Optimal Algorithms for Stochastic Complementary Composite Minimization

Clément Lezane [1] Alexandre d'Aspremont [2] Cristóbal Guzmán [3]

[1]University of Twente

[2]CNRS, INRIA

[3]Catholic University of Chile

September 21, 2023

Outline

Bregman Divergence

For $F : \mathcal{X} \to \mathbb{R}$ a continuously-differentiable convex function (a potential function), we can define :

Definition (Bregman Divergence)

$$\forall x, \epsilon \in \mathcal{X}, D^F(x+\epsilon, x) := F(x+\epsilon) - F(x) - \langle \nabla F(x), \epsilon \rangle$$

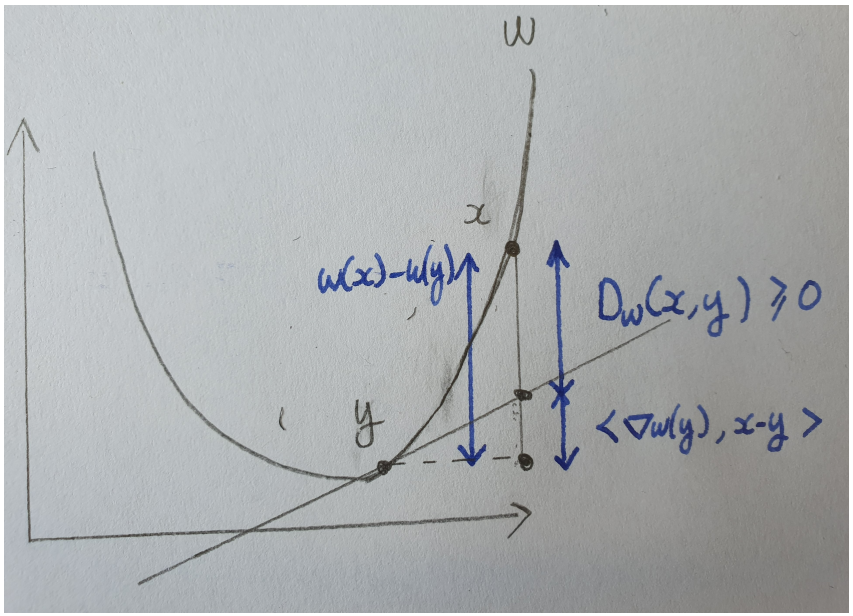


Figure: Bregman divergence

Smoothness and strong convexity

Definition (Bregman Divergence)

$$\forall x, \epsilon \in \mathcal{X}, D^F(x+\epsilon, x) := F(x+\epsilon) - F(x) - \langle \nabla F(x), \epsilon \rangle$$

Smoothness and strong convexity

Definition (Bregman Divergence)

$$\forall x, \epsilon \in \mathcal{X}, D^F(x+\epsilon, x) := F(x+\epsilon) - F(x) - \langle \nabla F(x), \epsilon \rangle$$

Definition (Smoothness,SC, condition number L/μ)

$$\forall x, \epsilon \in \mathcal{X}, D^F(x+\epsilon, x) \leq \frac{L}{2}\|\epsilon\|^2$$
$$\forall x, \epsilon \in \mathcal{X}, D^F(x+\epsilon, x) \geq \frac{\mu}{2}\|\epsilon\|^2$$

Smoothness and strong convexity

Definition (Bregman Divergence)

$$\forall x, \epsilon \in \mathcal{X}, D^F(x+\epsilon, x) := F(x+\epsilon) - F(x) - \langle \nabla F(x), \epsilon \rangle$$

