The interplay between geometry and convergence in Bregman proximal methods

Waïss Azizian (supervised by Franck Iutzeler, Jérôme Malick, Panayotis Mertikopoulos)

SMAI MODE 2024

Variational Inequality

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

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

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)

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

Classical methods

Gradient method: $\mathcal{K} = \mathbb{R}^d$

$$X_{t+1} = X_t - \gamma v(X_t)$$

Projected gradient method:

$$X_{t+1} = \mathrm{proj}_{\mathcal{K}}(X_t - \gamma v(X_t))$$

Multiplicative weight update: $\mathcal{K} = \text{simplex}$

$$X_{t+1,i} \propto X_{t,i} e^{-\gamma v(X_t)_i}$$

Classical methods

Gradient method: $\mathcal{K} = \mathbb{R}^d$

$$X_{t+1} = X_t - \gamma v(X_t)$$

Projected gradient method:

$$X_{t+1} = \mathrm{proj}_{\mathcal{K}}(X_t - \gamma v(X_t))$$

Multiplicative weight update: $\mathcal{K} =$ simplex

$$X_{t+1,i} \propto X_{t,i} e^{-\gamma v(X_t)_i}$$

Mirror Descent:

$$X_{t+1} = P_{X_t}(-\gamma v(X_t))$$

Bregman divergences, prox-mapping

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: in one dimension

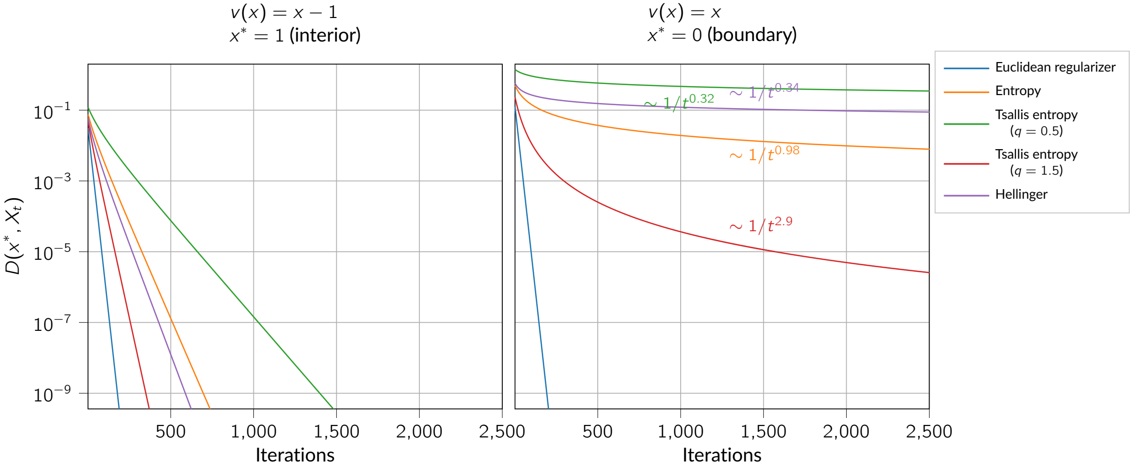
	$\mathcal{K}$	$h(x)$	$D(p, x)$	$P_x(y)$
Euclidean	$[0, +\infty)$	$\frac{x^2}{2}$	$\frac{(p-x)^2}{2}$	$(x+y)_+$
Entropy	$[0, +\infty)$	$x \log x$	$p \log \frac{p}{x} + p - x$	xe^y
Tsallis entropy, $q > 0$	$[0, +\infty)$	$\frac{-x^q}{q(1-q)}$	$\frac{(1-q)x^q - p(x^{q-1} - p^{q-1})}{q(1-q)}$	Explicit
Hellinger	$[-1, 1]$	$-\sqrt{1 - x^2}$	Explicit	Explicit

Mirror Descent:

$$X_{t+1} = P_{X_t}(-\gamma v(X_t))$$

What happens across divergences?

