Hybrid Block Successive Approximation for One-Sided Non-Convex Min-Max Problems: Algorithms and Applications

Songtao Lu, *Member, IEEE*, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen, *Member, IEEE*

***Abstract*—The min-max problem, also known as the saddle point problem, is a class of optimization problems which minimizes and maximizes two subsets of variables simultaneously. This class of problems can be used to formulate a wide range of signal processing and communication (SPCOM) problems. Despite its popularity, most existing theory for this class has been mainly developed for problems with certain special *convex-concave* structure. Therefore, it cannot be used to guide the algorithm design for many interesting problems in SPCOM, where various kinds of non-convexity arise. In this work, we consider a block-wise *one-sided* non-convex min-max problem, in which the minimization problem consists of multiple blocks and is non-convex, while the maximization problem is (strongly) concave. We propose a class of simple algorithms named Hybrid Block Successive Approximation (HiBSA), which alternatingly performs gradient descent-type steps for the minimization blocks and gradient ascent-type steps for the maximization problem. A key element in the proposed algorithm is the use of certain regularization and penalty sequences, which stabilize the algorithm and ensure convergence. We show that HiBSA converges to some properly defined first-order stationary solutions with quantifiable global rates. To validate the efficiency of the proposed algorithms, we conduct numerical tests on a number of problems, including the robust learning problem, the non-convex min-utility maximization problems, and certain wireless jamming problem arising in interfering channels.**

***Index Terms*—Min-max optimization, saddle point problems, block successive approximation, gradient descent and ascent.**

Manuscript received February 17, 2019; revised August 13, 2019, November 22, 2019, and January 8, 2020; accepted March 3, 2020. Date of publication April 17, 2020; date of current version June 26, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Weiyu Xu. The work of Songtao Lu, Ioannis Tsaknakis, and Mingyi Hong was supported by NSF under Grants CMMI-172775 and CIF-1910385 and in part by an ARO under Grant W911NF-19-1-0247; Yongxin Chen is supported by NSF under Grant 1901599. *(Songtao Lu and Ioannis Tsaknakis contributed equally to this work.) (Corresponding author: Mingyi Hong.)*

Songtao Lu was with the University of Minnesota, Minneapolis, MN 55455 USA. He is now with the IBM Research AI, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: songtao@ibm.com).

Ioannis Tsaknakis and Mingyi Hong are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: tsakn001@umn.edu; mhong@umn.edu).

Yongxin Chen is with the School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: yongchen@gatech.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the authors. The material includes an extension of the proposed algorithm and the respective convergence analysis. This material is five pages in size. This article was presented in part at the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, U.K., 2019 [1].

Digital Object Identifier 10.1109/TSP.2020.2986363

I. Introduction

CONSIDER the min-max (a.k.a. saddle point) problem below:

$$\min_{x}\max_{y} \quad f(x_1, x_2, \ldots, x_K, y) + \sum_{i=1}^{K} h_i(x_i) - g(y) \tag{1}$$
$$\text{s.t.} \quad x_i \in \mathcal{X}_i,\ y \in \mathcal{Y},\ i = 1, \ldots, K$$

where $f : \mathbb{R}^{NK+M} \to \mathbb{R}$ is a continuously differentiable function; $h_i : \mathbb{R}^N \to \mathbb{R}$ and $g : \mathbb{R}^M \to \mathbb{R}$ are some convex possibly non-smooth functions; $x := [x_1; \ldots; x_K] \in \mathbb{R}^{N\cdot K}$ and $y \in \mathbb{R}^M$ are the block optimization variables; $\mathcal{X}_i$'s and $\mathcal{Y}$ are some convex and compact feasible sets. We call the problem *one-sided* non-convex problem because we assume that $f(x, y)$ is non-convex with respect to (w.r.t.) x, and (strongly) concave w.r.t. y. For notational simplicity, we will use $\ell(x_1, x_2, \ldots, x_K, y)$ to denote the overall objective function for problem (1).

Problem (1) is quite generic, and it arises in a wide range of signal processing and communication (SPCOM) applications. We list of few of these applications below.

A. Motivating Examples in SPCOM

Distributed non-convex optimization: Consider a network of K agents defined by a connected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with $|\mathcal{V}| = K$, where each agent i can communicate with its neighbors. A generic problem formulation that captures many distributed machine learning and signal processing problems can be formulated as follows [2]–[5]:

$$\min_{\{x_i\}} \sum_{i=1}^{K} f_i(x_i) + h_i(x_i), \|x_i - x_j\| \le c_{i,j}, (i,j) \text{ are neighbors}$$

where each $f_i : \mathbb{R}^N \to \mathbb{R}$ is a non-convex, smooth function, $h_i : \mathbb{R}^N \to \mathbb{R}$ is a convex non-smooth regularizer, and $x_i \in \mathbb{R}^N$ is agent i's local variable. Each agent i has access only to f_i and h_i. The non-negative constants $c_{i,j}$'s are predefined, and they can be selected to represent different levels of agreement among the agents [6]. Despite the fact that there have been a number of recent works on distributed non-convex optimization [7]–[12], the above problem formulation cannot be covered by any of these due to two main reasons: *i)* the nonsmooth regularizers h_i's can be different across the nodes, invalidating the assumptions made in, e.g., [7] (which requires uniform regularizer across the nodes), and *ii)* the partial consensus constraints are considered rather than the exact consensus where $c_{i,j} \equiv 0,\ \forall i, j$.

The above problem can be equivalently expressed as:

$$\begin{aligned} \min_{x,\tilde{x}} \quad & f(x)+h(x) := \sum_{i=1}^{K} \left(f_i(x_i)+h_i(x_i)\right) \\ \text{s.t.} \quad & (A\otimes I_N)x-\tilde{x}=0,\ \tilde{x}\in\mathcal{Z}\subseteq\mathbb{R}^{|\mathcal{E}|\cdot N} \end{aligned} \tag{2}$$

where $x := [x_1;\ldots;x_K] \in \mathbb{R}^{KN}$; $A \in \mathbb{R}^{|\mathcal{E}|\times K}$ is the incidence matrix, i.e., assuming that the edge e is incident on vertices i and j, with $i > j$ we have that $A_{ei} = 1, A_{ej} = -1$ and $A_{e\cdot} = 0$ for all other vertices; $\otimes$ denotes the Kronecker product; $\tilde{x} \in \mathbb{R}^{|\mathcal{E}|N}$ is the auxiliary variable representing the difference between two neighboring local variables; the feasible set $\mathcal{Z}$ represents the bounds on the size of the differences. Using duality theory we can introduce the Lagrangian multiplier vector y and rewrite the above problem as:

$$\min_{x\in\mathbb{R}^{KN},\tilde{x}\in\mathcal{Z}} \max_{y\in\mathbb{R}^{|\mathcal{E}|\cdot N}} f(x)+h(x)+\langle y,(A\otimes I_N)x-\tilde{x}\rangle\,. \tag{3}$$

See Section III-A for detailed discussion on this reformulation and its relationship with (2). Clearly (3) is in the form of (1).

Robust learning over multiple domains: In [13] the authors introduce a robust learning framework, in which the training sets from M different domains are used to train a machine learning model. Let $\mathcal{S}_m = \{(s_i^m, t_i^m)\}$, $1 \le m \le M$ be the individual training sets with $s_i^m \in \mathbb{R}^N$, $t_i^m \in \mathbb{R}$; x be the parameter of the model we intent to learn, $\ell(\cdot)$ a non-negative loss function, and $f_m(x) = \frac{1}{|\mathcal{S}_m|}\sum_{i=1}^{|\mathcal{S}_m|} \ell(s_i^m, t_i^m, x)$ is the (possibly) non-convex empirical risk in the m-th domain. The following problem formulates the task of finding the parameter x that minimizes the empirical risk, while taking into account the worst possible distribution over the M different domains:

$$\min_x \max_{y\in\Delta}\ y^T F(x) - \frac{\lambda}{2} D(y||q) \tag{4}$$

where $F(x) := [f_1(x);\ldots;f_M(x)] \in \mathbb{R}^{M\times 1}$; y describes the adversarial distribution over the different domains; $\Delta := \{y \in \mathbb{R}^M \mid 0 \le y_i \le 1,\ i = 1,\ldots,M,\ \sum_{i=1}^M y_i = 1\}$ is the standard simplex; $D(\cdot)$ is some distance between probability distributions, q is some prior probability distribution, and $\lambda > 0$ is some constant. The last term in the objective function represents some regularizer that imposes structures on the adversarial distribution.

Power control and transceiver design problem: Consider a problem in wireless transceiver design, where K transmitter-receiver pairs transmit over N channels to maximize their *minimum* rates. User k transmits messages with power $x_k := [x_k^1;\ldots;x_k^N]$, and its rate is given by (assuming Gaussian signaling):

$$R_k(x_1,\ldots,x_K) = \sum_{n=1}^{N} \log\left(1+\frac{a_{kk}^n x_k^n}{\sigma^2+\sum_{\ell=1,\ell\neq k}^K a_{\ell k}^n x_\ell^n}\right),$$

which is a non-convex function on $x := [x_1;\ldots;x_K]$. Here $a_{\ell k}^n$'s denote the channel gain between the pair (ℓ, k) on the nth channel, and σ^2 is the noise power. Let $\bar{x}$ denote the power budget for each user, then the classical max-min fair power control problem is: $\max_{x\in\mathcal{X}} \min_k R_k(x)$, where $\mathcal{X} := \{x \mid 0 \le \sum_n x_k^n \le \bar{x}, \forall k\}$ denotes the feasible power allocations. The above max-min rate problem can be equivalently formulated as (1) (see Section III-A for details) :[1]

$$\min_{x\in\mathcal{X}} \max_{y\in\Delta} \sum_{k=1}^{K} -R_k(x_1,\ldots,x_K)\times y_k, \tag{5}$$

where the set $\Delta \subseteq \mathbb{R}^K$ is again the standard simplex.

A closely related problem is the coordinated beamforming design in a (multiple input single output) MISO interference channel. In this case the target is to find the optimal beamforming vector for each user in order to maximize some system utility function under the total power and outage probability constraints [14]. When the min-rate utility is used, this problem can be formulated as

$$\max_{x_i\in\mathbb{C}^{N_t},\forall i} \min_i R_i(\{x_k\}) \quad \text{s.t. } \|x_i\|^2 \le \bar{p}, \forall i \tag{6}$$

where x_i is the transmit beamformer, N_t is the number of antennas. Also, $R_i(\{x_k\}) = \log_2(1+\xi_i(\{x_k\}_{k\neq i}) x_i^H Q_{ii} x_i)$, where ξ_i incorporates the outage constraints and the cross-link interference, while Q_{ii} denotes the covariance matrix of the channel between the ith transmitter-receiver pair.

For other setups, similar min-max problems can be formulated, some of which can be solved optimally (e.g., power control [15]–[17], transmitter density allocation [18], or certain MISO beamforming [19], [20]). But for general multi-channel and/or MIMO interference channel, the corresponding problem is NP-hard [21]. Many heuristic algorithms are available for these problems [21]–[24], but they are all designed for special problems, and often require repeatedly invoking computationally expensive general purpose solvers. For computational tractability, a common approach is to perform the following approximation of the min-rate utility [25]:

$$\min_i r_i \approx -1/\gamma \log_2 \sum_{i=1}^{N} 2^{-\gamma r_i}. \tag{7}$$

However such an approximation procedure can introduce significant rate loss, as will be seen in Section IV.

Power control in the presence of a jammer: Consider an extension of the power control problem, where a jammer participates in a K-user N-channel interference channel transmission [26]. Differently from a regular user, the jammer's objective is to reduce the sum-rate of other users by properly transmitting noises. Let y^n denote the jammer's transmission on the nth channel, then one can formulate the following sum-rate maximization-minimization problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \sum_{(k,n)} -\log\left(1+\frac{a_{kk}^n x_k^n}{\sigma^2+\sum_{j=1,j\neq k}^K a_{jk}^n x_j^n + a_{0k}^n y^n}\right), \tag{8}$$

where x_k and y are the power allocations of user k and the jammer, respectively; the set $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$, where $\mathcal{X}_k$ is defined similarly as before.

B. Related Work

Motivated by these applications, it is of interest to develop efficient algorithms for solving these problems with theoretical convergence guarantees. In the optimization community, there has been a long history of studying min-max optimization problems. When the problem is convex in x and concave in y, algorithms

[1] A minus sign is added to equivalently transform to the min-max problem.

TABLE I
SUMMARY OF ALGORITHMS FOR $\min_x \max_y \ f(x,y)+h(x)-g(y)$, WHERE f IS A SMOOTH FUNCTION, WHILE h AND g ARE CONSIDERED CONVEX NON-SMOOTH FUNCTIONS UNLESS OTHERWISE STATED. NOTE THAT IN THE 3RD COLUMN WE CHARACTERIZE THE TYPE OF THE ALGORITHMS, I.E DETERMINISTIC (DET.) OR STOCHASTIC (ST.). MOREOVER, WE USE THE ABBREVIATIONS NC FOR NON-CONVEX, SP FOR STATIONARY POINT, STR. FOR STRONGLY, MIN. FOR MINIMIZATION AND MAX. FOR MAXIMIZATION.