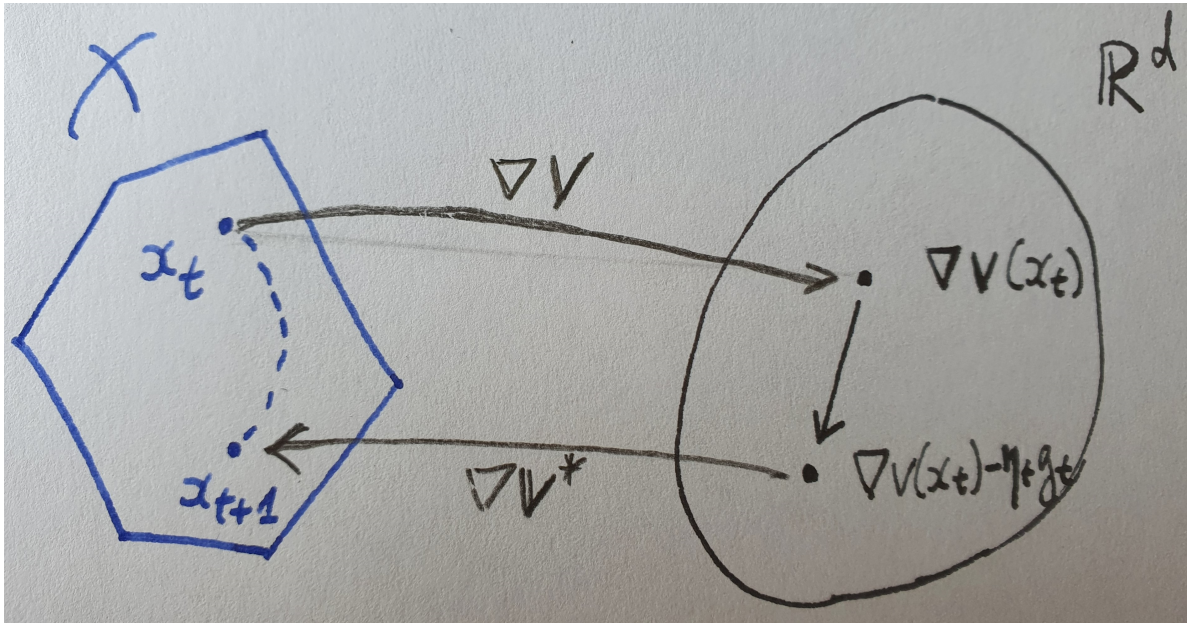
Definition (Smoothness,SC, condition number L/μ)

$$\forall x, \epsilon \in \mathcal{X}, D^F(x+\epsilon, x) \leq \frac{L}{2}\|\epsilon\|^2$$
$$\forall x, \epsilon \in \mathcal{X}, D^F(x+\epsilon, x) \geq \frac{\mu}{2}\|\epsilon\|^2$$

Definition (Mirror descent)

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \langle \nabla F(x_t), x \rangle + \eta_t D^{\Omega}(x, x_t) \right\}$$

Bregman divergence and mirror descent



Any question ?

Composite setting

$$\min_{x \in \mathcal{X}} \Psi(x) = F(x) + H(x)$$

- the loss function F is L-smooth and μ-strongly-convex
- the regularizer H is convex

An example would be Ridge regression :

$$\min_{x \in \mathbb{R}^d} \mathbb{E}_{(a,b)}[\|a^\top x - b\|^2] + \mu \|x\|_2^2. \tag{1}$$

Stochastic Mirror Descent (SMD)

We consider $G(x_t, \xi_t)$ to be an unbiased estimator of ∇F:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \alpha_t \Big(\langle G(x_t, \xi_t), x \rangle + H(x) \Big) + \gamma_t V(x, x_t) \right\}$$

For the stochastic error $\Delta_t(x_t) := G(x_t, \xi_t) - \nabla F(x_t)$, we assume that for all $t \geq 1$:

$$\begin{cases} \mathbb{E}\left[\Delta_t(x_t)\right] & = 0 \\ \mathbb{E}\left[\|\Delta_t(x_t)\|_*^2\right] & \leq \sigma^2, \end{cases} \tag{2}$$

