The Stochastic Polyak Stepsize

A fraudulent but interesting algorithm

Guillaume Garrigos

18 Février 2025 - Séminaire MILES

Université Paris Cité

A work (in progress) in collaboration with

Robert M. Gower
Flatiron Institute

K. Mishchenko
Meta

Nicolas Loizou
Johns Hopkins

Fabian Schaipp
Inria

The problem, the algorithm

Let $f_i : \mathbb{R}^N \to \mathbb{R}$ be convex, and minimize

$$\min_{x \in \mathbb{R}^N} f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x).$$

with the Stochastic Gradient Descent (SGD) algorithm

$$x_{t+1} = x_t - \gamma_t \nabla f_{i_t}(x_t), \quad \gamma_t > 0, \quad i_t \sim \mathcal{U}(1, \dots, m)$$

The problem, the algorithm

Let $f_i : \mathbb{R}^N \to \mathbb{R}$ be convex, and minimize

$$\min_{x \in \mathbb{R}^N} f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x).$$

with the Stochastic Gradient Descent (SGD) algorithm

$$x_{t+1} = x_t - \gamma_t \nabla f_{i_t}(x_t), \quad \gamma_t > 0, \quad i_t \sim \mathcal{U}(1, \ldots, m)$$

Rk: You can consider $f(x) = \mathbb{E}_\xi \left[f(\xi, x) \right]$ with $\xi \sim \mathcal{D}$ if you want

Rk: You can do minibatches if you want, the story will remain the same

The problem, the algorithm

Let $f_i : \mathbb{R}^N \to \mathbb{R}$ be convex, and minimize

$$\min_{x \in \mathbb{R}^N} f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x).$$

with the Stochastic Gradient Descent (SGD) algorithm

$$x_{t+1} = x_t - \gamma_t \nabla f_{i_t}(x_t), \quad \gamma_t > 0, \quad i_t \sim \mathcal{U}(1, \dots, m)$$

Rk: You can consider $f(x) = \mathbb{E}_\xi \left[f(\xi, x) \right]$ with $\xi \sim \mathcal{D}$ if you want

Rk: You can do minibatches if you want, the story will remain the same

Goal: How to tune properly the stepsize γ_t?

I : Stochastic Gradient Descent

I : Stochastic Gradient Descent

1 : The smooth case

Smooth case: Known results (1)

Theorem (Constant stepsize)

Let $f_i \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. If $\gamma_t \equiv \gamma \leqslant 1/4L$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{D^2}{\gamma T} + 2\gamma\sigma_*^2,$$

where $D := \|x^0 - x^*\|$ and $\sigma_*^2 := \mathbb{V}[\nabla f_i(x^*)]$ for $x^* \in \operatorname{argmin} f$.

Smooth case: Known results (1)

Theorem (Constant stepsize)

Let $f_i \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. If $\gamma_t \equiv \gamma \leqslant 1/4L$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{D^2}{\gamma T} + 2\gamma\sigma_*^2,$$

where $D := \|x^0 - x^*\|$ and $\sigma_*^2 := \mathbb{V}[\nabla f_i(x^*)]$ for $x^* \in \operatorname{argmin} f$.

- γ_t can go up to $\frac{2}{L}$, requires knowing L
- $\sigma_*^2 = 0$ in the deterministic case (not only!), we recover classic results

Interlude : Interpolation

Definition (Interpolation constants)

- $\sigma_*^2 := \mathbb{V}[\nabla f_i(x^*)]$ for $x^* \in \operatorname{argmin} f$,
- $\Delta_* := \inf f - \mathbb{E}\left[\inf f_i\right]$

Proposition

Assume that the f_i are convex and smooth. Then $\sigma_*^2, \Delta_* \geqslant 0$ and

$$\sigma_*^2 = 0 \;\Leftrightarrow\; \Delta_* = 0 \;\Leftrightarrow\; \bigcap_{i=1}^{m} \operatorname{argmin} f_i \neq \emptyset$$

