LARGE-SCALE NONCONVEX OPTIMIZATION: RANDOMIZATION, GAP ESTIMATION, AND NUMERICAL RESOLUTION*

J. FRÉDÉRIC BONNANS[†], KANG LIU[†‡], NADIA OUDJANE[§], LAURENT PFEIFFER[†], AND CHENG WAN[§]

Abstract. We address a large-scale and nonconvex optimization problem, involving an aggregative term. This term can be interpreted as the sum of the contributions of N agents to some common good, with N large. We investigate a relaxation of this problem, obtained by randomization. The relaxation gap is proved to converge to zeros as N goes to infinity, independently of the dimension of the aggregate. We propose a stochastic method to construct an approximate minimizer of the original problem, given an approximate solution of the randomized problem. McDiarmid's concentration inequality is used to quantify the probability of success of the method. We consider the Frank-Wolfe (FW) algorithm for the resolution of the randomized problem. Each iteration of the algorithm requires to solve a subproblem which can be decomposed into N independent optimization problems. A sublinear convergence rate is obtained for the FW algorithm. In order to handle the memory overflow problem possibly caused by the FW algorithm, we propose a stochastic Frank-Wolfe (SFW) algorithm, which ensures the convergence in both expectation and probability senses. Numerical experiments on a mixed-integer quadratic program illustrate the efficiency of the method.

Key words. Large-scale and nonconvex optimization, aggregative optimization, relaxation, decentralization, Frank-Wolfe algorithm, concentration inequalities, multi-agent optimization.

AMS subject classifications. 49M20, 49M27, 90C06, 90C26

1. Introduction.

Problem formulation. This article is devoted to the theoretical analysis and the numerical resolution of the following large-scale, aggregative, and nonconvex optimization problem:

$$\text{(P)} \qquad \inf_{x\in\mathcal{X}} J(x) \coloneqq f(G(x)), \qquad \text{where: } \begin{cases} G(x) = & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} g_i(x_i) \\ \mathcal{X} = & \prod_{i=1}^{N} \mathcal{X}_i. \end{cases}$$

Here, N can be seen as the number of agents and is assumed to be large. The main feature of this problem is the aggregative form of the function G, which is defined as the average of the N mappings g_i, each of which defined on some set $\mathcal{X}_i$ into a Hilbert space $\mathcal{E}$. These mappings are referred to as the contribution mappings. We will refer to the term $G(x)$ as the aggregate. Let q denote the dimension of $\mathcal{E}$ (possibly $q = +\infty$). While very few structural assumptions are made on the sets $\mathcal{X}_i$ and the mappings g_i, we will typically assume that f is convex, with a Lipschitz-continuous gradient. A central idea in this work is that the problem can be well approximated by a convex problem when N is large.

*J.F. Bonnans was partially supported by the FiME Lab Research Initiative (Institut Europlace de Finance). This article benefited from the support of the FMJH Program PGMO and from the support to this program from EDF.

[†] Université Paris-Saclay, CNRS, CentraleSupélec, Inria, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France (frederic.bonnans@inria.fr, laurent.pfeiffer@inria.fr).

[‡] Institut Polytechnique de Paris, CNRS, Ecole Polytechnique, CMAP, 91128 Palaiseau, France (kang.liu@polytechnique.edu).

[§] OSIRIS Department, EDF Lab, Paris-Saclay, and FiME, France (nadia.oudjane@edf.fr, cheng.wan.05@polytechnique.org).

Motivating examples. A particularly interesting instance of (P) is the following:

$$J(x) = f_0\left(\frac{1}{N}\sum_{i=1}^{N} h_i(x_i)\right) + \frac{1}{N}\sum_{i=1}^{N} l_i(x_i). \tag{1.1}$$

This is the setting investigated in a closely related paper by Mengdi Wang [21]. Following her terminology, the function h_i is the contribution of agent i to some common goods, f_0 is a social cost function of the common goods, and l_i describes the individual preference of agent i. There are various applications fitting into the framework of (1.1), see [21]. In particular, some power system management problems can be modeled as (1.1). Such a problem is investigated in [17]: x_i represents the production profile of the generator i, $l_i(x_i)$ is its individual production cost, f_0 denotes the demand elasticity or, equivalently, a penalty function that depends on the difference between the average production and some inflexible demand D (e.g. $f_0 := \|\cdot - D\|^2$) so as to penalize the deviation of the overall production from the inflexible demand.

The problem (1.1) also has applications in supervised learning. Let us consider a training set $(y_j^*, z_j^*)_{j=1}^J \in (\mathcal{Y}\times\mathcal{Z})^J$. The goal is to find parameters $(x_i)_{i=1}^N$ such that the inputs y_j^* are mapped to the outputs z_j^* according to the following relation: $z_j^* \approx \frac{1}{N}\sum_{i=1}^N \varphi(x_i, y_j^*)$, where φ is an activation function (ReLu, Sigmoid, etc). This leads to a problem of the form (1.1), where $h_i(x_i) = (\varphi(x_i, y_j^*))_{j=1,\dots,J} \in \mathcal{Z}^J$, f_0 is a fidelity function penalizing the difference between a given J-uplet $z \in \mathcal{Z}^J$ and $(z_j^*)_{j=1,\dots,J}$. The functions l_i can be chosen so as to regularize the parameters x_i. The framework of problem (1.1) therefore includes various applications of machine learning problems. Let us mention the "sharing problem" in [4], Lasso regression in [9], sparsity regression in [13, 12], and the training of neural networks with one hidden layer in [6].

Related works and methods. Let us return to the general problem (P). Classical Lagrangian relaxation (Chapter XII of [10]) methods can be relevant here because the dual problem is separable, thanks to the aggregative form of G. To see this, let us reformulate (P) as: $\inf_{(x,v)\in\mathcal{X}\times\mathcal{E}} f(v)$, subject to the constraint that $v = G(x)$. Its dual problem is:

$$\sup_{\lambda\in\mathcal{E}} \big(-f^*(\lambda) + \Phi(\lambda)\big), \tag{1.2}$$

where f^* is the Fenchel congugate function of f, and $\Phi(\lambda)$ is defined by

$$\Phi(\lambda) := \inf_{x\in\mathcal{X}} \langle \lambda, G(x)\rangle = \frac{1}{N}\sum_{i=1}^{N} \inf_{x_i\in\mathcal{X}_i} \langle \lambda, g_i(x_i)\rangle. \tag{1.3}$$

One sees that $\Phi(\lambda)$ can be evaluated by solving N independent sub-problems, one for each i in $\{1,\dots,N\}$. Solving these sub-problems can be much easier than addressing frontally the original problem with N coupled variables. This approach has been extensively employed in convex settings [17, 15]. However, the nonconvexity of the problem raises two major difficulties: the potentially large duality gap and the reconstruction of a primal solution from the dual optimal solution.

These two difficulties are addressed by Wang in [21]. She proposed a convex relaxation of the problem, based on a geometrical approach, that allows to obtain an estimate of the duality gap of order $\mathcal{O}(q^2/N^2)$. Her main tool was the Shapley-Folkman lemma [18], which allows to show that the image of G is close to a convex set. This idea was already present in the seminal work of Aubin and Ekeland in [1], dealing with a different setting involving a coupling constraint. We refer to [21] for a more exhaustive of mathematical works dedicated to the estimation of the duality gap, where a kind of convexification occurs. After having solved the dual problem by a cutting plane method and then found an approximate solution to the relaxed primal problem via a projection problem, Wang's method recovers an approximate solution to the original nonconvex problem, by computing a Shapley-Folkman decomposition of the aggregate with a standard linear programming approach.

There exist another important class of methods for large-scale optimization problems which are the block coordinate descent algorithm and its variants [3, 9]. These methods may not be applicable without additional assumptions on the sets $\mathcal{X}_i$ and the maps g_i (in the current framework, the sets $\mathcal{X}_i$ could be discrete). Even if we make additional regularity assumptions, they may be inefficient, in particular because the cost function J is not convex in general.

Contributions and organization of the paper. We first introduce in Section 2 a convex relaxation of the original problem (P). The relaxed problem is obtained by replacing the variables x_i by probability measures μ_i on $\mathcal{X}_i$. The contribution mappings $g_i(x_i)$ are replaced by $\int_{\mathcal{X}_i} g_i(x_i) d\mu_i(x_i)$; these terms are linear with respect to μ_i. The resulting relaxed cost function, denoted $\mathcal{J}$, is convex, and so is the relaxed problem. We give a first upper bound of the relaxation gap of order $\mathcal{O}(1/N)$. The relaxed problem has a stochastic interpretation: it amounts to replace the variables x_i by independent random variables X_i of probability distribution μ_i, and to replace $g_i(x_i)$ by the expectation of $g_i(X_i)$. To derive a good candidate (for (P)), given an approximate solution to the relaxed problem $\mu = (\mu_1, ..., \mu_N)$, we propose to simulate random variables X_i with probability distribution μ_i. We will call this technique the *selection method*. We give a sharp estimate of the probability of error for the selection method. More precisely, we estimate the probability that $J(X_1, \ldots, X_N) \geq \mathcal{J}(\mu) + \left(\frac{C}{N} + \epsilon\right)$, given $\epsilon > 0$. The proof relies on McDiarmid's inequality, a concentration inequality [14].

From a numerical point of view, our main contribution is a method combining the Frank-Wolfe (FW) algorithm [8, 11], applied to the relaxed problem, and the selection method described previously. The resulting algorithm, called stochastic Frank-Wolfe (SFW) algorithm, is described and analyzed in Section 3. Each iteration of the algorithm requires to solve a subproblem of the form (1.3), which is decomposable into N subproblems. Resorting to the selection method, we avoid to manipulate explicitely probability measures on the sets $\mathcal{X}_i$, which may otherwise cause memory issues. The SFW method is able to find an $\mathcal{O}(1/N)$-solution to problem P. In addition, we estimate the probability that the iterate x_k is $\left(\frac{C}{k} + \epsilon\right)$-optimal, for $k \leq 2N$. This result relies on concentration inequalities for martingales [7] which generalize McDiarmid's inequality.

Our last theoretical contribution is a sharp estimate of the relaxation gap, of order $\mathcal{O}(q \wedge N/N^2)$, where q is the (potentially infinite) dimension of the aggregate space $\mathcal{E}$. It is proved in Section 4. It relies on a geometrical relaxation of problem (P), shown to be equivalent to the relaxation by randomization. The relaxation gap is estimated with the help of a measure of nonconvexity for sets (introduced in [5]) and with the help of the Shapley-Folkman lemma [18]. We also give an estimate of the price of decentralization (as defined by Wang in [21]). We conclude the section with a detailed comparison of our approach and the one of [21].

Section 5 is dedicated to numerical tests for a mixed-integer linear-quadratic program.

1.1. Notations.

On sets. For two sets $\mathcal{A}$ and $\mathcal{B}$ in a normed vector space $\mathcal{X}$, we denote by $d(\mathcal{A}) := \sup_{x,y\in\mathcal{A}} \|x-y\|_{\mathcal{X}}$ the diameter of $\mathcal{A}$, by $\mathcal{A}+\mathcal{B} = \{x+y \mid x \in \mathcal{A}, y \in \mathcal{B}\}$ the Minkowski sum of A and B, by $\lambda\mathcal{A} = \{\lambda x \mid x \in \mathcal{A}\}$ the scalar multiplication of A with $\lambda \in \mathbb{R}$ and by $\text{conv}(\mathcal{A})$ the convex hull of $\mathcal{A}$. Note that $\text{conv}(\mathcal{A}+\mathcal{B}) = \text{conv}(\mathcal{A}) + \text{conv}(\mathcal{B})$.

For all $i \in \{1,\ldots,N\}$, we denote $\mathcal{X}_{-i} = \left(\prod_{i'=1}^{i-1} \mathcal{X}_{i'}\right) \times \left(\prod_{i'=i+1}^{N} \mathcal{X}_{i'}\right)$. Given $x \in \mathcal{X}$, we denote $x_{-i} = (x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_N) \in \mathcal{X}_{-i}$. From time to time, we represent x by the pair (x_i, x_{-i}).

On functions. Let $\mathcal{H}$ be a Hilbert space and let $F\colon \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$. The domain of F, denoted by $\text{dom}(F)$, is defined by $\text{dom}(F) = \{x \mid F(x) \neq +\infty\}$. The subgradient of F at some point $x \in \text{dom}(F)$ is denoted by $\partial F(x)$ and defined by

$$\partial F(x) = \{p \in \mathcal{H} \mid F(y) \geq F(x) + \langle p, y-x\rangle, \forall\, y \in \mathcal{H}\}.$$

The Fenchel's conjugate of F is denoted by $F^*\colon H \to \mathbb{R}$ and defined by $F^*(p) = \sup_{x\in\mathcal{H}} \langle p, x\rangle - F(x)$.

