Aggregative Nonconvex Optimization: Two Dual-Based Methods

Thibault Moquet

PhD supervisors: Guilherme Mazanti, Laurent Pfeiffer

L2S, CentraleSupélec, Université Paris-Saclay
Fédération de Mathématiques

February 2023

Summary

Summary

Objective

The objective is to solve the following problem:

$$\underset{x \in \mathcal{X}}{\text{minimize}} \; f\left(\frac{1}{N}\sum_{i=1}^{N} g_i(x_i)\right).$$

This is a multi-agent problem, where the choices $(x_1, \ldots, x_N)$ are only seen through the **aggregate term** $\frac{1}{N}\sum_{i=1}^{N} g_i(x_i)$.

References

- Vanishing Price of Decentralization in Large Coordnative nonconvex Optimization, M. Wang in SIOPT 2017.
- Large-scale nonconvex optimization: randomization, gap estimate and numerical resolution, J.F. Bonnans, K. Liu, N. Oudjane, L. Pfeiffer and C. Wan, ArXiv preprint 2022.

Example

- N power plants.
- x_i: decision variable associated to agent $i \in \{1, \ldots, N\}$
  $\rightarrow$ production $g_i(x_i)$.
- f penalizes the difference with a demand.

Summary

Setting

- $\mathcal{E}$ Hilbert space, with its inner product $\langle \cdot\,, \cdot \rangle_{\mathcal{E}}$ and its deriving norm $\| \cdot \|_{\mathcal{E}}$. Sometimes refered to as **aggregate space**.
- Function $f : \mathcal{E} \to \mathbb{R}$.
- Positive integer N.

Setting

- $\mathcal{E}$ Hilbert space, with its inner product $\langle \cdot , \cdot \rangle_{\mathcal{E}}$ and its deriving norm $\| \cdot \|_{\mathcal{E}}$. Sometimes refered to as **aggregate space**.
- Function $f : \mathcal{E} \to \mathbb{R}$.
- Positive integer N.
- For $i \in \{1, \ldots, N\}$, set $\mathcal{X}_i$ and function $g_i : \mathcal{X}_i \to \mathcal{E}$.
- Product set $\mathcal{X} := \prod_{i=1}^{N} \mathcal{X}_i$, with its elements denoted $x = (x_1, \ldots, x_N)$.

Setting

- $\mathcal{E}$ Hilbert space, with its inner product $\langle \cdot , \cdot \rangle_{\mathcal{E}}$ and its deriving norm $\| \cdot \|_{\mathcal{E}}$. Sometimes refered to as **aggregate space**.
- Function $f \colon \mathcal{E} \to \mathbb{R}$.
- Positive integer N.
- For $i \in \{1, \ldots, N\}$, set $\mathcal{X}_i$ and function $g_i \colon \mathcal{X}_i \to \mathcal{E}$.
- Product set $\mathcal{X} := \prod_{i=1}^{N} \mathcal{X}_i$, with its elements denoted $x = (x_1, \ldots, x_N)$.

The problem

$$\underset{x \in \mathcal{X}}{\text{minimize}}\ f\left(\frac{1}{N}\sum_{i=1}^{N} g_i(x_i)\right) . \qquad (P)$$

Assumptions

Assumption

- *For all $i \in \{1, \ldots, N\}$, the set $g_i(\mathcal{X}_i)$ is compact.*
- *The function f is convex, and is β-strongly smooth (which can be shown to be equivalent to f having β-Lipschitz gradient).*

Assumptions

Assumption

- *For all $i \in \{1, \ldots, N\}$, the set $g_i(\mathcal{X}_i)$ is compact.*
- *The function f is convex, and is β-strongly smooth (which can be shown to be equivalent to f having β-Lipschitz gradient).*

Remark

No assumptions on the sets $\mathcal{X}_i$ and the functions g_i *per se*, in particular in terms of structure or regularity.

Lemma

Under these assumptions, Problem (P) has a solution.