Algorithm	Optimality Criterion	Det./ St.	Assumptions		Gradient Complexity [1]
Multi-Step GDA [36], [37]	[2] 1st order Nash equilibrium	Det.	f NC in x/Polyak-Lojasiewicz in y $h(x)=0,\ g(y)=0$		$\mathcal{O}(\epsilon^{-2}\log(\frac{1}{\epsilon}))$
			f NC in x/concave in y $h(x)=0,\ g(y)=0$		$\mathcal{O}(\epsilon^{-3.5}\log(\frac{1}{\epsilon})))$
Robust optimization [13]	[3] 1st order SP for min. problem Optimality gap for max. problem	St.	f NC in x/linear in y $h(x)=0,\quad g$ convex, smooth		$\mathcal{O}(\epsilon^{-6})$
			f NC in x/linear in y $h(x)=0,\quad g$ str. convex, smooth		$\widetilde{\mathcal{O}}(\epsilon^{-2}+\epsilon^{-4})$
PG-SMD/ PGSVRG [34]	[4] 1st order SP for min. problem	St.	$f=\frac{1}{n}\sum_{i=1}^n y^T c_i(x)$, $c_i(x)$ NC	g str. convex	$\widetilde{\mathcal{O}}(n\epsilon^{-2}+\epsilon^{-4})$
			f NC in x, f concave in y	g convex	$\widetilde{\mathcal{O}}(\epsilon^{-6})$
			$f=\frac{1}{n}\sum_{i=1}^n f_i$ f NC in x f concave in y	g str. convex	$\widetilde{\mathcal{O}}(n\epsilon^{-2})$
				g convex	$\widetilde{\mathcal{O}}(n\epsilon^{-2}+\epsilon^{-6})$
HiBSA (our work)	[5] 1st order SP (def. (17))	Det.	f NC in x/strongly concave in y		$\mathcal{O}(\epsilon^{-2})$
			f NC in x/ linear in y		$\widetilde{\mathcal{O}}(\epsilon^{-4})$
			[6] f NC in x/concave in y		$\widetilde{\mathcal{O}}(\epsilon^{-4}\log(\frac{1}{\epsilon}))$

[1] Gradient complexity refers to the total number of gradient evaluations required to reach an ϵ-stationary solution, where the definition of stationarity can vary for different works.
[2] The approximate stationarity condition for $(\tilde{x},\tilde{y})$ is $-\min_x\langle\nabla_x f(\tilde{x},\tilde{y}),x-\tilde{x}\rangle\le\epsilon,\forall x\in\mathcal{X}:\|x-\tilde{x}\|\le 1$ and $\max_y\langle\nabla_y f(\tilde{x},\tilde{y}),y-\tilde{y}\rangle\le\epsilon,\forall y\in\mathcal{Y}:\|y-\tilde{y}\|\le 1$.
[3]For the minimization problem, convergence is established using the stationarity gap, i.e $\|\nabla_x f(x,y)\|\le\epsilon$, while for the maximization problem the optimality condition $\max_{y'\in\mathcal{Y}} f(x,y')-f(x,y)\le\epsilon$ is used.
[4]The optimality measure is the norm of the gradient of the Moreau envelope of the minimization problem, i.e $\|\nabla\phi_\gamma(x)\|\le\epsilon$ with $\phi_\gamma(x)=\min_z\{\phi(z)+(1/2\gamma)\|z-x\|^2\}$ and $\phi(z)=\max_y\{f(z,y)-g(y)\}$.
[5] A point $(\tilde{x},\tilde{y})$ such that $\|\mathcal{G}_\rho^\beta(\tilde{x},\tilde{y})\|\le\epsilon$ and [6] This complexity is obtained for the algorithm given in the supplementary document of the current work [40].

have been developed which can solve the convex-concave saddle problem optimally; see [27]–[31] and the references therein. However, when the problem is non-convex, the convergence behavior of such alternating type algorithms has not been well understood.

Although there are many recent works on the non-convex minimization problems [32], only a few of them have been focused on the non-convex min-max problems. An optimistic mirror descent algorithm is proposed in [33], and its convergence to a saddle point is established under certain strong coherence assumptions. In [13], algorithms for robust optimization are proposed, where the x problem is unconstrained, and y linearly couples with a non-convex function of x [cf. (4)]. In [34], a proximally guided stochastic mirror descent method (PG-SMD) is proposed, which provably converges to an approximate stationary point of the outer minimization problem. An oracle based non-convex stochastic gradient descent for generative adversarial networks (GAN) is proposed in [35], where the algorithms solve the maximization subproblem up to some small error. Moreover, in [36] a multi-step GDA scheme is introduced, where the maximization problem is approximately solved using a number of gradient ascent steps. In [37] the convergence of a primal-dual algorithm to a first-order stationary point is established for a class of GAN problems formulated as a special min-max optimization problem where the coupling term is linear w.r.t the discriminator. More recently, in [38] it has been shown that GDA can converge to a stationary point of the outer minimization problem in the (strongly) concave case under certain conditions. Under the same optimality criterion and assuming that the inner problem is concave, [39] proves convergence using a proximal dual implicit accelerated gradient method.

It is worth noting that, in the works discussed above, different optimality criteria are often utilized. Since these conditions are not equivalent to each other, one cannot directly compare the convergence guarantees of algorithms that reach these criteria. On the other hand, these optimality criteria often share some interesting implicit connections. For example, it can be shown that, no matter if the inner maximization problem is strongly concave or concave, as long as a point (x^*,y^*) is an (exact or approximate) stationary point defined in this current work [see (17)], then it is also an (exact or approximate, respectively) stationary point in the sense defined in [34]; see [38] for detailed discussions. In Table I we provide a summary of some algorithms discussed above, including the complexity and the respective optimality criterion.

C. Contribution of This Work

In this work, we design effective algorithms for the min-max problem by adopting the popular block alternating minimization/maximization strategy. The studied problems allow non-convexity and non-smoothness in the objective, as well as non-linear coupling between variables. The algorithm proposed in this work is named the *Hybrid Block Successive Approximation* (HiBSA) algorithm, because it updates the variables block by block, where each block is optimized using a strategy similar to the idea of successive *convex* approximation (SCA) [41] – except that to update the y block, a *concave* approximation is used (hence the name "hybrid"). Despite the fact that such a block-wise alternating optimization strategy is simple and easy to implement (for example it has been used in the popular block successive upper bound minimization (BSUM) framework [41], [42] for *minimization-only* problem), it turns out that having the

maximization subproblem invalidates all the previous analysis for *minimization-only* algorithms.

The main contributions of this paper are listed as follows. First, a number of applications in SPCOM have been formulated in the framework of non-convex, one-sided min-max problem (1). Second, based on different assumptions on how x and y variables are coupled, as well as whether the y problem is strongly concave or merely concave, three different types of min-max problems are studied. For each of the problem class, a simple single loop algorithm is presented, together with its convergence guarantees.[2] The major benefits of using the block successive approximation strategy are twofold: 1) each subproblem can be solved effectively, and 2) it is relatively easy to integrate many existing algorithms that are designed for only solving minimization problems (such as those based on the BSUM framework [41], [42]). Finally, extensive numerical experiments are conducted for selected applications from SPCOM to validate the proposed algorithms.

Overall, to the best of our knowledge this is the first time that the convergence of the alternating block successive approximation type algorithm is rigorously analyzed for the (one-sided) non-convex min-max problem (1).

Notation: The notation $\|\cdot\|$ denotes the vector 2-norm $\|\cdot\|_2$; $\otimes$ denotes the Kronecker product; I_N is the $N\times N$ identity matrix; $\langle\cdot,\cdot\rangle$ is the Euclidean inner product; $\mathcal{I}_{\mathcal{X}}(x)$ denotes the indicator function on set $\mathcal{X}$; in case the subscript is missing the set is implied by the context; $[K]:=\{1,\ldots,K\}$. Finaly, the notation $\tilde{\mathcal{O}}$ denotes big O notation $\mathcal{O}$ up to some logarithmic factor.

II. The Proposed Algorithms and Analysis

In this section, we present our main algorithm. Towards this end, we will first make a number of blanket assumptions on problem (1), and then present the HiBSA algorithm in its generic form. We will then discuss in detail about various algorithmic choices, as well as major challenges in the analysis.

Let the superscript r denote iteration number. For notational simplicity, we will define the following:

$$w_i^{r+1}:=[x_1^{r+1};\ldots,x_{i-1}^{r+1},x_i^r,\cdots x_K^r]\in\mathbb{R}^{NK}, \tag{9a}$$

$$w_{-i}^{r+1}:=[x_1^{r+1};\ldots,x_{i-1}^{r+1},x_{i+1}^r,\cdots x_K^r]\in\mathbb{R}^{N(K-1)}, \tag{9b}$$

$$x_{-i}:=[x_1;\ldots,x_{i-1},x_{i+1},\cdots x_K]\in\mathbb{R}^{N(K-1)}. \tag{9c}$$

Throughout the paper, we will assume that problem (1) satisfies the following blanket assumption.

Assumption A: The following conditions hold for (1):

A.1 $f:\mathbb{R}^{KN+M}\to\mathbb{R}$ is continuously differentiable; the feasible sets $\mathcal{X}=\mathcal{X}_1\times\cdots\times\mathcal{X}_K$ and $\mathcal{Y}\subseteq\mathbb{R}^M$ are convex and compact. Further $\ell(x,y)$ is lower bounded, that is, $\ell(x,y)\geq\underline{\ell},\ \forall x\in\mathcal{X},y\in\mathcal{Y}$.

A.2 $h_i(\cdot)$'s and $g(\cdot)$ are convex and non-smooth functions;

A.3 f has Lipschitz continuous gradient w.r.t. x_i for every i with constant L_{x_i}, that is:

$$\|\nabla_{x_i}f(\bar{z})-\nabla_{x_i}f(z)\|\leq L_{x_i}\|\bar{z}-z\|,\forall\bar{z},z\in\mathcal{X}\times\mathcal{Y}; \tag{10}$$

Furthermore, f has Lipschitz continuous gradient w.r.t. y with constant L_y, that is:

$$\|\nabla_y f(\bar{z})-\nabla_y f(z)\|\leq L_y\|\bar{z}-z\|,\forall\bar{z},z\in\mathcal{X}\times\mathcal{Y}. \tag{11a}$$

Next we describe the proposed HiBSA algorithm.

Hybrid Block Successive Approximation (HiBSA) Algorithm

At each iteration $r=1,2,3,\cdots$

[S1]. For $i=1,\ldots,K$, perform the following update:

$$x_i^{r+1}=\arg\min_{x_i\in\mathcal{X}_i}U_i(x_i;w_i^{r+1},y^r)+h_i(x_i)+\frac{\beta^r}{2}\|x_i-x_i^r\|^2. \tag{12}$$

[S2]. Perform the following update for the y-block:

$$y^{r+1}=\arg\max_{y\in\mathcal{Y}}U_y(y;x^{r+1},y^r)-g(y)-\frac{\gamma^r}{2}\|y\|^2. \tag{13}$$

[S3]. If converges, stop; otherwise, set $r=r+1$, go to **[S1]**.

Note that $\{\beta^r\geq 0\}$ and $\{\gamma^r\geq 0\}$ are some algorithm parameters, whose values will be specified shortly in the next section. Properly designing the regularization sequence $\{\gamma^r\}$ is the key to ensure that the algorithm works when the y problem is concave but not *strongly* concave. Further, each $U_i(\cdot;w,y):\mathbb{R}^N\to\mathbb{R}$ (resp. $U_y(\cdot,w,y)$) is some approximation function of $f(\cdot,x_{-i},y)$ (resp. $f(x,\cdot)$). For the $U_i(\cdot)$'s the following assumptions hold.

Assumption B: Each $U_i(\cdot)$ satisfies the following conditions:

B.1 **(Strong convexity):** Each $U_i(\cdot;w,y)$ is strongly convex with modulus $\mu_i>0$:

$$U_i(x_i;w,y)-U_i(z_i;w,y)\geq\langle\nabla_{z_i}U_i(z_i;w,y),x_i-z_i\rangle+\frac{\mu_i}{2}\|x_i-z_i\|^2,\forall w\in\mathcal{X},y\in\mathcal{Y},x_i,z_i\in\mathcal{X}_i.$$

B.2 **(Gradient consistency):** Each $U_i(\cdot;w,y)$ satisfies:

$$\nabla_{z_i}U_i(z_i;x,y)\Big|_{z_i=x_i}=\nabla_{x_i}f(x,y),\forall i,\forall x\in\mathcal{X},y\in\mathcal{Y}.$$

B.3 **(Tight upper bound):** Each $U_i(\cdot;w,y)$ satisfies:

$$U_i(z_i;x,y)\geq f(x,y),\text{and } U_i(x_i;x,y)=f(x,y),\ \forall x\in\mathcal{X},y\in\mathcal{Y},\ z_i\in\mathcal{X}_i.$$

B.4 **(Lipschitz gradient):** Each $U_i(\cdot;w,y)$ satisfies:

$$\|\nabla U_i(z_i;w,y)-\nabla U_i(v_i;w,y)\|\leq L_{u_i}\|v_i-z_i\|,\ \forall w\in\mathcal{X},y\in\mathcal{Y},v_i,z_i\in\mathcal{X}_i.$$

Clearly, the x update step **[S1]** closely resembles the BSUM algorithm [41], [43], which is designed for minimization problems. Similarly as in BSUM, approximation functions are used to simplify the update for each subproblem; see [41] for a number of such functions often used in signal processing applications.

However, a key difference from the BSUM, or for that matter, all successive convex approximation (SCA) based algorithms such as the inexact flexible parallel algorithm (FLEXA) [44]–[46], the concave-convex procedure (CCCP) [47], is the presence of the *ascent* step in **[S2]**. This step is needed to deal with the inner maximization problem, but unfortunately the use of it invalidates the existing analyses for SCA-type algorithms, because all of them critically depend on consistently achieving some form of descent as the algorithms progress. As a result,

[2]In addition to the algorithm presented in the main text, we also provide an alternative double-loop algorithm for the case where the y problem is concave, in the supplementary document of this article [40].

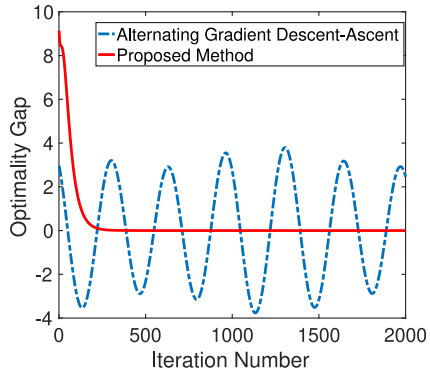


Fig. 1. Behavior of (14) and Alg. 1. The y-axis is the 1st-order optimality gap $\|Ax\|^2 + \|y^T A\|^2$, which ideally should go to zero. The solid line represents the HiBSA algorithm with $\gamma^r = 1/\sqrt{r},\ \beta^r = r,\ \forall r$.

how to properly implement and analyze the proposed algorithm represents a major challenge.

We note that it is not straightforward to design algorithms for one-sided non-convex min-max problems, as compared with non-convex minimization problems. For the former problem, simple algorithms like gradient descent-ascent can diverge (see Example 1 or [30]), but if we specialize such an algorithm to the latter problem (which becomes the well-known gradient descent), then it will converge to a second-order stationary solution [48]. We refer the readers to a few recent works [38] [39] for more discussions.

Example 1 [30]: Consider a special case of problem (1), where $K = 1$ (a single block variable), and A is a randomly generated matrix of size $N \times M$: $\min_{x\in\mathbb{R}^N} \max_{y\in\mathbb{R}^M}\ y^T Ax$. Let us apply a special case of the HiBSA algorithm by utilizing the following approximation function:

$$U_1(v; w, y) = y^T Av + \frac{\eta}{2}\|v - w\|^2,$$

$$U_y(u; x, y) = u^T Ax - \frac{1}{2\lambda}\|u - y\|^2.$$

Letting $\gamma^r = 0$ and $\beta^r = 0$ for all r, the HiBSA becomes an alternating gradient descent-ascent algorithm

$$x^r = x^{r-1} - \frac{1}{\eta}A^T y^{r-1}, \quad y^r = y^{r-1} + 2\lambda A x^r,\ \forall r. \tag{14}$$

Unfortunately, one can verify that for almost any A, *regardless* the choices of η, λ, (14) will not converge to the desired solution satisfying: $A^T y^* = 0$ and $Ax^* = 0$; see Fig. 1. This is because the linear system describing the dynamics of the vector (x^r, y^r) is always unstable. ■

The above example motivates us to introduce *both* the proximal term $\beta^r/2\|x - x^r\|^2$ in **[Step 1]** of HiBSA, and the penalty term $-\frac{\gamma^r}{2}\|y^r\|^2$ in **[Step 2]**. By properly selecting the sequences $\{\beta^r, \gamma^r\}$, we will show in the next section, that the HiBSA will converge for a wide class of problems (including Example 1 as a special case).