Interlude : Interpolation

$\sigma_*^2 := \mathbb{V}[\nabla f_i(x^*)]$, $\Delta_* := \inf f - \mathbb{E}\left[\inf f_i\right]$

Example (Linear model)

Suppose that we have a linear model (least squares problem):

$$f_i(x) = \frac{1}{2}\left(\langle \phi_i, x\rangle - y_i\right)^2, \quad f(x) = \frac{1}{2m}\|\Phi x - y\|^2, \quad \Phi = (\phi_i)_i$$

Interpolation means that there is an hyperplane supported by x^* which contains *every data point* $(\phi_i; y_i)_i$. Always true if Φ surjective.

Interlude : Interpolation

$$\sigma_*^2 := \mathbb{V}[\nabla f_i(x^*)],\ \Delta_* := \inf f - \mathbb{E}\left[\inf f_i\right]$$

Example (Neural Networks)

It is shown (Belkin et al.) that Neural Networks with a *very very* large number of parameters interpolate (conditions apply).
This is sometimes observed in *practice*.

Smooth case: Known results (1)

Theorem (Constant stepsize)

Let $f_i \in \Gamma_0(\mathbb{R}^N) \cap C^{1,1}_L(\mathbb{R}^N)$ and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. If $\gamma_t \equiv \gamma \leqslant 1/4L$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{D^2}{\gamma T} + 2\gamma\sigma_*^2,$$

where $D := \|x^0 - x^*\|$ and $\sigma_*^2 := \mathbb{V}[\nabla f_i(x^*)]$ for $x^* \in \operatorname{argmin} f$.

- SGD does *not* converge with constant stepsizes (complexity available)
- $\gamma \propto \frac{1}{\sqrt{T}}$ gives a *finite horizon* rate of $O(\frac{D^2+\sigma_*^2}{\sqrt{T}})$, not optimal
- $\gamma \propto \frac{1}{\sqrt{\sigma_*^2 T+1}}$ gives a better rate $O(\frac{D^2}{T} + \frac{\sigma_*^2}{\sqrt{T}})$ not *adaptive* to σ_*^2

Smooth case: Known results (2)

Theorem (Vanishing stepsize)

Let $f_i \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. If $\gamma_t \propto \frac{1}{\sqrt{t}} \leqslant 1/4L$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant O\left(\frac{D^2}{\sqrt{T}} + \frac{\log(T)}{T}\sigma_*^2\right),$$

where $D := \|x^0 - x^*\|$ and $\sigma_*^2 := \mathbb{V}[\nabla f_i(x^*)]$ for $x^* \in \operatorname{argmin} f$.

- This is an *asymptotic* convergence rate
- Still not optimal if $\sigma_* = 0$

Smooth case: what we really want

Ideally we want

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant O\left(\frac{D^2}{\sqrt{T}} + \frac{1}{T}\sigma_*^2\right)$$

where γ does not need to know σ_*^2. And possibly neither L.

- Adaptivity to L is standard for GD (linesearch) but uncommon for SGD
- Adaptivity to σ_*^2 is not really investigated (?)

I : Stochastic Gradient Descent

2 : The nonsmooth case

Nonsmooth case: Known results

Theorem (Constant stepsize)

Let $f_i \in \Gamma_0(\mathbb{R}^N)$ be G-Lipschitz and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. If $\gamma_t \equiv \gamma$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{D^2}{2\gamma T} + \frac{\gamma G^2}{2}.$$

- No conditions on γ_t
- Remains true if f_i are not differentiable (use subgradients)
- no interpolation story here

Nonsmooth case: Known results

Theorem (Constant stepsize)

Let $f_i \in \Gamma_0(\mathbb{R}^N)$ be G-Lipschitz and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. If $\gamma_t \equiv \gamma$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{D^2}{2\gamma T} + \frac{\gamma G^2}{2}.$$

- $\gamma = \frac{1}{\sqrt{T}}$ gives a finite horizon rate of $\frac{D^2+G^2}{2\sqrt{T}}$
- $\gamma = \frac{D}{G\sqrt{T}}$ gives an *optimal* rate of $\frac{DG}{2\sqrt{T}}$, requires knowing D, G
- Adaptive methods attempt to do this while ignoring D or G

Nonsmooth case: what we really want

Ideally we want to keep

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant O\left(\frac{D^2}{\sqrt{T}} + \frac{1}{T}\sigma_*^2\right)$$

where γ does not need to know D, G.

