SIAM J. OPTIM.
Vol. 27, No. 3, pp. 1977–2009

VANISHING PRICE OF DECENTRALIZATION IN LARGE COORDINATIVE NONCONVEX OPTIMIZATION*

MENGDI WANG†

Abstract. We focus on nonconvex multi-agent optimization where a large number of participants collaboratively optimize some social cost function and their individual preferences. We focus on the semidecentralized multi-agent best response setting where a central planner attempts to coordinate the agents and each participant pursues self-interest. By studying the duality framework, we provide geometric and analytic characterizations of the duality gap and the price of decentralization. We prove that the nonconvex problem becomes increasingly convex as the problem scales up in dimension. Upon appropriate coordination, the price of decentralization asymptotically vanishes to zero as the number of participants grows. We develop a duality-based coordination procedure for the central planner to adapt the price vector and select a particular best response for each participant. The coordination algorithm is able to induce individual best responses to dynamically converge to an approximate global optimum, regardless of the initial solution. A convergence rate and complexity analysis as well as numerical results are provided. In the case without any coordination, we provide counterexamples showing that the price of decentralization can be disastrously high.

Key words. nonconvex optimization, duality gap, price of decentralization, multi-agent optimization, cooperative optimization, cutting plane method

AMS subject classifications. 90C26, 90C46

DOI. 10.1137/16M1068207

1. Introduction.

Real-world social and engineering systems often involve a large number of participants driven by self-interests. The participants interact with one another by contributing to some common goods or societal factors. For example, market participants make individual decisions on their consumption or production of common goods, which aggregate to the overall demand and supply in the market. As another example, factory owners make decisions about carbon emissions, which aggregately influence the air quality of a city. The city as a whole pays a price for the carbon emissions generated by individuals. The overall social welfare is the aggregation of individual interests, as well as the utility or cost associated with the common factors. In this work, we study the social welfare optimization problem where the individual preferences are not necessarily convex.

Consider the optimization problem

$$\begin{aligned} &\text{minimize } \left\{ F(x) = \sum_{i=1}^{N} p_i(x_i) + f\left(\sum_{i=1}^{N} g_i(x_i) \right) \right\}, \\ &\text{subject to } x_i \in \mathcal{X}_i, \qquad i = 1, \dots, N, \end{aligned} \tag{1}$$

where $p_i : \Re^{n_i} \mapsto \Re$ are functions that describe individual preferences, $g_i : \Re^{n_i} \mapsto \Re^q$ are functions that describe individual impacts on the common goods, $f : \Re^q \mapsto \Re$ is the social cost function of the common goods, $\mathcal{X}_i$ is a compact subset of $\Re^{n_i}$, and $n = \sum_{i=1}^{N} n_i$ is the dimension of all decisions. Here N is the number of participants,

*Received by the editors March 29, 2016; accepted for publication (in revised form) July 3, 2017; published electronically September 7, 2017.
http://www.siam.org/journals/siopt/27-3/M106820.html
Funding: This work was funded by NSF grant DMS-1619818.
†Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540 (mengdiw@princeton.edu).

and M is the number of common goods/factors. We focus on the case where $N \gg M$. Throughout this paper, we assume that the social cost function f is convex, p_i, g_i are arbitrary continuous functions and may not be convex, and $\mathcal{X}_i$ are arbitrary compact sets.

An important special case of problem (1) is the constrained problem

$$\text{(2)} \quad \text{minimize } \sum_{i=1}^{N} p_i(x_i), \qquad \text{subject to } \sum_{i=1}^{N} g_i(x_i) \in \mathcal{A}, \quad x_i \in \mathcal{X}_i, \quad i = 1, \ldots, N.$$

It is a special case of problem (1) where $f(x) = 0$ if $x \in \mathcal{A}$, and $f(x) = +\infty$ if $x \notin \mathcal{A}$.

1.1. Applications and motivations. Problems of type (1)–(2) are very common in multi-agent optimization. Several examples are aircraft or vehicle coordination [35, 33, 1], robot navigation [25, 20, 5], smart grid control [3, 34], and communication and sensor networks [37, 39, 38].

One example of an application is the *network sharing problem.* Consider a distributed network that involves a large number of nodes. The nodes can cooperate with one another subject to network load sharing/balancing constraints. They attempt to collaborate and to reach a global consensus or maximize a common objective. Meanwhile, each node is likely to possess a private task. When the consensus or sharing balance is imposed as a soft penalty, we get an instance of problem (1). When there is a hard budget on the total resources or consensus, we get an instance of problem (2). Since each node or user often has a nonconvex preference function, the network sharing problem is often nonconvex.

Another motivating application, which is somewhat unconventional, is the *sparse graph learning problem.* A graphical model is a powerful machine learning tool to explore the interrelationship among a large number of random variables [23]. Estimating large graphical models under sparsity assumptions has wide applications in heterogeneous data analysis, e.g., learning brain network using fMRI data and analyzing relational networks of the stock market. The estimation problem of sparse graphical models can be formulated into an optimization problem mimicking problem (1) [19], given by

$$\min_{\beta_j, j=1,\ldots,N} \sum_{j=1}^{N} L_j(\beta_j), \quad \text{subject to } \sum_{j=1}^{N} \|\beta_j\|_0 \le M,$$

where L_j is the negative marginal likelihood function associated with coefficient vector β_j of the jth node, and $\|\beta_j\|_0$ is the number of nonzero coefficients of β_j (or in other words, the degree of the jth node).

Nonconvexity of p_i, g_i is inevitable in a majority of multi-agent applications taking the form (1). The nonconvex p_i is able to model arbitrary preference functions by individuals. The nonconvex g_i is able to model arbitrary impact functions, including the $\|\cdot\|_0$ norm. The nonconvex compact set $\mathcal{X}_i$ allows one to model discrete decisions. Note that we focus on the special case where the social cost function f is convex. Such f is able to model both the hard constraint in resource allocation problems (as an indicator function) and the soft penalty in more general problems. Due to the nonconvexity of p_i, g_i, problem (1) is generally NP-hard (see section 5.3 for more discussions on the complexity of problem (1)). It remains an open question whether there exist efficient approximation schemes, let alone how to design efficient algorithms to achieve global coordination.

Suppose that a central planner is able to collect information and set a price for the public goods. Meanwhile, each participant is charged for making an impact on the public goods. Let $\mu \in \Re^M$ be the price vector. Each participant is able to output his best response by selfishly solving its own problem:

$$x_i(\mu) \in \operatorname{argmin}\{p_i(x_i) + \mu^T g_i(x_i) \mid x_i \in \mathcal{X}_i\}, \quad i = 1, \ldots, N.$$

We assume that the best response solution $x_i(\mu)$ can be easily computed by the ith participant, which requires solving a nonconvex but much smaller-scale optimization problem. In the setting of free markets, this is equivalent to the assumption that all participants are rational and always pursue self-interests. We will not consider how individuals solve their small nonconvex problems, which is beyond the scope of the current paper.

In this paper, we focus on the *semidecentralized multi-agent best response setting*, where the participants act largely according to their own interests and the central planner applies some level of coordination. We ask the following question:

> Is there a fair price vector μ^* such that the individual best responses $(x_1(\mu^*), \ldots, x_N(\mu^*))$ achieve a social optimum, i.e., a global optimum of problem (1)? Can it be achieved automatically or in a coordinative manner?

In the case where problem (1) is convex, the answer is largely known to be yes. However, there is no simple answer in the lack of convexity. One may wonder why we care about the nonconvex case at all. In practical markets, each participant may face multiple alternative actions (discrete or continuous) to choose from. For example, a manufacturer seeks to reduce carbon emissions by installing either a high-efficiency device or a low-efficiency one. It is likely that the convex combination of both is always suboptimal, because it requires two lump-sum payments. For another example, the estimation of the sparse graphical model uses the $\|\cdot\|$ norm to impose the sparsity assumption; thus the optimization problem is fundamentally combinatorial. Indeed, nonconvexity is ubiquitous.

The price vector is naturally related to the multiplier of an appropriate dual problem. In the nonconvex case, strong duality fails to hold. There is a positive duality gap between the primal and dual problems. As a result, there is no guarantee that there exists a fair price at which the social optimum can be automatically achieved via individual best responses. It is possible that, regardless of the price, the overall welfare is far from the optimal value due to the participants' nonconvex self-interests. We refer to this loss of efficiency as the *price of decentralization*. Indeed, we will illustrate with examples that a high price of decentralization is inevitable with neither convexity nor coordination.

In what follows, we aim to provide an answer to this question by analyzing the nonconvex duality of problem (1). We will show that the price of decentralization is related to a form of duality gap between problem (1) and a suitable Fenchel dual problem. In order to quantify the duality gap, we study the dual geometry of problem (1). We discover that problem (1) exhibits a curious *convexification effect*, i.e., the problem becomes increasingly convex as the number of participants increases. This means that the duality gap can be made arbitrarily small when the number of participants becomes large. It also suggests that the highly nonconvex problem (1) can be approximately solved using an efficient duality-based approach. Under reasonable scaling of the problem, the price of decentralization vanishes to zero asymptotically.

1.2. Related works. The convexification effect is due to an intuitive fact from convex geometry: *the sum of a large number of nonconvex sets tends to be convex.*

The first result is the Shapley–Folkman lemma, established by the Nobel prize-winning economist Lloyd Shapley [40]. It is used to derive an upper bound on the distance between the sum of many sets and its convex hull, and can be viewed as a discrete counterpart to the Lyapunov theorem on nonatomic measures [21]. Many others have looked into the convexification result from the analytical, geometric, and probabilistic perspectives; see [14, 32, 15, 16, 51, 2, 41] for examples. The convexification effect has been widely studied in mathematical economy and game theory. It is used to show that central results of convex economic theory are good approximations to large economies with nonconvexities; see [40, 22, 48, 18, 13, 12, 42] for a selected few of these works. Most of the existing research focuses on proving the existence of quasi-equilibria in a variety of multiperson games and analyzing the economic impact.

In contrast to the economic literature, there exist only a handful of works that study the convexification effect from the optimization perspective. The pioneering work by Ekeland [17] studied a separable optimization problem:

$$\text{minimize } \sum_{i=1}^{N} p_i(x_i), \qquad \text{subject to} \quad \sum_{i=1}^{N} g_i(x_i) \le B,$$

and estimated an upper bound of the Langrangian duality gap. Later, Aubin and Ekeland [4], Bertsekas [6, 11], and Pappalardo [31] considered the same problem and proved sharper bounds on the duality gap. This idea of a small duality gap has been used in spectral management in signal processing [49, 27] as well as in supply chain management [47]. Some earlier works emphasized integer linear programming with special structures; see, e.g., [7, 46]. In addition, Lemaréchal and Renaud [26], Bertsekas, Nedić, and Ozdaglar [10], and Nedić and Ozdaglar [30] have studied the Langrangian duality gap of generic nonconvex optimization from a geometric perspective. A recent work by Udell and Boyd [44] considers a separable problem with linear equality and inequality constraints. It proposes a randomized approximation method that minimizes the convex envelope of nonconvex cost functions. Another recent work by Fang, Han, and Medgdi [19] in statistical learning proposes estimating large sparse graphical models by solving an ℓ_0-constrained optimization problem. It is shown that the duality gap is sufficiently smaller than the statistical error with high probability.

We note that the nonconvex problem (1) is closely related to integer programming. In fact, many integer linear programming problems can be seen as special cases of problem (1). An example is the knapsack problem given by

$$\min c^T x, \qquad \text{subject to} \quad Ax \le b,\ x_i = \{0,1\},\ i = 1, \dots, n,$$

where $c \in \Re^n$, $A \in \Re^{m \times n}$, and $b \in \Re^m$. This can be seen to take the form (1) if we let p_i, g_i be linear functions and let $\mathcal{X}_i$ be sets of discrete values. Although discrete constraint is not our focus in the current work, we keep in mind that our duality results do apply to these special cases. In contrast to the integer programming literature, our focus is on finding an approximate solution by leveraging the duality. We are interested in developing algorithms that can be applied in multi-agent settings.

1.3. Scope of this work. The main objective of this paper is to study the nonconvex duality of problem (1) from both the theoretical and algorithmic perspectives. By exploiting the convexification phenomenon of sums of nonconvex sets, we provide a mathematical characterization of the duality gap between the nonconvex problem (1) and its dual problem. We show that there exists a best response solution to the optimal dual multiplier that is nearly primal optimal and attains a near-minimal price

of decentralization. In other words, there indeed exist a nearly fair price and a particular combination of individual best responses, such that the price of decentralization vanishes to 0 as the market size increases to infinity (assuming that the participants are willing to cooperate).

From an algorithmic perspective, we propose an algorithm for the central planner, which coordinates individual best responses to dynamically converge to an approximate global optimum. The algorithm relies on best responses from individual participants, who are required to optimize their own objective functions given a price vector/multiplier. The algorithm is based on a dual cutting-plane method and can be applied in the multi-agent setting. Under mild conditions, it converges to the optimal multiplier at a rate of $\mathcal{O}(1/t)$ even if the dual function is nonsmooth. A key contribution is the algorithm's ability to identify a global near-optimal solution by coordinating individuals' behaviors. The coordination is cast into an approximate projection problem, which is related to finding the extreme point of a particular linear feasibility problem. We emphasize that coordination is the key to achieving an approximate global optimum for nonconvex optimization. Without coordination, we find examples to show that the price of decentralization can be disastrously high.

To the best of the author's knowledge, this is the first work that identifies a class of nonconvex problems (1) which bear a diminishing Fenchel duality gap (as the dimension of decision variable increases). This is also the first work that provides a tractable algorithmic solution to achieve the diminishing price of decentralization. It points out the necessity of enforcing coordination in multi-agent nonconvex optimization. The rest of the paper is summarized as follows.

Outline. In section 2, we introduce the duality framework of the nonconvex problem (1), illustrate the geometry of the duality gap, and introduce the price of decentralization. In section 3, we characterize the minimal price of decentralization and show that it can be achieved by a particular best response solution to the optimal multiplier. In section 4, we study how to coordinate individual best responses to achieve the approximate optimum, and we show by counterexamples that the lack of coordination may result in a high price of decentralization. In section 5, we propose a coordinative algorithm relying on individual best responses and show that it converges to the approximate global optimum at a favorable rate. In section 6, we conduct numerical experiments, and in section 7, we draw conclusions.

Notation. All vectors are considered as column vectors. For a vector $x \in \Re^n$, we denote by x^T its transpose, and we denote by $\|x\| = \sqrt{x^T x}$ its Euclidean norm. For two sequences $\{a_k\}, \{b_k\}$, we denote $a_k = \mathcal{O}(b_k)$ if there exists $c > 0$ such that $\|a_k\| \leq c\|b_k\|$ for all k, and we denote $a_k \to a$ if $\lim_{k\to\infty} a_k = a$. For a function $f(x)$, we denote by $\nabla f(x)$ its gradient at x if f is differentiable, and we denote by $\partial f(x)$ its subdifferential (the set of subgradients) or superdifferential at x if f is nondifferentiable. For convenience, we denote by $\tilde{\nabla} f(x)$ a particular subgradient of f at x, which will be specified in context. For a set $\mathcal{A}$, we denote by $\text{conv}(\mathcal{A})$ its convex hull, i.e., the set of all convex combinations of points in $\mathcal{A}$, and we denote by $|\mathcal{A}|$ its cardinality. For two sets $\mathcal{A}$ and $\mathcal{B}$, we denote by $\mathcal{A}/\mathcal{B}$ the set of their difference $\mathcal{A}/\mathcal{B} = \mathcal{A} \cap \mathcal{B}^c$, and we denote by $\mathcal{A} + \mathcal{B}$ their Minkowski sum (also known as vector sum), i.e., $\mathcal{A} + \mathcal{B} = \{a + b \mid a \in \mathcal{A}, b \in \mathcal{B}\}$. We denote by $\mathcal{X}$ the Cartesian product $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$, so the constraint of problem (1) can be written compactly as $x \in \mathcal{X}$.

