Aggregative optimization problems: relaxation and numerical resolution

Laurent Pfeiffer
Inria and CentraleSupélec, Université Paris-Saclay

Joint work with Frédéric Bonnans (Inria, L2S), Kang Liu (Polytechnique, L2S), Nadia Oudjane and Cheng Wan (EDF).

SIERRA Seminar
Inria Paris
November 9, 2022

Introduction

We investigate large scale **aggregative optimization problem**.

- Approximation by a convex mean-field optimization problem.
- Estimation of the relaxation gap.
- Numerical resolution with the **conditional gradient algorithm** (also called **Frank-Wolfe** algorithm).

Bonnans, Liu, Oudjane, Pfeiffer, Wan. Large-scale nonconvex optimization: randomization, gap estimation, and numerical resolution, *ArXiv preprint*, 2022.

Setting

Consider the problem

$$\inf_{x \in \mathcal{X}} J(x) = f\Bigg(\underbrace{\frac{1}{N}\sum_{i=1}^{N} g_i(x_i)}_{\text{aggregate}}\Bigg) + \frac{1}{N}\sum_{i=1}^{N} h_i(x_i), \qquad (\mathcal{P})$$

where $x = (x_1, \ldots, x_N) \in \mathcal{X} = \prod_{i=1}^{N} \mathcal{X}_i$.

Data:

- $\mathcal{E}$, a Hilbert space (the aggregate space)
- $g_i : \mathcal{X}_i \to \mathcal{E}$, $i = 1, \ldots, N$
- $h_i : \mathcal{X}_i \to \mathbb{R}$, $i = 1, \ldots, N$
- $f : \mathcal{E} \to \mathbb{R}$.

Interpretation

A **multi-agent** model:

- N: the number of agents
- $\mathcal{X}_i$: the decision set of agent i
- $h_i(x_i)$: individual cost function of agent i
- $g_i(x_i)$: contribution of agent i to a common good
- $\frac{1}{N}\sum_{i=1}^{N} g_i(x_i)$: a common good, referred to as aggregate
- f: a social cost associated with the aggregate.

Wang. Vanishing Price of Decentralization in Large Coordinative Nonconvex Optimization, *SIAM J. Optimization*, 2017.

Application

Applications in energy management problems:

- Set of agents: a (large) set of **small flexible consumptions units** (e.g. batteries, heating devices).
  Flexible: consumption can be shifted over time.
- Aggregate: the **total consumption**, at each time step of a given time interval.
- Social cost: **penalty function** for the difference between total consumption and a reference production level (typically highly variable because of the incorporation of renewable energy sources).

Séguret et al. Decomposition of high dimensional aggregative stochastic control problems, *ArXiv preprint*, 2021.

Applications

Our problem covers the case **training neural networks with a single hidden layer**.

- Social cost $\rightarrow$ fidelity function.
- Individual cost $\rightarrow$ regulizer.

We use the same kind of relaxation as in:

Chizat, Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport, *Advances in Neural Information Processing Systems*, 2018.

Assumptions

Assumptions:

- f is convex
- ∇f is D-Lipschitz continuous
- for all $i = 1, \ldots, N$, $\mathrm{diam}(g_i(\mathcal{X}_i)) \leq D$.

All constants appearing later on depend on D but not on N.
Another "numerical" assumption will be made later.

General difficulties:

- No convexity property of J.
- No regularity property for $\mathcal{X}_i$, g_i, h_i. In general, J is not differentiable.
- Large-scale (when N is large)... but N large actually helps!

