The Last-Iterate Convergence Rate of Optimistic Mirror Descent in Stochastic VI

Waïss Azizian[1,3], Franck Iutzeler[2,3], Jérôme Malick[2,3,4], Panayotis Mertikopoulos[2,4,5]

[1] ENS, Univ. PSL, [2] Univ. Grenoble Alpes, [3] LJK, [4] CNRS, [5] LIG

Contributions

Interplay between geometry, algorithm and convergence

- Introduce the Legendre exponent to describe the local geometry of a Bregman divergence
- Characterize the convergence of the last-iterate of Optimistic Mirror Descent near the solution
- Derive consequences for the tuning of step-sizes

Variational Inequality

For $\mathcal{K} \subset \mathbb{R}^d$, $v : \mathcal{K} \to \mathbb{R}^d$, find $x^* \in \mathcal{K}$ s.t.

$$\langle v(x^*), x - x^* \rangle \geq 0 \quad \text{for all } x \in \mathcal{K} \qquad \text{(VI)}$$

Example (Minimization)

$$\text{KKT points of } \min_{x \in \mathcal{K}} f(x) \iff \text{(VI) with } v = \nabla f$$

Example (Saddle-point)

$$\text{Stationary points of } \min_{x_1 \in \mathcal{K}_1} \max_{x_2 \in \mathcal{K}_2} \Phi(x_1, x_2) \iff \text{(VI) with } v = \begin{pmatrix} \nabla_{x_1} \Phi \\ -\nabla_{x_2} \Phi \end{pmatrix}$$

Bregman divergences

Bregman divergence: For $h \colon \mathcal{K} \subset \mathbb{R}^d \to \mathbb{R}$ 1-strongly convex

$$D(p, x) = h(p) - h(x) - \langle \nabla h(x), p - x \rangle, \qquad \text{for all } p \in \mathcal{K}, x \in \mathcal{K}$$

Prox-mapping: $P \colon \mathcal{K} \times \mathbb{R}^d \to \mathcal{K}$

$$P_x(y) = \arg\min_{x' \in \mathcal{K}} \{ \langle y, x - x' \rangle + D(x', x) \} \qquad \text{for all } x \in \mathcal{K}, y \in \mathbb{R}^d.$$

Example: on $\mathcal{K} = [0, +\infty)$

	$h(x)$	$D(p,x)$	$P_x(y)$
Euclidean	$\frac{x^2}{2}$	$\frac{(p-x)^2}{2}$	$(x+y)_+$
Entropy	$x \log x$	$p \log \frac{p}{x} + p - x$	$x e^y$
Tsallis entropy, $q > 0$	$\frac{-x^q}{q(1-q)}$	$\frac{(1-q)x^q - p(x^{q-1} - p^{q-1})}{q(1-q)}$	$\dots$

Optimistic Mirror Descent

$$X_{t+1/2} = P_{X_t}(-\gamma_t V_{t-1/2}) \qquad V_{t-1/2} = v(X_{t-1/2}) + \text{err}$$
$$X_{t+1} = P_{X_t}(-\gamma_t V_{t+1/2}) \qquad V_{t+1/2} = v(X_{t+1/2}) + \text{err}$$

Existing results:

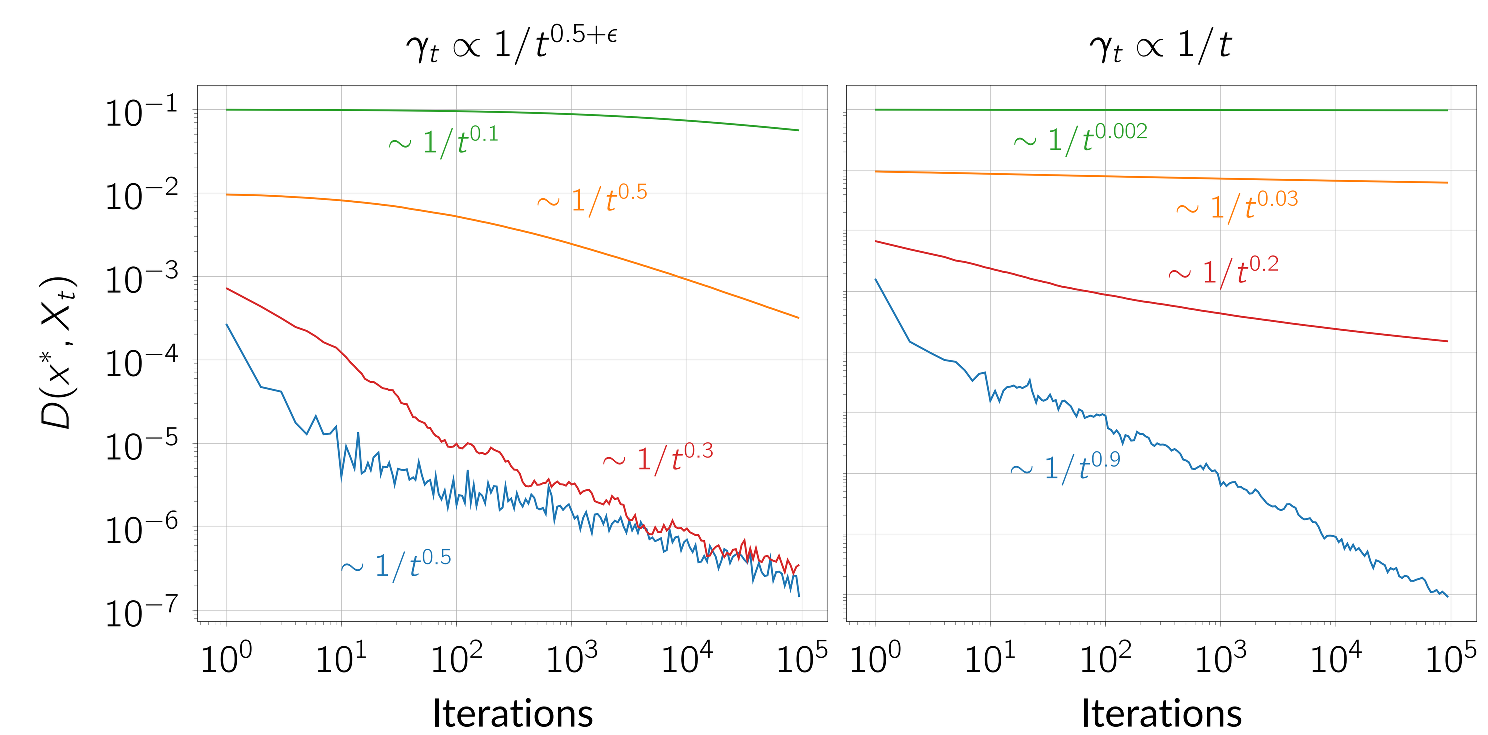
(VI)	Convergence	Setting	Stochastic
Mon.	Ergodic	Bregman	$O(1/\sqrt{t})$ with $\gamma_t \propto 1/\sqrt{t}$
Strongly Mon.	Last-iterate	Only Euclidean	$O(1/t)$ with $\gamma_t \propto 1/t$

(Nemirovski, 2004), (Juditsky et al., 2011, Gidel et al., 2019), (Hsieh et al., 2019)

What happens across divergences?

Example: $v(x) = x$ on $\mathcal{K} = [0, +\infty)$

— Euclidean — Tsallis entropy ($q = 0.5$)
— Entropy — Tsallis entropy ($q = 1.5$)



Question

How can we explain those differences in last-iterate convergence between divergences?