2. Duality, geometry, and price of decentralization. In this section, we introduce the duality framework for the nonconvex optimization problem (1). We

also illustrate the geometric intuition of our analysis.

2.1. Primal and dual problems. Let $f^\star$ be the convex conjugate of the function f, which is defined as $f^\star(\mu) = \sup_{y\in\Re^M}\{\mu^T y - f(y)\}$. By using the convexity of f, we also have $f(y) = \sup_{\mu\in\Re^M}\{\mu^T y - f^\star(\mu)\}$.

In order to leverage the separability with respect to $x_1,\dots,x_N$, we rewrite problem (1) using the Fenchel representation as

$$\text{(P)} \qquad \min_{x\in\mathcal{X}}\left\{\sum_{i=1}^N p_i(x_i) + \sup_{\mu\in\Re^q}\left\{-f^\star(\mu) + \mu^T\sum_{i=1}^N g_i(x_i)\right\}\right\},$$

which we refer to as the *primal problem*. We define the *dual problem* as the problem obtained by exchanging the "min" and "sup,"

$$\text{(D)} \qquad \sup_{\mu\in\Re^q}\left\{-f^\star(\mu) + \sum_{i=1}^N \min_{x_i\in\mathcal{X}_i}\left\{\mu^T g_i(x_i) + p_i(x_i)\right\}\right\}.$$

We let $L(x,\mu)$ be the Lagrangian function

$$L(x,\mu) = -f^\star(\mu) + \mu^T\sum_{i=1}^N g_i(x_i) + \sum_{i=1}^N p_i(x_i),$$

which is decomposable with respect to $x_1,\dots,x_N$. The primal function $F(x)$ can be represented using the Lagrangian function as

$$F(x) = \sup_{\mu\in\Re^M} L(x,\mu) = \sup_{\mu\in\Re^M}\left\{-f^\star(\mu) + \mu^T\sum_{i=1}^N g_i(x_i) + \sum_{i=1}^N p_i(x_i)\right\}.$$

Similarly, we define the dual function as

$$Q(\mu) = \min_{x\in\mathcal{X}} L(x,\mu) = -f^\star(\mu) + \sum_{i=1}^N \min_{x_i\in\mathcal{X}_i}\left\{\mu^T g_i(x_i) + p_i(x_i)\right\}.$$

We denote by F^* and Q^* the optimal values of problems (P) and (D), respectively. They satisfy

$$F^* = \min_{x\in\mathcal{X}} F(x) = \min_{x\in\mathcal{X}}\sup_{\mu\in\Re^M} L(x,\mu), \qquad Q^* = \sup_{\mu\in\Re^M} Q(\mu) = \sup_{\mu\in\Re^M}\min_{x\in\mathcal{X}} L(x,\mu).$$

By using weak duality (see [36]), we have

$$F^* \geq Q^*.$$

In earlier literature, the nonnegative difference $F^* - Q^*$ is often referred to as the duality gap of problem (1); see [9] for an example.

2.2. A geometric interpretation of duality gap. Let us try to understand the nonconvexity of problem (1) and its duality gap from a geometric point of view. Related earlier works such as [26, 10, 30] have only focused on Lagrangian duality. We define $\mathcal{W}_i \subset \Re^{M+1}$, $i = 1,\dots,N$, to be the set

$$\mathcal{W}_i = \{(z,y) \mid \exists\, x_i \in \mathcal{X}_i : z \geq p_i(x_i), y = g(x_i)\}.$$

If p_i, $\mathcal{X}_i$ are convex and g_i is linear, the set $\mathcal{W}_i$ is convex; otherwise $\mathcal{W}_i$ is not necessarily convex. The set $\mathcal{W}_i$ provides a joint characterization of the functions (p_i, g_i) and the set $\mathcal{X}_i$. Moreover, the set $\mathcal{W}_i$ provides an invariant representation of the ith participant's interest under change of variables. In subsequent analyses, we will use the convexity gap of $\mathcal{W}_1, \ldots, \mathcal{W}_N$ as a metric of the lack of convexity of problem (1).

We define $\mathcal{W}$ to be the *Minkowski sum* of the sets $\mathcal{W}_1, \ldots, \mathcal{W}_N$ given by

$$\mathcal{W} = \mathcal{W}_1 + \cdots + \mathcal{W}_N = \left\{ \sum_{i=1}^{N} w_i \;\middle|\; w_i \in \mathcal{W}_i, i = 1, \ldots, N \right\}.$$

Equivalently, we have

$$\mathcal{W} = \left\{ (z, y) \;\middle|\; \exists\, x \in \mathcal{X} : z \geq \sum_{i=1}^{N} p_i(x_i), y = \sum_{i=1}^{N} g_i(x_i), i = 1, \ldots, N \right\}.$$

Since $\mathcal{X}$ is compact and p_i, g_i are continuous, we can verify that $\mathcal{W}_1, \ldots, \mathcal{W}_N, \mathcal{W}$ are nonempty and compact.

Next we provide a heuristic analysis to argue that the primal problem (P) and the dual problem (D) can be represented using the set $\mathcal{W}$ and its convex hull $\mathrm{conv}(\mathcal{W})$, respectively. By using the definition of $\mathcal{W}$, we see that the primal problem (P) is equivalent to

$$F^* = \min_{w \in \mathcal{W}} \sup_{\mu \in \Re^M} \left\{ [1, \mu^T] w - f^\star(\mu) \right\}. \tag{3}$$

Similarly, we can rewrite the dual problem (D) as

$$Q^* = \sup_{\mu \in \Re^M} \min_{w \in \mathcal{W}} [1, \mu^T] w - f^\star(\mu).$$

Note that minimizing a linear function over a closed set is equivalent to minimizing over the convex hull. Therefore, we can replace $\mathcal{W}$ with its convex hull, and assuming that the minimax theorem holds (which would require additional technical assumptions), we further obtain

$$\begin{aligned} Q^* &= \sup_{\mu \in \Re^M} \min_{w \in \mathrm{conv}(\mathcal{W})} [1, \mu^T] w - f^\star(\mu) \\ &= \min_{w \in \mathrm{conv}(\mathcal{W})} \sup_{\mu \in \Re^M} [1, \mu^T] w - f^\star(\mu). \end{aligned} \tag{4}$$

Now let us compare (3) and (4). Clearly, the dualization can be viewed as a form of convexification. We emphasize that the dual problem is *not* obtained by replacing nonconvex functions in problem (1) with their convex envelops. It is the set $\mathcal{W}$ that is convexified in the dual problem, instead of the nonconvex function p_i, g_i and sets $\mathcal{X}_i$.

The fundamental nonconvexity of the primal problem is due to the nonconvex set $\mathcal{W}$. This implies that the duality gap between problems (P) and (D) can be estimated using the nonconvexity of $\mathcal{W}$. An important observation is that $\mathcal{W}$ is the sum of many sets $\mathcal{W} = \mathcal{W}_1 + \cdots + \mathcal{W}_N$. So the convexification phenomenon occurs (see Figure 1 for an illustration). When N is a large number, we can show that the set $\mathcal{W}$ does not differ much from its convex hull $\mathrm{conv}(\mathcal{W})$. This is the geometric motivation for our subsequent analysis.

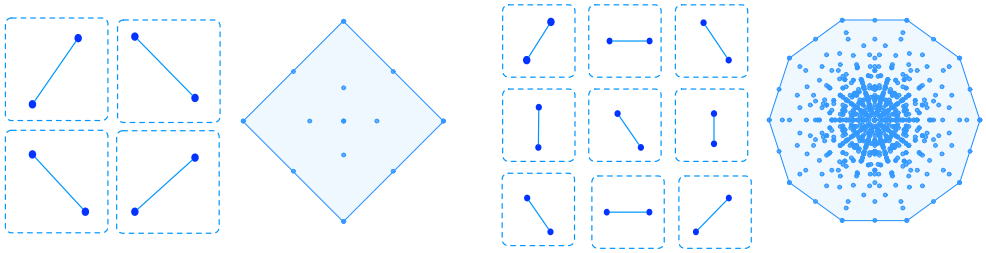

FIG. 1. *Illustration of the sums of nonconvex sets. On the left side, we plot four* 2*-point sets* $\mathcal{W}_1, \ldots, \mathcal{W}_4$*, their sum* $\mathcal{W}_1 + \cdots + \mathcal{W}_4$*, and their convex hulls. On the right side, we plot nine* 2*-point sets* $\mathcal{W}_1, \ldots, \mathcal{W}_9$*, their sum* $\mathcal{W}_1 + \cdots + \mathcal{W}_9$*, and their convex hulls. The convexity gap reduces as the number of sets increases.*

2.3. Price of decentralization. We argue that the conventional notion of duality gap $F^* - Q^*$ is not useful enough, especially in the context of computation. The difference between two optimal values does not quantify how the best response solution to the optimal multiplier performs in the primal problem. We need a new notion other than the duality gap to characterize the efficiency loss by solving the dual problem instead of the primal problem. Let μ^* be the optimal multiplier to the dual problem, which satisfies $\mu^* \in \operatorname{argmax}_{\mu \in \Re^M} Q(\mu)$.

DEFINITION 2.1. *We say that* $F(\check{x}) - F^*$ *is a price of decentralization of problem* (1) *if* $\check{x} = (\check{x}_1, \ldots, \check{x}_N)$ *is a best response to the optimal multiplier* μ^**, i.e.,*

$$\check{x}_i \in \operatorname{argmin}\{p_i(x_i) + {\mu^*}^T g_i(x_i) \mid x_i \in \mathcal{X}_i\}$$

for $i = 1, \ldots, N$*. Note that the price of decentralization may take multiple values because* $\check{x}$ *is not unique.*

We refer to the normalized difference $\frac{F(\check{x}) - F^*}{F^*}$ as the normalized price of decentralization. When the problem is convex, the price of decentralization is always zero by the strong duality. In lack of convexity, the best response $\check{x}$ to the optimal multiplier is often a suboptimal solution to the primal problem. We are interested in the positive difference $F(\check{x}) - F^*$, which is exactly the loss of total efficiency by applying $\check{x}$ to the primal objective.

Note that even if the optimal multiplier μ^* is unique, there may exist multiple choices of $\check{x}_i$ that are best responses to μ^* for the ith participant. Due to the nonconvex nature of p_i and g_i, the set of such best responses is also nonconvex. Thus we are interested in the "best" combination of individual best responses and the associated minimal price of decentralization, which is defined as follows.

DEFINITION 2.2. *The minimal price of decentralization of problem* (1) *is given by*

$$\begin{aligned} &\textit{minimize } F(\check{x}) - F^* \\ &\textit{subject to } \check{x}_i \in \operatorname{argmin}\left\{p_i(x_i) + {\mu^*}^T g_i(x_i) \mid x \in \mathcal{X}\right\}, \\ &\qquad i = 1, \ldots, N. \end{aligned} \tag{5}$$

Note that the price of decentralization $F(\check{x}) - F^*$ is related to but different from the traditional duality gap $F^* - Q^*$. Both of them are determined by the convexity gap of problem (1). In the rest of this paper, we will leverage the duality framework

and quantify the convexity gap of $\mathcal{W}$. Then we will be able to provide a complete mathematical characterization of the minimal price of decentralization and show that the minimal price of decentralization reduces to zero as the nonconvex problem scales up.

3. Analyzing the price of decentralization using duality. In this section, we analyze the price of decentralization of the multi-agent nonconvex optimization problem (1). First, we review some preliminaries about the convexity gap and the convexification of sums of sets. Second, we derive upper bounds for the minimal price of decentralization in two cases: the case with smooth penalty and the case with hard constraint. Finally, we show that the normalized price of decentralization vanishes to zero as the number of participants N grows to infinity.

3.1. Preliminaries. We state the Shapley–Folkman lemma, which was first proved in [40]. It implies that the sum of many nonconvex sets does not differ much from its convex hull. To illustrate the key idea, we provide a simplified proof based on linear programming.

LEMMA 3.1 (Shapley–Folkman lemma). *Let $\mathcal{S}_1, \dots, \mathcal{S}_n \subset \Re^m$, and let $x \in \operatorname{conv}(\mathcal{S}_1 + \cdots + \mathcal{S}_n)$. There exist $x_1, \dots, x_n$ and $\mathcal{I} \subset \{1, \dots, n\}$ such that $x = x_1 + \cdots + x_n$ and $|\mathcal{I}| \le m$ with*

$$x_i \in \begin{cases} \mathcal{S}_i & \text{if } i \notin \mathcal{I}, \\ \operatorname{conv}(\mathcal{S}_i)/\mathcal{S}_i & \text{if } i \in \mathcal{I}. \end{cases}$$

The Shapley–Folkman lemma implies that any point in $\operatorname{conv}(\mathcal{S}_1 + \cdots + \mathcal{S}_n)$ can be approximated by a point in $\mathcal{S}_1 + \cdots + \mathcal{S}_n$. In order to quantify the precision of approximation, we need a notion of a convexity gap for sets. Let $\mathcal{S}$ be an arbitrary set. One way to define the *convexity gap* of $\mathcal{S}$ is

$$\rho(\mathcal{S}) = \sup\{\|x - y\| \mid \lambda x + (1-\lambda)y \in \operatorname{conv}(\mathcal{S})/\mathcal{S} \ \forall \lambda \in (0,1)\},$$

which is the length of the longest line segment that lies in $\operatorname{conv}(\mathcal{S})/\mathcal{S}$.

Let us focus on the optimization problem (1). In order to prove tight duality gap results, we need to customize the notion of the convexity gap to our problem. Let

$$\rho_p(\mathcal{W}_i) = \sup\{|z_1 - z_2| \mid \lambda(z_1, y_1) + (1-\lambda)(z_2, y_2) \in \operatorname{conv}(\mathcal{W}_i)/\mathcal{W}_i \,\forall\, \lambda \in (0,1)\},$$

$$\rho_g(\mathcal{W}_i) = \sup\{\|y_1 - y_2\| \mid \lambda(z_1, y_1) + (1-\lambda)(z_2, y_2) \in \operatorname{conv}(\mathcal{W}_i)/\mathcal{W}_i \,\forall\, \lambda \in (0,1)\}.$$

These can be viewed as partial convexity gaps of set $\mathcal{W}_i$ with respect to the preference and impact functions p_i and g_i, respectively. We easily see that $\rho_p(\mathcal{W}_i) \le \rho(\mathcal{W}_i)$ and $\rho_g(\mathcal{W}_i) \le \rho(\mathcal{W}_i)$ for all $i = 1, \dots, N$. We further define

$$\delta_g = \max_{i=1,\dots,N} \rho_g(\mathcal{W}_i), \qquad \delta_p = \max_{i=1,\dots,N} \rho_p(\mathcal{W}_i).$$

These are the maximal nonconvex gaps of sets $\mathcal{W}_1, \dots, \mathcal{W}_N$, with respect to the impact and preference, respectively. They will be used frequently in subsequent analyses.

We say a convex function f is *β-strongly smooth* if it is continuously differentiable and satisfies for all x, y that

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2}\|y - x\|^2.$$

We say f is *σ-strongly convex* if for all x, y and $\gamma \in [0,1]$

$$\gamma f(x) + (1-\gamma) f(y) \geq f(\gamma x + (1-\gamma) y) + \frac{\sigma}{2}\gamma(1-\gamma)\|y-x\|^2,$$

and if in addition f is differentiable, we have for all x, y that

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\sigma}{2}\|y-x\|^2.$$

The next lemma states the duality between strong convexity and strong smoothness, which has been proved in [24] (see Theorem 6).