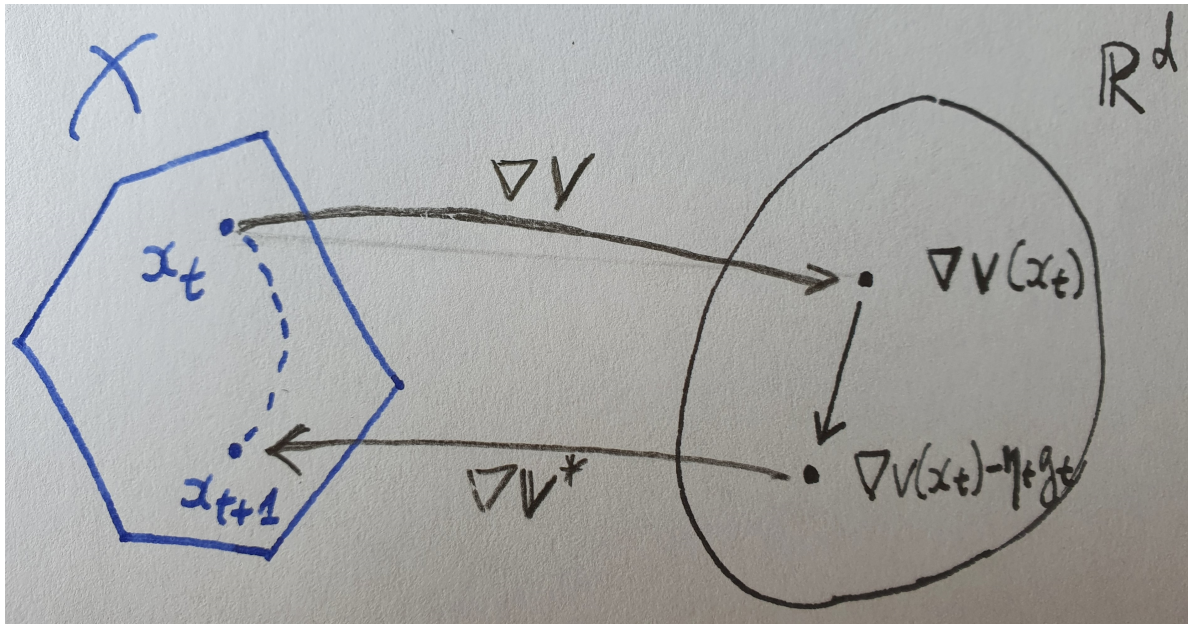
Remark Dual norm is defined as : $\|y\|_* = \sup_{\|x\| \leq 1} |\langle x, y \rangle|$

"Mirror" Interpretation

If $H = 0$, we notice that :

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \langle g_t, x \rangle + \frac{\gamma_t}{\alpha_t} V(x, x_t) \right\}$$

$$\implies \nabla V(x_{t+1}) = \nabla V(x_t) - \frac{\alpha_t}{\gamma_t} g_t.$$



Linear SA convergence

If we choose $V = D^F$ and the step sizes respecting the following criteria

$$\begin{cases} \gamma_t & \geq \frac{2L}{\mu}\alpha_t := 2\kappa\alpha_t \\ \alpha_t & \geq \gamma_{t+1} - \gamma_t, \end{cases} \tag{3}$$

then as we note $A_T = \sum_t \alpha_t$ and $x_{T+1}^{ag} = \frac{\sum \alpha_t x_t}{\sum_t \alpha_t}$:

$$\begin{aligned} & A_T[\Psi(x_{T+1}^{ag}) - \Psi(x_\star)] + \gamma_T V(x_\star, x_{T+1}) \\ \leq \quad & \gamma_1 V(x_\star, x_1) + \sum_{t=1}^T \alpha_t \langle \Delta_t(x_t) | x_\star - x_t \rangle + \sum_{t=1}^T \frac{2\|\alpha_t \Delta_t(x_t)\|_\star^2}{\mu\gamma_t} \end{aligned} \tag{4}$$

Linear SA convergence : Application

If we choose :

$$\begin{cases} \alpha_t & := t + 1 + 12\kappa \\ \gamma_t & := \frac{(t+12\kappa)^2}{2} \end{cases} \tag{5}$$

then for all $T \geq 1$:

$$\begin{aligned} & \mathbb{E}\left[\Psi(x_{T+1}^{ag}) - \Psi(x_\star) + V(x_\star, x_{T+1})\right] \\ \leq & \quad \frac{K_1 \kappa^2 V(x_\star, x_1)}{T^2} + \frac{K_2 \sigma^2}{\mu T} \end{aligned} \tag{6}$$

with $K_1 = 108$ and $K_2 = 20$.

Restarting algorithm 1/2

Algorithm Algorithm Multistage SA

Require: $n \geq 0$, $T \geq 0$, $x_1 \in \mathcal{X}$, K, algorithm $\mathcal{A}$

Consider initial start point : $x_1^0 = x_1$

for $1 \leq k \leq n$ **do**

Run Algorithm $\mathcal{A}$ with K iterations, start point $x_1^k \leftarrow x_{K+1}^{k-1}$.

Output x_{K+1}^k

end for

Run algorithm $\mathcal{A}$ with T iterations, start point $x_1^{n+1} = x_{K+1}^n$, compute the weighted output x_{T+1}^{ag}.

