BLOCK ALTERNATING OPTIMIZATION FOR NON-CONVEX MIN-MAX PROBLEMS: ALGORITHMS AND APPLICATIONS IN SIGNAL PROCESSING AND COMMUNICATIONS

Songtao Lu, Ioannis Tsaknakis, and Mingyi Hong

[†]Department of Electrical and Computer Engineering, University of Minnesota Twin Cities, Minneapolis, MN, 55455, USA

ABSTRACT

The min-max problem, also known as the saddle point problem, can be used to formulate a wide range of applications in signal processing and wireless communications. However, existing optimization theory and methods, which mostly deal with problems with certain convex-concave structure, are not applicable for the aforementioned applications, which oftentimes involve non-convexity. In this work, we consider a general 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 two simple algorithms, which alternatingly perform one gradient descent-type step for each minimization block and one gradient ascent-type step for the maximization problem. For the first time, we show that such simple alternating min-max algorithms converge to first-order stationary solutions. We conduct numerical tests on a robust learning problem, and a wireless communication problem in the presence of jammers, to validate the efficiency of the proposed algorithms.

1. INTRODUCTION

Consider the following min-max (a.k.a. saddle point) problem:

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

where f is a continuously differentiable function; h_i and g are some convex, not necessarily smooth functions; $x := [x_1;\cdots;x_K]$ and y are the block optimization variables; $\mathcal{X}_i$'s and $\mathcal{Y}$ are some convex feasible sets. We call the problem *one-sided* non-convex problem because we assume that $f(x,y)$ is non-convex w.r.t. x, and (strongly) concave w.r.t. y. The challenge is to design effective algorithms that can deal with the non-convexity in the problem.

Problem (1) is quite general. It arises in many signal processing, communication and networking applications, as listed below.

1.1. Motivating Examples

Distributed non-convex optimization: Consider a network of K agents defined by a connected graph $\mathcal{G}\triangleq\{\mathcal{V},\mathcal{E}\}$, where each agent i can communicate with its neighbors. The agents solve the following problem, $\min_{y\in\mathbb{R}^d}\sum_{i=1}^{K}(f_i(y)+h_i(y))$, where each $f_i(y):\mathbb{R}^d\to\mathbb{R}$ is a non-convex, smooth function, and $h_i(y):\mathbb{R}^d\to\mathbb{R}$ is a convex, not necessarily smooth function that plays the role of the regularizer. Each agent i has access only to f_i, h_i. Denote x_i as agent i's local copy of y, then we have the following equivalent formulation

$$\min_{x\in\mathbb{R}^{Kd}} f(x)+h(x)=\sum_{i=1}^{K}(f_i(x_i)+h_i(x_i))\quad \text{s.t. } (A\otimes I_d)x=0,$$

where $x=[x_1;\ldots;x_K]\in\mathbb{R}^{Kd}$ and $A\in\mathbb{R}^{|\mathcal{E}|\times K}$ is the incidence matrix, i.e., assuming that the edge e is incident on vertices i and j, with $i>j$ we have that $A_{ei}\triangleq 1, A_{ej}\triangleq -1$ and $A_{e\cdot}=0$ for all other vertices; $\otimes$ denotes the Kronecker product. This task captures the formulation of many problems that appear in distributed machine learning and signal processing (e.g., [1–4]).

Using duality theory we can rewrite the above problem as

$$\min_{x\in\mathbb{R}^{Kd}}\max_{\lambda\in\mathbb{R}^{|\mathcal{E}|d}} f(x)+h(x)+\langle\lambda,(A\otimes I_d)x\rangle\tag{2}$$

where λ are the multipliers. Clearly (2) is in the form of (1).

