

Scalable Min-Max Optimization via Primal-Dual Exact Pareto Optimization

Sangwoo Park
Imperial College London
London, United Kingdom
s.park@imperial.ac.uk

Stefan Vlaski
Imperial College London
London, United Kingdom
s.vlaski@imperial.ac.uk

Lajos Hanzo
University of Southampton
Southampton, United Kingdom
hanzo@soton.ac.uk

Abstract—In multi-objective optimization, minimizing the worst objective can be preferable to minimizing the average objective, as this ensures improved fairness across objectives. Due to the non-smooth nature of the resultant min-max optimization problem, classical subgradient-based approaches typically exhibit slow convergence. Motivated by primal-dual consensus techniques in multi-agent optimization and learning, we formulate a smooth variant of the min-max problem based on the augmented Lagrangian. The resultant Exact Pareto Optimization via Augmented Lagrangian (EPO-AL) algorithm scales better with the number of objectives than subgradient-based strategies, while exhibiting lower per-iteration complexity than recent smoothing-based counterparts. We establish that every fixed-point of the proposed algorithm is both Pareto and min-max optimal under mild assumptions and demonstrate its effectiveness in numerical simulations.

Index Terms—Multi-objective optimization, min-max optimization, exact Pareto optimality, primal-dual consensus, augmented Lagrangian.

I. INTRODUCTION

We consider a multi-objective optimization problem having K differentiable, positive objectives $J_1(w), \dots, J_K(w) > 0$ where $J_k(w)$ is the k -th objective evaluated at the model $w \in \mathbb{R}^d$. We wish to solve the *weighted min-max* problem [1], [2]:

$$\min_{w \in \mathbb{R}^d} \max_{k \in [K]} r_k J_k(w), \quad (1)$$

given a pre-determined *preference vector* $r = [r_1, \dots, r_K]^\top$ associated with $r_k > 0$ for $k = 1, \dots, K$. We focus our attention on the setting where the gradient information $\{\nabla J_k(w)\}_{k=1}^K$ is available. Solving the min-max problem (1) can be preferable to classical linear scalarization, i.e., $\min_{w \in \mathbb{R}^d} \sum_{k=1}^K r_k J_k(w)$, in applications where fairness is important [3]–[7], since it ensures that no individual objective $J_k(\cdot)$ is neglected in the interest of improving the average performance.

Perhaps the conceptually simplest approach for solving (1) is to consider the subgradient algorithm [10, Theorem 18.5]:

$$w_{i+1} = w_i - \mu \cdot r_{k_i^{\text{act}}} \nabla J_{k_i^{\text{act}}}(w_i) \quad (2)$$

where μ is a step-size and $k_i^{\text{act}} \triangleq \arg \max_k r_k J_k(w_i)$ denotes the index for the *active* objective that attains the maximum

This work was supported by EPSRC Grants EP/X04047X/1 and EP/Y037243/1.