On measures. Given a set Ω, we denote by δ_x the Dirac distribution at some point $x \in \Omega$. We denote by $\mathcal{P}_\delta(\Omega)$ the set of finitely supported probability distributions, defined by

$$\mathcal{P}_\delta(\Omega) \coloneqq \text{conv}(\{\delta_x \mid x \in \Omega\}).$$

Let $\mu = \sum_{j=1}^{J} \lambda_j \delta_{x_j} \in \mathcal{P}_\delta(\Omega)$. Given a Hilbert space $\mathcal{H}$ and a mapping $F\colon \Omega \to \mathcal{H}$, we denote

$$E_\mu\big[F\big] = \sum_{j=1}^{J} \lambda_j F(x_j), \qquad \sigma_\mu^2\big[F\big] = \sum_{j=1}^{J} \lambda_j \big\|F(x_j) - E_\mu\big[F\big]\big\|^2.$$

In other words, $E_\mu\big[F\big]$ is the integral of F with respect to the measure μ and $\sigma_\mu^2\big[F\big]$ is the variance of the probability measure $\sum_{j=1}^{J} \lambda_j \delta_{F(x_j)}$, in the sense of [20, Remark 7.5]. Finally, the Bernoulli distribution with parameter $\omega \in [0,1]$ is denoted by $\text{Bern}(\omega)$.

On numbers and real-valued random variables. We denote by $m \wedge n$ the minimum of the numbers m and n in $\mathbb{R}\cup\{+\infty\}$. Let X be a real-valued random variable. The expectation of X is denoted by $\mathbb{E}[X]$, the variance of X is denoted by $\text{Var}(X)$ and the conditional expectation of X w.r.t. some $\sigma-$algebra $\mathcal{F}$ is denoted by $\mathbb{E}[X \mid \mathcal{F}]$. Given $\mu \in \mathcal{P}_\delta(\Omega)$ and a random variable X in Ω, the notation $X \sim \mu$ indicates that μ is the probability distribution of X.

2. Relaxation by randomization and gap estimation. In this section we first make a structural assumption on the general problem of interest, problem (P). Next we introduce a relaxation of the problem, obtained by randomization. We give an upper bound of the randomization gap in Proposition 2.6. Finally we propose a method to recover an approximate solution to (P), given an approximate solution to the relaxed problem. Its performance is investigated in Theorem 2.9.

2.1. Problem formulation. We assume that the aggregate space $\mathcal{E}$ is the Cartesian product of M separable Hilbert spaces denotes $\mathcal{E}_j$, for $j = 1,\ldots,M$. The contribution mappings are of the form $g_i(x_i) = \big(g_{ij}(x_i)\big)_{j=1,\ldots,M}$, where $g_{ij}\colon \mathcal{X}_i \to \mathcal{E}_j$. Moreover we assume that f is additive: for all $y = (y_1,\ldots,y_M) \in \mathcal{E}$, we have

$f(y) = \sum_{j=1}^{M} f_j(y_j)$, where $f_j : \mathcal{E}_j \to \mathbb{R}$. Hence the criterion J of problem (P) writes

$$J(x) = f(G(x)) = \sum_{j=1}^{M} f_j\Big(\frac{1}{N}\sum_{i=1}^{N} g_{ij}(x_i)\Big). \tag{2.1}$$

Of course this structural assumption on $\mathcal{E}$ and f is not restrictive, since one can take $M = 1$. We will comment on its interest in Remark 2.7.

For any $i = 1, \dots, N$ and for any $j = 1, \dots, M$, we denote

$$S_{ij} \coloneqq \big\{g_{ij}(x_i) \mid x_i \in \mathcal{X}_i\big\} \qquad \text{and} \qquad S_j \coloneqq \frac{1}{N}\sum_{i=1}^{N} S_{ij}.$$

The following regularity assumption will be in force all along the article.

Assumption A. *For $i = 1, 2, \dots, N$ and $j = 1, 2 \dots, M$:*

1. *The range set S_{ij} in $\mathcal{E}_j$ has finite diameter $d_{ij} \coloneqq d(S_{ij})$.*
2. *The function f_j is L_j-Lipschitz on* $\operatorname{conv}(S_j)$.
3. *The function f_j is continuously differentiable on a neighborhood of* $\operatorname{conv}(S_j)$, *and ∇f_j is $\tilde{L}_j$–Lipschitz on* $\operatorname{conv}(S_j)$.

We next define two constants $C_0 > 0$ and $C_1 > 0$ by

$$C_0 = \sum_{j=1}^{M} \Big(L_j \max_{1 \le i \le N} \{d_{ij}\}\Big), \quad \text{and} \quad C_1 = \frac{1}{N}\sum_{j=1}^{M}\Big(\tilde{L}_j \sum_{i=1}^{N} d_{ij}^2\Big).$$

Remark 2.1. We will regularly employ notations of the form $O(h(N, q, k))$, where h is an explicit function of N, q (the dimension of $\mathcal{E}$), and k (some iteration number). We use it to express the fact that some variable is bounded by $C\,h(N, q, k)$, where the constant C only depends on $\big(\max_{1 \le i \le N} d_{ij}\big)_{j=1,\dots,M}$ and the Lipschitz moduli $(L_j)_{j=1,\dots,M}$ and $(\tilde{L}_j)_{j=1,\dots,M}$. With this convention in mind, we have

$$C_0 = O(1) \quad \text{and} \quad C_1 = O(1).$$

Remark 2.2. Our results can be applied to aggregative problems of the form

$$\inf_{x \in \mathcal{X}} \sum_{j=1}^{M} f_j\Big(\sum_{i=1}^{N} \hat{g}_{ij}(x_i)\Big),$$

i.e. of the same form as in (2.1), but without the coefficient $\frac{1}{N}$. Indeed, it suffices to define $g_{ij} = N\hat{g}_{ij}$ to come down to the formulation (2.1) and to use the fact that $d(g_{ij}(\mathcal{X}_i)) = Nd(\hat{g}_{ij}(\mathcal{X}_i))$. The introduction of the coefficient $\frac{1}{N}$ induces a natural scaling of the problem as N increases. It also enables to us to highlight the convexification of the problem as N becomes large, assuming that the coefficients d_{ij} are uniformly bounded.

We state in the following lemma a straightforward inequality, exhibiting the role of the constant C_0. Note that the role of the constant C_1 will be revealed in Lemma 2.6.

Lemma 2.3. *Let Assumption A be satisfied. For all $i \in \{1, \dots, N\}$, for all $x_{-i} \in \mathcal{X}_{-i}$, x_i and x_i' in $\mathcal{X}_i$, it holds:*

$$|J(x_i', x_{-i}) - J(x_i, x_{-i})| \le \frac{C_0}{N}.$$

2.2. The relaxed problem. The *randomized problem* is obtained by replacing each optimization variable x_i by a probability measure $\mu_i \in \mathcal{P}_\delta(\mathcal{X}_i)$. The contribution mappings $g_i(x_i)$ are replaced by their integral with respect to μ_i, $E_{\mu_i}[g_i]$. Denoting $\mathcal{P}_\delta = \prod_{i=1}^N \mathcal{P}_\delta(\mathcal{X}_i)$, we obtain

$$\text{(PR)} \qquad \inf_{\mu \in \mathcal{P}_\delta} \mathcal{J}(\mu) := f\Big(\frac{1}{N}\sum_{i=1}^N E_{\mu_i}[g_i]\Big) = \sum_{j=1}^M f_j\Big(\frac{1}{N}\sum_{i=1}^N E_{\mu_i}[g_{ij}]\Big).$$

The following equality justifies the denomination of the relaxed problem: given $\mu \in \mathcal{P}_\delta$ and given N random variables X_i in $\mathcal{X}_i$ such that $X_i \sim \mu_i$, we have

$$\mathcal{J}(\mu) = f\Big(\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[g_i(X_i)\big]\Big). \tag{2.2}$$

Remark 2.4. Working with probability measures with finite support, we do not need to equip the sets $\mathcal{X}_i$ with a topology and to consider regularity assumptions on the mappings g_i. Note that neither the original problem nor the relaxed one may have a solution under the standing assumptions of the article.

Let J^* denote the value of the primal problem (P) and $\mathcal{J}^*$ that of the randomized problem (PR). One is interested in comparing J^* and $\mathcal{J}^*$. The next lemma gives a direct result for one direction of this comparison.

LEMMA 2.5. *It holds:* $\mathcal{J}^* \leq J^*$.

Proof. Let $x \in \mathcal{X}$. Define $\mu = (\delta_{x_1}, \ldots, \delta_{x_N}) \in \mathcal{P}_\delta$. Then $\mathcal{J}(\mu) = J(x)$. The conclusion follows. ∎

The *randomization gap* is then defined as

$$\text{randomization gap} = J^* - \mathcal{J}^* \geq 0.$$

Next we prove a first upper bound of the randomization gap, of order $O(\frac{1}{N})$.

PROPOSITION 2.6. *Let Assumption A hold true. Let $\mu \in \mathcal{P}_\delta$ and let $(X_i)_{i=1,\ldots,N}$ denote N independent random variables such that $X_i \sim \mu_i$. Then,*

$$\mathbb{E}[J(X)] - \mathcal{J}(\mu) \leq \frac{1}{2N^2}\sum_{j=1}^M \Big(\tilde{L}_j \sum_{i=1}^N \sigma^2_{\mu_i}[g_{ij}]\Big) \leq \frac{C_1}{2N}, \tag{2.3}$$

where $X = (X_1, \ldots, X_N)$. As a consequence, $J^* - \mathcal{J}^* \leq \frac{C_1}{2N}$.

Proof. Let us define $Y_j = \frac{1}{N}\big(\sum_{i=1}^N g_{ij}(X_i)\big)$, for $j = 1, \ldots, M$. Let us set $Y = (Y_j)_{j=1,\ldots,M}$. We have

$$\mathbb{E}\big[J(X)\big] = \mathbb{E}\big[f(Y)\big] \qquad \text{and} \qquad \mathcal{J}(\mu) = f\big(\mathbb{E}[Y]\big).$$

Since the variables X_i are independent, the random variables $g_{ij}(X_i)$ are also independent (for fixed j). It follows that

$$\mathbb{E}\Big[\big\|Y_j - \mathbb{E}\big[Y_j\big]\big\|^2\Big] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\Big[\big\|g_{ij}(X_i) - \mathbb{E}[g_{ij}(X_i)]\big\|^2\Big] = \frac{1}{N^2}\sum_{i=1}^N \sigma^2_{\mu_i}[g_{ij}].$$

By Assumption A, we have

$$f(Y) \leq f\big(\mathbb{E}[Y]\big) + \big\langle \nabla f(\mathbb{E}[Y]), Y - \mathbb{E}[Y] \big\rangle + \frac{1}{2} \sum_{j=1}^{M} \Big(\tilde{L}_j \big\| Y_j - \mathbb{E}\big[Y_j\big] \big\|^2 \Big).$$

Taking the expectation of the above inequality and recalling the definition of C_1, we deduce (2.3). ∎

Remark 2.7. Let us briefly comment on the interest of considering an additive structure for f. It is easy to verify that the mapping ∇f is Lipschitz continuous with modulus $\big(\max_{j=1,\dots,M} \tilde{L}_j\big)$; this estimate is tight. If we do not take into account the additive structure of f in the proof of Proposition 2.6, we end up with the following estimate:

$$\mathbb{E}\big[J(X)\big] \leq \mathcal{J}(\mu) + \frac{1}{2N^2} \Big(\max_{j=1,\dots,M} \tilde{L}_j \Big) \sum_{i=1}^{N} \sum_{j=1}^{M} \sigma^2_{\mu_i}\big[g_{ij}\big],$$

which is less precise than inequality (2.3). The same kind of comment could be made for the constants appearing afterwards in the convergence results of our numerical method.

2.3. Selection method. Suppose that a minimizer or an approximate minimizer μ of the randomized problem (PR) has been obtained. We address in this subsection the issue of recovering an approximate minimizer of the original problem (P) from μ.

