Lecture 7: Stochastic algorithms

Luca Calatroni
CR CNRS, Laboratoire I3S
CNRS, UCA, Inria SAM, France

MSc DSAI - UCA

Inverse problems in image processing

March 3 2023

Table of contents

Bibliography

- Léon Bottou, Frank E. Curtis, Jorge Nocedal, *Optimization Methods for Large-Scale Machine Learning*, SIAM Review, available here `https://coral.ise.lehigh.edu/frankecurtis/files/papers/BottCurtNoce18.pdf`.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016, `http://www.deeplearningbook.org`.
- Guillaume Garrigos, Robert Gower, *Handbook of Convergence Theorems for (Stochastic) Gradient Methods*, `https://arxiv.org/abs/2301.11235`, 2023.
- Robert Gower, *Cornell lecture: Optimization for machine learning*, spring 2020, available here `https://gowerrobert.github.io/`

Stochastic gradient descent

Motivations

Back to smooth optimisation problem:

$$\boxed{\min_{x\in\mathbb{R}^n} \; f(x)}$$

for differentiable, proper f with L-Lipschitz gradient ∇f.

Motivations

Back to smooth optimisation problem:

$$\boxed{\min_{x\in\mathbb{R}^n} \; f(x)}$$

for differentiable, proper f with L-Lipschitz gradient ∇f.

In the context of learning approaches, very often f relates to empirical-risk minimisation function. For a set of examples $\{y_1,\ldots,y_n\}$, with (typically) large $n \gg 1$:

$$\min_{x\in\mathbb{R}^n} \; \left\{ f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) = \frac{1}{2n}\sum_{i=1}^n \|A_i x - y_i\|^2 \right\}$$

Recalling GD iteration, for $x_0 \in \mathbb{R}^n$ and $\tau \in (0, \frac{1}{L}]$, $k \geq 0$:

$$x_{k+1} = x_k - \tau\nabla f(x_k) = x_k - \frac{\tau}{n}\sum_{i=1}^n \nabla f_i(x_k)$$

Gradient evaluations may be costly (many matrix/vector products...)...

Motivations

Back to smooth optimisation problem:

$$\boxed{\min_{x\in\mathbb{R}^n} \; f(x)}$$

for differentiable, proper f with L-Lipschitz gradient ∇f.

In the context of learning approaches, very often f relates to empirical-risk minimisation function. For a set of examples $\{y_1,\ldots,y_n\}$, with (typically) large $n \gg 1$:

$$\min_{x\in\mathbb{R}^n} \; \left\{ f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \frac{1}{2n}\sum_{i=1}^{n} \|A_i x - y_i\|^2 \right\}$$

Recalling GD iteration, for $x_0 \in \mathbb{R}^n$ and $\tau \in (0, \frac{1}{L}]$, $k \geq 0$:

$$x_{k+1} = x_k - \tau\nabla f(x_k) = x_k - \frac{\tau}{n}\sum_{i=1}^{n} \nabla f_i(x_k)$$

Gradient evaluations may be costly (many matrix/vector products...)...

...can we use only one (or some) of the component(s) f_i to reduce computational costs?

Basic assumption

Use of $\nabla f_j(x) \approx \nabla f(x)$?

Unbiased gradient estimator

Let $j \in \{1, \dots, n\}$ be a random index selected **uniformly** at random (with probability $\frac{1}{n}$). Then:

$$\mathbb{E}_j[\nabla f_j(x_k)|x_k] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k) = \nabla f(x_k)$$

"On average" the use of a single component ∇f_j provides a good approximation of ∇f for uniformly sampled components j.

Basic assumption

Use of $\nabla f_j(x) \approx \nabla f(x)$?

Unbiased gradient estimator

Let $j \in \{1, \ldots, n\}$ be a random index selected **uniformly** at random (with probability $\frac{1}{n}$). Then:

$$\mathbb{E}_j[\nabla f_j(x_k)|x_k] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k) = \nabla f(x_k)$$

"On average" the use of a single component ∇f_j provides a good approximation of ∇f for uniformly sampled components j.