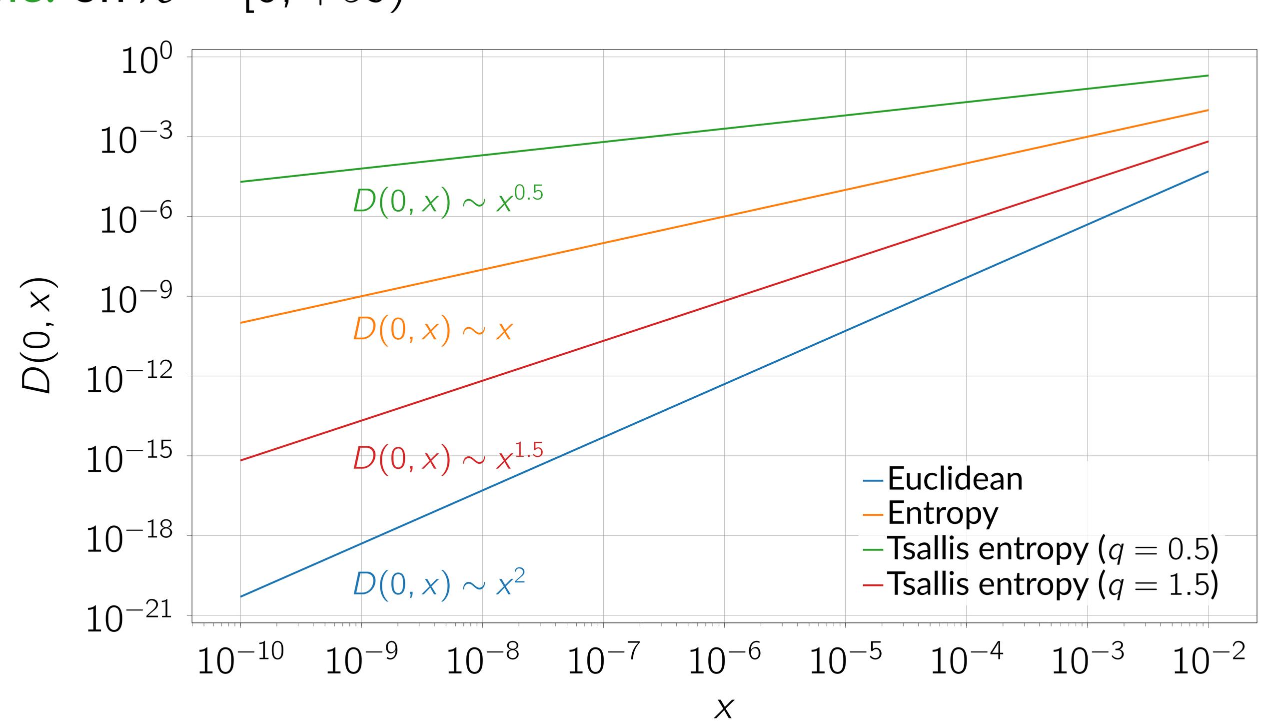
The Bregman topology

- Strong convexity of h: for all $p \in \mathcal{K}, x \in \mathcal{K}$

$$D(p, x) \geq \frac{1}{2} \|p - x\|^2$$

- Reverse does not hold in general:

Example: on $\mathcal{K} = [0, +\infty)$



Our proposal: quantify the deficit of regularity

Legendre exponent of h at $p \in \mathcal{K}$: $\beta \in [0, 1)$ s.t., for some $\kappa \geq 0$

$$\text{for all } x \text{ close to } p, \quad D(p, x) \leq \frac{1}{2} \kappa \|p - x\|^{2(1-\beta)}$$

Example: on $\mathcal{K} = [0, +\infty)$

	$p > 0$ (interior)	$p = 0$ (boundary)
Euclidean reg.	0	0
Entropy	0	1/2
Tsallis entropy $q \leq 2$	0	$1 - q/2$

Legendre exponent β

Assumptions and iterate stability

Oracle signal: $(U_t)_t$ zero-mean and with finite-variance,

$$V_t = v(X_t) + U_t$$

Lipschitz continuity:

$$\|v(x') - v(x)\|_* \leq L \|x' - x\| \quad \text{for all } x, x' \in \mathcal{K}.$$

Second-order sufficiency: there exists $\mu > 0$ s.t.,

$$\langle v(x), x - x^* \rangle \geq \mu \|x - x^*\|^2 \quad \text{for all } x \text{ close to } x^*.$$

Proposition

Take a step-size of the form $\gamma_t = \gamma/(t + t_0)^\eta$ with $\eta \in (1/2, 1]$ and $\gamma, t_0 > 0$ and fix any confidence level $\delta > 0$,

For every neighborhood $\mathcal{U}$ of x^*, if γ/t_0 is small enough and X_1 is close enough to x^*, then

$$\mathcal{E}_{\mathcal{U}} = \{X_t \in \mathcal{U} \text{ for all } t = 1, 2, \dots\}$$

happens with probability at least $1 - \delta$.

Last-iterate convergence

Legendre exponent: For all x close to x^*,

$$D(x^*, x) \leq \frac{1}{2} \kappa \|x^* - x\|^{2(1-\beta)}$$

Theorem

If $\mathcal{U}$ is small enough, with step-sizes of the form, $\gamma_t = \gamma/(t + t_0)^\eta$, $\mathbb{E}[D(x^*, X_t) \mid \mathcal{E}_{\mathcal{U}}]$ is bounded according to the following table:

Legendre exponent	Rate ($\eta = 1$)	Rate ($\frac{1}{2} < \eta < 1$)	Examples
$\beta = 0$	$\mathcal{O}(1/t)$	$\mathcal{O}(1/t^\eta)$	Euclidean, Interior
Conditions:	γ large enough	–	
$\beta \in (0, 1)$	$\mathcal{O}\left((\log t)^{-\frac{1-\beta}{\beta}}\right)$	$\mathcal{O}\left(t^{-\frac{(1-\eta)(1-\beta)}{\beta}} + t^{-\eta}\right)$	Entropy, Tsallis
Conditions:	γ small enough		

Optimal step-size:

Legendre exp.	η^*	Rate
$\beta \in [0, 1/2)$	$1 - \beta$	$\mathcal{O}\left(t^{-(1-\beta)}\right)$
$\beta \in [1/2, 1]$	$\approx 1/2$	$\mathcal{O}\left(t^{-\frac{1}{2\beta}}\right)$

Reduced bibliography

G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR '19*, 2019.

Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS '19*, 2019.

A. Juditsky, A. S. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 2011.

A. S. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 2004.