A naive approach would consist in *averaging* the measures μ_i, assuming that the sets $\mathcal{X}_i$ are convex. In such a case, one can define the point $x_i = E_{\mu_i}[\mathrm{Id}]$. Another approach, motivated by Proposition 2.6, consists in sampling μ, that is, in simulating N independent random variables $(X_1, \dots, X_N)$, with distributions $X_i \sim \mu_i$. This can be done without additional structural assumption on the sets $\mathcal{X}_i$, moreover, Proposition 2.6 ensures that for any $\varepsilon > 0$,

$$\mathbb{P}\Big[J(X_1, \dots, X_N) < \mathcal{J}(\mu) + \frac{C_1}{2N} + \varepsilon \Big] > 0. \tag{2.4}$$

Of course, one can realize several samplings of μ to increase the probability of finding a good candidate for the original problem. We will refer to this approach as the *selection method.*

Example 2.8. Consider the following instance of the problem (P), where N is a large even number:

$$\begin{cases} \text{minimize} \Big\{ J(x_1, x_2, \dots, x_N) = -\frac{1}{N} \sum_{i=1}^{N} x_i^2 + \Big(\frac{1}{N} \sum_{i=1}^{N} x_i \Big)^2 \Big\}; \\ \text{subject to } x_i \in [-1, 1], \quad i = 1, \dots, N. \end{cases} \tag{2.5}$$

It is easy to see that x^* is a minimizer of (2.5) if and only if x^* has $N/2$ coordinates equal to 1 and the others equal to -1. In this example, the originial and the relaxed problem have the same value, $J^* = \mathcal{J}^* = -1$. The relaxed problem does not have a unique solution. One of them is $\tilde{\mu}_i = \frac{1}{2}\big(\delta_{-1} + \delta_1\big)$. Averaging $\tilde{\mu}$ as suggested above yields $\tilde{x} = (0, \dots, 0)$ and $J(\tilde{x}) = 0$. Thus in this example, the averaging method yields a poor candidate, whatever the value of N.

On the other hand, the selection method yields good candidates when N is large. Indeed, assume that $\mathbb{P}[X_i = -1] = \mathbb{P}[X_i = 1] = 1/2$. When N is large, by the law of large numbers [19], nearly half of the random variables X_i are equal to 1 while the others are equal to -1, with probability close to 1. Then in such a case X is almost a minimizer of (2.5).

The next theorem provides a sharp estimate of the probability in (2.4) and confirms the interest of the selection method for large values of N. It relies on a concentration inequality, *McDiarmid's inequality* [14], and its variant [7] (cf. Corollary A.2) of "variance type". It is quite intuitive that if the probability measures μ_i have a small variance (in a sense to be specified), then the selection method will be more efficient. The interest of taking into account the variances of the probability distributions will be revealed in the analysis of the stochastic Frank-Wolfe algorithm in Subsection 3.3.

THEOREM 2.9. *Let Assumption A be satisfied. Let $\mu \in \mathcal{P}_\delta$ and let $X_1, \ldots, X_N$ be N independent random variables such that $X_i \sim \mu_i$. Let $X = (X_1, \ldots, X_N)$. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\left[J(X) < \mathcal{J}(\mu) + \frac{C_1}{2N} + \epsilon\right] \geq 1 - \exp\left(-\frac{2N\epsilon^2}{C_0^2}\right). \tag{2.6}$$

Assume further that for all $i = 1, \ldots, N$, there exists a constant v_i such that

$$\sigma^2_{\mu_i}\left[J(\cdot, x_{-i})\right] \leq v_i^2, \tag{2.7}$$

for all $x_{-i} \in X_{-i}$. Then (2.6) can be strengthened as:

$$\mathbb{P}\left[J(X) < \mathcal{J}(\mu) + \sum_{j=1}^{M}\sum_{i=1}^{N} \frac{\tilde{L}_j}{2N^2}\sigma^2_{\mu_i}\left[g_{ij}\right] + \epsilon\right] \geq 1 - \exp\left(-\frac{N\epsilon^2}{2\left(\sum_{i=1}^{N} Nv_i^2 + \frac{C_0\epsilon}{3}\right)}\right). \tag{2.8}$$

Proof. Combining Lemma 2.3 and McDiarmid's inequality [14], we obtain

$$\mathbb{P}\Big[J(X) < \mathbb{E}\big[J(X)\big] + \epsilon\Big] \geq 1 - \exp\left(-\frac{2N\epsilon^2}{C_0^2}\right).$$

Combining this estimate with the second inequality of Proposition 2.6, we obtain (2.6).

Estimate (2.8) is proved similarly, combining McDiarmid's inequality of "variance type" proved in Corollary A.2 and the first inequality of Proposition 2.6. ∎

We provide in the next lemma an explicit candidate for (2.7).

LEMMA 2.10. *Inequality (2.7) is satisfied with $v_i^2 = \frac{2}{N^2}\big(\sum_{j=1}^M L_j^2\big)\sigma^2_{\mu_i}(g_i)$.*

Proof. We first state a general following property: given a probability measure μ and two maps h_1 and h_2 suitably defined, we have the inequality $\sigma^2_\mu\big[h_1 \circ h_2\big] \leq 2L^2\sigma^2_\mu\big[h_2\big]$, assuming that h_1 is L-Lipschitz continuous. The proof of this property is left to the reader. It is easy to verify that the function f is L-Lipschitz continuous, with $L = \big(\sum_{j=1}^M L_j^2\big)^{1/2}$. Using the announced property, we conclude that

$$\sigma^2_{\mu_i}\left[J(\cdot, x_{-i})\right] \leq 2L^2\sigma^2_{\mu_i}\left[\frac{1}{N}g_i(\cdot) + C\right] = \frac{2L^2}{N^2}\sigma^2_{\mu_i}\left[g_i\right],$$

where $C = \frac{1}{N}\sum_{i' \neq i} g_{i'}(x_{i'})$ is regarded as a constant. The estimate follows. ∎

3. Stochastic Frank-Wolfe algorithm.

3.1. Assumptions.

We introduce two new assumptions, which will be in force until the end of the article.

ASSUMPTION B. *For all $j = 1, \dots, M$, the function $f_j : \mathcal{E}_j \to \mathbb{R}$ is lower semi-continuous and convex.*

Let μ^1 and μ^2 lie in $\mathcal{P}_\delta$. Take $\omega \in [0,1]$. Let $\mu = (\mu_1, \dots, \mu_N)$ be defined, for any $i = 1, \dots, N$, by $\mu_i = (1-\omega)\mu_i^1 + \omega \mu_i^2$. Here, the addition and the multiplication by a scalar are understood as usual in the set of signed measures. In the sequel, we simply denote $\mu = (1-\omega)\mu^1 + \omega\mu^2$. We have $\mu \in \mathcal{P}_\delta$; moreover, $E_{\mu_i}[g_i] = (1-\omega)E_{\mu_i^1}[g_i] + \omega E_{\mu_i^2}[g_i]$, for any $i = 1, \dots, N$. Then, Assumption B implies that $\mathcal{J}(\mu) \leq (1-\omega)\mathcal{J}(\mu^1) + \omega \mathcal{J}(\mu^2)$. In words, the relaxed problem (PR) is convex.

In this section, we address the numerical resolution of the relaxed problem (and the original problem) under Assumption B. Let us mention that this convexity assumption is natural for the application problems described in the introduction. It allows the application of the Frank-Wolfe algorithm (also called conditional gradient algorithm) [8], for which convergence can be established. The Frank-Wolfe algorithm requires to solve at each iteration a subproblem. Here, the subproblems can be decomposed in N optimization problems, which can be solved in parallel. This property is particularly interesting, since we aim at solving instances of (P) with large values of N. We do not detail here the practical resolution of the subproblems, which can only be investigated case by case. Instead, we make the following assumption.

ASSUMPTION C. *For all $i = 1, \dots, N$, for all $y \in \operatorname{conv}(G(\mathcal{X}))$, the problem*

$$\inf_{x_i \in \mathcal{X}_i} \langle \nabla f(y), g_i(x_i) \rangle \tag{3.1}$$

has at least a solution. For all $i = 1, \dots, N$, we fix a map $\mathbb{S}_i : \operatorname{conv}(G(\mathcal{X})) \mapsto \mathcal{X}_i$ such that, for any $y \in \operatorname{conv}(G(\mathcal{X}))$, $\mathbb{S}_i(y)$ is a solution to (3.1).

The map $\mathbb{S}_i$ can be understood as a best-response function corresponding to agent i. The involved cost function is a linear combination of the contribution mappings g_{ij}, with $j = 1, \dots, M$.

Remark 3.1. It is easy to find assumptions which ensure the existence of the map $\mathbb{S}_i$. For example, one can assume that $\mathcal{X}_i$ is a compact set in a topological vector space and that g_i is continuous. Let us emphasize that Assumption C is essentially an assumption of numerical nature: $\mathbb{S}_i$ should be understood as the output of an (efficient) numerical procedure for the resolution of (3.1). The algorithms described afterwards largely rely on evaluations of $\mathbb{S}_i$.

3.2. Basic Frank-Wolfe algorithm.

We first describe a rather direct application of the Frank-Wolfe algorithm, which is referred to as the basic Frank-Wolfe algorithm. The starting point of our numerical approach is the following lemma, the proof of which is straightforward.

LEMMA 3.2. *Let $y \in \operatorname{conv}(G(\mathcal{X}))$ and let $\bar{\mu} = (\bar{\mu}_1, \dots, \bar{\mu}_N) \in \mathcal{P}_\delta$. Then, $\bar{\mu}$ is a solution to*

$$\inf_{\mu \in \mathcal{P}_\delta} \Big\langle \nabla f(y), \frac{1}{N} \sum_{i=1}^N E_{\mu_i}(g_i) \Big\rangle. \tag{3.2}$$

if and only if for all $i = 1, \dots, N$, $\bar{\mu}_i$ is supported in $\operatorname{argmin}_{x_i \in \mathcal{X}_i} \langle \nabla f(y), g_i(x_i) \rangle$.

The cost function in (3.2) should be regarded as a linearization of $\mathcal{J}$, as needed in the abstract formulation of the Frank-Wolfe algorithm in [8]. An immediate consequence of Lemma 3.2 is that $(\delta_{\mathbb{S}_1(y)}, \ldots, \delta_{\mathbb{S}_N(y)})$ is a solution to (3.2). The resolution of problem (3.2) is a key step in the numerical procedures developed afterwards; let us emphasize that the maps $\mathbb{S}_i(y)$ can be evaluated independently from each other, i.e. the resolution of (3.2) can be parallelized.

Algorithm 1 Frank-Wolfe Algorithm

Initialization: $\mu^0 \in \mathcal{P}_\delta$.
for $k = 0, 1, \ldots$ **do**
  Step 1: Resolution of the subproblems.
  Set $y^k = \frac{1}{N}\sum_{i=1}^N E_{\mu_i^k}[g_i]$.
  for $i = 1, \ldots, N$ **do**
    Compute $\bar{x}_i^k = \mathbb{S}_i(y^k)$.
  end for
  Set $\bar{\mu}^k = (\delta_{\bar{x}_1^k}, \ldots, \delta_{\bar{x}_N^k})$.
  Step 2: Update.
  Choose $\omega_k \in [0,1]$.
  Set $\mu^{k+1} = (1-\omega_k)\mu^k + \omega_k \bar{\mu}^k$.
end for

The convergence analysis performed afterwards relies on standard arguments (compare our proof with [11]). We introduce the primal gap γ_k and the primal-dual gap β_k, defined by

$$\gamma_k = \mathcal{J}(\mu^k) - \mathcal{J}^*, \qquad \beta_k = \langle \nabla f(y^k), y^k - \bar{y}^k \rangle, \qquad \text{where: } \bar{y}^k = \frac{1}{N}\sum_{i=1}^N g_i(\bar{x}_i^k). \tag{3.3}$$

Note that β_k can be evaluated numerically. The following lemma shows that β_k is an upper bound of the primal gap γ_k.

LEMMA 3.3. *For all $k \in \mathbb{N}$, $\gamma_k \leq \beta_k$.*

Proof. Let $k \in \mathbb{N}$. Let $\mu \in \mathcal{P}_\delta$ and let $y = \frac{1}{N}\sum_{i=1}^N E_{\mu_i}[g_i]$. By Lemma 3.2, we have $\langle \nabla f(y^k), \bar{y}^k \rangle \leq \langle \nabla f(y^k), y \rangle$. Thus, using the convexity of f, we obtain

$$\beta_k = \langle \nabla f(y^k), y^k - \bar{y}^k \rangle \geq \langle \nabla f(y^k), y^k - y \rangle \geq f(y^k) - f(y) = \mathcal{J}(\mu^k) - \mathcal{J}(\mu). \tag{3.4}$$

Since μ is arbitrary, we deduce that $\beta_k \geq \mathcal{J}(\mu^k) - \mathcal{J}^* = \gamma_k$. ◻

We have the following convergence result.

PROPOSITION 3.4. *Let Assumptions A, B, and C hold. Consider the choice $\omega_k = \bar{\omega}_k := \frac{2}{k+2}$, for all $k \in \mathbb{N}$, in Algorithm 1. Then,*

$$\gamma_k \leq \frac{2C_1}{k},$$

for any $k \in \mathbb{N}^*$.

