

Highly-Smooth Zero-th Order Stochastic Optimization

Vianney Perchet (with F. Bach, INRIA-ENS)

PGMO Days

Paris, November 9 2016

ENS Paris-Saclay

Model, Motivations & Objectives

Motivations 1/2

- Final step in learning, statistics... minimization of loss $f(x)$

Structural assumption: **Convex optimization**

- Regularity of loss may vary
 - “**Non-smooth**”, with “kinks” (hinge loss)
 - “**smooth**”, if bounded second derivative
 - $f(x) = \log(1 + \exp(a^\top x))$, $f(x) = \|x - x^*\|_H^2$, much smoother

Regularity assumption: **High smoothness**

2nd, 3rd, ... derivatives bounded

- Noise in the data (i.i.d.) or in the program

Noise assumption: **Stochastic optimization**

$$f(\hat{\theta}) = \mathbb{E}\|\theta - \hat{\theta}\|^2, \text{ get to observe/compute } \|\theta - \hat{\theta}\|^2$$

Objectives

Stochastic Optimization of a convex function

- Unknown Mapping $f: \mathbb{R}^d \rightarrow \mathbb{R}$, constraint set $\mathcal{X} \subset \mathbb{R}^d$
- i) Query $x_1 \in \mathbb{R}^d$, feedback $f(x_1) + \xi_1$ (0-th order), $\xi_1 \sim \mathcal{N}(0, \sigma^2)$
computing gradients impossible/too expensive in high dimension
- i bis) Output $x_2^* \in \mathcal{X}$ the guessed “minimum” of f
- ii) Query $x_2 \in \mathbb{R}^d$, get $f(x_2) + \xi_2$, output $x_3^* \in \mathcal{X}$, etc.

Performance of an algorithm after T steps: $f(x_{T+1}^*) - f^*$
with $f^* = \min_{x \in \mathcal{X}} f(x) = f(x^*)$

Our Assumptions

Assumptions: strong convexity & smoothness

Strongly convex: Intuitively, $f''(x) \geq \mu$ or $\nabla^2(f)(x) \succeq \mu Id$

f is μ -strongly convex iff

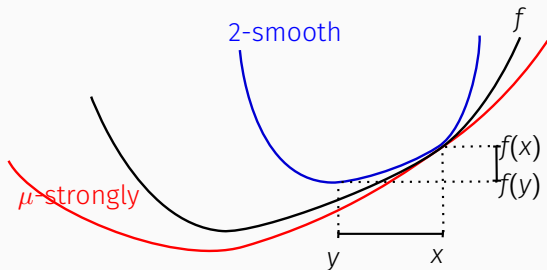
$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2} \|x - y\|^2$$

- Any convex mapping is 0-strongly convex
- f has “no flat part” (linear or 3rd order)

f is 2-smooth iff

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{M_2^2}{2} \|x - y\|^2$$

Picture strong cvxity & smoothness



strong cvx quadratic **lower-bound**
smooth quadratic **upper-bound**

High β -smoothness

$$\text{2-smooth} \quad \left| f(y) - \text{Taylor}_1[f](x) \right| \leq \frac{M_2^2}{2!} \|y - x\|^2$$

$$\beta\text{-smooth} \quad \left| f(y) - \text{Taylor}_{\beta-1}[f](x) \right| \leq \frac{M_\beta^\beta}{\beta!} \|y - x\|^\beta$$

f is β -smooth iff

$$\left| f(y) - \sum_{|m| \leq \beta-1} \frac{1}{m!} f^{(m)}(x) (y-x)^m \right| \leq \frac{M_\beta^\beta}{\beta!} \|y - x\|^\beta$$

$$\text{with } f^{(m)}(x)(y-x)^m = \frac{\partial^{m_1+\dots+m_d}}{\partial^{m_1} \dots \partial^{m_d}} f(x) (y_1 - x_1)^{m_1} \dots (y_d - x_d)^{m_d}$$

On the high-regularity

$$\left| f(y) - \sum_{|m| \leq \beta-1} \frac{1}{m!} f^{(m)}(x) (y-x)^m \right| \leq \frac{M_\beta^\beta}{\beta!} \|y-x\|^\beta$$