Stochastic Gradient Descent (SGD): constant step-size

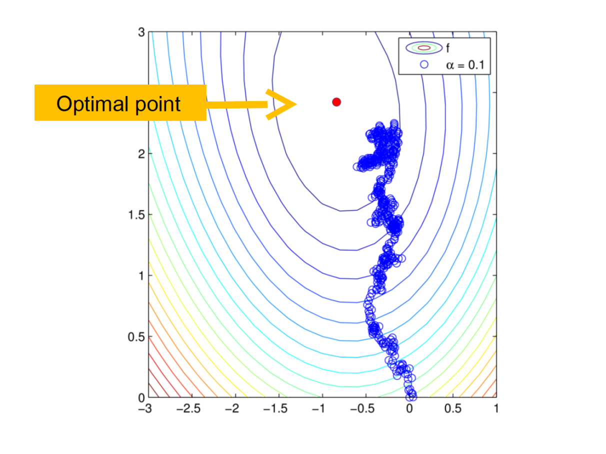
Input: $x_0 \in \mathbb{R}^n$ (initial guess), $\tau_k > 0$

Iterate for $k \geq 0$:

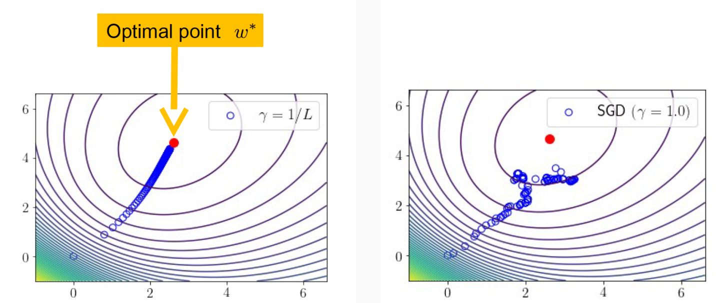
$$\text{sample uniformly } j \in \{1, \ldots, n\}$$
$$x_{k+1} = x_k - \tau_k \nabla f_j(x_k)$$

till **convergence**.

Intuitive behaviour



Intuitive behaviour



Why does this happen? Need of assumptions!

Convergence of SGD (constant step-size)

Convergence of SGD (constant τ

Let $x^* \in \arg\min f$, ∇f_i for $i = 1, \ldots, n$ be L_i-smooth and let $L_{max} := \max_i \; L_i$. Then, the sequence (x_k) generated by SGD with $0 < \tau_k \equiv \tau < \frac{1}{2L_{max}}$ satisfies:

$$\mathbb{E}[f(\bar{x}_k) - f(x^*)] \leq \frac{\|x_0 - x^*\|^2}{2\tau(1 - 2\tau L_{max})} \frac{1}{k} + \frac{\tau}{(1 - 2\tau L_{max})} \mathrm{Var}[\nabla f_j(x^*)],$$

with $\bar{x}_k := \frac{1}{k} \sum_{j=0}^{k-1} x_j$.

- First term is similar to what you get in non-stochastic GD
- Second term depends on

$$\mathrm{Var}[\nabla f_j(x^*)] = \mathbb{E}[\left\|\nabla f_j(x^*) - \nabla f(x^*)\right\|^2] = \mathbb{E}[\left\|\nabla f_j(x^*)\right\|^2]$$

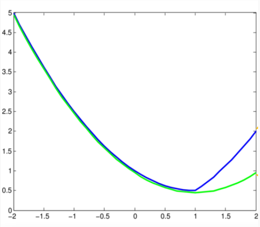
- Still $O(1/k)$ rate
- Possibly, very small step-size τ!

Improved condition for convergence: strong convexity

Strongly convex function

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper and differentiable function. We say that f is **strongly convex** of parameter $\mu > 0$ (or μ-strongly convex) if:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n.$$



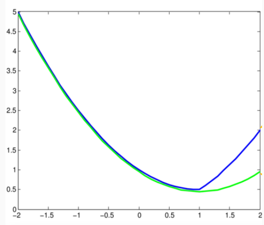
Strongly convex functions have a quadratic lower bound at every point.

Improved condition for convergence: strong convexity

Strongly convex function

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper and differentiable function. We say that f is **strongly convex** of parameter $\mu > 0$ (or μ-strongly convex) if:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n.$$



Strongly convex functions have a quadratic lower bound at every point.