Relaxation

General idea:

- Variable x_i replaced by a **probability distribution** $\mu_i \in \mathcal{P}(\mathcal{X}_i)$.
- The terms $g_i(x_i)$ and $h_i(x_i)$ are respectively replaced by

$$\mathbb{E}_{\mu_i}[g_i] := \int_{\mathcal{X}_i} g_i(x_i)\,\mathrm{d}\mu_i(x_i), \quad \mathbb{E}_{\mu_i}[h_i] := \int_{\mathcal{X}_i} h_i(x_i)\,\mathrm{d}\mu_i(x_i).$$

The relaxed problem:

$$\inf_{\mu} \ \tilde{J}(\mu) := f\Big(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mu_i}[g_i]\Big) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mu_i}[h_i], \qquad (\tilde{\mathcal{P}})$$

where $\mu = (\mu_1, \ldots, \mu_N) \in \prod_{i=1}^{N} \mathcal{P}(\mathcal{X}_i)$.

Remark: The cost function $\tilde{J}$ is **convex**.

Gap estimation

Theorem

There exists $C > 0$ (depending on D only) such that

$$\mathrm{Val}(\tilde{\mathcal{P}}) \leq \mathrm{Val}(\mathcal{P}) \leq \mathrm{Val}(\tilde{\mathcal{P}}) + \frac{C}{N}.$$

Proof. **Lower bound** of $\mathrm{Val}(\mathcal{P})$.
Let $x \in \mathcal{X}$. Let $\mu = (\delta_{x_1}, ..., \delta_{x_N})$. Then,

$$\mathrm{Val}(\tilde{\mathcal{P}}) \leq \tilde{J}(\mu) = J(x).$$

Minimizing with respect to x yields the result.

Gap estimation

Upper bound of $\mathrm{Val}(\mathcal{P})$. Let $\varepsilon > 0$. Let $\mu \in \prod_{i=1}^{N} \mathcal{P}(\mathcal{X}_i)$ be ε-optimal for the relaxed problem.
Let X_1,...,X_N be N independent random variables such that

$$\mathrm{Law}(X_i) = \mu_i, \quad i = 1, ..., N.$$

Then, setting $Y = \frac{1}{N} \sum_{i=1}^{N} g_i(X_i)$,

$$\begin{aligned} \tilde{J}(\mu) &= f\Big(\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[g_i(X_i)] \Big) + \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[h_i(X_i)], \\ &= f(\mathbb{E}[Y]) + \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[h_i(X_i)]. \end{aligned}$$

Therefore, $\mathbb{E}[J(X)] - \tilde{J}(\mu) = \mathbb{E}[f(Y)] - f(\mathbb{E}[Y])$.

Gap estimation

Using the Lipschitz continuity of ∇f, it is easy to show that:

$$\mathbb{E}[f(Y)] - f(\mathbb{E}[Y]) \leq \frac{L}{2}\mathbb{E}\Big[\|Y - \mathbb{E}[Y]\|^2\Big]$$

Since $Y = \frac{1}{N}\sum_{i=1}^N g_i(X_i)$ and since the X_i are independent,

$$\|Y - \mathbb{E}[Y]\|^2 = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\Big[\|g_i(X_i) - \mathbb{E}[g_i(X_i)]\|^2\Big] \leq \frac{D^2}{N}.$$

It finally follows that

$$\begin{aligned} \mathrm{Val}(\mathcal{P}) - \mathrm{Val}(\tilde{\mathcal{P}}) &\leq \mathbb{E}[J(X)] - \tilde{J}(\mu) + \varepsilon \\ &\leq \frac{L}{2}\mathbb{E}\Big[\|Y - \mathbb{E}[Y]\|^2\Big] + \varepsilon \leq \frac{D^2 L}{2N} + \varepsilon. \end{aligned}$$

Gap estimation

Theorem

Assume that $q := \dim \mathcal{E} + 1 \leq N$. There exists $C > 0$ (depending on D only) such that

$$\mathrm{Val}(\tilde{\mathcal{P}}) \leq \mathrm{Val}(\mathcal{P}) \leq \mathrm{Val}(\tilde{\mathcal{P}}) + \frac{Cq}{N^2}.$$

Proof. Let μ be as before. Using **Shapley-Folkman's** theorem, we can construct independent r.v. $\tilde{X}_i$, valued in $\mathcal{X}_i$ and such that

- $\tilde{J}(\mu) = f(\mathbb{E}[\tilde{Y}]) + \frac{1}{N}\sum_i \mathbb{E}[h_i(\tilde{X}_i)]$, where $\tilde{Y} = \frac{1}{N}\sum_{i=1}^N g_i(\tilde{X}_i)$,
- All r.v. $\tilde{X}_i$ are deterministic, except at most q of them.