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



Convergence of Mirror Methods

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)

Convergence of Mirror Methods

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 $\gamma_t \propto 1/t$

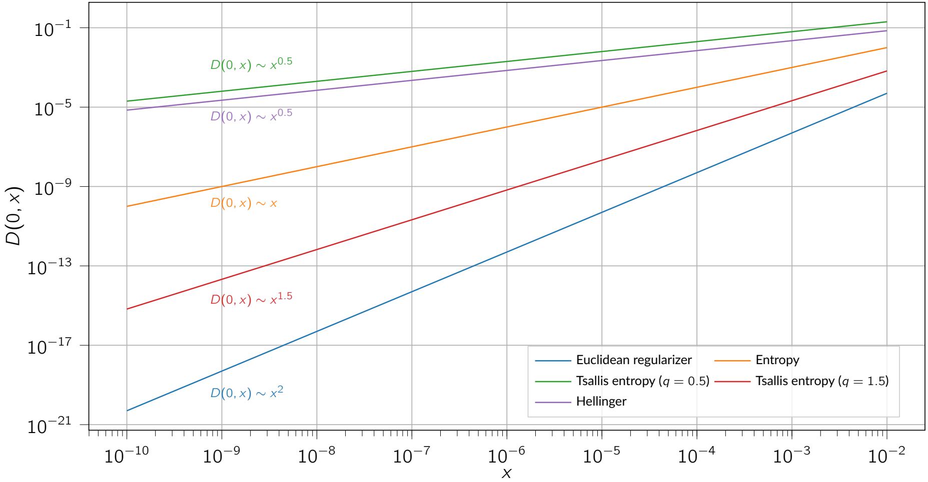
(Nemirovski, 2004), (Juditsky et al., 2011, Gidel et al., 2019), (Hsieh et al., 2019)

Our contribution

For locally strongly monotone (VI), *characterization of the last-iterate convergence of Mirror methods*

The topology of several standard divergences

Plot $D(0,x)$ on $[0,+\infty)$



“Degenerate” Bregman geometry

- Since h is strongly convex,

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

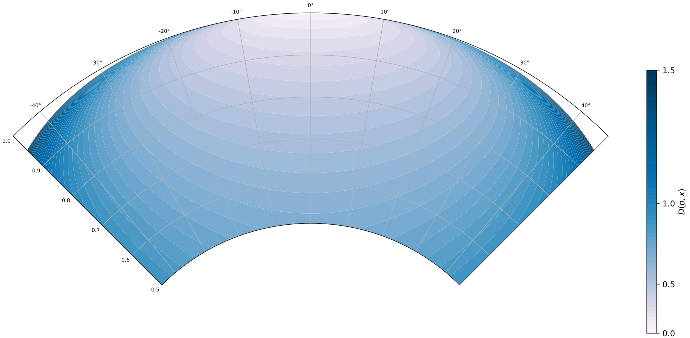
  Consequence: $D(p, x_t) \to 0 \implies \|x_t - p\| \to 0$.

- 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. $\|p\| = 1$



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

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

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

$\rightarrow$ *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	$1 - q/2$
Hellinger	0	$3/4$