III. Theoretical Properties of HiBSA

We start to present our main convergence results for the HiBSA. Our analysis will be divided into three cases according to the structure of the coupling term $f(x, y)$. Separately considering different cases of (1) is necessary, since the analysis and convergence guarantees could be different. Note that throughout this section, we will assume that $g(y)$ is convex, but *not* strongly convex. In case $\ell(x, y)$ is strongly concave in y, the strongly concave term will be absorbed into $f(x, y)$.

A. Optimality Conditions

First, let us elaborate on the type of solutions we would like to obtain for problem (1). Because of the non-convexity involved in the minimization problem, we will not be able to use the classical measure of optimality for saddle point problems (i.e., the distance to a saddle point). Instead, we will adopt some kind of first-order stationarity conditions. To precisely state our condition, let us define the *proximity operator* for x and y blocks as follows:

$$\text{Px}_i^\beta(v_i) := \arg\min_{x_i\in\mathcal{X}}\ h_i(x_i) + \frac{\beta}{2}\|x_i - v_i\|^2,\ \forall i \in [K]$$

$$\text{Py}^{1/\rho}(w) := \arg\max_{y\in\mathcal{Y}}\ -g(y) - \frac{1}{2\rho}\|y - w\|^2. \tag{15}$$

Moreover, we define the stationarity gap for problem (1) as:

$$\nabla\mathcal{G}_\rho^\beta(x, y) := \begin{bmatrix} \beta(x_1 - \text{Px}_1^\beta(x_1 - 1/\beta\nabla_{x_1} f(x, y))) \\ \vdots \\ \beta(x_K - \text{Px}_K^\beta(x_K - 1/\beta\nabla_{x_K} f(x, y))) \\ 1/\rho(y - \text{Py}^{1/\rho}(y + \rho\nabla_y f(x, y))) \end{bmatrix}. \tag{16}$$

We say that a tuple (x^*, y^*) is a first-order stationary solution for problem (1) if it holds that:

$$\|\nabla\mathcal{G}_\rho^\beta(x^*, y^*)\| = 0. \tag{17}$$

To see that (17) makes sense, first note that if $h \equiv 0, g \equiv 0, \mathcal{Y} = \mathbb{R}^M, \mathcal{X} = \mathbb{R}^{NK}$, then it reduces to the condition $\|[\nabla_x f(x^*, y^*); \nabla_y f(x^*, y^*)]\| = 0$, which is *independent* of the algorithm parameters (β, ρ). Further, we can check that if y is not present, then condition (17) is equivalent to the first-order stationary condition for the resulting non-convex minimization problem (see, e.g., [32]). Further, if x is not present, then condition (17) simply says that $y^* \in \arg\max_{y\in Y}\{f(\cdot, y) - g(y)\}$.

Following the above definition, we will say that (x^*, y^*) is an ϵ-stationary solution if the following holds

$$\|\nabla\mathcal{G}_\rho^\beta(x^*, y^*)\| \le \epsilon. \tag{18}$$

As mentioned in the introduction, the stationarity conditions (17) and (18) are related to the stationarity conditions utilized in [34], [38]. To illustrate this point, consider the case where $h \equiv 0, g \equiv 0, \mathcal{X} = \mathbb{R}^{NK}$; the stationarity condition (17) reduces to:

$$\|\nabla\mathcal{G}(x, y)\| = \|[\nabla_x f(x, y), y - \text{proj}_{\mathcal{Y}}(y + \nabla_y f(x, y))]^T\| = 0.$$

First, consider the case where f is strongly concave in y. Then, it can be shown that if (x^*, y^*) is an ϵ-stationary point in the sense of (18), then it is also an $\mathcal{O}(\epsilon)$ stationary point in the sense defined in [34], that is:

$$\|\nabla F(x^*)\| \le \mathcal{O}(\epsilon), \quad F(x) := \max_{y\in\mathcal{Y}} f(x, y). \tag{19}$$

Moreover, in the case where f is concave in y, for an ϵ-stationary point according to definition (18), it holds that $\|\nabla F_{1/2\ell}(x^*)\| \le \mathcal{O}(\epsilon)$, where $F_{1/2\ell}$ is the Moreau envelope of F defined as

$$F_{1/2\ell}(x) = \min_w F(w) + \ell\|w - x\|^2. \tag{20}$$

For more details the readers can refer to [38].

Based on the above definition of the first-order stationarity, we establish the equivalence between a few optimization formulations discussed in Section I.

Proposition 3.1: Problems (2) and (3) are equivalent, in the sense that every KKT point for problem (2) is a first-order stationary solution of (3) (in the sense of (17)), and vice versa.

Proof: For simplicity of notation we assume $N = 1$. Consider the following KKT conditions for problem (2)

$$\begin{aligned}&\langle \nabla_x f(x^*) + \xi^*(x^*) + A^T y^*, x - x^* \rangle \\ &\quad - \langle y^*, \tilde{x} - \tilde{x}^* \rangle \geq 0, \forall \text{ feasible } (x, \tilde{x}) \\ &Ax^* = \tilde{x}^*,\end{aligned} \tag{21}$$

where $\xi^*(x^*) \in \partial h(x^*)$ and y is the Lagrange multiplier. Now consider a stationary point $(x^*, \tilde{x}^*, y^*)$ of problem (3). Then the stationarity condition (17) implies that

$$x^* = \arg\min_x \ \langle A^T y^* + \nabla_x f(x^*), x - x^* \rangle + h(x) + \frac{\beta}{2}\|x - x^*\|^2, \tag{22a}$$

$$\tilde{x}^* = \arg\min_{\tilde{x} \in \mathcal{Z}} \ \langle -y^*, \tilde{x} - \tilde{x}^* \rangle + \frac{\beta}{2}\|\tilde{x} - \tilde{x}^*\|^2, \tag{22b}$$

$$y^* = \arg\max_y \ \langle Ax^* - \tilde{x}^*, y - y^* \rangle - \frac{1}{2\rho}\|y - y^*\|^2. \tag{22c}$$

The optimality conditions for these problems imply

$$A^T y^* + \nabla_x f(x^*) + \xi(x^*) = 0 \tag{23}$$

$$\langle -y^*, \tilde{x} - \tilde{x}^* \rangle \geq 0, \ \forall \tilde{x} \in \mathcal{Z}, \quad Ax^* - \tilde{x}^* = 0. \tag{24}$$

Clearly, the conditions (23)–(24) imply (21).

Conversely, suppose (21) is true. By setting $x = x^*$ in (21) we obtain condition (24). Moreover, in order to obtain condition (23) we set $\tilde{x} = \tilde{x}^*$ in (21) and take into account the fact that (21) holds $\forall x \in \mathbb{R}^{KN}$. The proof is completed. **Q.E.D.**

Proposition 3.2: Consider the problem:

$$\max_{x \in \mathcal{X}} \min_k R_k(x),$$

and its reformulation (5). They are equivalent in the sense that, an equivalent smooth reformulation of the former has the same first-order stationary solutions as those of the latter [in the sense of (17)].

Proof: A well-known equivalent smooth formulation of the min-utility maximization problem is given below (equivalent in that the global optima of these two problems are the same)

$$\max_{\lambda, x \in \mathcal{X}} \ \lambda, \quad \text{s.t. } R_k(x) \geq \lambda, \ \forall k. \tag{25}$$

The partial KKT conditions of the above problem are

$$\left\langle \sum_{i=1}^K \hat{y}_i \nabla_x R_i(\hat{x}), x - \hat{x} \right\rangle \leq 0, \ \forall x \in \mathcal{X}, \tag{26}$$

$$\sum_i \hat{y}_i = 1, \ \hat{y}_i \geq 0, \ \hat{y}_i(\hat{\lambda} - R_i(\hat{x})) = 0, \ R_i(\hat{x}) \geq \hat{\lambda}, \ \forall i,$$

where $\{\hat{y}_i\}_{i=1}^K$ are the respective Lagrange multipliers, which together with $(\hat{\lambda}, \hat{x})$ satisfy the KKT conditions.

Now consider a stationary point (x^*, y^*) of problem (5). Then the optimality conditions (17) imply that

$$x^* = \arg\min_{x_i \in \mathcal{X}} \left\langle \sum_{i=1}^K -y_i^* \nabla_x R_i(x^*), x - x^* \right\rangle + \frac{\beta}{2}\|x - x^*\|^2 \tag{27a}$$

$$y^* = \arg\max_{y \in \Delta} \ \langle -R(x^*), y - y^* \rangle - \frac{1}{2\rho}\|y - y^*\|^2, \tag{27b}$$

where $R(x^*) := [R_1(x^*); \ldots; R_K(x^*)]$. Let us define

$$\lambda^* = \min_{i=1,\ldots,K} \{R_i(x^*)\}$$

so it holds that $R_i(x^*) \geq \lambda^*, \ \forall i$.

Plugging (x^*, y^*) into the optimality conditions of (27a), (27b), we obtain:

$$\left\langle \sum_{i=1}^K -y_i^* \nabla_{x_i} R_i(x^*), x - x^* \right\rangle \geq 0, \ \forall x \in \mathcal{X} \tag{28a}$$

$$\langle -R(x^*), y - y^* \rangle \leq 0, \ \forall y \in \Delta, \ y^* \in \Delta. \tag{28b}$$

For all i such that $R_i(x^*) = \lambda^*$ obviously it holds that $y_i^*(\lambda^* - R_i(x^*)) = 0$. Let i, j be indices such that $R_i(x^*) > \lambda^*$ and $R_j(x^*) = \lambda^*$. Then, plugging $y_i = 0, y_j = y_i^* + y_j^*$ and $y_k = y_k^*, k \neq i, j$ into (28b) yields $y_i^*(R_j(x^*) - R_i(x^*)) \geq 0$. Because $R_j(x^*) - R_i(x^*) < 0$ and $y_i^* \geq 0$ it must necessarily hold $y_i^* = 0$ and thus $y_i^*(\lambda^* - R_i(x^*)) = 0$. As a result the conditions (26) are satisfied.

Conversely, assume (x^*, y^*) satisfies conditions (26). Note that $R_i(x^*) y_i \geq \lambda^* y_i$ for any $y \in \Delta$, so

$$\langle R(x^*), y - y^* \rangle = \sum_{i=1}^K R_i(x^*)(y_i - y_i^*) \geq \sum_{i=1}^K \lambda^*(y_i - y_i^*) = 0,$$

for all $y \in \Delta, y^* \in \Delta$. It is not difficult to see that (x^*, y^*) satisfies the rest of the conditions in (28a)–(28b). As a result the opposite direction also holds. **Q.E.D.**

B. Convergence Analysis: $f(x, y)$ Strongly Concave in y

Starting this subsection, we will analyze the convergence of the HiBSA algorithm or HiBSA. For the ease of presentation, we relegate all the details of the proof to the appendix.

We will first consider a subset of problem (1), where $f(x, y)$ is strongly concave in y. Specifically, we assume the following.

Assumption C-1: For any $x \in X$, $f(\cdot)$ satisfies the following:

$$f(x, z) - f(x, y) \leq \langle \nabla_y f(x, y), z - y \rangle - \frac{\theta}{2}\|z - y\|^2, \forall y, z \in \mathcal{Y},$$

where $\theta > 1$ is the strong concavity constant. Further assume:

$$U_y(u; x, y) = \langle \nabla_y f(x, y), u - y \rangle - \frac{1}{2\rho}\|u - y\|^2, \tag{29}$$

where $\rho > 0$ is some fixed constant. ■

We note that it can be verified that the jamming problem (8) satisfies Assumption C-1. Next we will present a series of lemmas which lead to our main result in this subsection. The detailed proof can be found in Appendix sections A – D.

Lemma 1 (Descent Lemma on x): Suppose that Assumptions A, B and C-1 hold. Let (x^r, y^r) be a sequence generated by HiBSA, with $\gamma^r = 0$, and $\beta^r = \beta > 0, \ \forall r$. Then we have the following descent estimate:

$$\begin{aligned}&\ell(x^{r+1}, y^r) - \ell(x^r, y^r) \\ &\leq -\sum_{i=1}^K \left(\beta + \mu_i - \frac{L_{x_i}}{2}\right) \|x_i^{r+1} - x_i^r\|^2.\end{aligned}$$

Lemma 2 (Descent Lemma on y): Suppose that Assumptions A, B and C-1 hold. Let (x^r, y^r) be a sequence generated by HiBSA, with $\gamma^r = 0$, and $\beta^r = \beta > 0, \ \forall r$. Then we have the following descent estimate:

$$\ell(x^{r+1}, y^{r+1}) - \ell(x^{r+1}, y^r) \leq \frac{1}{\rho}\|y^{r+1} - y^r\|^2$$

$$-\left(\theta-\left(\frac{1}{2\rho}+\frac{\rho L_y^2}{2}\right)\right)\|y^r-y^{r-1}\|^2+\frac{\rho L_y^2}{2}\|x^{r+1}-x^r\|^2.$$

Lemma 3: Suppose that Assumptions A, B and C-1 hold. Let (x^r,y^r) be a sequence generated by HiBSA, with $\gamma^r=0$, and $\beta^r=\beta>0,\ \forall r$. Let us define a potential function as

$$\mathcal{P}^{r+1}:=\ell(x^{r+1},y^{r+1})+\left(\frac{2}{\rho^2\theta}+\frac{1}{2\rho}-4\left(\frac{1}{\rho}-\frac{L_y^2}{2\theta}\right)\right)\|y^{r+1}-y^r\|^2. \tag{30}$$

When the following conditions are satisfied:

$$\rho<\frac{\theta}{4L_y^2},\quad \beta>L_y^2\left(\frac{2}{\theta^2\rho}+\frac{\rho}{2}\right)+\frac{L_{x_i}}{2}-\mu_i,\ \forall i \tag{31}$$

then there exist positive constants $c_1,\{c_{2i}\}_{i=1}^N$ such that:

$$\mathcal{P}^{r+1}-\mathcal{P}^r<-c_1\|y^{r+1}-y^r\|^2-\sum_{i=1}^K c_{2i}\|x_i^{r+1}-x_i^r\|^2. \tag{32}$$

Combining the above analysis, we can obtain the following convergence guarantee for the HiBSA algorithm.

