2021 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, OCT. 25–28, 2021, GOLD COAST, AUSTRALIA

MODEL-FREE LEARNING OF OPTIMAL DETERMINISTIC RESOURCE ALLOCATIONS IN WIRELESS SYSTEMS VIA ACTION-SPACE EXPLORATION

Hassaan Hashmi and Dionysios S. Kalogerias
Department of Electrical Engineering, Yale University, New Haven, USA
{hassaan.hashmi,dionysis.kalogerias}@yale.edu

ABSTRACT

Wireless systems resource allocation refers to perpetual and challenging nonconvex constrained optimization tasks, which are especially timely in modern communications and networking setups involving multiple users with heterogeneous objectives and imprecise or even unknown models and/or channel statistics. In this paper, we propose a technically grounded and scalable primal-dual deterministic policy gradient method for efficiently learning optimal parameterized resource allocation policies. Our method not only efficiently exploits gradient availability of popular universal policy representations, such as deep neural networks, but is also truly model-free, as it relies on consistent zeroth-order gradient approximations of the associated random network services constructed via low-dimensional perturbations in action space, thus fully bypassing any dependence on critics. Both theory and numerical simulations confirm the efficacy and applicability of the proposed approach, as well as its superiority over the current state of the art in terms of both achieving near-optimal performance and scalability.

***Index Terms*—** Wireless Systems, Resource Allocation, Reinforcement Learning, Zeroth-order Optimization, Deterministic Policy Gradients, Deep Learning.

1. INTRODUCTION

Optimally allocating resources in modern wireless systems presents three major challenges [1,2]: *Infinite dimensionality of policies*; *nonconvex constraints*; *system and channel model unavailability*. Building on now classical dual-domain techniques for globally optimal model-based resource allocation [1, 3], as well as effective heuristics [4, 5], *machine learning for wireless communications* has developed as a rapidly expanding area since, aside from nonconvexity (which results in natural algorithmic limitations), it can effectively address both challenges of infinite dimensionality and (partially) model availability; see, e.g., [2, 6–11].

Reinforcement Learning (RL), and in particular *policy gradient algorithms* are attractive for wireless systems resource allocation, because they can naturally accommodate continuous channel responses and allocation decisions, and can produce parameterized optimal resource allocation policies that operate in online, even adaptive settings. Methods involving deterministic and stochastic off-policy learning have been developed recently [10, 12]. These methods employ some form of a gradient function approximator (i.e., a critic), making them essentially model-based, since the choice of such an approximator depends on the structure/nature of the underlying problem. A model-free primal-dual learning method was presented in [9], based on the standard stochastic policy gradient method. However, the method in [9] injects noise both in training and implementation phases due to the use of stochastic policies, which is undesirable in the resource allocation setting. Moreover, although one-point policy gradient estimates might be improved with the use of baselines, the latter are computationally expensive, sample inefficient, and cannot be implemented in an online fashion. Recently, [2] introduced two-point zeroth-order policy gradient representations, optimizing purely deterministic resource allocation policies. However, similar to [9], the approach in [2] presents scalability issues, as the corresponding zeroth-order gradient representations are evaluated directly in the policy parameter space, which can be prohibitively large.

To tackle the limitations in terms of applicability and scalability, we propose a new primal-dual zeroth-order deterministic policy gradient method (referred to as PD-ZDPG+), which relies on consistent zeroth-order two-point gradient approximations of the associated random network services, constructed via *low-dimensional perturbations in action space*, rather than in parameter space (as in [2]). The advantages of our approach are *three-fold*, as follows:

- First, as the policies are deterministic, they directly conform with the standard formulation of resource allocation problems in wireless settings, in which randomized policies are undesirable and potentially redundant (this is not achieved in [9], but it is the case in actor-critic-based methods, such as those in [10, 12]).
- Second, our algorithm is completely model-free, as it depends only on direct service function evaluations (obtainable during the operation of the system), thus effectively bypassing dependence on critics (this is obviously not the case in actor-critic-based methods, such as those in [10, 12], but is naturally achieved in [9] and [2]).
- Third, the complexity of action-space exploration is independent of the dimensionality of the employed policy parameterization and thus, in conjunction with availability of the gradient of the latter (in any off-the-shelf machine learning software library), our approach scales substantially better with respect to the complexity of the problem, compared with the state of the art (this not the case in [2], but is achieved in [9]).

Consequently, the proposed method evidently combines the best elements among competing approaches in the current literature.

In terms of related work, our primal-dual method resembles general purpose zeroth-order deterministic policy gradient methods in multi-stage unconstrained RL [13, 14]; in this work, we essentially consider a single-stage RL formulation, however within a constrained learning setting, leading to a challenging minimax problem formulation.

Lastly, we evaluate the performance of the proposed method against competing policy gradient methods in [2, 9, 10, 12, 15], adapted to the setting studied herein, and with respect to the points made above. Source code for our experiments may be found at: https://github.com/hassaanhashmi/pd_zdpg_plus. Certain proofs are omitted due to lack of space, to be presented in a follow-up work.

978-1-7281-6338-3/21/$31.00 ©2021 IEEE

2. PROBLEM FORMULATION

We consider a generic wireless systems resource allocation problem of the form [1, 2, 9]

$$\begin{aligned} \underset{\mathbf{x},\boldsymbol{\theta}}{\text{maximize}} \quad & g^o(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \leq \mathbb{E}\left\{\mathbf{f}(\boldsymbol{\phi}(\mathbf{H},\boldsymbol{\theta}),\mathbf{H})\right\} \\ & \mathbf{g}(\mathbf{x}) \geq \mathbf{0}, (\mathbf{x},\boldsymbol{\theta}) \in \mathcal{X}\times\Theta \end{aligned} , \tag{1}$$

where $\mathbf{H} \in \mathcal{H} \subseteq \mathbb{R}^{N_H}$ are random fading channels of the network with a joint distribution $\mathcal{M}_{\mathbf{H}}$, $\mathbf{f}(\boldsymbol{\phi}(\mathbf{H},\boldsymbol{\theta}),\mathbf{H})$ are the instantaneous service level metrics, which are functions of the channel vector $\mathbf{H}$ and a parameterized and differentiable policy funtion $\boldsymbol{\phi}: \Theta \times \mathcal{H} \to \mathcal{A} \subseteq \mathbb{R}^{N_R}$, with $\mathcal{A}$ closed, and $g^o : \mathcal{X} \to \mathbb{R}$ and $\mathbf{g} : \mathcal{X} \to \mathbb{R}^{N_g}$ are concave utility functions, usually chosen by the designer. The expected value of $\mathbf{f}(\boldsymbol{\phi}(\mathbf{H},\boldsymbol{\theta}),\mathbf{H})$ bounds the ergodic (in a sense, long-term) service level metrics $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{N_S}$.

