

NO-REGRET CRITERIA IN LEARNING, GAMES AND CONVEX OPTIMIZATION

Sylvain Sorin
sylvain.sorin@imj-prg.fr

Institut de Mathématiques Jussieu-PRG
Sorbonne Université, UPMC - Paris 6
CNRS UMR 7586

Operator Splitting Methods in Data Analysis
Center for Computational Mathematics
Flatiron Institute, New York
March 20-22, 2019

Abstract

The purpose of this work is to underline links between no-regret algorithms used in learning, games and convex optimization. In particular we will study continuous and discrete time versions and their connections.

We will comment on recent advances in on:

- Euclidean and non-euclidean approaches
- speed of convergence of the evaluation
- convergence of the trajectories

Model

V normed vector space, finite dimensional
dual V^* and duality map $\langle \cdot | \cdot \rangle$.

$X \subset V$ compact convex.

The aim is to study properties of algorithms that associate to a process $\{u_t \in V^*, t \geq 0\}$, a process $\{x_t \in X, t \geq 0\}$, where x_t is function of $\{(x_s, u_s), 0 \leq s < t\}$, satisfying:

$$R_t(x) = \int_0^t \langle u_s | x - x_s \rangle ds \leq o(t), \quad t \geq 0, \forall x \in X \quad (1)$$

or in discrete time with $\{x_m\}$ depending on $\{x_1, u_1, \dots, x_{m-1}, u_{m-1}\}$ with:

$$R_n(x) = \sum_{m=1}^n \langle u_m | x - x_m \rangle \leq o(n), \quad \forall x \in X. \quad (2)$$

Case 1 : general bounded process $\{u_t\}$ or $\{u_n\}$
no-regret learning

Model

V normed vector space, finite dimensional
dual V^* and duality map $\langle \cdot | \cdot \rangle$.

$X \subset V$ compact convex.

The aim is to study properties of algorithms that associate to a process $\{u_t \in V^*, t \geq 0\}$, a process $\{x_t \in X, t \geq 0\}$, where x_t is function of $\{(x_s, u_s), 0 \leq s < t\}$, satisfying:

$$R_t(x) = \int_0^t \langle u_s | x - x_s \rangle ds \leq o(t), \quad t \geq 0, \forall x \in X \quad (1)$$

or in discrete time with $\{x_m\}$ depending on $\{x_1, u_1, \dots, x_{m-1}, u_{m-1}\}$ with:

$$R_n(x) = \sum_{m=1}^n \langle u_m | x - x_m \rangle \leq o(n), \quad \forall x \in X. \quad (2)$$

Case 1 : general bounded process $\{u_t\}$ or $\{u_n\}$
no-regret learning

Properties

Case 2 : vector field $g : X \rightarrow V^*$

$u_t = g(x_t)$ or $u_n = g(x_n)$

VI or game framework

Consider a game with a finite set of players I where equilibria are solution of variational inequalities:

$$\langle g^i(x) | x^i - y^i \rangle \geq 0, \quad \forall y^i \in X^i, \forall i \in I$$

$X^i \subset V^i$ is the strategy set of player i , $X = \prod_i X^i$, and $g^i : X \rightarrow V^{i*}$ is his evaluation function.

At stage n each player i chooses x_n^i , this defines a profile $x_n \in X$ and the reference process for player i is $u_n^i = g^i(x_n)$.

$$\langle g(x) | y - x \rangle = \sum_i \langle g^i(x) | y^i - x^i \rangle$$

S' is the set of solutions of the variational inequality:

$$\langle g(x)|y-x\rangle \leq 0, \quad \forall y \in X \quad (3)$$

Lemma

If g is continuous and $x_s \rightarrow x$ then $x \in S'$.

S is the set of solutions of the variational inequality:

$$\langle g(y)|y-x\rangle \leq 0, \quad \forall y \in X. \quad (4)$$

g *dissipative* ($\langle g(x) - g(y)|x - y\rangle \leq 0, \forall x, y \in X$) implies $S' \subset S$ and if moreover g is continuous: $S \subset S'$.

