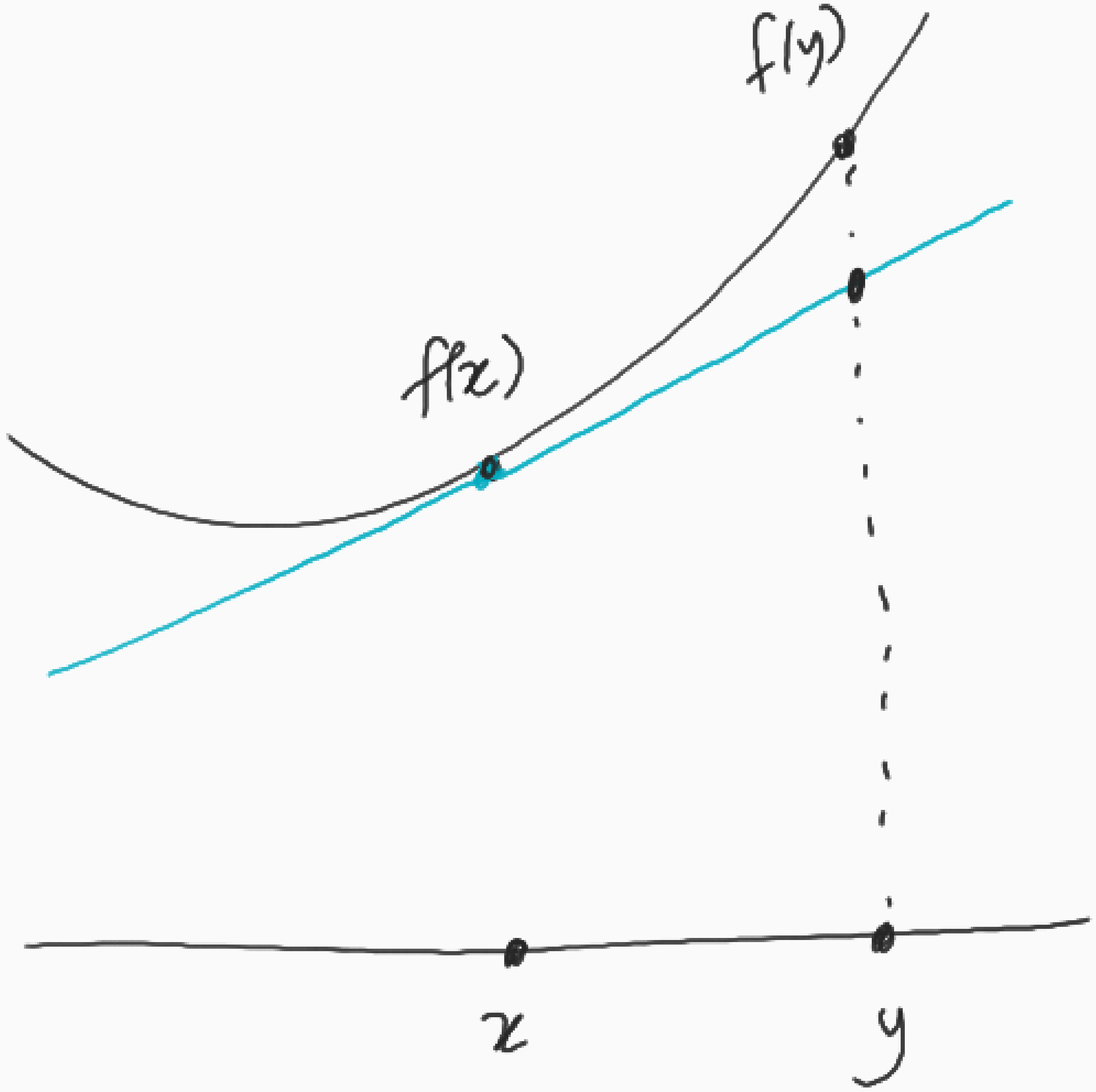
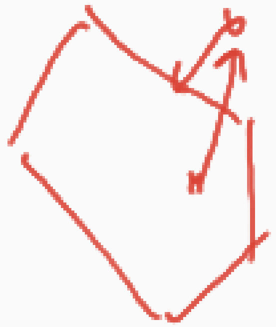
Lecture 19 : Mirror Descent

- recap : and online optimization
  - Bregman distances
  - two views
    - proximal point
    - mirror map.
    - preconditioned GD
    - follow the regularized leader.