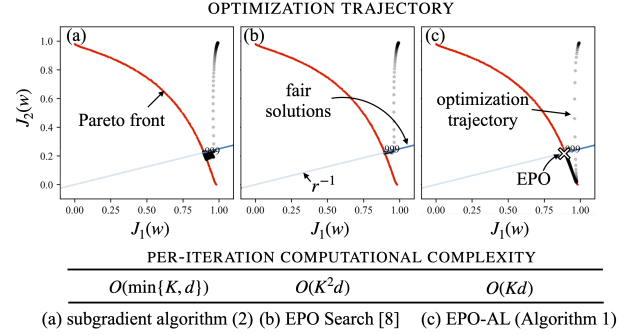


Fig. 1. Multi-objective optimization trajectories (top) for subgradient algorithm (2) in (a), EPO Search [8] in (b), and the proposed approach via augmented Lagrangian in (c), referred to as EPO-AL; the table shows the per-iteration computational complexities (bottom). Optimization trajectories are obtained for $K = 2$ non-convex objectives $J_1(w) = 1 - e^{-\|w-1/\sqrt{d}\|^2}$ and $J_2(w) = 1 - e^{-\|w+1/\sqrt{d}\|^2}$ [9] with $w \in \mathbb{R}^3$ (see Sec. V-A for details). The intersection between the Pareto front (red arc, see Def. 1) and the fair solutions (blue line, see Def. 2) is an exact Pareto optimal (EPO) [8] solution (white cross, see Def. 3), which satisfies the min-max optimality (1) under mild assumptions (see Prop. 1). Observe that the proposed strategy first finds the Pareto front, and then searches for the Pareto solution that is min-max optimal according to (1). The subgradient algorithm (2) exhibits oscillations around the min-max optimal solution Pareto front due to the non-smooth behavior of maximum operator in (1), unlike both EPO-based approaches that smoothly converge to the min-max optimal solution.

in (1) at iteration i , i.e., $\max_k r_k J_k(w_i) = r_{k_i^{\text{act}}} J_{k_i^{\text{act}}}(w_i)$. Note that, if more than one objective achieves the maximum for a particular model w , i.e., the set $\mathcal{A}(w) = \{k' \in [K] : J_{k'}(w) = \max_k J_k(w)\}$ contains more than one element, then any convex combination of $\{\nabla J_k(w)\}_{k \in \mathcal{A}(w)}$ is a subgradient of the min-max objective (1) and can be utilized in (2) [11].

However, the iterative update rule (2) generally suffers from slow convergence rate [11], primarily due to ignoring the gradient information gleaned from inactive objectives. This observation has motivated the development of smoothing-based alternatives [4], [11]–[13]. Existing approaches for (1) either rely on (i) directly replacing the max-function by a smooth approximation [5], [12]–[15], or (ii) designing a smooth saddle point problem [4], [7], [11], i.e.,

$$w^* \in \arg \min_{w \in \mathbb{R}^d} \max_{y \in \Delta^K} \sum_{k=1}^K r_k J_k(w) y_k, \quad (3)$$

where y_k is the k -th element of the $(K-1)$ -simplex $y \in \Delta^K$

given by $\Delta^K = \{y \in \mathbb{R}_+^K : y^\top \mathbb{1}_K = 1\}$ with $\mathbb{1}_K$ being the $K \times 1$ all-one vector.

II. PRELIMINARIES

In this paper, we consider an alternative smoothing approach for min-max optimization problem (1) inspired by classical results in multi-objective optimization capturing the relationship between min-max and Pareto optimization [16]. To formalize the discussion, we introduce the following definitions.

Definition 1 (Weak Pareto optimality [1]). *A model w is weakly Pareto optimal, if there exists no $w' \neq w$ for which all objectives are reduced. Accordingly, the collection $\mathcal{P}$ of all weakly Pareto optimal points is given by*

$$\mathcal{P} = \{w \in \mathbb{R}^d : \nexists w' \in \mathbb{R}^d \text{ s.t. } J_k(w') < J_k(w) \text{ for all } k\}. \quad (4)$$

The set of (weakly) Pareto optimal points is illustrated as a red arc in Fig. 1. Second, we introduce the set of *fair* points [7], [17].

Definition 2 (Fairness [7], [17]). *A model w is called fair with respect to the preference vector r if the weighted objectives are equal, i.e., $r_1 J_1(w) = \dots = r_K J_K(w)$. The collection of all such fair models is denoted by the set:*

$$\mathcal{F}_r = \{w \in \mathbb{R}^d : r_1 J_1(w) = \dots = r_K J_K(w)\}. \quad (5)$$

The set of fair points is illustrated as a blue line in Fig. 1. When the set of (weakly) Pareto optimal solutions and the set of fair solutions intersect, this gives rise to the set of *exact Pareto optimal* solutions (white cross in Fig. 1).

Definition 3 (Exact Pareto optimality [8]). *The set of exact Pareto optimal solutions is given by:*

$$\mathcal{E}_r = \mathcal{P} \cap \mathcal{F}_r. \quad (6)$$

The set $\mathcal{E}_r$ is given by the intersection of the red arc (the weakly Pareto optimal solutions) and the blue line (the fair solutions) in Fig. 1. When this intersection is non-empty, we refer to r as *being Pareto feasible*. We remark that compared to [8, Eq. (8)] we define $\mathcal{E}_r$ in terms of global (weak) Pareto optimality, rather than merely local Pareto optimality. This allows us to establish the following proposition.