Proof. As we will see, the result is a consequence of Lemma A.3, with $C = \frac{C_1}{2}$ and $u_k = 0$. By Assumption A,

$$f(y^{k+1}) \leq f(y^k) + \langle \nabla f(y^k), y^{k+1} - y^k \rangle + \sum_{j=1}^M \frac{\tilde{L}_j}{2} \|y_j^{k+1} - y_j^k\|^2.$$

We have $y^{k+1} - y^k = \omega_k(\bar{y}^k - y^k)$. Therefore, by definition of β_k,

$$f(y^{k+1}) \leq f(y^k) - \omega_k \beta_k + \omega_k^2 \sum_{j=1}^{M} \frac{\tilde{L}_j}{2} \|\bar{y}_j^k - y_j^k\|^2. \tag{3.5}$$

By definition, $\|\bar{y}_j^k - y_j^k\|^2 = \frac{1}{N^2} \big\| \sum_{i=1}^N E_{\mu_i^k}\big[g_{ij}(\bar{x}_i^k) - g_{ij}(\cdot)\big] \big\|^2$, thus by Cauchy-Schwarz inequality,

$$\|\bar{y}_j^k - y_j^k\|^2 \leq \frac{1}{N} \sum_{i=1}^{N} \big\| E_{\mu_i^k}\big[g_{ij}(\bar{x}_i^k) - g_{ij}(\cdot)\big] \big\|^2 \leq \frac{1}{N} \sum_{i=1}^{N} d_{ij}^2.$$

Combining the above estimate with (3.5) and using the inequality $\gamma_k \leq \beta_k$ proved in Lemma 3.3, we obtain that $\gamma_{k+1} \leq (1-\omega_k)\gamma_k + \frac{C_1}{2}\omega_k^2$. Thus Lemma A.3 applies, which concludes the proof. ∎

Remark 3.5. For any $k \in \mathbb{N}$, denote $h_k(\omega) = -\omega\beta_k + \frac{C_k}{2}\omega^2$, where the constant C_k is defined by $C_k = \sum_{j=1}^M \tilde{L}_j \|\bar{y}_j^k - y_j^k\|^2$. In view of inequality (3.5), the result of Proposition 3.4 remains true if the sequence $(\omega_k)_{k\in\mathbb{N}}$ is chosen such that for any $k \in \mathbb{N}$, $h(\omega_k) \leq h(\bar{\omega}_k)$. The result remains in particular true for

$$\omega_k = \underset{\omega\in[0,1]}{\operatorname{argmin}}\ h(\omega) = \min\Big(\frac{\beta_k}{C_k}, 1\Big). \tag{3.6}$$

The above proposition shows the convergence of the Frank-Wolfe algorithm. Yet the algorithm only provides a relaxed solution. In order to get a solution to the original problem, one can use the selection method introduced in Subsection 2.3. A first direct application of Proposition 2.6 yields the following. Let $(X_1, \ldots, X_N)$ be N independent random variables such that $X_i \sim \mu_i^k$, for all i. Then,

$$\mathbb{E}\big[J(X)\big] \leq J^* + \frac{2C_1}{k} + \frac{C_1}{2N}.$$

Therefore, from a theoretical point of view, there is no guaranty of improvements when $k \gg N$ since, then, the error term $\frac{2C_1}{k}$ becomes negligible in comparison with $\frac{C_1}{2N}$. The following lemma provides a convergence result (in probability) for the combination of the Frank-Wolfe algorithm and the selection method, for a number of iterations $k \leq N$.

Lemma 3.6. *Let $(\mu_k)_{k\in\mathbb{N}}$ be the output of Algorithm 1, for $\omega_k = \bar{\omega}_k$. Let $k \leq N$. Let $\zeta \in (0,1)$. Let $n \in \mathbb{N}^*$ and let $(X_i^j)_{i=1,\ldots,N}^{j=1,\ldots,n}$ be Nn independent random variables such that $X_i^j \sim \mu_i^k$. Let $X^j = (X_1^j, \ldots, X_N^j)$. Then,*

$$\mathbb{P}\left[\min_{j=1,\ldots,n} J(X^j) < \mathcal{J}^* + \frac{3C_1}{k}\right] \geq 1 - \zeta, \quad \text{if} \quad n \geq \frac{2C_0^2}{C_1^2}\,\frac{k^2}{N}\,\ln\Big(\frac{1}{\zeta}\Big). \tag{3.7}$$

Proof. Since $k \leq N$, we have $\frac{C_1}{2N} \leq \frac{C_1}{2k}$. Therefore, by Theorem 2.9,

$$\mathbb{P}\left[\min_{j=1,\ldots,n} J(X^j) < \mathcal{J}^* + \frac{2C_1}{k} + \frac{C_1}{2k} + \epsilon\right] \geq 1 - \exp\Big(-\frac{2N\epsilon^2 n}{C_0^2}\Big),$$

for any $\epsilon > 0$. Take $\epsilon = \frac{C_1}{2k}$. If n satisfies (3.7), then $\exp\big(-\frac{2N\epsilon^2 n}{C_0^2}\big) \leq \zeta$. ∎

3.3. Stochastic Frank-Wolfe algorithm. At each iteration of Algorithm 1, a new point $\bar{x}_i^k$ is added to the support of each distribution μ_i^k. Therefore, if at iteration K, the points $(\bar{x}_i^k)_{k=0,\dots,K-1}$ are distinct from each other (for each i), then KN places are needed to store the iterate μ^K, which can be prohibitive as K becomes large. We propose in this subsection a variant of Algorithm 1 which significantly mitigates the risk of memory overflow. We call it the *Stochastic Frank-Wolfe* (SFW) algorithm, it is given in Algorithm 2 below.

Algorithm 2 Stochastic Frank-Wolfe Algorithm

Initialization: $x^0 \in \mathcal{X}$
for $k = 0, 1, 2, \dots$ **do**
  Step 1: Resolution of the subproblems.
  Compute $y^k = \frac{1}{N}\sum_{i=1}^N g_i(x_i^k)$.
  for $i = 1, 2, \dots, N$ **do**
    Compute $\bar{x}_i^k = \mathbb{S}_i(y^k)$.
  end for
  Step 2: Update.
  Choose $n_k \in \mathbb{N}^*$ and $\omega_k \in [0,1]$.
  for $j = 1, 2, \dots, n_k$ **do**
    for $i = 1, 2, \dots, N$ **do**
      Simulate $\lambda_i^{k,j} \sim \text{Bern}(\omega_k)$, independently of all previously defined random variables.
      Set $\hat{x}_i^{k,j} = (1 - \lambda_i^{k,j})x_i^k + \lambda_i^{k,j}\bar{x}_i^k$.
    end for
    Define $\hat{x}^{k,j} = (\hat{x}_i^{k,j})_{i=1,\dots,N}$.
  end for
  Find $x^{k+1} \in \operatorname{argmin}\big\{J(x)\,\big|\, x \in \{\hat{x}^{k,j},\, j = 1, 2, \dots, n_k\}\big\}$.
end for

Starting from an initialization $x^0 \in \mathcal{X}$, Algorithm 2 generates a sequence $(x^k)_{k\in\mathbb{N}}$ in $\mathcal{X}$. Let us emphasize that there is no probability distribution involved in the practical implementation of Algorithm 2. However, for the analysis of the algorithm and for its description, it is convenient to introduce $\mu^k = (\delta_{x_1^k}, \dots, \delta_{x_N^k})$. With this notation at hand, we first observe that y^k, as defined in Step 1 of Algorithm 2, satisfies $y^k = \frac{1}{N}\sum_{i=1}^N E_{\mu_i^k}[g_i]$. Thus the Steps 1 of Algorithms 1 and 2 play exactly the same role. Let us focus next on Step 2 of Algorithm 2 and let us define $\bar{\mu}^k = (\delta_{\bar{x}_1^k}, \dots, \delta_{\bar{x}_N^k})$ and $\hat{\mu}^k = (1-\omega_k)\mu^k + \omega_k\bar{\mu}^k$. In contrast with Algorithm 1, we do not directly use $\hat{\mu}^k$ at the next iteration but instead employ our selection method so that $\hat{\mu}^k$ is reduced to an N-uplet of Dirac measures. The application of the selection method is here simple since $\hat{\mu}_i^k = (1-\omega_k)\delta_{x_i^k} + \omega_k\delta_{\bar{x}_i^k}$. Thus, to simulate a random variable with distribution $\hat{\mu}_i^k$, it suffices to simulate a random variable λ with Bernoulli distribution Bern(ω_k) and to consider $(1-\lambda)x_i^k + \lambda\bar{x}_i^k$. Using this method, Step 2 consists in simulating n_k random variables $(\hat{x}^{k,j})_{j=1,\dots,n_k}$ such that their probability distribution is equal to $\hat{\mu}^k$ (to be rigorous, their probability distribution conditionally to x^k). Finally, Step 2 selects a random variable $\hat{x}^{k,j}$ which minimizes J.

It is important to keep in mind that all variables involved in the algorithm (x^k, $\bar{x}^k$, $\hat{x}^{k,j}$) and all variables defined above (μ^k, $\bar{\mu}^k$, $\hat{\mu}^k$) are themselves random variables, since they depend on the Bernoulli random variables $\lambda_i^{k,j}$. For the analysis of the algorithm, we need to consider the filtration generated by the Bernoulli random variables. We introduce the set of indices $\mathcal{I}$ defined by

$$\mathcal{I} = \Big\{(k,j,i) \,|\, k \in \mathbb{N},\, j \in \{1,\dots n_k\},\, i \in \{1,\dots,N\}\Big\} \cup \big\{(0,0,0)\big\}.$$

We equip the set $\mathcal{I}$ with the lexicographic order: given (k_1,j_1,i_1) and (k_2,j_2,i_2) in $\mathcal{I}$, we write $(k_1,j_1,i_1) < (k_2,j_2,i_2)$ if and only if

$$[k_1 < k_2] \quad \text{or} \quad [k_1 = k_2 \text{ and } j_1 < j_2] \quad \text{or} \quad [(k_1,j_1) = (k_2,j_2) \text{ and } i_1 < i_2].$$

We further write $(k_1,j_1,i_1) \leq (k_2,j_2,i_2)$ if and only if $(k_1,j_1,i_1) < (k_2,j_2,i_2)$ or $(k_1,j_1,i_1) = (k_2,j_2,i_2)$. Note that this order coincides with the simulation order of the random variables $\lambda_i^{k,j}$ in the algorithm. The relation $\leq$ defines a total order with minimal element $(0,0,0)$. For any $(k,j,i) \neq (0,0,0)$, we denote by $(k,j,i)-1$ the maximal element of the set $\{(k',j',i') \in \mathcal{I} \,|\, (k',j',i') < (k,j,i)\}$. Finally, we consider the filtration $(\mathcal{G})_{(k,j,i)\in\mathcal{I}}$ defined by

$$\mathcal{G}_{(k,j,i)} = \begin{cases} \text{trivial } \sigma\text{−algebra}, & \text{if } (k,j,i) = (0,0,0), \\ \sigma\big(\mathcal{G}_{(k,j,i)-1}, \lambda_i^{k,j}\big), & \text{otherwise}, \end{cases}$$

where $\sigma\big(\mathcal{G}_{(k,j,i)-1}, \lambda_i^{k,j}\big)$ denotes the σ-algebra generated by $\mathcal{G}_{(k,j,i)-1}$ and $\lambda_i^{k,j}$. Note that $\hat{x}_i^{k,j}$ is $\mathcal{G}_{(k,j,i)}$-adapted and that x^k and $\bar{x}^k$ are $\mathcal{G}_{(k,1,1)-1}$-adapted.

THEOREM 3.7. *Let Assumptions A, B, and C hold true. Assume that* $\omega_k = \bar{\omega}_k = \frac{2}{k+2}$, *for all* $k \in \mathbb{N}$ *in Algorithm 2. Then, for all* $K = 1,\dots,2N$,

$$\mathbb{E}[\gamma_K] \leq \frac{4C_1}{K}, \quad \textit{where } \gamma_K = J(x^K) - \mathcal{J}^*.$$

Moreover, for all $\epsilon > 0$,

$$\mathbb{P}\Big[\gamma_K < \frac{4C_1}{K} + \epsilon\Big] \geq 1 - \exp\left(\frac{-\varepsilon^2 N}{2(v_K + \epsilon m_K/3)}\right), \tag{3.8}$$

where $v_K = \frac{2C_0^2}{K^2(K+1)^2}\left(\sum_{k=1}^{K-1}\frac{k(k+1)^2}{n_k}\right)$ *and* $m_K = \frac{C_0}{K(K+1)}\left(\max_{k=1,\dots,K-1}\frac{(k+1)(k+2)}{n_k}\right)$. *Finally, the following estimates quantify the variability of* γ_K:

$$\mathrm{Var}\big[\gamma_K\big] \leq \frac{16C_1^2}{K^2} + \frac{v_K}{N} \quad \textit{and} \quad \mathbb{E}\Big[\Big(\max\Big(\gamma_K - \frac{4C_1}{K}, 0\Big)\Big)^2\Big] \leq \frac{v_K}{N}. \tag{3.9}$$

The proof is postponed to Section 3.4. Let us note that the constants m_K and v_K involved in the theorem depend on the sequence $(n_k)_{k=0,1,\dots}$ but do not depend on N.