Recap:

$$\min_{x \in K} f(x) \quad \text{convex set}$$



PGD: → OCO

$x_1 = \text{start}$

$$x'_{t+1} \leftarrow x_t - \eta \underbrace{\nabla f_t(x_t)}_{\overset{\text{def}}{=} \nabla_t}$$

$$x_{t+1} \leftarrow \text{Proj}_K(x'_{t+1})$$

OCO. at time t:

you play x_t

adv plays $f_t(\ )$

pay/loss $= \boxed{f_t(x_t)}$

Thm:

$$\sum_{t=1}^{T} f_t(x_t) \le \sum_{t=1}^{T} f_t(x^*) + \underbrace{\sum_{t=1}^{T} \|\nabla_t\|^2 \frac{\eta}{2} + \frac{\|x_1 - x^*\|^2}{2\eta}}_{\text{regret against } x^*}$$

Pf. $\Phi_t = \dfrac{\|x_t - x^*\|^2}{2\eta}$

Claim: $f_t(x_t) + \Phi_{t+1} - \Phi_t \le f_t(x^*) + \text{junk}_t$

Pf:

$$f_t(x_t) + \frac{1}{2\eta}\left[\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right]$$

$$= f_t(x_t) + \frac{1}{2\eta}\left[\|x_{t+1} - x_t\|^2 + 2\langle x_{t+1} - x_t, x_t - x^*\rangle\right]$$

$$= f_t(x_t) + \langle -\nabla_t, x_t - x^*\rangle + \frac{\eta^2}{2\eta}\|\nabla_t\|^2$$

$$\le f_t(x^*) + \frac{\eta}{2}\|\nabla_t\|^2$$

$$\Rightarrow \sum_t f_t(x_t) \le \sum_t f_t(x^*) + \sum_t \text{junk}_t + \Phi_0$$

[Zinkevich 02]

Thm: $\sum f_t(x_t) \le \sum f_t(x^*) + \underbrace{\sum \|\nabla_t\|^2 \cdot \frac{\eta}{2} + \frac{\|x_1 - x^*\|^2}{2\eta}}_{\text{regret } = \varepsilon}$

$\underbrace{T = \left(\frac{GD}{\varepsilon}\right)^2}$ upper bd on $\|\nabla f_t\|$; diam $\le \|x_1 - x^*\|$