Geometric formulation, relaxation

Geometric formulation:

$$\mathcal{Y}_i = g_i(\mathcal{X}_i) \quad \text{and} \quad \mathcal{Y} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{Y}_i$$

$$(P) \longleftrightarrow \underset{y\in\mathcal{Y}}{\text{minimize}}\ f(y) \longleftrightarrow \underset{y_i\in\mathcal{Y}_i}{\text{minimize}}\ f\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)$$

Geometric formulation, relaxation

Geometric formulation:

$$\mathcal{Y}_i = g_i\big(\mathcal{X}_i\big) \quad \text{and} \quad \mathcal{Y} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{Y}_i$$

$$(P) \longleftrightarrow \underset{y\in\mathcal{Y}}{\text{minimize}}\ f(y) \longleftrightarrow \underset{y_i\in\mathcal{Y}_i}{\text{minimize}}\ f\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)$$

Convex relaxation:

$$\underset{a^i\in\mathcal{P}(\mathcal{Y}_i)}{\text{minimize}}\ f\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[a^i]\right), \qquad (\tilde{P})$$

where, for all $a^i \in \mathcal{M}(\mathcal{Y}_i)$, $\mathbb{E}[a^i] := \int_{\mathcal{Y}_i} y_i da^i(y_i)$.

Advantages of the relaxation

Relaxation gap: quantity $\mathrm{Val}(P) - \mathrm{Val}(\tilde{P}) \geq 0$. May be positive.

In [Bonnans et al. '22]

- Estimate of the relaxation gap (tends to 0 when $N \to +\infty$).
- Method to **recover** an approximate solution of Problem (P) from an approximate solution of Problem $(\tilde{P})$ (efficient when N large).

We want N large
+ focus on the resolution of Problem $(\tilde{P})$, with dual-based methods.

Lagragian function

$$\tilde{\mathcal{Y}} = \prod_{i=1}^{N} \mathcal{Y}_i, \qquad \tilde{\mathcal{M}}(\tilde{\mathcal{Y}}) = \prod_{i=1}^{N} \mathcal{M}(\mathcal{Y}_i), \qquad \tilde{\mathcal{P}}(\tilde{\mathcal{Y}}) = \prod_{i=1}^{N} \mathcal{P}(\mathcal{Y}_i)$$

There is a bounded linear operator $K : \mathcal{E} \to \prod_{i=1}^{N} \mathcal{C}(\mathcal{Y}_i)$ such that

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[a^i] = -K^* a.$$

Lagragian function

$$\tilde{\mathcal{Y}} = \prod_{i=1}^{N} \mathcal{Y}_i, \qquad \tilde{\mathcal{M}}(\tilde{\mathcal{Y}}) = \prod_{i=1}^{N} \mathcal{M}(\mathcal{Y}_i), \qquad \tilde{\mathcal{P}}(\tilde{\mathcal{Y}}) = \prod_{i=1}^{N} \mathcal{P}(\mathcal{Y}_i)$$

There is a bounded linear operator $K \colon \mathcal{E} \to \prod_{i=1}^{N} \mathcal{C}(\mathcal{Y}_i)$ such that

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[a^i] = -K^* a.$$

Lagrangian function $\hat{L} \colon \tilde{\mathcal{M}}(\tilde{\mathcal{Y}}) \times \mathcal{E} \to \bar{\mathbb{R}}$.

$$\hat{L}(a, \mu) = \begin{cases} -f^*(\mu) + \langle \mu, -K^* a \rangle_{\mathcal{E}} + \iota_{\tilde{\mathcal{P}}(\tilde{\mathcal{Y}})}(a) & \text{if } \mu \in \operatorname{dom}(f^*). \\ -\infty & \text{otherwise.} \end{cases}$$

where $f^* \colon \mathcal{E} \to \bar{\mathbb{R}}$ is the Fenchel transform of f, defined as

$$f^*(\mu) = \sup_{\lambda \in \mathcal{E}} \langle \lambda, \mu \rangle_{\mathcal{E}} - f(\lambda).$$