COROLLARY 3.8. *Let* $A > 0$, *assume that* $\omega_k = \bar{\omega}_k$, *and* $n_k \geq \max\big(\frac{Ak^2}{N}, 1\big)$. *Then, for all* $k = 1,\dots,2N$,

$$\mathbb{P}\Big[\gamma_K < \frac{4C_1 + C_0}{K}\Big] \geq 1 - \exp\left(-\frac{A}{12}\right).$$

Proof. Using $k+1 \leq 2k$, we obtain

$$\begin{aligned} v_K &\leq \frac{2C_0^2}{K^2(K+1)^2}\left(\sum_{k=1}^{K-1}\frac{Nk(k+1)^2}{Ak^2}\right) \leq \frac{8NC_0^2}{AK^2(K+1)^2}\left(\sum_{k=1}^{K-1} k\right) \\ &= \frac{4NC_0^2(K-1)K}{AK^2(K+1)^2} \leq \frac{4NC_0^2}{AK^2} \end{aligned}$$

and $m_K \leq \frac{C_0}{K(K+1)}\Big(\max_{k=1,\ldots,K-1}\frac{N(k+1)(k+2)}{Ak^2}\Big) \leq \frac{6NC_0}{AK^2}$. Applying Theorem 3.7 with $\epsilon = \frac{C_0}{K}$, we obtain that $\mathbb{P}\Big[\gamma_K < \frac{4C_1+C_0}{K}\Big] \geq 1-p$, with

$$p \leq \exp\left(\frac{-(C_0/K)^2 N}{2\Big(\frac{4NC_0^2}{AK^2}+\frac{6NC_0^2}{3AK^3}\Big)}\right) \leq \exp\left(\frac{-A}{12}\right),$$

as was to be proved. ∎

Remark 3.9. A variant of Algorithm 2 consists in setting $x^{k+1} = x^k$ if $J(\hat{x}^{k,j}) \geq J(x^k)$ for all $j = 1,\ldots,n_k$. Theorem 3.7 is still satisfied under this modification.

3.4. Proof of Theorem 3.7 and comments.

Step 1: proof of the convergence in expectation. We make use of the notations μ^k, $\bar{\mu}^k$, and $\hat{\mu}^k$, introduced right after Algorithm 2. We also introduce $\beta_k = \langle \nabla f(y^k), y^k - \bar{y}^k\rangle$, where $\bar{y}^k = \frac{1}{N}\sum_{i=1}^N g_i(\bar{x}_i^k)$. By construction, we have

$$J(x^{k+1}) = \min_{j=1,\ldots,n_k} J(\hat{x}^{k,j}) \leq \frac{1}{n_k}\sum_{j=1}^{n_k} J(\hat{x}^{k,j}).$$

Recalling that $\mathcal{J}(\mu^k) = J(x^k)$, we deduce that $\gamma_{k+1} \leq \gamma_k + a_k + b_k + c_k$, where

$$\begin{aligned} a_k &= \frac{1}{n_k}\sum_{j=1}^{n_k}\Big(J(\hat{x}^{k,j}) - \mathbb{E}\big[J(\hat{x}^{k,j})\,|\,\mathcal{G}_{(k,1,1)-1}\big]\Big), \\ b_k &= \frac{1}{n_k}\sum_{j=1}^{n_k}\Big(\mathbb{E}\big[J(\hat{x}^{k,j})\,|\,\mathcal{G}_{(k,1,1)-1}\big] - \mathcal{J}(\hat{\mu}^k)\Big), \\ c_k &= \mathcal{J}(\hat{\mu}^k) - \mathcal{J}(\mu^k) = \mathcal{J}(\hat{\mu}^k) - J(x^k). \end{aligned}$$

The term a_k does not play a significant role at the moment since its expectation is null. The term b_k must be understood as a relaxation cost, induced by the use of the selection method. The term c_k is estimated exactly as in Proposition 3.4: as was seen in its proof, we have $c_k \leq -\omega_k\beta_k + \omega_k^2\frac{C_1}{2}$. A direct adaptation of Proposition 2.6 shows that

$$b_k \leq \frac{1}{2N^2}\sum_{j=1}^M\sum_{i=1}^N \tilde{L}_j \sigma^2_{\hat{\mu}_i^k}[g_{ij}] \leq \frac{1}{2N^2}\sum_{j=1}^M\sum_{i=1}^N \tilde{L}_j\omega_k(1-\omega_k)d_{ij}^2 = \omega_k(1-\omega_k)\frac{C_1}{2N}.$$

Combining the above estimates, we obtain

$$\gamma_{k+1} \leq \gamma_k + a_k + \Big(-\omega_k\beta_k + \omega_k^2\frac{C_1}{2}\Big) + \omega_k(1-\omega_k)\frac{C_1}{2N}. \tag{3.10}$$

For the choice $\omega_k = \bar{\omega}_k$, we have $(1-\omega_k)/N = k/(N(k+2)) \leq \omega_k$, since $k \leq 2N$. It follows that

$$\omega_k(1-\omega_k)\frac{C_1}{2N} \leq \omega_k^2 \frac{C_1}{2}$$

and finally, since $\gamma_k \leq \beta_k$, we have $\gamma_{k+1} \leq (1-\omega_k)\gamma_k + \omega_k^2 C_1 + a_k$. Next by Lemma A.3,

$$\gamma_K \leq \frac{4C_1}{K} + S_K, \quad \text{where: } S_K = \sum_{k=0}^{K-1} \frac{(k+1)(k+2)}{K(K+1)} a_k. \tag{3.11}$$

We have $\mathbb{E}[a_k] = 0$, thus $\mathbb{E}[S_K] = 0$ and finally $\mathbb{E}\big[\gamma_K\big] \leq \frac{4C_1}{K}$.

Step 2: proof of the probability and variance estimates. We next need to find an estimate of $\mathbb{P}[S_K \geq \epsilon]$. For this purpose, we need to further decompose the term a_k as a sum of random variables. A first observation is the following equality: $\mathbb{E}\big[J(\hat{x}^{k,j}) \mid \mathcal{G}_{(k,1,1)-1}\big] = \mathbb{E}\big[J(\hat{x}^{k,j}) \mid \mathcal{G}_{(k,j,1)-1}\big]$, which easily follows from Lemma A.5. As a consequence,

$$J(\hat{x}^{k,j}) - \mathbb{E}\big[J(\hat{x}^{k,j}) \mid \mathcal{G}_{(k,1,1)-1}\big] = \sum_{i=1}^{N} U_{(k,j,i)},$$

where

$$U_{(k,j,i)} = \mathbb{E}\big[J(\hat{x}^{k,j}) \mid \mathcal{G}_{(k,j,i)}\big] - \mathbb{E}\big[J(\hat{x}^{k,j}) \mid \mathcal{G}_{(k,j,i)-1}\big].$$

We obtain the following decomposition of S_K:

$$S_K = \sum_{k=1}^{K-1} \sum_{j=1}^{n_k} \sum_{i=1}^{N} \frac{(k+1)(k+2)}{n_k K(K+1)} U_{(k,j,i)}.$$

Note that the index k starts at 1. Indeed, $\bar{\omega}_0 = 1$, thus $\hat{x}^{0,j} = \bar{x}^0$ and then $a_0 = 0$. Let us apply Proposition A.1 to S_K. We have $\mathbb{E}\big[U_{(k,j,i)} \mid \mathcal{G}_{(k,j,i)-1}\big] = 0$. Viewing the term $J(\hat{x}^{k,j})$ as a function F of the random variables $A := (\lambda_{i'}^{k',j'})_{(k',j',i')<(k,j,i)}$, $B := \lambda_i^{k,j}$, and $C := (\lambda_{i'}^{k',j'})_{(k,j,i)<(k',j',i')}$, we can apply Lemma A.4 to $U_{(k,j,i)}$, with $\delta = C_0/N$ (by Lemma 2.3). This yields

$$U_{(k,j,i)} \leq \frac{C_0}{N} \qquad \text{and} \qquad \mathbb{E}\big[U_{(k,j,i)}^2 \mid \mathcal{G}_{(k,j,i)-1}\big] \leq \frac{\omega_k(1-\omega_k)C_0^2}{N^2}.$$

Therefore, Proposition A.1 applies to $\mathbb{P}[S_K \geq \epsilon]$, where the constants m and v are given by

$$m = \max_{k=1,\ldots,K-1} \frac{(k+1)(k+2)}{n_k K(K+1)} \frac{C_0}{N} = \frac{m_K}{N},$$

$$v = \sum_{k=1}^{K-1} \sum_{j=1}^{n_k} \sum_{i=1}^{N} \Big(\frac{(k+1)(k+2)}{n_k K(K+1)}\Big)^2 \frac{2kC_0^2}{(k+2)^2 N^2} = \frac{v_K}{N}.$$

This proves estimate (3.8). Recalling that $\gamma_K \leq \frac{4C_1}{K} + S_K$ a.s., we obtain

$$\mathrm{Var}\big[\gamma_K\big] \leq \mathbb{E}\big[\gamma_K^2\big] \leq \mathbb{E}\Big[\Big(\frac{4C_1}{K} + S_K\Big)^2\Big] = \frac{16C_1^2}{K^2} + \mathbb{E}\big[S_K^2\big].$$

Next by Proposition A.1, $\mathbb{E}\big[S_K^2\big] \leq v_K/N$. The first inequality in (3.9) follows. The second inequality follows from the inequality: $\max\big(\gamma_K - \frac{4C_1}{K}, 0\big)^2 \leq S_K^2$.

Remark 3.10. Let us set $h_k(\omega) = -\omega\beta_k + \omega^2\frac{C_1}{2} + \omega(1-\omega)\frac{C_1}{2N}$. If for all $k \in \mathbb{N}$, we have $h_k(\omega_k) \leq h_k(\bar{\omega}_k)$, then the convergence in expectation of Theorem 3.7 still holds, i.e. $\mathbb{E}\big[\gamma_K\big] \leq 4C_1/K$, in view of inequality (3.10). In particular, one can take

$$\omega_k = \operatorname*{argmin}_{\omega\in[0,1]} h_k(\omega) = \max\left(\min\left(\frac{\beta_k - C_1/2N}{C_1(1-1/N)}, 1\right), 0\right). \tag{3.12}$$

3.5. A speed-up of the SFW algorithm. Step 1 of Algorithm 2 requires to solve N independent subproblems. It turns out that only a subset of those subproblems need to be solved for the implementation of Step 2. At iteration k consider the following set:

$$I_k = \bigcup_{j=1,2,\ldots,n_k} \Big\{ i \in \{1,\ldots,N\} \,|\, \lambda_i^{k,j} = 1 \Big\}.$$

If $i \notin I_k$, then $\hat{x}_i^{k,j} = x_i^k$, in other words, for such an index i, it is not necessary to evaluate $\mathbb{S}_i(y^k)$. A speed-up of the SFW algorithm can therefore be obtained by simulating the Bernoulli random variables before Step 1, next by evaluating $\mathbb{S}_i(y^k)$ only for the indices i in I_k, and finally by computing $\hat{x}^{k,j}$ and x^{k+1} as before. The expectation of the number of subproblems to be solved at iteration k is given by

$$\begin{aligned}\mathbb{E}\big[|I_k|\big] &= \sum_{i=1}^N \mathbb{P}\big[i \in I_k\big] = N\big(1 - \mathbb{P}\big[1 \notin I_k\big]\big) = N\Big(1 - \mathbb{P}\big[\lambda_1^{k,j} = 0,\ \forall j = 1,\ldots,n_k\big]\Big) \\ &= N\Big(1 - \Big(\frac{k}{k+2}\Big)^{n_k}\Big).\end{aligned}$$

3.6. Stopping time strategy. In Algorithm 2, the number of samplings n_k is chosen at the beginning of Step 2. We consider here a variant: we generate a sequence of random variables $\hat{x}^{k,j}$ with probability distribution equal to $\hat{\mu}_k$ (conditionally to $\mathcal{G}_{(k,1,1)-1}$); the variables are constructed via Bernoulli variables independent from each other. We define n_k as the first index j such that

$$J(\hat{x}^{k,j}) \leq \mathcal{J}(\hat{\mu}^k) + \Big(\frac{C_1}{2} + C_0\Big)\omega_k^2. \tag{3.13}$$

The next iterate is defined by $x^{k+1} = \hat{x}^{k,n_k}$. Note that $\mathcal{J}(\hat{\mu}^k) = J\big((1-\omega_k)x^k + \omega_k\bar{x}^k\big)$, thus the right-hand side of (3.13) is easy to evaluate numerically. We consider here the stepsize rule $\omega_k = \bar{\omega}_k$.