Proposition 1 (Exact Pareto optimality implies min-max optimality). *Assume that $w^{\text{EPO}} \in \mathcal{E}_r = \mathcal{P} \cap \mathcal{F}_r$ is weakly Pareto optimal and fair. Then w^{WPO} is also min-max optimal as defined in (1):*

$$w^{\text{EPO}} = \arg \min_{w \in \mathbb{R}^d} \max_{k \in [K]} r_k J_k(w). \quad (7)$$

Proof. We prove the statement by contradiction. Suppose that $w^{\text{EPO}} \in \mathcal{P} \cap \mathcal{F}_r$ does not minimize (1). Then:

$$\exists w : \max_{k \in [K]} r_k J_k(w) < \max_{k \in [K]} r_k J_k(w^{\text{EPO}}). \quad (8)$$

We first simplify the right-hand side by using the fairness condition (5). From $w^{\text{EPO}} \in \mathcal{F}_r$, we have

$$\max_{k \in [K]} r_k J_k(w) < r_\ell J_\ell(w^{\text{EPO}}) \text{ for any } \ell \in [K]. \quad (9)$$

Since each element of a set is upper bounded by its maximum:

$$r_\ell J_\ell(w) \leq \max_{k \in [K]} r_k J_k(w) < r_\ell J_\ell(w^{\text{EPO}}) \text{ for all } \ell \in [K]. \quad (10)$$

After cancelling $r_\ell > 0$ on both sides of (10), we conclude that w^{EPO} cannot be weakly Pareto optimal. Hence $w^{\text{EPO}} \notin \mathcal{P} \cap \mathcal{F}_r$, leading to a contradiction. $\square$

Proposition 1 provides a sufficient condition to ensure that a solution for the min-max problem (1) can be pursued by instead searching for a point on the (weak) Pareto frontier $\mathcal{P}$ that is also fair $\mathcal{F}_r$. This fact does not hold in general — see [16] for counter-examples. As long as $\mathcal{E}_r = \mathcal{P} \cap \mathcal{F}_r$ is non-empty, however, it follows that any exact Pareto optimal point is also min-max optimal. Motivated by this consideration, in the sequel we will propose a new algorithm for min-max optimization via exact Pareto optimization. Compared to existing algorithms in the literature, the proposed strategy will rely on a single-time scale, and exhibit a reduced per-iteration complexity of $O(Kd)$ as opposed to $O(K^2d)$ [8], [18], [19]. This results in better scaling with the number of objectives K .

III. EXACT PARETO OPTIMALITY VIA AUGMENTED LAGRANGIAN

In this section, we develop an algorithm for exact Pareto optimization via the augmented Lagrangian, inspired by classical primal-dual consensus techniques in multi-agent optimization and learning — see [20], [21] for early examples and [22] for a recent survey. To this end, note that an exact Pareto optimal solution can be pursued via:

$$\min_{w \in \mathbb{R}^d} \frac{1}{K} \sum_{k=1}^K J_k(w) \quad (11a)$$

$$\text{s.t. } r_k J_k(w) = r_\ell J_\ell(w) \quad \forall k, \ell. \quad (11b)$$

Here, the relationship (11a) encourages Pareto optimality, while (11b) ensures the fairness condition (5). In light of (11b), the objective (11a) can be replaced by any linear combination of $J_k(\cdot)$ having non-negative weights. The choice of equal weighting given by $\frac{1}{K}$ is merely a matter of simplicity. In analogy to [20]–[22], we replace the collection of constraints (11b) by a single constraint involving the aggregate constraint violations:

$$\min_{w \in \mathbb{R}^d} \frac{1}{K} \sum_{k=1}^K J_k(w) \quad (12a)$$

$$\text{s.t. } \frac{1}{2K} \sum_{k=1}^K \sum_{\ell=1}^K \|r_k J_k(w) - r_\ell J_\ell(w)\|^2 = 0. \quad (12b)$$