Legendre exponent β

Last-iterate convergence

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.,

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

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

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

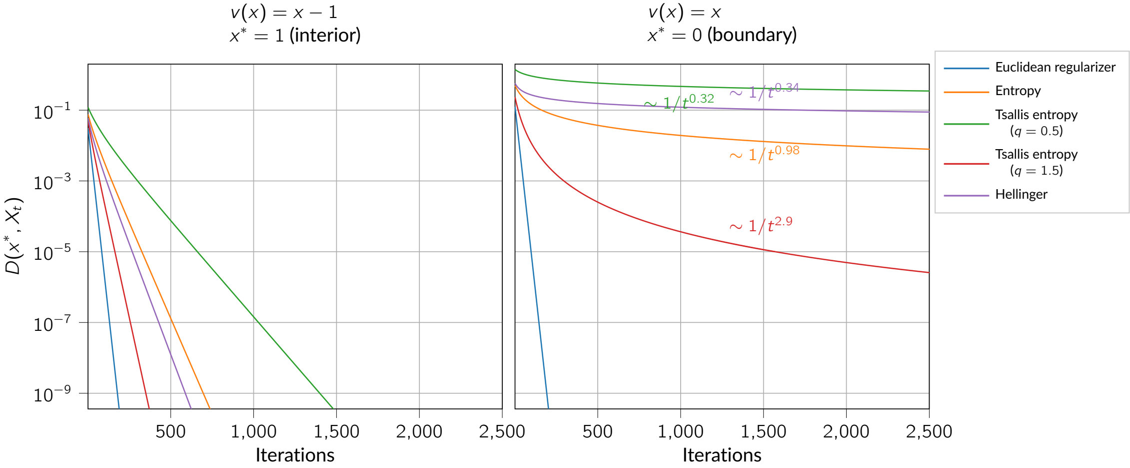
Theorem 1

With classical step-sizes, if X_1 is close enough to x^*,

$$D(x^*, X_t) \leq \begin{cases} \mathcal{O}\left(e^{-\frac{\gamma\mu t}{2K}}\right) & \text{if } \beta = 0 \\ \mathcal{O}\left(1/t^{1/\beta - 1}\right) & \text{if } \beta \in (0, 1) \end{cases}$$

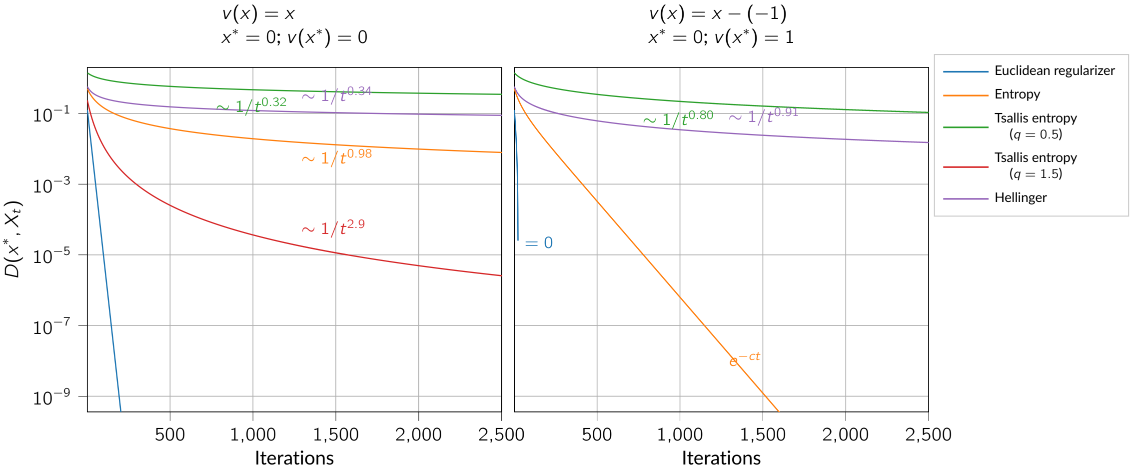
What happens across divergences?

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



What happens across divergences? Bis

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



Finer rates for linearly constrained problems

(Standard form) Polyhedron:

$$\mathcal{K} = \{x \in \mathbb{R}^d : x \geq 0,\ Ax = b\}$$

Decomposable Bregman regularizer:

$$h(x) = \sum_{i=1}^{d} \theta(x_i) \quad \text{with} \quad \theta : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$$

Key quantity: dual multipliers $\lambda_i \geq 0$ associated to the constraints $x_i \geq 0$.

