Bregman Gradient Methods for Relatively-Smooth Optimization

PhD defence

Transfagarasan road, Romania

Radu-Alexandru Dragomir,
Université Toulouse 1 Capitole,
D.I. Ecole normale supérieure.

Directed by Jérôme Bolte and Alexandre d'Aspremont.

Joint work with Adrien Taylor, Dmitrii Ostrovskii, Hadrien Hendrikx, Mathieu Even.

September 14, 2021

Large-scale optimization

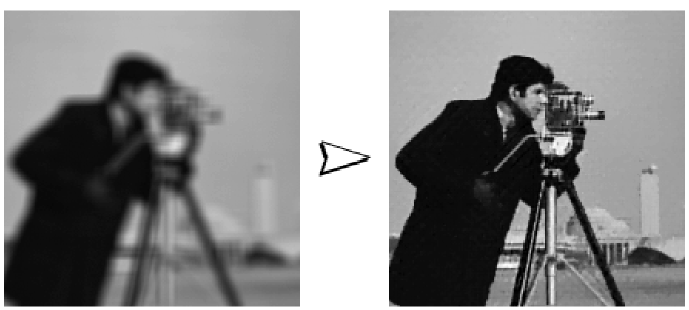
We want to solve

$$\min_{x \in \mathcal{C}} f(x) \tag{P}$$

where $\mathcal{C}$ is a convex set of $\mathbb{R}^d$, $d \gg 1$.

Signal processing

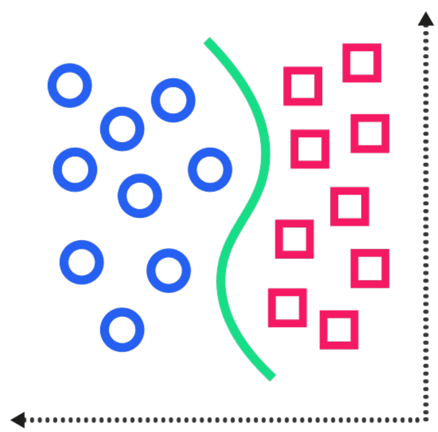
Recovery of unknown signal from partial and noisy observations



Source: LASIP toolbox

Machine learning

Learning a prediction function from training data



Source: ipullrank.com

Our objective

$$\min_{x \in \mathcal{C}} f(x) \qquad \text{(P)}$$

- **Iterative methods:** solve a series of subproblems to compute a sequence

$$x_0, x_1, x_2, \dots x_k \dots$$

  which approaches the solution x_*.

- **First-order methods:** for large-scale problems, the algorithm has only cheap access to **first-order oracle**

$$x \mapsto \Big(f(x), \nabla f(x)\Big).$$

- In practice, f is not a **black box**: use problem structure to devise **efficient** algorithms, with **theoretical guarantees**.

- **Our approach:** Bregman methods and relatively-smooth optimization.

$$\nabla^2 f \preceq L \nabla^2 h \quad \text{(Bauschke, Bolte, Teboulle, 2017)}$$

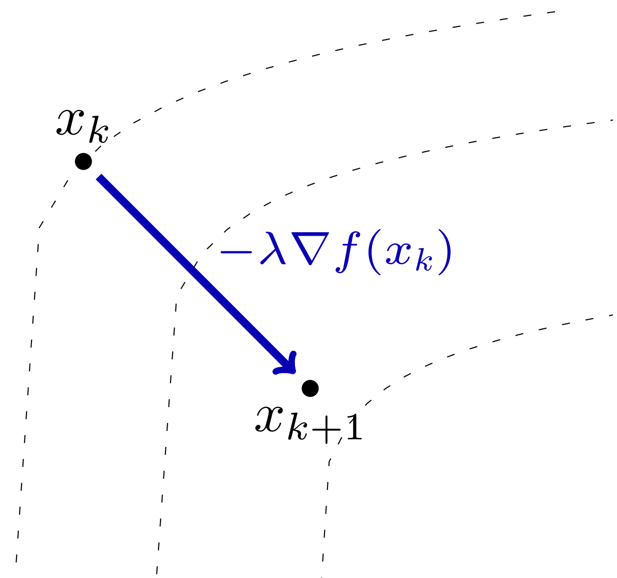
Outline

- **Bregman gradient methods and relative smoothness**
- Application to low-rank minimization
- Theoretical complexity: lower bound and computer-aided analyses
- Stochastic variants

Gradient descent

$$x_{k+1} = \Pi_{\mathcal{C}}\left[x_k - \lambda \nabla f(x_k)\right] \quad \text{(GD)}$$

λ is the step size, $\Pi_{\mathcal{C}}$ denotes projection on $\mathcal{C}$.



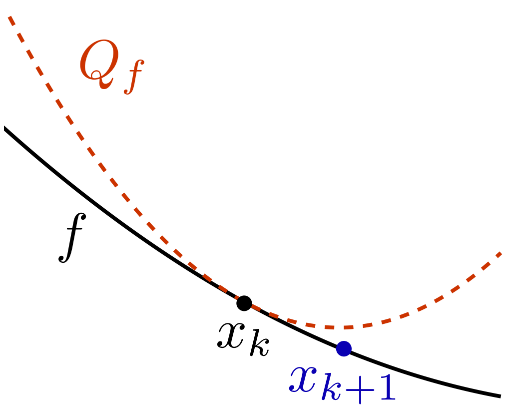
Smoothness

$$x_{k+1} = \operatorname*{argmin}_{u \in \mathcal{C}} f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{2\lambda}\|u - x_k\|^2 \qquad \text{(GD)}$$

GD iteratively minimizes a **quadratic approximation** of f: when is it accurate?

Smoothness assumption: if f has a L-Lipschitz continuous gradient, then for every $\lambda \in (0, 1/L]$,

$$f(u) \leq f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{2\lambda}\|u - x_k\|^2.$$



The quadratic model is an upper approximation of f.