Theorem 1: Suppose that Assumptions A, B, C-1 hold. Let (x^r,y^r) be a sequence generated by HiBSA, with $\gamma^r=0$, and $\beta^r=\beta>0,\ \forall r$, satisfying (31). For a given $\epsilon>0$, let $T(\epsilon)$ denote the first iteration index, such that the following holds:

$$T(\epsilon)\ =\min\{r\mid\|\nabla\mathcal{G}_\rho^\beta(x^{r+1},y^{r+1})\|\le\epsilon,r\ge1\}.$$

Then, $T(\epsilon)=\mathcal{O}(\frac{1}{\epsilon^2})$.

C. Convergence Analysis: $f(x,y)$ Concave in y

Next, we consider the following assumptions for (1).

Assumption C-2: Assume that $f(\cdot)$ in (1) satisfies:

$$f(x,y)-f(x,z)\le\langle\nabla_y f(x,z),y-z\rangle,\ \forall y,z\in\mathcal{Y},x\in\mathcal{X}.$$

That is, it is *concave* in y. Further, assume that

$$U_y(u;x,y)=f(x,u)-\frac{1}{2\rho}\|u-y\|^2. \tag{33}$$

That is, the y update directly maximizes a regularized version of the objective function. Note that $U_y(u;x;y)$ is strongly concave in u, which satisfies the counterpart of Assumption B.1 for $U_y(\cdot)$. ■

Despite the fact that $f(x,y)$ is no longer strongly concave in y, the y-update in **[S2]** is still relatively easy since it maximizes a strongly concave function. However, the absence of strong concavity of $f(x,y)$ in y poses significant challenges in the analysis. In fact, from Example 1 it is clear that directly utilizing the alternating gradient type algorithm may fail to converge to any interesting solutions. Towards resolving this issue, we specialize the HiBSA algorithm, by using a novel *diminishing regularization* plus *increasing penalty* strategy to regularize the y and x update, respectively (by using a sequence of diminishing $\{\gamma^r\}$, and increasing $\{\beta^r\}$).

We have the following convergence analysis. The proofs of the results below can be found in Appendix Section E – G.

Lemma 4 (Descent lemma): Suppose that Assumptions A, B and C-2 hold. Let (x^r,y^r) be a sequence generated by HiBSA, with $\gamma^r>0$ and $\beta^r>L_{x_i},\ \forall r,i$. Then we have:

$$\ell(x^{r+1},y^{r+1})-\ell(x^r,y^r)\le\frac{1}{2\rho}\|y^r-y^{r-1}\|^2-\left(\frac{\beta^r}{2}+\mu-\frac{\rho L_y^2}{2}\right)\|x^{r+1}-x^r\|^2-\left(\frac{\gamma^{r-1}}{2}-\frac{1}{\rho}\right)\|y^{r+1}-y^r\|^2+\frac{\gamma^r}{2}\|y^{r+1}\|^2-\frac{\gamma^{r-1}}{2}\|y^r\|^2+\frac{\gamma^{r-1}-\gamma^r}{2}\|y^{r+1}\|^2. \tag{34}$$

Next we show that there exists a potential function, given below, which decreases consistently

$$\mathcal{P}^{r+1}:=\left(\frac{1}{2\rho}+\frac{2}{\rho^2\gamma^r}+\frac{2}{\rho}\left(\frac{1}{\rho\gamma^{r+1}}-\frac{1}{\rho\gamma^r}\right)\right)\|y^{r+1}-y^r\|^2+\ell(x^{r+1},y^{r+1})-\frac{\gamma^r}{2}\|y^{r+1}\|^2-\frac{2}{\rho}\left(\frac{\gamma^{r-1}}{\gamma^r}-1\right)\|y^{r+1}\|^2. \tag{35}$$

Lemma 5: Suppose that Assumptions A, B and C-2 are satisfied. Let (x^r,y^r) be a sequence generated by HiBSA. Suppose the following conditions are satisfied for all r,

$$\beta^r>\rho L_y^2+\frac{4L_y^2}{\rho(\gamma^r)^2}-2\mu,\ \ \beta^r>L_{x_i},\forall i,\ \ \frac{1}{\gamma^{r+1}}-\frac{1}{\gamma^r}\le\frac{\rho}{5}, \tag{36}$$

then the change of potential function can be bounded through

$$\mathcal{P}^{r+1}\le\mathcal{P}^r-\left(\frac{\beta^r}{2}+\mu-\left(\frac{\rho L_y^2}{2}+\frac{2L_y^2}{\rho(\gamma^r)^2}\right)\right)\|x^{r+1}-x^r\|^2-\frac{1}{10\rho}\|y^{r+1}-y^r\|^2+\frac{\gamma^{r-1}-\gamma^r}{2}\|y^{r+1}\|^2+\frac{2}{\rho}\left(\frac{\gamma^{r-2}}{\gamma^{r-1}}-\frac{\gamma^{r-1}}{\gamma^r}\right)\|y^r\|^2. \tag{37}$$

Before proving the main result in this section, we make the following assumptions on the parameter choices.

Assumption C-3: Suppose that the following conditions hold:

1) The sequence $\{\gamma^r\}$ satisfies

$$\gamma^r-\gamma^{r+1}\ge0,\quad \gamma^r\to0,\qquad \sum_{r=1}^\infty(\gamma^r)^2=\infty,\quad \frac{1}{\gamma^{r+1}}-\frac{1}{\gamma^r}\le\frac{\rho}{5}. \tag{38}$$

2) The sequence β^r satisfies

$$\beta^r>\rho L_y^2+\frac{4L_y^2}{\rho(\gamma^r)^2}-2\mu,\quad \beta^r>L_{x_i},\ \forall i. \tag{39}$$

Note that the above assumption on $\{\gamma^r\}$ can be satisfied, for example, when $\gamma^r=\frac{1}{\rho r^{1/4}}$; see the discussion after (69). ■

Theorem 2: Suppose that Assumptions A, B, C-2 and C-3 hold. Let (x^r,y^r) be a sequence generated by HiBSA. For a given $\epsilon>0$, let $T(\epsilon)$ be defined similarly as in Theorem 1. Then, $T(\epsilon)=\widetilde{\mathcal{O}}(\frac{1}{\epsilon^4})$.

It is important to note that, when the problem is only concave in y, the condition (33) asserts that in each step a strongly concave problem has to be solved exactly. However, for a generic objective function, this step does not involve a closed-form solution. In the supplementary material accompanying this paper [40], we extend this algorithm to the case where the maximization problem is solved by performing a finite number of gradient ascent steps.

D. Convergence Analysis: $f(x, y)$ Linear in y

Finally, we briefly discuss the case where the coupling term in (1) is linear in y. The derivation of the results in this section largely follows from what we have presented in Section III-C, therefore we choose to omit it.

Assumption C-4: Assume that problem (1) simplifies to:

$$\min_x \max_y \quad y^T F(x_1, x_2, \ldots, x_K) + \sum_{i=1}^{K} h_i(x_i) - g(y) \qquad (40)$$
$$\text{s.t.} \quad x_i \in \mathcal{X}_i,\ y \in \mathcal{Y},\ i = 1, \ldots, K$$

where $F(\cdot) : \mathbb{R}^{NK} \to \mathbb{R}^M$ is a vector function. Further assume that (33) holds for $U_y(\cdot)$. ■

Note that (40) contains the robust learning problem (4), the min utility maximization problem (5), and Example 1 as special cases. It is worth noting that, due to the use of the strongly concave approximation function $U_y(u; x, y)$ as defined in (33), we are able to perform a simple gradient step to update y, while in the algorithm proposed in the previous section, each iteration has to solve an optimization problem involving y.

It is worth mentioning that, in this case the analysis steps are similar to those in Section III-C. In particular, we can show that the potential function (35) has the same behavior as in Lemma 5. Therefore, we state our convergence result in the following corollary.

Corollary 3.1: Suppose that Assumptions A, B, C-3 and C-4 hold. Let (x^r, y^r) be a sequence generated by HiBSA. For a given $\epsilon > 0$, let $T(\epsilon)$ be defined as in Theorem 1. Then, $T(\epsilon) = \widetilde{\mathcal{O}}(\frac{1}{\epsilon^4})$.

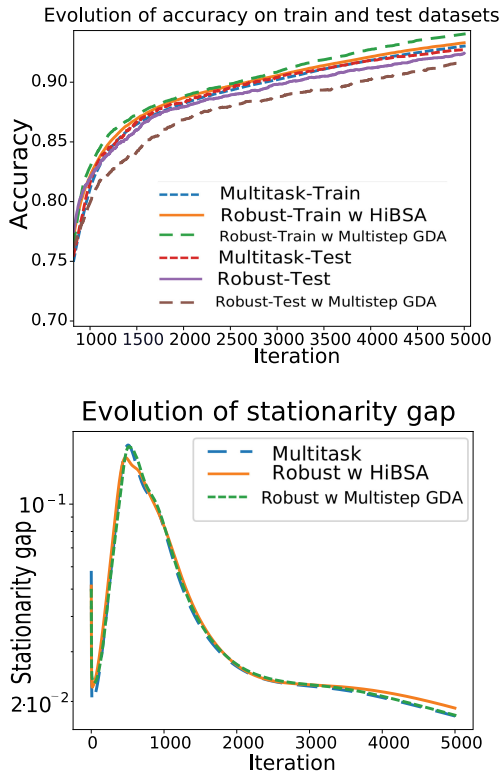


Fig. 2. The results on the experiments performed on the MNIST dataset [50]. The top figure depicts training and testing accuracies, while the second one depicts the convergence behavior of the two algorithms.

IV. Numerical Results

We test our algorithms on three applications: a robust learning problem, a rate maximization problem in the presence of a jammer and a coordinated beamforming problem.

Robust learning over multiple domains. Consider a scenario where we have datasets from two different domains and adopt a neural network model in order to solve a multi-class classification problem. The neural network consists of two hidden layers with 50 neurons, each endowed with sigmoid activations, except from the output layer where we adopt the softmax activation. We aim to learn the model parameters using the following two approaches:

[1] *Robust Learning:* Apply the robust learning model (4) and optimize the cost function using the HiBSA algorithm with $\gamma^r = \frac{1}{r^{1/4}}$ and the Multi-step GDA algorithm [36] with one gradient descent and five gradient ascent steps per iteration. Note that we treat the minimization variable as one block and use the first-order Taylor expansion of the cost function as the approximation function.

[2] *Mutltitask Learning:* Apply a multitask learning model [49], where we optimize the sum of the respective empirical risks correspsonding to the two domains/tasks; the weights associated with each task are fixed to 1/2. The problem is optimized using gradient descent.

Moreover, we evaluate the above algorithms by using the minimum accuracy across the two domains, over both training and test datasets. That is, accuracy $= \min\{$accuracy on domain 1, accuracy on domain 2$\}$.

In our experiments we use the MNIST [50] dataset whose data points are images of handwritten digits of dimensions 28×28. We select two different parts of the MNIST dataset as the two different domains we mentioned above. The first part consists of the digits from 0 to 4, while the second one contains the rest. Moreover, for the first domain we use 5000 images for training and 1000 for testing, while in the second one we employ 25000 and 5000 images respectively. Finally, we average the results over 5 iterations.

Note that we do not perform extensive parameter tuning, since the purpose of this experiment is not to support the superiority of the robust model, but merely to illustrate that the proposed HiBSA computes a reasonable model similar to what can be computed by multistep GDA, and to what can be obtained by multi-task learning. Indeed, the results presented in Fig. 2 support this view, since different approaches achieve approximately the same accuracy on the test set.

Power control in the presence of a jammer: Consider the multi-channel and multi-user formulation (8) where there are N channels, K collaborative users and one jammer. We can verify that the jammer problem (i.e., the maximization problem over y) has a strongly concave objective function over the feasible set.

We compare HiBSA with the classic interference pricing algorithm [51], [52], and the WMMSE algorithm [53], which are designed for solving the sum-rate optimization problem *without* the jammer. Our problem is tested using the following setting. We construct a network with $K = 10$, and the interference channel among the users and the jammer is generated using the uncorrelated fading channel model with channel coefficients generated from the complex zero-mean Gaussian distribution with unit covariance [53]. All users' power budget is fixed at $P = 10^{\text{SNR}/10}$. For test cases without a jammer, we set $\sigma_k^2 = 1$ for all k. For test cases with a jammer, we set $\sigma_k^2 = 1/2$ for

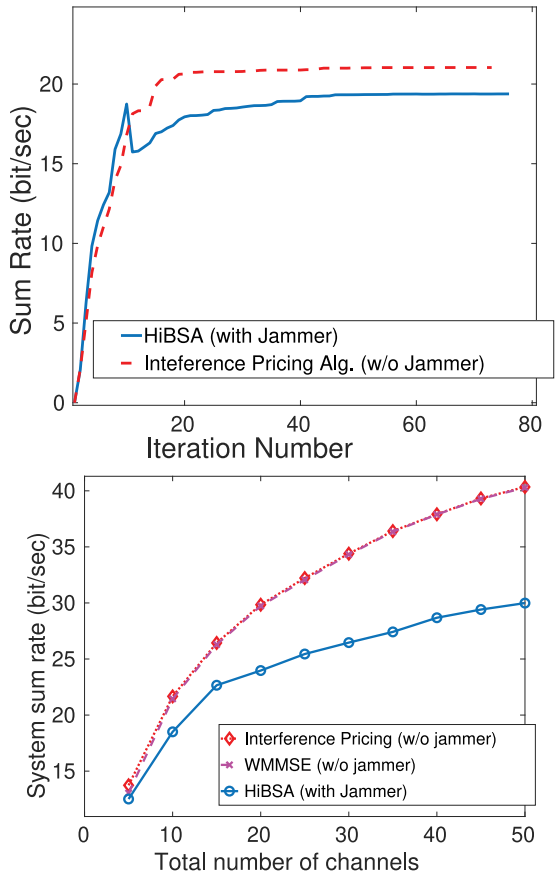


Fig. 3. The convergence curves and total averaged system performance comparing three algorithms: WMMSE, Interference Pricing, and HiBSA. The first figure shows a single realization of the algorithms, and in the second figure, each point represents an average of 50 realizations. The total number of the users is 10, and $\mathrm{SNR} = 1$.

all k, and let the jammer have the rest of the noise power, i.e., $p_{0,\max} = N/2$. Note that by splitting the noise power we intend to achieve some fair comparison between the cases with and without the jammer. However, it is not possible to be completely fair because even though the total noise budgets are the same, the noise power transmitted by the jammer has to go through the random channel, so the total received noise power could be different. Nevertheless, this setting is sufficient to demonstrate the behavior of the HiBSA algorithm.

From the Fig. 3 (top), it is clear that the pricing algorithm monotonically increases the sum rate (as is predicted by theory), while HiBSA behaves differently: after some initial oscillation, the algorithm converges to a value that has a lower sum-rate. Further in Fig. 3 (bottom), we do see that by using the proposed algorithm, the jammer is able to effectively reduce the total sum rate of the system.