The fairness condition (12b) can be written more compactly as:

$$\frac{1}{K} \sum_{k=1}^K \sum_{\ell=1}^K \|r_k J_k(w) - r_\ell J_\ell(w)\|^2 = \mathcal{J}(w)^\top L_r \mathcal{J}(w) = 0, \quad (13)$$

where $\mathcal{J}(w) = [J_1(w), \dots, J_K(w)]^\top$ is a vector containing the collection of objectives evaluated for the model w and L_r is given by:

$$L_r = \text{diag}(r) \left(I_{K \times K} - \frac{1}{K} \mathbb{1}_K \mathbb{1}_K^\top \right) \text{diag}(r). \quad (14)$$

Here, $\text{diag}(r)$ is a diagonal matrix that contains the elements of r on its diagonal. Since L_r is symmetric and positive semi-definite, it has a square root $\sqrt{L_r}$ that satisfies $\sqrt{L_r} \sqrt{L_r} = L_r$ and hence (12b) is equivalent to:

$$\left\| \sqrt{L_r} \mathcal{J}(w) \right\|^2 = 0 \iff \sqrt{L_r} \mathcal{J}(w) = 0. \quad (15)$$

We can then define the corresponding augmented Lagrangian [23, Sec. 4] as

$$\mathcal{L}(w, \lambda) = \frac{1}{K} \mathbb{1}_K^\top \mathcal{J}(w) + \lambda^\top \sqrt{L_r} \mathcal{J}(w) + \frac{\eta}{2} \left\| \sqrt{L_r} \mathcal{J}(w) \right\|^2, \quad (16)$$

where $\eta > 0$ is a penalty parameter and λ is the corresponding Lagrangian multiplier. We then update the *primal* variable w and the *dual* variable λ in an iterative first-order approach as in [24], [25]:

$$w_i = w_{i-1} - \mu \nabla_w \mathcal{L}(w_{i-1}, \lambda_{i-1}) \quad (17a)$$

$$= w_{i-1} - \mu G(w_{i-1}) \left[\frac{1}{K} \mathbb{1}_K + \sqrt{L_r} \lambda_{i-1} + \eta L_r \mathcal{J}(w_{i-1}) \right]$$

$$\lambda_i = \lambda_{i-1} + \mu \nabla_\lambda \mathcal{L}(w_{i-1}, \lambda_{i-1}) = \lambda_{i-1} + \mu \sqrt{L_r} \mathcal{J}(w_{i-1}), \quad (17b)$$

where $G(w) = [\nabla J_1(w), \dots, \nabla J_K(w)]$ is a $d \times K$ matrix that collects the gradients from the K objectives evaluated for the model w and $\mu > 0$ is the step size. Multiplying (17b) by $\sqrt{L_r}$ from the left and by defining $p_i \triangleq (1/K) \mathbb{1}_K + \sqrt{L_r} \lambda_i$, we obtain the following equivalent formulation:

$$w_i = w_{i-1} - \mu G(w_{i-1}) [p_{i-1} + \eta L_r \mathcal{J}(w_{i-1})] \quad (18a)$$

$$p_i = p_{i-1} + \mu L_r \mathcal{J}(w_{i-1}). \quad (18b)$$

From the initialization $\lambda_0 = 0$, we find the initial condition $p_0 = (1/K) \mathbb{1}_K$. Lastly, by we apply the positivity operator $[\cdot]_+ = \max\{\cdot, 0\}$ element-wise to p_{i-1} in (18a), which yields Algorithm 1.

The per-iteration complexity of EPO-AL scales well with the number of objectives, as summarized in the following remark.

Remark 1 (Per-iteration computational complexity). *Each iteration of EPO-AL requires $O(Kd)$ computations, resulting from the evaluation of the $d \times K$ gradient matrix $G(w_{i-1})$ and multiplication with the $K \times 1$ vector $([p_{i-1}]_+ + \eta L_r \mathcal{J}(w_{i-1}))$.*

Note that typical multi-objective optimization algorithms [26], including the ones that aim for finding exact Pareto optimal solutions [8], [18], [19] generally involve the evaluation of $G(w_{i-1})^\top G(w_{i-1})$, which requires on the order of $O(K^2 d)$ computations; subgradient-based approaches (2) require $O(K)$ computations to identify the active objective, and $O(d)$ computations to evaluate the corresponding subgradient.