Lemma

If g is dissipative the accumulation points of $\{\bar{x}_t\}$ or $\{\bar{x}_n\}$ are in S .

Proof:

$$\frac{R_t(y)}{t} = \frac{1}{t} \int_0^t \langle g(x_s)|y - x_s \rangle \geq \frac{1}{t} \int_0^t \langle g(y)|y - x_s \rangle = \langle g(y)|y - \bar{x}_t \rangle$$

where $\bar{x}_t = \frac{1}{t} \int_0^t x_s ds$

S' is the set of solutions of the variational inequality:

$$\langle g(x)|y-x \rangle \leq 0, \quad \forall y \in X \quad (3)$$

Lemma

If g is continuous and $x_s \rightarrow x$ then $x \in S'$.

S is the set of solutions of the variational inequality:

$$\langle g(y)|y-x \rangle \leq 0, \quad \forall y \in X. \quad (4)$$

g *dissipative* ($\langle g(x) - g(y)|x - y \rangle \leq 0, \forall x, y \in X$) implies $S' \subset S$ and if moreover g is continuous: $S \subset S'$.

Lemma

If g is dissipative the accumulation points of $\{\bar{x}_t\}$ or $\{\bar{x}_n\}$ are in S .

Proof:

$$\frac{R_t(y)}{t} = \frac{1}{t} \int_0^t \langle g(x_s)|y - x_s \rangle \geq \frac{1}{t} \int_0^t \langle g(y)|y - x_s \rangle = \langle g(y)|y - \bar{x}_t \rangle$$

where $\bar{x}_t = \frac{1}{t} \int_0^t x_s ds$

Case 3 : $u_t = -\nabla f(x_t)$, f convex $\mathcal{C}^1$
convex optimization

$$\langle \nabla f(x_t) | y - x_t \rangle \leq f(y) - f(x_t)$$

gives:

$$\int_0^t [f(x_s) - f(y)] ds \leq \int_0^t \langle -\nabla f(x_s) | y - x_s \rangle ds = R_t(y)$$

which implies by Jensen's inequality:

$$f(\bar{x}_t) - f(y) \leq \frac{1}{t} \int_0^t [f(x_s) - f(y)] ds \leq \frac{R_t(y)}{t} \quad (5)$$

Lemma

The accumulation points of $\{\bar{x}_t\}$ or $\{\bar{x}_n\}$ belong to $S = \operatorname{argmin}_X f$.

One can also deal with the case $u_m = -\lambda_m \nabla f(x_m)$ with

$\lambda_m \geq 0$, $\sigma_n = \sum_{m \leq n} \lambda_m$ and $\sigma_n \hat{x}_n = \sum_{m \leq n} \lambda_m x_m$:

$$\sigma_n [f(\hat{x}_n) - f(y)] \leq \sum_{m \leq n} \lambda_m [f(x_m) - f(y)] \leq \sum_{m \leq n} \langle \lambda_m \nabla f(x_m) | x_m - y \rangle \leq R_n(y).$$

Continuous time

Potential function $P(t;y) \geq 0$ satisfying:

$$\langle u_t, y - x_t \rangle \leq -\frac{d}{dt}P(t;y), \quad \text{hence}$$

$$R_t(y) = \int_0^t \langle u_s | y - x_s \rangle ds \leq P(0;y) - P(t;y)$$

(1) rate of convergence $1/t$.

(2) Assume $y \in S$, then $P(t;y)$ is decreasing:

$$\frac{d}{dt}P(t;y) \leq \langle g(x_t), x_t - y \rangle \leq 0$$

(3) If $\{x_t\}$ is a descent procedure ($\frac{d}{dt}f(x_t) \leq 0$),

$$E(t;y) = t(f(x_t) - f(y)) + P(t;y)$$

is decreasing, for all $y \in X$.

$$\begin{aligned} \frac{d}{dt}E(t;y) &= f(x_t) - f(y) + t \frac{d}{dt}f(x_t) + \frac{d}{dt}P(t;y) \\ &\leq f(x_t) - f(y) + \langle \nabla f(x_t), y - x_t \rangle \leq 0 \end{aligned}$$

Accumulation point of $\{x_t\}$ are in S .