Special Case: $f_t(x) = \langle l_t, x\rangle$ $\quad l_t \in [-1,1]^n$

$K = \Delta_n$ prob simplex

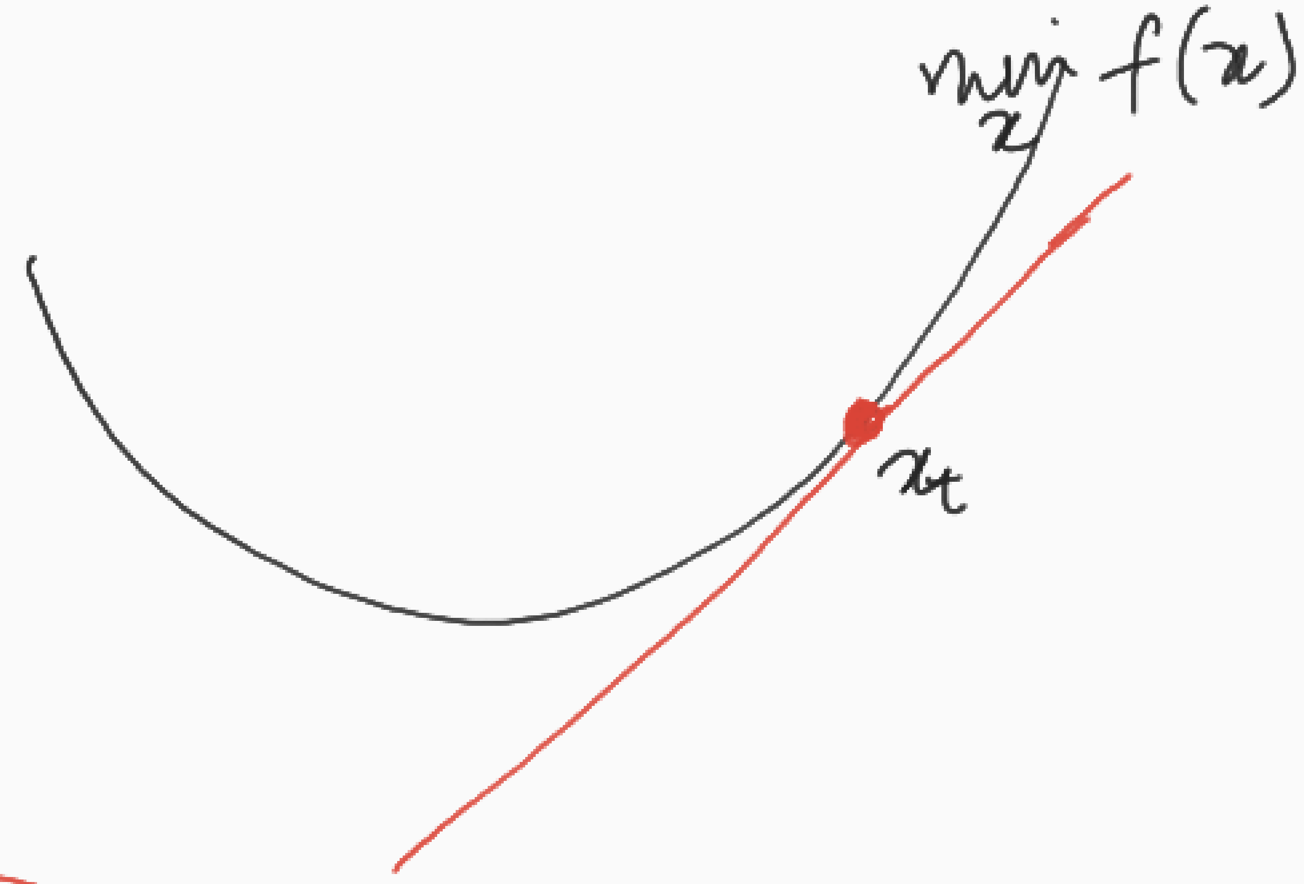
[MW | Hedge got regret ε after $T = 4\frac{\log n}{\varepsilon^2}$ steps

[GD ———— $T = \left(\frac{2 \cdot n}{\varepsilon^2}\right)$ ☹

$f_t = \langle l^t, \cdot\rangle \Rightarrow \nabla f_t = l^t \Rightarrow \| \quad \|_2 \le \sqrt{n}$

$D = \max_{x,y \in \Delta_n} \|x - y\|_2 \le \sqrt{2}$

Proximal Point View



$$f(x) \simeq f(x_t) + \langle \nabla f(x_t), x - x_t \rangle$$

$$x_{t+1} \leftarrow \arg\min_x \left\{ \langle \nabla f(x_t), x \rangle + \underbrace{\overbrace{\frac{1}{\eta} \frac{\|x - x_t\|^2}{2}}^{\text{regularizer}}} \right\}$$

$$\nabla \frac{\|y\|^2}{2} = y$$

$\nabla RHS = 0$ at x_{t+1}

$$\nabla f(x_t) + \frac{1}{\eta}(x_{t+1} - x_t) = 0$$

$$\iff x_{t+1} = x_t - \eta \nabla f(x_t)$$

(GD update rule!)

$$x_{t+1} \leftarrow \underset{x}{\arg\min} \left\{ \langle \nabla f(x_t), x \rangle + \frac{1}{\eta} \underbrace{KL(x \| x_t)} \right\}$$

$$\| \quad \|^2 \quad l_2^2$$

$$\sum_i x_i^{1/2}$$

$\nabla RHS = 0$ at optimizer

$$0 = \nabla f(x_t) + \frac{1}{\eta}\left[\left(\log \frac{x_1}{(x_t)_1} + 1, \quad \ldots \ldots, \log \frac{x_n}{(x_t)_n} + 1\right)\right]$$

$$(\nabla_t)_1 + \frac{1}{\eta}\left(\log \frac{x_1}{(x_t)_1} + 1\right) = 0$$

Hedge!