Important remark: strong convexity $\Rightarrow$ strict convexity $\Rightarrow$ convexity

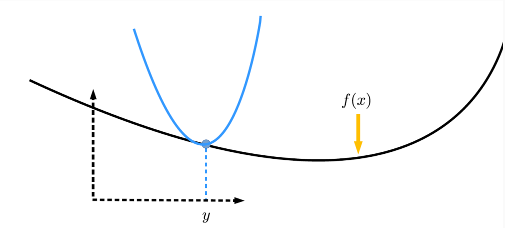
Exercise: $f(x) = x^4$ is a strictly convex function which is not strongly convex.

From lower to upper bounds via L-Lipschitz smoothness

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper and differentiable function with L-Lipschitz gradient. Then:

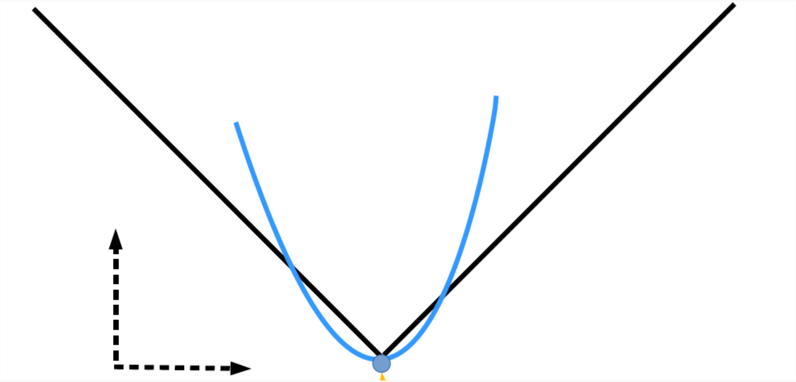
$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n.$$



L-smooth functions have a quadratic **upper** bound at every point.

Test this property, convex counter-example

$f(x) = \|x\|_1$, convex.



Can't define a quadratic upper bound in $x = 0$

Theorem (Convergence of SGD for strongly convex objectives))

If f is μ-strongly convex and all ∇f_i are L_i-Lipschitz continuous $i = 1, \ldots, n$, denoting by $x^* = \arg\min \; f(x)$ and defining:

$$L_{max} := \max_{i=1,\ldots,n} L_i, \qquad \mathrm{Var}[\nabla f_j(x^*)] := \mathbb{E}\left(\|\nabla f_j(x^*)\|^2\right),$$

the iterates (x_k) of SGD with $\tau \leq \frac{1}{2L_{max}}$ satisfy:

$$\mathbb{E}\left(\|x_k - x^*\|^2\right) \leq (1 - \tau\mu)^k \|x_0 - x^*\|^2 + \frac{2\tau}{\mu}\mathrm{Var}[\nabla f_j(x^*)]$$

Convergence result

Theorem (Convergence of SGD for strongly convex objectives))

If f is μ-strongly convex and all ∇f_i are L_i-Lipschitz continuous $i = 1, \ldots, n$, denoting by $x^* = \arg\min\ f(x)$ and defining:

$$L_{max} := \max_{i=1,\ldots,n} L_i, \qquad \mathrm{Var}[\nabla f_j(x^*)] := \mathbb{E}\left(\|\nabla f_j(x^*)\|^2\right),$$

the iterates (x_k) of SGD with $\tau \leq \frac{1}{2L_{max}}$ satisfy:

$$\mathbb{E}\left(\|x_k - x^*\|^2\right) \leq (1 - \tau\mu)^k \|x_0 - x^*\|^2 + \frac{2\tau}{\mu}\mathrm{Var}[\nabla f_j(x^*)]$$

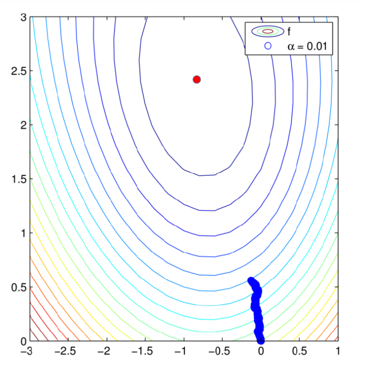
Note that in order to have $\mathbb{E}\left(\|x_k - x^*\|^2\right) \to 0$, we need:

- To get $(1 - \tau\mu)^k \to 0$ fast, we hope $1 - \tau\mu \approx 0$, i.e. $\tau = 1/\mu \gg 1$
- To get $\frac{2\tau}{\mu}\mathrm{Var}[\nabla f_j(x^*)] \to 0$, I need τ small

Not really compatible. . .