- 0-smooth = bounded (by M_0 if \mathcal{X} compact)
- 1-smooth = M_1 -Lipschitz

Lemma

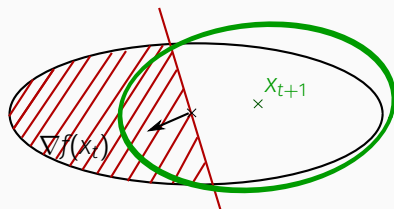
f is β_1 - and β_2 -smooth $\Rightarrow \beta$ -smooth for all $\beta \in [\beta_1, \beta_2]$

$$\alpha_\beta M_\beta^\beta \leq 2(\alpha_{\beta_1} M_{\beta_1}^{\beta_1})^{\frac{\beta_2-\beta}{\beta_2-\beta_1}} (\alpha_{\beta_2} M_{\beta_2}^{\beta_2})^{\frac{\beta-\beta_1}{\beta_2-\beta_1}}, \text{ and } \alpha_k \text{ fct only of } k$$

- **Logistic Regression** $f(x) = \mathbb{E}_a \log(1 + \exp(-a^\top x))$, random a .
 $\|a\| \leq R \Rightarrow f$ is ∞ -smooth and $M_\beta \leq \beta R$
- **Quadratic Loss** $f(x) = \mathbb{E}_a \|x - a\|^2$ for random a , ∞ -smooth.

Classic Optimization Methods

Optim without noise: Ellipsoid method



- Decrease the volumes of Ellipsoids (by a constant factor $\exp(-\frac{1}{2d})$)
- Exponential decay [Yudin & Nemirovski]

$$\min_t f(x_t) - f^* \leq O\left(M_0 R \exp\left(-\frac{1}{2} \frac{T}{d^2}\right)\right)$$

Gradient (1st order) methods

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

1-smooth $f(\bar{x}_T) - f^* \leq \frac{1}{\sqrt{T}} R M_1$, with $\eta = \frac{\|\mathcal{X}\|}{M_0 \sqrt{T}}$

2-smooth $f(x_T) - f^* \leq \frac{1}{T} \|x_1 - x^*\|^2 M_2^2$, with $\eta = \frac{1}{M_2^2}$

with acceleration $f(x_T) - f^* \leq \frac{1}{T^2} \|x_1 - x^*\|^2 M_2^2$

1-smth + strg $f(\hat{x}_T) - f^* \leq \frac{1}{\mu T} 2M_1^2$, with $\eta = \frac{2}{\mu(t+1)}$

2-smth + strg $f(x_T) - f^* \leq \left(\frac{\kappa-1}{\kappa+1}\right)^{2(T-1)} M_2^2 \|x_1 - x^*\|^2$,

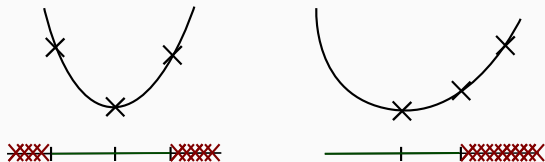
where $\kappa = M_2^2/\mu$ is the condition number

0th-order methods

- Using $2d$ queries to get $\nabla f(\cdot)$, same rate except $t \rightsquigarrow \frac{t}{2d}$

$$\frac{f(x + \delta \vec{u}) - f(x - \delta \vec{u})}{2\delta} \simeq \nabla f(x)^\top \vec{u}$$

- True 0-th order method (“Pyramids” in higher dimensions)



Convergence: $O\left(d \exp\left(-\frac{T}{d^2}\right)\right)$ [Nemirovski-Yudin]

From 1st to 0th order: Rates multiplied by d

With noise, $\nabla f(x_t) + \vec{\xi}_t$

- Cutting algo (ellipsoid, pyramids...)
Query d times the same points (lower the variance)
- Stochastic gradient

$$x_{t+1} = x_t - \eta \left(\nabla f(x_t) + \vec{\xi}_t \right), \mathbb{E} \|\xi_t\|^2 \simeq d\sigma^2$$

non-strongly rates $O\left(\frac{1}{\sqrt{T}}(M_1 R + \sqrt{d}\sigma)\right) = O\left(\sqrt{\frac{d}{T}}\right)$

μ -strongly rates $O\left(\frac{1}{\mu T}(M_1^2 + d\sigma^2)\right) = O\left(\frac{d}{\mu T}\right)$

Without noise to noisy: Lose a factor d

[Shamir, Tang], [Hazan et al], [Bach, Moulines], [Bach et al]...