Coordinated MISO beamforming design: Consider the coordinated beamforming design problem [14] described in Section I over a MISO interference channel. In this problem we experiment with the scenario where there are $K = 10$ transmitter-receiver pairs, each transmitter is equipped with $N = 6$ antennas. We adopt the min-rate utility, i.e., $U(\{R_i(x)\}_{i=1}^K) = \min_{i=1,\ldots,K}\{R_i(x)\}$. Moreover, the transmission is performed over a complex Gaussian channel, and we set the power budget to be $\bar{p} = 1$. The channel covariance matrices $\{C_{ij}\}, i, j = 1, \ldots, 10$ are generated at random and their maximum eigenvalues are normalized to 1, if $i = j$, and to some constant $\lambda > 0$, if $i \neq j$. Thus, the parameter λ quantifies the level of intereference.

The problem of interest is to design the users' beamformers in order to maximize the system's utility function under constraints in power and outage probability. We approach the solution of the problem using two different algorithms:

[1] *BSUM-LSE [14]:* Substitute the min-rate utility function with a popular log-sum-exp approximation, i.e.,

$$\min_{i=1,\ldots,K}\{R_i(x)\} := r_{\min}(x) \approx \frac{1}{\nu}\log_2\left(\sum_{i=1}^{K} 2^{-\nu R_i}\right).$$

Note that ν specifies the level of approximation with higher ν's corresponding to tighter bounds for the approximation error. Then following what is suggested in [[14], Section C], we formulate the respective problem using the surrogate function, and solve the resulting problem iteratively using the projected gradient descent.

[2] *HiBSA:* We apply the HiBSA to solve the formulation in (5). The x-subproblem is solved similarly as in BSUM-LSE. Moreover, in the maximization problem we use $\gamma^r = 1/r^{1/4}$.

We run both algorithms for 1000 complete iterations (one complete iteration involves one update of *all the block variables*, and 1000 iterations are sufficient for both algorithms to converge in all scenarios) and set the stepsizes β and ρ of HiBSA and the respective stepsize of BSUM-LSE all equal to 10^{-2}. We also average the final results over 10 independent random problem instances. Moreover, in order to evaluate the effect of the log-sum approximation we show the achieved min-rate utility of BSUM-LSE, by using 3 different values of $\nu \in \{1, 5, 7\}$.

In Fig. 4 we plot the min-rate utility for 7 different values of the noise variance and 2 different levels of interference. Notice that the HiBSA algorithm achieves higher utility than BSUM-LSE, while as expected the larger the value of ν the higher the utility achieved by the latter algorithm.

Furthermore, since large values of the parameter ν lead to low approximation error bounds, it is of interest to consider experiments with large ν for the BSUM-LSE algorithm. Intuitively, we expect the resulting objective to be very close to the min-rate utility and thus the achieved min-rate of the BSUM-LSE algorithm should approach the respective min-rate of HiBSA. In order to determine the behavior of BSUM-LSE in that range of ν's we consider an experiment with $K = 10, N = 6$ and $\lambda = 0.6$. Regarding the stepsizes we keep them constant across the different values of $1/\sigma^2$, however an effort was made to select the optimal ones for all algorithms in order to ensure fair comparisons. Moreover, we terminate both algorithms when the relative successive differences of the min-rate utility becomes small, i.e., $|r_{\min}(x^{r+1}) - r_{\min}(x^r)|/|r_{\min}(x^r)| \leq 10^{-7}$, or the number of iterations becomes larger than 5,000. Finally, the results are provided in Fig. 5.

Note that, for large values of ν, i.e., $\nu = 50, 100, 1000$, the achieved rate of BSUM-LSE is close but still inferior to that of HiBSA. Additionally, for the same ν's the HiBSA is faster than BSUM-LSE; in fact the larger the ν is the longer the runtime will be. On the other hand, the former algorithm is in general slower than BSUM-LSE with $\nu = 5$, however in that case HiBSA achieves higher min-rate utility. Overall, note that even though large ν leads (in most cases) to improvements in the attained min-rate utility, it also incurs longer runtimes. This can be attributed to the fact that for high ν the log-sum-exp

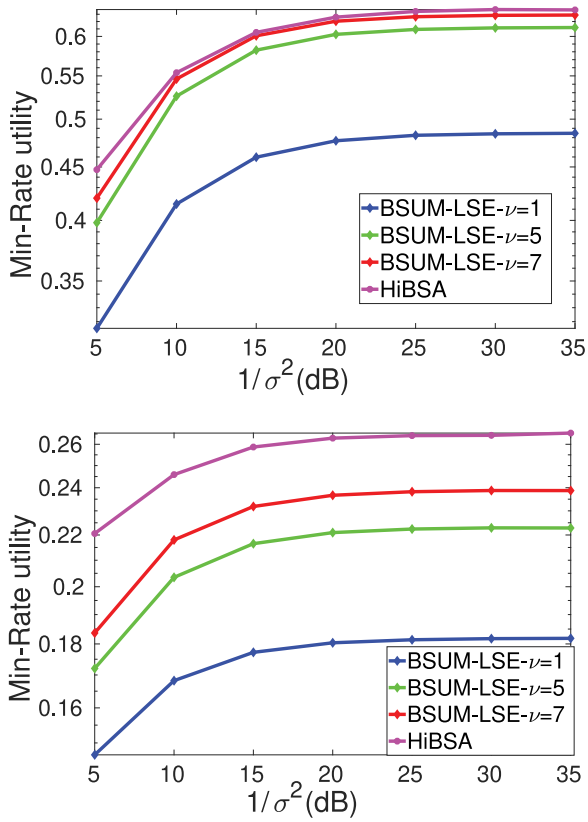


Fig. 4. The min-rate utility achieved using the HiBSA and the BSUM-LSE algorithm [14] for two different interference levels in a scenario where we have $K = 10$ users equipped with $N = 6$ antennas. Note that the top figure corresponds to a lower interference level ($\lambda = 0.6$) than the bottom one ($\lambda = 1$). For each interference level we experiment with 3 different values of the level of approximation (ν). Finally, σ^2 is the noise power at the receivers.

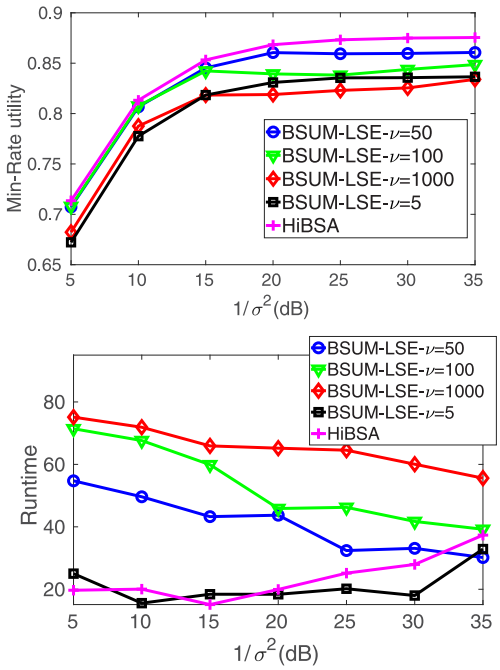


Fig. 5. The achieved min-rate utility (top) and runtime (bottom) for the BSUM-LSE algorithm, with $\nu = 5, 50, 100, 1000$, and the HiBSA algorithm. The results are averaged over 30 independent runs.

objective approaches a non-smooth function, which is difficult to optimize. In conclusion, HiBSA in general outperforms BSUM-LSE in terms of runtime and attained min-rate utility.

V. Conclusion

In this paper, motivated by the min-max problems arising in the areas of signal processing and wireless communications, we propose an algorithm called HiBSA. By leveraging the (strong) concavity of the maximization problem, we conduct analysis on the convergence behavior of the proposed algorithm. Numerical results show the effectiveness of the proposed algorithms for solving the min-max problems in robust machine learning and wireless communications. There are many potential future research directions we plan to explore. For example, it will be interesting to develop algorithms for more challenging problems where the y problem is also non-convex. Further, it will also be interesting to establish some lower complexity bounds for non-convex and/or non-concave min-max problems, which characterize the *best* performance one can achieve when optimizing such a family of problems.

Appendix

A. Proof of Lemma 1

By using the assumption that f has Lipschitz gradient, h_i is convex (cf. Assumption A), and by noticing that $w_i^{r+1} = (x_i^r, w_{-i}^{r+1})$, we obtain the following:

$$\ell(x_i^{r+1}, w_{-i}^{r+1}, y^r) - \ell(x_i^r, w_{-i}^{r+1}, y^r)$$
$$\leq \langle \nabla_{x_i} f(w_i^{r+1}, y^r) + \vartheta_i^{r+1}, x_i^{r+1} - x_i^r \rangle + \frac{L_{x_i}}{2} \|x_i^{r+1} - x_i^r\|^2 \tag{41}$$

for some $\vartheta_i^{r+1} \in \partial h_i(x_i^{r+1})$.

Second, the optimality condition for the x_i update step (12) is

$$\langle \nabla_{x_i} U_i(x_i^{r+1}; w_i^{r+1}, y^r) + \vartheta_i^{r+1} + \beta(x_i^{r+1} - x_i^r), x_i^r - x_i^{r+1} \rangle \geq 0. \tag{42}$$

So adding and subtracting $\langle \nabla_{x_i} U_i(x_i^r; w_i^{r+1}, y^r), x_i^r - x_i^{r+1} \rangle$ in (42), and by applying assumptions B.1 (strong convexity) and B.2 (gradient consistency), we obtain the following:

$$\langle \nabla_{x_i} f(w_i^{r+1}, y^r), x_i^{r+1} - x_i^r \rangle + \langle \vartheta_i^{r+1}, x_i^{r+1} - x_i^r \rangle$$
$$\leq -\mu_i \|x_i^{r+1} - x_i^r\|^2 - \beta \|x_i^{r+1} - x_i^r\|^2.$$

Then, combining the above expression with (41) results in

$$\ell(x_i^{r+1}, w_{-i}^{r+1}, y^r) - \ell(x_i^r, w_{-i}^{r+1}, y^r)$$
$$\leq \left(-\beta - \mu_i + \frac{L_{x_i}}{2} \right) \|x_i^{r+1} - x_i^r\|^2.$$

Summing over $i \in [K]$ we obtain the desired result. **Q.E.D.**

B. Proof of Lemma 2

For notational simplicity, let us define

$$\ell'(x^{r+1}, y) = f(x^{r+1}, y) + \sum_{i=1}^{K} h_i(x_i^{r+1}) - \mathcal{I}_{\mathcal{Y}}(y) - g(y).$$

Notice that for any $y \in \mathcal{Y}$, we have $\ell'(x^{r+1}, y) = \ell(x^{r+1}, y)$. The optimality condition of the y-step (13) becomes

$$0 = \nabla_y f(x^{r+1}, y^r) - \frac{1}{\rho}(y^{r+1} - y^r) - \xi^{r+1}, \tag{43}$$

where $\xi^{r+1} \in \partial(\mathcal{I}_{\mathcal{Y}}(y^{r+1}) + g(y^{r+1}))$ is a subgradient vector. Since $\ell'(x, y)$ is concave with w.r.t. y, we have

$$\begin{aligned}
&\ell'(x^{r+1}, y^{r+1}) - \ell'(x^{r+1}, y^r) \\
&\leq \langle \nabla_y f(x^{r+1}, y^r) - \xi^r, y^{r+1} - y^r \rangle \\
&\overset{(43)}{=} \frac{1}{\rho}\|y^{r+1} - y^r\|^2 + \langle \xi^{r+1} - \xi^r, y^{r+1} - y^r \rangle \\
&\overset{(a)}{=} \frac{1}{\rho}\|y^{r+1} - y^r\|^2 + \langle \nabla_y f(x^{r+1}, y^r) \\
&\quad - \nabla_y f(x^r, y^{r-1}), y^{r+1} - y^r \rangle \\
&\quad - \frac{1}{\rho}\langle y^{r+1} - y^r - (y^r - y^{r-1}), y^{r+1} - y^r \rangle \\
&\overset{(b)}{=} \frac{1}{\rho}\|y^{r+1} - y^r\|^2 + \langle \nabla_y f(x^{r+1}, y^r) \\
&\quad - \nabla_y f(x^r, y^{r-1}), y^{r+1} - y^r \rangle \\
&\quad + \frac{1}{2\rho}\|y^r - y^{r-1}\|^2 - \frac{1}{2\rho}\|y^{r+1} - y^r\|^2 - \frac{1}{2\rho}\|v^{r+1}\|^2 \\
&\overset{(c)}{\leq} \frac{1}{\rho}\|y^{r+1} - y^r\|^2 + \frac{\rho L_y^2}{2}\|x^{r+1} - x^r\|^2 + \frac{1}{2\rho}\|y^r - y^{r-1}\|^2 \\
&\quad - \frac{1}{2\rho}\|v^{r+1}\|^2 + \langle \nabla_y f(x^r, y^r) \\
&\quad - \nabla_y f(x^r, y^{r-1}), y^{r+1} - y^r \rangle
\end{aligned} \tag{44}$$

where (a) follows from the optimality conditions of the y-step (13) at iterations $r+1$ and r; in (b) we apply the following identity:

$$\begin{aligned}
&\langle v^{r+1}, y^{r+1} - y^r \rangle \\
&= \frac{1}{2}\left(\|y^{r+1} - y^r\|^2 + \|v^{r+1}\|^2 - \|y^r - y^{r-1}\|^2\right),
\end{aligned} \tag{45}$$

where we have defined

$$v^{r+1} := y^{r+1} - y^r - (y^r - y^{r-1}); \tag{46}$$

in (c) we add and subtract a term $\langle \nabla_y f(x^r, y^r), y^{r+1} - y^r \rangle$, and apply the Young's inequality and obtain:

$$\begin{aligned}
&\langle \nabla_y f(x^{r+1}, y^r) - \nabla_y f(x^r, y^r), y^{r+1} - y^r \rangle \\
&\leq \frac{\rho L_y^2}{2}\|x^{r+1} - x^r\|^2 + \frac{1}{2\rho}\|y^{r+1} - y^r\|^2,
\end{aligned} \tag{47}$$

where L_y is defined in (11a). By applying the strong concavity of $f(x, y)$ in y, the Young's inequality and the Lipschitz condition w.r.t y, we can have the following bound for the inner product term in (44):

$$\begin{aligned}
&\langle \nabla_y f(x^r, y^r) - \nabla_y f(x^r, y^{r-1}), y^{r+1} - y^r \rangle \\
&\leq \langle \nabla_y f(x^r, y^r) - \nabla_y f(x^r, y^{r-1}), v^{r+1} + y^r - y^{r-1} \rangle \\
&\leq \frac{\rho L_y^2}{2}\|y^r - y^{r-1}\|^2 + \frac{1}{2\rho}\|v^{r+1}\|^2 - \theta\|y^r - y^{r-1}\|^2.
\end{aligned} \tag{48}$$

Combining the above with (44) completes the proof. **Q.E.D.**

At this point, by simply combining Lemmas 1–2, it is not clear how the objective value behaves after each x and y update. To capture the essential dynamics of the algorithm, the key is to identify a proper potential function, which decreases after each round of x and y updates.