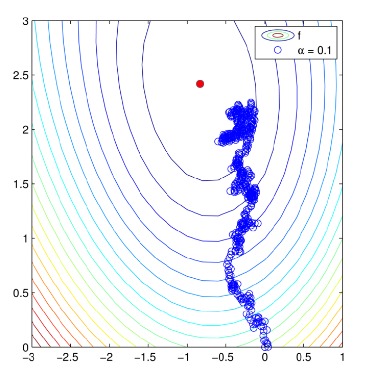
Convergence of SGD for fixed step-size

Small $\tau \approx 0$



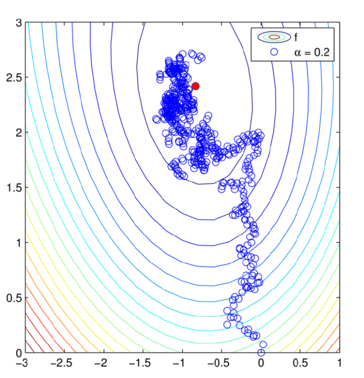
Convergence of SGD for fixed step-size

$\tau = 0.1$



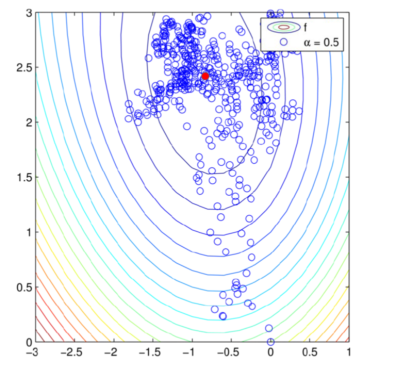
Convergence of SGD for fixed step-size

$\tau = 0.2$



Convergence of SGD for fixed step-size

$\tau = 0.5$



Idea: choose large τ_k in early iterations to get close to the minimiser and then reduce it.

SGD with varying step-sizes

Stochastic Gradient Descent (SGD): varying step-size

Input: $x_0 \in \mathbb{R}^n$ (initial guess), (τ_k) s.t. $\sum_{k=1}^{\infty} \tau_k = +\infty$, $\tau_k \to 0$

Iterate for $k \geq 0$:

$$\text{sample uniformly } j \in \{1, \ldots, n\}$$
$$x_{k+1} = x_k - \tau_k \nabla f_j(x_k)$$

till **convergence**.

SGD with varying step-size: convergence

Theorem (decaying step-sizes)

If f is μ-strongly convex, all ∇f_i are L_i-Lipschitz continuous, let $x^* = \arg\min \ f(x)$ and:

$$L_{max} := \max_{i=1,\dots,n} L_i, \quad \mathrm{Var}[\nabla f_j(x^*)] := \mathbb{E}\left(\|\nabla f_j(x^*)\|^2\right), \quad \kappa := \lceil L_{max}/\mu \rceil,$$

for the following choice of step-sizes:

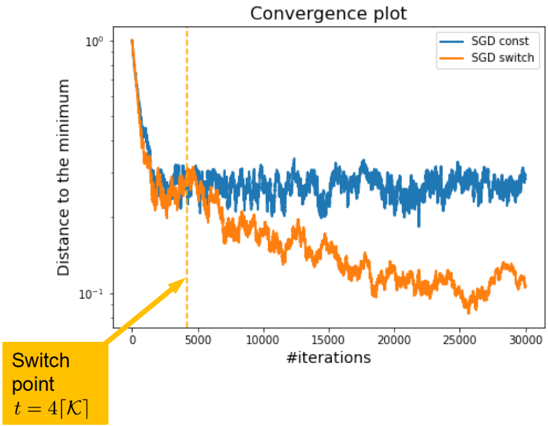
$$\tau_k = \begin{cases} \frac{1}{2L_{max}} & \text{if } k \leq 4\kappa \\ \frac{2k+1}{(k+1)^2\mu} & \text{if } k > 4\kappa \end{cases}$$