Sum up - Objectives

Principal rates (without smoothness for the 1st column)

| | 1st | noisy | 0th | 0/noise | Our results |
|-------------|----------------------|----------------------|----------------------|--------------------------|---|
| cvx | $\sqrt{\frac{1}{T}}$ | $\sqrt{\frac{d}{T}}$ | $\sqrt{\frac{d}{T}}$ | $\sqrt{\frac{d^2}{T}} ?$ | $(\sqrt{\frac{d^2}{T}})^{\frac{\beta-1}{\beta}}$ & $(\sqrt{\frac{d^2}{T}})^{\frac{\beta}{\beta+1}}, \beta \leq 2$ |
| μ -strg | $\frac{1}{\mu T}$ | $\frac{d}{\mu T}$ | $\frac{d}{\mu T}$ | $\frac{d^2}{\mu T} ?$ | $(\frac{d^2}{\mu T})^{\frac{\beta-1}{\beta+1}}$ or $\frac{1}{\mu^2} (\frac{d^2}{T})^{\frac{\beta-1}{\beta}}$ |

Existing results

- Minimax speed (not strg) $\frac{\text{poly}(d)}{\sqrt{T}}$ [Bubeck-Eldan],[Rakhlin et al.]
- Strongly AND smooth, algo $\sqrt{\frac{d^2}{\mu T}}$ [Hazan, Levy]
- Strongly OR smooth, algo $T^{-1/3}$ [Agarwal et al][Saha,Tewari]
- Only Convex $T^{-1/4}$ [Flaxman et al][Kleinberg]
- Only Convex $\frac{\log^d(T)}{\sqrt{T}}$ [Yu-Hazan]
- Strongly and β -smooth, $T^{-\frac{\beta-1}{\beta}}$ [Polyak,Tsybakov]

Tricks and Algorithms

The tricks 1/2: Stoch. Gradient

- Build an estimate of $f'(x)$ based on the values of $f(x) + \xi$

$$f'(x) \simeq \frac{f(x+\delta) - f(x-\delta)}{2\delta} = f'_\delta(x), f_\delta(x) := \frac{1}{2} \int_{-1}^1 f(x + \delta v) dv = \mathbb{E}_{|v| \leq 1} f(x + \delta v)$$

- Draw $\varepsilon = \pm 1$ with proba 1/2, $g(z) = \frac{f(x + \varepsilon \delta)}{\delta} \varepsilon$

Unbiased: $\mathbb{E}_\varepsilon[g(z)] = f'_\delta(x)$, $\mathbb{E}_\varepsilon[g^2(z)] \leq \frac{1}{\delta^2}$

- Stochastic gradient descent (w.r.t. f_δ which **is convex**):

$$f_\delta(x_T^*) - f_\delta(x_\delta^*) \lesssim \frac{1}{\delta\sqrt{T}}, \text{ so } f(x_T^*) - f(x^*) \lesssim \frac{1}{\delta\sqrt{T}} + \delta M_1 \lesssim \frac{1}{T^{1/4}}$$

- In higher dimension

$$\mathbb{E}_{\|u\|=1} \frac{d}{\delta} f(x + \delta u) u = \nabla f_{\delta}(x) \text{ with } f_{\delta}(x) = \mathbb{E}_{\|v\| \leq 1} f(x + \delta v)$$

$$- \mathbb{E} \|g(\xi)\|^2 \leq \frac{d^2}{\delta^2} \text{ and } |f_{\delta}(x) - f(x)| \leq M_1 \delta$$

$$f(x_T^*) - f^* \lesssim \frac{d}{\delta \sqrt{T}} + \delta \lesssim \left(\frac{d^2}{T}\right)^{1/4} \quad [\text{Nemirovski-Yudin, Flaxman et al, Hazan et al.}]$$