LEMMA 3.11. *Let $(x^k)_{k\in\mathbb{N}}$ denote the sequence obtained with the stopping rule (3.13) and $\omega_k = \bar{\omega}_k$. Then*

$$J(x^{k+1}) - \mathcal{J}^* \leq \frac{4(C_1 + C_0)}{k}, \quad \forall k = 1,\ldots 2N, \quad a.s.$$

Moreover,

$$\mathbb{E}\big[n_k\big] \leq \Big(1 - \exp\Big(-\frac{4N}{(k+2)^3}\Big)\Big)^{-2}, \quad \forall k = 1, \dots, 2N.$$

Proof. Let $\hat{x}$ be a random variable with probability distribution equal to $\hat{\mu}^k$, conditionally to $\mathcal{G}_{(k,1,1)-1}$. Then, for all $\epsilon > 0$, estimate (2.8) of Theorem 2.9 yields:

$$\mathbb{P}\Big[J(\hat{x}) \geq \mathcal{J}(\hat{\mu}^k) + \frac{C_1}{2N}\omega_k(1-\omega_k) + \epsilon \,\Big|\, \mathcal{G}_{(k,1,1)-1}\Big] \leq p_\epsilon \tag{3.14}$$

where $p_\epsilon = \exp\Big(\frac{-N\epsilon^2}{2(\omega_k(1-\omega_k)C_0^2 + \frac{C_0}{3}\epsilon)}\Big)$. For $\epsilon = C_0\omega_k^2$, we have

$$p_\epsilon = \exp\left(\frac{-NC_0^2\omega_k^4}{2\big(\omega_k C_0^2 - \frac{2}{3}\omega_k^2 C_0^2\big)}\right) \leq p := \exp\left(\frac{-N\omega_k^3}{2}\right) = \exp\left(\frac{-4N}{(k+2)^3}\right).$$

Recalling that $\frac{C_1}{2N}\omega_k(1-\omega_k) \leq \frac{C_1}{2}\omega_k^2$, we deduce that

$$\mathbb{P}\Big[J(\hat{x}) \geq \mathcal{J}(\hat{\mu}^k) + \Big(\frac{C_1}{2} + C_0\Big)\omega_k^2 \,\Big|\, \mathcal{G}_{(k,1,1)-1}\Big] \leq p.$$

Now, let us consider a sequence of independent random variables $(\hat{x}^{k,j})_{j=1,\dots}$ (conditionally to $\mathcal{G}_{(k,1,1)-1}$), with conditional probability distribution $\hat{\mu}^k$. By estimate (3.14),

$$\mathbb{P}\big[n_k = j\big] \leq \mathbb{P}\Big[J(\hat{x}^{k,j'}) \geq \mathcal{J}(\hat{\mu}^k) + \Big(\frac{C_1}{2} + C_0\Big)\omega_k^2,\ \forall j' \,\Big|\, \mathcal{G}_{(k,1,1)-1}\Big] \leq p^{j-1}.$$

We finally deduce that $\mathbb{E}\big[n_k\big] \leq \sum_{n=1}^\infty jp^{j-1} = \frac{1}{(1-p)^2}$, which proves the second part of the lemma. For the first part of the lemma, it suffices to observe that

$$J(x^{k+1}) \leq \mathcal{J}(\hat{\mu}^k) + \Big(\frac{C_1}{2} + C_0\Big)\omega_k^2 \leq J(x^k) - \beta_k\omega_k + (C_1 + C_0)\omega_k^2,$$

and to conclude with Lemma A.3. ∎

4. Refined gap estimates.

4.1. Nonconvexity measure and gap estimate. We give in this subsection a refinement of the randomization gap obtained in Proposition 2.6. Our analysis relies on the concept of nonconvexity measure, introduced in [5].

Definition 4.1. *Given a subset $\mathcal{K}$ of $\mathcal{E}$, we call nonconvexity measure of $\mathcal{K}$ the number $\rho(\mathcal{K})$ defined by*

$$\rho(\mathcal{K}) = \left(\sup_{y \in \mathrm{conv}(\mathcal{K})} \inf_{\substack{\mu \in \mathcal{P}_\delta,\\ E_\mu[\mathrm{Id}] = y}} \sigma_\mu\big[\mathrm{Id}\big]^2\right)^{1/2},$$

where $\mathrm{Id}\colon \mathcal{E} \to \mathcal{E}$ *denotes the identity mapping.*

The terminology nonconvexity measure is motivated by the following: if $\mathcal{K}$ is convex, then obviously $\rho(\mathcal{K}) = 0$ and conversely, if $\rho(\mathcal{K}) = 0$, then $\mathcal{K}$ is dense into $\mathrm{conv}(\mathcal{K})$. We have the following two properties, easily verified. The map ρ is homogeneous in the following sense: given $a \in \mathbb{R}$, we have $\rho(a\mathcal{K}) = |a|\rho(\mathcal{K})$. Moreover

$\rho(\mathcal{K}) \leq d(\mathcal{K})$, where $d(\mathcal{K})$ is the diameter of $\mathcal{K}$. Another particularly interesting property for our aggregative problem is the sub-additivity of $\rho(\cdot)^2$: given two subsets $\mathcal{K}_1$ and $\mathcal{K}_2$, we have $\rho(\mathcal{K}_1 + \mathcal{K}_2)^2 \leq \rho(\mathcal{K}_1)^2 + \rho(\mathcal{K}_2)^2$, see [5, Theorem 1]. We will use an improvement of this inequality in the proof of Theorem 4.4, based on the Shapley-Folkman theorem.

The next lemma provides a general relaxation estimate based on the nonconvexity measure of the feasible set. Let us emphasize that the central idea behind this result is the same as the one in the proof of Proposition 2.6. The only difference is the point of view, which is here geometric while it was previously probabilistic.

Lemma 4.2. *Let $\mathcal{K}$ be a subset of $\mathcal{E}$. Let F be a differentiable real-valued function defined on some neighborhood of* $\operatorname{conv}(\mathcal{K})$. *Assume that ∇F is $\tilde{L}$-Lipschitz continuous over* $\operatorname{conv}(\mathcal{K})$. *Then,*

$$\inf_{y \in \mathcal{K}} F(y) \leq \Big(\inf_{y \in \operatorname{conv}(\mathcal{K})} F(y) \Big) + \frac{\tilde{L}}{2} \rho(\mathcal{K})^2.$$

Proof. Let $y \in \operatorname{conv}(\mathcal{K})$. Let $\mu \in \mathcal{P}_\delta(\mathcal{K})$ be such that $E_\mu[\mathrm{Id}] = y$. Then, since ∇F is $\tilde{L}$-Lipschitz continuous, we have

$$\inf_{y' \in \mathcal{K}} F(y') \leq E_\mu[F] \leq F(y) + \frac{\tilde{L}}{2} \sigma_\mu^2[\mathrm{Id}].$$

Minimizing the right-hand side with respect to μ, we obtain that

$$\inf_{y' \in \mathcal{K}} F(y') \leq F(y) + \frac{\tilde{L}}{2} \rho(\mathcal{K})^2.$$

Minimizing the result with respect to y yields the announced estimate. □

Some notations are needed for the application of Lemma 4.2 to (P). We set

$$\tilde{g}_{ij}(x_i) = \sqrt{\tilde{L}_j}\, g_{ij}(x_i), \qquad \tilde{g}_i(x_i) = \big(\tilde{g}_{ij}(x_i)\big)_{j=1,\dots,M}$$

$$\tilde{f}_j(y_j) = f_j\Big(\frac{y_j}{\sqrt{\tilde{L}_j}}\Big), \qquad \tilde{f}(y) = \sum_{j=1}^M \tilde{f}_j(y_j).$$

Obviously, $J(x) = \tilde{f}\big(\frac{1}{N} \sum_{i=1}^N \tilde{g}_i(x_i)\big) = \sum_{j=1}^M \tilde{f}_j\big(\frac{1}{N}\sum_{i=1}^N \tilde{g}_{ij}(x_i)\big)$. Finally we denote

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

We give next two new formulations of problems (P) and (PR), revealing the geometric nature of the relaxation technique employed so far.

Lemma 4.3. *We have*

(PG) $$J^* = \inf_{y \in \mathcal{Y}} \tilde{f}(y),$$

(PGR) $$\mathcal{J}^* = \inf_{y \in \operatorname{conv}(\mathcal{Y})} \tilde{f}(y).$$

Proof. The first equality is straightforward. For the second one, it suffices to observe that $\operatorname{conv}(\mathcal{Y}) = \frac{1}{N}\sum_{i=1}^N \operatorname{conv}(\mathcal{Y}_i)$ and that $\operatorname{conv}(\mathcal{Y}_i) = \big\{ E_{\mu_i}[\tilde{g}_i] \mid \mu_i \in \mathcal{P}_\delta(\mathcal{X}_i) \big\}$. □

We introduce the following constants:

$$D_i = \sum_{j=1}^{M} \tilde{L}_j d_{ij}^2, \qquad D[k] = \max_{\substack{K \subseteq \{1,\dots,N\} \\ |K| = k}} \sum_{i \in K} D_i.$$

THEOREM 4.4. *Let Assumption A hold true. It holds:*

$$J^* - \mathcal{J}^* \leq \frac{1}{2N^2} \left(\max_{\substack{Q \subseteq \{1,\dots,N\} \\ |Q| = q \wedge N}} \sum_{i \in Q} \rho(\mathcal{Y}_i)^2 \right) \leq \frac{D[q \wedge N]}{2N^2}. \tag{4.1}$$

Note that $D[N] = NC_1$, thus the new gap estimate is the same as the one obtained in Proposition 2.6 when $q \geq N$ and it is strictly better when $q < N$.

Proof of Theorem 4.4. We let the reader verify that $\nabla \tilde{f}$ is 1-Lipschitz. Then Lemma 4.3 and the homogeneity of ρ yield

$$J^* - \mathcal{J}^* \leq \frac{1}{2} \rho(\mathcal{Y})^2 \leq \frac{1}{2N^2} \rho\left(\sum_{i=1}^{N} \mathcal{Y}_i \right)^2.$$

Applying [5, Theorem 2], we obtain that

$$\rho\left(\sum_{i=1}^{N} \mathcal{Y}_i \right)^2 \leq \max_{\substack{Q \subseteq \{1,\dots,N\} \\ |Q| = q \wedge N}} \sum_{i \in Q} \rho(\mathcal{Y}_i)^2,$$

which proves the first inequality. Observing that

$$\rho(\mathcal{Y}_i)^2 \leq d(\mathcal{Y}_i)^2 \leq \sum_{j=1}^{M} d\big(\tilde{g}_{ij}(\mathcal{X}_i)\big)^2 = \sum_{j=1}^{M} \tilde{L}_j d\big(g_{ij}(\mathcal{X}_i)\big)^2 = D_i,$$

we obtain the second inequality. ∎

4.2. Duality and price of decentralization. In this subsection we introduce a dual problem (we work again under Assumption B) and investigate its connection with the geometric relaxed problem (PGR). This allows us to obtain a last refinement of the randomization gap. For all $i = 1, \dots, N$ and for all $\lambda \in \mathcal{E}$, we introduce

$$\Phi_i(\lambda) = \inf_{x_i \in \mathcal{X}_i} \langle \lambda, \tilde{g}_i(x_i) \rangle, \qquad \mathcal{Y}_i(\lambda) = \operatorname*{argmin}_{y_i \in \mathcal{Y}_i} \langle \lambda, y_i \rangle, \qquad \mathcal{X}_i(\lambda) = \operatorname*{argmin}_{x_i \in \mathcal{X}_i} \langle \lambda, \tilde{g}_i(x_i) \rangle.$$

We refer to the following problem as the dual problem:

$$\sup_{\lambda \in \mathcal{E}} \Big(- \tilde{f}^*(\lambda) + \frac{1}{N} \sum_{i=1}^{N} \Phi_i(\lambda) \Big). \tag{D}$$

Let $\mathcal{D}^*$ denote the value of Problem (D).