C. Proof of Lemma 3

According to (43), the optimality condition of y-problem (13) at iterations $r+1$ and r are given by:

$$-\nabla_y f(x^{r+1}, y^r) + \xi^{r+1} + \frac{1}{\rho}(y^{r+1} - y^r) = 0, \tag{49}$$

$$-\nabla_y f(x^r, y^{r-1}) + \xi^r + \frac{1}{\rho}(y^r - y^{r-1}) = 0, \tag{50}$$

where $\xi^r \in \partial(\mathcal{I}_{\mathcal{Y}}(y^r) + g(y^r))$. We subtract these two equalities, multiply both sides by $y^{r+1} - y^r$, utilize the defining property of subgradient vectors: $\langle \xi^{r+1} - \xi^r, y^{r+1} - y^r \rangle \geq 0$, and we obtain:

$$\begin{aligned}
&\frac{1}{\rho}\langle v^{r+1}, y^{r+1} - y^r \rangle \\
&\leq \langle \nabla_y f(x^{r+1}, y^r) - \nabla_y f(x^r, y^r), y^{r+1} - y^r \rangle \\
&\quad + \langle \nabla_y f(x^r, y^r) - \nabla_y f(x^r, y^{r-1}), y^{r+1} - y^r \rangle,
\end{aligned}$$

where v^{r+1} is defined in (46). Applying (45) to the LHS to the above expression, and using similar techniques as in (47), (48) for the RHS of the above expression (note that this time we use a constant θ instead of ρ, when applying (47)), we obtain the following:

$$\begin{aligned}
&\frac{1}{2\rho}\|y^{r+1} - y^r\|^2 \\
&\leq \frac{1}{2\rho}\|y^r - y^{r-1}\|^2 - \frac{1}{2\rho}\|v^{r+1}\|^2 + \frac{L_y^2}{2\theta}\|x^{r+1} - x^r\|^2 \\
&\quad + \frac{\theta}{2}\|y^{r+1} - y^r\|^2 + \frac{\rho L_y^2}{2}\|y^r - y^{r-1}\|^2 + \frac{1}{2\rho}\|v^{r+1}\|^2 \\
&\quad - \theta\|y^r - y^{r-1}\|^2 \\
&= \frac{1}{2\rho}\|y^r - y^{r-1}\|^2 + \frac{L_y^2}{2\theta}\|x^{r+1} - x^r\|^2 + \frac{\theta}{2}\|y^{r+1} - y^r\|^2 \\
&\quad - \left(\theta - \frac{\rho L_y^2}{2}\right)\|y^r - y^{r-1}\|^2.
\end{aligned} \tag{51}$$

By combining Lemmas 1 and 2 we obtain

$$\begin{aligned}
&\ell(x^{r+1}, y^{r+1}) - \ell(x^r, y^r) \\
&\leq -\sum_{i=1}^{K}\left(\beta + \mu_i - \frac{L_{x_i}}{2} - \frac{\rho L_y^2}{2}\right)\|x_i^{r+1} - x_i^r\|^2 \\
&\quad + \frac{1}{\rho}\|y^{r+1} - y^r\|^2 - \left(\theta - \left(\frac{1}{2\rho} + \frac{\rho L_y^2}{2}\right)\right)\|y^r - y^{r-1}\|^2.
\end{aligned}$$

Multiplying both sides of (51) by $4/(\theta\rho)$, and adding the resulting inequality to the above expression, we have

$$\ell(x^{r+1}, y^{r+1}) + \frac{2}{\rho^2\theta}\|y^{r+1} - y^r\|^2$$

$$\leq \ell(x^r, y^r) + \frac{2}{\rho^2\theta}\|y^r - y^{r-1}\|^2 + \frac{3}{\rho}\|y^{r+1} - y^r\|^2$$
$$- \sum_{i=1}^{K}\left(\beta + \mu_i - \frac{L_{x_i}}{2} - \left(\frac{2L_y^2}{\theta^2\rho} + \frac{\rho L_y^2}{2}\right)\right)\|x_i^{r+1} - x_i^r\|^2$$
$$- \left(\theta - \left(\frac{1}{2\rho} + \frac{\rho L_y^2}{2}\right)\right)\|y^r - y^{r-1}\|^2$$
$$- \frac{4}{\theta\rho}\left(\theta - \frac{\rho L_y^2}{2}\right)\|y^r - y^{r-1}\|^2.$$

Finally, adding in both sides the term $(\frac{1}{2\rho} - 4(\frac{1}{\rho} - \frac{L_y^2}{2\theta}))\|y^{r+1} - y^r\|^2$ and using the definition of the potential function (30), we obtain the following

$$\mathcal{P}^{r+1} \leq \mathcal{P}^r + \left(\frac{3}{\rho} + \frac{1}{2\rho} - 4\left(\frac{1}{\rho} - \frac{L_y^2}{2\theta}\right)\right)\|y^{r+1} - y^r\|^2$$
$$- \sum_{i=1}^{K}\left(\beta + \mu_i - \frac{L_{x_i}}{2} - \left(\frac{2L_y^2}{\theta^2\rho} + \frac{\rho L_y^2}{2}\right)\right)\|x_i^{r+1} - x_i^r\|^2.$$

In the inequality above we do not include $\theta - \rho L_y^2/2$ (from RHS of the descent estimate in Lemma 1) because by the choice of ρ this term is positive. Therefore, when

$$\rho < \frac{\theta}{4L_y^2}, \quad \beta > L_y^2\left(\frac{2}{\theta^2\rho} + \frac{\rho}{2}\right) + \frac{L_{x_i}}{2} - \mu_i, \ \forall i \qquad (52)$$

we have sufficient descent of the potential function $\mathcal{P}^{r+1}$. This completes the proof. **Q.E.D.**

D. Proof of Theorem 1

We first bound the ith block of the optimality gap (16) by

$$\|(\nabla\mathcal{G}_\rho^\beta(x^r, y^r))_i\|$$
$$\leq \beta\|x_i^{r+1} - x_i^r\| + \beta\|x_i^{r+1} - \mathrm{Px}_i^\beta(x_i^r - 1/\beta\nabla_{x_i}f(x^r, y^r))\|$$
$$\overset{(a)}{\leq} \beta\|x_i^{r+1} - x_i^r\| + \beta\|\mathrm{Px}_i^\beta\left(x_i^r - \left(\frac{1}{\beta}\nabla_{x_i}U_i(x_i^{r+1}; w_i^{r+1}, y^r)\right)\right)$$
$$- \mathrm{Px}_i^\beta(x_i^r - \frac{1}{\beta}\nabla_{x_i}f(x^r, y^r))\|$$
$$\overset{(b)}{\leq} \beta\|x_i^{r+1} - x_i^r\| + L_{u_i}\|x_i^{r+1} - x_i^r\| + L_{x_i}\|w_i^{r+1} - x^r\|$$
$$\leq (\beta + L_{u_i} + L_{x_i})\|x^{r+1} - x^r\|,$$

where in (a) we use the optimality conditions w.r.t to x_i in (12); in (b) we use the nonexpansiveness of the proximal operator, $\nabla_{x_i}U_i(x_i^r; w_i^{r+1}, y^r) = \nabla_{x_i}f_i(w_i^{r+1}, y^r)$ (Assumption B2), Assumption B4 (Lipschitz gradient), as well as the following identity

$$\nabla_{x_i}U_i(x_i^{r+1}; w_i^{r+1}, y^r) - \nabla_{x_i}f(x^r, y^r)$$
$$= \nabla_{x_i}U_i(x_i^{r+1}; w_i^{r+1}, y^r) - \nabla_{x_i}U_i(x_i^r; w_i^{r+1}, y^r)$$
$$+ \nabla_{x_i}U_i(x_i^r; w_i^{r+1}, y^r) - \nabla_{x_i}f(x^r, y^r).$$

Moreover, utilizing the same argument for the optimality condition w.r.t to y problem (13), we obtain:

$$\|(\nabla\mathcal{G}_\rho^\beta(x^r, y^r))_{K+1}\|$$
$$\overset{(a)}{\leq} \frac{1}{\rho}\|y^{r+1} - y^r\| + \frac{1}{\rho}\|\mathrm{Py}^{1/\rho}(y^r + \rho\nabla_y f(x^{r+1}, y^r))$$
$$- \mathrm{Py}^{1/\rho}(y^r + \rho\nabla_y f(x^r, y^r))\|$$
$$\overset{(b)}{\leq} \frac{1}{\rho}\|y^{r+1} - y^r\| + \|\nabla_y f(x^{r+1}, y^r) - \nabla_y f(x^r, y^r)\|$$
$$\overset{(c)}{\leq} L_y\|x^{r+1} - x^r\| + \frac{1}{\rho}\|y^{r+1} - y^r\|,$$

where in (a) we use the optimality conditions w.r.t y, in (b) we use the nonexpansiveness of the proximal operator and finally in (c) the Assumption A.3. Combining (32) and the above two inequalities, we see that there exist constants $\sigma_1 > 0$ and $\sigma_2 > 0$ such that the following holds:

$$\|\nabla\mathcal{G}_\rho^\beta(x^r, y^r)\|^2 \leq \frac{\sigma_2}{\sigma_1}(\mathcal{P}^r - \mathcal{P}^{r+1}). \qquad (53)$$

Summing the above inequality over $r \in [T]$, we have

$$\sum_{r=1}^{T}\|\nabla\mathcal{G}_\rho^\beta(x^r, y^r)\|^2 \leq \frac{\sigma_2}{\sigma_1}(\mathcal{P}^1 - \mathcal{P}^{T+1}) \leq \frac{\sigma_2}{\sigma_1}(\mathcal{P}^1 - \underline{\ell}), \qquad (54)$$

where in the last inequality we have used the fact that $\mathcal{P}^r$ is decreasing (by Lemma 3) and lower bounded by $\underline{\ell}$. The latter fact is because, when condition (31) holds true, the coefficient in front of $\|y^{r+1} - y^r\|^2$ is positive, therefore $\mathcal{P}^{r+1}$ is lower bounded by $\ell(x^{r+1}, y^{r+1})$, according to Assumption A1. By utilizing the definition $T(\epsilon)$, the above inequality becomes $T(\epsilon)\epsilon^2 \leq \frac{\sigma_2}{\sigma_1}(\mathcal{P}^1 - \underline{\ell})$.

Dividing both sides by ϵ^2, the desired result is obtained. **Q.E.D.**

E. Proof of Lemma 4

Following similar steps as in Lemma 1 and using the assumption $\beta^r > L_{x_i}, \forall i$ we obtain

$$\ell(x^{r+1}, y^r) - \ell(x^r, y^r) \leq -\left(\frac{\beta^r}{2} + \mu\right)\|x^{r+1} - x^r\|^2, \qquad (55)$$

where $\mu := \min_{i\in[K]}\mu_i$. To analyze the y update, define

$$\ell'(x^{r+1}, y) = f(x^{r+1}, y) + \sum_{i=1}^{K}h_i(x_i^{r+1}) - \mathcal{I}_\mathcal{Y}(y) - g(y).$$

The optimality condition for the y update is

$$\xi^{r+1} - \nabla_y f(x^{r+1}, y^{r+1}) + \frac{1}{\rho}(y^{r+1} - y^r) + \gamma^r y^{r+1} = 0, \qquad (56)$$

where $\xi^{r+1} \in \partial(\mathcal{I}_\mathcal{Y}(y^{r+1}) + g(y^{r+1}))$. Using this, we have the following series of inequalities:

$$\ell'(x^{r+1}, y^{r+1}) - \ell'(x^{r+1}, y^r)$$
$$\overset{(a)}{\leq} \langle\nabla_y f(x^{r+1}, y^r), y^{r+1} - y^r\rangle - \langle\xi^r, y^{r+1} - y^r\rangle$$
$$\overset{(b)}{=} \langle\nabla_y f(x^{r+1}, y^r) - \nabla_y f(x^{r+1}, y^{r+1}), y^{r+1} - y^r\rangle$$
$$+ \frac{1}{\rho}\|y^{r+1} - y^r\|^2 + \gamma^r\langle y^{r+1}, y^{r+1} - y^r\rangle$$
$$+ \langle\xi^{r+1} - \xi^r, y^{r+1} - y^r\rangle$$

$$\begin{aligned}
&\overset{(c)}{=} \gamma^{r-1}\langle y^r, y^{r+1}-y^r\rangle + \frac{1}{\rho}\|y^{r+1}-y^r\|^2 - \frac{1}{\rho}\langle v^{r+1}, y^{r+1}-y^r\rangle \\
&\quad + \langle \nabla_y f(x^{r+1}, y^r) - \nabla_y f(x^r, y^r), y^{r+1}-y^r\rangle \\
&\overset{(d)}{\leq} \frac{1}{2\rho}\|y^r - y^{r-1}\|^2 + \frac{\rho L_y^2}{2}\|x^{r+1}-x^r\|^2 \\
&\quad - \left(\frac{\gamma^{r-1}}{2} - \frac{1}{\rho}\right)\|y^{r+1}-y^r\|^2 \\
&\quad + \frac{\gamma^r}{2}\|y^{r+1}\|^2 - \frac{\gamma^{r-1}}{2}\|y^r\|^2 + \frac{\gamma^{r-1}-\gamma^r}{2}\|y^{r+1}\|^2,
\end{aligned} \tag{57}$$

where (a) uses the concavity of $\ell'(x,y)$; in (b) we use (56); (c) follows from (56), the optimality condition for y at iteration r and plugging the resulting $\xi^{r+1}-\xi^r$; in (d) we use the quadrilateral identity (45) for the term involving v and ignore the resulting negative term $-\frac{1}{2\rho}\|v^{r+1}\|^2$, the Lipschitz continuity of $\nabla_y f$ (cf. Assumption A. 3), the Young's inequality, as well as the following identity:

$$\begin{aligned}
&\gamma^{r-1}\langle y^r, y^{r+1}-y^r\rangle \\
&= \frac{\gamma^{r-1}}{2}\left(\|y^{r+1}\|^2 - \|y^r\|^2 - \|y^{r+1}-y^r\|^2\right) \\
&= \frac{\gamma^r}{2}\|y^{r+1}\|^2 - \frac{\gamma^{r-1}}{2}(\|y^r\|^2 + \|y^{r+1}-y^r\|^2) \\
&\quad + \left(\frac{\gamma^{r-1}-\gamma^r}{2}\right)\|y^{r+1}\|^2.
\end{aligned}$$

Combining (55) and (57), we obtain the desired result. **Q.E.D.**

F. Proof of Lemma 5

To simplify notation, define $f^{r+1} := f(x^{r+1}, y^{r+1})$. The optimality conditions of y problem are given by

$$\langle \nabla_y f^{r+1} - \frac{1}{\rho}(y^{r+1}-y^r) - \gamma^r y^{r+1} - \vartheta^{r+1}, y^{r+1}-y\rangle \geq 0 \tag{58a}$$

$$\langle \nabla_y f^r - \frac{1}{\rho}(y^r - y^{r-1}) - \gamma^{r-1}y^r - \vartheta^r, y^r - y\rangle \geq 0, \tag{58b}$$

for all $y \in \mathcal{Y}$, where $\vartheta^{r+1} \in \partial g(y^{r+1})$.