Any reasonable iterative optimization algorithm for directly solving (1) requires evaluations of the gradients $\nabla g^o(\mathbf{x})$, $\nabla \mathbf{g}(\mathbf{x})$ and $\nabla\,\mathbb{E}\{\mathbf{f}(\boldsymbol{\phi}(\mathbf{H},\boldsymbol{\theta}),\mathbf{H})\}$. As the exact form of the services $\mathbf{f}$ and the channel distribution $\mathcal{M}_{\mathbf{H}}$ are most often not known a priori (incomplete model knowledge due to dependence on propagation physics and increasing complexity of interference and multiple access management models), it is not possible to evaluate the gradients of these functions [2, 9]. To address this limitation, we focus on a *model-free setting*, i.e., we develop a stochastic approximation approach (bypassing the need for knowledge of $\mathcal{M}_{\mathbf{H}}$), in which we are able to approximate the above gradients using appropriate function evaluations, which may be obtained by merely probing the actual wireless system during policy training (also bypassing the need for gradient evaluations –and thus model information–).

3. SMOOTHED SURROGATES IN ACTION SPACE

We first introduce Gaussian-smoothed versions of the functions involved in problem (1). Such functions demonstrate, under some feasible assumptions, certain desirable properties, particularly pertaining to their gradient evaluations, which we use to develop the proposed method.

3.1. Gaussian Smoothed Functions

To begin, let us define

$$g^o_{\mu_S}(\mathbf{x}) \triangleq \mathbb{E}\left\{g^o(\Pi_{\mathcal{X}}\{\mathbf{x}+\mu_S\boldsymbol{U}_S\})\right\},\ \mathbf{x}\in\mathcal{X}, \tag{2}$$

$$\mathbf{g}_{\mu_S}(\mathbf{x}) \triangleq \mathbb{E}\left\{\mathbf{g}(\Pi_{\mathcal{X}}\{\mathbf{x}+\mu_S\boldsymbol{U}_S\})\right\},\ \mathbf{x}\in\mathcal{X} \text{ and} \tag{3}$$

$$\bar{\mathbf{f}}^{\boldsymbol{\phi}}_{\mu_R}(\boldsymbol{\theta}) \triangleq \mathbb{E}\left\{\mathbf{f}(\Pi_{\mathcal{A}}\left\{\boldsymbol{\phi}(\mathbf{H},\boldsymbol{\theta})+\mu_R\boldsymbol{U}_R\right\},\mathbf{H})\right\},\boldsymbol{\theta}\in\Theta, \tag{4}$$

where $\mu_S \geq 0$ and $\mu_R \geq 0$ are smoothing parameters, $\boldsymbol{U}_S \sim \mathcal{N}(\mathbf{0},\mathbf{I}_S)$ and $\boldsymbol{U}_R \sim \mathcal{N}(\mathbf{0},\mathbf{I}_R)$ are random Gaussian vectors, and $\Pi_{\mathcal{Y}}\{\cdot\}$ denotes projection onto a closed set $\mathcal{Y}$. Also, we assume $\boldsymbol{U}_R$ and $\mathbf{H}$ to be independent. This formulation is new and interesting in that we are *smoothing the function* $\mathbf{f}$ *in the action space* $\mathcal{A}$, rather than the parameter space Θ, as was previously done in [2]. Then, we may introduce a *smoothed surrogate to problem* (1), defined as

$$\begin{aligned} \underset{\mathbf{x},\boldsymbol{\theta}}{\text{maximize}} \quad & g^o_{\mu_S}(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x}+\mathtt{S}(\mu_R) \leq \bar{\mathbf{f}}^{\boldsymbol{\phi}}_{\mu_R}(\boldsymbol{\theta}) \\ & \mathbf{g}_{\mu_S}(\mathbf{x}) \geq \mathbf{0}, (\mathbf{x},\boldsymbol{\theta}) \in \mathcal{X}\times\Theta \end{aligned} , \tag{5}$$

where the non-negative feasibility slack $\mathtt{S} : \mathbb{R}_+ \to \mathbb{R}^{N_S}_+$ prevents constraint violations in (5) relative to (1) and is assumed to have certain properties, as explained later (see Assumption 2). Here, we note that functions g^o and $\mathbf{g}$ are concave. Hence, their smoothed surrogates $g^o_{\mu_S}$ and $\mathbf{g}_{\mu_S}$ are their underestimators respectively [16], and thus do not require feasibility slacks.

Formulation of the surrogate (5) enables zeroth-order gradient representations of g^o_μ, $\mathbf{g}_{\mu_S}$ and $\bar{\mathbf{f}}^{\boldsymbol{\phi}}_{\mu_R}$ (i.e., based only on corresponding function evaluations), whenever they appropriately demonstrate certain properties. In the following subsections, we formally define those properties for the surrogate functions and ascertain the feasibility of the surrogate program (5), in relation to the original resource allocation program (1).

3.2. Properties of Surrogate Functions

We impose appropriate structures on the i-th entries g^i and $g^i_{\mu_S}$ of $\mathbf{g}$ and $\mathbf{g}_{\mu_S}$ where $i \in \mathbb{N}^+_{N_g}$, and on the entries f^i and $\bar{f}^{\boldsymbol{\phi},i}_{\mu_R}$ of $\mathbf{f}$ and $\bar{\mathbf{f}}^{\boldsymbol{\phi}}_{\mu_R}$ where $i \in \mathbb{N}^+_{N_S}$, respectively.

Assumption 1. The following conditions are satisfied:
C1 For every $i \in \{o, \mathbb{N}^+_{N_\mathbf{g}}\}$, $g^i(\mathbf{x})$ is globally Lipschitz with constant $L^i_g < \infty$.
C2 For every $i \in \mathbb{N}^+_{N_S}$, $f^i(\cdot,\mathbf{H})$ is Lipschitz on $\mathcal{A}$ with constant $L^i_f(\mathbf{H})$, such that $\|L^i_f(\mathbf{H})\|_{\mathcal{L}_2} < \infty$.
C3 For every $\boldsymbol{\theta}$ and almost every $\mathbf{H}$, the parameterization $\boldsymbol{\phi}(\mathbf{H},\cdot)$ is Lipschitz in a neighborhood of $\boldsymbol{\theta}$ with constant $L^{\boldsymbol{\theta}}_{\boldsymbol{\phi}}(\mathbf{H})$ such that $\|L^{\boldsymbol{\theta}}_{\boldsymbol{\phi}}(\mathbf{H})\|_{\mathcal{L}_2} < \infty$.