Restarting algorithm 2/2

Consider (x_{T+1}, x_{T+1}^{ag}) the output of an algorithm, such that there exist K_1, K_2, K_3 and $\alpha_1 \geq \alpha_2, \alpha_3$ such that for all T:

$$\mathbb{E}\left[\Psi(y_{T+1}) - \Psi^\star + D^\omega(x_\star, x_{T+1})\right] \leq \frac{K_1 D^\omega(x_\star, x_1)}{T^{\alpha_1}} + \frac{K_2}{T^{\alpha_2}} + \frac{K_3}{T^{\alpha_3}} \tag{7}$$

Then for the output Y_{T+1} of the restarting algorithm, we have:

$$\mathbb{E}\left[\Psi(Y_{T+1}) - \Psi^\star\right] \leq \frac{D^\omega(x_\star, x_1)}{2^{n+1}} + \frac{2K_2}{T^{\alpha_2}} + \frac{2K_3}{T^{\alpha_3}}$$

Accelerated methods 1/2

$$\begin{cases} x_t^{md} & = \frac{A_{t-1}}{A_t} x_t^{ag} + \frac{\alpha_t}{A_t} x_t \\ x_{t+1} & = \arg\min_x \{\alpha_t [\langle G(x_t^{md}, \xi_t), x\rangle] + \gamma_t V(x, x_t)\} \\ x_{t+1}^{ag} & = \frac{A_{t-1}}{A_t} x_t^{ag} + \frac{\alpha_t}{A_t} x_{t+1} \end{cases}$$

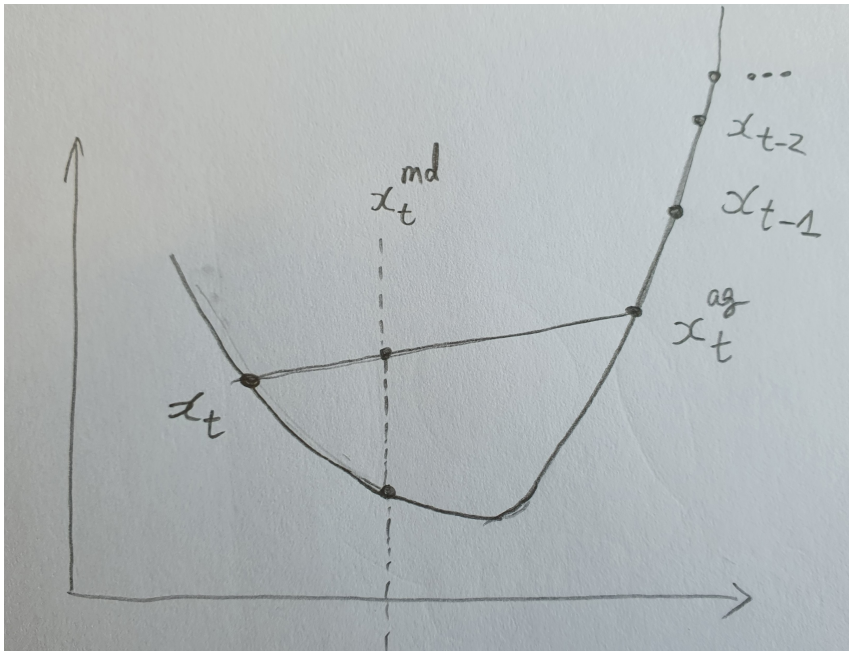


Figure: Accelerated methods

Accelerated methods 2/2

If we choose :

$$\begin{cases} \alpha_t & := t+1 \\ \gamma_t & := \frac{t^2}{2} + 8\kappa \end{cases} \tag{8}$$