Then $\mathbb{E}\big[\|\tilde{Y} - \mathbb{E}[\tilde{Y}]\|^2\big] \leq Cq/N^2$.

Frank-Wolfe algorithm

Consider the following problem:

$$\inf_{x \in \mathbb{R}^n} F(x), \quad \text{subject to: } x \in K. \tag{$\mathcal{P}$}$$

Assumptions:

- $F\colon \mathbb{R}^n \to \mathbb{R}$ is convex, continuously differentiable, with Lipschitz-continuous gradient.
- $K \subseteq \mathbb{R}^n$ is convex and compact.

The **linearized problem** at $\tilde{x}$ is defined by

$$\inf_{x \in \mathbb{R}^n} \langle \nabla F(\tilde{x}), x \rangle, \quad \text{subject to: } x \in K. \tag{$\mathcal{P}_{\text{lin}}(\tilde{x})$}$$

We assume that it is easy to solve numerically, for any $\tilde{x}$.

Frank-Wolfe algorithm

Algorithm 1: Frank-Wolfe algorithm

Input: $\bar{x}_0 \in K$;

for $k = 0, 1, \ldots$ **do**

Find a solution x_k to $\mathcal{P}_{\text{lin}}(\bar{x}_k)$;

Set $\delta_k = 2/(k+2)$;

Set $\bar{x}_{k+1} = (1-\delta_k)\bar{x}_k + \delta_k x_k$;

end

Lemma

There exists a constant C such that

$$f(\bar{x}_k) \leq f(\bar{x}) + \frac{C}{k}, \quad \forall k > 0,$$

where $\bar{x}$ denotes a solution of $(\mathcal{P})$.

The subproblem

We call any map $\mathbb{S} \colon \lambda \in \mathcal{E} \mapsto \big(\mathbb{S}_1(\lambda), \ldots, \mathbb{S}_N(\lambda)\big) \in \mathcal{X}$ a **best-response** function if for any $\lambda \in \mathcal{E}$,

$$\mathbb{S}_i(\lambda) \in \underset{x_i \in \mathcal{X}_i}{\text{argmin}} \; \langle \lambda, g_i(x_i) \rangle + h_i(x_i), \quad \text{for } i = 1, \ldots, N.$$

The variable λ can be here interpreted as a **price** for the contribution to the aggregate.

Numerical assumption. We assume that such a function can be easily constructed numerically. The evaluation of $\mathbb{S}$ relies on the resolution of N **independent** optimization problems.

The subproblem

Lemma

Let $\tilde{\mu} \in \prod_{i=1}^N \mathcal{P}(\mathcal{X}_i)$. *Let* $\lambda = \nabla f\left(\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\tilde{\mu}_i}[g_i]\right)$. *Define*

$$\hat{\mu} = \left(\delta_{\mathbb{S}_1(\lambda)}, \dots, \delta_{\mathbb{S}_N(\lambda)}\right).$$

Then $\hat{\mu}$ *is a solution to*

$$\inf_{\mu \in \prod_{i=1}^N \mathcal{P}(\mathcal{X}_i)} D\tilde{J}(\tilde{\mu}).\mu. \qquad (\tilde{\mathcal{P}}_{\mathrm{lin}}(\tilde{\mu}))$$

Proof. Straightforward calculations yield:

$$D\tilde{J}(\tilde{\mu}).\mu = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mu_i}\Big[\langle \lambda, g_i(\cdot)\rangle + h_i(\cdot)\Big].$$