ASSUMPTION D. *The set* conv$(\mathcal{Y})$ *is closed.*

Remark 4.5. Assume that $\mathcal{E}$ is finite-dimensional. If the sets $\mathcal{X}_i$ are compact and the maps $\tilde{g}_i$ continuous, then the sets $\mathcal{Y}_i = \tilde{g}_i(\mathcal{X}_i)$ are also compact. It is then easy to verify with Carathéodory's theorem that conv$(\mathcal{Y}_i)$ is also compact, thus closed, which finally implies Assumption D.

LEMMA 4.6. *The problem* (PGR) *has a solution.*

Proof. This is a direct application of [2, Theorem 11.9]. ∎

The next lemma provides a duality result and a characterization of optimal solutions for problem (PGR).

LEMMA 4.7. *Let Assumptions A, B, C, and D hold true. Then,* $\mathcal{J}^* = \mathcal{D}^*$ *and the dual problem* (D) *has at least one solution. Fix a solution* λ *to Problem* (D)*. Let* $y \in \mathcal{E}$*. Then,* y *is a solution to* (PGR) *if and only if* $y \in \partial \tilde{f}^*(\lambda)$ *and* $y \in \frac{1}{N}\sum_{i=1}^N \operatorname{conv}\big(\mathcal{Y}_i(\lambda)\big)$.

Proof. Let h denote the indicatrix function of $\operatorname{conv}(\mathcal{Y})$. By Assumption A, the domain of $\tilde{f}$ contains a neighborhood of $\operatorname{conv}(\mathcal{Y})$. By Assumption D, h is lower semi-continuous. Therefore, the Fenchel-Rockafellar theorem [16] applies and yields

$$\mathcal{J}^* = \inf_{y\in\mathcal{E}} \Big(f(y) + h(y)\Big) = \sup_{\lambda\in\mathcal{E}} \Big(-\tilde{f}^*(\lambda) - h^*(-\lambda)\Big).$$

Moreover, the supremum in the right-hand side is a maximum. We have

$$-h^*(-\lambda) = \inf_{y\in\operatorname{conv}(\mathcal{Y})} \langle \lambda, y\rangle = \inf_{y\in\mathcal{Y}} \langle \lambda, y\rangle = \frac{1}{N}\sum_{i=1}^N \Phi_i(\lambda).$$

As a consequence, $\mathcal{J}^* = \mathcal{D}^*$ and problem (D) has at least one solution.

Now let us fix a solution λ to the dual problem (D). Let $y \in \mathcal{E}$. Then y is a solution if and only if (i) $\tilde{f}(y) + \tilde{f}^*(\lambda) = \langle \lambda, y\rangle$ and (ii) $h(y) + h^*(-\lambda) = -\langle \lambda, y\rangle$. The condition (i) is equivalent to $y \in \partial \tilde{f}(\lambda)$. The condition (ii) is equivalent to

$$y \in \operatorname{conv}(\mathcal{Y}) \text{ and } \langle \lambda, y\rangle = -h^*(-\lambda) = \inf_{y'\in\mathcal{Y}} \langle \lambda, y'\rangle.$$

Thus (ii) $\Longleftrightarrow y \in Y$, where $Y = \operatorname*{argmin}_{y'\in\operatorname{conv}(\mathcal{Y})} \langle \lambda, y'\rangle$. We further have

$$Y = \operatorname{conv}\left(\operatorname*{argmin}_{y'\in\mathcal{Y}} \langle \lambda, y'\rangle\right) = \operatorname{conv}\Big(\frac{1}{N}\sum_{i=1}^N \mathcal{Y}_i(\lambda)\Big) = \frac{1}{N}\sum_{i=1}^N \operatorname{conv}\big(\mathcal{Y}_i(\lambda)\big),$$

which concludes the proof. ∎

Remark 4.8. If $\tilde{f}$ is differentiable on $\mathcal{E}$, with a Lipschitz-continous gradient, then $\tilde{f}^*$ is strongly convex (see [2, Theorem 18.15]), which implies that (D) has a unique solution.

Let us fix a solution λ to the dual problem until the end of the subsection. Let us consider

$$J_{\text{dec}} = \inf_{x\in\mathcal{X}} J(x), \quad \text{subject to: } x_i \in \mathcal{X}_i(\lambda), \quad \forall i = 1,\dots,N.$$

In words, we restrict $\mathcal{X}_i$ to the best-responses corresponding to the dual variable λ. Following the terminology of [21], we call price of decentralization the real number $p = J_{\text{dec}} - J^*$.

PROPOSITION 4.9. *Let Assumptions A, B, C, and D hold true. It holds:*

$$p \le J_{\text{dec}} - \mathcal{J}^* \le \frac{1}{2N^2}\left(\max_{\substack{Q\subseteq\{1,\dots,N\}\\ |Q| = q\wedge N}} \sum_{i\in Q} \rho\big(\mathcal{Y}_i(\lambda)\big)^2 \right).$$

Proof. The definition of J_{dec} and Lemma 4.7 respectively yield:

$$J_{\text{dec}} = \inf_{y \in \frac{1}{N}\sum_{i=1}^N \mathcal{Y}_i(\lambda)} \tilde{f}(y) \qquad \text{and} \qquad \mathcal{J}^* = \inf_{y \in \frac{1}{N}\sum_{i=1}^N \operatorname{conv}(\mathcal{Y}_i(\lambda))} \tilde{f}(y).$$

The announced estimate follows then from Lemma 4.2 and [5, Theorem 2], as in the proof of Theorem 4.4. ∎

Remark 4.10. The randomization gap is bounded from above by $J_{\text{dec}} - \mathcal{J}^*$. Moreover, one can show that $\rho(\mathcal{Y}_i(\lambda)) \leq \rho(\mathcal{Y}_i)$. Thus Proposition 4.9 provides a last refinement of the gap estimate (4.1).

4.3. Literature comparison. Let us compare our results and our method with the work of Wang [21]. Our gap estimate, as well as our estimate of the price of decentralization, are of order $\mathcal{O}(\min(q,N)/N^2)$, while the estimates obtained by applying [21, Theorem 3.5] are of order $\mathcal{O}(q^2/N^2)$. We emphasize that our first gap estimate, of order $\mathcal{O}(1/N)$, already improves [21] when $q \gg \sqrt{N}$. Note that the geometric relaxation employed in Section 4.1 is the same as the one used in [21].

Let us compare our algorithmic approaches. At a general level, one can observe that we have a primal approach, while Wang solves the dual problem to the relaxed problem. Our approach is restricted to the case where f is differentiable, while the dual approach allows to tackle the case of hard constraints (for example when f is the indicator function of some convex set). Both approaches leverage the decomposability of the problem into N problems and require that the subproblems can be easily solved. Let us emphasize however that we only need to be able to compute a single solution for those problems, while [21, Algorithm 2] requires to compute the full set of ξ-optimal solutions, which may be much more difficult. Our algorithm does not require to perform Shapley-Folkman decompositions, contrary to [21]. This is a major advantage when the dimension of the aggregate q is very large. Also, we do not need to evaluate f^*. As a counterpart, we are only able to find $\mathcal{O}(1/N)$-optimal solutions, while the algorithm of [21] can find $\mathcal{O}(q^2/N^2)$-optimal solutions. The design of a method for the computation of $\mathcal{O}(q \wedge N/N^2)$-solutions will be the topic of future research.

5. Numerical test. In this section we provide numerical results for a mixed-integer linear quadratic problem of the form (P).

Let A be a real $M \times N$ matrix and $\bar{y} \in \mathbb{R}^M$. Consider the following problem:

$$\text{(MIQP)} \qquad \min_{x \in \{0,1\}^N} J(x) := \frac{1}{N^2}\|Ax - \bar{y}\|^2_{\mathbb{R}^M} = \sum_{j=1}^M \left(\frac{1}{N}\sum_{i=1}^N A_{ji}x_i - \frac{\bar{y}_j}{N} \right)^2.$$

Problem (MIQP) has the form (P), with $f_j(y_j) = \left(y_j - \frac{\bar{y}_j}{N}\right)^2$ for $1 \leq j \leq M$, and $g_{ij}(x_i) = A_{ji}x_i$ for $1 \leq i \leq N$, $1 \leq j \leq M$. Moreover, Assumption A is satisfied with $\tilde{L}_j = 2$ and $d_{ij} = |A_{ji}|$. Thus $C_1 = \frac{2}{N}\sum_{i=1}^N \sum_{j=1}^M |A_{ji}|$. Due to the linearity of g_{ij}, the randomized problem coincides with the minimization problem of J on $[0,1]^N$, which is a convex linear-quadratic program that can be solved with independent methods; thus it is easy here to obtain a precise estimate of $\mathcal{J}^*$.

In the numerical simulation, we set $M = N = 100$, and draw the parameters A_{ji} according to the uniform distribution on the interval $[0,1]$ while y_j is drawn according to the uniform distribution on $[0, N/2]$. Thus $C_1 \approx M$ and the gap estimate is given by $\frac{C_1}{2N} \approx 0.5$. Figure 1 shows the outcome of the basic Frank-Wolfe algorithm 1 with 200 iterations. The left sub-figure shows the evolution of γ_k for $\omega_k = 2/(k+2)$

(green curve) and for ω_k determined by line search (3.6) (red curve). A sub-linear rate of convergence is observed (note that logarithmic scales are employed for both axes). The right sub-figure represents the evolution of $J(X^k) - \mathcal{J}^*$, where X^k is a random variable with distribution μ^k. For both choices of ω_k, approximate solutions to the problems are simulated, with a gap smaller than 10^{-3}, significantly smaller than the gap estimate $\frac{C_1}{2N}$. The line search approach is quicker than the approach with $\omega_k = \frac{2}{k+2}$.

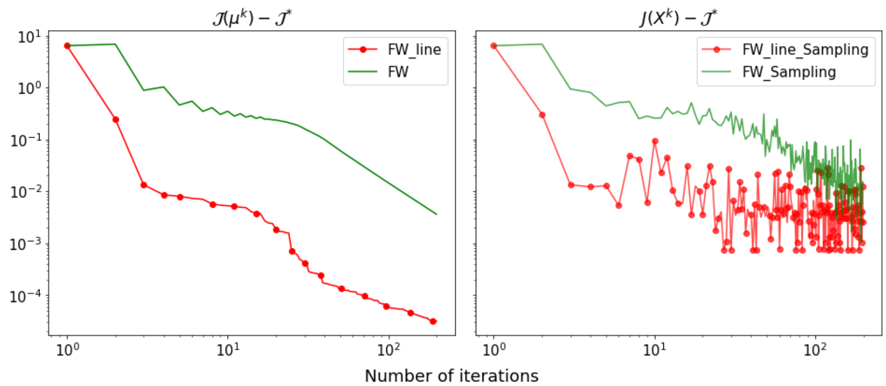

Fig. 1: MIQP by Algorithm 1, 200 iterations, with $\omega_k = 2/(k+2)$ and line search (3.6).

Figure 2 shows the outcome of Algorithm 2 (with the modification suggested in Remark 3.9), for different (constant) choices of n_k with 200 iterations, for two different stepsize rules ($\omega_k = 2/(k+2)$ on the left, line search on the right). Since the algorithm is stochastic, we have tested it 50 times to evaluate its efficiency; the curves represent the average value of γ_k. The standard deviation (for these 8 instances of the SFW method) is displayed on Figure 3. In all cases, an average value of the gap significantly smaller than $\frac{C_1}{2N}$ can be reached; the standard deviation is also significantly smaller than $\frac{C_1}{2N}$ at the last iterations. There is a benefit (both in expectation and standard deviation) in increasing the number of simulations n_k (note that the choice $n_k = 1000$ is much smaller the rule suggested by Corollary 3.8). Yet the convergence is slower in comparison with the basic Franck-Wolfe algorithm, which can be explained by the use of the selection method at each iteration.

6. Conclusion. We have investigated a large-scale and aggregative optimization problem and its relaxation. New error bounds for the relaxation gap have been obtained. We have proposed a tractable algorithm for its resolution with a detailed convergence analysis relying on concentration inequalities. Assuming that an efficient method for the resolution of the subproblems is available, the implementation of our stochastic Frank-Wolfe method is easy.

Future research will focus on refinements of the selection method, allowing the computation of $\mathcal{O}(q \wedge N/N^2)$-solutions. We also aim at working on more complex problems, involving for example convex constraints on the aggregate. Finally, we intend to apply our method to large-scale optimal control problems, such as nonconvex variants of the problem investigated in [17].

Appendix A. Concentration inequalities and other technical lemmas.

