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

Waïss Azizian, Franck Iutzeler, Jérôme Malick, Panayotis Mertikopoulos

ICCOPT 2022

Variational Inequality

For $\mathcal{K} \subset \mathbb{R}^d$, $v : \mathcal{K} \to \mathbb{R}^d$,

$$\text{Find } x^* \in \mathcal{K} \text{ such that } \langle v(x^*), x - x^* \rangle \geq 0 \text{ for all } x \in \mathcal{K}. \tag{VI}$$

Example (Minimization)

$$\text{Karush–Kuhn–Tucker (KKT) points of } \min_{x \in \mathcal{K}} f(x) \iff \text{(VI) with } v = \nabla f\,.$$

Example (Saddle-point)

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

Example

In particular : games, adversarial training in ML

Classical methods in the unconstrained case $\mathcal{K} = \mathbb{R}^d$

Gradient method:

$$X_{t+1} = X_t - \gamma_t V_t \qquad\qquad V_t = v(X_t)$$

$\rightarrow$ Good convergence properties for large classes of VI, but fails on e.g., bilinear games

Extragradient (Korpelevich, 1976):

$$\begin{aligned} X_{t+1/2} &= X_t - \gamma_t V_t & V_t &= v(X_t) \quad , \\ X_{t+1} &= X_t - \gamma_t V_{t+1/2} & V_{t+1/2} &= v(X_{t+1/2}) \end{aligned}$$

$\rightarrow$ Better convergence properties, but requires two evaluations of v per iteration

Optimistic Gradient Method (Popov, 1980):

$$\begin{aligned} X_{t+1/2} &= X_t - \gamma_t V_{t-1/2} & V_{t-1/2} &= v(X_{t-1/2}) \\ X_{t+1} &= X_t - \gamma_t V_{t+1/2} & V_{t+1/2} &= v(X_{t+1/2}) \end{aligned}$$

Classical methods in the unconstrained case $\mathcal{K} = \mathbb{R}^d$

Gradient method:

$$X_{t+1} = X_t - \gamma_t V_t \qquad\qquad V_t = v(X_t) + \mathrm{err.}$$

$\to$ Good convergence properties for large classes of VI, but fails on e.g., bilinear games

Extragradient (Korpelevich, 1976):

$$\begin{aligned} X_{t+1/2} &= X_t - \gamma_t V_t & V_t &= v(X_t) + \mathrm{err.}\,, \\ X_{t+1} &= X_t - \gamma_t V_{t+1/2} & V_{t+1/2} &= v(X_{t+1/2}) + \mathrm{err.} \end{aligned}$$

$\to$ Better convergence properties, but requires two evaluations of v per iteration

Optimistic Gradient Method (Popov, 1980):

$$\begin{aligned} X_{t+1/2} &= X_t - \gamma_t V_{t-1/2} & V_{t-1/2} &= v(X_{t-1/2}) + \mathrm{err.} \\ X_{t+1} &= X_t - \gamma_t V_{t+1/2} & V_{t+1/2} &= v(X_{t+1/2}) + \mathrm{err.} \end{aligned}$$

$+$ Stochastic error err, e.g., in large scale ML

Bregman divergences

Constraint set: $\mathcal{K} \neq \mathbb{R}^d$, e.g., $\mathcal{K} =$ simplex in games

Bregman divergence: For $h\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ 1-strongly convex with $\operatorname{dom} h = \mathcal{K}$

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

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

$$P_x(y) = \operatorname*{arg\,min}_{x' \in \mathcal{K}} \{\langle y, x - x' \rangle + D(x', x)\} \qquad \text{for all } x \in \mathcal{K}, y \in \mathcal{Y}.$$

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	$x \log x$	$p \log \frac{p}{x} + p - x$	xe^y
Tsallis entropy, $q > 0$	$\frac{-x^q}{q(1-q)}$	$\frac{(1-q)x^q - p(x^{q-1} - p^{q-1})}{q(1-q)}$	Explicit