- Adaptivity to G (knowing D) is achieved e.g. by Adagrad $\gamma_t = \frac{\gamma D}{\sqrt{\sum_s \|g_s\|^2}}$
- Adaptivity to D (knowing G) is achieved with coin-betting (online)
- Interesting recent litterature in the deterministic setting [D-adaptation, DoG, DoWG]
- Not yet mature in the stochastic setting?

II : Stochastic Polyak Stepsize

II : Stochastic Polyak Stepsize

1 : Warm-up : Deterministic Polyak Stepsize

Polyak Stepsize

In the deterministic setting ($m = 1$) Polyak proposed the following rule

$$\gamma_t := \frac{f(x^t) - \inf f}{\|\nabla f(x^t)\|^2}$$

- Updates are scale-invariant: $\gamma_t \nabla f(x_t)$ has no units (Adam, Adagrad, ..)
- We need to know $\inf f$!!
  - In the worst cases, this is as hard as minimizing f
  - In some cases (think interpolation) we know that $\inf f = 0$
  - this is in general *unreasonable*

Polyak Stepsize

In the deterministic setting ($m = 1$) Polyak proposed the following rule

$$\gamma_t := \frac{f(x^t) - \inf f}{\|\nabla f(x^t)\|^2}$$

Theorem (Polyak - 1987 & Hazad, Kakade - 2019)

When using the Polyak stepsize, we can guarantee:

1. $f(x^T) - \inf f \leqslant \frac{2LD^2}{T}$ in the smooth case
2. $f(x^T) - \inf f \leqslant \frac{DG}{\sqrt{T}}$ in the nonsmooth case

Bounds are “optimal” and adaptive to L, D, G!

Polyak Stepsize

In the deterministic setting ($m = 1$) Polyak proposed the following rule

$$\gamma_t := \frac{f(x^t) - \inf f}{\|\nabla f(x^t)\|^2}$$

Where does this come from? The analysis of the Lyapunov energy:

$$\|x^{t+1} - x^*\|^2 - \|x^t - x^*\|^2 \leqslant \gamma^2 \|\nabla f(x_t)\|^2 - 2\gamma \left(f(x^t) - f(x^*) \right)$$

Upper bound is *minimized* if γ_t is the Polyak stepsize.

II : Stochastic Polyak Stepsize

2 : Our proposal for SGD

The Stochastic Polyak Stepsize (SPS)

The same Lyapunov analysis leads to

$$\|x^{t+1} - x^*\|^2 - \|x^t - x^*\|^2 \leqslant \gamma^2 \|\nabla f_{i_t}(x_t)\|^2 - 2\gamma \left(f_{i_t}(x^t) - f_{i_t}(x^*) \right)$$

The upper bound is minimized if:

$$\boxed{\gamma_t := \frac{(f_{i_t}(x^t) - f_{i_t}(x^*))_+}{\|\nabla f_{i_t}(x^t)\|^2}}$$

- $f_{i_t}(x^*)$ is impossible to know exactly ... except if there is interpolation
- γ_t can be 0 if x^t is too good at minimizing f_{i_t}
- the distance to minimizers is *decreasing* which is unheard of for SGD

SPS : An alternative definition

$$\gamma_t := \frac{(f_{i_t}(x^t) - f_{i_t}(x^*))_+}{\|\nabla f_{i_t}(x^t)\|^2}$$