Condition **C2** of Assumption 1 has the following consequences on the behavior of $\mathbb{E}\{f^i(\cdot,\mathbf{H})\}$ for every $i \in \mathbb{N}^+_{N_S}$.

Proposition 1. *Let condition* **C2** *of Assumption 1 be in effect. Then, for every* $i \in \mathbb{N}^+_{N_S}$, $\mathbb{E}\{f^i(\cdot,\mathbf{H})\}$ *is* $\mathbb{E}\{L^i_f(\mathbf{H})\}$*-Lipschitz on* $\mathcal{A}$*. Moreover, it is true that, for every* $(\boldsymbol{\theta},\boldsymbol{u}) \in \Theta \times \mathbb{R}^{N_R}$*,*

$$\begin{aligned} &\mathbb{E}\{|f^i(\Pi_{\mathcal{A}}\left\{\boldsymbol{\phi}(\mathbf{H},\boldsymbol{\theta})+\boldsymbol{u}\right\},\mathbf{H})|\} \\ &\quad\leq \mathbb{E}\{L^i_f(\mathbf{H})\}\|\boldsymbol{u}\|_2+\mathbb{E}\{|f^i(\boldsymbol{\phi}(\mathbf{H},\boldsymbol{\theta}),\mathbf{H})|\}. \end{aligned} \tag{6}$$

Proof. We start with condition **C2** and employ Jensen's inequality and the triangle inequality, respectively. □

We now use Assumption 1 and Proposition 1 to establish well-definiteness and basic properties of $g^o_{\mu_S}$, $\mathbf{g}_{\mu_S}$ and $\bar{\mathbf{f}}^{\boldsymbol{\phi}}_{\mu_R}$: For $\mathbf{x} \in \mathcal{X}$, $\mu_S > 0$ and for every $i \in \{o, \mathbb{N}^+_{N_\mathbf{g}}\}$, let us define finite differences

$$\Delta^i_g(\mathbf{x},\mu_S,\boldsymbol{U}_S) \triangleq \frac{g^i(\Pi_{\mathcal{X}}\{\mathbf{x}+\mu_S\boldsymbol{U}_S\})-g^i(\mathbf{x})}{\mu_S}. \tag{7}$$

Similarly, for all $\boldsymbol{\theta} \in \Theta$, $\mu_R > 0$, and for every $i \in \mathbb{N}^+_{N_S}$ we define

$$\begin{aligned} &\Delta^i_f(\boldsymbol{\theta},\mu_R,\boldsymbol{U}_R,\mathbf{H}) \\ &\triangleq \frac{f^i(\Pi_{\mathcal{A}}\left\{\boldsymbol{\phi}(\mathbf{H},\boldsymbol{\theta})+\mu_R\boldsymbol{U}_R\right\},\mathbf{H})-f^i(\boldsymbol{\phi}(\mathbf{H},\boldsymbol{\theta}),\mathbf{H})}{\mu_R}. \end{aligned} \tag{8}$$

Using Assumption 1 and Proposition 1, we now formally define the set of properties which enable us to evaluate zeroth-order gradients of the smoothed functions defined in Subsection 3.1.

Lemma 2. *Let condition **C1** of Assumption 1 be in effect. Then, for every $i \in \{o, \mathbb{N}^+_{N_\mathbf{g}}\}$ and for every $\mu_S > 0$, each $g^i_{\mu_S}$ is a well-defined, finite, concave and everywhere differentiable underestimator of g^i on $\mathcal{X}$, such that*

$$\sup_{\mathbf{x}\in\mathcal{X}} |g^i_{\mu_S}(\mathbf{x}) - g^i(\mathbf{x})| \leq \mu_S L^i_g \sqrt{N_S}, \tag{9}$$

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\{\|\Delta^i_g(\mathbf{x}, \mu_S, \boldsymbol{U}_S)\boldsymbol{U}_S\|_2^2\} \leq (L^i_g)^2 (N_S + 4)^2, \tag{10}$$

$$\text{and} \quad \mathbb{E}\{\Delta^i_g(\mathbf{x}, \mu_S, \boldsymbol{U}_S)\boldsymbol{U}_S\} \equiv \nabla g^i_{\mu_S}(\mathbf{x}) \tag{11}$$

for all $\mathbf{x} \in \mathcal{X}$.

Lemma 3. *Let condition **C2** of Assumption 1 be in effect. Then, for every $\mu_R > 0$ and for all $i \in \mathbb{N}^+_{N_S}$, each $\bar{f}^{\boldsymbol{\phi},i}_{\mu_R}$ and each $\mathbb{E}\{f^i(\boldsymbol{\phi}(\mathbf{H}, \cdot), \mathbf{H})\} \equiv \bar{f}^{\boldsymbol{\phi},i}(\cdot)$ are such that*

$$\sup_{\boldsymbol{\theta}\in\Theta} |\bar{f}^{\boldsymbol{\phi},i}_{\mu_R}(\boldsymbol{\theta}) - \bar{f}^{\boldsymbol{\phi},i}(\boldsymbol{\theta})| \leq \mu_R \mathbb{E}\{L^i_f(\mathbf{H})\}\sqrt{N_R} \tag{12}$$

$$\sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\{\|\Delta^i_f(\boldsymbol{\theta}, \mu_R, \boldsymbol{U}_R, \mathbf{H})\boldsymbol{U}_R\|_2^2\} \leq \mathbb{E}\{(L^i_f(\mathbf{H}))^2\}(N_R+4)^2. \tag{13}$$

Additionally, it is true that

$$\begin{aligned}&\mathbb{E}\{\Delta^i_f(\boldsymbol{\theta}, \mu_R, \boldsymbol{U}_R, \mathbf{H})\boldsymbol{U}_R|\mathbf{H}\}\\ &\quad\equiv \nabla_{\boldsymbol{a}} \mathbb{E}\{f^i(\Pi_{\mathcal{A}}\{\boldsymbol{a} + \mu_R \boldsymbol{U}_R\}, \mathbf{H})|\mathbf{H}\}\big|_{\boldsymbol{a}=\boldsymbol{\phi}(\mathbf{H},\boldsymbol{\theta})}.\end{aligned} \tag{14}$$

Lemmata 2 and 3 are similar to ([2], Lemmata 3 and 4, respectively). However, in Lemma 3, owing to condition **C2** of Assumption 1, there is a key difference in that we are concerned with zeroth-order gradient evaluations on $\mathcal{A}$ rather than Θ. Thus, the dimension of policy parameter $N_{\boldsymbol{\phi}}$ in the respective result in [2] is replaced by the action space dimension N_R, where $N_{\boldsymbol{\phi}} \gg N_R$.