Bregman gradient descent

Are we limited to a quadratic model? A more general method is

$$x_{k+1} = \operatorname*{argmin}_{u \in \mathcal{C}} f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\lambda} D_h(u, x_k) \tag{BGD}$$

where

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle \geq 0$$

is the **Bregman divergence** induced by some strictly convex **kernel** function h adapted to $\mathcal{C}$.

Examples:

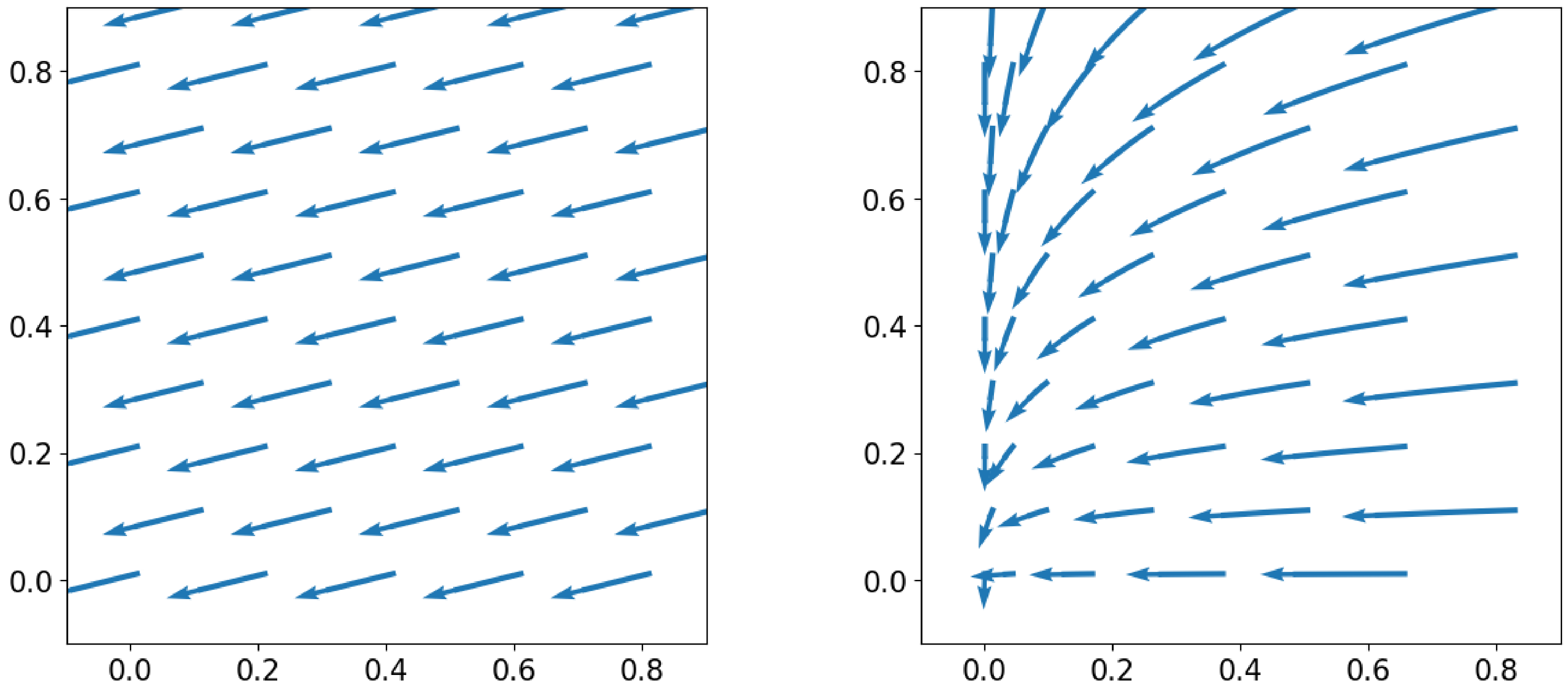
- **Euclidean:** $h(x) = \frac{1}{2}\|x\|^2$: then $D_h(x, y) = \frac{1}{2}\|x - y\|^2$,
- **Entropy:** $h(x) = \sum_{i=1}^{d} x^i \log(x^i) - x^i$, then $D_h = D_{\mathrm{KL}}$ and (BGD) writes

$$x_{k+1} = x_k \cdot \exp[-\lambda \nabla f(x_k)],$$

Also called Mirror descent / NoLips...

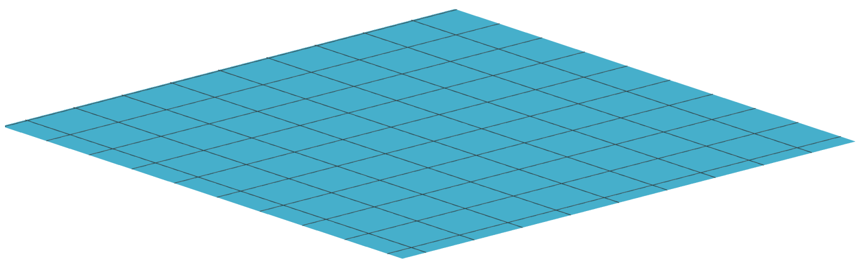
Effect of Bregman divergence

Comparing the Bregman update with $\nabla f(x_k) = (4, 1)$ from different starting points and kernel functions:

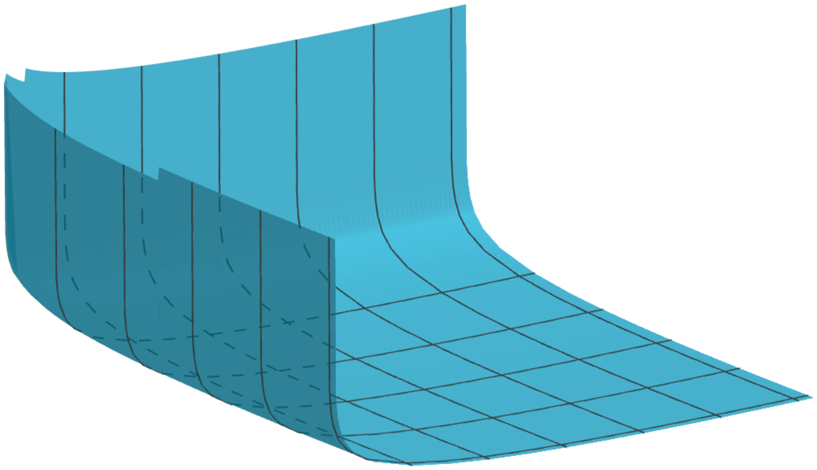


(a) Euclidean (b) Entropy

Effect of Bregman divergence



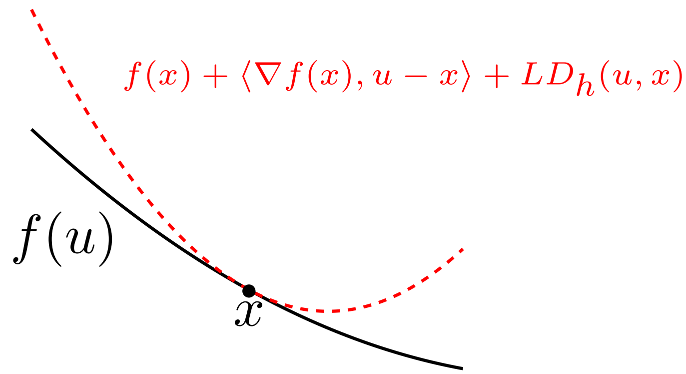
(c) Euclidean



(d) Entropy

Relative smoothness

(Bauschke, Bolte, Teboulle, 2017)



f is **L-smooth relative** to the kernel function h if

$$f(u) \le f(x) + \langle \nabla f(x), u - x \rangle + LD_h(u, x).$$

For C^2 functions, equivalent to

$$\nabla^2 f(x) \preceq L \nabla^2 h(x).$$

Similarly, **relative strong convexity** is defined as (Lu, Freund, Nesterov, 2018):

$$\mu \nabla^2 h(x) \preceq \nabla^2 f(x).$$

Example of relatively-smooth function

Linear inverse problems with Poisson noise (Bauschke et al., 2017): let $b \in \mathbb{R}^m, A \in \mathbb{R}_+^{m \times d}$,

$$\min_{x \in \mathbb{R}_+^n} D_{KL}(b, Ax) = \sum_{j=1}^{m} b_j \log\big(\frac{b_j}{A_j x}\big) - A_j x + b_j$$

Applications in medical imaging, astronomy...

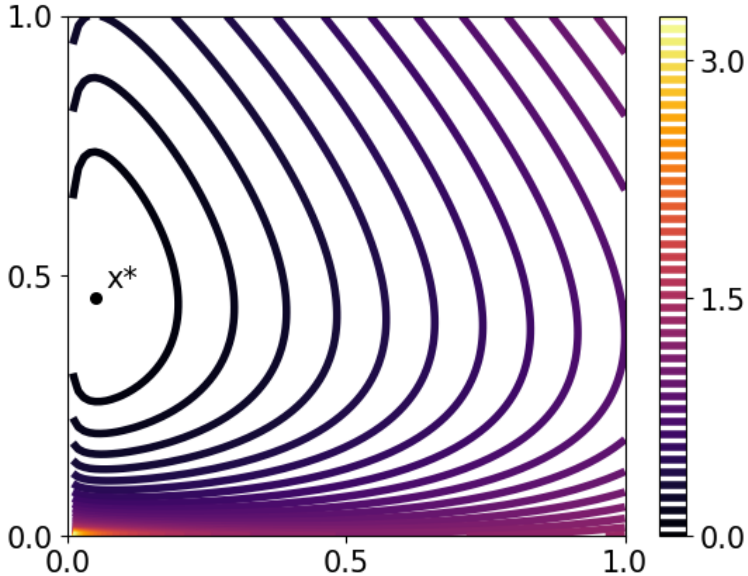


Figure 1: Example for $d = 2$

Standard smoothness does not hold as the Hessian is singular when $A_j x \to 0$, but relative smoothness holds with

$$h(x) = \sum_{i=1}^{d} -\log(x^i).$$

Convergence guarantees

If f is L-smooth relative to h, then BGD with step size $\lambda = 1/L$ satisfies:

- If f is **convex** (Bauschke, Bolte, Teboulle, 2017):

$$f(x_N) - f(x_*) \leq \frac{L D_h(x_*, x_0)}{N}$$

- If f is μ-**strongly convex relative to** h (Lu, Freund, Nesterov 2018):

$$f(x_N) - f(x_*) \leq L \left(1 - \frac{\mu}{L}\right)^N D_h(x_*, x_0)$$

- If f is **non-convex** (Bolte et al., 2018):
  - the sequence $\{f(x_k)\}$ is nonincreasing,
  - if $\mathcal{C} = \mathbb{R}^d$ and f satisfies the *Kurdyka–Lojasiewicz property:* the sequence $\{x_k\}$ converges to a critical point.

How to choose the kernel in practice?

$$x_{k+1} = \underset{u \in \mathcal{C}}{\operatorname{argmin}} \, f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\lambda} D_h(u, x_k) \qquad \text{(BGD)}$$

We seek h such that

- the inner objective in (BGD) is a **good approximation** of f, the inequality

$$\nabla^2 f(x) \preceq L \nabla^2 h(x)$$

  holds as tightly as possible;

- the inner minimization problem can be solved easily.

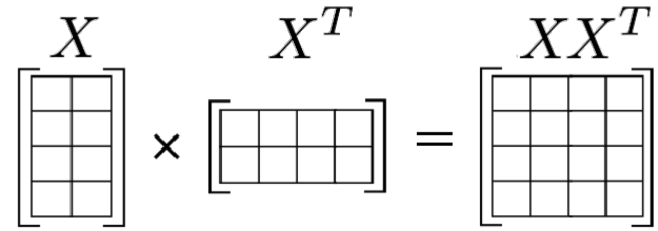
There is often a tradeoff between these two goals!