Dual problem: Saddle-point

We have

$$(\tilde{P}) \longleftrightarrow \underset{a \in \tilde{\mathcal{M}}(\tilde{\mathcal{Y}})}{\text{minimize}} \quad \sup_{\mu \in \mathcal{E}} \quad \hat{L}(a, \mu).$$

Dual Problem:

$$\underset{\mu \in \mathcal{E}}{\text{maximize}} \quad \inf_{a \in \tilde{\mathcal{M}}(\tilde{\mathcal{Y}})} \hat{L}(a, \mu) \tag{D}$$

General Lemma

Lemma

We have

1. *Problems (P) and $(\tilde{P})$ have same dual (D).*
2. *There is no duality gap between Problems $(\tilde{P})$ and (D), i.e. Problems $(\tilde{P})$ and (D) have same value V. Moreover, Problem (D) has a solution μ^*.*
3. *Problem $(\tilde{P})$ has a solution.*

Other formulation of duality

Dual function $Q \colon \mathcal{E} \to \bar{\mathbb{R}}$

$$Q(\mu) = -f^*(\mu) + \frac{1}{N} \sum_{i=1}^{N} \left(\inf_{y_i \in \mathcal{Y}_i} \langle \mu, y_i \rangle_{\mathcal{E}} \right)$$

$$(D) \longleftrightarrow \underset{\mu \in \mathcal{E}}{\text{maximize}} \; Q(\mu)$$

Q is **strongly concave** (since f^* is **strongly convex**)
+ decomposed along the sets $\mathcal{Y}_i$.

Summary

Approach

In [Wang '17], the following general approach is investigated:

- Find an approximate solution of Problem (D).
- **Reconstruct** an approximate solution for Problem $(\tilde{P})$ from it.

Approach

In [Wang '17], the following general approach is investigated:

- Find an approximate solution of Problem (D).
- **Reconstruct** an approximate solution for Problem $(\tilde{P})$ from it.

The first method we propose is as follows:

- Same Cutting-plane algorithm as in [Wang '17] to solve Problem (D).
- Primal reconstruction different from [Wang '17], easier to implement in our opinion.

General algorithm

Algorithm Cutting-Plane Algorithm

Require: $\mu^0 \in \mathcal{E}$.

for t=0, 1, . . . **do**

At *Iteration t*, we have $\mu^t \in \mathcal{E}$ and $\left(a^k\right)_{k\in\{0,\dots,t-1\}} \in \tilde{\mathcal{P}}(\tilde{\mathcal{Y}})^t$.

Step 1: Find $a^t \in \operatorname*{argmin}_{a\in\tilde{\mathcal{P}}(\tilde{\mathcal{Y}})} \hat{L}(a, \mu^t)$.

Step 2: Set the approximated dual function

$$Q_{t+1} := \min_{k\in\{0,\dots,t\}} \hat{L}(a^k, \cdot\,) = \min_{a\in\mathrm{conv}\{a^k, k\in\{0,\dots,t\}\}} \hat{L}(a, \cdot\,).$$

Step 3: Find a dual candidate $\mu^{t+1} \in \operatorname*{argmax}_{\mu\in\mathcal{E}} Q_{t+1}(\mu)$.

Step 4: Find $\tilde{a}^{t+1}$ a primal candidate.

end for

Interpretation

Definition

We call **cut** a function of the form $\mu \in \mathcal{E} \mapsto \hat{L}(\bar{a}, \mu)$, for a given $\bar{a}$.
We say that it is **exact** at μ if $\hat{L}(\bar{a}, \mu) = Q(\mu) = \inf_{a \in \tilde{\mathcal{P}}(\tilde{\mathcal{Y}})} \hat{L}(a, \mu)$.