Frank-Wolfe algorithm

Algorithm 2: Frank-Wolfe algorithm

Input: $\bar{\mu}^{(0)}$;

for $k = 0, 1, \ldots$ **do**

  Find a solution $\mu^{(k)}$ to $\tilde{\mathcal{P}}_{\text{lin}}(\bar{\mu}^{(k)})$;

  Set $\delta_k = \frac{2}{k+2}$;

  Set $\bar{\mu}^{(k+1)} = (1 - \delta_k)\bar{\mu}^{(k)} + \delta_k \mu^{(k)}$;

end

Difficulties:

- The support of $\bar{\mu}_i^{(k)}$ possibly is of cardinality k.
- How to deduce an **approximate solution** to $(\mathcal{P})$ from $\bar{\mu}^{(k)}$?

Selection

Selection: A simple **stochastic method** for constructing $x \in \mathcal{X}$ out of $\mu \in \prod_{i=1}^{N} \mathcal{P}(\mathcal{X}_i)$.

Algorithm 3: Selection algorithm

Input: μ, $n \in \mathbb{N}$;

Construct a random variable $X = (X_1, ..., X_N)$ such that

$$X_1,...,X_N \text{ are independent}, \qquad \mathrm{Law}(X_i) = \mu_i.$$

for $j = 1, ..., n$ **do**

  Draw samples $\hat{x}^j = (x_1^j, ..., x_N^j)$ of $(X_1, ..., X_N)$.

end

Output: $\hat{x} = \underset{x \in \{\hat{x}^1,...,\hat{x}^n\}}{\operatorname{argmin}} J(x)$.

Selection

Lemma

Let $\mu \in \prod_{i=1}^{N} \mathcal{P}(\mathcal{X}_i)$ and let $n \in \mathbb{N}$. There exists a constant $C > 0$ such that for any $\varepsilon > 0$,

$$\mathbb{P}\Big[J(\hat{x}) \geq \tilde{J}(\mu) + \frac{C}{N} + \varepsilon\Big] \leq \exp\Big(-\frac{nN\varepsilon^2}{C}\Big).$$

Proof. Let X be as in the selection algorithm. We know that

$$\tilde{J}(\mu) - \mathbb{E}[J(X)] \leq \frac{C}{N}.$$

Concentration inequality: by McDiarmid's inequality, there exists $C > 0$ such that for any $\varepsilon > 0$,

$$\mathbb{P}\Big[J(X) \geq \mathbb{E}[J(X)] + \varepsilon\Big] \leq \exp\Big(-\frac{N\varepsilon^2}{C}\Big).$$

Stochastic Frank-Wolfe algorithm

Algorithm 4: Stochastic Frank-Wolfe algorithm

Input: $\bar{\mu}^{(0)}$, a sequence $(n_k)_{k\in\mathbb{N}}$;

for $k = 0, 1, \ldots$ **do**

Find a solution $\mu^{(k)}$ to $\tilde{\mathcal{P}}_{\text{lin}}(\bar{\mu}^{(k)})$;

Set $\delta_k = \frac{2}{k+2}$;

Set $\tilde{\mu}^{(k+1)} = (1-\delta_k)\bar{\mu}^{(k)} + \delta_k \mu^{(k)}$;

Set $\hat{x}^{(k+1)} = \text{Selection}(\tilde{\mu}^{(k+1)}, n_{k+1})$;

Set $\bar{\mu}^{(k+1)} = \left(\delta_{\hat{x}_1^{(k+1)}}, \ldots, \delta_{\hat{x}_N^{(k+1)}}\right)$.

end

Practical comments

- At iteration k, $\bar{\mu}^{(k)}$ and $\mu^{(k)}$ are N-uplets of Dirac measures:

$$\bar{\mu}^{(k)} = (\delta_{\hat{x}_1^{(k)}}, \dots, \delta_{\hat{x}_N^{(k)}}), \quad \mu^{(k)} = (\delta_{x_1^{(k)}}, \dots, \delta_{x_N^{(k)}}).$$