We now present a deterministic policy gradient theorem which exploits our two-point zeroth-order gradient evaluations in action space, which will be used in the proposed primal-dual learning algorithm presented in Section 4.

Theorem 4 (A Deterministic Policy Gradient Theorem). *Let conditions **C2** and **C3** of Assumption 1 be in effect. Then, for every $\mu_R > 0$, for all $\boldsymbol{\theta} \in \Theta$, and for all $i \in \mathbb{N}^+_{N_S}$, each $\bar{f}^{\boldsymbol{\phi},i}_{\mu_R}(\boldsymbol{\theta})$ is such that*

$$\nabla_{\boldsymbol{\theta}} \bar{f}^{\boldsymbol{\phi},i}_{\mu_R}(\boldsymbol{\theta}) \equiv \mathbb{E}\big\{(\Delta^i_f(\boldsymbol{\theta}, \mu_R, \boldsymbol{U}_R, \mathbf{H})\boldsymbol{U}_R)^T \nabla_{\boldsymbol{\theta}} \boldsymbol{\phi}(\mathbf{H}, \boldsymbol{\theta})\big\}. \tag{15}$$

Proof. For $i \in \mathbb{N}^+_{N_S}$, consider the expectation function $\bar{f}^{\boldsymbol{\phi},i}_{\mu_R}(\cdot) \equiv \mathbb{E}\left\{f^i(\Pi_{\mathcal{A}}\{\boldsymbol{\phi}(\mathbf{H}, \cdot) + \mu_R \boldsymbol{U}_R\}, \mathbf{H})\right\}$. Then, from the law of total expectation, the Leibniz integral rule, and (14), we verify (15) as

$$\begin{aligned}&\nabla_{\boldsymbol{\theta}} \bar{f}^{\boldsymbol{\phi},i}_{\mu_R}(\boldsymbol{\theta})\\ &= \nabla_{\boldsymbol{\theta}} \mathbb{E}\{\mathbb{E}\{f^i(\Pi_{\mathcal{A}}\{\boldsymbol{\phi}(\mathbf{H}, \boldsymbol{\theta}) + \mu_R \boldsymbol{U}_R\}, \mathbf{H})|\mathbf{H}\}\}\\ &\equiv \mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathbb{E}\{f^i(\Pi_{\mathcal{A}}\{\boldsymbol{\phi}(\mathbf{H}, \boldsymbol{\theta}) + \mu_R \boldsymbol{U}_R\}, \mathbf{H})|\mathbf{H}\}\}\\ &\equiv \mathbb{E}\big\{\nabla_{\boldsymbol{a}} \mathbb{E}\{f^i(\Pi_{\mathcal{A}}\{\boldsymbol{a} + \mu_R \boldsymbol{U}_R\}, \mathbf{H})|\mathbf{H}\}^T\big|_{\boldsymbol{a}=\boldsymbol{\phi}(\mathbf{H},\boldsymbol{\theta})}\\ &\qquad\qquad \times \nabla_{\boldsymbol{\theta}} \boldsymbol{\phi}(\mathbf{H}, \boldsymbol{\theta})\big\}\\ &\equiv \mathbb{E}\big\{(\Delta^i_f(\boldsymbol{\theta}, \mu_R, \boldsymbol{U}_R, \mathbf{H})\boldsymbol{U}_R)^T \nabla_{\boldsymbol{\theta}} \boldsymbol{\phi}(\mathbf{H}, \boldsymbol{\theta})\big\},\end{aligned} \tag{16}$$

for all $\boldsymbol{\theta} \in \Theta$. □

3.3. Feasible Solutions with Surrogate

Before proceeding with exploiting our smoothed surrogate (5) together with Theorem 4, we put forth conditions that ensure feasibility of the surrogate, specifically on the feasible set of the *original* parameterized problem (1), which is the one we can initially specify.

The premise of our formulation is to show the existence of at least one strictly feasible point for (1) and (5) *simultaneously*. Moreover, as we also show, (5) can be made strictly feasible at will (due to the feasibility slack $\mathtt{S}(\mu_R)$). We first define Lipschitz constant vectors

$$\boldsymbol{c}_S \triangleq \left[L^1_g \dots L^{N_\mathbf{g}}_g\right]^T \text{ and } \boldsymbol{c}_R \triangleq \left[\mathbb{E}\{L^1_f\} \dots \mathbb{E}\{L^{N_S}_f\}\right]^T, \tag{17}$$

and consider the following assumption.

Assumption 2. The *feasibility slack* $\mathtt{S}_\mathbf{f}$ is increasing around the origin, and $\lim_{\mu_R \to 0} \mathtt{S}_\mathbf{f}(\mu_R) \equiv \mathtt{S}_\mathbf{f}(0) \equiv \mathbf{0}$.

We now state results which guarantee feasibility of (5) relative to (1). We analyze both the strict feasibility of (5) and constraint violations in (1) for every feasible solution of (5).

Theorem 5. *Let Assumptions 1 and 2 be in effect, and let $(\mathbf{x}^*, \boldsymbol{\theta}^*) \in \mathbb{R}^{N_S} \times \mathbb{R}^{N_{\boldsymbol{\phi}}}$ be a strictly feasible point of the problem (1). Then, there exist $\mu^*_S > 0$ and $\mu^*_R > 0$, possibly dependent on $(\mathbf{x}^*, \boldsymbol{\theta}^*)$, such that, for every $0 \leq \mu_S \leq \mu^*_S$ and $0 \leq \mu_R \leq \mu^*_R$, the point $(\mathbf{x}^*, \boldsymbol{\theta}^*)$ is strictly feasible for (5).*

Proof. Similar to that of Theorem 6 in [2], with key differences of N_R being in place of $N_{\boldsymbol{\phi}}$ and the Lipschitz vector $\boldsymbol{c}_R$ being on $\mathcal{A}$ rather than Θ. □

Theorem 6. *Let Assumption 1 be in effect. Then, for every $\mu_R \geq 0$ such that*

$$\mathtt{S}(\mu_R) - \mu_R \boldsymbol{c}_R \sqrt{N_R} \geq \mathbf{0} \tag{18}$$

and for every $\mu_S \geq 0$, every feasible point of (5) is also feasible for (1). Otherwise, negative LHS values in (18) correspond to the respective levels of maximal constraint violations for (1).