Interpretation

Definition

We call **cut** a function of the form $\mu \in \mathcal{E} \mapsto \hat{L}(\bar{a}, \mu)$, for a given $\bar{a}$. We say that it is **exact** at μ if $\hat{L}(\bar{a}, \mu) = Q(\mu) = \inf_{a \in \tilde{\mathcal{P}}(\tilde{\mathcal{Y}})} \hat{L}(a, \mu)$.

- Step 1: find an exact cut at μ^t.
  Best-response procedure (we can take a^t as a N-tuple of Dirac measures).
- Step 2: update the approximation of Q.
- Step 4, in [Wang '17]: almost-projection algorithm.
  We follow a different idea. Results from the study of Step 3.

Convergence results

Theorem

There exists C such that the following assertions hold:

1. *We have the following dual convergence speeds:*

$$V - Q\left(\hat{\mu}^t\right) \leq \frac{C}{t} \quad \text{and} \quad \|\mu^* - \hat{\mu}^t\|_{\mathcal{E}} \leq \frac{C}{\sqrt{t}}.$$

where $\hat{\mu}^t \in \underset{k \in \{0,\dots,t\}}{\operatorname{argmax}} \; Q(\mu^k)$.

2. *We have the following primal convergence speed:*

$$f\left(-K^* \tilde{a}^t\right) - V \leq \frac{C}{\sqrt{t}}.$$

Convergence results

Theorem

There exists C such that the following assertions hold:

1. *We have the following dual convergence speeds:*

$$V - Q\left(\hat{\mu}^t\right) \leq \frac{C}{t} \quad \text{and} \quad \|\mu^* - \hat{\mu}^t\|_{\mathcal{E}} \leq \frac{C}{\sqrt{t}}.$$

where $\hat{\mu}^t \in \operatorname*{argmax}_{k \in \{0,\dots,t\}} Q(\mu^k)$.

2. *We have the following primal convergence speed:*

$$f\left(-K^* \tilde{a}^t\right) - V \leq \frac{C}{\sqrt{t}}.$$

Improves [Wang '17] (no assumption of strong convexity of f).

Primal algorithm

Formulate the previous algorithm only in terms of a^t and $\tilde{a}^{t+1}$.

Algorithm Equivalent Fully-Corrective Frank-Wolfe algorithm

Require: $\tilde{a}^0 \in \tilde{\mathcal{P}}(\tilde{\mathcal{Y}})$.

for t=0, 1, ... **do**

Find $a^t \in \operatorname*{argmin}_{a \in \tilde{\mathcal{P}}(\tilde{\mathcal{Y}})} \langle \underbrace{\nabla f\left(-K^* \tilde{a}^t\right)}_{\mu^t}, -K^* a \rangle_{\mathcal{E}}$.

Find $\tilde{a}^{t+1}$ solution of

$$\underset{\tilde{a} \in \operatorname{conv}\{a^k,\, k \in \{0,\ldots,t\}\}}{\text{minimize}} \quad f\left(-K^* \tilde{a}\right)$$

end for

Better primal convergence result

Yields a better primal convergence speed.

Theorem

There exists C such that for all t,

$$f\left(-K^{*}\tilde{a}^{t}\right) - V \leq \frac{C}{t}.$$

Summary

Context

We have

$$(D) \longleftrightarrow \underset{\mu \in \mathcal{E}}{\text{minimize}} \sup_{a \in \tilde{\mathcal{M}}(\tilde{\mathcal{Y}})} -\hat{L}(a, \mu).$$

We use [Chambolle, Pock '16], Algorithm 4. It is the best-suited algorithm from [Chambolle, Pock '16] for our problem.

- Makes use of the **strong convexity** of f^*.
- Allows for the use of a **nonlinear proximity operator** D_a. We use one deriving from the **Kullback-Leibler** divergence (well-suited for $\tilde{\mathcal{P}}(\tilde{\mathcal{Y}})$).