- Thus $\text{supp}(\tilde{\mu}_i^{(k+1)}) = \{\hat{x}_i^{(k)}, x_i^{(k)}\}$.
- We need to simulate a random variable X_i such that

$$\mathbb{P}\Big[X_i = \hat{x}_i^{(k)}\Big] = \frac{k}{k+2}, \quad \mathbb{P}\Big[X_i = x_i^{(k)}\Big] = \frac{2}{k+2}.$$

Simulate a uniformly distributed r.v. W_i in $[0,1]$ and define

$$X_i = \begin{cases} \hat{x}_i^{(k)} & \text{if } W_i \leq k/(k+2) \\ x_i^{(k)} & \text{otherwise.} \end{cases}$$

- **Speed up:** we can simulate W_i before computing $x_i^{(k)}$ and thus figure out whether it is necessary to compute $x_i^{(k)}$ **or not**.

Convergence result

Theorem

There exists a constant $C > 0$ *such that for all* $K \leq 2N$, *for all* $\varepsilon > 0$, *it holds:*

$$\mathbb{P}\Big[J(\hat{x}^K) \geq \mathrm{Val}(\tilde{P}) + \frac{C}{K} + \varepsilon\Big] \leq \exp\Big(-\frac{N\varepsilon^2}{C_1(K) + \varepsilon C_2(K)}\Big),$$

where

$$C_1(K) = C \sum_{k=1}^{K-1} \frac{k(k+1)^2}{n_k K^2 (K+1)^2},$$

$$C_2(K) = C \max_{k \leq K-1} \frac{(k+1)(k+2)}{n_k K(K+1)}.$$

Remark. We can find a C/N-optimal solution with arbitrarily small probability if $n_k \geq Ak^2/N$, with A large enough.

Numerical example

Let $A \in \mathbb{R}^{M\times N}$ and let $\bar{y} \in \mathbb{R}^{M}$. Consider:

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

Data: $M = N = 100$.

Remark: Problem (MIQP) is a discrete problem, over a set of cardinality 2^{100}.

Numerical example

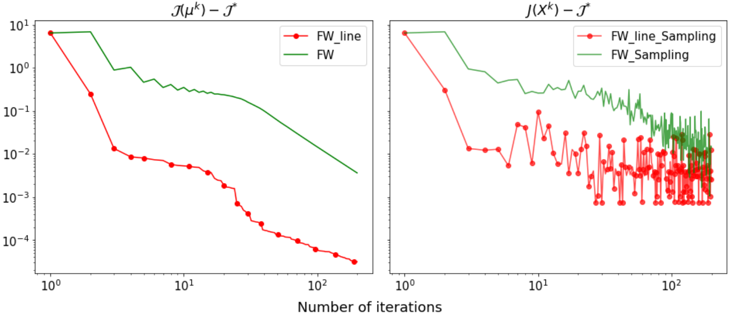

Figure: Convergence of the relaxed optimality gap.

Left: Frank-Wolfe for the relaxed problem.
Right: Selection algorithm applied to the iterates.

Numerical example

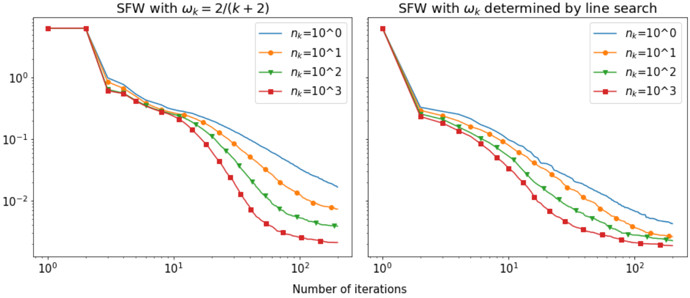

Figure: Relaxed optimality gap for Stochastic Frank-Wolfe algorithm.

Left: Stepsize $\delta_k = 2/(k+2)$.
Right: Stepsize determined by line-search.