The tricks 2/2: Kernels

- β -regularity on f (in 1 dimension)

$$\left| f(x+r) - \sum_{m=0}^{\beta-1} \frac{r^m}{m!} f^{(m)}(x) \right| \leq \frac{M_\beta^\beta}{\beta!} r^\beta$$

- $k(\cdot) : [-1, 1] \rightarrow \mathbb{R}$ such that
 - $\int_{-1}^1 r k(r) dr = 1$
 - $\int_{-1}^1 r^m k(r) dr = 0$ for all $m \in \{0, 2, \dots, \beta\}$

$$\left| \int_{-1}^1 f(x+r\delta) r k(r) dr - f(x) \right| \leq \frac{\delta^\beta M_\beta^\beta}{\beta!} \int_{-1}^1 |k(r) r^{\beta+1}| dr$$

- Explicit forms for $k(\cdot) = k_\beta(\cdot)$ (Legendre Polynomial)
 - $k_1(r) = k_2(r) = 3r$
 - $k_3(r) = k_4(r) = \frac{15r}{4}(5 - 7r^3)$
 - $k_5(r) = k_6(r) = \frac{195r}{64}(99r^4 - 126r^2 + 35)$

Both tricks combined “smoothing”

- Smoothened β -smooth mapping f (with unbiased estimate)

$$f_{r,\delta}(x) = \mathbb{E}_r \mathbb{E}_{\|v\| \leq 1} f(x + r\delta v) k(r)$$

- $\nabla f_{r,\delta}(x) = \mathbb{E}_r \mathbb{E}_{\|u\|=1} \frac{d}{\delta} f(x + r\delta u) k(r) u$

$$|f_{r,\delta}(x) - f(x)| \leq \frac{M_\beta^\beta}{\beta!} \delta^\beta \mathbb{E}_r |k(r) r^{\beta+1}|$$

$$|\nabla f_{r,\delta}(x) - \nabla f(x)| \leq \frac{M_\beta^\beta}{\beta-1!} \delta^{\beta-1} \mathbb{E}_r |k(r) r^{\beta+1}|$$

- $\mathbb{E}_r |k(r)|^2 \leq 3\beta^3$
 $\mathbb{E}_r |k(r)|^2 r^2 \leq 8\beta^2$
 $\mathbb{E}_r |k(r) r^{\beta+1}| \leq 2\sqrt{2}\beta$
- $f_{r,\delta}$ is $\frac{\mu}{2}$ -strongly-convex if f is μ -strongly convex (and δ small)
 $f_{r,\delta}$ is convex if f convex and $\beta \leq 2$

Two meta-algorithms

Constrained \mathcal{X} compact, one query $f(\xi) + \varepsilon$ per iteration

One Point algorithm

$$x_t = \Pi_{\mathcal{X}} \left(x_{t-1} - \gamma_t \frac{d}{\delta_t} \left[f(x_{t-1} + \delta_t r_t u_t) + \varepsilon_t \right] k(r_t) u_t \right)$$

$$r_t \sim \mathcal{U}([-1, 1]), u_t \sim \mathcal{U}(\mathbb{S}^d)$$

γ_t, δ_t are deterministic sequences

Unconstrained $\mathcal{X} = \mathbb{R}^d$, $f(\xi_t^1)$ and $f(\xi_t^2)$ (independent noises)

Two points algorithm

$$x_t = x_{t-1} - \gamma_t \frac{d}{2\delta_t} \left[f(x_{t-1} + \delta_t r_t u_t) - f(x_{t-1} - \delta_t r_t u_t) + \varepsilon_t \right] k(r_t) u_t$$

μ -Strongly convex + Constrained

- 1-point meta-algo, f is β -smooth for $\beta \geq 2$

$$x_t = \Pi_{\mathcal{X}} \left(x_{t-1} - \gamma_t \frac{d}{\delta_t} \left[f(x_{t-1} + \delta_t r_t u_t) + \varepsilon_t \right] k(r_t) u_t \right)$$

- Choice of parameters
 - $\gamma_t = \frac{1}{\mu t}$ (classic choice for μ -strongly)
 - $\delta_t = \left(2 \frac{d^2 \beta^2 \beta!}{t \mu M_{\beta}^{\beta}} \right)^{\frac{1}{\beta+2}}$ (remark $\beta!^{1/\beta} \sim \beta/e$)
- Output. Averaging $x_T^* = \frac{1}{T} \sum_{t=1}^T x_t$
- Convergence guarantee (leading term)

$$\mathbb{E}f(x_T^*) - f^* \leq 12\beta^2 \left(\frac{2d^2 M_{\beta}^2}{\mu T} \right)^{\frac{\beta-1}{\beta+1}} ((M_0 + M_1)^2 + \sigma^2 + 1)$$

The Proofs!!