Context

We have

$$(D) \longleftrightarrow \underset{\mu \in \mathcal{E}}{\text{minimize}} \sup_{a \in \tilde{\mathcal{M}}(\tilde{\mathcal{Y}})} -\hat{L}(a, \mu).$$

We use [Chambolle, Pock '16], Algorithm 4. It is the best-suited algorithm from [Chambolle, Pock '16] for our problem.

- Makes use of the **strong convexity** of f^*.
- Allows for the use of a **nonlinear proximity operator** D_a. We use one deriving from the **Kullback-Leibler** divergence (well-suited for $\tilde{\mathcal{P}}(\tilde{\mathcal{Y}})$).

Assumption

- *The sets $\mathcal{Y}_i$ are finite, of cardinal n_i.*
- *For all $\tau > 0$, we can compute*

$$prox_{\tau f} := \underset{\mu \in \mathcal{E}}{\text{argmin}} \; f + \frac{1}{2\tau} \|\mu - \cdot\|_{\mathcal{E}}^2.$$

Accelerated primal-dual algorithm

Algorithm Accelerated primal-dual algorithm

Require: : Initial guesses $\mu^{-1} = \mu^0 \in \mathcal{E}$ and $a^0 \in \text{ri}\left(\tilde{\mathcal{P}}(\tilde{\mathcal{Y}})\right)$, sequences $(\theta_t)_{t\in\mathbb{N}}, (\tau_t)_{t\in\mathbb{N}}, (\sigma_t)_{t\in\mathbb{N}} \in \left(\mathbb{R}_+^*\right)^{\mathbb{N}}$ well-chosen.

for $t \in \mathbb{N}$ **do**

$$a^{t+1} = \operatorname*{argmin}_{a\in\tilde{\mathcal{P}}(\tilde{\mathcal{Y}})} -\langle \mu^t + \theta_t(\mu^t - \mu^{t-1}), K^* a\rangle_{\mathcal{E}} + \frac{1}{\sigma_t} D_a(a, a^t).$$

$$\mu^{t+1} = \operatorname*{argmin}_{\mu\in\mathcal{E}} \langle \mu, K^* a\rangle_{\mathcal{E}} + f^*(\mu) + \frac{1}{2\tau_t}\|\mu - \mu^t\|_{\mathcal{E}}^2.$$

end for

Convergence result

Theorem

For well-chosen sequences $(\theta_t)_{t\in\mathbb{N}}, (\tau_t)_{t\in\mathbb{N}}, (\sigma_t)_{t\in\mathbb{N}} \in \left(\mathbb{R}_+^*\right)^{\mathbb{N}}$, *there exists* C, *which depends on the data of the problem such that, for all* $T > 1$

- $f\left(-K^* A^T\right) - V \leq \frac{C}{T^2}$
- $V - Q\left(M^T\right) \leq \frac{C}{T^2}$
- $\|M^T - \mu^*\|_{\mathcal{E}} \leq \frac{\sqrt{2\beta C}}{T}$

with for all $T \in \mathbb{N}^*$:

$$S^T = \sum_{t=1}^{T} \frac{\sigma_{t-1}}{\sigma_0}, \quad M^T = \frac{1}{S^T}\sum_{t=1}^{T} \frac{\sigma_{t-1}}{\sigma_0}\mu^t, \quad A^T = \frac{1}{S^T}\sum_{t=1}^{T} \frac{\sigma_{t-1}}{\sigma_0}a^t.$$

Convergence result II

More specifically, the constant C depends on

- The initial guesses (μ^0, a^0).
- The Lipschitz constant β of ∇f.
- The operator norm $\|K\| = \max_{i \in \{1,\dots,N\}} \max_{y_i \in \mathcal{Y}_i} \|y_i\|_{\mathcal{E}}$.
- The cardinals of the sets $\mathcal{Y}_i$, or more precisely $\max_{i \in \{1,\dots,N\}} n_i$.