Proof. Similar to that of Theorem 7 in [2], with key differences of N_R being in place of $N_{\boldsymbol{\phi}}$ and the Lipschitz vector $\boldsymbol{c}_R$ being on $\mathcal{A}$ rather than Θ. □

From the statements of Theorems 5 and 6, we see that both of these can hold simultaneously. These, however, are different from the respective results in [2], since that the slack $\mathtt{S}(\mu_R)$ deals with perturbations in the action space $\mathcal{A}$ rather than the parameter space Θ, which is much more flexible in terms of implementation.

4. MODEL-FREE PRIMAL-DUAL LEARNING IN RESOURCE ALLOCATION SPACES

To efficiently tackle constrained problem (1) within a fully model-free setting, the idea is to exploit the zeroth-order representations of the smoothed versions of the functions involved in (1), as introduced in Section 3.2. To do this, we first define the *Lagrangian* of the smoothed surrogate (5) as

$$\begin{aligned}\mathcal{L}_{\phi,\mu}(\boldsymbol{\theta}, \mathbf{x}, \boldsymbol{\lambda}_S, \boldsymbol{\lambda}_R) \triangleq\ & g^o_{\mu_S}(\mathbf{x}) + \boldsymbol{\lambda}^T_S \mathbf{g}_{\mu_S}(\mathbf{x})\\ &+ \boldsymbol{\lambda}^T_R [\bar{\mathbf{f}}^{\boldsymbol{\phi}}_{\mu_R}(\boldsymbol{\theta}) - \mathbf{x} - \mathtt{S}(\mu_R)].\end{aligned} \tag{19}$$

Algorithm 1: PD-ZDPG+

Initialize $\mathbf{x}^0$, $\boldsymbol{\theta}^0$, $\boldsymbol{\lambda}_S^0$, $\boldsymbol{\lambda}_R^0$, $\alpha_{\mathbf{x}}^0$, $\alpha_{\boldsymbol{\theta}}^0$, $\alpha_{\boldsymbol{\lambda}_S}^0$, $\alpha_{\boldsymbol{\lambda}_R}^0$, μ_S, μ_R.
for $k = 0, N-1$ **do**
  Sample $\boldsymbol{U}_S^{k+1}$ and $\boldsymbol{U}_R^{k+1}$, and measure $\mathbf{H}^{k+1}$.
  Evaluate $g^o(\mathbf{x}^k)$, $\mathbf{g}(\mathbf{x}^k)$, $g^o(\Pi_{\mathcal{X}}\{\mathbf{x}^k+\mu_S\boldsymbol{U}_S^{k+1}\})$ and $\mathbf{g}(\Pi_{\mathcal{X}}\{\mathbf{x}^k+\mu_S\boldsymbol{U}_S^{k+1}\})$.
  Probe the system for $\mathbf{f}(\boldsymbol{\phi}(\mathbf{H}^{k+1},\boldsymbol{\theta}^k),\mathbf{H}^{k+1})$ and $\mathbf{f}(\Pi_{\mathcal{A}}\{\boldsymbol{\phi}(\mathbf{H}^{k+1},\boldsymbol{\theta}^k)+\mu_R\boldsymbol{U}_R^{k+1}\},\mathbf{H}^{k+1})$.
  Update $\mathbf{x}^{k+1}$ and $\boldsymbol{\theta}^{k+1}$ using (21) and (22).
  Evaluate $\mathbf{g}(\Pi_{\mathcal{X}}\{\mathbf{x}^{k+1}+\mu_S\boldsymbol{U}_S^{k+1}\})$ and probe the system for $\mathbf{f}(\Pi_{\mathcal{A}}\{\boldsymbol{\phi}(\mathbf{H}^{k+1},\boldsymbol{\theta}^{k+1})+\mu_R\boldsymbol{U}_R^{k+1}\},\mathbf{H}^{k+1})$.
  Update $\boldsymbol{\lambda}_S^{k+1}$ and $\boldsymbol{\lambda}_R^{k+1}$ using (23) and (24).
end

Then, driven by both classical and state-of-the-art approaches, see, e.g., [1,2,9,12], we are interested in the minimax problem

$$\begin{array}{ll}\underset{\boldsymbol{\lambda}_S,\boldsymbol{\lambda}_R}{\text{minimize}} & \sup_{(\mathbf{x},\boldsymbol{\theta})\in\mathcal{X}\times\Theta}\ \mathcal{L}_{\phi,\mu}(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\lambda}_S,\boldsymbol{\lambda}_R)\\ \text{subject to} & \boldsymbol{\lambda}_S\geq\mathbf{0},\boldsymbol{\lambda}_R\geq\mathbf{0}\end{array}. \tag{20}$$

We now develop our proposed zeroth-order primal-dual learning algorithm to solve (20). By Lemma 2 and Theorem 4, gradients of $g_{\mu_S}^o$, $\mathbf{g}_{\mu_S}$ and $\bar{\mathbf{f}}_{\mu_R}^{\boldsymbol{\phi}}$ are given by expectation functions in (11) and (15). Given standard Gaussian i.i.d. sequences $\{\boldsymbol{U}_S^k\}_{k\in\mathbb{N}^+}$, $\{\boldsymbol{U}_R^k\}_{k\in\mathbb{N}^+}$, and a mutually independent channel state observation sequence $\{\mathbf{H}^k\}_{k\in\mathbb{N}^+}$, we define parameter update rules employing stochastic approximation as

$$\mathbf{x}^{k+1}\equiv\Pi_{\mathcal{X}}\{\mathbf{x}^k+\alpha_{\mathbf{x}}^k\circ(\Delta_g^i(\mathbf{x}^k,\mu_S,\boldsymbol{U}_S^{k+1})\boldsymbol{U}_S^{k+1} +\boldsymbol{U}_S^{k+1}\boldsymbol{\Delta}_{\mathbf{g}}(\mathbf{x}^k,\mu_S,\boldsymbol{U}_S^{k+1})^T\boldsymbol{\lambda}_S^k-\boldsymbol{\lambda}_R^k)\}, \tag{21}$$