$$\Rightarrow x_1 = (x_t)_1 \cdot \underline{\underline{e^{-(\eta(\nabla_t)_1 \cancel{+1})}}} = x_1^{old} \cdot e^{-\eta l_1}$$

$$\boxed{KL(x \| q) = \sum_{i=1}^{n} x_i \log \frac{x_i}{q_i}}$$

$$\frac{\partial KL(x\|q)}{\partial x_i} = \frac{d}{dx_i}\left(x_i \log \frac{x_i}{q_i}\right) = \frac{d}{dx_i}\left(x_i \log x_i - x_i \log q_i\right) = \log x_i + x_i \frac{1}{x_i} - \log q_i = \left(\log \frac{x_i}{q_i} + 1\right)$$

Mirror Descent:

- Choose h

$$\Rightarrow D_h(\cdot \| \cdot)$$

- $$x_{t+1} \leftarrow \underset{x}{\operatorname{argmin}} \left\{ \cancel{\langle \nabla f(x_t), x \rangle} + \frac{1}{\eta} D_h(x \| x_t) \right\}$$

$f(x_t) + \langle \nabla f(x_t), x - x_t \rangle$

$$\boxed{\nabla h(x_{t+1}) = \nabla h(x_t) - \eta \nabla f(x_t)}$$

$$\frac{1}{2}\|x-y\|^2 = D_{\frac{1}{2}\|x\|^2}(x \| y)$$

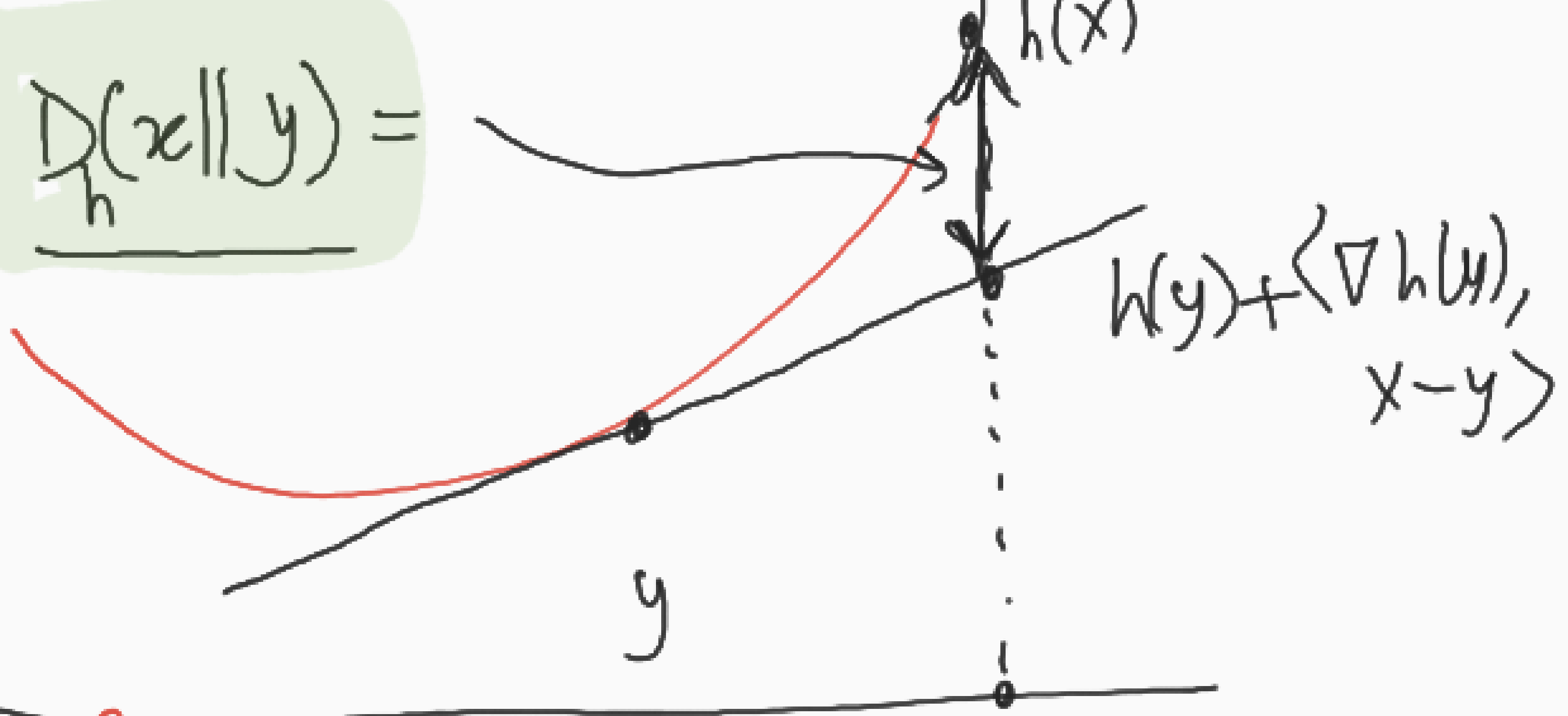
$$KL(x \| y) = D_{H(x)}(x \| y), \quad H(x) = \sum_i x_i (\log x_i - 1)$$