A. Projected gradient

V Hilbert, $X \subset V$, convex closed.

Dynamics

(Projected) gradient descent is defined by:

$$\langle u_t - \dot{x}_t, y - x_t \rangle \leq 0, \forall y \in X. \quad (6)$$

which is:

$$\dot{x}_t = \Pi_{T_X(x_t)}(u_t) \quad (7)$$

where Π_C is the projection on the closed convex set C and $T_C(x)$ is the tangent cone to C at x .

Potential

Let:

$$V(t; y) = \frac{1}{2} \|x_t - y\|^2, \quad y \in X. \quad (8)$$

$$\langle u_t, y - x_t \rangle \leq \langle \dot{x}_t, y - x_t \rangle = -\frac{d}{dt} V(t; y)$$

Trajectories

Lemma

Assume $S \neq \emptyset$ and g dissipative.

$\{\bar{x}_t\}$ converges weakly to a point in S .

Proof:

- $\{\bar{x}_t\}$ is bounded hence has weak accumulation points.
- The weak limit points of $\{\bar{x}_t\}$ are in S
- $\|x_t - y^*\|$ converges when $y^* \in S$

Hence by Opial's lemma, $\bar{x}_t$ converges weakly to a point in S . ■

Descent properties

Consider case 3: $u_t = -\nabla f(x_t)$.

Lemma

$f(x_t)$ is decreasing

Proof:

$$\frac{d}{dt}f(x_t) = \langle \nabla f(x_t), \dot{x}_t \rangle = -\|\dot{x}_t\|^2$$

since $\langle u_t - \dot{x}_t, \dot{x}_t \rangle = 0$ (Moreau's decomposition). ■

Lemma

$\{x_t\}$ weakly converges to a point in S

Proof:

Weak accumulation points of $\{x_t\}$ are in S . Then Opial's lemma applies. ■

to summarize :

- R_t is bounded
- in addition in case 2 $\{\bar{x}_t\}$ weakly converges to a point in S
- in case 3 $f(x_t)$ is decreasing thus $f(x_t)$ converges to f^* at speed $1/t$ and $\{x_t\}$ weakly converges to a point in S .

B. Mirror descent

Dynamics

Continuous version of “Mirror descent algorithm”

Nemirovski and Yudin [?], Beck and Teboulle [?]

Alvarez, Bolte and Brahic, Attouch and Teboulle, Bolte and Teboulle

H strictly convex, $\mathcal{C}^1$

X , compact, convex, $\subset \text{dom}H$.

The continuous time process satisfies:

$$\langle u_t - \frac{d}{dt} \nabla H(x_t) | y - x_t \rangle \leq 0, \forall y \in X. \quad (9)$$

The previous analysis corresponds to the case: $H(x) = \frac{1}{2} \|x\|^2$.

Potential

Bregman distance associated to H

$$D_H(y, x) = H(y) - H(x) - \langle \nabla H(x) | y - x \rangle (\geq 0).$$

$$\frac{d}{dt} D_H(y, x_t) = \langle -\frac{d}{dt} \nabla H(x_t) | y - x_t \rangle \quad (10)$$

so that (9) implies

$$\langle u_t | y - x_t \rangle \leq -\frac{d}{dt} D_H(y, x_t)$$

The use of a special functions H adapted to X allows to get rid of the normal cone and to produce a trajectory that remains in $\text{int}X$.

This leads to:

$$\frac{d}{dt} \nabla H(x_t) = u_t \quad (11)$$

$$\dot{x}_t = \nabla^2 H(x_t)^{-1} u_t. \quad (12)$$

This corresponds to a Riemannian metric.