LEMMA 3.2. *Assume that f is a closed and convex function. Then f is β-strongly smooth if and only if its convex conjugate $f^\star$ is $1/\beta$-strongly convex.*

In addition, we need the following lemma, which characterizes the superdifferential of the dual function.

LEMMA 3.3. *The superdifferential of the dual function is*

$$\partial Q(\mu) = \operatorname{conv}\left\{ \sum_{i=1}^N g_i(x_i) - y \;\middle|\; x \in \check{\mathcal{X}}^*(\mu), y \in \partial f^\star(\mu) \right\},$$

where

$$\check{\mathcal{X}}^*(\mu) = \operatorname{argmin}\left\{ (\mu)^T \left(\sum_{i=1}^N g_i(x_i) \right) + \sum_{i=1}^N p_i(x_i) \;\middle|\; x \in \mathcal{X} \right\}.$$

Proof. We have $Q(\mu) = -f^\star(\mu) + \min_{x\in\mathcal{X}}\{\mu^T \sum_{i=1}^N g_i(x_i) + \sum_{i=1}^N p_i(x_i)\}$. Since $-Q(\mu)$ is the pointwise maximum of multiple convex functions, its subdifferential is the convex hull of all subgradients of functions that are currently active (see page 737 of [8] for the Danskin theorem). ∎

3.2. Upper bounds for the price of decentralization. Our first main result establishes the existence of a "reasonable" best response solution to the optimal multiplier. We show that one can construct a best response solution $\check{x}$ to the optimal multiplier μ^* such that the strong duality condition is approximately achieved, i.e.,

$$\check{x} \in \operatorname{argmin}\, L(x, \mu^*), \qquad \tilde{\nabla}_\mu L(\check{x}, \mu^*) \approx 0.$$

This implies that $(\check{x}, \mu^*)$ is an approximate saddle point of the Lagrangian function.

THEOREM 3.4 (near-optimality condition). *Suppose there exists an optimal dual multiplier μ^* to problem* (D). *Then there exists a best response solution $\check{x}$ to the optimal dual multiplier and a subgradient $\tilde{\nabla} f^\star(\mu^*) \in \partial f^\star(\mu^*)$ such that*

$$\left\| \tilde{\nabla} f^\star(\mu^*) - \sum_{i=1}^N g_i(\check{x}_i) \right\| \leq M\delta_g, \qquad (\mu^*)^T \left(\tilde{\nabla} f^\star(\mu^*) - \sum_{i=1}^N g_i(\check{x}_i) \right) \leq M\delta_p. \tag{6}$$

Proof. Let $\mu^* \in \operatorname{argmax} Q(\mu)$ be an optimal multiplier to the dual problem (D). Define $\check{\mathcal{X}}^*$ to be the set of best response solutions to the optimal multiplier μ^*, given by

$$\check{\mathcal{X}}^* = \operatorname{argmin}_{x\in\mathcal{X}}\left\{ (\mu^*)^T \left(\sum_{i=1}^N g_i(x_i) \right) - f^\star(\mu^*) + \sum_{i=1}^N p_i(x_i) \right\}.$$

Since the Langrangian function is decomposable with respect to $x_1, \ldots, x_N$, we can rewrite $\check{\mathcal{X}}^*$ as the Cartesian product of many simpler sets, $\check{\mathcal{X}}^* = \prod_{i=1}^N \check{\mathcal{X}}_i^*$, where we define

$$\check{\mathcal{X}}_i^* = \operatorname{argmin}_{x_i \in \mathcal{X}_i} \left\{ (\mu^*)^T g_i(x_i) + p_i(x_i) \right\}.$$

Since $\mathcal{X}_i$ is compact and g_i, p_i are continuous functions, the set $\check{\mathcal{X}}_i^*$ is nonempty and compact (but not necessarily convex).

By using the definition of μ^* and the concavity of $Q(\mu)$, we have $0 \in \partial Q(\mu^*)$. Then, by applying Lemma 3.3, we obtain

$$0 \in \partial Q(\mu^*) = \operatorname{conv} \left\{ \sum_{i=1}^N g_i(x_i) - y \;\middle|\; x \in \check{\mathcal{X}}^*, y \in \partial f^\star(\mu^*) \right\}.$$

Since $\partial f^\star$ is convex, it follows that there exists a subgradient $\tilde{\nabla} f^\star(\mu^*) \in \partial f^\star(\mu^*)$ such that

$$\begin{aligned} \tilde{\nabla} f^\star(\mu^*) &\in \operatorname{conv} \left\{ \sum_{i=1}^N g_i(x_i) \;\middle|\; x \in \check{\mathcal{X}}^* \right\} \\ &= \operatorname{conv} \left\{ \sum_{i=1}^N g_i(x_i) \;\middle|\; x_i \in \check{\mathcal{X}}_i^*, i = 1, \ldots, N \right\} \\ &= \operatorname{conv} \left\{ \mathcal{G}_1 + \cdots + \mathcal{G}_N \right\}, \end{aligned}$$

where we define

$$\mathcal{G}_i = \{ g_i(x_i) \mid x_i \in \check{\mathcal{X}}_i^* \}, \qquad i = 1, \ldots, N.$$

By using the Shapley–Folkman lemma, Lemma 3.1, there exists $y_1, \ldots, y_N$ such that

$$\tilde{\nabla} f^\star(\mu^*) = y_1 + \cdots + y_N,$$

where among $y_1, \ldots, y_N$, at most M out of them satisfy $y_i \in \operatorname{conv}(\mathcal{G}_i)/\mathcal{G}_i$ and the rest satisfy $y_i \in \mathcal{G}_i$.

Now let us construct a solution $\check{x} \in \check{\mathcal{X}}^*$ that satisfies (6). The construction is index by index. There are two cases:

- Consider an index $i \in \{1, \ldots, N\}$ such that $y_i \in \mathcal{G}_i$; we take $\check{x}_i \in \check{\mathcal{X}}_i^*$ and $z_i \in \Re$ to be such that

  $$y_i = g_i(\check{x}_i), \qquad z_i = p_i(\check{x}_i).$$

  This case happens at least $N - M$ times.
- Consider an index $i \in \{1, \ldots, N\}$ such that $y_i \in \operatorname{conv}(\mathcal{G}_i)/\mathcal{G}_i$. We let z_i be such that $(z_i, y_i) \in \operatorname{conv}(\check{\mathcal{W}}_i)/\check{\mathcal{W}}_i$, where $\check{\mathcal{W}}_i = \{ (p_i(x), g_i(x)) \mid x \in \check{\mathcal{X}}_i^* \}$. Then we take $\check{x}_i$ to be

  $$\check{x}_i \in \operatorname{argmin}_{x_i \in \check{\mathcal{X}}_i^*} \|y_i - g_i(x_i)\|^2, \tag{7}$$

  so $g_i(\check{x}_i)$ and y_i form an open line segment in $\operatorname{conv}(\mathcal{G}_i)/\mathcal{G}_i$. Such $\check{x}_i$ exists because $\check{\mathcal{X}}_i^*$ is nonempty and compact and g_i, p_i are continuous. We see that the two points (z_i, y_i) and $(p_i(\check{x}_i), g_i(\check{x}_i))$ form an open line segment in $\operatorname{conv}(\check{\mathcal{W}}_i)/\check{\mathcal{W}}_i$. We claim that these two points also form an open line segment in $\operatorname{conv}(\mathcal{W}_i)/\mathcal{W}_i$. If it is not true, there would exist another $\hat{x}_i \in \mathcal{X}_i$

such that $(p_i(\hat{x}_i), g_i(\hat{x}_i))$ lies on the open line segment between (z_i, y_i) and $(p_i(\check{x}_i), g_i(\check{x}_i))$. Such a point $\hat{x}_i$ is also a minimizer of $(\mu^*)^T g_i(x_i) + p_i(x_i)$. So we have $\hat{x}_i \in \check{\mathcal{X}}_i^*$ and $(p_i(\hat{x}_i), g_i(\hat{x}_i)) \in \check{\mathcal{W}}_i$, which conflicts with the fact that $\big((z_i, y_i), (p_i(\check{x}_i), g_i(\check{x}_i))\big)$ is an open line segment outside $\check{\mathcal{W}}_i$. Therefore, the two points (z_i, y_i) and $(p_i(\check{x}_i), g_i(\check{x}_i))$ form an open line segment in $\mathrm{conv}(\mathcal{W}_i)/\mathcal{W}_i$. By using the definition of the convexity gap, we have

$$\|y_i - g(\check{x}_i)\| \le \rho_g(\mathcal{W}_i), \qquad |z_i - p_i(\check{x}_i)| \le \rho_p(\mathcal{W}_i).$$

This case happens at most M times.

So far we have constructed a best response solution $\check{x} \in \check{\mathcal{X}}^*$ to the optimal dual multiplier. We can show that

$$\begin{aligned}\left\|\nabla f^\star(\mu^*) - \sum_{i=1}^N g_i(\check{x}_i)\right\| &= \left\|y_1 + \cdots + y_N - \sum_{i=1}^N g_i(\check{x}_i)\right\| \\ &\le \|y_1 - g(\check{x}_1)\| + \cdots + \|y_N - g(\check{x}_N)\| \\ &\le M \max_{i=1,\ldots,N} \rho_g(\mathcal{W}_i) \\ &= M\delta_g.\end{aligned}$$

By using the definition of $\check{\mathcal{X}}_i^*$ and $\check{\mathcal{W}}_i = \{(g_i(x), p_i(x)) \mid x \in \check{\mathcal{X}}_i^*\}$, we see that ${\mu^*}^T g(x_i) + p(x_i)$ takes constant value for all $x_i \in \check{\mathcal{X}}_i^*$. Since $(z_i, y_i) \in \mathrm{conv}(\check{\mathcal{W}}_i)/\check{\mathcal{W}}_i$ and $\check{x}_i \in \check{\mathcal{X}}_i^*$, we have

$${\mu^*}^T y_i + z_i = {\mu^*}^T g(\check{x}_i) + p(\check{x}_i), \qquad i = 1, \ldots, N.$$

We use the preceding equality and obtain that

$$\begin{aligned}(\mu^*)^T \left(\tilde{\nabla} f^\star(\mu^*) - \sum_{i=1}^N g_i(\check{x}_i)\right) &= (\mu^*)^T \left(\sum_{i=1}^N y_i - \sum_{i=1}^N g_i(\check{x}_i)\right) \\ &= \sum_{i=1}^N (p_i(\check{x}_i) - z_i) \\ &\le M\delta_p. \end{aligned}$$

∎

Theorem 3.4 shows that the near-optimality condition (6) is achievable by a particular best response solution $\check{x}$ to μ^*. We remark that the construction can be simplified to achieve a slightly relaxed error bound. For $i \in \{1, \ldots, N\}$ such that $z_i \in \mathrm{conv}(\mathcal{G}_i)/\mathcal{G}_i$, we may omit the nonconvex projection step (7) and choose an arbitrary best response $\check{x}_i \in \check{\mathcal{X}}_i^*$ instead. As a result, we obtain a slightly relaxed near-optimality condition:

$$\text{(8)} \qquad \left\|\tilde{\nabla} f^\star(\mu^*) - \sum_{i=1}^N g_i(\check{x}_i)\right\| \le M\gamma_g, \qquad (\mu^*)^T \left(\tilde{\nabla} f^\star(\mu^*) - \sum_{i=1}^N g_i(\check{x}_i)\right) \le M\gamma_p,$$

where γ_g and γ_p are the diameters of the sets $\{g_i(x) \mid x \in \mathcal{X}_i\}$ and $\{p_i(x) \mid x \in \mathcal{X}_i\}$, respectively. The relaxed bound uses the maximal diameters of the sets $\mathcal{W}_1, \ldots, \mathcal{W}_N$ rather than the maximum convexity gap. In the case where $p_i, g_i, \mathcal{X}_i$ are uniformly bounded across all participants, the relaxed condition (8) has a similar asymptotic

property as the one given in Theorem 1. If we normalize the error bounds by $1/F^* = \mathcal{O}(1/N)$, both error bounds diminish to zero as $N/M \to \infty$.

By leveraging the convexification phenomenon, Theorem 3.4 states that although there does not exist $x \in \check{\mathcal{X}}^*$ such that $\tilde{\nabla}_\mu L(x, \mu^*) = 0$, there exists a solution $\check{x} \in \check{\mathcal{X}}^*$ such that $\tilde{\nabla}_\mu L(\check{x}, \mu^*) \approx 0$. In other words, it is possible to find some $\check{x}$ such that strong duality is "nearly" satisfied. In what follows, we will use this fact to show that $\check{x}$ is indeed an approximate optimum to the primal problem. Our next result concerns the case where the penalty function f is sufficiently smooth.

THEOREM 3.5 (price of decentralization of smooth penalized problems). *Let f be β-strongly smooth. Then there exists a best response solution $\check{x}$ to the optimal dual multiplier that is a nearly optimal solution to problem* (1) *such that*

$$F^* \leq F(\check{x}) \leq F^* + 2\beta \left(q\delta_g\right)^2.$$

Proof. Given that f is β-strongly smooth, we use the duality between conjugate functions and obtain that $f^\star$ is $\frac{1}{\beta}$-strongly convex (see Lemma 3.2). The dual function $Q(\mu)$ is the infimum of multiple $\frac{1}{\beta}$-strongly concave functions; therefore, it is also $\frac{1}{\beta}$-strongly concave. It follows that there exists at least one optimal multiplier $\mu^* \in \operatorname{argmin} Q(\mu)$. We apply Theorem 3.4 and obtain that there exists a solution $\check{x}$ to problem (D) such that $\|\tilde{\nabla} f^\star(\mu^*) - \sum_{i=1}^N g_i(\check{x}_i)\| \leq M\delta_g$, where $\tilde{\nabla} f^\star(\mu^*)$ is a subgradient of $f^\star$ at μ^*.

In what follows, we show that $\check{x}$ satisfies the error bound stated in the theorem. We apply $\check{x}$ to the primal problem (1) and obtain

$$F(\check{x}) = \sum_{i=1}^N p_i(\check{x}_i) + \sup_{\mu \in \Re^q} \left\{ \mu^T \left(\sum_{i=1}^N g_i(\check{x}_i) \right) - f^\star(\mu) \right\}.$$

We also apply $\check{x}$ to the dual problem (D) and obtain

$$Q^* = \sum_{i=1}^N p_i(\check{x}_i) + (\mu^*)^T \left(\sum_{i=1}^N g_i(\check{x}_i) \right) - f^\star(\mu^*).$$

Next, we compare the values of $F(\check{x})$ and Q^*.