and $k \geq 4\kappa$, the following convergence result holds true:

$$\mathbb{E}\left(\|x_k - x^*\|^2\right) \leq \frac{8\sigma^2}{\mu^2 k} + \frac{16\kappa}{e^2 k^2}\|x_0 - x^*\|^2$$

Note: RHS$\to 0$. Note that the decreasing step is $\approx O\left(\frac{1}{k+1}\right)$. Practically, often a slowest decay $\tau_k = \frac{C}{\sqrt{k+1}}$, for tuned $C > 0$ is chosen.

Convergence comparisons



How to avoid oscillations?

Stochastic gradient descent with averaging

SGD with averaging

SGD with varying step-size and (late) averaging

Input: $x_0 \in \mathbb{R}^n$ (initial guess), (τ_k) s.t. $\sum_{k=1}^{\infty} \tau_k = +\infty$, $\tau_k \to 0$, $s_0 \in \mathbb{N}$

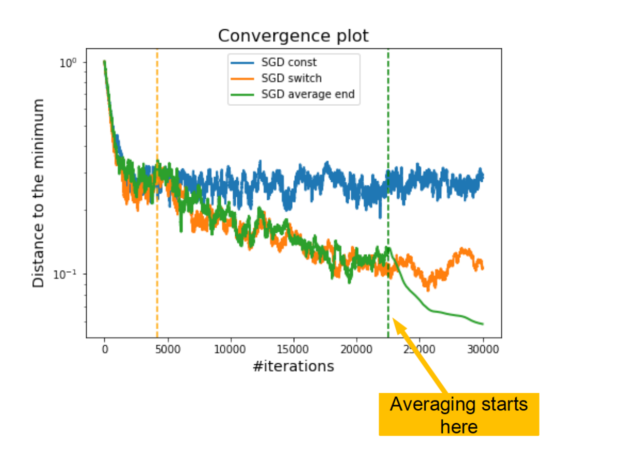
Iterate for $k \geq 0$:

$$
\begin{aligned}
&\text{sample uniformly } j \in \{1, \ldots, n\} \\
&x_{k+1} = x_k - \tau_k \nabla f_j(x_k) \\
&\text{if } k > s_0 \\
&\qquad \bar{x} := \frac{1}{k - s_0} \sum_{i=s_0}^{k} x_k \\
&\text{else} \\
&\qquad \bar{x} = x_k \\
&x_k = \bar{x}
\end{aligned}
$$

till **convergence**.

Strategy employed in: Polyak, Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM Journal on Control and Optimization, 1992.

Further convergence comparisons

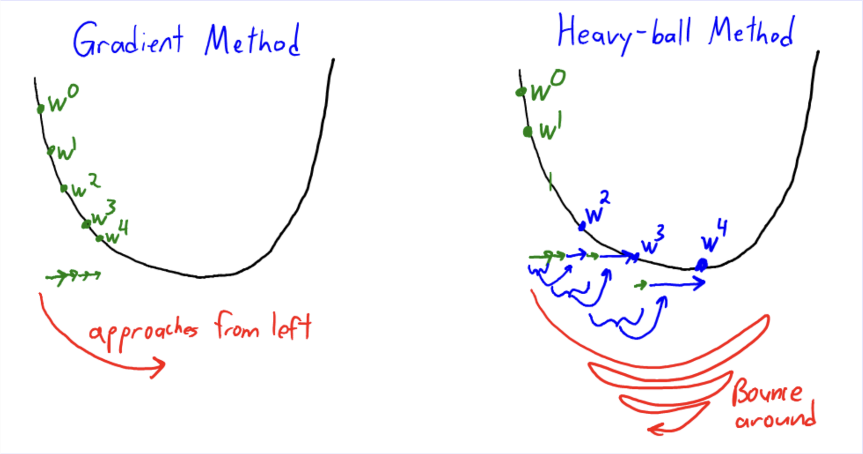


Acceleration strategies

We have seen already the idea of "inertia" (à la FISTA) to improve convergence speed.