In this case one has a descent algorithm for the gradient since:

$$\langle \nabla f(x_t) | \dot{x}_t \rangle = -\langle \nabla f(x_t) | \nabla^2 H(x_t)^{-1} \nabla f(x_t) \rangle \leq 0$$

To prove convergence the steps are:

1) $\{x_t\}$ has accumulation points (sublevels of $D_H(x^*, \cdot)$ bounded)

2) If $x_{t_k} \rightarrow x^*$ then $x^* \in S$

3) H1 if $z^k \rightarrow y$ then $D_H(y, z^k) \rightarrow 0$

For example H L -smooth and then:

$$0 \leq D_H(x, y) \leq \frac{L}{2} \|x - y\|^2$$

4) H2 if $D_H(y, z^k) \rightarrow 0$ then $z^k \rightarrow y$

For example H β -strongly convex and then:

$$D_H(x, y) \geq \frac{\beta}{2} \|x - y\|^2$$

C. Dual averaging

Continuous version of dual averaging Nesterov [?], “Lazy gradient mirror descent”, Kwon and Mertikopoulos [?].

Dynamics

Assume h bounded strictly convex sci with $\text{dom } h = X \subset V$ convex compact.

Let $U_t = \int_0^t u_s ds$ and x_t be the argmax of:

$$\langle U_t | x \rangle - h(x).$$

Let $h^*(w) = \sup_{x \in V} \langle w | x \rangle - h(x)$ be the Fenchel conjugate. h^* is differentiable.

The dynamics is given by:

$$x_t = \nabla h^*(U_t) \in X \tag{13}$$

Potential

Define, for $y \in X$:

$$W(t; y) = h^*(U_t) - \langle U_t | y \rangle + h(y) \quad (\geq 0). \quad (14)$$

$$\frac{d}{dt} h^*(U_t) = \langle u_t | \nabla h^*(U_t) \rangle = \langle u_t | x_t \rangle \quad (15)$$

thus:

$$\frac{d}{dt} W(t; y) = \langle u_t | x_t - y \rangle$$

Trajectories

Lemma

$f(x_t)$ is decreasing.

Proof:

$$\frac{d}{dt}f(x_t) = \langle \nabla f(x_t) | \nabla^2 h^*(U_t)(u_t) \rangle$$

with $u_t = -\nabla f(x_t)$. ■

Hence the accumulation points of x_t are in S .

Discrete time

A. Projected gradient

Dynamics

Levitin and Polyak [?], Polyak [?]

$$x_{m+1} = \operatorname{argmin}_X \{ \langle \nabla f(x_m), x \rangle + (1/2\eta_m) \|x - x_m\|^2 \}, \quad (16)$$

(η_m decreasing) which corresponds to:

$$x_{m+1} = \Pi_X[x_m + \eta_m u_m], \quad (17)$$

or with variational characterization:

$$\langle x_m + \eta_m u_m - x_{m+1}, y - x_{m+1} \rangle \leq 0, \forall y \in X. \quad (18)$$

Values

Let $m(X)$ be the diameter of X . Assume $\|u_m\|_* \leq M$.

Proposition

$$R_n(x) \leq \frac{1}{2\eta_n} m(X)^2 + \frac{M^2}{2} \sum_{m=1}^n \eta_m$$

hence with $\eta_n = 1/\sqrt{n}$, $R_n(x) \leq O(\sqrt{n})$.

Trajectories

Assume $S \neq \emptyset$.

Lemma

For $x^* \in S$, $\|x_m - x^*\|$ converges if $\eta_n \in \ell^2$.

Lemma

If $\eta_n \in \ell^2$ and g is dissipative, $\{\bar{x}_n\}$ converges to a point in S .

B. Mirror descent

*Assumption: H L -strongly convex for some norm $\|\cdot\|$ on $V = \mathbb{R}^n$.
 $\|u_n\|_* \leq M$.*

Dynamics

Nemirovski and Yudin [?], Beck and Teboulle [?]

The usual MD algorithm is given by :

$$x_{m+1} = \operatorname{argmin}_X \{ \langle \nabla f(x_m) | x \rangle + (1/\eta_m) D_H(x, x_m) \}, \quad (19)$$

General formulation:

$$\langle \nabla H(x_m) + \eta_m u_m - \nabla H(x_{m+1}) | x - x_{m+1} \rangle \leq 0, \forall x \in X. \quad (20)$$

Values

Proposition

$$R_n(x) \leq \frac{D_H(x, x_1)}{\eta} + n\eta \frac{M^2}{2L}.$$

Then $\eta = 1/\sqrt{n}$.