Robust learning over multiple domains: In [5] the authors introduce a robust optimization framework, in which training sets from K different domains are used to train a machine learning model. Let $\mathcal{S}_k\triangleq\{(x_i^k,y_i^k)\}, 1\le k\le K$ be the individual training sets with $x_i^k\in\mathbb{R}^n$, $y_i^k\in\mathbb{R}$; w be the parameter of the model we intent to learn, $l(\cdot)$ a non-negative loss function and $f_k(w)=\frac{1}{|\mathcal{S}_k|}\sum_{i=1}^{|\mathcal{S}_k|}l(x_i^k,y_i^k,w)$ the non-convex (in general) empirical risk in the k-th domain. The following problem formulates the task of finding the parameter w that minimizes the empirical risk, while taking into account the worst possible distribution over the K different domains:

$$\min_{w}\max_{p\in\Delta}p^TF(w),\tag{3}$$

where $F(w):=[f_1(w),\ldots,f_K(w)]^T$, p describes the adversarial distribution over the different domains and $\Delta:=\{p\in\mathbb{R}^K\mid 0\le p_i\le 1,\ i=1,\ldots,K,\ \sum_{i=1}^{K}p_i=1\}$ is the standard simplex. Furthermore, it is also common to add a regularizer that imposes structures on the adversarial distribution ($\lambda>0$ is some parameter):

$$\min_{w}\max_{p\in\Delta}p^TF(w)-\frac{\lambda}{2}D(p||q)\tag{4}$$

where D is some distance between probability distributions and q is some prior probability distribution. The objective function consists of a linear coupling between p and the non-convex function $F(w)$, and a convex or strongly convex regularizer.

Power control problem: Consider a problem in wireless transceiver design, where K transmitter-receiver pairs transmit over N channels aiming to maximize their minimum rates. User k transmits with power $p_k:=[p_k^1;\cdots;p_k^N]$, and its rate is given by: $R_k(p_1,\ldots,p_K)=\sum_{n=1}^{N}\log\left(1+\dfrac{a_{kk}^np_k^n}{\sigma^2+\sum_{\ell=1,\ell\neq k}^{K}a_{\ell k}^np_\ell^n}\right)$ (assuming Gaussian signaling), which is a non-convex function on $p:=[p_1;\cdots;p_K]$. Here $a_{\ell k}^n$'s denote the channel gain between pair (ℓ,k) on nth channel, and σ^2 is the noise power. The classical max-min fairness power control problem is: $\max_{p\in P}\min_{k}R_k(p)$, where $\mathcal{P}:=\{p\mid 0\le p_k\le\bar{p},\forall\ k\}$ denotes the feasible power allocations. The above max-min rate problem can be equivalently formulated as: (1):

$$\min_{p\in\mathcal{P}} \max_{y\in\Delta} \quad \sum_{k=1}^{N} -R_k(p_1,\cdots,p_K)\times y_k, \tag{5}$$

where the set Δ is again the standard simplex. Note that $R_k(p)$ is a non-convex function in p, and there is a linear coupling between $\mathbf{y}$ and the set of functions $\{R_k(p)\}_{k=1}^{K}$ in the objective function.

Power control in the presence of a jammer: Consider an extension of the above scenario (which is first described in [6]), where a jammer participates in a K-user N-channel interference channel transmission. Differently from a regular user, the jammer's objective is to reduce the total sum-rate of the other users by properly transmitting noises. Because there are N channels, we use p_k^n to denote k user's transmission on nth channel. The corresponding sum-rate maximization-minimization problem can be formulated as:

$$\min_{p\in\mathcal{P}} \max_{p_0\in\mathcal{P}_0} \sum_{(k,n)=(1,1)}^{K,N} -\log\left(1+\frac{a_{kk}^n p_k^n}{\sigma^2+\sum_{j=1,j\neq k}^{K} a_{jk}^n p_j^n + a_{0k}^n p_0^n}\right), \tag{6}$$

where p_k and p_0 are the power allocation of user k and the jammer, respectively; the set $\mathcal{P} := \mathcal{P}_1 \times \ldots \times \mathcal{P}_K$, where $\mathcal{P}_k$ are defined similarly as before.

1.2. Related Work

Motivated by these applications, it is of interest to develop efficient algorithms for solving these problems with theoretical convergence guarantees. There has been a long history of studying the min-max optimization problems. When the problem is convex-concave, previous works [7–10] and the references therein have shown that certain primal-dual type algorithms, which alternate between the update of x and y variables, can solve the convex-concave saddle problem optimally. However, when the problem is non-convex, the convergence behavior of the primal-dual algorithm has not been well understood. One challenge is that the primal and dual variables are coupled in a general nonlinear way, therefore it is difficult to to characterize the progress of the iterates as the algorithm proceeds.