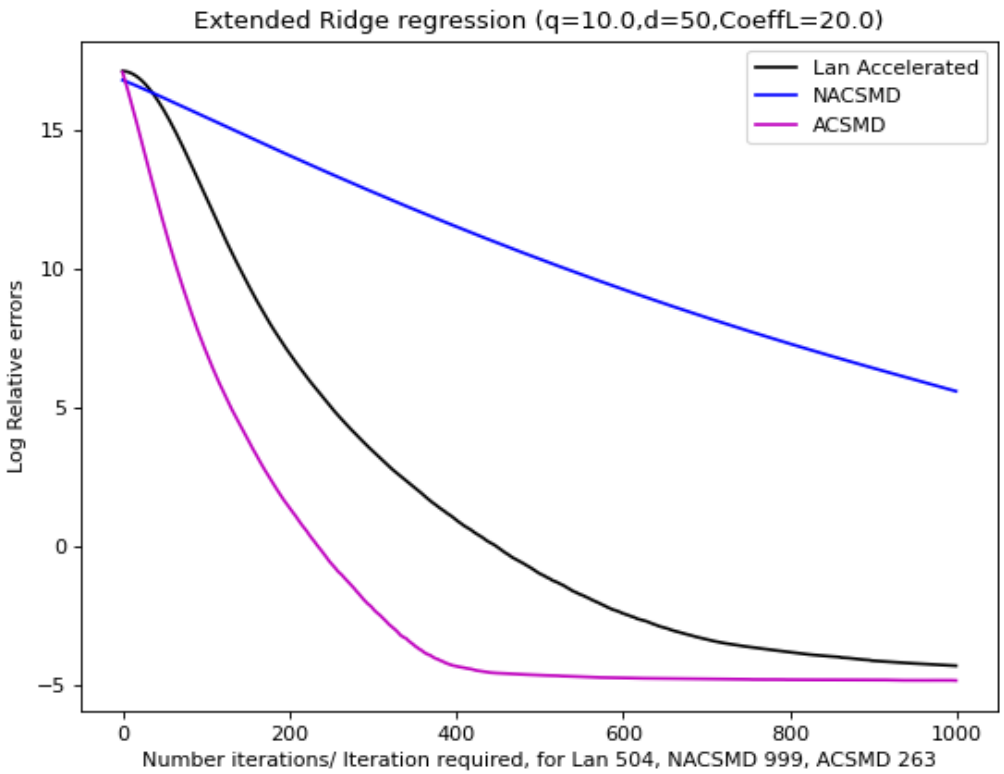
then for all $T \geq 1$:

$$\begin{aligned} & \mathbb{E}\left[\Psi(x_{T+1}^{ag}) - \Psi(x_\star) + V(x_\star, x_{T+1})\right] \\ \leq & \quad \frac{K_1' \kappa V(x_\star, x_1)}{T^2} + \frac{K_2' \sigma^2}{\mu T} \end{aligned} \tag{9}$$

with restarting algorithm, the complexity becomes:

$$O(1)\left(\sqrt{\kappa}\log\left(\frac{V(x_\star, x_1)}{\epsilon}\right) + \frac{\sigma^2}{\mu\epsilon}\right)$$

State of the art



Any question ?

Problem : condition number in high dimension

Lemma (NY83, AGJ18)

Let $\mathcal{X} \subset \mathbb{R}^d$ with an interior not empty, if $F : \mathcal{X} \to \mathbb{R}$ is L-smooth, μ-strongly convex w.r.t the norm $\|\cdot\|_q$, then :

$$\kappa := \frac{L}{\mu} \geq \frac{d^{1-\frac{2}{q}}}{2}$$

Recall For $x \in \mathbb{R}^d$, $\|x\|_q := \left(\sum_{i=1}^d |x_i|^q\right)^{\frac{1}{q}}$

Proof 1/2

We consider $e_1 = \frac{\nabla F(x_0)}{\|\nabla F(x_0)\|}$ and we complete it to make an orthogonal base $(e_1, \dots, e_d)$, with some weights $\delta_k \in \{-1, 1\}$, the idea is to build $x_k = x_0 + \frac{1}{d^{1/q}} \sum_{i=1}^{k} \delta_i e_i$ progressively.
We choose δ_{k+1} such that :

$$\delta_{k+1} = sign(\langle \nabla F(x_k), e_{k+1} \rangle)$$

Proof 1/2

We consider $e_1 = \frac{\nabla F(x_0)}{\|\nabla F(x_0)\|}$ and we complete it to make an orthogonal base $(e_1, \dots, e_d)$, with some weights $\delta_k \in \{-1, 1\}$, the idea is to build $x_k = x_0 + \frac{1}{d^{1/q}} \sum_{i=1}^{k} \delta_i e_i$ progressively.
We choose δ_{k+1} such that :

$$\delta_{k+1} = sign(\langle \nabla F(x_k), e_{k+1} \rangle)$$

The strong convexity makes :

$$D^F(x_{k+1}, x_k) \geq \frac{\mu}{2} \|x_{k+1} - x_k\|_q^2$$

$$F(x_{k+1}) - F(x_k) - \langle \nabla F(x_k), \delta_{k+1} e_{k+1} \rangle \geq \frac{\mu}{2d^{2/q}}$$

Proof 2/2

As we obtain for all k,

$$F(x_{k+1}) - F(x_k) - \langle \nabla F(x_k), \delta_{k+1} e_{k+1} \rangle \geq \frac{\mu}{2d^{2/q}}$$

We sum up the previous equation d times,

$$F(x_{d+1}) - F(x_0) - \langle \nabla F(x_0), \delta_1 e_1 \rangle \geq \frac{\mu d}{2d^{2/q}}$$