Trajectories

Assume $S \neq \emptyset$.

Lemma

For $x^* \in S$, $D_H(x^*, x_n)$ converges if $\{\eta_n\} \in \ell^2$.

C: Dual averaging

Assumption: h L -strongly convex for some norm $\|\cdot\|$ on $V = \mathbb{R}^n$.

Dynamics

Nesterov [?]

The algorithm is given by:

$$x_{m+1} = \nabla h^*(\eta_m U_m).$$

and $\{\eta_m\}$ is decreasing.

Values

Nesterov [?] or discrete approximation of (13) Kwon and Mertikopoulos [?]:

Proposition

$$R_n(x) = \sum_{m=1}^n \langle u_m | x - x_m \rangle \leq \frac{r_X(h)}{\eta_n} + \frac{\sum_{m=1}^n \eta_{m-1} \|u_m\|_*^2}{2L} \quad (21)$$

Assume: $\|u_m\|_* \leq M$.

Hence the convergence rate $O(\sqrt{n})$ with time varying parameters $\eta_m = 1/\sqrt{m}$.

Smooth case

Assume that f is β smooth:

$$|f(y) - f(x) - \langle \nabla f(x) | y - x \rangle| \leq \frac{\beta}{2} \|x - y\|^2 \quad (22)$$

A: Projected gradient

Let $x_{m+1} = \Pi_X(y_{m+1})$, $y_{m+1} = x_m + \eta u_m$ and $u_m = -\nabla f(x_m)$.

Take $\eta = 1/\beta$ and define $v_n = \beta(x_{n+1} - x_n)$

$$f(x_{n+1}) - f(y) \leq \langle v_n, y - x_n \rangle - \frac{1}{2\beta} \|v_n\|^2$$

in particular $f(x_n)$ decreasing and $\{\|v_n\|\} \in \ell^2$.

Values

$$n[f(x_{n+1}) - f(y)] \leq R_n^v(y) - \frac{1}{2\beta} \left\| \sum_{m=1}^n v_m \right\|^2 = \frac{\beta}{2} \|y - x_1\|^2$$

Hence convergence rate of the order $\frac{1}{n}$.

Trajectories

Lemma

Let $y^* \in S$. Then $\|x_n - y^*\|$ decreases.

Lemma

$\{x_n\}$ weakly converge to a point in S .

B: Mirror descent

We follow Bauschke, Bolte and Teboulle [?].

$$\langle \nabla H(x_n) - \lambda \nabla f(x_n) - \nabla H(x_{n+1}) | x - x_{n+1} \rangle \leq 0, \forall x \in X$$

Hypothesis 1:

$$LD_H - D_f \geq 0$$

($LH - f$ convex) If H is strongly convex and f is smooth, there exist L such that this holds.

Values

One has, by H1:

$$f(x) \leq f(y) + \langle \nabla f(z) | x - y \rangle + LD_h(x, z) - D_f(y, z)$$

(the last term is ≤ 0 when f is convex). Take $2\lambda L = 1$

Theorem

Assume f convex, lower bounded.

- 1) $f(x_n)$ is decreasing.
- 2) $\sum D_H(x_{n+1}, x_n) < +\infty$.

$$f(x_n) - f(y) \leq \frac{2L}{n} D_H(y, x_1)$$

Trajectories

Theorem

Assume f convex, S compact $\neq \emptyset$.

1) $y \in S$ implies $D_H(y, x_n)$ decreases.

2) Assume

$H2 : x^k \rightarrow x^* \in S \Rightarrow D_H(x^*, x^k) \rightarrow 0$

$H3 : x^* \in S, D_H(x^*, x^k) \rightarrow 0 \Rightarrow x^k \rightarrow x^*$

Then $\{x_n\}$ converges to a point in S .

C: Dual averaging

Similar results for the values in case 3.

Lu, Freund and Nesterov (2018)

D: Mirror prox

Nemirovski (2004)

Assume g to be β Lipschitz.