Although there are lots of recent works on the non-convex minimization problems [11], only few works focus on the non-convex min-max problems. For example, a robust optimization problem from multiple distributions is proposed recently in [5], where the coupling between the iterates in the minimization and maximization problems is linear and the variable of the minimization problem is unconstrained. A simple min-max algorithm in the stochastic settings is also considered, which updates the iterates of the both problems independently and converges to the stationary point of the min-max problems; please see [5, Theorem 1] for details about the convergence rate. In [12], a proximally guided stochastic mirror descent method (PG-SMD) is proposed, which also optimizes the minimization and maximization components separately and updates the corresponding iterates at the same time. From the convergence analysis, it turns out that PG-SMD converges provably to an approximate stationary point of the minimization problem. Finally, an oracle based non-convex stochastic gradient descent for generative adversarial networks was proposed in [13], where the algorithm assumes that the maximization subproblem can be solved up to some small error. None of the above works are of the alternating type we consider in this work, and none of them are for deterministic problems.

1.3. Contribution of this work

In this work, we consider the min-max problem from an alternating minimization/maximization perspective. The studied problems are general, allowing non-convexity and non-smoothness in the objective, non-convex coupling between variables as well. The proposed algorithm solves the maximization and minimization problems in an alternating way. The main contributions of this paper are listed as follows:

1) A class of machine learning and resource allocations problems is formulated in the framework of non-convex min-max problems.

2) Two types of min-max problems are studied, and two simple algorithms which alternate between the minimization and maximization steps, are presented with provable convergence guarantees.

3) To the best of our knowledge, it is the first time that the convergence rate of the alternating min-max algorithm is quantified for the (one-sided) non-convex min-max problem (1).

2. THE PROPOSED ALGORITHMS

In this section, we formulate the problems into two cases according to the structure of the coupling between variable x and y.

2.1. Maximization problem is strongly concave

First, we consider the following min-max problem,

$$\min_{x_i\in\mathcal{X}_i,\forall i} \max_{y\in\mathcal{Y}} \; f(x_1,\ldots,x_K,y) \tag{7}$$

where $x_i \in \mathcal{X}_i$ with $\mathcal{X}_i, \mathcal{Y}$ convex compact feasible sets and $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_K$. Moreover, x and y are coupled in a general way by function $f(x,y)$, the objective function $f(x,y)$ is non-convex with respect to x and it has Lipschitz continuous gradient w.r.t to x_i with constants $L_{x_i}, \forall i$. Finally, $f(x,y)$ is strongly concave with respect to y with modulus $\theta > 0$, that is

$$f(x,y) - f(x,z) \le \langle \nabla f(x,z), y-z\rangle - \frac{\theta}{2}\|y-z\|^2, \; \forall\, y\in\mathcal{Y}.$$

For the above problem, we propose the following algorithm:

$$x_i^{r+1} = \arg\min_{x_i\in\mathcal{X}_i} \langle \nabla_{x_i} f(w_i^r, y^r), x_i - x_i^r\rangle + \frac{\beta}{2}\|x_i - x_i^r\|^2, \quad i = 1,\cdots,K, \tag{8a}$$

$$y^{r+1} = \arg\max_{y\in\mathcal{Y}} \langle \nabla_y f(x^{r+1}, y^r), y - y^r\rangle - \frac{1}{2\rho}\|y - y^r\|^2 \tag{8b}$$

where r is the iterate's index, $w_i^r := [x_1^{r+1};\cdots x_{i-1}^{r+1}; x_i^r;\cdots]$, β and ρ are the regularization parameters and will be discussed further in Lemma 2.