Sketch of proof - 6 steps

1) Definition of x_t by algo

$$\begin{aligned}\|x_t - x\|^2 &\leq \|x_{t-1} - x\|^2 - 2\gamma_t \frac{d}{\delta_t} [f(x_{t-1} + \delta_t r_t u_t) + \varepsilon_t] k(r_t) u_t^\top (x_{t-1} - x) \\ &\quad + 2|k(r_t)|^2 \frac{\gamma_t^2 d^2}{\delta_t^2} [|f(x_{t-1} + \delta_t r_t u_t)|^2 + |\varepsilon_t|^2]\end{aligned}$$

2) $\mu/2$ -strong convexity of f_δ

$$\mathbb{E}\|x_t - x\|^2 \leq \mathbb{E}\|x_{t-1} - x\|^2 - 2\gamma_t \left(\mathbb{E}f_{\delta_t}(x_{t-1}) - \mathbb{E}f_{\delta_t}(x) + \frac{\mu}{4} \mathbb{E}\|x_{t-1} - x\|^2 \right) + \frac{\gamma_t^2 d^2}{\delta_t^2} C$$

3) Rearranging $\mathbb{E}f_{\delta_t}(x_{t-1}) - \mathbb{E}f_{\delta_t}(x) \leq \mathbb{E}\|x_{t-1} - x\|^2 \left(\frac{1}{2\gamma_t} - \frac{\mu}{2} \right) - \mathbb{E}\|x_t - x\|^2 \frac{1}{2\gamma_t} + \frac{\gamma_t d^2}{\delta_t^2} C$

4) The choice of $\gamma_t = 1/\mu t$

$$\mathbb{E}f_{\delta_t}(x_{t-1}) - \mathbb{E}f_{\delta_t}(x) \leq \frac{(t-1)\mu}{2} \mathbb{E}\|x_{t-1} - x\|^2 - \frac{t\mu}{2} \mathbb{E}\|x_t - x\|^2 + \frac{d^2}{t\mu\delta_t^2} C$$

5) Summing over t and averaging

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[f_{\delta_t}(x_{t-1}) - f_{\delta_t}(x)] + \frac{\mu}{2} \mathbb{E}\|x_T - x\|^2 \leq \frac{d^2}{\mu} \frac{1}{T} \sum_{t=1}^T \frac{1}{t\delta_t^2} C$$

6) Plugging back f and balancing

$$\mathbb{E}f(x_T^*) - f^* \lesssim \frac{d^2}{\mu} \frac{1}{T} \sum_{t=1}^T \frac{1}{t\delta_t^2} + \frac{1}{T} \sum_{t=1}^T \delta_t^\beta \lesssim \frac{1}{T} \sum_{t=1}^T \left(\frac{d^2}{\mu t} \right)^{\frac{\beta}{\beta+2}} \simeq \left(\frac{d^2}{\mu T} \right)^{\frac{\beta}{\beta+2}}$$

Strongly convex + Unconstrained

- 2-point meta-algo, f is β -smooth for $\beta \geq 2$

$$x_t = x_{t-1} - \gamma_t \frac{d}{2\delta_t} \left[f(x_{t-1} + \delta_t r_t u_t) - f(x_{t-1} - \delta_t r_t u_t) + \varepsilon_t \right] k(r_t) u_t$$

