$x_{t+1} = \arg\min_x \{ \eta\, g_t \cdot x + \| x - x_t \|^2 \}$ fenchel dual $| 2.5 \times 10^{-3} - 50 \times 10^{-3} |$

$\| \cdot - x_t \|^2 \, [\,\eta g_t\,] = x_t - \eta g_t$

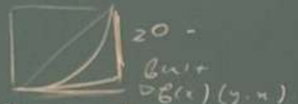
$\| x_{t+1} - x^* \|_2^2 - \| x_t - x^* \|_2^2 =$ 'drop'

$y = \arg\min_{x \in K} f(x)$

f convex, K convex body $\subseteq \mathbb{R}^P$.

$\downarrow$ $\quad \forall x, y$.

$f(y) \geq f(x) + \nabla f(x)(y - x)$



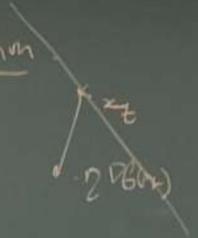
$-\nabla f(x)(y - x) \geq f(x) - f(y)^x \geq 0.$

$\forall x$

$-\nabla f(x) \cdot (x - y) \geq 0$

"gradient good choice."

GD algorithm



x^*

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

THM.

f convex (on K).

$$\boxed{f\left(\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \text{range} + \text{variance}.}$$

$$\| x_t - \eta \nabla f(x_t) - x^* \|^2 - \| x_t - x^* \|^2 = \| a - b \|^2 - \| b \|^2 = -2 a \cdot b + \| a \|^2.$$

$$= -2 (x_t - x^*) \cdot (\eta \nabla f(x_t)) + \eta^2 \| \nabla f(x_t) \|^2.$$

$f(x_t) - f(x^*) \leq \nabla f(x_t)(x_t - x^*)$

$$\leq -2\eta\, (x_t - x^*) \nabla f(x_t) \underbrace{\leq}_{f \text{ convex}} 2 \left(f(x^*) - f(x_t)\right) \eta = -2 \delta_t \eta \quad \text{current loss}$$

$$\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{1}{2\eta} \sum_{t=1}^{T} \left(\| x_{t+1} - x^* \|^2 + \| x_t - x^* \|_2^2 \right).$$

$$\leq \text{Jensen Ineq.} = \frac{1}{2\eta} \| x_1 - x^* \|_2^2 - \| x_{T+1} - x^* \|^2$$

$f\left(\frac{1}{T} \sum x_t\right) - f(x^*)$

$$= \frac{\| x_1 - x^* \|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \| \nabla f(x_t) \|^2.$$

$x_{t+1} = \arg\min_x \{ \eta g_t \cdot x + \|x - x_t\|^2 \}$ fenchel dual $|2.5\times 10^{-3} - 50\times 10^{-3}|$

$\partial \|\cdot - x_t\|^2 [\ldots] = x_t - \eta g_t$

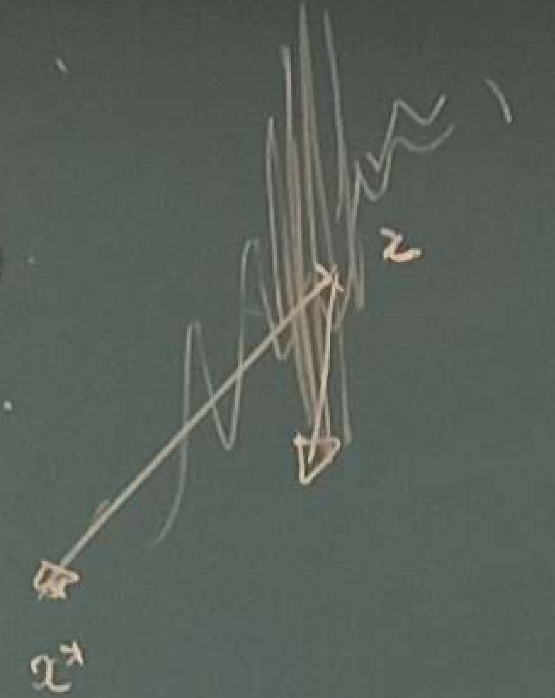
1) vitesse $\frac{1}{\sqrt{T}}$ / $O(1/\varepsilon^2)$ / $\sqrt{dT}$ regret.

Rm

2) $\arg\min_{x \in \mathbb{R}^p} \{ \eta g_t \cdot x + \frac{\|x - x_t\|_2^2}{2} \}$

$\eta g_t + (x^* - x_t) = 0$

$x^* = x_t - \eta g_t$



3) THM valable pour tout $g_t \in \mathbb{R}^p$, $t = 1, \dots, T$.