Dynamics

x_n gives y_{n+1} via usual MD i.e. $v_n = g(x_n)$

$$\langle \nabla H(x_n) + \lambda g(x_n) - \nabla H(y_{n+1}) - |x - y_{n+1} \rangle \leq 0, \forall x \in X$$

x_n gives x_{n+1} via translated MD i.e. $u_n = g(y_{n+1})$

$$\langle \nabla H(x_n) + \lambda g(y_{n+1}) - \nabla H(x_{n+1}) | x - x_{n+1} \rangle \leq 0, \forall x \in X$$

Values

If H is α strongly convex and $\alpha \geq \lambda \beta$

$$\lambda \sum_{m=1}^n \langle g(y_m) | u - y_m \rangle \leq D_H(u, x_1) - D_H(u, x_n)$$

Acceleration: from discrete to continuous

Nesterov (1983)

$$x_{k+1} = y_k - s \nabla f(y_k)$$

$$y_{k+1} = x_{k+1} + \frac{k}{k+3}(x_{k+1} - x_k)$$

f with Lip gradient L and $s \leq 1/L$

convergence of $f(x_k)$ of the order $O(1/k^2)$ (best bound)

Su, Boyd, Candes (NIPS 2014, JMLR 2016)

$$\ddot{x}_t + \frac{r}{t} \dot{x}_t + \nabla f(x_t) = 0,$$

$r = 3$: continuous version of Nesterov discrete algorithm.

Lyapounov function

$$E(t; y) = \frac{t^2}{r-1} [f(x_t) - f(y)] + \frac{r-1}{2} \|x_t + \frac{t}{r-1} \dot{x}_t - y\|^2$$

For $r = 3$, $E(t; y)$ is decreasing for all y .

If $r > 3$, $E(t; y)$ is decreasing for $y \in S$. In particular

$$f(x_t) - f^* \leq O\left(\frac{1}{t^2}\right)$$

Acceleration: from discrete to continuous

Nesterov (1983)

$$x_{k+1} = y_k - s \nabla f(y_k)$$

$$y_{k+1} = x_{k+1} + \frac{k}{k+3}(x_{k+1} - x_k)$$

f with Lip gradient L and $s \leq 1/L$

convergence of $f(x_k)$ of the order $O(1/k^2)$ (best bound)

Su, Boyd, Candes (NIPS 2014, JMLR 2016)

$$\ddot{x}_t + \frac{r}{t} \dot{x}_t + \nabla f(x_t) = 0,$$

$r = 3$: continuous version of Nesterov discrete algorithm.

Lyapounov function

$$E(t; y) = \frac{t^2}{r-1} [f(x_t) - f(y)] + \frac{r-1}{2} \|x_t + \frac{t}{r-1} \dot{x}_t - y\|^2$$

For $r = 3$, $E(t; y)$ is decreasing for all y .

If $r > 3$, $E(t; y)$ is decreasing for $y \in S$. In particular

$$f(x_t) - f^* \leq O\left(\frac{1}{t^2}\right)$$

Attouch, Chbani, Peypouquet, Redont (Math Pro 2018) extend the analysis

$r \geq 3$ Hilbert space $H + L^1$ perturbation

same speed of convergence for the values (with the same Lyapunov function)

if $r > 3$ weak convergence of the trajectory x_t using energy functions of the form (with real parameters a, b)

$$F(t) = \frac{t^2}{r-1} [f(x_t) - f^*] + \frac{r-1}{2} \|a(x_t - x^*) + \frac{t}{r-1} \dot{x}_t\|^2 + b \|x_t - x^*\|^2$$

leading (for some specific b) to

$$F'(t) \leq (2-a)t[f(x_t) - f^*] - (r-a-1)t\|\dot{x}(t)\|^2$$

in fact for $r > 3$ speed of cv $o(\frac{1}{t^2})$ (May, 2017)

extension non euclidean

Krichene, Bayen, Bartlett (NIPS 2015)

$$F(t; y) = \frac{t^2}{q} [f(x_t) - f(y)] + q[h^*(z_t) - \langle y, z_t \rangle + h(y)]$$