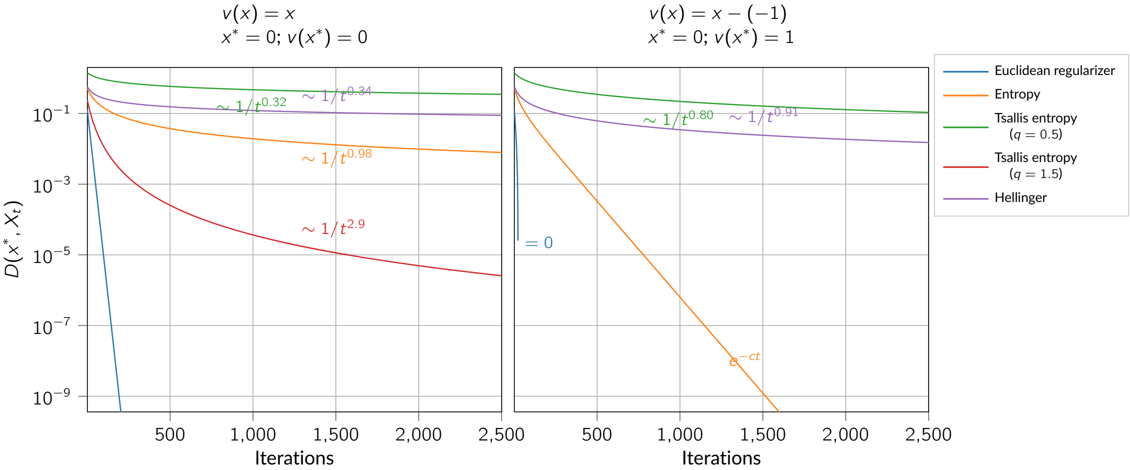
Theorem 2

For i such that $x_i^* = 0$ and $\lambda_i > 0$,

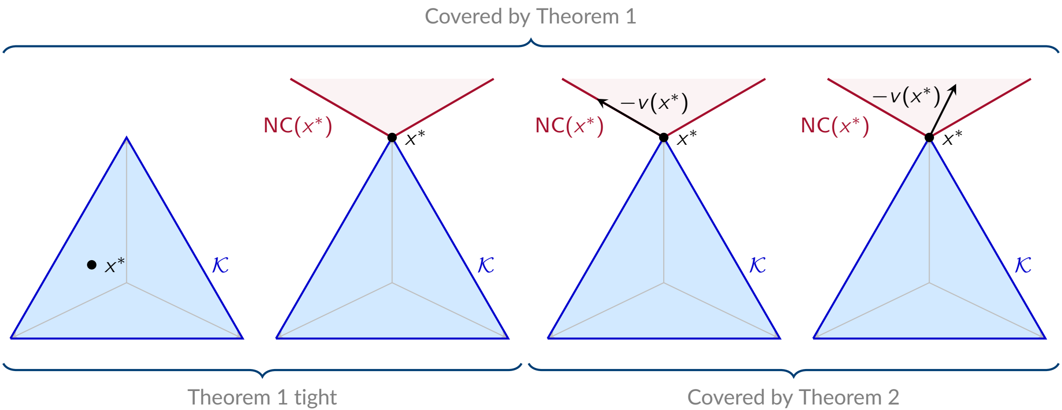
$$X_{t,i} = \begin{cases} 0 \text{ in finite time} & \text{if } \theta'(s) = \Omega(1) \text{ at } 0 \\ \mathcal{O}\left(e^{-\nu t}\right) & \text{if } \theta'(s) = \log s + \Omega(1) \text{ at } 0 \\ \mathcal{O}\left(1/t^{1/p}\right) & \text{if } \theta'(s) = \Omega(s^p),\ p \in (0,1) \text{ at } 0 \end{cases}$$

What happens across divergences? Bis

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



Different situations



Conclusion

Take-home message: Convergence of Mirror methods is more complex than just $\mathcal{O}(1/t)$, depends on

- Local behavior of the Bregman divergence
- Structrure of the constraints and of the solution

Azizian, W., Iutzeler, F., Malick, J. and Mertikopoulos, P., 2021, July. The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In Conference on Learning Theory (pp. 326-358). PMLR.

Azizian, W., Iutzeler, F., Malick, J. and Mertikopoulos, P., 2022. The rate of convergence of Bregman proximal methods: Local geometry vs. regularity vs. sharpness. arXiv preprint arXiv:2211.08043.

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.

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.