$$\boldsymbol{\theta}^{k+1}\equiv\Pi_{\Theta}\{\boldsymbol{\theta}^k+\alpha_{\boldsymbol{\theta}}^k\circ(\nabla_{\boldsymbol{\theta}}\boldsymbol{\phi}(\mathbf{H}^{k+1},\boldsymbol{\theta}^k)^T\boldsymbol{U}_R^{k+1} \times\boldsymbol{\Delta}_{\mathbf{f}}(\boldsymbol{\theta}^k,\mu_R,\boldsymbol{U}_R^{k+1},\mathbf{H}^{k+1})^T\boldsymbol{\lambda}_R^k)\}, \tag{22}$$

$$\boldsymbol{\lambda}_S^{k+1}\equiv[\boldsymbol{\lambda}_S^k-\alpha_{\boldsymbol{\lambda}_S}^k\circ\mathbf{g}(\Pi_{\mathcal{X}}\{\mathbf{x}^{k+1}+\mu_S\boldsymbol{U}_S^{k+1}\})]_+ \text{ and} \tag{23}$$

$$\boldsymbol{\lambda}_R^{k+1}\equiv[\boldsymbol{\lambda}_R^k-\alpha_{\boldsymbol{\lambda}_R}^k\circ(\mathbf{f}(\Pi_{\mathcal{A}}\{\boldsymbol{\phi}(\mathbf{H}^{k+1},\boldsymbol{\theta}^{k+1})+\mu_R\boldsymbol{U}_R^{k+1}\},\mathbf{H}^{k+1}) -\mathbf{x}^{k+1}-\mathbb{S}(\mu_R))]_+, \tag{24}$$

where, dropping dependencies, the vectors of finite differences $\boldsymbol{\Delta}_{\mathbf{g}}\in\mathbb{R}^{N_{\mathbf{g}}}$ and $\boldsymbol{\Delta}_{\mathbf{f}}\in\mathbb{R}^{N_S}$ are defined as

$$\boldsymbol{\Delta}_{\mathbf{g}}\triangleq\left[\Delta_g^1\dots\Delta_g^{N_{\mathbf{g}}}\right]^T \quad\text{and}\quad \boldsymbol{\Delta}_{\mathbf{f}}\triangleq\left[\Delta_f^1\dots\Delta_f^{N_S}\right]^T \tag{25}$$

respectively. A complete description of our model-free primal-dual method is provided in Algorithm 1.

As problem (5) is non-convex (due to $\bar{\mathbf{f}}_{\mu_R}^{\boldsymbol{\phi}}(\boldsymbol{\theta})$), it is difficult to guarantee convergence of Algorithm 1 theoretically. Nevertheless, it is true that if the algorithm converges, then it discovers a feasible (though possibly suboptimal) solution of the problem. In the next section, we evaluate the performance of Algorithm 1 empirically instead, on two indicative examples.

5. SIMULATIONS

We consider two basic wireless models, an Additive White Gaussian Noise (AWGN) channel, and a Multiple Access Inference (MAI) channel. All experiments that follow were rerun five times and plotted in terms of mean performances and bootstrap confidence bands without any *cherry-picking*.

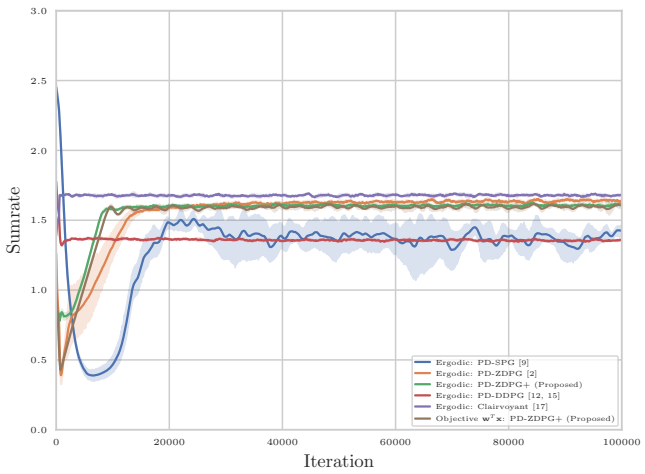

Fig. 1: Sumrates (*in nats per unit of time*) achieved by the proposed PD-ZDPG+, PD-SPG [9], PD-ZDPG [2], PD-DDPG [12,15] and the clairvoyant policy [17] for the AWGN channel case.

For the AWGN channel case, we consider a scenario where multiple users are given dedicated channels to communicate, without interference. The goal is to achieve optimal power allocation given an average total power budget p_{max}. This problem can be formally stated as

$$\begin{array}{ll}\underset{x^i,\boldsymbol{\theta}^i}{\text{maximize}} & \sum_i w^i x^i\\ \text{subject to} & x^i\leq\mathbb{E}\left\{\log\left(1+\dfrac{H^i\phi^i(H^i,\boldsymbol{\theta}^i)}{v^i}\right)\right\},\\ & \mathbb{E}\left\{\sum_i\phi^i(H^i,\boldsymbol{\theta}^i)\right\}\leq p_{max}\\ & (x^i,\boldsymbol{\theta}^i)\in\mathbb{R}_+\times\mathbb{R}^{N_\phi^i},\forall i\in\mathbb{N}_{N_S}^+\end{array} \tag{26}$$

where all user weights w^i are positive, randomly generated, and sum to 1, and where the policy is uncoupled among users. This is due to the simplicity of problem (26), for which a specially-structured strictly optimal *clairvoyant* solution requiring complete system model information is available [17]; this also defines an ultimate performance benchmark under the AWGN setup.

We consider a 10-user case for the AWGN channel problem and set $p_{max}\equiv 20$. Also, as in all experiments in this section, noise variance $v^i\equiv 1$ and H^i is exponentially distributed with parameter $\lambda\equiv 1/2$. We then compare the proposed method (PD-ZDPG+) with other primal-dual methods including the stochastic (actor-only) policy gradient (PG) method of [9] (PD-SPG), the deterministic zeroth order actor-only PG method of [2] (PD-ZDPG), and a customized primal-dual version of a deep deterministic actor-critic policy gradient method [15] (PD-DDPG).