2.2. Maximization problem is linear

Second, we consider the following min-max problem,

$$\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} \quad \langle f(x), y\rangle \tag{9}$$

where $\mathcal{X}, \mathcal{Y}$ are still convex compact sets, the objective function $f(x)$ is non-convex, and it is Lipschitz and has Lipschitz gradient, with constants L_x and L'_x. Because the maximization problem over variable y is linear rather than strongly concave, we propose a novel algorithm which adds some (diminishing) perturbation term in the maximization problem and uses adaptive regularization parameters (i.e., β^r and γ^r shown in the following section):

$$x^{r+1} = \arg\min_{x\in\mathcal{X}} \langle \nabla_x \langle f(x^r), y^r\rangle, x - x^r\rangle + \frac{\beta^r}{2}\|x - x^r\|^2, \tag{10a}$$

$$y^{r+1} = \arg\max_{y\in\mathcal{Y}} \langle f(x^{r+1}), y - y^r\rangle - \frac{1}{2\rho}\|y - y^r\|^2 - \frac{\gamma^r}{2}\|y\|^2, \tag{10b}$$

where β^r is the regularizer of the minimization problem, while ρ and γ^r are the regularizers of the maximization problem, with γ^r being a decreasing sequence.

3. THEORETICAL PROPERTIES AND DISCUSSIONS

In this section, we present our main convergence results for the proposed algorithms. Due to space limitations, we will omit the proof details and instead will be presenting a few key lemmas leading to the main results. The detailed proofs are provided in [14].

Lemma 1 *(Descent lemma) Assume that $f(x,y)$ has Lipschitz continuous gradient w.r.t y with constant L_y and is strongly concave with respect to y with modulus $\theta > 0$. Let (x^r, y^r) be a sequence generated by* (8). *Then we have the following descent estimate*

$$f(x^{r+1}, y^{r+1}) - f(x^{r+1}, y^r) \leq \frac{1}{\rho}\|y^{r+1} - y^r\|^2$$
$$- \big(\theta - (\frac{1}{2\rho} + \frac{\rho L_y^2}{2})\big)\|y^r - y^{r-1}\|^2 + \frac{\rho L_y^2}{2}\|x^{r+1} - x^r\|^2.$$

From Lemma 1, it is not clear whether the objective function is decreasing or not, since the minimization step will consistently decrease the objective value while the maximization step will increase the objective value. The key in our analysis is to identify a proper potential function, which can capture the essential dynamics of the algorithm.

Lemma 2 *Let (x^r, y^r) be a sequence generated by* (8). *When the following conditions are satisfied,*

$$\rho < \frac{\theta}{4L_y^2}, \quad \beta \geq \max\left\{L_y^2\left(\frac{4}{\theta^2\rho} + \rho\right), L_{x_{\max}}\right\}, \tag{11}$$

then there exist $c_1, c_2 > 0$ such that the potential function will monotonically decrease, i.e.,

$$P^{r+1} - P^r \leq -c_1\|y^{r+1} - y^r\|^2 - c_2\|x^{r+1} - x^r\|^2, \tag{12}$$

where $P^{r+1} := f(x^{r+1}, y^{r+1}) + \big(\frac{2}{\rho^2\theta} + \frac{1}{2\rho} - 4(\frac{1}{\rho} - \frac{L_y^2}{2\theta^2})\big)\|y^{r+1} - y^r\|^2$ and $L_{x_{\max}} = \max_{i=1,\ldots,K} L_{x_i}$.

3.1. Convergence rate of algorithm (8)

To state our main result, let us define the proximal gradient of the objective function as

$$\nabla\mathcal{L}(x,y) \triangleq \left[\begin{array}{c} x - P_{\mathcal{X}}[x - \nabla_x f(x,y)] \\ y - P_{\mathcal{Y}}[y + \nabla_y f(x,y)] \end{array}\right] \tag{13}$$

where $P_{\mathcal{X}}$ denotes the projection operator on convex set $\mathcal{X}$. Clearly, when $\nabla\mathcal{L}(x,y) = 0$, then a first-order stationary solution of the problem (1) is obtained. We have the following convergence rate for the proposed algorithm.