Let us define the function $h(\cdot)$ and its maximizer $\hat{\mu}$ to be

$$h(\mu) = \mu^T \left(\sum_{i=1}^N g_i(\check{x}_i) \right) - f^\star(\mu), \qquad \hat{\mu} = \operatorname*{argmax}_{\mu \in \Re^q} h(\mu).$$

We can see that $F(\check{x}) - Q^* = h(\hat{\mu}) - h(\mu^*)$ and $\sum_{i=1}^N g_i(\check{x}_i) - \tilde{\nabla} f^\star(\mu^*) \in \partial h(\mu^*)$. On one hand, by using the concavity of h, we have

$$h(\hat{\mu}) - h(\mu^*) \leq \left(\sum_{i=1}^N g_i(\check{x}_i) - \tilde{\nabla} f^\star(\mu^*) \right)^T (\hat{\mu} - \mu^*) \leq \left\| \sum_{i=1}^N g_i(\check{x}_i) - \tilde{\nabla} f^\star(\mu^*) \right\| \|\hat{\mu} - \mu^*\|.$$

By applying Theorem 1, we obtain that $h(\hat{\mu}) - h(\mu^*) \leq q\delta_g \|\hat{\mu} - \mu^*\|$. On the other hand, by using the $1/\beta$-strong convexity of $f^\star$ and $-h$ and the optimality of $\hat{\mu}$, we further obtain $h(\hat{\mu}) - h(\mu^*) \geq \frac{1}{2\beta}\|\hat{\mu} - \mu^*\|^2$. Combining the preceding two relations, we obtain $\|\hat{\mu} - \mu^*\| \leq 2\beta q \delta_g$. As a result, we have

$$F(\check{x}) - Q^* = h(\hat{\mu}) - h(\mu^*) \leq q\delta_g \|\hat{\mu} - \mu^*\| \leq 2\beta \left(q\delta_g\right)^2.$$

Finally, by using weak duality, we have $F(\check{x}) - F^* \leq F(\check{x}) - Q^* \leq 2\beta \left(q\delta_g\right)^2$. ∎

Next, we focus on the constrained version of problem (1) given by

$$\text{(9)}\quad \text{minimize } \sum_{i=1}^{N} p_i(x_i), \qquad \text{subject to } \sum_{i=1}^{N} g_i(x_i) \in \mathcal{A}, \quad x_i \in \mathcal{X}_i, \quad i = 1, \ldots, N.$$

Problem (9) can be viewed as a special case of problem (1) by letting $f(\cdot)$ be the indicator function of the set $\mathcal{A}$, i.e.,

$$f(y) = \begin{cases} 0, & y \in \mathcal{A}, \\ +\infty, & y \notin \mathcal{A}. \end{cases}$$

When $\mathcal{A}$ is convex and nonempty, the convex conjugate of f and its subdifferential are

$$f^\star(\mu) = \sup_{y \in \mathcal{A}} \mu^T y, \qquad \partial f^\star(\mu) = \operatorname*{argmax}_{y \in \mathcal{A}} \mu^T y.$$

It is easy to see that $\partial f^\star(\mu) \subset \mathcal{A}$ for all μ. We have the following result.

THEOREM 3.6 (price of decentralization of constrained problems). *Assume that $\mathcal{A}$ is a convex set and there exists $x \in \mathcal{X}$ such that $\sum_{i=1}^N g_i(x_i)$ is an interior point of $\mathcal{A}$. Then there exists a best response solution $\check{x}$ to the optimal dual multiplier μ^* that is a nearly feasible optimal solution to problem* (9) *such that*

$$\sum_{i=1}^{N} g_i(\check{x}_i) \in \mathcal{A} + \mathcal{B}(0, M\delta_g), \qquad F^* \le \sum_{i=1}^{N} p_i(\check{x}_i) \le F^* + M\delta_p,$$

where $\mathcal{B}(0, \epsilon)$ denotes the closed ball centered at 0 with radius ϵ.

Proof. We claim that there exists at least one optimal multiplier to the dual problem. Since the primal problem is feasible, we have $F^* < \infty$. By using weak duality, we have $Q^* \le F^* < \infty$. Since there is at least one μ such that $Q(\mu) > -\infty$, we obtain that Q^* is a finite value. Then there exists a sequence $\{\mu_k\}$ such that $\lim_{k\to\infty} Q(\mu_k) = Q^*$. We assume to the contrary that $\{\mu_k\}$ is unbounded. Let $\{\frac{\mu_k}{\|\mu_k\|}\}$ be the projections of $\{\mu_k\}$ on the unit ball, so it has at least one convergent subsequence. We assume without loss of generality that $\{\frac{\mu_k}{\|\mu_k\|}\}$ converges to some vector d on the unit ball.

Let $\bar{\mathcal{A}}$ be a compact subset of $\mathcal{A}$ that contains some $\sum_{i=1}^N g_i(x_i)$, where $x \in \mathcal{X}$, as an interior point. On one hand, we have

$$\begin{aligned}
\lim_{k\to\infty} \frac{Q(\mu_k)}{\|\mu_k\|} &= \lim_{k\to\infty} \inf_{x\in\mathcal{X}, y\in\mathcal{A}} \frac{1}{\|\mu_k\|} \sum_{i=1}^{N} p_i(x_i) + \left(\left(\frac{\mu_k}{\|\mu_k\|} - d\right) + d\right)^T \left(\sum_{i=1}^{N} g_i(x_i) - y\right) \\
&\le \lim_{k\to\infty} \inf_{x\in\mathcal{X}, y\in\bar{\mathcal{A}}} \frac{1}{\|\mu_k\|} \sum_{i=1}^{N} p_i(x_i) + \left(\left(\frac{\mu_k}{\|\mu_k\|} - d\right) + d\right)^T \left(\sum_{i=1}^{N} g_i(x_i) - y\right) \\
&= \inf_{x\in\mathcal{X}, y\in\bar{\mathcal{A}}} d^T \left(\sum_{i=1}^{N} g_i(x_i) - y\right),
\end{aligned}$$

where the inequality uses the fact $\bar{\mathcal{A}} \subset \mathcal{A}$, and last equality uses the boundedness of $\sum_{i=1}^N p_i(x_i)$ and $\sum_{i=1}^N g_i(x_i) - y$ over the compact set $\{x \in \mathcal{X}, y \in \bar{\mathcal{A}}\}$ and the facts $\frac{1}{\|\mu_k\|} \downarrow 0$ and $\frac{\mu_k}{\|\mu_k\|} - d \downarrow 0$. On the other hand, we have $\lim_{k\to\infty} \frac{Q(\mu_k)}{\|\mu_k\|} = 0$ because $Q(\mu^k) \to Q^*$. Now we have $\inf_{x\in\mathcal{X}, y\in\bar{\mathcal{A}}} d^T(\sum_{i=1}^N g_i(x_i) - y) \ge 0$. By using the interior

point feasibility condition, we know that $(\sum_{i=1}^N g_i(x_i) - y)$ can take arbitrary value in a small open ball centered at the origin. This implies that $d = 0$, which contradicts the fact that d is on the boundary of the unit ball. As a result, $\{\mu_k\}$ is a bounded sequence so it has at least one limit point μ^* attaining the optimal dual value Q^*.

Now that there exists an optimal multiplier μ^*, we may apply Theorem 3.4. It follows that there exists $\check{x} \in \check{\mathcal{X}}^*$, a best response solution to μ^*, such that

$$(10) \quad (\mu^*)^T \left(\tilde{\nabla} f^\star(\mu^*) - \sum_{i=1}^N g_i(\check{x}_i) \right) \le M\delta_p, \qquad \left\| \tilde{\nabla} f^\star(\mu^*) - \sum_{i=1}^N g_i(\check{x}_i) \right\| \le M\delta_g,$$

where $\tilde{\nabla} f^\star(\mu^*)$ is a subgradient of $f^\star$. Now it remains to show that $\check{x}$ is nearly feasible and nearly optimal.

First, we prove the near-feasibility. Note that $\tilde{\nabla} f^\star(\mu^*) \in \mathcal{A}$ by using the property of $f^\star$. So we apply (10) and obtain $\sum_{i=1}^N g_i(\check{x}_i) \in \mathcal{A} + \mathcal{B}(0, M\delta_g)$. Second, we prove the near-optimality. Note that since f is the indicator function of $\mathcal{A}$, we have $f^\star(\mu) = \sup_y y^T\mu$ and $\tilde{\nabla} f^\star(\mu^*) \in \operatorname{argmax}_{y\in\mathcal{A}} y^T\mu^*$; therefore, $f^\star(\mu^*) = (\mu^*)^T \tilde{\nabla} f^\star(\mu^*)$. By applying the preceding equality and the inequality (10), we have

$$\begin{aligned} Q^* &= \sum_{i=1}^N p_i(\check{x}_i) + (\mu^*)^T \left(\sum_{i=1}^N g_i(\check{x}_i) \right) - f^\star(\mu^*) \\ &= \sum_{i=1}^N p_i(\check{x}_i) + (\mu^*)^T \left(\sum_{i=1}^N g_i(\check{x}_i) - \tilde{\nabla} f^\star(\mu^*) \right) \\ &\ge \sum_{i=1}^N p_i(\check{x}_i) - M\delta_p. \end{aligned}$$

Finally, we apply weak duality and obtain $\sum_{i=1}^N p_i(\check{x}_i) \le Q^* + M\delta_p \le F^* + M\delta_p$. ∎

In Theorem 3.6, the feasibility violation is often inevitable. When the diameter of $\mathcal{A}$ is smaller than the duality gap, there is no guarantee that the constructed dual solution lies in the small feasible set. When $\mathcal{A}$ is sufficiently large, we may consider a modified problem in which $\mathcal{A}$ is replaced with its subset. We let $\mathcal{A}_\epsilon$ be a subset such that $\mathcal{A}_\epsilon + \mathcal{B}(0, \epsilon) \subset \mathcal{A}$ and solve the new problem instead. By choosing ϵ to be the $M\delta_p$ and applying Theorem 3.6, we can show that the dual solution to the modified problem is a feasible solution to the original problem.

Remarks. Theorems 3.5 and 3.6 provide upper bounds for the minimal price of decentralization. We do not intend to achieve the exact minimal price of decentralization, which would require solving the combinatorial problem (5). Instead, we attempt to provide a tractable solution with reasonable approximation guarantees, i.e., the upper bounds given by Theorems 3.5 and 3.6. These bounds are invariant with the number of participants N. As a result, they are tight with respect to N. One may wonder whether these bounds are tight with respect to parameters such as M, δ_p, δ_g. This question is substantially more difficult to answer, because getting a tight upper bound would require explicit characterization of the minimal price of decentralization, which further requires investigating the combinatorial structure of nonconvex sets. We leave this question open for future research. Within the scope of this paper, we are interested in establishing useful upper bounds for analyzing the asymptotic price of decentralization as N scales up. In the next section, we will see that these upper bounds are indeed quite small and become asymptotically negligible as $N \to \infty$.

3.3. Vanishing price of decentralization as $N \to \infty$. We analyze the asymptotic price of decentralization as the multi-agent optimization problem (1) scales up. Take the penalty case, for example, and let the smoothness parameter β be a fixed constant value. Suppose in addition that $\sum_{i=1}^{N} p_i(x_i)$ and $\sum_{i=1}^{N} g_i(x_i)$ scale up on the order of $\mathcal{O}(N)$; thus there exists $c > 0$ such that

$$F^* = \min_{x \in \mathcal{X}} \sum_{i=1}^{N} p_i(x_i) + f\left(\sum_{i=1}^{N} g_i(x_i)\right) \geq c \cdot (N + \beta N^2).$$

In other words, we suppose that the optimal value increases at least linearly as the number of participants increases.

Suppose that the functions p_i, g_i are uniformly bounded across all participants as the number of participants N increases. In this case, the maximal nonconvexity gaps δ_p, δ_g remain bounded as N increases. Then Theorem 3.5 implies that

$$\frac{F(\check{x}) - F^*}{F^*} \leq \mathcal{O}\left(\frac{\beta M^2}{N + \beta N^2}\right) \to 0 \qquad \text{as } N/M \to \infty.$$

Suppose that the preferences of the participants are drawn independently from an unbounded distribution. Under some distributional assumptions, we conjecture that the maximum convexity gaps satisfy $\delta_p = \mathcal{O}(\log N)$ and $\delta_g = \mathcal{O}(\log N)$ with high probability. Then we can show that

$$\mathbf{P}\left(\frac{F(\check{x}) - F^*}{F^*} \leq \mathcal{O}\left(\frac{\beta M^2 \log N}{N + \beta N^2}\right)\right) \to 1 \qquad \text{as } N^2/(M^2 \log N) \to \infty.$$

The vanishing price of decentralization essentially requires that $N/M \to \infty$, i.e., the number of participants be substantially larger than the number of common factors. In contrast, if N and M are on the same order, the convexification effect becomes very minor, resulting in a potentially high price of decentralization.

We have performed numerical experiments on the asymptotic price of decentralization. In section 6, we present numerical results on randomly generated instances of problem (1). We compute and plot the sample values of price of decentralization in Figure 3. The observed mean price of decentralization is inversely related to the number N, where M, β are kept constant. This validates our theory. To summarize, the minimal price of decentralization asymptotically converges to zero as the multi-agent problem scales up in proper ways. As long as the number of participants is much larger than the number of common goods, there exists a fair price and a particular best response solution with negligible price of decentralization.

4. The role of coordination. In this section, we aim to understand the role of coordination among multiple participants. From the computational perspective, we study how to find an approximate optimum $\check{x}$ when the optimal multiplier μ^* is known. Identifying such a solution out of many best responses can be viewed as coordinating the participants in a centralized manner. We discover that the coordination problem is essentially an approximate projection problem, and we provide an algorithmic solution to it. We use examples to illustrate the necessity of coordination, without which the price of decentralization can be disastrously high.

4.1. Finding the approximate global optimum when μ^* is given. We have proved the existence of a best response solution to the optimal multiplier $\check{x} \in \check{\mathcal{X}}^*$ that is nearly primal optimal. However, the set of all best response solutions $\check{\mathcal{X}}^*$

may contain as many as $\mathcal{O}(2^N)$ solutions, each of them being a best response to the multiplier μ^*. Now we consider how to identify a good solution $\check{x}$ out of the many candidates in $\check{\mathcal{X}}^*$. In other words, we need a coordination mechanism to select a decision $\check{x}_i$ for each participant i from its best responses.

Suppose that the optimal multiplier μ^* is given; the remaining problem is to find a solution $\check{x}$ that satisfies the near-optimality inequalities given in the former theory. Identifying an approximate optimum is equivalent to finding a point $\check{x} \in \check{\mathcal{X}}^*$ and a subgradient $\tilde{\nabla} f^\star(\mu^*) \in \partial f^\star(\mu^*)$ such that

$$\tilde{\nabla}_\mu L(\check{x}, \mu) = \sum_{i=1}^N g_i(\check{x}_i) - \tilde{\nabla} f^\star(\mu^*) \approx 0.$$

This can be viewed as a projection problem, in which we try to find $y \in \sum_{i=1}^N \{g_i(x_i) \mid x_i \in \check{\mathcal{X}}_i^*\}$ such that the distance between y and the subdifferential $\partial f(\mu^*)$ is minimized.

Consider a more general nonconvex projection problem. Suppose that we are given the sets $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$ and $\mathcal{S}$. The projection problem from $\mathcal{S}$ onto $\mathcal{Y}_1 + \cdots + \mathcal{Y}_N$ is given by

$$\begin{array}{ll} \text{minimize} & \|\bar{y} - (y_1 + \cdots + y_n)\| \\ \text{subject to} & \bar{y} \in \mathcal{S}, y_1 \in \mathcal{Y}_1, \ldots, y_N \in \mathcal{Y}_N. \end{array} \tag{11}$$

In Algorithm 1, we develop a computation method to approximately solve problem (11). It has three steps: solving a quadratic optimization problem, finding the extreme point solution to a linear feasibility problem, and recovering the approximate projection from the extreme point.

Algorithm 1 Approximate Projection onto Sums of Nonconvex Finite Sets

Input: $\mathcal{Y}_1, \ldots, \mathcal{Y}_N, \mathcal{S}$

1: Let $A^{(i)}$ be the matrix consisting of column vectors in $\mathcal{Y}_i$, $i = 1, \ldots, N$.