In all deterministic policy gradient methods, each policy features a ReLU-activated three-layer single-input single-p_{max}-sigmoid-output feed-forward DNN with eight and four neurons in the hidden layers (i.e., ten $1-8-4-1$ DNNs). For the stochastic policies, we define the same networks, but with two sigmoid outputs scaled by p_{max} and $\sqrt{p_{max}}$, corresponding to the mean and variance of a truncated normal distribution from which actions are sampled [9]. For the global *critic* in PD-DDPG, we consider a ReLU-activated three-layer feed-forward DNN with twenty and forty neurons in the

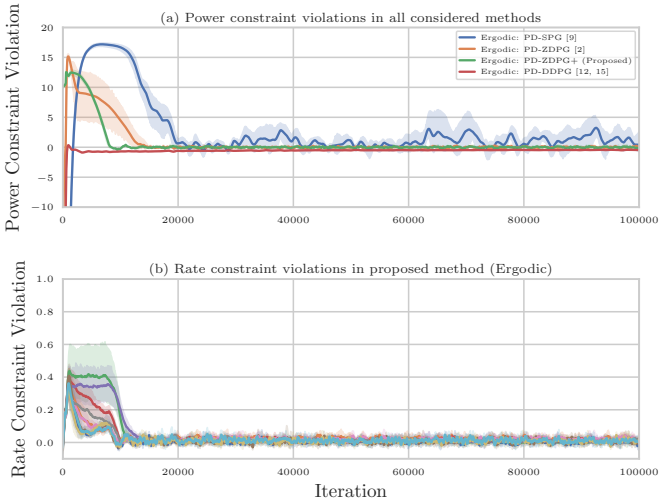

Fig. 2: (a) Power constraint violation exhibited by the proposed PD-ZDPG+, PD-SPG [9], PD-ZDPG [2] and PD-DDPG [12, 15], for the AWGN channel case.(b) Rate constraint violations of PD-ZDPG+.

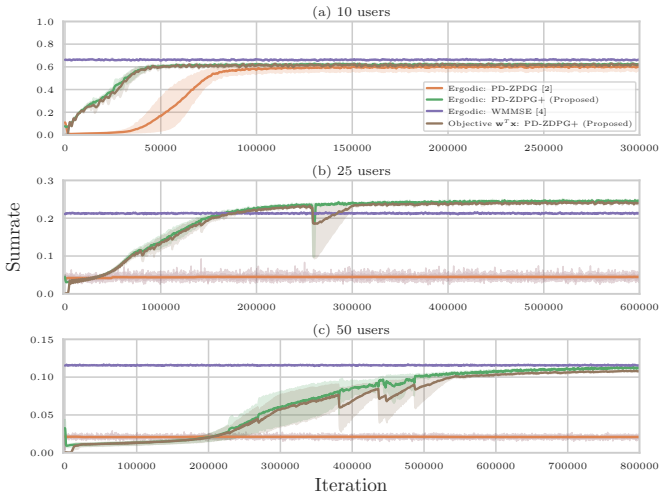

Fig. 3: Sumrates (*in nats per unit of time*) achieved by the proposed PD-ZDPG+, PD-ZDPG [2] and the WMMSE policy [4] for the MAI channel case with (a) 10 users (b) 25 users and (c) 50 users.

hidden layers with all parameters initialized at 10^{-6}. The inputs to the *critic* are instantaneous channel values, whereas all policy action values are concatenated with the output of the first hidden as input to the second hidden layers. All actor parameters $\boldsymbol{\theta}$ are initialized to 0's while all metrics $\mathbf{x}$ and dual variables are initialized to 1's.

Given that the objective description is known, we smooth only the constraints of the problem (cf. (5)). The exact learning rates for all the methods can be found in Table 1. For PD-SPG, we use the Adam optimizer with β_1 and β_2 set to 0.9 and 0.999 respectively with batch size of 32 where as all other methods use SGD optimization. We show the convergence of all methods over 10^5 iterations in Figure 1 and report that the proposed method converges both faster and to a near-optimal solution as compared with other methods. We also show a similar behavior for constraint violations in Figure 2.

Method	$\alpha^k_{\mathbf{x}}$	$\alpha^k_{\boldsymbol{\theta}}$	$\alpha^k_{\boldsymbol{\lambda}_R}$	$\alpha^k_{\boldsymbol{\lambda}_R}*$
PD-SPG [9]	0.01	0.01	0.0001	0.08
PD-ZDPG [2]	0.001	0.0008	0.008	0.0001
PD-ZDPG+(Proposed)	0.001	0.02	0.008	0.0001
PD-DDPG [12, 15]	0.001	0.002/0.001	0.01	0.0001

Table 1: Learning rates for the AWGN channel ($*$ Power constraint)

Optimizing stochastic policies proves to be a noisy procedure. In fact, these policies not only converge to suboptimal solutions, but are also not desirable as they inject extra noise into the system during implementation, inducing unnecessary statistical variability in performance. Interestingly, PD-DDPG also converges to the same solution as PD-SPG showing that using a *critic* is indeed a heuristic. As such, we witness the credit assignment problem in terms of the individual policies becoming apparent in PD-DDPG as the critic inaccurately approximates the functions in the constraints. PD-ZDPG+, on the other hand, approaches the globally optimal solution to the AWGN problem at hand rather faithfully, achieving similar performance to PD-ZDPG, which relies on high-dimensional parameter space exploration.

Comparable performances of PD-ZDPG [2] and our PD-ZDPG+ on the simple AWGN channel problem shows that our method performs as well as the state-of-the-art. To further compare the scalability of these algorithms, we consider a MAI channel model setting, in which N_S transmitters communicate simultaneously with a central node. Thus, a signal transmitted by each user will interfere with the signals transmitted by all the other users. Again, the task is to optimize power allocation per user given an average power budget of p_{max}. In this case, our resource allocation problem is formulated as

$$\begin{aligned} &\underset{x^i,\boldsymbol{\theta}}{\text{maximize}} \sum_i w^i x^i \\ &\text{subject to } x^i \leq \mathbb{E}\left\{\log\left(1+\frac{H^i\phi^i(H,\boldsymbol{\theta})}{v^i+\sum_{i\neq j} H^j\phi^j(H,\boldsymbol{\theta})}\right)\right\}. \\ &\mathbb{E}\left\{\sum_i \phi^i(H,\boldsymbol{\theta})\right\} \leq p_{max} \\ &(x^i,\boldsymbol{\theta}) \in \mathbb{R}_+ \times \mathbb{R}^{N_\phi}, \forall i \in \mathbb{N}^+_{N_S} \end{aligned} \tag{27}$$

In the following, we consider the MAI problem for 10, 25 and 50 users, where, again, the total allocated power budget $p_{max} \equiv 20$ and all the user weights w^i are positive, randomly generated, and sum to 1. For our implementations, we define global policies via ReLU activated three layer neural networks having 64 and 32 neurons in the hidden layers respectively (a single $\#-64-32-\#$ DNN, where "#" is the number of users). All policy parameters $\boldsymbol{\theta}$ and all metrics $\mathbf{x}$ are initialized to 0's while all dual variables are initialized to 1's.