Convergence result II

More specifically, the constant C depends on

- The initial guesses (μ^0, a^0).
- The Lipschitz constant β of ∇f.
- The operator norm $\|K\| = \max_{i \in \{1,\dots,N\}} \max_{y_i \in \mathcal{Y}_i} \|y_i\|_{\mathcal{E}}$.
- The cardinals of the sets $\mathcal{Y}_i$, or more precisely $\max_{i \in \{1,\dots,N\}} n_i$.

In particular, it depends *directly* on N only through the dependence in N of the functions g_i and of the sets $\mathcal{X}_i$, if there is one.

Summary

Problem

We tested our algorithm on a problem presented in [Bonnans et al. '22], called the MIQP. The objective is to solve the following problem:

$$\underset{x\in\{0,1\}^N}{\text{minimize}} \; \frac{1}{N^2}\|Ax-\bar{y}\|^2 = \left\| \frac{1}{N}\sum_{i=1}^{N}\left(A_i x_i - \frac{\bar{y}}{N}\right)\right\|^2$$

with N, M positive integers, $A \in \mathcal{M}_{M,N}(\mathbb{R})$ and $\bar{y} \in \mathbb{R}^M$.

Problem

We tested our algorithm on a problem presented in [Bonnans et al. '22], called the MIQP. The objective is to solve the following problem:

$$\underset{x\in\{0,1\}^N}{\text{minimize}} \ \frac{1}{N^2}\|Ax-\bar{y}\|^2 = \left\| \frac{1}{N}\sum_{i=1}^{N}\left(A_i x_i - \frac{\bar{y}}{N}\right)\right\|^2$$

with N, M positive integers, $A \in \mathcal{M}_{M,N}(\mathbb{R})$ and $\bar{y} \in \mathbb{R}^M$.
In our computations, we took

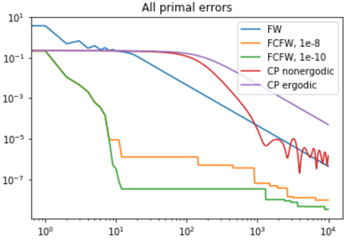
- $N = 100$ and $M = 50$.
- A with coefficients taken randomly, following a uniform law over $[0, 1]$.
- $\bar{y}$ with coordinates taken randomly, following a uniform law over $[0, N/2]$.

Results

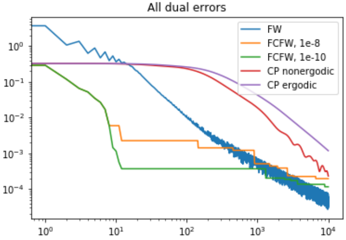
We compare three algorithms.

- The Frank-Wolfe algorithm, investigated in [Bonnans et al. '22].
- The Fully-Corrective Frank-Wolfe algorithm, which is equivalent to our cutting-plane method, for two tolerances for Problem $(\tilde{P}_t)$
- The Algorithm from Chambolle-Pock, for both the ergodic and nonergodic sequences.

Comparing the results



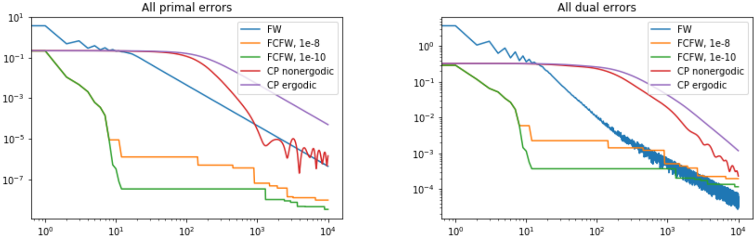
(a) Primal errors



(b) Dual errors

Figure: Comparison of the primal and dual errors for all algorithms, in log-log scale

Comparing the results



(a) Primal errors (b) Dual errors

Figure: Comparison of the primal and dual errors for all algorithms, in log-log scale

- On this example, the FCFW is clearly much more effective, at least on the first iterations.
- However, these curves can be misleading. An iteration of the FCFW can indeed become very difficult to compute.