Outline

- Bregman gradient methods and relative smoothness
- **Application to low-rank minimization**
- Theoretical complexity: lower bound and computer-aided analyses
- Stochastic variants

Non-convex low-rank minimization

$$\min_{X \in \mathbb{R}^{n \times r}} \underbrace{\mathcal{L}(XX^T)}_{\text{differentiable error function}} + \underbrace{g(X)}_{\text{nonsmooth penalty}}$$



$r \in \mathbb{N}$ is the **target rank**, $\mathcal{L}$ is a L_1-smooth error function (typically a quadratic),

- **Example:** symmetric nonnegative matrix factorization

$$\min_{X \in \mathbb{R}^{n \times r}} \|XX^T - M\|^2 \quad \text{subject to } X \geq 0.$$

- $f(X) = \mathcal{L}(XX^T)$ is **not globally smooth** (typically quartic) $\rightarrow$ standard Euclidean methods might not be adapted.

Objective

Design kernels h adapted to f by leveraging the **quartic** structure, and apply Bregman *proximal* gradient method

$$X_{k+1} = \underset{U \in \mathcal{C}}{\text{argmin}}\, f(X_k) + \langle \nabla f(X_k), U - X_k \rangle + \frac{1}{\lambda} D_h(U, X_k) + g(U) \qquad \text{(BPG)}$$

Two different kernels

The "simple" norm kernel

$$h_n(X) = \frac{\alpha}{4}\|X\|^4 + \frac{\sigma}{2}\|X\|^2.$$

Proposition (D., d'Aspremont, Bolte, 2021): f is 1-smooth relative to h_n for α, σ high enough.

- **Bregman update:** easy (computing $\nabla F(X_k)$ + simple scalar equation).

The "more refined" Gram kernel

$$h_G(X) = \frac{\alpha}{4}\|X\|^4 + \frac{\beta}{4}\|X^T X\|^2 + \frac{\sigma}{2}\|X\|^2.$$

Proposition (D., d'Aspremont, Bolte, 2021): f is 1-smooth relative to h_G for α, β, σ high enough.

- **Better approximation** of f than h_n for well-conditionned $\mathcal{L}$;
- **Bregman update:** harder. Computable only for unpenalized problems $(g = 0)$ and requires solving a subproblem of dimension r (the target rank).

Experiments: Distance Matrix Completion

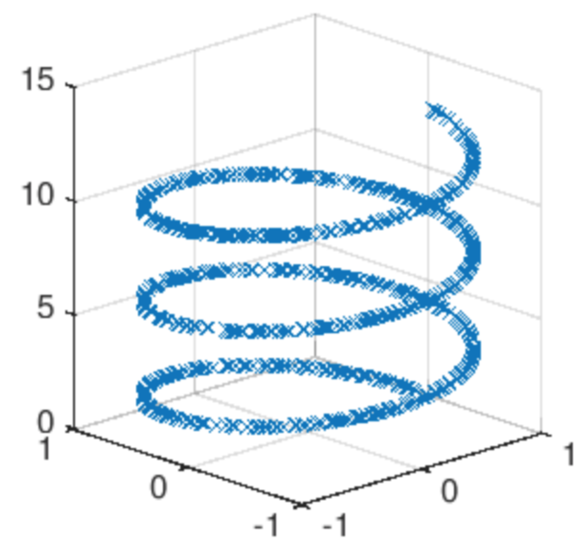
Recover the position of n points $X_1^*, \ldots, X_n^*$ in $\mathbb{R}^r$ from an incomplete set of pairwise distances

$$\{d_{ij} = \|X_i^* - X_j^*\|^2 \,|\, (i,j) \in \Omega\}.$$

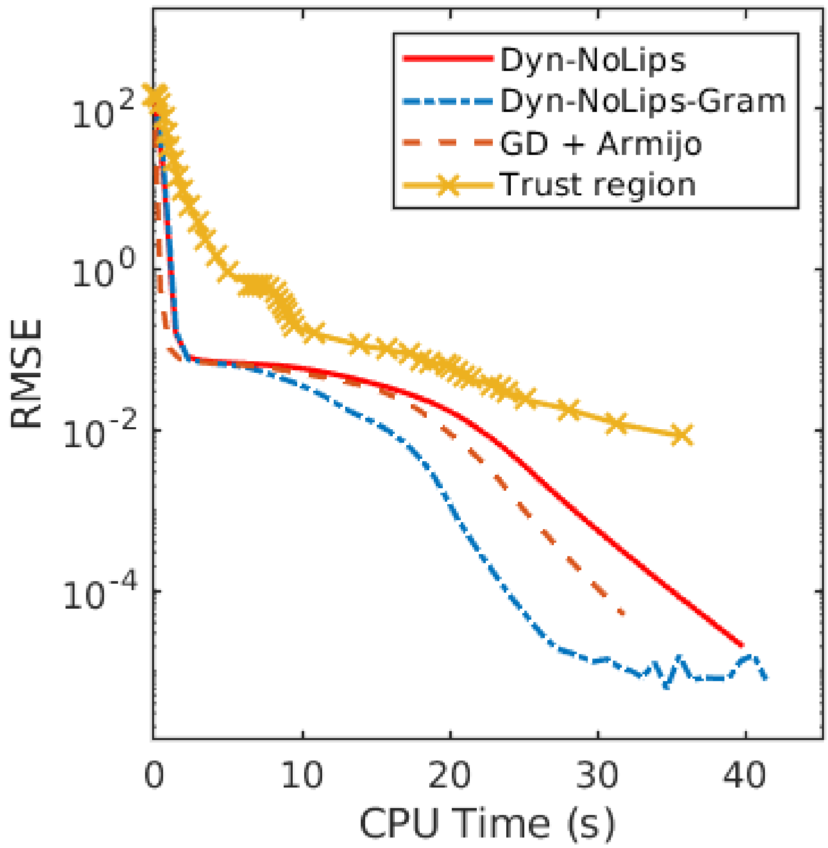
$$\min_{X \in \mathbb{R}^{n \times r}} f(X) = \sum_{(i,j) \in \Omega} \left(\|X_i - X_j\|^2 - d_{ij}\right)^2 \tag{EDMC}$$

Unconstrained problem: we compare the norm kernel h_n with the Gram kernel h_G.