Method	$\alpha^k_{\mathbf{x}}$	$\alpha^k_{\boldsymbol{\theta}}$	$\alpha^k_{\boldsymbol{\lambda}_R}$	$\alpha^k_{\boldsymbol{\lambda}_R}*$
PD-ZDPG [2]	0.001	0.00005	0.004	0.0001
PD-ZDPG+(Proposed)	0.001	0.04	0.008	0.0001

Table 2: Learning rates for the MAI channel ($*$ Power constraint)

We compare these methods with the well-known WMMSE policy [4] (at saturation) as a benchmark upper bound. We use the same parameterization and learning rates in all experiments as given in Table 2. PD-ZDPG consistently converged to the optimal solution in all cases (even occasionally outperforming WMMSE), whereas for the 25- and 50-user cases, PD-ZPDG did not converge and we were unable to find effective learning rates to make the method convergent.

Finally, we observed that PD-ZDPG+ is also much faster in terms of training times (in Python), as shown in Figure 4. These comparisons are justified according to the documentation of Python's PyTorch library, which we have used in our experiments.

6. CONCLUSION

In this work, we have proposed a new primal-dual method for learning resource allocations in wireless systems by exploiting zeroth-order deterministic policy gradients via low-dimensional action space exploration. Our method works with powerful policy parameterizations such as DNNs, and outperforms the current state-of-the-art, both in terms of optimal convergence and scalability. This is due to the fact the our method optimizes deterministic policies, is completely model-free, and is not limited, in terms of scalability, by the dimensionality of employed policy parameterizations. Although we have only tested feed-forward neural networks in our numerics, PD-ZDPG+ works equally fine with other deep learning models like convolutional and recurrent neural networks, as well as architectures which exploit intrinsic network structure, such as random edge graph neural networks. Therefore, PD-ZDPG+ can be applied to a wide range of constrained resource allocation problems. In the future, we would like to evaluate the behavior of PD-ZDPG+ for such more elaborate policy parameterizations. We are also interested to see how our proposed method extends and performs on problems which are outside the scope of wireless systems resource allocation, such as problems arising in general purpose constrained RL.

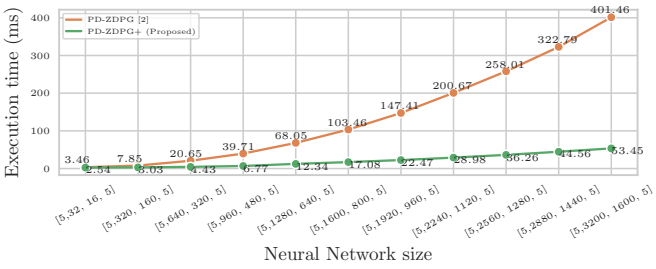

Fig. 4: Best training execution times per iteration (in ms) of the proposed PD-ZDPG+ and PD-ZDPG [2] for the MAI channel case with 5 users and multiple neural network sizes, using 2 virtual Intel Xeon CPUs @ 2.30GHz, 13GB of RAM, and an 8 core TPU v3.

7. REFERENCES

[1] Alejandro Ribeiro, "Optimal resource allocation in wireless communication and networking," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, pp. 1–19, 2012.

[2] Dionysios S Kalogerias, Mark Eisen, George J Pappas, and Alejandro Ribeiro, "Model-free learning of optimal ergodic policies in wireless systems," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6272–6286, 2020.

[3] Wei Yu and Raymond Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Transactions on communications*, vol. 54, no. 7, pp. 1310–1322, 2006.

[4] Qingjiang Shi, Meisam Razaviyayn, Zhi-Quan Luo, and Chen 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.

[5] Xinzhou Wu, Saurabha Tavildar, Sanjay Shakkottai, Tom Richardson, Junyi Li, Rajiv Laroia, and Aleksandar Jovicic, "Flashlinq: A synchronous distributed scheduler for peer-to-peer ad hoc networks," *IEEE/ACM Transactions on networking*, vol. 21, no. 4, pp. 1215–1228, 2013.

[6] Haoran Sun, Xiangyi Chen, Qingjiang Shi, Mingyi Hong, Xiao Fu, and Nicholas D Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5438–5453, 2018.

[7] Hoon Lee, Sang Hyun Lee, and Tony QS Quek, "Deep learning for distributed optimization: Applications to wireless resource management," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2251–2266, 2019.

[8] Yasar Sinan Nasir and Dongning Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2239–2250, 2019.

[9] Mark Eisen, Clark Zhang, Luiz FO Chamon, Daniel D Lee, and Alejandro Ribeiro, "Learning optimal resource allocations in wireless systems," *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2775–2790, 2019.

[10] Fan Meng, Peng Chen, Lenan Wu, and Julian Cheng, "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches," *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6255–6267, 2020.

[11] Ahmad Ali Khan and Raviraj S Adve, "Centralized and distributed deep reinforcement learning methods for downlink sum-rate optimization," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 8410–8426, 2020.

[12] Xiaoming Wang, Yuhan Zhang, Ruijuan Shen, Youyun Xu, and Fu-Chun Zheng, "Drl-based energy-efficient resource allocation frameworks for uplink noma systems," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7279–7294, 2020.

[13] Harshat Kumar, Dionysios S Kalogerias, George J Pappas, and Alejandro Ribeiro, "Zeroth-order deterministic policy gradient," *arXiv preprint arXiv:2006.07314*, 2020.

[14] Anirudh Vemula, Wen Sun, and J. Bagnell, "Contrasting exploration in parameter and action space: A zeroth-order optimization perspective," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[15] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra, "Continuous control with deep reinforcement learning," in *4th International Conference on Learning Representations*, 2016.

[16] Yurii Nesterov and Vladimir Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[17] Xin Wang and Na Gao, "Stochastic resource allocation over fading multiple access and broadcast channels," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2382–2391, 2010.