2: The first step is to let

$$(b^*, y^*) \in \operatorname{argmin}_{b \in \mathcal{S}, y} \left\{ \|b - y\|^2 \;\middle|\; \sum_{i=1}^N A^{(i)} z^{(i)} = y,\ (e^{(i)})^T z^{(i)} = 1,\ z^{(i)} \geq 0 \right\},$$

where $e^{(i)}$ is the vector with all 1's whose dimension is equal to the column dimension of $A^{(i)}$.

3: The second step is to find a basic feasible solution z to the linear feasibility problem

$$\sum_{i=1}^N A^{(i)} z^{(i)} = y^*, \qquad (e^{(i)})^T z^{(i)} = 1, \qquad z^{(i)} \geq 0, \qquad i = 1, \ldots, N.$$

4: The third step is to let

$$y_i = \begin{cases} A^{(i)} z^{(i)} & \text{if } z^{(i)} \text{ is an integer vector,} \\ \operatorname{argmin}_{y \in \mathcal{Y}_i} \|y - A^{(i)} z^{(i)}\| & \text{otherwise.} \end{cases}$$

Output: $y = (y_1, \ldots, y_N)$

Next we show that Algorithm 1 indeed finds an approximate solution to the projection problem (11). It can be directly applied to the optimization problem (1)

as a coordination procedure. When the optimal multiplier μ^* is given, we can use Algorithm 1 to identify a solution $\check{x}$ that achieves the small price of decentralization.

THEOREM 4.1. (a) *Let $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$ be nonconvex sets, and let $\mathcal{S}$ be convex. Then the vector $y = (y_1, \ldots, y_N)$ generated by Algorithm* 1 *satisfies*

$$\begin{aligned}\min_{\bar{y}\in\mathcal{S}} \|\bar{y} - (y_1 + \cdots + y_n)\| &\le \min_{\bar{y}\in S, y\in \operatorname{conv}(\mathcal{Y}_1+\cdots+\mathcal{Y}_n)} \|\bar{y} - y\| + M \max_{i=1,\ldots,N} \rho(\mathcal{Y}_i) \\ &\le \min_{\bar{y}\in S, y\in \mathcal{Y}_1+\cdots+\mathcal{Y}_n} \|\bar{y} - y\| + M \max_{i=1,\ldots,N} \rho(\mathcal{Y}_i).\end{aligned}$$

(b) *Under the assumptions of Theorems* 3.5 *and* 3.6*, let y be generated by Algorithm* 1 *with input*

$$\left(\{g_1(x_1) \mid x_1 \in \check{\mathcal{X}}_1^*\}, \ldots, \{g_N(x_N) \mid x_N \in \check{\mathcal{X}}_N^*\}, \partial f^\star(\mu^*)\right),$$

and let $\check{x}$ be such that $g_i(\check{x}_i) = \hat{y}_i$ for $i = 1, \ldots, N$. Then $\check{x}$ is an approximate optimum to problem (1) *that satisfies the inequalities given by Theorems* 3.5 *and* 3.6.

Proof. (a) According to the first step, $b^* \in \mathcal{S}$ achieves the minimal distance between $\mathcal{S}$ and $\sum_{i=1}^N \operatorname{conv}(\mathcal{Y}_i)$. Let y^* be the projection of b^* on $\sum_{i=1}^N \operatorname{conv}(\mathcal{Y}_i)$. By following an analysis similar to that of Lemma 3.1 and Theorem 3.4, we see that the second and third steps generate y such that $\|y - b^*\| \le \|y^* - b^*\| + M \max_{i=1,\ldots,N} \rho(\mathcal{Y}_i) = \operatorname{dist}(\mathcal{S}, \sum_{i=1}^N \operatorname{conv}(\mathcal{Y}_i)) + M \max_{i=1,\ldots,N} \rho(\mathcal{Y}_i) \le \operatorname{dist}(\mathcal{S}, \operatorname{conv}(\sum_{i=1}^N \mathcal{Y}_i)) + M \max_{i=1,\ldots,N} \rho(\mathcal{Y}_i)$, where the last inequality is due to the fact that $\sum_{i=1}^N \mathcal{Y}_i \subset \operatorname{conv}(\sum_{i=1}^N \mathcal{Y}_i) \subset \sum_{i=1}^N \operatorname{conv}(\mathcal{Y}_i)$.

(b) It follows from part (a) that the constructed solution $\check{x}$ satisfies the approximate optimality condition (6). Then the analysis of Theorems 3.5 and 3.6 follows directly. ∎

Implementation and computational complexity of Algorithm 1. We suppose that $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$ are finite sets. When they are not, we need to use finite discretization as an approximation. In the first step, computing b^* requires solving a least squares problem over linear inequality constraints. This can be solved efficiently using convex optimization solvers. When applying Algorithm 1 to problem (1), we take $\mathcal{S} = \partial f^\star(\mu^*)$. If f is a strictly convex function, the conjugate function $f^\star$ is differentiable. Then we can omit this step and simply take b^* to be the unique gradient $\nabla f^\star(\mu^*)$.

In the second step, we need to find a basic feasible solution to a linear program. To do this, one approach is to apply the simplex method and directly obtain an optimal basic feasible solution. Another approach is to apply the interior point algorithm [29] that finds a basic feasible solution to any linear program in polynomial time.

In the third step, we need to perform the nonconvex projection $\operatorname{argmin}_{y\in\mathcal{Y}_i} \|y - A^{(i)} z^{(i)}\|$ at most M times. This step can be interpreted as "each individual participant finds his closest feasible solution to the mixed decision $A^{(i)} z^{(i)}$." Since each individual set $\mathcal{Y}_i$ is a low-dimensional set, finding the nonconvex projection onto $\mathcal{Y}_i$ is considered computationally efficient and easy to conduct.

Alternatively, we can skip the nonconvex projection step and simply return an arbitrary $y_i \in \mathcal{Y}_i$. In particular, we may replace step 3 with the following:

$$y_i = \begin{cases} A^{(i)} z^{(i)} & \text{if } z^{(i)} \text{ is an integer vector,} \\ A_1^{(i)} & \text{otherwise.} \end{cases}$$

In this case, we obtain results analogous to Theorem 4.1 where the maximal convexity gap $\max_{i=1,\ldots,N} \rho(\mathcal{Y}_i)$ is replaced with the maximal diameter of $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$. When

applying the alternative algorithm to solve problem (1), we obtain a price of decentralization bounds similar to that of Theorems 3.5 and 3.6 with δ_p, δ_g replaced by γ_p, γ_g. In this way, the constructed solution is a slightly relaxed approximate global optimum. Its asymptotic price of decentralization remains of the same order.

When $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$ are infinite and compact sets, we need to use their finite discretizations as approximations. The cardinality of the finite discretization is often exponential with respect to the dimension of the set. Thus this approach is tractable only if each $\mathcal{Y}_i$ is small in dimension. When finite discretizations are used in Algorithm 1, the sum of discretization errors for $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$ should be added to the error bound of Theorem 4.1.

We have presented a computational method for finding an approximate global optimum out of many best response solutions, as long as $\mu^*, \check{\mathcal{X}}^*$ are given. It involves examining all candidate best responses in a centralized manner. The key step is to find an extreme point solution to a linear feasibility problem. It can be viewed as a form of coordination that selects a particular solution for each user in order to optimize the overall objective.

4.2. Price of decentralization without coordination. Recall that our main motivation is to answer the following question: Is there a fair price μ such that a global social optimum can be attained, as long as each participant reacts optimally to his own problem? So far, we have shown the dual optimal multiplier μ^* acts as a "nearly fair" price vector. We have shown the existence of a best response solution $\check{x} = (\check{x}_1, \ldots, \check{x}_N)$ to μ^* that is nearly primal optimal. Such a solution $\check{x}$ consists of individual best responses to the nearly fair price μ^*. As the number of participants increases, the normalized price of decentralization vanishes to zero, implying that $\check{x}$ becomes asymptotically primal optimal. This provides a partial answer to the earlier question.

Due to the lack of convexity, individual users may have multiple optimal actions to choose from even if they face the fair price μ^*. Any combination of individual decisions constitutes a best response solution to μ^*—there are many of them. It is not enough to know that one out of the large number of solutions achieves the small price of decentralization. We are also interested in how other solutions perform. Let us consider two examples.

Example 4.2 (efficiency loss without coordination). Consider the unconstrained optimization problem

$$\min_{x_1,\ldots,x_N} (x_1 + \cdots + x_N)^2 + |x_1^2 - 1| + \cdots + |x_N^2 - 1|.$$

The optimal multiplier to the dual problem is $\mu^* = 0$, and the corresponding solution sets are $\check{\mathcal{X}}_i^* = \{-1, 1\}$. Each user is indifferent about two equally optimal decisions -1 and 1. In the case without coordination, the objective value can be as large as $\mathcal{O}(N^2)$, while the optimal value is 1 if N is odd. In this case, without coordinating $x_1, \ldots, x_N$ to achieve the duality gap, the worst-case price of decentralization can be as large as $\mathcal{O}(N^2)$ and increases to infinity as N increases.

Example 4.3 (violation of feasibility without coordination). Consider the constrained optimization problem

$$\begin{aligned} &\min |x_1^2 - 1| + \cdots + |x_N^2 - 1| \\ &\text{subject to } |x_1 + \cdots + x_N| \le N/2. \end{aligned} \tag{12}$$

Similar to the earlier example, the optimal multiplier to the dual problem is $\mu^* = 0$, and the corresponding solution sets are $\check{\mathcal{X}}_i^* = \{-1, 1\}$. Again, each user is indifferent about two optimal decisions -1 and 1. Without coordination, the worst-case constraint violation can be as large as $N/2$. In this case, the price of decentralization does not improve as N increases.

As demonstrated in the examples, the price of decentralization can be far greater than the small upper bound if we let users choose their best responses arbitrarily. Although there exists a nearly optimal solution $\check{x}$ consisting of best responses to μ^*, there is no guarantee of reaching it without any coordination. Without convexity, a fully autonomous system will not work to its best capability. In short, coordination is critical to the success of multi-agent nonconvex optimization.

5. Dynamic convergence to approximate global optimum. In this section, we study how to dynamically coordinate individual participants to achieve an approximate global optimum of problem (1). We will develop a duality-based coordinative method, which is a semidecentralized method to be used by the central planner. The method relies on best responses from individual participants. Even without convexity, it achieves dynamic convergence to an approximate global optimum, as long as the participants are cooperative and a central decision maker properly coordinates the individuals' behaviors. We provide a rigorous convergence and rate of convergence analysis, and we comment on the algorithm's complexity.

5.1. A coordinative best response algorithm. We propose a coordinative algorithm for finding an approximate global solution of problem (1). The algorithm is based on a cutting plane method applied to the dual problem $\max_\mu Q(\mu)$. It is known that the cutting plane method works for the dual problem, because $Q(\cdot)$ is concave, regardless of the nonconvexity of the primal problem. However, even if an optimal multiplier μ^* is known, arbitrary best responses from individual participants may result in a high price of decentralization. This requires the algorithm to coordinate the participants in order to achieve an approximate social optimum. The coordination feature makes the proposed algorithm different from the traditional cutting plane method for convex optimization.

We present the coordinative dual algorithm in Algorithm 2. For simplicity, we use $\langle \cdot, \cdot \rangle$ to denote the inner product between two vectors, and we use $\epsilon\text{-argmin}_x f(x)$ to denote the set $\{x \mid f(x) \le \inf_z f(z) + \epsilon\}$.

Algorithm 2 involves a central planner that coordinates the behaviors of multiple participants. In each iteration, the coordinator updates the multiplier and an upper approximation of the dual function. Moreover, the coordinator collects the sets of selfish best responses from all participants and selects a particular decision for each of them. Steps 2–4 of Algorithm 2 are the dual updates and guarantee the convergence of the multiplier. Steps 5–10 are the coordination steps and select a nearly best solution out of the many candidates. The algorithm does not require any tuning stepsize, which makes it preferable to typical dual ascent methods.

Implementation and computational complexity of Algorithm 2. Algorithm 2 is a semidecentralized method. On one hand, it decomposes a complex task into multiple smaller-scale problems for the participants to solve on their own. On the other hand, it also involves centralized coordination and picks a particular best response for each participant. Step 2 of Algorithm 2 relies on individual participants to solve their own nonconvex problems. When the dimension M is a fixed small value, we assume that each individual's nonconvex problem is small-scale and can be solved efficiently in

Algorithm 2 Coordinative Multi-agent Nonconvex Optimization via Dual Cutting Plane

Input: $\mu^0 \in \Re^M, R \in \Re, \beta \in \Re$, precision parameter $\epsilon > 0$.

1: **repeat**

2: Given the price vector μ^t, each participant $i = 1, \ldots, N$ finds a best response

$$x_i^t \in \operatorname{argmin}\{p_i(x_i) + \langle \mu^t, g_i(x_i)\rangle \mid x_i \in \mathcal{X}_i\}.$$

3: The central planner updates the approximate dual problem by

$$Q_{t+1}(\mu) = \min_{k \le t} \left\{ \langle \mu, \bar{g}^k \rangle + \bar{p}^k - f^\star(\mu) \right\},$$

where $\bar{p}^t = \sum_{i=1}^N p_i(x_i^t), \bar{g}^t = \sum_{i=1}^N g_i(x_i^t)$.

4: The central planner sets the new test price and estimated gap as

$$\mu^{t+1} \in \operatorname*{argmax}_{\mu \in \Re^M} \; Q_{t+1}(\mu), \qquad \varepsilon_t = Q_t(\mu^t) - \max_{k \le t}\{\langle \mu^k, \bar{g}^k \rangle + \bar{p}^k - f^\star(\mu^k)\}.$$

5: **if** $\langle \mu^t, \bar{g}^t \rangle + \bar{p}^t - f^\star(\mu^t) > \max_{k \le t-1}\{\langle \mu^k, \bar{g}^k \rangle + \bar{p}^k - f^\star(\mu^k)\}$, **then**

6: Each participant $i = 1, \ldots, N$ finds the ξ_t-optimal responses

$$\mathcal{X}_i^t = \xi_t\text{-argmin}\{p_i(x_i) + \langle \mu^t, g_i(x_i)\rangle \mid x_i \in \mathcal{X}_i\}, \qquad \mathcal{G}_i^t = \{g_i(x_i) \mid x_i \in \mathcal{X}_i^t\},$$

where $\xi_t = R\sqrt{\beta \varepsilon_t}$.

7: The central planner applies Algorithm 1 with input $(\mathcal{G}_1^t, \ldots, \mathcal{G}_N^t, \partial f^\star(\mu^t))$ and obtains $y^t = (y_1^t, \ldots, y_n^t)$.

8: Set $\hat{\mu}^t = \mu^t$, $\hat{x}^t = (\hat{x}_1^t, \ldots, \hat{x}_n^t)$ such that $g_i(\hat{x}_i^t) = y_i^t$ for $i = 1, \ldots, N$.

9: **else**

10: Set $\hat{\mu}^t = \hat{\mu}^{t-1}$, $\hat{x}^t = \hat{x}^{t-1}$.

11: **end if**

12: **until** $\varepsilon_t \le \epsilon$.

constant time. Steps 3 and 4 require solving simple linear programs. Step 6 relies on each participant to return a set of near-optimal best responses, i.e., the ξ_t-argmin operation. This step also requires each participant to solve a small-scall nonconvex optimization and return a possibly infinite set. When the ξ_t-argmin set is infinite, one has to use discrete approximation instead and the discretization error will enter the final error bound. Step 7 calls the nonconvex projection Algorithm 1, where the major computation overhead is used to find the basic feasible solution of a linear program (see section 4.1 for discussions on its complexity). To sum up, when individual best responses are easy to compute, each iteration of Algorithm 2 mainly involves solving linear programs and is computationally efficient.