Theorem 1 *Suppose that the sequence (x^r, y^r) is generated by* (8) *and ρ, β satisfy the conditions* (11). *For a given small constant ϵ, let $T(\epsilon)$ denote the first iteration index, such that the following inequality is satisfied: $T(\epsilon) \triangleq \min\{r \mid \|\nabla\mathcal{L}(x^r, y^r)\|^2 \leq \epsilon, r \geq 1\}$. Then there exists some constant $C > 0$ such that $\epsilon \leq C(P^1 - \underline{P})/T(\epsilon)$ where $\underline{P}$ denotes the lower bound of P^r.*

3.2. Convergence rate of perturbed algorithm (10)

We have the following convergence analysis for the algorithm (10).

Lemma 3 *(Descent lemma) Assume that $f(x,y)$ is Lipschitz continuous with respect to x, with constant L_x. Assume that it is also linear with respect to y. Let (x^r, y^r) be a sequence generated by* (10). *Then we have*

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

Lemma 4 *Suppose (x^r, y^r) is generated by* (10) *and we choose $\beta^r \geq \rho L_x^2 + 6\alpha^r$, where $\alpha^r = \frac{L_x^2}{\rho(\gamma^r)^2}$. Then if the following conditions are satisfied,*

$$\frac{1}{\gamma^{r+1}} - \frac{1}{\gamma^r} \leq \frac{\rho}{5}, \ \sum_{r=1}^{\infty}\frac{1}{\alpha^r} = \infty, \alpha^r - \alpha^{r+1} \leq 0, \ \frac{1}{\alpha^r} \to 0, \tag{14}$$

then there exits a sequence $\{c^r\}$ bounded away from zero, such that the potential function will monotonically decrease, i.e.,

$$\widetilde{P}^{r+1} - \widetilde{P}^r \leq -\frac{1}{10\rho}\|y^{r+1} - y^r\|^2 - c^r\|x^{r+1} - x^r\|^2$$
$$+ \frac{1}{2}(\gamma^{r-1} - \gamma^r)\|y^{r+1}\|^2 + \frac{2}{\rho}(\frac{\gamma^{r-2}}{\gamma^{r-1}} - \frac{\gamma^{r-1}}{\gamma^r})\|y^r\|^2,$$

where $\widetilde{P}^{r+1} := f(x^{r+1}, y^{r+1}) - \left(\frac{\gamma^r}{2} + \frac{2}{\rho}(\frac{\gamma^{r-1}}{\gamma^r} - 1)\right)\|y^{r+1}\|^2 + \left(\frac{2}{\rho^2\gamma^r} + \frac{1}{2\rho} + \frac{2}{\rho^2}(\frac{1}{\gamma^{r+1}} - \frac{1}{\gamma^r})\right)\|y^{r+1} - y^r\|^2$.

Then, let us define the proximal gradient of the objective function (9) as the following to quantify the optimality gap of the algorithm.

$$\nabla\widetilde{\mathcal{L}}(x,y) \triangleq \left[\begin{array}{c} x - P_{\mathcal{X}}[x - \nabla_x\langle f(x), y\rangle] \\ y - P_{\mathcal{Y}}[y + f(x)] \end{array}\right]. \tag{15}$$

After applying the telescope sum and boundedness assumption on y, we have the following convergence analysis result.

Theorem 2 *Suppose that the sequence (x^r, y^r) is generated by* (10) *and γ^r, β^r satisfy the conditions* (14). *For a given small constant ϵ, let $T(\epsilon)$ denote the first iteration index, such that the following inequality is satisfied: $T(\epsilon) \triangleq \min\{r \mid \|\nabla\widetilde{\mathcal{L}}(x^r, y^r)\|^2 \leq \epsilon, r \geq 1\}$. Then there exists some constant $\widetilde{C} > 0$ such that $\epsilon \leq \widetilde{C}\log(T(\epsilon))/\sqrt{T(\epsilon)}$.*

Remark 1. Note that the two min-max problems have different structures, e.g., the coupling between x and y and convexity assumptions on the maximization problem over y, resulting in the different algorithms and corresponding convergence rates.