$x_{t+1} = x_t - \eta g_t$. $\quad \forall y, \quad \sum_{t=1}^T g_t \cdot (x_t - y) \le \mathbb{E}\left(\frac{1}{2\eta} \|x_1 - x^*\|^2 + \eta \sum_{t=1}^T \|g_t\|^2 \right)$

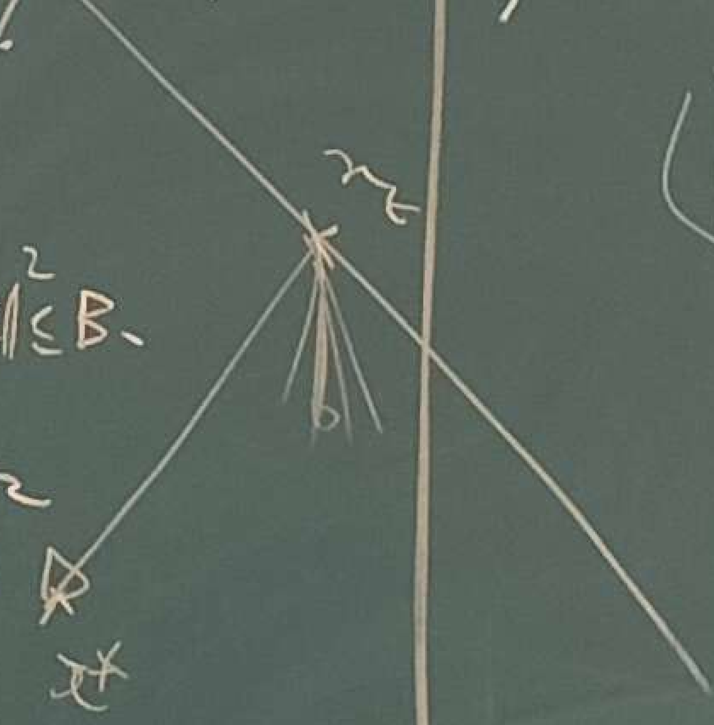
$\eta B^2 T$

ROBUSTNESS!

$\mathbb{E}\|g_t\|^2 \le B^2$ $\quad \{\mathbb{E} \le \mathbb{E}\}$

Soit g_t stochastic. $\mathbb{E}_g(g_t | x_t) = \nabla f(x_t)$ $\quad (g_t = \nabla f(x_t) + \varepsilon)$.

$\le \mathbb{E}\|\varepsilon\|^2 \le B$.



$$\mathbb{E}_g \sum_{t=1}^T (f(x_t) - f(x^*)) \le \frac{1}{2\eta}\|x_1 - x^*\|^2 + \eta \sum_{t=1}^T \|\nabla f(x_t)\|^2 \sigma^2$$

$$\le \mathbb{E}_g \sum_{t=1}^T \nabla f(x_t)(x_t - x^*) = \mathbb{E}_g \sum_{t=1}^T \mathbb{E}(g_t | x_t)(x_t - x^*)$$

$$= \mathbb{E}_g \sum_t g_t (x_t - x^*) \le$$

Stochastic GD

$\boxed{\theta_{t+1} = \theta_t - \eta \nabla_\theta \ell(x_t, \theta)}$ $\quad R(\hat\theta) = \mathbb{E}_P \ell(\theta, x)$

one-pass SGD. $\quad \theta \in \mathbb{R}^p$, $\quad x_1 \dots x_n$ iid $\sim P$ unknown.

$g_t = \nabla_\theta \ell(x_t, \theta) \leftrightarrow \frac{1}{n} \sum_{i=1}^n \nabla_\theta \ell(x_i, \theta)$ $\quad \theta^* = \min \mathbb{E}_P \ell(\theta, x)$

$\mathbb{E}_P(g_t | \text{past}) = \mathbb{E}_P \nabla_\theta \ell(X, \theta)$ $\quad \nabla_\theta \mathbb{E}_P \ell(\theta, X)$

int deg

$\Rightarrow \mathbb{E}_P \frac{1}{T} \sum f(x_t) - f(x^*) \le$ $\quad P$ $\quad$ variance est.

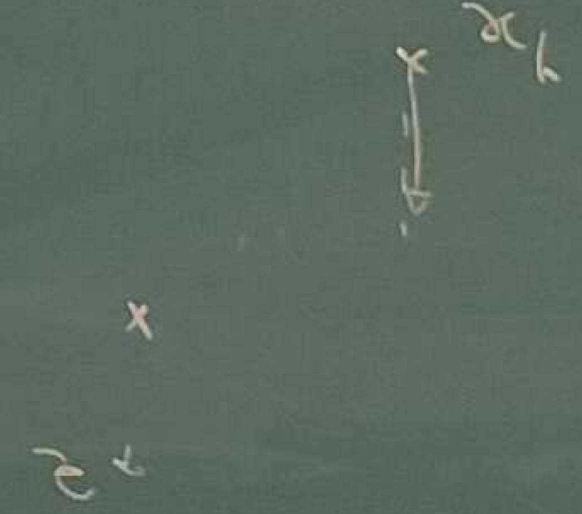
"

$\mathbb{E}_P(\ell(\theta, x) - \ell(\theta^*, x)) \le \frac{\|\theta_1 - \theta^*\|^2}{\eta} + \eta B^2 T$

(!) epochs "multi pass" SGD.

$\theta_{t+1} = \theta_t - \eta \nabla_\theta \ell(x_{(t)}, \theta)$, $\quad x_{(t)} \sim \mathcal{U}\{x_1, \dots, x_T\}$.