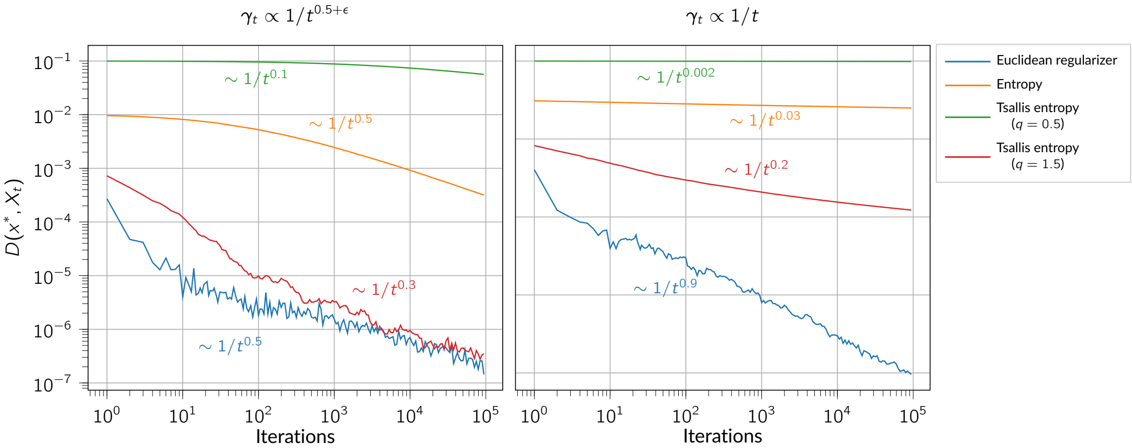
Optimistic Mirror Descent:

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

What happens across divergences?

Example

$v(x) = x$ on $\mathcal{K} = [0, +\infty)$ and $V_t = v(X_t) + \mathcal{N}(0, \sigma^2 I_d)$



Convergence of Optimistic Mirror Descent/Mirror-Prox

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

Existing results:

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

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

The Bregman topology

- By the strong convexity of h,

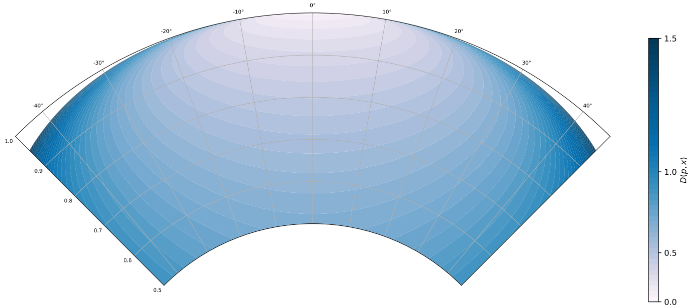
$$D(p,x) = h(p) - h(x) - \langle \nabla h(x), p - x \rangle \geq \tfrac{1}{2}\|p - x\|^2 \quad \text{for all } p \in \mathcal{K}, x \in \mathcal{K}.$$

  Consequence: $D(p, x_t) \to 0 \implies \|x_t - p\| \to 0$.
- Conversely consider,

$$\mathcal{K} = \{x \in \mathbb{R}^2 : \|x\|_2 \leq 1\}, \quad h(x) = -\sqrt{1 - \|x\|_2^2}.$$

  There exists $(x_t)_t$ s.t. $\|x_t - p\| \to 0$ but $D(p, x_t) \not\to 0$

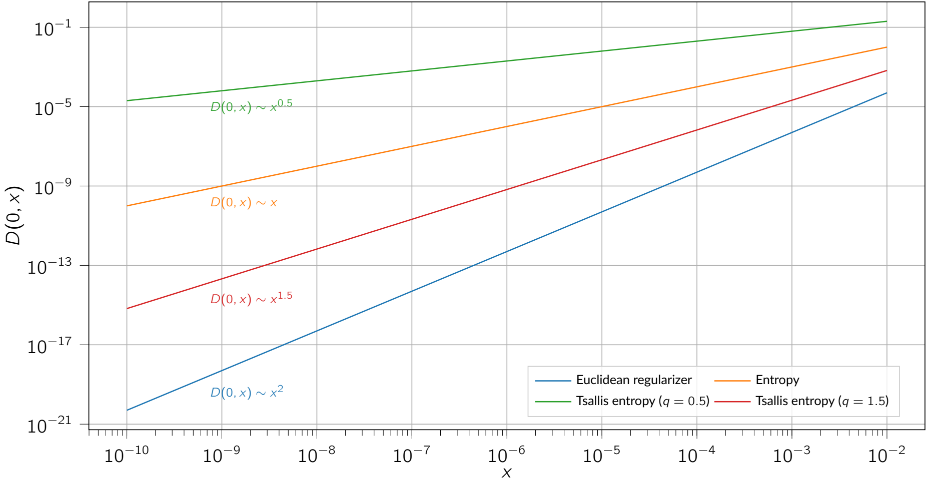
$D(p,x)$ for fixed p s.t. $\|p\| = 1$



The topology of several standard divergences

Example

On $\mathcal{K} = [0, +\infty)$.