Setting

Mathematical setting:

- Each agent optimize a deterministic and discrete dynamical system.
- The aggregate is a time function.

Application:

- The agents are **batteries**, whose state is the state-of-charge of the battery and the control is the loading speed.
- The aggregate is the total energy consumption induced by the loading of the batteries.

Dynamics of the agents

Let $T \in \mathbb{N}$ be the time horizon, let

$$\mathcal{T} = \{0, 1, \ldots, T-1\} \quad \text{and} \quad \bar{\mathcal{T}} = \{0, 1, \ldots, T\}.$$

For each agent i, consider:

- A family of finite **state** sets $(S_i^t)_{t \in \bar{\mathcal{T}}}$.
- A family of finite **control** sets $(U_i^t(s_i^t))_{t \in \mathcal{T}, s_i^t \in S_i^t}$.
- For any $t \in \mathcal{T}$, for any $s_i^t \in S_i^t$, a **transition function**

$$u_i^t \in U_i^t(s_i^t) \mapsto \pi_i^t(s_i^t, u_i^t) \in S_i^{t+1}.$$

We call an element $x_i = \big((s_i^t)_{t \in \bar{\mathcal{T}}}, (u_i^t)_{t \in \mathcal{T}}\big)$ a **state-control trajectory**. We denote by $\mathcal{X}_i$ the set of elements x_i such that

$$\begin{cases} s_i^t \in S_i^t, \quad \forall t \in \bar{\mathcal{T}} \\ u_i^t \in U_i(s_i^t) \quad \text{and} \quad s_i^{t+1} = \pi_i^t(s_i^t, u_i^t), \quad \forall t \in \mathcal{T}. \end{cases}$$

Cost and contribution

Let $\mathcal{E} = \prod_{t\in\mathcal{T}} \mathcal{E}_t$, where every $\mathcal{E}_t$ is a Hilbert space.
For each $i = 1, \ldots, N$ and for each $t \in \mathcal{T}$, fix two functions

$$g_i^t : (s_i^t, u_i^t) \mapsto g_i^t(s_i^t, u_i^t) \in \mathcal{E}_t$$
$$h_i^t : (s_i^t, u_i^t) \mapsto h_i^t(s_i^t, u_i^t) \in \mathbb{R},$$

for any $s_i^t \in S_i^t$ and for any $u_i^t \in U_i^t(s_i^t)$.

For an agent i, the individual cost h_i and the contribution function g_i are defined by

$$g_i(x_i) = \left(g_i^t(s_i^t, u_i^t)\right)_{t\in\mathcal{T}} \in \mathcal{E}$$
$$h_i(x_i) = \sum_{t\in\mathcal{T}} h_i^t(s_i^t, u_i^t).$$

Social cost and problem

For any $t \in \mathcal{T}$, let $f_t \colon \mathcal{E}_t \to \mathbb{R}$ and let $f \colon \mathcal{E} \to \mathbb{R}$ be defined by

$$f(y_0, \ldots, y_{T-1}) = \sum_{t \in \mathcal{T}} f_t(y_t).$$

We consider the corresponding aggregative problem:

$$\inf_{x \in \prod_{i=1}^N \mathcal{X}_i} f\Big(\frac{1}{N} \sum_{i=1}^N g_i(x_i) \Big) + \frac{1}{N} \sum_{i=1}^N h_i(x_i).$$

which be equivalently written as...

Aggregative control problem

$$\inf_{\substack{(s_i^t)_{t\in\bar{\mathcal{T}},\, i=1,\dots,N}\\(u_i^t)_{t\in\mathcal{T},\, i=1,\dots,N}}} \sum_{t\in\mathcal{T}} f_t\left(\frac{1}{N}\sum_{i=1}^N g_i^t(s_i^t,u_i^t)\right) + \frac{1}{N}\sum_{i=1}^N\sum_{t\in\mathcal{T}} h_i^t(s_i^t,u_i^t),$$

$$\text{subject to: } \begin{cases} s_i^t \in S_i^t, & \forall i=1,\dots,N,\ \forall t\in\bar{\mathcal{T}} \\ s_i^{t+1} = \pi_i^t(s_i^t,u_i^t), & \forall i=1,\dots,N,\ \forall t\in\mathcal{T} \\ u_i^t \in U_i^t(s_i^t), & \forall i=1,\dots,N,\ \forall t\in\mathcal{T}. \end{cases}$$

This is a discrete-time, finite-space optimal control problem with state set $\Pi_{i=1}^N S_i^t$, at time t.