4. NUMERICAL RESULTS

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

Robust learning over multiple domains. Consider a learning scenario where we have datasets from two different domains and adopt a regularized (non-convex) logistic regression model in order to solve a binary classification problem. 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 algorithm (8).

2) Mutltitask Learning : Apply a multitask learning model, where the weights associated with each loss function/task are fixed to 1/2. The problem is optimized using gradient descent.

We evaluate the accuracy of the above algorithms as the worst hit rate across the two domains, i.e.,

accuracy = $\min\{$ hit rate on domain 1, hit rate on domain 2 $\}$.

We run two sets of experiments. In the first we generate randomly two datasets using the Gaussian and the Laplacian distribution respectively. The first dataset contains 200 training and 40 testing data points, while in the second one the number is 1000 and 200 respectively. At each dataset the data points are generated in a way that the two classes are approximately linearly separable. In Fig. 1 the accuracy of the two algorithms is illustrated.

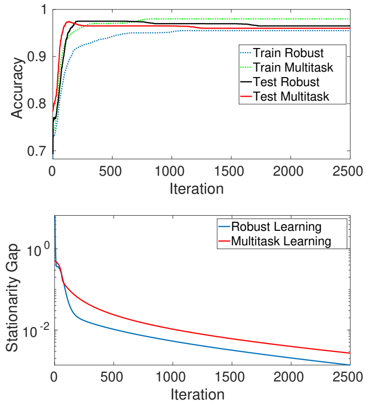

Fig. 1. The results on the experiments performed on synthetic data. The 1st figure depicts the accuracy of the two algorithms for both training and testing sets, while in the 2nd one the convergence behavior is shown.

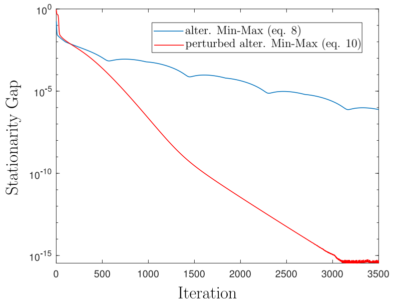

Fig. 2. Convergence comparison with alter-min algorithm (8) and perturbed algorithm (10) for the min-max problem where the maximization problem is concave rather than strongly concave.

In the second set of experiments we use two different parts of the MNIST dataset [15] as the two different domains. For the 1st domain we use 200 images for training and 50 for testing, and in the second 800 and 200 images respectively Also, we contaminate the second dataset with Gaussian noise, in order to differentiate it from the first one. We train the two aforementioned models on those datasets to classify correctly the digit '1' from the rest of them. The results are presented in Fig. 3. It is apparent that the performance of the robust model is comparable to that of the multitask learning model, which shows that the proposed algorithm is capable of attaining good solutions of min-max optimization problems.

Power control in the presence of a jammer. We consider the multi-channel and multi-user formulation (6) where there are N parallel channels available, and there are K normal users and one jammer in the system. For this problem it is easy to verify that the jammer problem has a strongly concave objective function, therefore we can directly apply Algorithm (8).

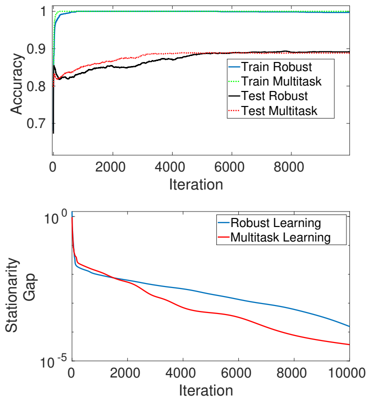

Fig. 3. The results on the experiments performed on the MNIST dataset [15]. The top figure depicts training and testing accuracies, while the bottom figure depicts the convergence behaviors of the two algorithms.