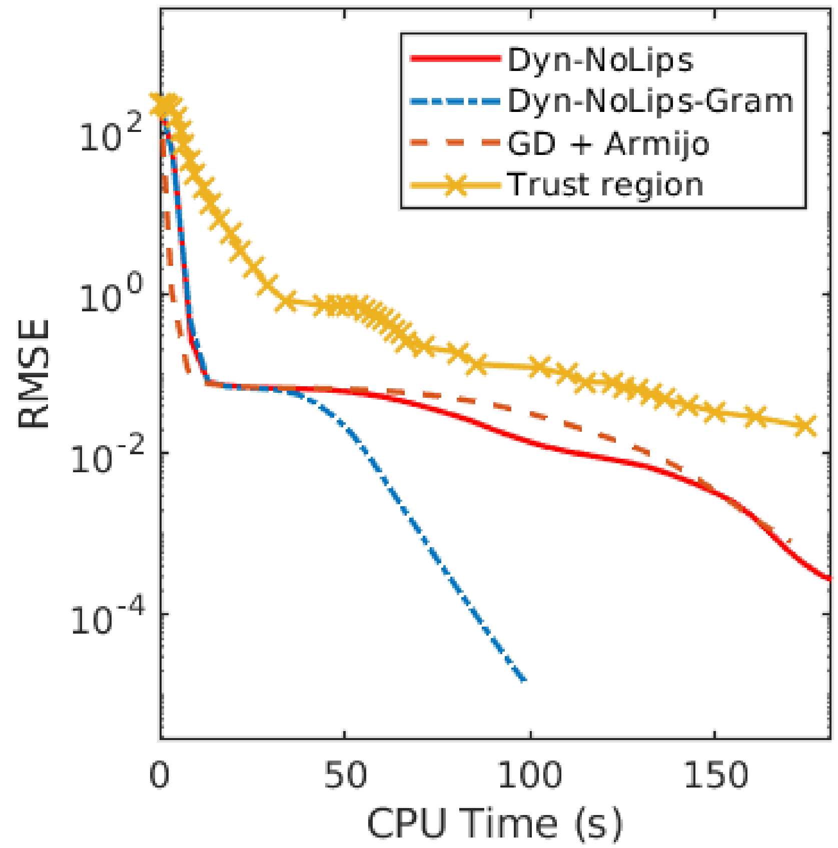
Experiments on synthetic `Helix` dataset with 10% known distances, dimension $r = 3$.



Experiments: Distance Matrix Completion



(a) $n = 2000$



(b) $n = 5000$

Figure 2: Experiments on `Helix` dataset

Outline

- Bregman gradient methods and relative smoothness
- Application to low-rank minimization
- **Theoretical complexity: lower bound and computer-aided analyses**
- Stochastic variants

The question of acceleration

We recall the convergence rate of BGD for relatively smooth convex functions

$$f(x_N) - f(x_*) \leq \frac{LD_h(x_*, x_0)}{N}.$$

Is there an algorithm that does better ?

Algorithm	**Supplementary assumptions**	**Convergence rate**
Accelerated gradient descent (Nesterov, 1983)	$h(x) = \frac{1}{2}\|x\|^2$	$O(1/N^2)$
Accelerated BGD (Auslender and Teboulle, 2006)	h is μ-strongly convex and f is L-smooth	$O(1/N^2)$
Accerated BGD (Hendrikx et al., 2020; Hanzely et al., 2021)	h satisfies *triangle scaling inequality*	Improved asymptotically

These assumptions are quite restrictive... What about the general case?

A lower bound for relatively-smooth convex minimization

In the general case, the $\mathcal{O}(1/N)$ rate of BGD is **optimal.**

Theorem (D., Taylor, d'Aspremont, Bolte, 2021)

For every $N \geq 1$, there exists functions $f_N, h_N : \mathbb{R}^{2N+1} \to \mathbb{R}$ and $x_0 \in \mathbb{R}^{2N+1}$ such that

- f_N is L-smooth relative to h_N,
- for **any Bregman first-order method** $\mathcal{A}$ initialized at x_0, after N iterations we have

$$f_N(x_N) - f_N(x_*) \geq \frac{LD_{h_N}(x_*, x_0)}{4N+1}.$$

- **Bregman first-order method:** uses $\nabla f, \nabla h, \nabla h^*$ and linear operations.
- Additional assumptions are needed to achieve acceleration.
- Worst-case functions f_N, h_N are "nearly" nondifferentiable.

Computer-aided analyses

Performance estimation: computing the **worst-case** behavior of a first-order through optimization (Drori and Teboulle, 2014; Taylor et al., 2017).

Recall the convergence rate of BGD for f **convex** and L-**smooth relative to** h:

$$f(x_N) - f(x_*) \leq \frac{LD_h(x_*, x_0)}{N}$$

Is this the best possible bound for **generic** f and h ? What are the corresponding worst-case functions ?

Performance Estimation Problem

maximize

subject to h is a kernel (differentiable and strictly convex),

f is convex and L-smooth relative to h,

$x_1, \ldots, x_N$ are generated from x_0 by BGD with step size $1/L$,

in the variables $x_0, \ldots, x_N, x_*, f, h$.

How to solve the PEP?

- Reduction to a finite-dimensional problem by replacing f, h with their discrete representations at $x_0, \ldots x_N$ (Drori and Teboulle, 2014):

$$(f_i, g_i) = \big(f(x_i), \nabla f(x_i)\big),$$
$$(h_i, s_i) = \big(h(x_i), \nabla h(x_i)\big).$$

- Equivalence with original problem is guaranteed by **interpolation conditions** (Taylor et al., 2017), which we extend to the relatively smooth setting.

$$x_i \neq x_j \implies h_i - h_j - \langle s_j, x_i - x_j \rangle > 0, \qquad \text{(strict convexity of h)}$$
$$s_i \neq s_j \implies x_i \neq x_j, \qquad \text{(differentiability of h)}$$
$$\vdots$$

- The PEP is then equivalent to a finite-dimensional problem in $\{(x_i, f_i, g_i, h_i, s_i)\}$, with quadratic constraints: can be solved via **semidefinite programming**.