- Choice of parameters
 - $\gamma_t = \frac{2}{\mu(t+1)}$ (classic choice for μ -strongly & 2-smooth)
 - $\delta_t = \delta = \left(2 \frac{d^2 \beta^2 \beta!}{t \mu M_\beta^\beta} \right)^{\frac{1}{\beta+2}}$ (constant step size)
- Output. Uniform averaging $x_T^* = \frac{1}{T} \sum_{t=1}^T x_t$
- Convergence guarantee (leading term)

$$\mathbb{E}f(x_T^*) - f^* \leq 27\beta \left(\frac{2d^2 M_\beta^2}{\mu T} \right)^{\frac{\beta}{\beta+2}} (\sigma^2 + 2)$$

Remarks on strongly convex

- Constrained & unconstrained, same rates $\left(\frac{2d^2M_\beta^2}{\mu T}\right)^{\frac{\beta}{\beta+2}}$
- In the proof, Step 5 yields

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} [f_{\delta_t}(x_{t-1}) - f_{\delta_t}(x)] + \frac{\mu}{2} \mathbb{E} \|x_T - x\|^2 \lesssim \frac{d^2}{\mu} \frac{1}{T} \sum_{t=1}^T \frac{1}{t\delta_t^2}$$

- Using this for $\frac{\mu}{2} \mathbb{E} \|x_T - x\|^2$, improve rates into

$$\mathbb{E} \|x_T - x^*\|^2 \leq 2\mathbb{E} \|x_T - x_\delta^*\|^2 + 2\mathbb{E} \|x_\delta^* - x^*\|^2 \lesssim \frac{1}{\mu^2} \left(\frac{d^2}{T}\right)^{\frac{\beta-1}{\beta}}$$

Convex + constrained

- 1-point meta-algo, f is β -smooth for $\beta > 2$

$$x_t = \Pi_{\mathcal{X}} \left(x_{t-1} - \gamma_t \frac{d}{\delta_t} \left[f(x_{t-1} + \delta_t r_t u_t) + \varepsilon_t \right] k(r_t) u_t \right)$$

- Choice of parameters
 - $\gamma_t = \frac{\delta_t R}{\beta^{3/2} d \sqrt{t}}$ (classic choice for non-strongly)
 - $\delta_t = \left(\frac{dR\sqrt{\beta}\beta!}{\sqrt{t}M_\beta^\beta} \right)^{\frac{1}{\beta}}$
- Output. Averaging $x_T^* = \frac{1}{T} \sum_{t=1}^T x_t$
- Convergence guarantee (leading term)

$$\mathbb{E}f(x_T^*) - f^* \leq 19\sqrt{\beta}\beta^2 \left(\sqrt{\frac{d^2 M_\beta^2 R^2}{T}} \right)^{\frac{\beta-1}{\beta}} ((M_0 + M_1)^2 + \sigma^2 + 1)$$

Remarks on convex, $\beta = 2$

- Rates for $\beta > 2$, $\mathbb{E}f(x_T^*) - f^* \lesssim \left(\sqrt{\frac{d^2}{T}}\right)^{\frac{\beta-1}{\beta}}$
- For $\beta = 2$, we can get $\mathbb{E}f(x_T^*) - f^* \lesssim \left(\sqrt{\frac{d^2}{T}}\right)^{\frac{\beta}{\beta+1}} = \left(\frac{d^2}{T}\right)^{\frac{1}{3}}$

The mapping f_δ is convex for $\beta = 2$, **not necessarily** $\beta > 2$.

- The first lines of proof are modified
- Instead stoch. gradient **descent of f_δ** + **approximation of f by f_δ**
- **Approximation of $\nabla f(x)$ by $\nabla f_\delta(x)$** + stoch. grad **descent of f** .
- Loose 1 in the regularity number ($\beta \rightsquigarrow \beta - 1$)

Convex + Unconstrained

- 2-point meta-algo, f is β -smooth for $\beta \geq 2$

$$x_t = x_{t-1} - \gamma_t \frac{d}{2\delta_t} \left[f(x_{t-1} + \delta_t r_t u_t) - f(x_{t-1} - \delta_t r_t u_t) + \varepsilon_t \right] k(r_t) u_t$$