→ Huge cardinality, dynamic programming is **intractable**.

Best-response function

We focus now on the computation of the **best-response** function.

Let $\lambda = (\lambda_t)_{t\in\mathcal{T}} \in \mathcal{E}$. The associated subproblem, for agent i, reads:

$$\inf_{\substack{(s_i^t)_{t\in\bar{\mathcal{T}}} \\ (u_i^t)_{t\in\mathcal{T}}}} \sum_{t\in\mathcal{T}} \Big(\langle \lambda_t, g_i^t(s_i^t, u_i^t) \rangle + h_i^t(s_i^t, u_i^t) \Big),$$

$$\text{subject to: } \begin{cases} s_i^t \in S_i^t, & \forall i = 1,\ldots,N,\ \forall t \in \bar{\mathcal{T}} \\ s_i^{t+1} = \pi_i^t(s_i^t, u_i^t), & \forall i = 1,\ldots,N,\ \forall t \in \mathcal{T} \\ u_i^t \in U_i^t(s_i^t), & \forall i = 1,\ldots,N,\ \forall t \in \mathcal{T}. \end{cases}$$

It is a discrete-time, finite-space optimal control problem with state set S_i^t, at time t.

$\rightarrow$ Tractable by **dynamic programming**, if $|S_i^t|$ is not too large!

Best-response function

The resolution is done in two steps.

1. **Backward phase.** We compute a value function,

$$(t, s_i^t) \mapsto V_i^t(s_i^t) \in \mathbb{R}.$$

It models the optimal cost for the initial time t and initial condition s_i^t.

- Define $V_i^T(s_i^T) = 0$.
- For $t = T-1, T-2, \dots, 0$, compute

$$V_i^t(s_i^t) = \min_{u_i^t \in U_i^t(s_i^t)} \Big(\langle \lambda_t, g_i^t(s_i^t, u_i^t) \rangle + h_i^t(s_i^t, u_i^t) + V_i^{t+1}\Big(\pi_i^t(s_i^t, u_i^t)\Big)\Big),$$

for all $s_i^t \in S_i^t$.

Best-response function

2. Forward phase.

- Find $\bar{s}_i^0 \in \operatorname{argmin} V_i^0(\cdot)$.
- For $t = 0, 1, \ldots, T-1$, compute

$$\bar{u}_i^t \in \underset{u_i^t \in U_i^t(\bar{s}_i^t)}{\operatorname{argmin}} \Big(\langle \lambda_t, g_i^t(\bar{s}_i^t, u_i^t) \rangle + h_i^t(\bar{s}_i^t, u_i^t) + V_i^{t+1}\big(\pi_i^t(\bar{s}_i^t, u_i^t)\big)\Big),$$

and set $\bar{s}_i^{t+1} = \pi_i^t(\bar{s}_i^t, u_i^t)$.

Then $\big((\bar{s}_i^t)_{t \in \bar{\mathcal{T}}}, (\bar{u}_i^t)_{t \in \mathcal{T}}\big)$ is a solution to the subproblem.

Dual problem

Joint work with Thibault Moquet (Optimization Master, Paris-Saclay University, now at L2S).

Without loss of generality, we assume that $h_i = 0$, $\forall i = 1, \ldots, N$.