Proof 2/2

As we obtain for all k,

$$F(x_{k+1}) - F(x_k) - \langle \nabla F(x_k), \delta_{k+1} e_{k+1} \rangle \geq \frac{\mu}{2d^{2/q}}$$

We sum up the previous equation d times,

$$F(x_{d+1}) - F(x_0) - \langle \nabla F(x_0), \delta_1 e_1 \rangle \geq \frac{\mu d}{2d^{2/q}}$$

As $D_F(x_{d+1}, x_0) = F(x_{d+1}) - F(x_0) - \langle \nabla F(x_0), \delta_1 e_1 \rangle$,

$$\frac{L}{2}\|x_{d+1} - x_0\|_q^2 \geq D_F(x_{d+1}, x_0) \geq \frac{\mu d^{1-\frac{2}{q}}}{2}$$

□

Complementary composite setting

$$\min_{x \in \mathcal{X}} \Psi(x) = F(x) + H(x)$$

with F being (L, κ)-weakly-smooth and H being (μ, q)-uniformly convex w.r.t. the same norm $\|\cdot\|$.

Definition ((L, κ) weak-smoothness)

$$\forall x, y \in \mathcal{X}, D^F(x, y) \leq \frac{L}{\kappa}\|x - y\|^\kappa$$

Definition ((μ, q) uniform convexity)

$$\forall x, y \in \mathcal{X}, D^H(x, y) \geq \frac{\mu}{q}\|x - y\|^q$$

Assumptions

We consider $G(x_t, \xi_t)$ to be an unbiased estimator of ∇F and $\frac{1}{p} + \frac{1}{q} = 1$, we assume that for all $t \geq 1$:

$$\begin{cases} \Delta_t(x_t) := G(x_t, \xi_t) - \nabla F(x_t) \\ \mathbb{E}\left[\Delta_t(x_t)\right] = 0 \\ \mathbb{E}\left[\|\Delta_t(x_t)\|_*^p\right] \leq \sigma^p, \end{cases} \tag{10}$$

and

$$\begin{cases} \alpha_t & \geq \gamma_{t+1} - \gamma_t \\ \gamma_t & \geq \frac{2L}{\mu} \frac{\alpha_t^q}{A_t^{q-1}}. \end{cases}$$

Convergence rate

Theorem

Under good assumptions, we have for all $T \geq 1$*:*

$$\mathbb{E}\left[\Psi(x_{T+1}^{ag}) - \Psi^\star + D^\omega(x_\star, x_{T+1})\right]$$
$$\leq O_{q,\kappa}\left(\frac{L^{\frac{n+1}{q}} V_0}{(\mu^{1/q} T)^{n+1}} + \left(\frac{L}{\mu}\right)^{\frac{1}{r}} \frac{L}{T^{\frac{q-r}{r}}} + \frac{\sigma^p}{(\mu T)^{p/q}}\right)$$

where $V_0 = D^\omega(x_\star, x_1)$ *and* $O_{q,\kappa}$ *omits absolute constants that depend on* q, κ*.*

Extended Regression problem

$$\min_{x \in \mathbb{R}^d} \mathbb{E}_{(a,b)}[(a^\top x - b)^2] + \mu \|x\|_q^q. \tag{11}$$

We generate synthetic data from a uniform distribution $\mathcal{U}$ and Gaussian noise:

$$\begin{cases} a \sim \mathcal{U}([-1,1]^d) \\ b = a^\top x_\star + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma_b). \end{cases}$$

then :

$$\begin{cases} D^F(x_\star, x) & = \mathbb{E}_{(a,b)}\left[(a^\top (x - x_\star))^2\right] = \frac{1}{3}\|x - x_\star\|_2^2. \\ \frac{1}{3}\|x - x_\star\|_q^2 & \leq D^F(x_\star, x) \leq \frac{d^{1-\frac{2}{q}}}{3}\|x - x_\star\|_q^2. \end{cases}$$

Simulation result

We run the simulation 50 times and take the average of the expected error. We note the required iterations to first hit the precision $\epsilon = 0.01$:

Iteration required	Lan	NACSMD	ACSMD
d=20	91	81	32
d=50	110	66	26
d=100	145	82	33
d=200	138	76	26

Table: Simulation results with different dimensions

Take home messages

- Role of condition number for smooth problems
- Condition number in non-Euclidean space, different tricks to reduce the related complexity
- Role of moment order in Stochastic optimization