Plugging in $y = y^r$ in (58a), $y = y^{r+1}$ in (58b), adding them together and utilizing the defining property of subgradient vectors, i.e $\langle \vartheta^{r+1} - \vartheta^r, y^{r+1}-y^r\rangle \geq 0$, we obtain

$$\begin{aligned}
&\frac{1}{\rho}\langle v^{r+1}, y^{r+1}-y^r\rangle + \langle \gamma^r y^{r+1} - \gamma^{r-1}y^r, y^{r+1}-y^r\rangle \\
&\leq \langle \nabla_y f^{r+1} - \nabla_y f^r, y^{r+1}-y^r\rangle,
\end{aligned} \tag{59}$$

where v^{r+1} is defined in (46). In the following, we will use the above inequality to analyze the recurrence of the size of the difference between two consecutive iterates. First, we have

$$\begin{aligned}
&\langle \gamma^r y^{r+1} - \gamma^{r-1}y^r, y^{r+1}-y^r\rangle \\
&= \langle \gamma^r y^{r+1} - \gamma^r y^r + \gamma^r y^r - \gamma^{r-1}y^r, y^{r+1}-y^r\rangle \\
&= \gamma^r\|y^{r+1}-y^r\|^2 + (\gamma^r - \gamma^{r-1})\langle y^r, y^{r+1}-y^r\rangle \\
&= \gamma^r\|y^{r+1}-y^r\|^2 \\
&\quad + \frac{\gamma^r - \gamma^{r-1}}{2}\left(\|y^{r+1}\|^2 - \|y^r\|^2 - \|y^{r+1}-y^r\|^2\right) \\
&= \frac{\gamma^r + \gamma^{r-1}}{2}\|y^{r+1}-y^r\|^2 - \frac{\gamma^{r-1}-\gamma^r}{2}\left(\|y^{r+1}\|^2 - \|y^r\|^2\right).
\end{aligned} \tag{60}$$

Substituting (60) and (45) into (59), we have

$$\begin{aligned}
&\frac{1}{2\rho}\|y^{r+1}-y^r\|^2 - \frac{\gamma^{r-1}-\gamma^r}{2}\|y^{r+1}\|^2 \\
&\leq \frac{1}{2\rho}\|y^r - y^{r-1}\|^2 - \frac{1}{2\rho}\|v^{r+1}\|^2 - \frac{\gamma^{r-1}-\gamma^r}{2}\|y^r\|^2 \\
&\quad - \frac{\gamma^{r-1}+\gamma^r}{2}\|y^{r+1}-y^r\|^2 + \langle \nabla_y f^{r+1} - \nabla_y f^r, y^{r+1}-y^r\rangle \\
&\overset{(a)}{\leq} \frac{1}{2\rho}\|y^r - y^{r-1}\|^2 - \gamma^r\|y^{r+1}-y^r\|^2 - \frac{\gamma^{r-1}-\gamma^r}{2}\|y^r\|^2 \\
&\quad + \langle \nabla_y f(x^{r+1}, y^r) - \nabla_y f(x^r, y^r), y^{r+1}-y^r\rangle \\
&\overset{(b)}{\leq} \frac{1}{2\rho}\|y^r - y^{r-1}\|^2 - \frac{\gamma^{r-1}-\gamma^r}{2}\|y^r\|^2 \\
&\quad + \frac{L_y^2}{2\gamma^r}\|x^{r+1}-x^r\|^2 - \frac{\gamma^r}{2}\|y^{r+1}-y^r\|^2
\end{aligned}$$

where (a) is true because of the fact that $0 < \gamma^r < \gamma^{r-1}$, which implies that $-\frac{\gamma^{r-1}-\gamma^r}{2} < 0$ and $-\frac{\gamma^{r-1}+\gamma^r}{2} < -\gamma^r$, and the concavity of function $f(x,y)$ in y; in (b) we use the Young's inequality. Next, let us define

$$\mathcal{F}^{r+1} := \frac{1}{2\rho}\|y^{r+1}-y^r\|^2 - \frac{\gamma^{r-1}-\gamma^r}{2}\|y^{r+1}\|^2.$$

Then we have

$$\begin{aligned}
\frac{4\mathcal{F}^{r+1}}{\rho\gamma^r} &\leq \frac{2}{\rho^2\gamma^r}\|y^r - y^{r-1}\|^2 - \frac{2}{\rho}\left(\frac{\gamma^{r-1}}{\gamma^r} - 1\right)\|y^r\|^2 \\
&\quad - \frac{2}{\rho}\|y^{r+1}-y^r\|^2 + \frac{2L_y^2}{\rho(\gamma^r)^2}\|x^{r+1}-x^r\|^2 \\
&\leq \frac{4\mathcal{F}^r}{\rho\gamma^{r-1}} + \frac{2}{\rho^2}\left(\frac{1}{\gamma^r} - \frac{1}{\gamma^{r-1}}\right)\|y^r - y^{r-1}\|^2 \\
&\quad + \frac{2}{\rho}\left(\frac{\gamma^{r-2}}{\gamma^{r-1}} - \frac{\gamma^{r-1}}{\gamma^r}\right)\|y^r\|^2 \\
&\quad - \frac{2}{\rho}\|y^{r+1}-y^r\|^2 + \frac{2L_y^2}{\rho(\gamma^r)^2}\|x^{r+1}-x^r\|^2.
\end{aligned} \tag{61}$$

Furthermore, adding (34) and (61), and ignoring the negative term $-\frac{\gamma^{r-1}}{2}\|y^{r+1}-y^r\|^2$ we have

$$\begin{aligned}
&\ell(x^{r+1}, y^{r+1}) - \frac{\gamma^r}{2}\|y^{r+1}\|^2 + \frac{4\mathcal{F}^{r+1}}{\rho\gamma^r} \\
&\leq \ell(x^r, y^r) - \frac{\gamma^{r-1}}{2}\|y^r\|^2 + \frac{4\mathcal{F}^r}{\rho\gamma^{r-1}} - \frac{1}{\rho}\|y^{r+1}-y^r\|^2 \\
&\quad - \left(\frac{\beta^r}{2} + \mu - \left(\frac{\rho L_y^2}{2} + \frac{2L_y^2}{\rho(\gamma^r)^2}\right)\right)\|x^{r+1}-x^r\|^2 \\
&\quad + \frac{\gamma^{r-1}-\gamma^r}{2}\|y^{r+1}\|^2 \\
&\quad + \frac{2}{\rho}\left(\frac{1}{4} + \frac{1}{\rho\gamma^r} - \frac{1}{\rho\gamma^{r-1}}\right)\|y^r - y^{r-1}\|^2
\end{aligned}$$

$$+\frac{2}{\rho}\left(\frac{\gamma^{r-2}}{\gamma^{r-1}}-\frac{\gamma^{r-1}}{\gamma^{r}}\right)\|y^r\|^2.$$

Finally, by adding to both sides the term $\frac{2}{\rho}(\frac{1}{\rho\gamma^{r+1}}-\frac{1}{\rho\gamma^r})\|y^{r+1}-y^r\|^2$, using the definition $\{\mathcal{P}^r\}$ in (35), we obtain

$$\mathcal{P}^{r+1}-\mathcal{P}^r \le -\frac{1}{2\rho}\|y^{r+1}-y^r\|^2+\frac{\gamma^{r-1}-\gamma^r}{2}\|y^{r+1}\|^2$$
$$-\left(\frac{\beta^r}{2}+\mu-\left(\frac{\rho L_y^2}{2}+\frac{2L_y^2}{\rho(\gamma^r)^2}\right)\right)\|x^{r+1}-x^r\|^2$$
$$+\frac{2}{\rho}\left(\frac{1}{\rho\gamma^{r+1}}-\frac{1}{\rho\gamma^r}\right)\|y^{r+1}-y^r\|^2+\frac{2}{\rho}\left(\frac{\gamma^{r-2}}{\gamma^{r-1}}-\frac{\gamma^{r-1}}{\gamma^r}\right)\|y^r\|^2.$$

According to the above, to achieve descent in $\|y^{r+1}-y^r\|^2$ we need to ensure that the following holds:

$$-1/2\rho+2/\rho^2(1/\gamma^{r+1}-1/\gamma^r)<0. \tag{62}$$

Note that, (62) is equivalent to the condition $\frac{1}{\gamma^{r+1}}-\frac{1}{\gamma^r}\le\rho/4$, which holds by condition (36). This completes the proof. **Q.E.D.**

G. Proof of Theorem 2

For simplicity, let $\mathcal{G}_i^r:=(\mathcal{G}_\rho^{\beta^r}(x^r,y^r))_i$. Similarly as in the proof of Theorem 1, we have

$$\|\mathcal{G}_i^r\|\le(\beta^r+L_{u_i}+L_{x_i})\|x^{r+1}-x^r\|,\ \forall i\in[K].$$

For the corresponding bound for y we have

$$\begin{aligned}
&\|\nabla\mathcal{G}_{K+1}^r\|\\
&\le\frac{1}{\rho}\|y^{r+1}-y^r\|+\frac{1}{\rho}\|y^{r+1}-\mathrm{Py}^{1/\rho}(y^r+\rho\nabla_y f(x^r,y^r))\|\\
&\overset{(a)}{=}\frac{1}{\rho}\|y^{r+1}-y^r\|\\
&\quad+\frac{1}{\rho}\|\mathrm{Py}^{1/\rho}(y^r+\rho\nabla_y f(x^{r+1},y^{r+1})-\rho\gamma^r y^{r+1})\\
&\quad-\mathrm{Py}^{1/\rho}(y^r+\rho\nabla_y f(x^r,y^r))\|\\
&\overset{(b)}{\le}L_y\|x^{r+1}-x^r\|+\left(\frac{1}{\rho}+L_y\right)\|y^{r+1}-y^r\|+\gamma^r\|y^{r+1}\|,
\end{aligned}$$

where in (a) we use the optimality conditions w.r.t y; in (b) we use the nonexpansiveness of the proximal operator, as well as the the Lipschitz gradient condition w.r.t y two times. Combining the above two bounds we obtain

$$\begin{aligned}
\|\nabla\mathcal{G}^r\|^2&\le\sum_{i=1}^K(\beta^r+L_{u_i}+L_{x_i})^2\|x^{r+1}-x^r\|^2\\
&+3(\gamma^r)^2\|y^{r+1}\|^2+3L_y^2\|x^{r+1}-x^r\|^2\\
&+3\left(\frac{1}{\rho}+L_y\right)^2\|y^{r+1}-y^r\|^2\\
&\le\left(K(L+\beta^r)^2+3L_y^2\right)\|x^{r+1}-x^r\|^2\\
&+3\left(\frac{1}{\rho}+L_y\right)^2\|y^{r+1}-y^r\|^2+3(\gamma^r)^2\|y^{r+1}\|^2,
\end{aligned}\tag{63}$$

where we defined $L:=\max_{i\in[K]}(L_{u_i}+L_{x_i})$. Moreover, we choose

$$\beta^r=\rho L_y^2+\frac{2\kappa L_y^2}{\rho(\gamma^r)^2}-2\mu, \tag{64}$$

where κ is chosen to satisfy $\kappa>2,\quad\beta^0>L_{x_i},\ \forall i$.

By condition (38), it is clear that $\beta^{r+1}\ge\beta^r$. Combining this with the choice of κ we have: $\beta^r\ge\beta^0>L_{x_i},\ \forall i,r$. Thus, this choice of β^r satisfies Assumption C-3.

Moreover, such a choice implies that

$$\alpha^r:=\frac{\beta^r}{2}+\mu-\left(\frac{\rho L_y^2}{2}+\frac{2L_y^2}{\rho(\gamma^r)^2}\right)=\frac{(\kappa-2)L_y^2}{\rho(\gamma^r)^2}. \tag{65}$$

Using these properties in (63), the constants in front of $\|x^{r+1}-x^r\|^2$ becomes

$$\begin{aligned}
K(L+\beta^r)^2+3L_y^2&=K\left(L+\rho L_y^2+\frac{2\kappa L_y^2}{\rho(\gamma^r)^2}-2\mu\right)^2+3L_y^2\\
&\overset{(a)}{=}\left(K^2L+\rho K^2L_y-2\mu K^2+K^2\frac{2\kappa}{\kappa-2}\alpha^r\right)^2+3L_y^2\\
&\overset{(b)}{\le}(d_1\alpha^r)^2
\end{aligned}\tag{66}$$

in (a) we use the identity shown in (65); (b) always holds for some $d_1>1$ (which are both independent of r), since α^r is an increasing sequence, and α^0 is bounded away from zero. Note that since y lies in a bounded set, there exists σ_y such that $\|y^{r+1}\|^2\le\sigma_y^2,\forall r$. Using (66), setting $z:=3(L_y+\frac{1}{\rho})^2$, we obtain

$$\begin{aligned}
\|\nabla\mathcal{G}^r\|^2&\le(d_1\alpha^r)^2\|x^{r+1}-x^r\|^2\\
&\quad+z\|y^{r+1}-y^r\|^2+3(\gamma^r)^2\sigma_y^2.
\end{aligned}\tag{67}$$

Furthermore, when $\beta^r=\rho L_y^2+\frac{2\kappa L_y^2}{\rho(\gamma^r)^2}-2\mu$ and since $\frac{1}{\gamma^{r+1}}-\frac{1}{\gamma^r}\le\frac{\rho}{5}$, the bound of the potential function (37) becomes

$$\begin{aligned}
\mathcal{P}^{r+1}&\le\mathcal{P}^r-\frac{1}{10\rho}\|y^{r+1}-y^r\|^2-\alpha^r\|x^{r+1}-x^r\|^2\\
&\quad+\frac{\gamma^{r-1}-\gamma^r}{2}\|y^{r+1}\|^2+\frac{2}{\rho}\left(\frac{\gamma^{r-2}}{\gamma^{r-1}}-\frac{\gamma^{r-1}}{\gamma^r}\right)\|y^r\|^2.
\end{aligned}$$