Saddle-point formulation of the problem:

$$\inf_{\mu \in \prod_{i=1}^N \mathcal{M}(\mathcal{X}_i)} \sup_{\lambda \in \mathcal{E}} L(\mu, \lambda),$$

where the Lagrangian L is defined by

$$L(\mu, \lambda) = \left\langle \lambda, \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\mu_i}[g_i] \right\rangle - f^*(\lambda) + \sum_{i=1}^N \iota_{\mathcal{P}(\mathcal{X}_i)}(\mu_i),$$

where f^* denotes the Fenchel conjugate of f, $\iota_{\mathcal{P}(\mathcal{X}_i)}$ the indicatrix function, and $\mathcal{M}(\mathcal{X}_i)$ the space of finite signed measures.

Dual problem

The **dual problem** reads:

$$\sup_{\lambda \in \mathcal{E}} Q(\lambda) := \left(\inf_{\mu \in \prod_{i=1}^{N} \mathcal{M}(\mathcal{X}_i)} L(\mu, \lambda) \right).$$

We have

$$\begin{aligned} Q(\lambda) &= \inf_{\mu} \left\langle \lambda, \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mu_i}[g_i] \right\rangle - f^*(\lambda) + \sum_{i=1}^{N} \iota_{\mathcal{P}(\mathcal{X}_i)}(\mu_i) \\ &= \underbrace{\frac{1}{N} \sum_{i=1}^{N} \left(\inf_{x_i \in \mathcal{X}_i} \langle \lambda, g_i(x_i) \rangle \right)}_{=:-\sigma(x)} - f^*(\lambda). \end{aligned}$$

Optimality conditions

Theorem

1. *The dual problem has a unique solution* $\bar{\lambda}$.
2. *Any* μ *is a solution to the relaxed problem if and only if for* $\lambda = \nabla f\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mu_i}[g_i]\right)$, *it holds that*

$$\mathit{supp}(\mu_i) \subseteq \underset{x_i \in \mathcal{X}_i}{\mathit{argmin}}\ \langle \lambda, g_i(x_i) \rangle.$$

3. *If* μ *is optimal, then* $\bar{\lambda} = \nabla f\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mu_i}[g_i]\right)$.

Remark. The conditions in the second statement can be interpreted as the **equilibrium conditions** for a **game** with N groups of infinitely many agents.

Numerical methods

- The saddle point formulation of the dual problem can be addressed with **Chambolle-Pock**'s algorithm (using Kullback-Leibler distance in the non-linear proximity operators), when the sets $\mathcal{X}_i$ are finite, with small cardinality. When ∇f is Lipschitz continuous, f^* is strongly convex and convergence with rate $\mathcal{O}(1/k^2)$ can be achieved. Empirically less efficient than Frank-Wolfe.

- Mengdi Wang proposed a **cutting-plane** type algorithm for solving the dual problem. At each iteration, a piecewise affine approximation of σ is utilized to approximate Q. The algorithm turns out to be equivalent to the **fully corrective** variant of the Frank-Wolfe algorithm.

Extensions

What can we do if f is convex, but **not continuously differentiable** ? Example: $f = f_0 + \iota_\Omega$.

- **Gap estimate**: replace f by its Moreau enveloppe

$$f_\varepsilon(y) = \inf_{z \in \mathcal{E}} \frac{1}{2\varepsilon}\|y - z\|^2 + f(z).$$

  Then bound $\|f - f_\varepsilon\|_\infty$, use the $1/\varepsilon$-Lipschitz continuity of ∇f_ε for a suitable value of ε.

- **Numerics:** if the sets $\mathcal{X}_i$ are finite, then Chambolle-Pock can be used again (convergence rate of $\mathcal{O}(1/k)$).

- **Numerics:** use an appropriate extension of Frank-Wolfe to for non-smooth cost (not tested yet...).

References

Chambolle, Pock. On the ergodic convergence rates of a first-order primal–dual algorithm, *Mathematical Programming*, 2016.

Silveti-Falls, Molinari, Fadili. Generalized conditional gradient with augmented lagrangian for composite minimization, *SIAM Journal on Optimization*, 2020.

Thank you for your attention!