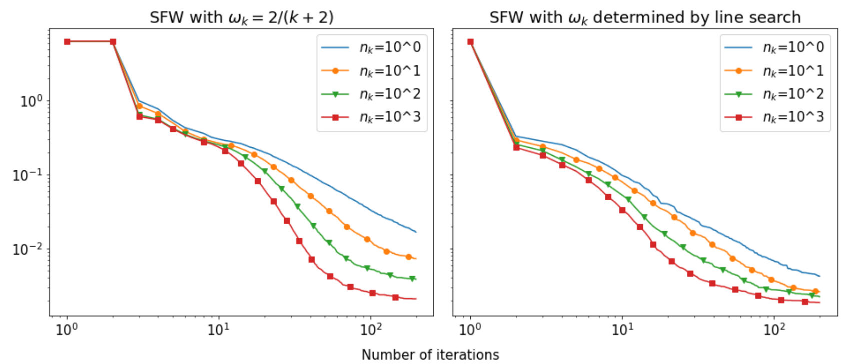

Fig. 2: MIQP by Algorithm 2 with 200 iterations, expectation of the gap.

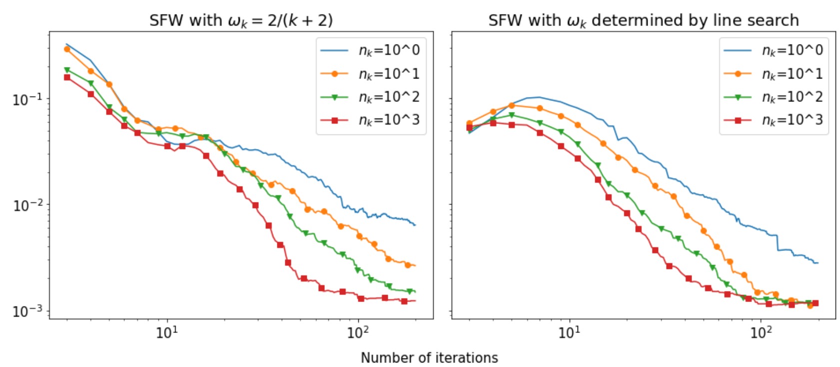

Fig. 3: MIQP by Algorithm 2 with 200 iterations, standard deviation of the gap.

Proposition A.1. *Consider T real-valued random variables $(Y_t)_{t=1,\dots,T}$. Let $(\mathcal{F}_t)_{t=1,\dots,T}$ denote the associated filtration ($\mathcal{F}_0$ is the trivial σ-algebra). Let $Z_t = \mathbb{E}[Y_t^2|\mathcal{F}_{t-1}]$ and let $S_T = \sum_{t=1}^T Y_t$. Assume the following:*

$$(\mathrm{A}.1) \qquad (i) \quad \mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = 0, \qquad (ii) \quad Y_t \le m, \qquad (iii) \quad \sum_{t'=1}^{T} Z_{t'} \le v, \quad a.s.$$

for all $t = 1, \dots, T$ and for some constants m and v. Then, $\mathbb{E}[S_T^2] \le v$. Moreover, for any $\epsilon > 0$,

$$(\mathrm{A}.2) \qquad \mathbb{P}\big[S_T \ge \epsilon\big] \le \exp\Big(-\frac{\epsilon^2}{2\,(v + \epsilon m/3)}\Big).$$

Proof. The estimate of $\mathbb{E}[S_T^2]$ can be easily obtained by induction. For the estimate of $\mathbb{P}\big[S_T \ge \epsilon\big]$, see [7, Theorem 7]. ∎

As a corollary, we obtain the following McDiarmid's inequality of "variance type".

Corollary A.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(\Omega_i)_{i=1,\dots,n}$ be n measurable subsets of Ω. Let $X = (X_i)_{i=1,\dots,n}$ be n independent random variables valued respectively in $(\Omega_i)_{i=1,\dots,n}$. Consider a measurable function $f\colon \prod_{i=1}^n \Omega_i \to \mathbb{R}$*

and real constants $m \in \mathbb{R}$ and $(v_i)_{i=1,...,n}$ such that

$$\mathrm{Var}\big[f(X_i, x_{-i})\big] \leq v_i^2, \quad a.s., \qquad \big|f(X_i, x_{-i}) - \mathbb{E}\big[f(X_i, x_{-i})\big]\big| \leq m, \qquad a.s.,$$

for all $i = 1, ..., n$ and for all $x_{-i} \in \Big(\prod_{j=1}^{i-1} \Omega_i\Big) \times \Big(\prod_{j=i+1}^{n} \Omega_j\Big)$. Then, for any $\epsilon > 0$,

$$\mathbb{P}\Big[f(\mathbf{x}) - \mathbb{E}\big[f(\mathbf{x})\big] \geq \epsilon\Big] \leq \exp\Big(-\frac{\epsilon^2}{2\left(\sum_{i=1}^n v_i^2 + \frac{m\epsilon}{3}\right)}\Big). \tag{A.3}$$

Proof. Define $Y_t = \mathbb{E}\left[f(X) \mid X_1, \ldots, X_t\right] - \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{t-1}\right]$ and apply Proposition A.1. □

LEMMA A.3. *For all $k \in \mathbb{N}$, denote $\bar{\omega}_k = \frac{2}{k+2}$. Let $(u_k)_{k\in\mathbb{N}}$ and $(\gamma_k)_{k\in\mathbb{N}}$ be two sequences of real numbers. Assume that there exists a positive number C such that*

$$\gamma_{k+1} \leq (1 - \bar{\omega}_k)\gamma_k + C\bar{\omega}_k^2 + u_k, \tag{A.4}$$

for all $k \in \mathbb{N}$. Then, for all $k \in \mathbb{N}^*$,

$$\gamma_k \leq \frac{4C}{k} + \sum_{k'=0}^{k-1} \frac{(k'+1)(k'+2)}{k(k+1)} u_{k'}. \tag{A.5}$$

Proof. We proof this lemma by induction on k. We have $\bar{\omega}_0 = 1$, thus taking $k = 0$ in (A.4), we obtain that $\gamma_1 \leq C + u_0$, which proves the claim for $k = 1$. Let us assume that the claim holds true for some $k \in \mathbb{N}^*$. We deduce from (A.4) that

$$\begin{aligned}
\gamma_{k+1} &\leq \Big(\frac{1}{k+2} + \frac{1}{(k+2)^2}\Big)4C + \frac{k}{k+2}\Big(\sum_{k'=0}^{k-1} \frac{(k'+1)(k'+2)}{k(k+1)} u_{k'}\Big) + u_k \\
&\leq \frac{4C}{k+1} + \sum_{k'=0}^{k} \frac{(k'+1)(k'+2)}{(k+1)(k+2)} u_{k'}.
\end{aligned}$$

Therefore the claim holds for $k + 1$. This concludes the proof. □

LEMMA A.4. *Let A, B, and C be three random variables. Assume that B is independent of (A, C) and that $B \sim \mathrm{Bern}(\omega)$ for some $\omega \in [0, 1]$. Let F be a real-valued function of (A, B, C). Assume that $|F(A, 1, C) - F(A, 0, C)| \leq \delta$, a.s. Finally, define $U = \mathbb{E}[F(A, B, C) \mid A, B] - \mathbb{E}[F(A, B, C) \mid A]$. Then,*

$$\mathbb{E}[U \mid A] = 0, \qquad U \leq \delta, \qquad \mathbb{E}[U^2 \mid A] \leq \omega(1-\omega)\delta^2, \quad a.s.$$

Proof. The equality $\mathbb{E}[U \mid A] = 0$ is trivial. We have $U = \mathbb{E}[Z \mid A, B]$, where

$$Z = F(A, B, C) - \mathbb{E}[F(A, B, C) \mid A, C].$$

It is easy to verify that $Z \leq \delta$, a.s., which implies that $\mathbb{E}[U \mid A] = \mathbb{E}[Z \mid A] \leq \delta$. The first inequality is proved. For the second inequality, we first note that

$$\mathbb{E}[Z^2 \mid A, C] = \omega(1-\omega)(F(A, 1, C) - F(A, 0, C))^2,$$

as can be easily verified. Thus $\mathbb{E}[Z \mid A] \leq \omega(1-\omega)\delta^2$. Next by Jensen's inequality, we have $U^2 \leq \mathbb{E}[Z^2 \mid A, B]$. Therefore,

$$\mathbb{E}[U^2 \mid A] \leq \mathbb{E}\big[\mathbb{E}[Z^2 \mid A, B] \mid A\big] = \mathbb{E}[Z^2 \mid A] \leq \omega(1-\omega)\delta^2,$$

as was to be proved. □

The following lemma is an elementary property of the conditional expectation. For the sake of simplicity, we only state it (and prove it) with discrete random variables.

LEMMA A.5. *Let X, Y, and Z be three random variables. Assume that Y and Z are discrete and that Z is independent of (X, Y). Then,* $\mathbb{E}\big[X \mid Y, Z\big] = \mathbb{E}\big[X \mid Y\big]$.

Proof. By definition, $\mathbb{E}\big[X \mid Y, Z\big] = \phi(Y, Z)$, where ϕ is defined as follows: for any pair (y, z) such that $\mathbb{P}\big[Y = y \text{ and } Z = z\big] \neq 0$,

$$\phi(y,z) = \frac{\mathbb{E}\big[X\mathbf{1}_{Y=y}\mathbf{1}_{Z=z}\big]}{\mathbb{P}\big[Y = y \text{ and } Z = z\big]} = \frac{\mathbb{E}\big[X\mathbf{1}_{Y=y}\big]}{\mathbb{P}\big[Y = y\big]},$$

since Z is independent of (X, Y). Thus ϕ does not depend on Z and the result follows. ∎

REFERENCES

[1] J. AUBIN AND I. EKELAND, *Estimates of the duality gap in nonconvex optimization*, Mathematics of Operations Research, 1 (1976), pp. 225–245.

[2] H. BAUSCHKE AND P. COMBETTES, *Convex analysis and monotone operator theory in Hilbert spaces*, vol. 408, Springer, 2011.

[3] A. BECK AND L. TETRUASHVILI, *On the convergence of block coordinate descent type methods*, SIAM Journal on Optimization, 23 (2013), pp. 2037–2060.

[4] S. BOYD, N. PARIKH, AND E. CHU, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Now Publishers Inc, 2011.

[5] J. CASSELS, *Measures of the non-convexity of sets and the Shapley–Folkman–Starr theorem*, in Mathematical Proceedings of the Cambridge Philosophical Society, vol. 78, Cambridge University Press, 1975, pp. 433–436.

[6] L. CHIZAT AND F. BACH, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, in Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc., 2018.

[7] B. DELYON, *Exponential inequalities for dependent processes*, tech. report, Oct. 2015.

[8] J. DUNN AND S. HARSHBARGER, *Conditional gradient algorithms with open loop step size rules*, Journal of Mathematical Analysis and Applications, 62 (1978), pp. 432–444.

[9] O. FERCOQ AND P. RICHTÁRIK, *Optimization in high dimensions via accelerated, parallel, and proximal coordinate descent*, Siam review, 58 (2016), pp. 739–771.

[10] J. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex analysis and minimization algorithms I: Fundamentals*, vol. 305, Springer science & business media, 1996.

[11] M. JAGGI, *Revisiting Frank-Wolfe: Projection-free sparse convex optimization*, in International Conference on Machine Learning, PMLR, 2013, pp. 427–435.

[12] J. MAIRAL, F. BACH, AND J. PONCE, *Sparse modeling for image and vision processing*, arXiv preprint arXiv:1411.3230, (2014).

[13] S. MALLAT, *A wavelet tour of signal processing*, Elsevier, 2009.

[14] C. MCDIARMID, *On the method of bounded differences*, Surveys in combinatorics, 141 (1989), pp. 148–188.

[15] F. PACAUD, *Decentralized optimization for energy efficiency under stochasticity*, PhD Thesis, Université Paris-Est, Oct. 2018.

[16] R. T. ROCKAFELLAR, *Convex analysis*, Princeton university press, 2015.

[17] A. SEGURET, C. ALASSEUR, J. F. BONNANS, A. DE PAOLA, N. OUDJANE, AND V. TROVATO, *Decomposition of high dimensional aggregative stochastic control problems*, arXiv preprint arXiv:2008.09827, (2020).

[18] R. STARR, *Quasi-equilibria in markets with non-convex preferences*, Econometrica: journal of the Econometric Society, (1969), pp. 25–38.

[19] A. TSYBAKOV, *Introduction to Nonparametric Estimation*, Springer, 2009.

[20] C. VILLANI, *Topics in optimal transportation*, vol. 58, American Mathematical Soc., 2003.

[21] M. WANG, *Vanishing price of decentralization in large coordinative nonconvex optimization*, SIAM Journal on Optimization, 27 (2017), pp. 1977–2009.