Because $\{\alpha^r\}$ is increasing and $\|y^r\|^2\le\sigma_y^2$, the above relation implies the following

$$\begin{aligned}
&\frac{1}{10\rho}\|y^{r+1}-y^r\|^2+\alpha^r\|x^{r+1}-x^r\|^2\\
&\le\mathcal{P}^r-\mathcal{P}^{r+1}+\frac{\gamma^{r-1}-\gamma^r}{2}\sigma_y^2+\frac{2}{\rho}\left(\frac{\gamma^{r-2}}{\gamma^{r-1}}-\frac{\gamma^{r-1}}{\gamma^r}\right)\sigma_y^2.
\end{aligned}\tag{68}$$

Let us define

$$d_2^r:=\min\left\{\frac{1}{10\rho},1\right\}/\max\left\{z,d_1^2\alpha^r\right\}.$$

Then by combining (68) and (67), we obtain

$$\begin{aligned}
&\|\nabla\mathcal{G}^r\|^2\times d_2^r\le\mathcal{P}^r-\mathcal{P}^{r+1}\\
&+\frac{\gamma^{r-1}-\gamma^r}{2}\sigma_y^2+\frac{2}{\rho}\left(\frac{\gamma^{r-2}}{\gamma^{r-1}}-\frac{\gamma^{r-1}}{\gamma^r}\right)\sigma_y^2+3(\gamma^r)^2\sigma_y^2\times d_2^r.
\end{aligned}$$

Summing both sides from $r=1$ to T, and noting that condition (38) implies $\frac{\gamma^r}{\gamma^{r+1}}\le1.2,\ \forall r$, we obtain

$$\sum_{r=1}^{T} d_2^r \|\nabla \mathcal{G}^r\|^2 \le \sum_{r=1}^{T} d_2^r \frac{3(\kappa-2)L_y^2\sigma_y^2}{\rho\alpha^r} +$$

$$+ \mathcal{P}^1 - \underline{\mathcal{P}} + \sigma_y^2\left(\frac{\gamma^0-\gamma^T}{2} + \frac{2}{\rho}\left(\frac{\gamma^{-1}}{\gamma^0} - \frac{\gamma^{T-1}}{\gamma^T}\right)\right),$$

where we have defined

$$d_3 := \mathcal{P}^1 - \underline{\mathcal{P}} + \sigma_y^2\left(\frac{\gamma^0-\gamma^T}{2} + \frac{2}{\rho}\left(\frac{\gamma^{-1}}{\gamma^0} - \frac{\gamma^{T-1}}{\gamma^T}\right)\right);$$

$\underline{\mathcal{P}}$ is a lower bound of $\mathcal{P}$, which is a finite number due to the lower boundness assumption of $\underline{\ell}$ and the compactness of $\mathcal{Y}$ (see Assumption A.1). Notice that since $d_1 > 1$, we have

$$d_2^r \le \frac{d_4}{d_1^2 a^r} \le \frac{d_4}{a^r},$$

where $d_4 := \min\{\frac{1}{10\rho}, 1\}$. Also, there exists $d_5 > \max\{\frac{d_1^2}{d_4}, \frac{z}{d_4 a^0}\}$ such that $d_2^r \ge \frac{1}{d_5\alpha^r}$.

By utilizing the definition of $T(\epsilon)$ and the above bounds, we know that

$$\epsilon^2 \le \frac{d_3 d_5 + \frac{3d_4 d_5(\kappa-2)L_y^2\sigma_y^2}{\rho}\sum_{r=1}^{T(\epsilon)}\frac{1}{(\alpha^r)^2}}{\sum_{r=1}^{T(\epsilon)}\frac{1}{\alpha^r}}. \tag{69}$$

Moreover, when $\gamma^r = \frac{1}{\rho r^{1/4}}$, it can be verified that the following holds:

$$\frac{1}{\gamma^{r+1}} - \frac{1}{\gamma^r} \le 0.19\rho, \quad \forall r \ge 1,$$

because $(r+1)^{1/4} - (r)^{1/4}$ is a monotonically decreasing function and its maximum value is achieved at $r = 1$. We can plug in this choice of γ^r into (65), and obtain

$$\alpha^r = (\kappa-2)\rho L_y^2\sqrt{r}.$$

Using these choices of $\{\gamma^r, \alpha^r\}$, and by utilizing the bounds that $\sum_{r=1}^{T} 1/r \le c\ln(T)$ (for some $c > 0$), and $\sum_{r=1}^{T} 1/\sqrt{r} \ge \sqrt{T}$, the relation (69) becomes:

$$\epsilon^2 \le \frac{C\log(T(\epsilon))}{\sqrt{T(\epsilon)}}, \tag{70}$$

where $C > 0$ is some constant independent of the iteration. Then, the desired result follows directly from (70). **Q.E.D.**

ACKNOWLEDGMENT

The authors would like to thank Dr. Tsung-Hui Chang and Dr. Wei-Chiang Li for helpful discussion on the MISO beamforming problem, and for providing their codes.

REFERENCES

[1] S. Lu, R. Singh, X. Chen, Y. Chen, and M. Hong, "Alternating gradient descent ascent for nonconvex min-max problems in robust learning and GANs," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, 2019, pp. 680–684.

[2] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5262–5276, Oct. 2010.

[3] W. Liao, M. Hong, H. Farmanbar, and Z.-Q. Luo, "Semi-asynchronous routing for large scale hierarchical networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 2894–2898.

[4] D. Hajinezhad and M. Hong, "Perturbed proximal primal dual algorithm for nonconvex nonsmooth optimization," *Math. Program.*, vol. 176, pp. 207–245, 2019.

[5] G. B. Giannakis, Q. Ling, G. Mateos, I. D. Schizas, and H. Zhu, "Decentralized learning for wireless communications and networking," in, *Splitting Methods in Communication and Imaging*. New York, NY, USA: Springer, 2015.

[6] A. Koppel, B. M. Sadler, and A. Ribeiro, "Proximity without consensus in online multiagent optimization," *IEEE Trans. Signal Process.*, vol. 65, no. 12, pp. 3062–3077, Jun. 2017.

[7] P. Di Lorenzo and G. Scutari, "Next: In-network nonconvex optimization," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 2, pp. 120–136, Jun. 2016.

[8] Y. Tian, Y. Sun, B. Du, and G. Scutari, "ASY-SONATA: Achieving geometric convergence for distributed asynchronous optimization," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput.*, 2018, pp. 543–551.

[9] Y. Sun, A. Daneshmand, and G. Scutari, "Convergence rate of distributed optimization algorithms based on gradient tracking," 2019, *arXiv:1905.02637*.

[10] Z. Jiang, A. Balu, C. Hegde, and S. Sarkar, "Collaborative deep learning in fixed topology networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5904–5914.

[11] X. Lian, C. Zhang, H. Zhang, C. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Proc. Adv. Neural Inf. Process. Syst*, 2017, pp. 5336–5346.

[12] S. Lu, X. Zhang, H. Sun, and M. Hong, "GNSD: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization," in *Proc. IEEE Data Sci. Workshop*, Jun. 2019, pp. 315–321.

[13] Q. Qian, S. Zhu, J. Tang, R. Jin, B. Sun, and H. Li, "Robust optimization over multiple domains," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 4739–4746.

[14] W. Li, T. Chang, and C. Chi, "Multicell coordinated beamforming with rate outage constraint—Part II: Efficient approximation algorithms," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2763–2778, Jun. 2015.

[15] G. J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence," *IEEE Trans. Veh. Technol.*, vol. 42, no. 4, pp. 641–646, Nov. 1993.

[16] J. Zander, "Performance of optimal transmitter power control in cellular system," *IEEE Trans. Veh. Technol.*, vol. 41, no. 1, pp. 57–63, Feb. 1992.

[17] J. Zander, "Distributed cochannel interference control in cellular radio systems," *IEEE Trans. Veh. Technol.*, vol. 41, no. 3, pp. 305–311, Aug. 1992.

[18] S. Lu and Z. Wang, "Spatial transmitter density allocation for frequency-selective wireless ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 473–486, Jan. 2019.

[19] M. Bengtsson and B. Ottersten, "Optimal and suboptimal transmit beamforming,", in *Handbook of Antennas in Wireless Communications*. Boca Raton, FL, USA: CRC Press, 2001.

[20] A. Wiesel, Y. C. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, Jan. 2006.

[21] Y.-F Liu, Y.-H. Dai, and Z.-Q. Luo, "Max-min fairness linear transceiver design for a multi-user MIMO interference channel," *IEEE Trans. Signal Process.*, vol. 61, no. 9, pp. 2413–2423, May 2013.

[22] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "Linear transceiver design for a MIMO interfering broadcast channel achieving max-min fairness," *Signal Process.*, vol. 93, no. 12, pp. 3327–3340, 2013.

[23] R. Sun, M. Hong, and Z.-Q. Luo, "Joint downlink base station association and power control for max-min fairness: Computation and complexity," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1040–1054, Jun. 2015.

[24] S. Lu and Z. Wang, "Training optimization and performance of single cell uplink system with massive-antennas base station," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1570–1585, Feb. 2019.

[25] W. Li, T.-H. Chang, and C. Chi, "Multicell coordinated beamforming with rate outage constraint—Part I: Complexity analysis," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2749–2762, Jun. 2015.

[26] R. H. Gohary, Y. Huang, Z.-Q. Luo, and J.-S. Pang, "A generalized iterative water-filling algorithm for distributed power control in the presence of a jammer," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2660–2674, Jul. 2009.

[27] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *J. Optim. Theory Appl.*, vol. 142, no. 1, pp. 205–228, Jul. 2009.

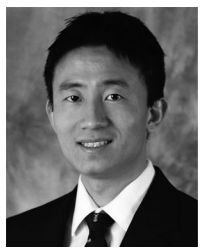
[28] K. T. L. Hien, R. Zhao, and W. B. Haskell, "An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems," 2018, *arXiv:1711.03669*.

[29] Y. Chen, G. Lan, and Y. Ouyang, "Optimal primal-dual methods for a class of saddle point problems," *SIAM J. Optim.*, vol. 24, no. 4, pp. 1779–1814, 2014.

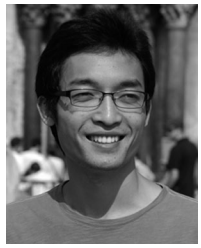
[30] C. Daskalakis and I. Panageas, "The limit points of (optimistic) gradient descent in min-max optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9256–9266.
[31] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, "Training GANs with optimism," in *Proc. Int. Con. Learn. Representations*, 2018.
[32] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.
[33] P. Mertikopoulos, H. Zenati, B. Lecouat, C.-S. Foo, V. Chandrasekhar, and G. Piliouras, "Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–23.
[34] H. Rafique, M. Liu, Q. Lin, and T. Yang, "Non-convex min-max optimization: Provable algorithms and applications in machine learning," 2018, *arXiv:1810.02060*.
[35] M. Sanjabi, B. Jimmy, M. Razaviyayn, and J. D. Lee, "On the convergence and robustness of training GANs with regularized optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7088–7098.
[36] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn, "Solving a class of non-convex min-max games using iterative first order methods," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14905–14916.
[37] S. Lu, R. Singh, X. Chen, Y. Chen, and M. Hong, "Alternating gradient descent ascent for nonconvex min-max problems in robust learning and GANs," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, 2019, pp. 680–684.
[38] T. Lin, C. Jin, and M. I. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," 2019, *arXiv:1906.00331*, 2019.
[39] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh, "Efficient algorithms for smooth minimax optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 12659–12670.
[40] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen, "Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications, supplementary material," *IEEE Trans. Signal Process.*, to be published.
[41] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, Jan. 2016.
[42] M. Razaviyayn, "Successive convex approximation: Analysis and applications," Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Minnesota, Minneapolis, MN, USA, 2014.
[43] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.
[44] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 641–656, Feb. 2014.
[45] F. Facchinei, S. Sagratella, and G. Scutari, "Flexible parallel algorithms for big data optimization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 7208–7212.
[46] F. Facchinei, G. Scutari, and S. Sagratella, "Parallel selective algorithms for nonconvex big data optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1874–1889, Apr. 2015.
[47] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Comput.*, vol. 15, no. 4, pp. 915–936, Apr. 2003.
[48] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Proc. 29th Annu. Conf. Learn. Theory*, 2016, vol. 49, pp. 1246–1257.
[49] Y. Zhang and Q. Yang, "An overview of multi-task learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 30–43, 2018.
[50] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[51] D. A. Schmidt, C. Shi, R. A. Berry, M. L. Honig, and W. Utschick, "Comparison of distributed beamforming algorithms for MIMO interference networks," *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3476–3489, Jul. 2013.
[52] C. Shi, R. A. Berry, and M. L. Honig, "Monotonic convergence of distributed interference pricing in wireless networks," in *Proc. IEEE Int. Conf. Symp. Inf. Theory*, 2009, pp. 1619–1623.
[53] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

Songtao Lu (Member, IEEE) received the Ph.D. degree in electrical engineering from Iowa State University in 2018. Currently, he is an AI resident at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY. He was a Postdoctoral Associate with the Department of Electrical and Computer Engineering, University of Minnesota Twin Cities, Minneapolis. He was a recipient of the Graduate and Professional Student Senate Research Award from Iowa State University in 2015, the Research Excellence Award from the Graduate College of Iowa State University in 2017, and the Student Travel Awards from the 20th International Conference on Artificial Intelligence and Statistics and the thirty-sixth International Conference on Machine Learning. His primary research interests include signal processing, optimization, artificial intelligence, and machine learning.

Ioannis Tsaknakis received the B.Sc. degree in electrical and computer Engineering in 2015 and the M.Sc. degree in applied mathematics in 2017 from the National Technical University of Athens. He is currently a Ph.D Student in the Department of Electrical and Computer Engineering at the University of Minnesota, Twin Cities.



Mingyi Hong received the Ph.D. degree from the University of Virginia, Charlottesville, in 2011. He is an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Minnesota, Minneapolis. He serves on the IEEE Signal Processing for Communications and Networking and Machine Learning for Signal Processing Technical Committees. His research interests include optimization theory and applications in signal processing and machine learning. He is a Member of the IEEE.



Yongxin Chen (Member, IEEE) received the B.Sc. degree from Shanghai Jiao Tong University in 2011 and Ph.D. from the University of Minnesota in 2016, both in mechanical engineering. He is currently an Assistant Professor in the School of Aerospace Engineering at Georgia Institute of Technology. He has served on the faculty at Iowa State University (2017-2018). He received the George S. Axelby Best Paper Award in 2017 for his joint work with Tryphon Georgiou and Michele Pavon. His current research focuses on the intersection of control theory, machine learning, robotics and optimization.