θ?



$x_{t+1} = \arg\min \{ \eta g_t x + \|x - x_t\|^2 \}$ Euclidean $|2.5\times 10^{-3} - 50 \times 10^{-3}|$

$\|x - x_t\|^2 \, [\eta g_t] = x_t - \eta g_t$

1a/ $x_{t+1} = \arg\min \left\{ \eta g_t \cdot x + \frac{\|x_t - x\|^2}{2} \right\}$

no "localisation"

no geometry

no preconditioning

no a priori

2c $\phi(x) = \frac{\|x\|^2}{2}$

$\nabla\phi(x) = x$

$\nabla^2\phi(x) = Id$ 2.d

2a $x_{t+1} = \arg\min_x \left\{ \eta g_t x + \frac{D_\phi(x, x_t)}{2} \right\}$

2.b $D_\phi(x, x_t) = \phi(x) - \phi(x_t) - \nabla\phi(x_t)(x - x_t) = \|x_t - x\|^2_{x_t}$ 2.e

ϕ potential

$(x_t - x)\nabla^2\phi(x_t)(x_t - x)^T$.

$\phi(\cdot) - \nabla\phi(\cdot)(x_t - x)$

$-\frac{\|x_t\|^2}{2} + \frac{\|x\|^2}{2} - x_t(x - x_t)$

$= \frac{\|x\|^2}{2} + \frac{\|x_t\|^2}{2} - x \cdot x_t$

$= \frac{\|x - x_t\|^2}{2}$

$\nabla_x \{ \eta g_t x + D_\phi(x, x_t) \} = 0$.

$\eta g_t + \nabla\phi(x) - \nabla\phi(x_t) = 0$

Mirror descent algorithm

$\nabla\phi(x_{t+1}) = \nabla\phi(x_t) - \eta g_t$ dual

$x_{t+1} = x_t - \eta g_t$. primal

gradient descent.

1) $x + \nabla f(x)$

$\boxed{\text{THM}}$

$$\sum_{t=1}^{T} f(x_t) - f(x^*) \le \underbrace{\frac{D_\phi(x_1, x^*)}{\eta}}_{\text{'range'}} + \underbrace{\eta \sum_{t=1}^{T} D_{\phi^*}\big(\nabla\phi(x_t) - \eta g_t, \nabla\phi(x_t)\big)}_{\text{"variance"}}$$

drop = $D_\phi(x_{t+1}, y) - D_\phi(x_t, y)$

+ telescoping term

$\max_{x: \|x\| \le 1} \langle g_t, x \rangle_{\omega_t} = g_t \nabla^2\phi(x_t)^{-1} g_t = \|g_t\|^2_{\omega_t, *}$

Rmq. link with PEA.

$\{x_t, \dots\}$ $\quad \{e_{k,t},\ k \in [N],\ t \in [T]\}$.

$\boxed{\text{THM}}$ $\displaystyle \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \min_k \sum_{t=1}^{T} \ell(e_{k,t}, y_t) \le \sqrt{\frac{T \log N}{2}}$.

$\displaystyle \hat{y}_t = \sum_{k=1}^{N} \hat{\omega}_k e_{k,t}$. $\quad \displaystyle \hat{\omega}_k = \exp\Big(-\sum_{u=1}^{t-1} \ell(e_{k,u}, y_u)\Big) \Big/ W_k$.

$w_{t+1} \sim e^{-\eta g_t} w_t$.

$\displaystyle \min_{x \in \Delta_N} \sum_{t=1}^{T} \ell(x \cdot e, y_t)$

$\Updownarrow$ mirror descent with mirror map $\displaystyle \phi(x) = \sum_{i=1}^{N} x_i \log x_i$ entropy!

$\nabla\phi(x) = \log x$. $\quad \nabla^2\phi(x) = \begin{pmatrix} 1/x_1 & & \\ & \ddots & \\ & & 1/x_N \end{pmatrix}$ $\quad \| \quad e^{-\eta g_t + \log w_t}$

$w_{t+1} = x_{t+1} = \nabla\phi^{-1}\big(\nabla\phi(x_t) - \eta g_t\big)$

$\displaystyle \eta \sum_{t=1}^{T} \|g_t\|^2_{w_t,*} \le \eta \frac{T}{2}$

loss bounded.

$\log N / \eta$

$\le$

$\phi(w_1) - \phi(x^*)$

D

$\boxed{\text{THM}}$

'range'

$\displaystyle \sum_{t=1}^{T} f(x_t) - f(x^*) \le \frac{D_\phi(x_1, x^*)}{\eta} + \eta \sum_{t=1}^{T} D_{\phi^*}\big(\nabla\phi(x_t) - \eta g_t, \nabla\phi(x_t)\big)$

"variance"

$\mathrm{op} = D_\phi(x_{t+1}, y) - D_\phi(x_t, y)$

telescoping term

$\displaystyle \max_{x : \|x\| \le 1} \langle g_t, x \rangle_{w_t} = g_t^\top \nabla^2\phi(w_t)^{-1} g_t = \|g_t\|^2_{w_t,*}$