5.2. Convergence rate of the coordinative algorithm. We analyze the iteration complexity of Algorithm 2. We can prove the following convergence and rate of convergence results.

THEOREM 5.1. *Assume that f is α-strongly convex and β-strongly smooth. Let $R > 0$ be such that $\sup\{\|\sum_{i=1}^N g_i(x_i) - \sum_{i=1}^N g_i(y_i)\| \mid x, y \in \mathcal{X}\} \le R$. Then the following hold:*

(a) *The multiplier* $\hat{\mu}^t$ *converges to* μ^* *as* $t \to \infty$ *and satisfies for all* t *that*

$$Q^* - Q(\hat{\mu}^t) \leq \frac{\beta R^2 + \alpha^{-2}\beta^2(Q^* - Q(\mu^0))}{t/2}.$$

(b) *Every limit point of* $\hat{x}^t$ *is an approximate global optimum of problem* (1) *and satisfies for all* t *that*

$$F(\hat{x}^t) - F^* \leq \beta\left(M\delta_g\right)^2 + NR\sqrt{\frac{\beta^2R^2 + \alpha^{-2}\beta^3(Q^* - Q(\mu^0))}{t/2}} + \frac{\beta^3\alpha^{-2}R^2 + (Q^* - Q(\mu^0))}{t/2}.$$

(c) *With the stopping criterion* $\varepsilon_t \leq \epsilon$, *the algorithm terminates within* $T = \frac{\beta R^2 + \alpha^{-2}\beta^2(Q^* - Q(\mu^0))}{\epsilon/2}$ *iterations and satisfies*

$$F(\hat{x}^T) - F^* \leq \beta\left(M\delta_g\right)^2 + NR\sqrt{\beta\epsilon} + \frac{\beta^2}{\alpha^2}\epsilon. \tag{13}$$

We consider two typical choices of the precision parameter ϵ. Let us work out the iteration complexity in each of the cases.

(i) One may pick ϵ to be sufficiently small such that the error bound (13) is dominated by the constant gap $\beta\left(M\delta_g\right)^2$. In particular, it follows from Theorem 5.1 that

$$F(\hat{x}^T) - F^* \leq 2\beta\left(M\delta_g\right)^2$$

if ϵ is picked to satisfy $NR\sqrt{\beta\epsilon} + \frac{\beta^2}{\alpha^2}\epsilon \leq \beta\left(M\delta_g\right)^2$ and the number of iterations satisfies

$$T = \frac{\beta R^2 + \alpha^{-2}\beta^2(Q^* - Q(\mu^0))}{\min\{\alpha^2\beta^{-1}(M\delta_g)^2, \beta N^{-2}R^{-2}(M\delta_g)^4\}} = \Theta(N^2).$$

However, the preceding iteration complexity scales quadratically as N increases.

(ii) Alternatively, let us assume that $F^* \geq CN$ for some $C > 0$; in other words, the optimal value increases at least linearly as the number of participants N increases. In this case, one may pick an approximation ratio $1 + \varepsilon$ for some $\varepsilon \in (0, 1)$ a priori. Then one may pick $\epsilon = \mathcal{O}(C^2R^{-2}\beta^{-1}\varepsilon^2)$ such that the error bound (13) of Theorem 5.1 becomes

$$\frac{F(\hat{x}^T) - F^*}{F^*} \leq \varepsilon$$

as long as the number of iterations satisfies

$$T = \Theta\left(\frac{\beta R^2 C^{-2}(\beta R^2 + \alpha^{-2}\beta^2(Q^* - Q(\mu^0)))}{\varepsilon^2}\right).$$

As a result, Algorithm 2 yields an $(1 + \varepsilon)$-approximation to the global optimum using $\Theta(\frac{1}{\varepsilon^2})$ iterations. Note that this approximation guarantee and the corresponding iteration complexity are *invariant with respect to* N.

To sum up, Theorem 5.1 asserts that the nonconvex N-agent problem (1) can be solved efficiently up to a certain approximation guarantee. According to case (ii), the approximation ratio can be made arbitrarily close to 1, where the iteration complexity is invariant with the number of participants N. This makes it possible to solve large-scale coordinative optimization problems using individual best responses.

Due to nonconvexity, the primal and dual convergences of Algorithm 2 result from different mechanisms. The formal proof of Theorem 5.1 is developed through a series of lemmas, which are deferred to the appendices.

The dual convergence $\mu^t \to \mu^*$ is guaranteed by the concavity of the dual problem. Note that the dual problem is almost always nonsmooth because it is the pointwise minimum of a number of functions. We remark that our dual convergence result applies to the more general nonsmooth convex optimization problem

$$\min_{w} \left\{ \rho(w) + \max_{\theta \in \Theta} \ell(w; \theta) \right\}, \tag{14}$$

where $\rho(\cdot)$ is a strongly convex regularization function and $\ell(w; \theta)$ is a convex loss function in w for all θ. Problem (14) is very common in empirical risk minimization; see, e.g., [43] for an application in machine learning. In Theorem 5.1 part (a), we show that the cutting plane method converges at a rate of $\mathcal{O}(1/t)$ for problem (14). This result contains the earlier results in [43] as a special case. In fact, it has been shown in [50] that the $\mathcal{O}(1/t)$ convergence rate of the cutting plane method is nonimprovable for problem (14).

The primal convergence is not automatically guaranteed by the dual convergence. Without the coordination steps, the primal functions may not converge at all, even if $\mu^t \to \mu^*$. There are two difficulties. First, even if μ^* is known, finding the approximate optimum requires the selection of a good solution out of a large number of best responses. Otherwise, as demonstrated in section 4.2, the price of decentralization can be very high. Second, the set of best responses is not continuous with respect to the multiplier. When $\mu^t \to \mu^*$ but $\mu^t \neq \mu^*$ for all t, it is possible that many best responses to μ^* are never best responses to any μ^t. This suggests that the primal solutions will never be close to optimal, even if $\mu^t \to \mu^*$. To induce convergence, we require that the participants submit their ξ_t-optimal responses. Here ξ_t is a diminishing error tolerance that induces continuity in the nearly best response set with respect to the multiplier μ^t. After a careful balance between the error tolerance and the convergence error, we can construct a sequence of $\hat{x}^t$ that converges into the approximate social optima.

Other than the cutting plane method, alternative dual methods may apply too. One example is the dual ascent method, which only requires the dual subgradient and does not require the dual function value. This means that participants only need to report their $g_i(x_i)$ values, i.e., their individual impacts on the public goods. They can keep their preference values $p_i(x_i)$ private. However, use of the dual ascent method may result in slower convergence and additional tuning stepsizes. Another open question is how to solve problem (1) when f is not strongly convex/smooth. An important special case is where f is the indicator function. We conjecture that duality-based methods will still work but the rate of convergence will deteriorate. These topics are left for future research.

5.3. Discussions on the complexity of problem (1). Let us comment on the complexity of the nonconvex problem (1). Although problem (1) is continuous, it

bears strong connection to the multirow knapsack problem given by

$$\min c^T x, \qquad \text{subject to} \quad Ax \le b,\ x_i = \{0,1\},\ i = 1,\dots,n,$$

where $c \in \Re^n$, $A \in \Re^{m\times n}$, and $b \in \Re^m$. In fact, we believe that they admit polynomial-time reductions to each other. Let us describe the main ideas without getting into technical details.

On one hand, the knapsack problem is a special case of problem (1). To see this, we let p_i and g_i be linear functions determined by c_i and the ith column of A, respectively. Also we let $\mathcal{X}_i = \{0,1\}$ and $\mathcal{A} = \{y \le b\}$. In this way, we have constructed an instance of the constrained problem (2) that is equivalent to the knapsack problem. Note that problem (2) is a special case of problem (1). So we have obtained an immediate reduction from the knapsack problem to problem (1).

On the other hand, problem (1) can be approximated by the knapsack problem via discretization. Given a problem instance with functions $p_i, g_i, \mathcal{X}_i, f$, we add a dummy variable to get a constrained problem

$$\text{minimize } \sum_{i=1}^N p_i(x_i) + f\left(z\right), \quad \text{subject to } \sum_{i=1}^N g_i(x_i) = z, \quad x_i \in \mathcal{X}_i, \quad i = 1,\dots,N.$$

We discretize the functions p_i, f into vectors, the mappings g_i into matrices, and the sets $\mathcal{X}_i$ into grid points. Then we change the variables from x_i to indices of the discrete vectors. The resulting problem is a multiple-choice multirow knapsack problem, which can be further reduced to the multirow $\{0,1\}$-knapsack problem. Under additional continuity assumptions on the functions p_i, g_i, f, we conjecture that the reduction is polynomial-time. Note that the discretization disregards any structure of the continuous functions and results in a huge-dimensional discrete problem. So it is not a practical computation method.

The two-way reduction implies that problem (1) and the knapsack problem belong to the same complexity class (requiring a rigorous definition of complexity of continuous problem). The knapsack problem is known to be $\mathcal{NP}$-complete. When it has at least two row constraints, the knapsack problem admits no fully polynomial-time approximation scheme unless $\mathcal{P} = \mathcal{NP}$ (see [28]). Because problem (1) contains the knapsack problem as a special case, the more general problem (1) is at least as hard. In other words, we cannot develop an efficient algorithm to achieve an ϵ-optimal solution to problem (1) for arbitrarily small $\epsilon > 0$. In contrast, we have developed Algorithm 2, which finds a constant-error approximate solution in a polynomial number of iterations. The constant error is the price of decentralization that diminishes to zero as N/M increases. A rigorous complexity analysis and proof of reduction are beyond the scope of this work but will be an interesting topic for future investigation.

6. Numerical experiments. We study the performance of Algorithm 2 and the asymptotic price of decentralization by generating random instances of problem (1). We let p_i, g_i be piecewise constant functions defined on $[0,1]$, where the jump points are independent random vectors generated from the multivariate uniform distribution for all $i = 1,\dots,N$. We let $f(\cdot)$ be a quadratic function, where the Hessian is a random symmetric and positive definite matrix with mean $I_{N\times N}$. According to our random generation, the maximal convexity gaps δ_p, δ_g are bounded for all N with probability 1.

For each sample instance of problem (1), we apply Algorithm 2 to find an approximate optimal value $F(\check{x})$. The primal and dual trajectories are plotted in Figure 2.

We observe that the dual variables converge very quickly, usually within 20–50 iterations. In addition, the dual iterate seems to converge at a geometric rate during the initial iterations and slow down afterward. This is consistent with our analysis that the dual error diminishes according to $\varepsilon_{t+1} \leq \varepsilon_t - \frac{\varepsilon_t^2}{\beta R^2 + \alpha^{-2}\beta^2 \varepsilon_t}$, which is a contraction when the error is large. On the other hand, the primal convergence is much slower and has a constant error due to the convexity gap. By comparing the trajectories for different values of N, M, we see that the normalized price of decentralization becomes small when the ratio N/M is large. We also notice that the convergence is faster when M is small. This is due to the random generation of the problem, because problems with small M are more likely to be well conditioned (resulting in small β/α).

For comparison, we compute the global optimum x^* and the optimal value F^* by discretizing the optimization problem and using an exhaustive search. We compute the price of decentralization $\frac{F(\tilde{x}) - F^*}{F^*}$ for each sampled problem, where N varies and M is fixed. The samples and means of the price of decentralization are illustrated in Figure 3. We observe that the price of decentralization is very close to zero with high probability. Moreover, the inverse of the mean price of decentralization is linearly related to the dimension N. This validates our theory that the price of decentralization vanishes to 0 at a rate of $\mathcal{O}(1/N)$ as the number of participants increases.

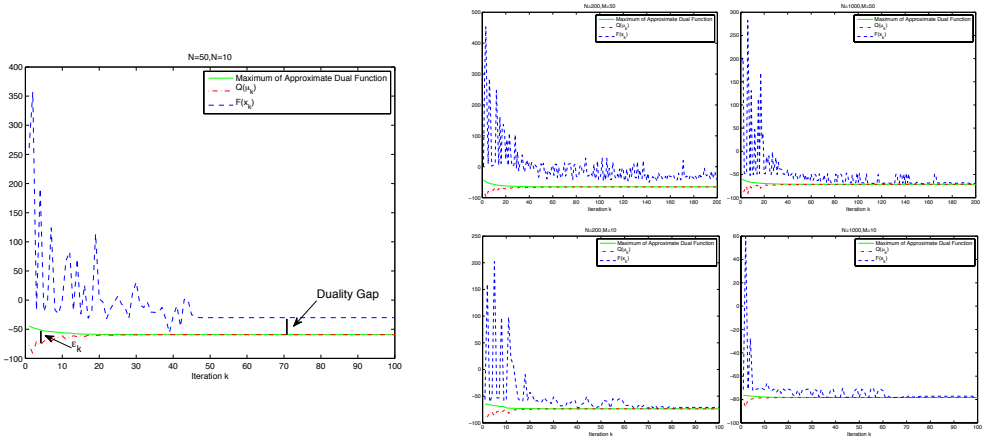

FIG. 2. *Primal and dual convergence of Algorithm* 2. *We plot trajectories of the primal objective values, dual objective values, and upper bounds of dual objective values generated by Algorithm* 2. *The estimated dual convergence error* ϵ_t *diminishes to* 0 *very quickly, while the primal objective value converges to a constant gap.*

7. Concluding remarks. We have studied a nonconvex coordinative optimization problem where N participants minimize a common objective. We consider a Fenchel dual of the nonconvex problem in order to decompose it with respect to the individual decisions. We show that the dual problem becomes increasingly convex in a geometric sense as N increases. We have mathematically characterized the duality gap and the price of decentralization. We prove that the minimal price of decentralization, which is due to the nonconvexity, asymptotically diminishes to zero as N grows.

Algorithmically, we show how to achieve the approximate social optimum using a coordination procedure. Without coordination, we show by examples that the price

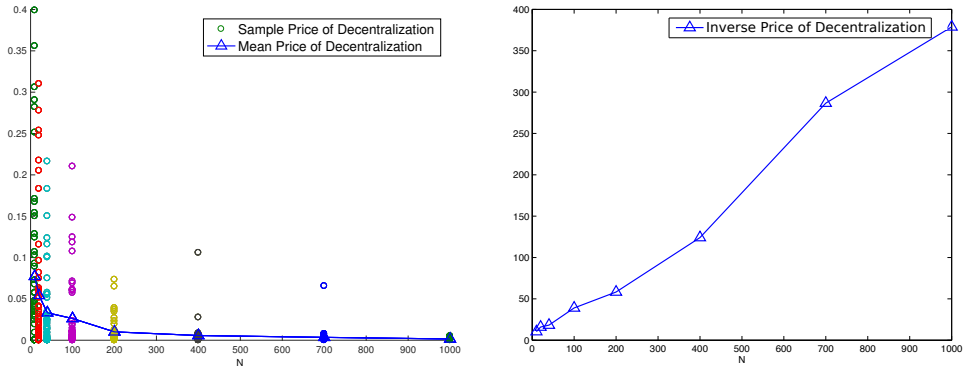

FIG. 3. *Asymptotic price of decentralization. We generate* 50 *random instances of problem* (1) *for each value of* N *and find the approximate optimum using Algorithm* 2. *In the left figure, we plot the samples and means of the price of decentralization* $\frac{F(\hat{x})-F^*}{F^*}$ *for different values of* N. *In the right figure, we plot the inverse of the mean price of decentralization against the values of* N.

of decentralization can be arbitrarily high. We propose a duality-based coordinative algorithm that relies on individual best responses to find the approximate optimum. The algorithm has an important coordination step, which involves finding an extreme point of a linear feasibility problem. It ensures that both the primal and dual variables are convergent, regardless of the initial solution. We analyze the convergence and rate of convergence of the proposed algorithm, and we test the algorithm on randomly generated examples.