Bregman divergences

fix function h convex fn

distance defining

$D_h(x \| y) =$



$h(x) = x^2/2$

$$D(x \| y) = h(x) - \left[h(y) + \langle \nabla h(y), x-y \rangle \right] = \frac{1}{2}x^2 - \frac{1}{2}y^2 - y(x-y) = \frac{1}{2}(x-y)^2$$

Mirror Map View [Nemirovki - Yudin]

$$x_{t+1} \leftarrow x_t - \eta \nabla f(x_t)$$

$\in \mathbb{R}^n$ $\quad$ $\in \mathbb{R}^n$

linear map from $\mathbb{R}^n \to \mathbb{R}$

dual space

$\text{gradient}_{x_t}(y)$

$= \langle g_{x_t}, y \rangle$

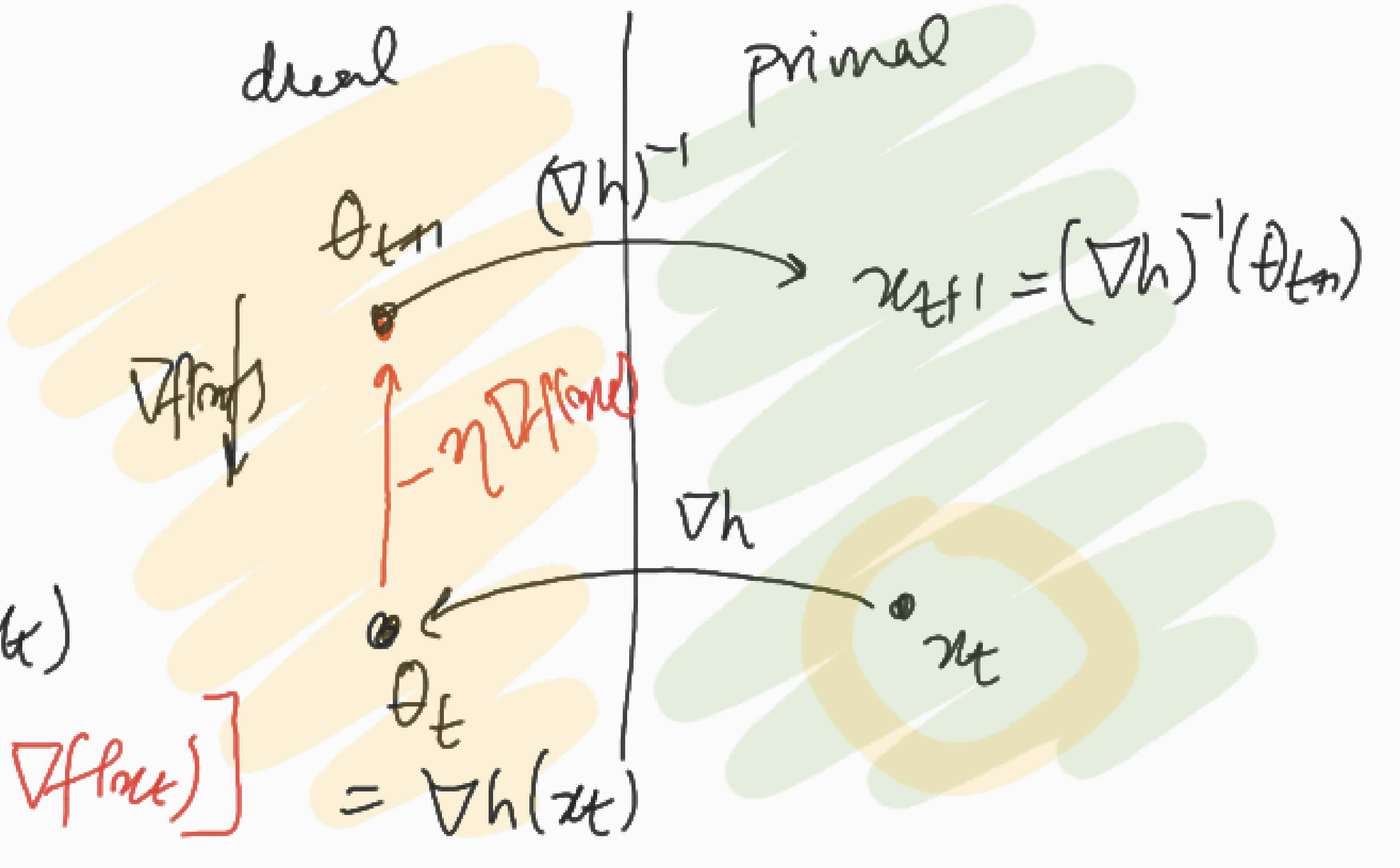
$\nabla f(x_t)$

fix a fn $h : \mathbb{R}^n \to \mathbb{R}$

"mirror map" : $\underline{\nabla h} : \mathbb{R}^n \to \mathbb{R}^n$

$$\nabla h(x_{t+1}) = \nabla h(x_t) - \eta \cdot \nabla f(x_t)$$

$$x_{t+1} = \underline{(\nabla h)}^{-1}\left[\nabla h(x_t) - \eta \nabla f(x_t)\right]$$