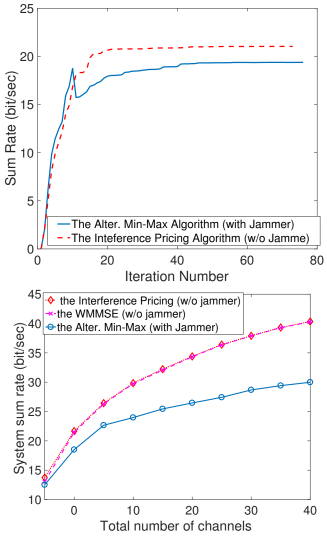

Fig. 4. The convergence curves and total averaged system performance comparing three algorithms: WMMSE, Interference Pricing and the proposed algorithm with Jammer. The first figure shows a single realization of the algorithms, and in the second figure, each point represents an average of 50 realizations. The total number of users is 10, and SNR $= 1$. The rest of the parameters are described in the main text.

In our test we will compare the proposed method with the well-known interference pricing method [16, 17], and the WMMSE algorithm [18], both of which are designed to solve the K-user N-channel sum-rate optimization problem *without* the jammer. Our problem is tested using the following setting. Consider a network with $K = 10$, and the interference channel among the users and the jammer is generated using uncorrelated fading channel model with channel coefficients generated from the complex $\mathcal{CN}(0,1)$ [18]. The users' power budget is assumed to be P for all transmitters, where $P = 10^{\text{SNR}/10}$. 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 all k, and let the jammer has the rest of the noise power, i.e., $p_{0,\max} = N/2$. Note that the consideration behind splitting the noise power is mostly for fairness comparison between the cases with and without the jammer. However, we do note that 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 also has to go through the channel, therefore the total received noise power could still be different.

From the Fig. 4 (top), we see that the pricing algorithm monotonically increases the sum rate (as is predicted by theory), while our proposed algorithm behaves very differently. Further in Fig. 4 (bottom), we do see that by using the proposed algorithm, the jammer is able to effectively reduce the total sum rate of the system.

5. REFERENCES

[1] G. Mateos, J. A. Bazerque, and G. B. Giannakis, “Distributed sparse linear regression,” *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5262–5276, 2010.

[2] W. Liao, M. Hong, H. Farmanbar, and Z.-Q. Luo, “Semi-asynchronous routing for large scale hierarchical networks,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2894–2898.

[3] D. Hajinezhad and M. Hong, “Perturbed proximal primal dual algorithm for nonconvex nonsmooth optimization,” 2017, Submitted for publication.

[4] 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*. Springer New York, 2015.

[5] Q. Qian, S. Zhu, J. Tang, R. Jin, B. Sun, and H. Li, “Robust optimization over multiple domains,” *arXiv preprint arXiv:1805.07588*, 2018.

[6] 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 Transactions on Signal Processing*, vol. 57, no. 7, pp. 2660–2674, 2009.

[7] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[8] A. Nedić and A. Ozdaglar, “Subgradient methods for saddle-point problems,” *Journal of Optimization Theory and Applications*, vol. 142, no. 1, pp. 205–228, jul 2009.

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

[10] Y. Chen, G. Lan, and Y. Ouyang, “Optimal primal-dual methods for a class of saddle point problems,” *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 1779–1814, 2014.

[11] D. Bertsekas, *Nonlinear Programming, 2nd ed*, Athena Scientific, Belmont, MA, 1999.

[12] H. Rafique, M. Liu, Q. Lin, and T. Yang, “Non-convex min-max optimization: Provable algorithms and applications in machine learning,” *arXiv preprint arXiv:1810.02060*, 2018.

[13] M. Sanjabi, B. Jimmy, M. Razaviyayn, and J. D. Lee, “On the convergence and robustness of training GANs with regularized optimal transport,” in *Proceedings of Advances in Neural Information Processing Systems*, 2018, pp. 7088–7098.

[14] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen, “Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications,” *IEEE Transactions on Signal Processing*, 2019, Submitted.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] D. A. Schmidt, C. Shi, R. A. Berry, M. L. Honig, and W. Utschick, “Comparison of distributed beamforming algorithms for MIMO interference networks,” *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3476–3489, July 2013.

[17] C. Shi, R. A. Berry, and M. L. Honig, “Monotonic convergence of distributed interference pricing in wireless networks,” in *Proceedings of IEEE international conference on Symposium on Information Theory*, 2009.

[18] 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 Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331–4340, 2011.