Finally, we are able to answer the question raised in section 1: Is there a fair price such that individual best responses automatically achieve the social optimum? The answer is largely yes. Indeed, there exists a nearly fair price at which a particular best response solution is an approximate social optimum. Such an approximate optimum can only be obtained via coordination.

In future work, a theoretical challenge is to generalize the results to more abstract settings, such as the minimax problems and variational inequalities. Another potential topic of future work is the design and study of more efficient algorithms with better error-complexity guarantees. From the practical perspective, an important future task is to identify practical instances of multi-agent optimization and tailor the analysis and algorithms to specific applications.

Appendix A. Proofs of technical lemmas.

Proof of Lemma 3.1. We start with the case where $\mathcal{S}_1 + \cdots + \mathcal{S}_n$ is a finite set, which implies that each $\mathcal{S}_i$ is also a finite set.

We let $A^{(i)}$ be the matrix whose columns are equal to vectors in $\mathcal{S}_i$ and let $e^{(i)}$ be the unit vector whose dimension is equal to the column dimension of $A^{(i)}$, where $i = 1, \ldots, n$. Consider the following linear feasibility problem for variables $z^{(i)}, i = 1, \ldots, n$:

$$\sum_{i=1}^{n} A^{(i)} z^{(i)} = x, \quad \left(e^{(i)}\right)^T z^{(i)} = 1, \quad z^{(i)} \geq 0, \quad i = 1, \ldots, n. \tag{15}$$

We see that (15) takes the standard form of linear programming and has $m + n$

equalities (m from the row dimension of $A(i)$). We claim that there exists at least one feasible solution. Since $x \in \mathrm{conv}(\mathcal{S}_1 + \cdots + \mathcal{S}_n)$, there exists $\lambda_1, \ldots, \lambda_m$ such that

$$x = \sum_{j=1}^{m} \lambda_j \left(\sum_{i=1}^{n} x_{ij} \right), \qquad \text{where } x_{ij} \in \mathcal{S}_i,$$

and $\sum_{j=1}^{m} \lambda_j = 1, \lambda_j \geq 0$ for all $i = 1, \ldots, n, j = 1, \ldots, m$. We let $z^{(i)}$ be the vector whose entry takes value λ_j when the corresponding column in $A^{(i)}$ is equal to x_{ij} for some j (there might be multiple values of such j) and takes value zero otherwise. We can verify that the constructed solution $(z^{(1)}, \ldots, z^{(n)})$ is a feasible solution to (15).

Because (15) is feasible and has $m + n$ equality constraints, there exists a basic feasible solution $z = (z^{(1)}, \ldots, z^{(n)})$ to (15) with at most $m + n$ nonzero elements (see [45]). According to the constraint $(e^{(i)})^T z^{(i)} = 1, z^{(i)} \geq 0$, each $z^{(i)}$ has at least one nonzero element. This implies that $z = (z^{(1)}, \ldots, z^{(n)})$ has at most m components $z^{(i)}$ with more than one nonzero entry. We take $x_i = A^{(i)} z^{(i)}$ for each i, and $\mathcal{I} = \{i \mid \|z^{(i)}\|_0 > 1\}$. Clearly, we have $x_i \in \mathcal{S}_i$ if $i \notin \mathcal{I}$ and $x_i \in \mathrm{conv}(\mathcal{S}_i)/\mathcal{S}_i$ if $i \in \mathcal{I}$. Moreover, the cardinality of the set $\mathcal{I}$ is at most m.

Finally, we consider the general case where $\mathcal{S}_1 + \cdots + \mathcal{S}_n$ is an infinite set. By Carathéodory's theorem, any point $x \in \mathrm{conv}(\mathcal{S}_1 + \cdots + \mathcal{S}_n)$ can be represented by the convex combination of at most $m + 1$ points in $\mathcal{S}_1 + \cdots + \mathcal{S}_n$. Then it is sufficient to focus on the remaining $m + 1$ points and apply the preceding analysis. ∎

LEMMA A.1. *Let $f_1, f_2 : \Re^n \mapsto \Re$ be continuous σ-strongly convex functions, and let $\gamma \in \Re, g \in \Re^n$ be such that $\min_w f_1(w) = f_1(0) = 0, f_2(0) = \gamma, g \in \partial f_2(0)$. Then*

$$\min_w \max\{f_1(w), f_2(w)\} \geq \max\left\{ \gamma - \frac{\|g\|^2}{2\sigma}, \frac{2\gamma^2\sigma}{\|g\|^2} \right\}.$$

Proof of Lemma A.1. We use the minimax duality for the convex-concave function to obtain

$$\begin{aligned} \min_w \max\{f_1(w), f_2(w)\} &= \min_w \max_{\lambda \in [0,1]} (1-\lambda) f_1(w) + \lambda f_2(w) \\ &= \max_{\lambda \in [0,1]} \min_w (1-\lambda) f_1(w) + \lambda f_2(w). \end{aligned} \tag{16}$$

By the assumptions, we obtain that $(1-\lambda)f_1(w) + \lambda f_2(w)$ is also σ-strongly convex, and moreover, $(1-\lambda)f_1(0) + \lambda f_2(0) = \lambda\gamma$, and $\lambda g \in \partial\big((1-\lambda)f_1 + \lambda f_2\big)(0)$. By using the definition of strongly convex functions, we obtain for all w that

$$(1-\lambda) f_1(w) + \lambda f_2(w) \geq \frac{\sigma}{2}\|w\|^2 + \lambda g^T w + \lambda\gamma.$$

By applying the preceding relation to the minimax equality (16) and using basic calculations, we obtain

$$\begin{aligned} \min_w \max\{f_1(w), f_2(w)\} &\geq \max_{\lambda \in [0,1]} \min_w \frac{\sigma}{2}\|w\|^2 + \lambda g^T w + \lambda\gamma \geq \max_{\lambda \in [0,1]} \lambda\gamma - \frac{\lambda^2\|g\|^2}{2\sigma} \\ &\geq \max\left\{ \gamma - \frac{\|g\|^2}{2\sigma}, \frac{2\gamma^2\sigma}{\|g\|^2} \right\}. \end{aligned}$$ ∎

LEMMA A.2. *Let $\{a_t\}$ be a sequence of positive scalars, and let η be a positive scalar, such that $a_{t+1} \leq a_t - \eta a_t^2$ for all $t \geq 0$ and $a_0 \leq \frac{1}{2\eta}$. Then $a_t \leq \frac{a_0}{\eta a_0 t + 1}$ for all $t \geq 0$.*

Proof of Lemma A.2. We prove by induction. The statement is clearly true when $t = 0$. Now suppose that it is true for some $t > 0$. Then we have

$$\begin{aligned} a_{t+1} &\le a_t - \eta a_t^2 \le \frac{a_0}{\eta a_0 t + 1} - \eta \left(\frac{a_0}{\eta a_0 t + 1} \right)^2 \\ &= a_0 \cdot \frac{\eta a_0 (t-1) + 1}{(\eta t a_0 + 1)^2} \le a_0 \cdot \frac{1}{\eta (t+1) a_0 + 1}, \end{aligned}$$

where the second inequality uses the monotonicity of $f(x) = x - \eta x^2$ when $x \le \frac{1}{2\eta}$ and the fact $a_t \le a_0 \le \frac{1}{2\eta}$. Thus, we have shown by induction that the inequality holds for all $t \ge 0$. ∎

Appendix B. Proof of convergence for Algorithm 2. This subsection is devoted to the convergence analysis of Algorithm 2. For readers who are not interested, it can be safely skipped. The proof of Theorem 5.1 is developed through the series of lemmas stated and proved in Appendix A.

The complete proof of Theorem 5.1 is divided into two parts: the dual convergence and the primal convergence. We remark that the dual convergence analysis of the cutting plane method applies to the more general risk minimization problem (14), where the objective is the sum of a strongly convex function and the maximum of a number of convex functions.

Proof of Theorem 5.1(a). Since f is α-strongly convex and β-strongly smooth, it follows from Lemma 3.2 that $f^\star$ has $1/\alpha$-Lipschitz continuous gradient and is $1/\beta$-strongly convex. It follows that $Q(\mu), Q_t(\mu), L(x, \mu)$ are all $1/\beta$-strongly concave functions with respect to μ.

According to the update rule of Q_t, we have

$$Q_{t+1}(\mu) = \min\{Q_t(\mu), \langle \mu, \bar{g}^t \rangle + \bar{p}^t - f^\star(\mu)\} = \min\{Q_t(\mu), L(x^t, \mu)\}.$$

Note that each $L(x^t, \mu)$ is an upper bound on the dual function $Q(\mu)$. Therefore, Q_t is a decreasing sequence of functions which approximate Q from above; i.e., $Q(\mu) \le Q_{t+1}(\mu) \le Q_t(\mu)$ for all t and μ. According to the update rule of μ^t, we have $Q_t(\mu^t) = \max_\mu Q_t(\mu) \ge \max_\mu Q(\mu) = Q(\mu^*)$ for all t.

Recall that the estimated gap at time t is

$$\varepsilon_t = Q_t(\mu^t) - \max_{k \le t} \{\langle \mu^k, \bar{g}^k \rangle + \bar{p}^k - f^\star(\mu^k)\} = Q_t(\mu^t) - \max_{k \le t} Q(\mu^k).$$

We have

$$\begin{aligned} \varepsilon_{t+1} - \varepsilon_t &= Q_{t+1}(\mu^{t+1}) - \max_{k \le t+1} Q(\mu^k) - (Q_t(\mu^t) - \max_{k \le t} Q(\mu^k)) \\ &\le Q_{t+1}(\mu^{t+1}) - Q_t(\mu^t) \\ &= \max_\mu Q_{t+1}(\mu) - Q_t(\mu^t) \\ &= \max_\mu \min\{Q_t(\mu), L(x^t, \mu)\} - Q_t(\mu^t), \end{aligned} \tag{17}$$

where the inequality uses the fact that $\max_{k \le t} Q(\mu^k) \le \max_{k \le t+1} Q(\mu^k)$, the second equality uses the definition of μ^t, and the last relation uses the definition of Q_{t+1}.

Note that $\max_\mu Q_t(\mu) = Q_t(\mu^t)$ and $\bar{g}_t - \nabla f^\star(\mu^t) \in \partial_\mu L(x^t, \mu^t)$. Moreover, $L(x^t, \mu)$ and $Q_t(\mu)$ are $1/\beta$-strong concave with respect to μ. Therefore, we are

able to apply Lemma A.1, yielding

$$\min\{Q_t(\mu), L(x^t,\mu)\} - Q_t(\mu^t) \le -\frac{2(L(x^t,\mu^t) - Q_t(\mu^t))^2}{\beta\|\bar{g}_t - \nabla f^\star(\mu^t)\|^2}. \tag{18}$$

Note that $L(x^t,\mu^t) - Q(\mu^t) = Q(\mu^t) - Q_t(\mu^t) \le \max_{k\le t} Q(\mu^k) - Q_t(\mu^t) = -\varepsilon_t$. Also note that

$$\|\bar{g}_t - \nabla f^\star(\mu^t)\| \le \|\bar{g}_t - \nabla f^\star(\mu^*)\| + \|\nabla f^\star(\mu^*) - \nabla f^\star(\mu^t)\| \le R + \frac{\sqrt{\beta\epsilon_t}}{\alpha},$$

where $\|\bar{g}_t - \nabla f^\star(\mu^*)\| \le R$ because $\nabla f^\star(\mu^*)$ is a convex combination of $\sum_{i=1}^N g_i(x_i)$, the set $\{\sum_{i=1}^N g_i(x_i) \mid x \in \mathcal{X}\}$ has a radius smaller than R, and $\|\nabla f^\star(\mu^*) - \nabla f^\star(\mu^t)\| \le \frac{\sqrt{\beta\epsilon_t}}{\alpha}$ because of the strong convexity and strong smoothness assumptions. So it follows from (18) that

$$\min\{Q_t(\mu), L(x^t,\mu)\} - Q_t(\mu^t) \le -\frac{2\varepsilon_t^2}{\beta(R + \frac{\sqrt{\beta\epsilon_t}}{\alpha})^2} \le -\frac{\varepsilon_t^2}{2(\beta R^2 + \alpha^{-2}\beta^2\varepsilon_t)}. \tag{19}$$

By combining (17) and (19), we obtain for all $t \ge 0$

$$\varepsilon_{t+1} \le \varepsilon_t - \frac{\varepsilon_t^2}{2(\beta R^2 + \alpha^{-2}\beta^2\varepsilon_t)} \le \varepsilon_t - \frac{\varepsilon_t^2}{2(\beta R^2 + \alpha^{-2}\beta^2\varepsilon_0)}.$$

Note that $\varepsilon_t \le \varepsilon_0 \le \alpha^{-2}\beta^2\varepsilon_0$. So we apply Lemma A.2 and get $\varepsilon_t \le \frac{\beta R^2 + \alpha^{-2}\beta^2\varepsilon_0}{t/2 + (\beta R^2 + \alpha^{-2}\beta^2\varepsilon_0)\beta\varepsilon_0^{-1}}$ $\le \frac{\beta R^2 + \alpha^{-2}\beta^2\varepsilon_0}{t/2}$ for all $t \ge 0$. Note that $Q^* = \max_\mu Q(\mu) \le \max_\mu Q_t(\mu) = Q_t(\mu^t)$ and $\max_{k\le t} Q(\mu^k) = Q(\hat{\mu}^t)$. So we have

$$Q^* - Q(\hat{\mu}^t) = Q^* - \max_{k\le t} Q(\mu^k) \le Q_t(\mu^t) - \max_{k\le t} Q(\mu^k) = \varepsilon_t \le \frac{\beta R^2 + \alpha^{-2}\beta^2\varepsilon_0}{t/2}.$$

Finally, we use the definition of $\hat{\mu}^t$ and the $1/\beta$-strong concavity of Q to obtain

$$\|\mu^* - \hat{\mu}^t\|^2 \le \beta\left(Q^* - Q(\hat{\mu}^t)\right) \le \beta\varepsilon_t \le \frac{\beta^2 R^2 + \alpha^{-2}\beta^3\varepsilon_0}{t/2}.$$

Thus, we have shown that $\hat{\mu}^t$ converges to μ^* as $t \to \infty$ at a rate of $\|\hat{\mu}^t - \mu^*\|^2 = \mathcal{O}(R^2\beta^2/t)$. ∎

Now we consider the convergence of the primal function values.