$$\sum f_t(x_t) \le \sum f_t(x^*) + \sum \|\nabla f(x_t)\|^2 \cdot \frac{\eta}{2} + \frac{\|x_1 - x^*\|^2}{2\eta}$$

① Chose norm $\| \|_1$

② $h : \sum (x_i \lg x_i - x_i)$

$$\underbrace{\frac{\eta}{2}\sum \boxed{\|\nabla f(x_t)\|_*^2} + \frac{D_h(x^* \| x_1)}{2\eta}}$$

if h is strongly convex wrt $\| \|_1$ [

Hedge regret $\frac{\eta}{2}\underbrace{\sum \|\nabla\|_\infty^2}_{\le \frac{\eta}{2}T + \frac{\lg n}{2\eta}} + \frac{KL(x^* \| x_1)}{2\eta}$

unif

$$\underline{\underline{x_{t+1}}} \leftarrow \underline{\underline{x_t}} - \eta \nabla f(x_t) \qquad \text{(GD)}$$

$$x - \eta [H_h]^{-1} \nabla f(x_t) \qquad \text{(MD)}$$

$$x - \eta [H_f]^{-1} \nabla f(x_t) \qquad \text{[Newton]}$$

condition # of f

$K \log \frac{1}{\varepsilon} \rightarrow \sqrt{K} \log \frac{1}{\varepsilon}$

<u>Smooth fns:</u> GD: $O(1/\varepsilon)$

Acceleration: $O(1/\sqrt{\varepsilon})$ Nesterov $x_{t+1} = g(x_t, x_{t-1})$