Our proposal: quantify the deficit of regularity w.r.t. ambient norm

Definition

The Legendre exponent of h at $p \in \mathcal{K}$ is the smallest $\beta \in [0, 1)$ such, for some $\kappa \geq 0$ and for all x close enough to p,

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

$\rightarrow$ *Local* notion around p in $\mathcal{K}$

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

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

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

Proof: using tools from Hsieh et al. (2019)

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

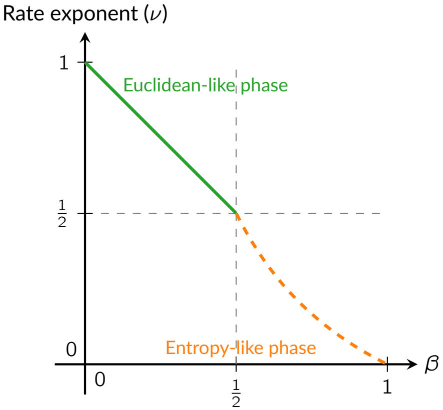
If $\mathcal{U}$ is small enough, with step-sizes of the form, $\gamma_t = \gamma/(t+t_0)^\eta$, $\mathbb{E}[D(x^*, X_t) \mid \mathcal{E}_\mathcal{U}]$ is bounded according to the following table and conditions:

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*		

Best step-size schedule

Two regimes:

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

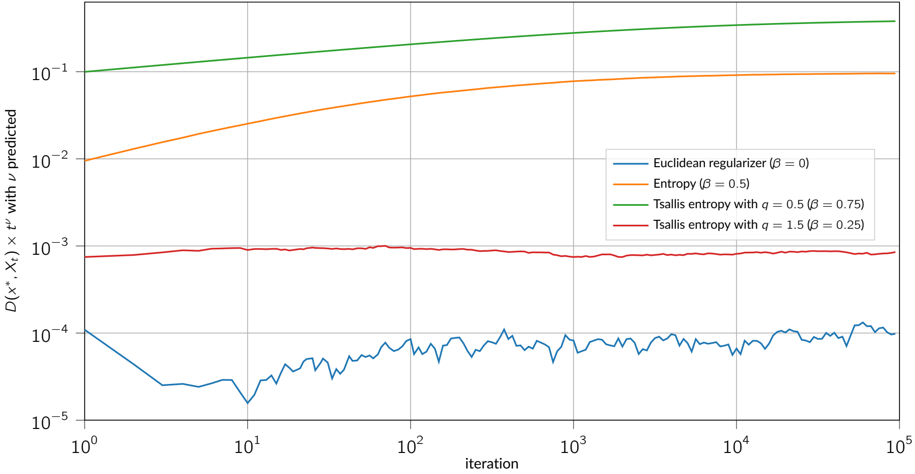


Predicted rate $\mathcal{O}(1/t^{\nu})$ vs. β

Predicted rates vs. observed rates on the simple example

Example

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



$D(x^*, X_t) \times t^\nu$ with ν predicted

Conclusion

Take-home message: Interplay between geometry, algorithm and convergence

- Introduce the Legendre exponent which characterizes the local geometry of the Bregman near a solution
- Characterize the convergence of the last-iterate near the solution
- Derive consequence for the tuning of the step-size

Perspectives: Can we refine the analysis of this interplay?

- Using the structures of the constraints?
- Deterministic setting?
- Other algorithms?

…

Preview of our current work in the deterministic setting

Deterministic setting: broader variety of behaviors!

1. General convergence result:

Legendre exponent	Rate	Examples
$\beta = 0$	Linear	Euclidean, Interior
$\beta \in (0, 1)$	$\mathcal{O}\left(t^{1/\beta - 1}\right)$	Entropies on the boundary

2. When x^* on the boundary of $\mathcal{K}$ and linear constraints, finer guarantees on the convergence of active constraints.

Bibliography I

G. Gidel, R. A. Hemmat, M. Pezehski, R. L. Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. In *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6936–6946, 2019.

A. Juditsky, A. S. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody*, 12:747–756, 1976.

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*, 15(1):229–251, 2004.

L. D. Popov. A modification of the Arrow–Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.