Proof of Theorem 5.1(b),(c). Note that $L(x,\mu)$ is $1/\beta$-strongly concave with respect to μ and that $\sum_{i=1}^N g_i(x_i) - \nabla f^\star(\hat{\mu}^t) \in \partial_\mu L(x,\hat{\mu}^t)$ for all x. So we have for all x,μ that

$$L(x,\mu) \le L(x,\hat{\mu}^t) + \left\langle \sum_{i=1}^N g_i(x_i) - \nabla f^\star(\hat{\mu}^t), \mu - \hat{\mu}^t \right\rangle - \frac{1}{2\beta}\|\mu - \hat{\mu}^t\|^2.$$

Letting $x = \hat{x}^t$ and taking the supreme over μ, we have

$$F(\hat{x}^t) = \sup_\mu L(\hat{x}^t,\mu) \le L(\hat{x}^t,\hat{\mu}^t) + \sup_\mu \left\{ \left\langle \sum_{i=1}^N g_i(\hat{x}_i^t) - \nabla f^\star(\hat{\mu}^t), \mu - \mu^t \right\rangle - \frac{1}{2\beta}\|\mu - \hat{\mu}^t\|^2 \right\}.$$

By the definition of $\hat{x}^t$ in Step 6 of Algorithm 2, we have

$$L(\hat{x}^t, \hat{\mu}^t) \le \inf_{x \in \mathcal{X}} L(x, \hat{\mu}^t) + N\xi_t = Q(\hat{\mu}^t) + N\xi_t.$$

We also have $Q(\hat{\mu}^t) \le \max_\mu Q(\mu) = Q^* \le F^*$ and $\langle \sum_{i=1}^N g_i(\hat{x}_i^t) - \nabla f^\star(\hat{\mu}^t), \mu - \mu^t \rangle - \frac{1}{2\beta}\|\mu - \hat{\mu}^t\|^2 \le \frac{\beta}{2}\|\sum_{i=1}^N g_i(\hat{x}_i^t) - \nabla f^\star(\hat{\mu}^t)\|^2$ for all μ. It follows from the preceding relations that

$$F(\hat{x}^t) \le F^* + N\xi_t + \frac{\beta}{2}\left\| \sum_{i=1}^N g_i(\hat{x}_i^t) - \nabla f^\star(\hat{\mu}^t) \right\|^2. \tag{20}$$

It remains to show that $\|\sum_{i=1}^N g_i(\hat{x}_i^t) - \nabla f^\star(\hat{\mu}^t)\|^2$ diminishes to zero at a suitable rate.

We claim that $\check{\mathcal{X}}_i^* \subset \mathcal{X}_i^t$ for all $i = 1, \dots, N, t \ge 0$. To see this, we let x_i^* be any best response by the ith participant to the optimal multiplier μ^* and let x_i^t be any best response by the ith participant to the optimal multiplier $\hat{\mu}^t$. We have

$$\begin{aligned} p_i(x_i^*) + \langle \hat{\mu}^t, g_i(x_i^*) \rangle &= p_i(x_i^*) + \langle \mu^*, g_i(x_i^*) \rangle + \langle \hat{\mu}^t - \mu^*, g_i(x_i^*) \rangle \\ &\le p_i(x_i^t) + \langle \mu^*, g_i(x_i^t) \rangle + \langle \hat{\mu}^t - \mu^*, g_i(x_i^*) \rangle \\ &= p_i(x_i^t) + \langle \hat{\mu}^t, g_i(x_i^t) \rangle + \langle \hat{\mu}^t - \mu^*, g_i(x_i^*) - g_i(x_i^t) \rangle \\ &= \min_x \{ p_i(x) + \langle \hat{\mu}^t, g_i(x) \rangle \} + \langle \hat{\mu}^t - \mu^*, g_i(x_i^*) - g_i(x_i^t) \rangle, \end{aligned}$$

where the inequality uses the optimality of x_i^*, and the last relation uses the optimality of x_i^t. By using the bounded radius R of $\{g_i(x_i) \mid x_i \in \mathcal{X}_i\}$, we also have

$$\langle \hat{\mu}^t - \mu^*, g_i(x_i^*) - g_i(x_i^t) \rangle \le \|\hat{\mu}^t - \mu^*\| \cdot \|g_i(x_i^*) - g_i(x_i^t)\| \le \sqrt{\beta\varepsilon_t} \cdot R = \xi_t.$$

According to the definition of $\mathcal{X}_i^t$, it provides sufficient error tolerance so that $x_i^* \in \mathcal{X}_i^t$ for all t. Thus we have proved the claim.

Because $\check{\mathcal{X}}_i^* \subset \mathcal{X}_i^t$, we have

$$\mathcal{G}_i^* = \{ g_i(x_i) \mid x_i \in \check{\mathcal{X}}_i^* \} \subset \mathcal{G}_i^t, \qquad i = 1, \dots, N.$$

So we have $\nabla f^\star(\mu^*) \in \operatorname{conv}(\mathcal{G}_1^* + \cdots + \mathcal{G}_N^*) \subset \operatorname{conv}(\mathcal{G}_1^t + \cdots + \mathcal{G}_N^t)$. According to the construction of $\hat{x}^t$, we use Theorem 4.1 to obtain

$$\begin{aligned} \left\| \sum_{i=1}^N g_i(\hat{x}_i^t) - \nabla f^\star(\hat{\mu}^t) \right\| &\le \min_{y \in \operatorname{conv}(\mathcal{G}_1^t + \cdots + \mathcal{G}_N^t)} \|y - \nabla f^\star(\hat{\mu}^t)\| + M\delta_g \\ &\le \|\nabla f^\star(\mu^*) - \nabla f^\star(\hat{\mu}^t)\| + M\delta_g. \end{aligned}$$

By using the $1/\alpha$-Lipschitz continuity of $\nabla f^\star$ and the $1/\beta$-strong convexity of f, we have $\|\nabla f^\star(\mu^*) - \nabla f^\star(\hat{\mu}^t)\| \le \frac{1}{\alpha}\|\hat{\mu}^t - \mu^*\| \le \frac{\sqrt{\beta\varepsilon_t}}{\alpha}$. Applying the preceding two inequalities to (20), we obtain that

$$F(\hat{x}^t) - F^* \le N\xi_t + \frac{\beta}{2}\left\| \sum_{i=1}^N g_i(\hat{x}_i^t) - \nabla f^\star(\hat{\mu}^t) \right\|^2 \le NR\sqrt{\beta\varepsilon_t} + \beta\left(M\delta_g\right)^2 + \frac{\beta^2}{\alpha^2}\varepsilon_t.$$

Since F is continuous and $\varepsilon_t \to 0$, any limit point of $\hat{x}^t$ is an approximate optimum. Finally, we apply part (a) and use $\varepsilon_t \le \frac{\beta R^2 + \alpha^{-2}\beta^2\varepsilon_0}{t/2}$. Then we obtain

$$F(\hat{x}^t) - F^* \le NR\sqrt{\frac{\beta^2 R^2 + \alpha^{-2}\beta^3(Q^* - Q(\mu^0))}{t/2}} + \beta\left(M\delta_g\right)^2 + \frac{\beta^3\alpha^{-2}R^2 + (Q^* - Q(\mu^0))}{t/2},$$

and parts (b) and (c) follow immediately. ∎

REFERENCES

[1] P. Abichandani, H. Y. Benson, and M. Kam, *Decentralized multi-vehicle path coordination under communication constraints*, in Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Washington, DC, 2011, pp. 2306–2313.

[2] K. J. Arrow, F. Hahn, et al., *General Competitive Analysis*, Holden-Day, San Francisco, CA, 1971.

[3] I. Atzeni, L. G. Ordóñez, G. Scutari, D. P. Palomar, and J. R. Fonollosa, *Noncooperative and cooperative optimization of distributed energy generation and storage in the demand-side of the smart grid*, IEEE Trans. Signal Process., 61 (2013), pp. 2454–2472.

[4] J.-P. Aubin and I. Ekeland, *Estimates of the duality gap in nonconvex optimization*, Math. Oper. Res., 1 (1976), pp. 225–245.

[5] J. Bento, N. Derbinsky, J. Alonso-Mora, and J. S. Yedidia, *A message-passing algorithm for multi-agent trajectory planning*, in Proceedings of the Conference on Advances in Neural Information Processing Systems, Lake Tahoe, NV, 2013, pp. 521–529.

[6] D. Bertsekas, *Convexification procedures and decomposition methods for nonconvex optimization problems*, J. Optim. Theory Appl., 29 (1979), pp. 169–197.

[7] D. Bertsekas, G. Lauer, N. Sandell, Jr., and T. A. Posbergh, *Optimal short-term scheduling of large-scale power systems*, IEEE Trans. Automat. Control, 28 (1983), pp. 1–11.

[8] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.

[9] D. P. Bertsekas, *Convex Optimization Theory*, Athena Scientific, Belmont, MA, 2009.

[10] D. P. Bertsekas, A. Nedić, and A. Ozdaglar, *Min common/max crossing duality: A simple geometric framework for convex optimization and minimax theory*, Rep. LIDS-P-2536, 2002; available online from http://www.ifp.illinois.edu/~angelia/min_common.pdf.

[11] D. P. Bertsekas and N. Sandell, *Estimates of the duality gap for large-scale separable nonconvex optimization problems*, in Proceedings of the 21st IEEE Conference on Decision and Control (Orlando, FL, 1982), IEEE, Washington, DC, 1982, pp. 782–785.

[12] E. Budish, *The combinatorial assignment problem: Approximate competitive equilibrium from equal incomes*, J. Political Economy, 119 (2011), pp. 1061–1103.

[13] G. Carmona and K. Podczeck, *On the existence of pure-strategy equilibria in large games*, J. Econom. Theory, 144 (2009), pp. 1300–1319.

[14] 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, Cambridge, UK, 1975, pp. 433–436.

[15] R. Cerf, *Large deviations for sums of i.i.d. random compact sets*, Proc. Amer. Math. Soc., 127 (1999), pp. 2431–2436.

[16] N. Dyn and E. Farkhi, *Set-valued approximations with Minkowski averages—convergence and convexification rates*, Numer. Funct. Anal. Optim., 25 (2005), pp. 363–377.

[17] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, corrected reprint of the 1976 English ed., Classics Appl. Math. 28, SIAM, Philadelphia, 1999, https://doi.org/10.1137/1.9781611971088.

[18] I. V. Evstigneev and S. Flam, *Sharing nonconvex costs*, J. Global Optim., 20 (2001), pp. 257–271.

[19] E. X. Fang, L. Han, and W. Mengdi, *Blessing of Massive Scale: Spatial Graphical Model Inference with a Total Cardinality Constraint*, Working paper, 2015.

[20] P. Fiorini and Z. Shiller, *Motion planning in dynamic environments using velocity obstacles*, Internat. J. Robotics Res., 17 (1998), pp. 760–772.

[21] B. Grodal, *The equivalence principle*, in Optimization and Operation Research, Vol. III, Encyclopedia of Life Support Systems (EOLSS), EOLSS Publications, Abu, Dhabi, U.A.E., 2009, pp. 248–273.
[22] W. Hildenbrand, D. Schmeidler, and S. Zamir, *Existence of approximate equilibria and cores*, Econometrica, 41 (1973), pp. 1159–1166.
[23] M. I. Jordan, *Learning in Graphical Models*, Nato Sci. Ser. D 89, Springer, Dordrecht, The Netherlands, 1998.
[24] S. Kakade, S. Shalev-Shwartz, and A. Tewari, *On the Duality of Strong Convexity and Strong Smoothness: Learning Applications and Matrix Regularization*, Unpublished manuscript; available online from http://ttic.uchicago.edu/~shai/papers/KakadeShalevTewari09.pdf, 2009.
[25] O. Khatib, *Real-time obstacle avoidance for manipulators and mobile robots*, Internat. J. Robotics Res., 5 (1986), pp. 90–98.
[26] C. Lemaréchal and A. Renaud, *A geometric study of duality gaps, with applications*, Math. Program., 90 (2001), pp. 399–427.
[27] Z.-Q. Luo and S. Zhang, *Duality gap estimation and polynomial time approximation for optimal spectrum management*, IEEE Trans. Signal Process., 57 (2009), pp. 2675–2689.
[28] M. J. Magazine and M.-S. Chern, *A note on approximation schemes for multidimensional knapsack problems*, Math. Oper. Res., 9 (1984), pp. 244–247.
[29] N. Megiddo, *On finding primal and dual optimal bases*, ORSA J. Comput., 3 (1991), pp. 63–65.
[30] A. Nedić and A. Ozdaglar, *A geometric framework for nonconvex optimization duality using augmented Lagrangian functions*, J. Global Optim., 40 (2008), pp. 545–573.
[31] M. Pappalardo, *On the duality gap in nonconvex optimization*, Math. Oper. Res., 11 (1986), pp. 30–35.
[32] M. L. Puri and D. A. Ralescu, *Limit theorems for random compact sets in Banach space*, in Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 97, Cambridge University Press, Cambridge, UK, 1985, pp. 151–158.
[33] R. L. Raffard, C. J. Tomlin, and S. P. Boyd, *Distributed optimization for cooperative agents: Application to formation flight*, in Proceedings of the 43rd IEEE Conference on Decision and Control, Vol. 3, IEEE, Washington, DC, 2004, pp. 2453–2459.
[34] K. Rahbar, C. C. Chai, and R. Zhang, *Real-time energy management for cooperative microgrids with renewable energy integration*, in Proceedings of the IEEE International Conference on Smart Grid Communications (SmartGridComm), IEEE, Washington, DC, 2014, pp. 25–30.
[35] A. Richards, J. Bellingham, M. Tillerson, and J. How, *Coordination and control of multiple UAVs*, in AIAA Guidance, Navigation, and Control Conference, Monterey, CA, 2002.
[36] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[37] G. Scutari, F. Facchinei, and J.-S. Pang, *Equilibrium selection in MIMO communication games*, in Proceedings of the 13th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), IEEE, Washington, DC, 2012, pp. 80–84.
[38] A. Simonetto and G. Leus, *Distributed Maximum Likelihood Sensor Network Localization*, preprint, https://arxiv.org/abs/1309.2502, 2013.
[39] S. Srirangarajan, A. H. Tewfik, and Z.-Q. Luo, *Distributed sensor network localization using SOCP relaxation*, IEEE Trans. Wireless Commun., 7 (2008), pp. 4886–4895.
[40] R. M. Starr, *Quasi-equilibria in markets with non-convex preferences*, Econometrica, 37 (1969), pp. 25–38.
[41] R. M. Starr, *Approximation of points of the convex hull of a sum of sets by points of the sum: An elementary approach*, J. Econom. Theory, 25 (1981), pp. 314–317.
[42] R. M. Starr, *General Equilibrium Theory: An Introduction*, Cambridge University Press, Cambridge, UK, 2011.
[43] C. H. Teo, S. Vishwanthan, A. J. Smola, and Q. V. Le, *Bundle methods for regularized risk minimization*, J. Mach. Learn. Res., 11 (2010), pp. 311–365.
[44] M. Udell and S. Boyd, *Bounding Duality Gap for Separable Problems with Linear Constraints*, preprint, https://arxiv.org/abs/1410.4158, 2014.
[45] R. J. Vanderbei, *Linear Programming: Foundations and Extensions*, 3rd ed., Internat. Ser. Oper. Res. Management Sci. 37, Kluwer Academic Publishers, Boston, MA, 2001.
[46] R. Vujanic, P. M. Esfahani, P. Goulart, S. Mariéthoz, and M. Morari, *Vanishing Duality Gap in Large Scale Mixed-Integer Optimization: A Solution Method with Power System Applications*, J. Math. Program., submitted.

[47] R. Vujanic, P. M. Esfahani, P. Goulart, and M. Morari, *Large scale mixed-integer optimization: A solution method with supply chain applications*, in Proceedings of the 22nd Mediterranean Conference on Control and Automation (MED), IEEE, Washington, DC, 2014, pp. 804–809.

[48] A. Yamazaki, *An equilibrium existence theorem without convexity assumptions*, Econometrica, 46 (1978), pp. 541–555.

[49] W. Yu and R. Lui, *Dual methods for nonconvex spectrum optimization of multicarrier systems*, IEEE Trans. Commun., 54 (2006), pp. 1310–1322.

[50] X. Zhang, A. Saha, and S. Vishwanathan, *Lower bounds on rate of convergence of cutting plane methods*, in Proceedings of the Conference on Advances in Neural Information Processing Systems, Vancouver, Canada, 2010, pp. 2541–2549.

[51] L. Zhou, *A simple proof of the Shapley-Folkman theorem*, Econom. Theory, 3 (1993), pp. 371–372.