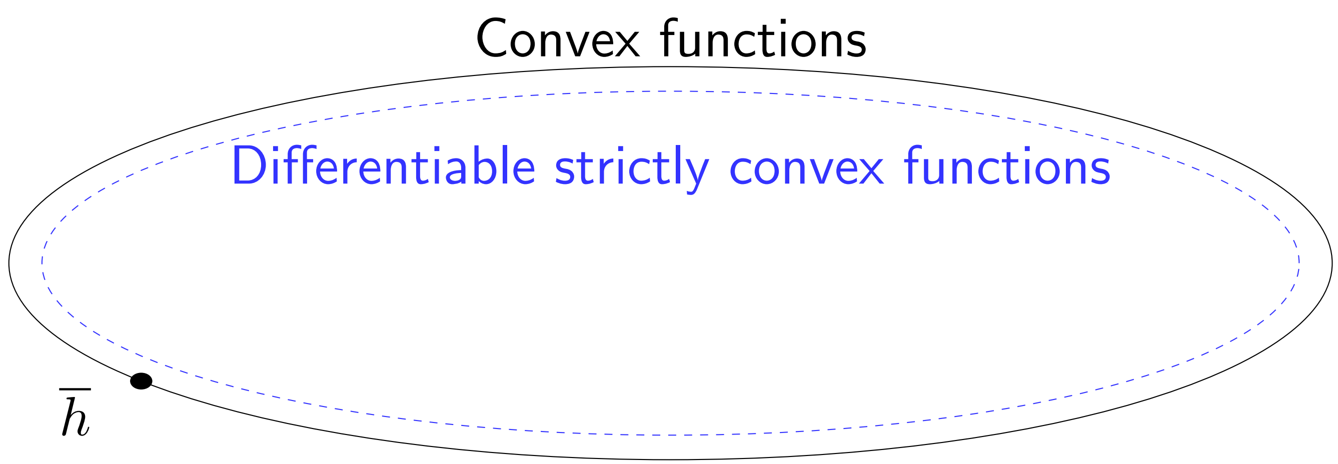
Results and insights

- The numerical value of the PEP is **exactly** L/N: the bound

$$f(x_N) - f(x_*) \leq \frac{L D_h(x_*, x_0)}{N}$$

  is tight in the worst case for BGD.

- **Limiting nonsmooth behavior:** the feasible set is not closed; the supremum is reached as (f, h) approach some nonsmooth limiting functions $(\overline{f}, \overline{h})$.



- With some modifications, discovered worst-case functions which are hard for **any Bregman method** $\rightarrow$ general lower bound

The case of entropy

Joint work with D. Ostrovskii

The case of **generic** h is too hard: let us now focus on a particular kernel, the entropy

$$h_e(x) = \sum_{i=1}^{d} x^i \log x^i - x^i$$

Performance Estimation Problem - entropic case

maximize	$\big(f(x_N) - f(x_*)\big)/D_{h_e}(x_*, x_0)$
subject to	f is convex and L-smooth relative to h_e (*entropic-smooth*),
	$x_1, \ldots, x_N$ are generated from x_0 by BGD with step size $1/L$,

in the variables $x_0, \ldots, x_N, x_*, f$.

Not solvable yet (convex program on *cone of pairwise Kullback-Leibler matrices*)

Outline

- Bregman gradient methods and relative smoothness
- Application to low-rank minimization
- Theoretical complexity: lower bound and computer-aided analyses
- **Stochastic variants**

Bregman stochastic gradient descent

Joint work with Hadrien Hendrikx and Mathieu Even

$$\min_{x \in \mathcal{C}} f(x) := \mathbb{E}_{\xi}\left[f_{\xi}(x)\right] \qquad \text{(P)}$$

where functions f_ξ are L-smooth and μ-strongly convex relative to h.

Bregman SGD

$$x_{k+1} = \operatorname*{argmin}_{u \in \mathcal{C}} \langle g_k, u - x_k \rangle + \frac{1}{\lambda} D_h(u, x_k),$$

$$g_k = \nabla f_{\xi_k}(x_k) \text{ for } \xi_k \text{ such that } \mathbb{E}\left[g_k\right] = \nabla f(x_k).$$

Convergence rate: with $\lambda = 1/(2L)$,

$$\mathbb{E}\left[D_h(x^\star, x_k)\right] \leq \underbrace{\left(1 - \frac{\mu}{2L}\right)^k D_h(x^\star, x_0)}_{\text{linear convergence}} + \underbrace{\lambda \frac{\sigma^2}{\mu}}_{\text{noise}}.$$

Noise assumption: σ^2 is the variance of $\nabla f_\xi(x^*)$ "with respect to Bregman divergence".

Variance reduction

We now assume that the problem is a finite sum:

$$\min_{x \in \mathcal{C}} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x),$$

where f_i are L-smooth and μ-strongly convex relative to h.

Variance reduction methods leverage the finite sum assumption to obtain *fast* convergence rates (Schmidt et al., 2013; Johnson and Zhang, 2013; Defazio et al., 2014).

Bregman-SAGA

$$x_{k+1} = \operatorname*{argmin}_{u \in \mathcal{C}} \langle \tilde{g}_k, u - x_k \rangle + \frac{1}{\lambda} D_h(u, x_k)$$

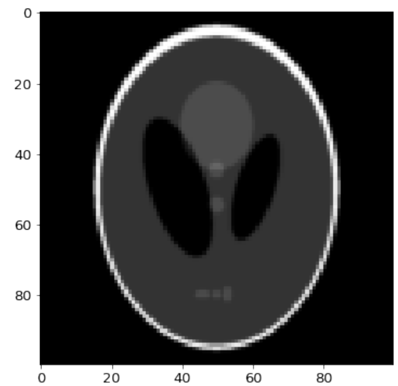
$$\tilde{g}_k = \nabla f_{i_k}(x_k) - \underbrace{\sum_{i=1}^{n} \beta_i \nabla f_i(\phi_i)}_{\text{contains previously computed gradients}}$$

Same situation as for acceleration: **asymptotical** convergence result under **additional regularity of** h.