Heavy ball acceleration (deterministic):

$$x_{k+1} = x_k - \tau \nabla f(x_k) + \beta_k (x_k - x_{k-1}), \quad \beta_k \in (0,1)$$



Back to GD with momentum

We have seen already the idea of “inertia” (à la FISTA) to improve convergence speed.

Heavy ball acceleration (deterministic):

$$x_{k+1} = x_k - \tau \nabla f(x_k) + \beta_k (x_k - x_{k-1}), \quad \beta_k \in (0,1)$$

Momentum method (stochastic):

$$x_{k+1} = x_k - \tau \nabla f_{i_k}(x_k) + \beta_k (x_k - x_{k-1}), \quad \beta_k \in (0,1)$$

with $i_k \in \{1, \dots, n\}$ drawn uniformly.

Convergence of momentum method

Convergence of momentum method

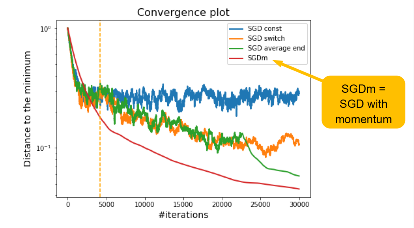
By taking the parameters as:

$$\tau_k = \frac{2\eta}{k+3}, \quad \beta_k = \frac{k}{k+1}, \quad \eta \leq \frac{1}{4L_{max}},$$

then the following result holds for SGD with momentum:

$$\mathbb{E}[f(x_k) - f(x^*)] \leq \frac{\|x_0 - x^*\|^2}{\eta(k+1)} + 2\eta \text{Var}[\nabla f_i(x^*)]$$

Not **faster** result, but **stronger** result as convergence on sequence (not on average)!



Nods on variants and non-smooth alternatives

Variants of SGD

- **Averaging**: close relations ($\approx$ equivalence) with SGD with momentum approaches.
- **AdaGrad**, **RMSProp**: SGD with adaptive (component-dependent) learning rate.
- **ADAM**: adaptive algorithm using both estimation of first and second moment of the gradients.

Generalisation to non-smooth problems (proximal algorithms?)

$$\min_{x \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} f_i(x) + g(x)$$

Stochastic proximal gradient algorithm

$$\min_{x \in \mathbb{R}^n} \; F(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + g(x)$$

- f_i is differentiable, has L-Lipschitz gradient
- g is proper, convex and l.s.c
- $x^* \in \arg\min \; F$

Stochastic proximal gradient descent (constant τ)

$x_0 \in \mathbb{R}^n$, $\tau < \frac{1}{4L_{max}}$

$$i_k \in \{1, \dots, n\} \quad \text{with probability } \frac{1}{n}$$

$$x_{k+1} = \operatorname{prox}_{\tau g}(x_k - \tau \nabla f_{i_k}(x_k))$$

Convergence of stochastic proximal gradient algorithm

Convergence of SPGD (constant τ)

Let (x_k) be the sequence generated by SPGD with constant step-size $\tau < \frac{1}{4L_{max}}$.
Then for all $k \geq 1$:

$$\mathbb{E}[F(\bar{x}_k) - F(x^*))] \leq \frac{\|x_0 - x^*\|^2 + 2\tau(F(x_0) - F(x^*))}{2(1 - 4\tau L_{max})\tau k} + \frac{2\tau \mathrm{Var}[\nabla f_i(x^*)]}{(1 - 4\tau L_{max})},$$

with $\bar{x}_k = \frac{1}{k}\sum_{j=0}^{k-1} x_j$.

Again, "average" result.

- In the case of strong convexity, convergence condition is $\tau < \frac{1}{2L_{max}}$ and rate is of the form $(1 - \mu\tau)^k$
- Acceleration (momentum) in stochastic contexts is very tricky, not very clear how to perform it
- Stochastic proximal operators?

See A. Khaled et al., 2020 for more variants.

Questions?

calatroni@i3s.unice.fr