SPS : An alternative definition

$$\gamma_t := \frac{(f_{i_t}(x^t) - f_{i_t}(x^*))_+}{\|\nabla f_{i_t}(x^t)\|^2}$$

Given a solution x^*, our problem is equivalent to find x such that

$$(\forall i \in \{1, \dots, m\}) \quad f_i(x) \leqslant f_i(x^*)$$

SPS : An alternative definition

$$\gamma_t := \frac{(f_{i_t}(x^t) - f_{i_t}(x^*))_+}{\|\nabla f_{i_t}(x^t)\|^2}$$

Given a solution x^*, our problem is equivalent to find x such that

$$(\forall i \in \{1, \ldots, m\}) \quad f_i(x) \leqslant f_i(x^*)$$

Newton-Raphson : sample & project onto linearization of the constraints

$$x^{t+1} = \text{argmin } \|x - x^t\|^2 \text{ s.t. } f_{i_t}(x^t) + \langle \nabla f_{i_t}(x^t), x - x^t \rangle \leqslant f_{i_t}(x^*)$$

SPS : An alternative definition

$$\gamma_t := \frac{(f_{i_t}(x^t) - f_{i_t}(x^*))_+}{\|\nabla f_{i_t}(x^t)\|^2}$$

Given a solution x^*, our problem is equivalent to find x such that

$$(\forall i \in \{1, \dots, m\}) \quad f_i(x) \leqslant f_i(x^*)$$

Newton-Raphson : sample & project onto linearization of the constraints

$$x^{t+1} = \text{argmin } \|x - x^t\|^2 \text{ s.t. } f_{i_t}(x^t) + \langle \nabla f_{i_t}(x^t), x - x^t \rangle \leqslant f_{i_t}(x^*)$$

- Those iterates are *exactly* the ones of SGD+SPS
- No convexity needed for this formulation (but $\neq$ problem)

SPS : the smooth case

Theorem

Let $f_i \in \Gamma_0(\mathbb{R}^N)$ be locally smooth and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. Then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{4LD^2}{T} + \frac{2D\sigma_*}{\sqrt{T}},$$

where L is the worst Lipschitz constant of ∇f_i over $\mathbb{B}(x^*, D)$.

SPS : the smooth case

Theorem

Let $f_i \in \Gamma_0(\mathbb{R}^N)$ be locally smooth and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. Then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{4LD^2}{T} + \frac{2D\sigma_*}{\sqrt{T}},$$

where L is the worst Lipschitz constant of ∇f_i over $\mathbb{B}(x^*, D)$.

- This is an *asymptotic* $\frac{1}{\sqrt{T}}$ rate, with no log terms
- Nearly optimal in the interpolation regime, adaptive to σ_*^2, L, D
- No need for global smoothness!
- If $f = \mathbb{E}f_\xi$, ask locally smooth to be uniform in ξ

SPS : the nonsmooth case

Theorem

Let $f_i \in \Gamma_0(\mathbb{R}^N)$ be locally lipschitz and $\bar{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. Then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant \frac{DG}{\sqrt{T}},$$

where G is the worst Lipschitz constant of f_i over $\mathbb{B}(x^*, D)$.

- This is an *asymptotic* $\frac{1}{\sqrt{T}}$ rate
- Nearly optimal for this class of problem, adaptive to G, D
- No need for global Lipschitz!
- If $f = \mathbb{E}f_\xi$, ask instead local expected lipschitz

II : Stochastic Polyak Stepsize

3 : Can we run this in practice ???

SPS for large NNs

For large (enough) models, interpolation holds, meaning that $f_i(x^*) = \inf f_i$

So it is worth trying the cheap rule $\gamma_t = \frac{f_{i_t}(x^t) - \inf f_i}{\|\nabla f_{i_t}(x^t)\|^2}$

Usually $\inf f_i = 0$ except for regularized problems where the extra $\|x\|^2$ perturbs the minima. But we can still compute the infimum[Loizou et al.]