$$F'(t; y) = \frac{2t}{q} [f(x_t) - f(y)] + \frac{t^2}{q} \langle \nabla f(x_t), \dot{x}_t \rangle + q \langle \nabla h^*(z_t) - y, \dot{z}_t \rangle$$

choose

$$\dot{z}_t = -\frac{t}{q} \nabla f(x_t), \quad x_t + \frac{t}{q} \dot{x}_t = \nabla h^*(z_t)$$

$$\begin{aligned} F'(t; y) &= \frac{2t}{q} [f(x_t) - f(y)] - t \langle \nabla f(x_t), -\frac{t}{q} \dot{x}_t + \nabla h^*(z_t) - y \rangle \\ &= \frac{2t}{q} [f(x_t) - f(y)] - t \langle \nabla f(x_t), x_t - y \rangle \leq \frac{2t}{q} [f(x_t) - f(y)] - t [f(x_t) - f(y)] \end{aligned}$$

which is non positive if $q = 2$ or $q > 2$ and $y^* \in S$.

Note: **no condition on** ∇f .

For the euclidean unconstrained case take $h(x) = \frac{1}{2}\|x\|^2$ so that $\nabla h^* = Id$ and one has

$$\frac{d}{dt}\left[x_t + \frac{t}{q}\dot{x}_t\right] = -\frac{t}{q}\nabla f(x_t)$$

which is the SBC equation with $r = q + 1$.

The second equation can be written

$$t^q x_t = q \int_0^t s^{q-1} \nabla h^*(z_s) ds$$

so that x_t is an average of the previous $\nabla h^*(z_s)$

For the euclidean unconstrained case take $h(x) = \frac{1}{2}\|x\|^2$ so that $\nabla h^* = Id$ and one has

$$\frac{d}{dt}\left[x_t + \frac{t}{q}\dot{x}_t\right] = -\frac{t}{q}\nabla f(x_t)$$

which is the SBC equation with $r = q + 1$.

The second equation can be written

$$t^q x_t = q \int_0^t s^{q-1} \nabla h^*(z_s) ds$$

so that x_t is an average of the previous $\nabla h^*(z_s)$

More extension **KBB, (NIPS 2016)**

$$\dot{z}_t = -\eta_t \nabla f(x_t), \quad x_t = \frac{1}{W_t} \int_0^t w_s \nabla h^*(z_s) ds$$

with η and w positive.

new Lyapounov function is of the form

$$E(t) = a_t [f(x_t) - f(y)] + [h^*(z_t) - \langle y, z_t \rangle]$$

and speed of cv $1/a_t$, with compatibility conditions between η , w and a (standard case $a_t = t^2$)

$$E'(t) \leq [f(x_t) - f(y)](a'_t - \eta_t) + \langle \nabla f(x_t), \dot{x}_t \rangle (a_t - \frac{\eta_t W_t}{w_t})$$

Discrete properties

no natural discretization

2 first order equations: choice of coefficients

$$x_{k+1} = y_k - s \nabla f(y_k)$$

$$y_{k+1} = x_{k+1} + \frac{k}{k+r}(x_{k+1} - x_k)$$

discrete Lyapounov function (SBC)

$$E(k) = \frac{2(k+r-2)^2 s}{r-1} [f(x_k) - f^*] + (r-1) \|w_k - x^*\|^2$$

with

$$w_k = \frac{k+r-1}{r-1} y_k - \frac{k}{r-1} x_k$$

satisfies

$$E(k) + \frac{2s[(r-3)(k+r-2)+1]}{r-1} [f(x_k) - f^*] \leq E(k-1)$$

Similar computations in BBK

In addition for $r > 3$:

weak convergence of x_n , Chambolle and Dossal (2015), ACPR (2018)

Attouch and Peypouquet (2016) cv of the value with rate $o(\frac{1}{n^2})$

Open pb:

- link between continuous and discrete:

property of the curve

property of the approximation

- cv of the trajectory in the non euclidean setting

- similar procedure for smooth learning ??