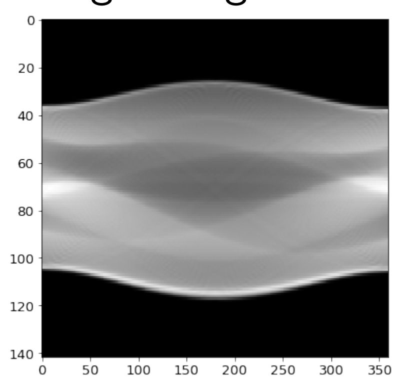
Experiments: tomographic reconstruction problem

Inverse problem with Poisson noise

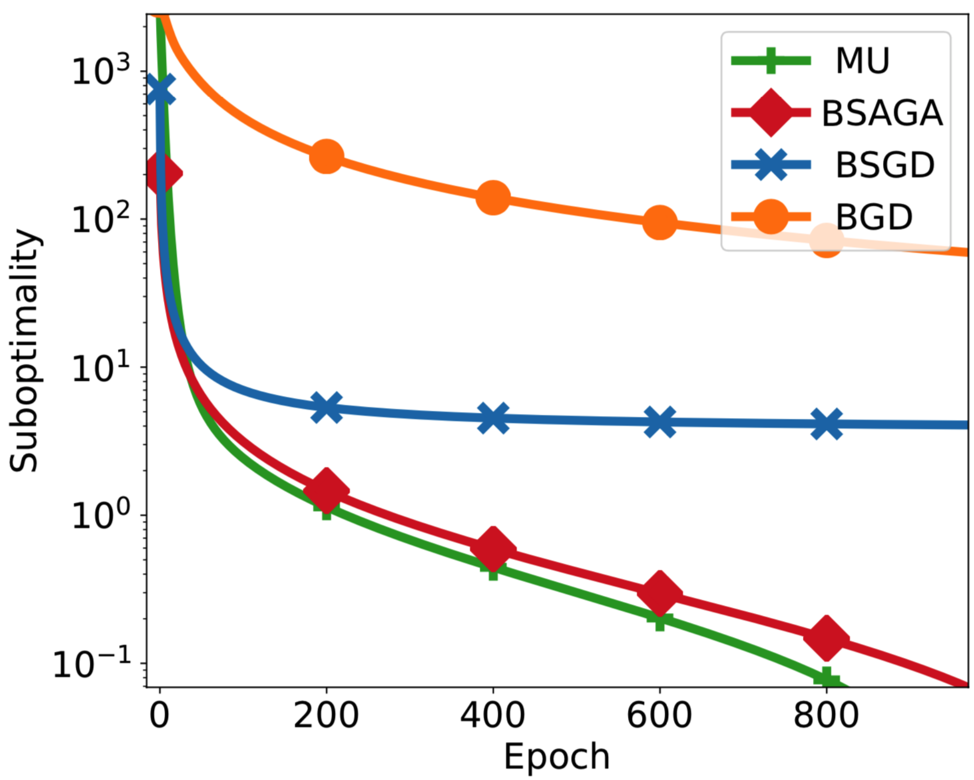
$$f(x) = D_{KL}(b, Ax), \quad h(x) = \sum_{i=1}^{d} -\log x^i.$$



Original signal x^*

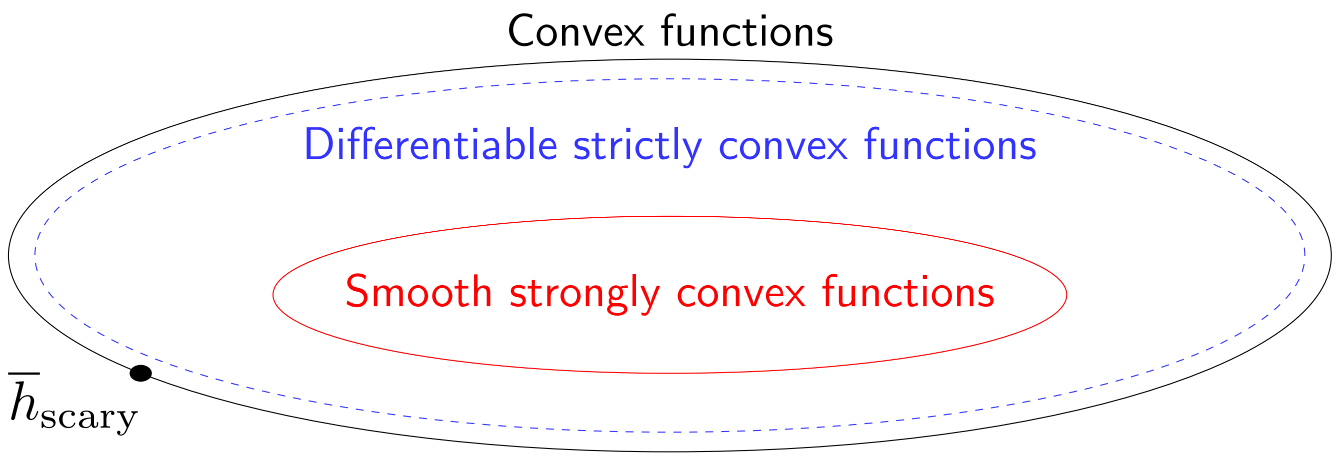


Sinogram Ax^*



Perspectives

- **Relatively-smooth optimization:** emerging subject, with many applications left to be explored;



- **Algorithmic extensions** (acceleration, variance reduction...): find the right regularity properties;

- **Adaptivity** to improve practical performance.

Thank you!

*

References

Alfred Auslender and Marc Teboulle. Interior Gradient and Proximal Methods for Convex and Conic Optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First Order Methods Beyond Convexity and Lipschitz Gradient Continuity with Applications to Quadratic Inverse Problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. pages 1–15, 2014.

Yoel Drori and Marc Teboulle. Performance of First-Order Methods for Smooth Convex Minimization: A Novel Approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

Filip Hanzely, Peter Richtarik, and Lin Xiao. Accelerated Bregman Proximal Gradient Methods for Relatively Smooth Convex Optimization. *Computational Optimization and Applications*, 2021.