SPS with approximation : optimistic version

A reasonable approach consists in replacing $f_i(x^*)$ with an approximation

One could *hope* that interpolation holds, and use $\inf f_i$ even if it is illegal

Theorem (Loizou et al. - 2021)

Let f_i be convex and L-smooth. If $\gamma_t := \max\left\{\frac{f_{i_t}(x^t)-\inf f_i}{\|\nabla f_{i_t}(x^t)\|^2}; \bar{\gamma}\right\}$ then

$$\mathbb{E}\left[f(\bar{x}^T) - \inf f\right] \leqslant O\left(\frac{D^2}{\bar{\gamma}T} + \Delta^*\right)$$

where $\Delta^* = \mathbb{E}\left[f_i(x^*) - \inf f_i\right] = \inf f - \mathbb{E}\left[\inf f_i\right]$.

We pay the error we make on estimating $f_i(x^*)$

SPS with approximation : educated version

A reasonable approach consists in replacing $f_i(x^*)$ with an approximation

~~One could *hope* that interpolation holds, and use $\inf f_i$ even if it is not legal~~

We could build a more precise estimation of $f_i(x^*)$

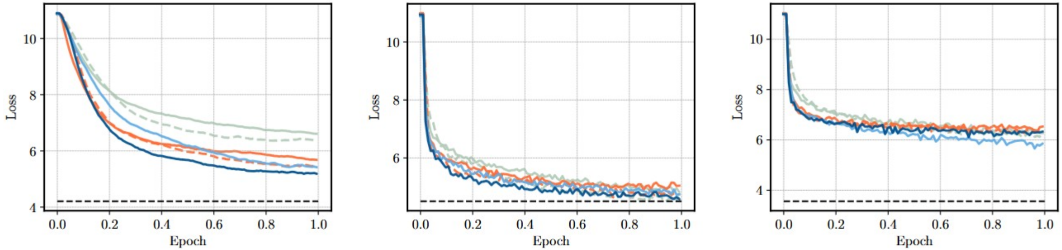
Example (Black-box model distillation)

Train a small model (student) with a pretrained bigger model (teacher).
If the weights of the teacher are good, it should happen that

$$f_i(x^*) \simeq f_i(x^{\text{tea}}) \leqslant f_i(x^{\text{stu}})$$

So we can use $f_i(x^{\text{tea}})$ as a surrogate for $f_i(x^*)$.

Numerical experiment : Distillation



tinyShakespeare PTB Wikitext2

SGD (gray), Adam (Orange), SPS+Mom (Blue), SPS+Adam (Dark blue)

Dotted lines : scheduler (warmup + cosine decay)

Datasets: 300K, 1M, 2M tokens

SPS with approximation : on-the-fly version

Remember that SGD+SPS is Newton applied to the feasability problem

$$(\forall i \in \{1, \dots, m\}) \quad f_i(x) \leqslant f_i(x^*)$$

We could be creative and introduce an other problem without $f_i(x^*)$ such as:

$$\min_{x \in \mathbb{R}^n, s \in \mathbb{R}^m} \frac{1}{m} \sum_{i=1}^{m} s_i \text{ s.t. } f_i(x) \leqslant s_i.$$

Not only do we need to project on linearization of the constraints, but also take into account the objective function. This leads to a stochastic proximal method:

$$x^{t+1}, s^{t+1} = \text{argmin } s_i + \|(x, s) - (x^t, s^t)\|^2 \text{ s.t. } f_i(x^t) + \langle \nabla f_i(x^t), x - x^t \rangle \leqslant s_i$$

SPS with approximation : on-the-fly version

$$x^{t+1}, s^{t+1} = \operatorname{argmin}\ s_i + \|(x,s) - (x^t, s^t)\|^2 \text{ s.t. } f_i(x^t) + \langle \nabla f_i(x^t), x - x^t \rangle \leqslant s_i$$

- This algorithm (FUVAL) admits a closed form solution (nasty)
- s_i^t tries to converge to $f_i(w^*)$
- Extra parameters needed for the theory to work (boooo)
- Can garantee a dirty $O\left(\frac{1}{\sqrt{T}}\right)$ rate in the nonsmooth case
- Can guarantee a $O\left(\frac{1}{T} + \Delta_*\right)$ rate in the smooth case
- Numerics aren't great
- To be improved! Stochastic prox methods are not well understood...