- Choice of parameters
 - $\gamma_t = \frac{\delta}{\beta^2 d M_2 \sqrt{T}}$ (classic choice for μ -strongly & 2-smooth)
 - $\delta_t = \delta = \left(\frac{d \beta \beta!}{\sqrt{t} M_{\beta}^{\beta} M_2} \right)^{\frac{1}{\beta}}$ (constant step size)
- Output. Non-uniform averaging $x_T^* = \frac{2}{T(T+1)} \sum_{t=1}^T t x_t$
- Convergence guarantee (leading term)

$$\mathbb{E}f(x_T^*) - f^* \leq \beta \left(\sqrt{\frac{2d^2 M_{\beta}^2}{T M_2^2}} \right)^{\frac{\beta-1}{\beta}} (96M_2^2 \|x_0 - x^*\|^2 + \sigma^2 + 20)$$

Sum up - Objectives

Principal rates (without smoothness for the 1st column)

| | 1st | noisy | 0th | 0/noise | Our results |
|-------------|----------------------|----------------------|----------------------|--------------------------|---|
| cvx | $\sqrt{\frac{1}{T}}$ | $\sqrt{\frac{d}{T}}$ | $\sqrt{\frac{d}{T}}$ | $\sqrt{\frac{d^2}{T}} ?$ | $\left(\sqrt{\frac{d^2}{T}}\right)^{\frac{\beta-1}{\beta}}$ & $\left(\sqrt{\frac{d^2}{T}}\right)^{\frac{\beta}{\beta+1}}, \beta \leq 2$ |
| μ -strg | $\frac{1}{\mu T}$ | $\frac{d}{\mu T}$ | $\frac{d}{\mu T}$ | $\frac{d^2}{\mu T} ?$ | $\left(\frac{d^2}{\mu T}\right)^{\frac{\beta-1}{\beta+1}}$ or $\frac{1}{\mu^2} \left(\frac{d^2}{T}\right)^{\frac{\beta}{\beta+1}}$ |

Existing results

- Minimax speed (not strg) $\frac{\text{poly}(d)}{\sqrt{T}}$ [Bubeck-Eldan],[Rakhlin et al.]
- Strongly AND smooth, algo $\sqrt{\frac{d^2}{\mu T}}$ [Hazan, Levy]
- Strongly OR smooth, algo $T^{-1/3}$ [Agarwal et al][Saha,Tewari]
- Only Convex $T^{-1/4}$ [Flaxman et al][Kleinberg]
- Strongly and β -smooth, $T^{-\frac{\beta}{\beta+1}}$ [Polyak,Tsybakov]

Online Optimization

Stochastic Online Optimization of convex functions

- Sequence of mappings $f_t : \mathbb{R}^d \rightarrow \mathbb{R}$, constraint set $\mathcal{X} \subset \mathbb{R}^d$
 - Query $x_1 \in \mathbb{R}^d$, feedback $f_1(x_1) + \xi_1$ (0-th order), $\xi_1 \sim \mathcal{N}(0, \sigma^2)$
 - bis) Output $x_2^* \in \mathcal{X}$ the guessed “minimum” of f_2
 - ii) Query $x_2 \in \mathbb{R}^d$, get $f_2(x_2) + \xi_2$, output $x_3^* \in \mathcal{X}$, etc.

Performance of an algorithm after T steps,

$$\text{“Regret”}: \frac{1}{T} \sum_{t=1}^T f_t(x_t^*) - \min_{x^* \in \mathcal{X}} \frac{1}{T} \sum_{t=1}^T f_t(x^*)$$

Not bandit ! $x_k^* \neq x_k$ and $x_k \notin \mathcal{X}$

- Remember Step 5 of the proof in off-line

5) Summing over t and averaging

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[f_{\delta_t}(x_{t-1}) - f_{\delta_t}(x)] + \frac{\mu}{2} \mathbb{E}\|x_T - x\|^2 \leq \frac{d^2}{\mu} \frac{1}{T} \sum_{t=1}^T \frac{1}{t\delta_t^2} C$$

The first term is the regret !

All results hold in online optimization !
with an additional $\log(T)$ factor sometimes

- In bandit learning, x_t^* must be equal to x_t (the query point),
 - Theoretical results $\text{poly}(d)/\sqrt{n}$ for convex mappings
 - Algo: \sqrt{T} for $\beta = 2$ and strongly convex
 - Algo: $T^{-1/4}$ for convex, $T^{-1/3}$ for smooth or strongly convex