Hadrien Hendrikx, Lin Xiao, Sébastien Bubeck, Francis Bach, and Laurent Massoulié. Statistically preconditioned accelerated gradient method for distributed optimization. In *International Conference on Machine Learning*, number 119, pages 4203—-4227, 2020.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 2013.

Yurii Nesterov. A Method of Solving A Convex Programming Problem With Convergence rate O(1/k^2). *Soviet Mathematics Doklady*, 27 (2):372–376, 1983.

Mark Schmidt, Nicolas Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162, 09 2013.

Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Smooth Strongly Convex Interpolation and Exact Worst-Case Performance of First-Order Methods. *Mathematical Programming*, 161(1-2):307–345, 2017.

Supplementary material

How to solve the PEP?

Performance Estimation Problem

$$
\begin{aligned}
\text{maximize} \quad & (f_N - f_*)/\big(h_* - h_0 - \langle s_0, x_* - x_0\rangle\big) \\
\text{subject to} \quad & h \text{ is a kernel (differentiable and strictly convex)}, \\
& f \text{ is convex and } L\text{-smooth relative to } h, \\
& f(x_i) = f_i,\ h(x_i) = h_i,\ \nabla f(x_i) = g_i,\ \nabla h(x_i) = s_i \quad \forall i \in I, \\
& x_1, \ldots, x_N \text{ are generated from } x_0 \text{ by BGD with step size } 1/L,
\end{aligned}
$$

in the variables $\{x_i, f_i, h_i, g_i, s_i\}_{i\in I}, f, h$.

- Reduction to a finite-dimensional problem (Drori and Teboulle, 2014);

How to solve the PEP?

Performance Estimation Problem

maximize $\quad (f_N - f_*)/\big(h_* - h_0 - \langle s_0, x_* - x_0\rangle\big)$

subject to $\quad$ there exist f, h such that h is a kernel,

f is convex and L-smooth relative to h,

$f(x_i) = f_i,\ h(x_i) = h_i,\ \nabla f(x_i) = g_i,\ \nabla h(x_i) = s_i \quad \forall i \in I,$

$x_1, \ldots, x_N$ are generated from x_0 by BGD with step size $1/L$,

in the variables $\{x_i, f_i, h_i, g_i, s_i\}_{i\in I}$.

- Reduction to a finite-dimensional problem (Drori and Teboulle, 2014);
- Equivalence with original problem is guaranteed by interpolation conditions;

How to solve the entropic PEP ?

- Reduction to a finite-dimensional problem by replacing f with its discrete representation

$$\{(f_i, g_i)\}_{1\le i\le N} = \left\{\left(f(x_i), \nabla f(x_i)\right)\right\}_{1\le i\le N}.$$

- Equivalence with original problem is guaranteed by **interpolation conditions**, which we extend to the entropic-smooth setting:

$$f_i - f_j - \langle g_j, x_i - x_j\rangle \ge L D_{\mathrm{KL}}\left[x_i, x_i \circ \exp\left(\frac{g_j - g_i}{L}\right)\right] \quad \forall i, j.$$

- The PEP is then equivalent to a finite-dimensional problem on a convex cone, the **Kullback-Leibler cone** with **log-linear constraints**:

$$\mathcal{K}_m(A) = \left\{ \left[D_{\mathrm{KL}}(x_i, x_j)\right]_{1\le i,j\le m} \;\middle|\; \begin{array}{l} d \in \mathbb{N} \text{ and } x_1, \dots x_m \in \mathbb{R}^d \\ \text{such that } \sum_{j=1}^m A_{ij} \log(x_j) = 0, \\ \qquad i = 1 \dots q \end{array} \right\}.$$

... no known solver yet

Bregman SGD - theoretical guarantees

Assume

- **Sampling:** $g_k = \nabla f_{\xi_k}(x_k)$ for some ξ_k and $\mathbb{E}_{\xi_k}[g_k] = \nabla f(x_k)$,
- **Variance:**
$$\mathbb{E}_{\xi_k}\left[P_{x_k}\left(\nabla f_{\xi_k}(x^*)\right)\right] \leq \sigma^2$$
where $P_x(v)$ is the Bregman counterpart of $\|v\|^2$:
$$P_x(v) = \frac{1}{4\lambda^2} D_{h^*}\left[\nabla h(x) - 2\lambda v, \nabla h(x)\right]$$
- **Regularity:** functions f_ξ are L-smooth and μ-strongly convex relative to h.

Theorem (D., Hendrikx, Even, 2021)

The iterates of Bregman SGD with step size $\lambda = 1/(2L)$ satisfy

$$\mathbb{E}\left[D_h(x^\star, x_k)\right] \leq \underbrace{(1 - \frac{\mu}{2L})^k D_h(x^\star, x_0)}_{\text{linear convergence}} + \underbrace{\lambda\frac{\sigma^2}{\mu}}_{\text{noise}}.$$

Bregman-SAGA, theoretical guarantees

Assumption: gain function

There exists a gain function G such that for any $x, y, v \in \mathbb{R}^d$ and $\lambda \in [-1, 1]$,

$$D_{h^*}(x + \lambda v, x) \leq G(x, y, v)\lambda^2 D_{h^*}(y + v, y)\,.$$

G determines the step size and convergence rate.

- h **is quadratic:** then $G = 1$, Bregman-SAGA rate is

$$\mathcal{O}\left(1 - \min\left(\frac{\mu}{8L}, \frac{1}{2n}\right)\right)^k.$$

- h^* **has Lipschitz Hessian** (and extra local smoothness): with the right choice of step size, Bregman-SAGA rate is

$$\mathcal{O}\left(1 - \min\left(\frac{\mu}{8G_k L}, \frac{1}{2n}\right)\right)^k \quad \text{with } G_k \to 1 \text{ as } k \to \infty.$$

Asymptotical rate under additional regularity: same situation as for accelerated BGD (Hendrikx et al., 2020; Hanzely et al., 2021)