II : Stochastic Polyak Stepsize

4 : Would you want some momentum?

Adam is SGD + AdaGrad + Momentum. Replace Adagrad with SPS?

Adam is SGD + AdaGrad + Momentum. Replace Adagrad with SPS?

Momentum (v1 : Heavy Ball)

$$
\begin{aligned}
y_t &= x_t + \beta_t(x_t - x_{t-1}) \\
x_{t+1} &= y_t - \gamma \nabla f_{i_t}(x_t)
\end{aligned}
$$

Momentum (v2 : Classic)

$$
\begin{aligned}
m_t &= \beta_t m_{t-1} + \nabla f_{i_t}(x_t) \\
x_{t+1} &= x_t - \gamma m_t
\end{aligned}
$$

Momentum (v3 : Iterative Moving Average)

$$
\begin{aligned}
z_t &= z_{t-1} - \eta_t \nabla f_{i_t}(x_t) \\
x_{t+1} &= (1-\alpha_t)x_t + \alpha_t z_t
\end{aligned}
$$

Momentum : Known result

Theorem

Let $f_i \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ and run IMA with $\eta_t \equiv \eta \leqslant \frac{1}{4L}$ and $\alpha_t = \frac{2}{2+t}$.

$$\mathbb{E}\left[f(x^T) - \inf f\right] \leqslant \frac{D^2}{\eta T} + 2\eta\sigma_*^2,$$

- Exact same bound as SGD constant stepsize
- Momentum provides *last iterate* bounds
- No known acceleration

Momentum + SPS : Smooth case

Theorem

Let f_i convex and locally smooth, and run IMA with $\eta_t = \text{SPS}$ and $\alpha_t = \frac{1}{1+t}$.

$$\mathbb{E}\left[f(x^T) - \inf f\right] \leqslant \frac{2LD^2 \log(T)}{T} + \frac{2\sqrt{L\Delta_*}D}{\sqrt{T}},$$

- Like SGD+SPS but with *last iterates*
- Spurious log term (boooo)

Momentum + SPS : Nonsmooth case

Theorem

Let f_i convex and locally Lipschitz, run IMA with $\eta_t =$ SPS and $\alpha_t = \frac{1}{1+t}$:

$$\mathbb{E}\left[f(x^T) - \inf f\right] \leqslant \frac{GD}{\sqrt{T}}.$$

- Like SGD+SPS but with *last iterates*
- No spurious log term (yay)

What is SPS for momentum?

If you really want to know:

Definition (Momentum + SPS)

$$\begin{cases} \eta_t & = \frac{\left(f_{i_t}(x^t) - f_{i_t}(x^*) + \langle \nabla f_{i_k}(x_k), z_{t-1} - x_t \rangle\right)_+}{\|\nabla f_{i_k}(x_k)\|^2} \\ z_k & = z_{t-1} - \eta_t \nabla f_{i_k}(x_k) \\ x_{k+1} & = (1-\alpha_k)x_k + \alpha_k z_k \end{cases}$$

We can also do SPS for Adam!

Conclusion

Conclusion on SPS

- Theory: great
  - Nearly optimal rates in both smooth and nonsmooth
  - Adaptivity to all parameters (except $f_i(x^*)$)
- Practice: disputable
  - Can't be used as is in every scenario
  - Some promising edge cases (interpolation, distillation)
  - Need for more analysis when approximating $f_i(x^*)$
  - Need more algorithms like FUVAL with on-the-fly tracking of $f_i(x^*)$

Thanks for your attention !

Any questions?