Algorithm 1: Exact Pareto Optimization via Augmented Lagrangian (EPO-AL)

Input: K positive, differentiable, objectives $J_1(\cdot), \dots, J_K(\cdot)$ with $J_k : \mathbb{R}^d \rightarrow \mathbb{R}_+$; step size $\mu > 0$; penalty parameter $\eta > 0$.

initialize $w_0 \in \mathbb{R}^d, p_0 = \mathbb{1}_K/K, z_t = \mathbf{0}_K$, where $\mathbf{0}_K$ is the all-zero vector of size K .

for $i = 1, 2, \dots$ **do**

$$w_i = w_{i-1} - \mu G(w_{i-1}) ([p_{i-1}]_+ + \eta L_r \mathcal{J}(w_{i-1})) \quad (19a)$$

$$p_i = p_{i-1} + \mu L_r \mathcal{J}(w_{i-1}) \quad (19b)$$

end

IV. FIXED POINT ANALYSIS

Before analyzing the fixed-point behavior of Algorithm 1, we first recall the notion of *Pareto stationarity* [27].

Definition 4 (Pareto stationarity). *A model is called Pareto stationary if one can find a convex combination of the gradients $\{\nabla J_k(w)\}_{k=1}^K$ that yields an all-zero vector. Hence, the collection of all Pareto stationary points $\mathcal{P}^{st}$ is given by:*

$$\mathcal{P}^{st} = \left\{ w \in \mathbb{R}^d : \min_{p \in \Delta^K} \|G(w)p\| = 0 \right\}. \quad (20)$$

Note that weak Pareto optimality implies Pareto stationarity, i.e., $\mathcal{P} \subseteq \mathcal{P}^{st}$ [28, Lemma 2.2]. We now characterize the fixed-point behavior of EPO-AL.

Theorem 1 (Fixed point analysis). *Assume that Algorithm 1 converges to a pair of fixed-points w_∞ and p_∞ . Then w_∞ is both Pareto stationary and fair.*

Proof. We begin the proof by substituting w_∞ for w_i and w_{i-1} as well as p_∞ for p_i and p_{i-1} in (19a)–(19b), which yields:

$$G(w_\infty) ([p_\infty]_+ + \eta L_r \mathcal{J}(w_\infty)) = 0; \quad (21)$$

$$L_r \mathcal{J}(w_\infty) = 0. \quad (22)$$

From (22), (21) simplifies to:

$$G(w_\infty) [p_\infty]_+ = \sum_{k=1}^K [p_{k,\infty}]_+ \nabla J_k(w_\infty) = 0, \quad (23)$$

which ensures that w_∞ is Pareto stationary, provided that $[p_\infty]_+$ contains at least one non-zero entry, which we will investigate further below. In light of (13), the second condition (22) implies that w_∞ yields $r_1 J_1(w_\infty) = \dots = r_K J_K(w_\infty)$, and hence $w_\infty \in \mathcal{F}_r$ satisfies the fairness condition.

The only remaining step is to show that p_∞ contains at least one strictly positive element. To this end, observe from (14) that $r^{-1} \triangleq [r_1^{-1}, \dots, r_K^{-1}]$ is in the nullspace of L_r :

$$\begin{aligned} L_r r^{-1} &= \text{diag}(r) \left(I_{K \times K} - \frac{1}{K} \mathbb{1}_K \mathbb{1}_K^\top \right) \text{diag}(r) r^{-1} \\ &= \text{diag}(r) \left(I_{K \times K} - \frac{1}{K} \mathbb{1}_K \mathbb{1}_K^\top \right) \mathbb{1}_K = 0. \end{aligned} \quad (24)$$

Hence, by taking the inner product of (19b) with r^{-1} , we have:

$$(r^{-1})^\top p_i = (r^{-1})^\top p_{i-1} + \mu(r^{-1})^\top L_r \mathcal{J}(w_{i-1}) = (r^{-1})^\top p_{i-1}. \quad (25)$$

Upon iterating all the way back to p_0 , we find that:

$$\sum_{k=1}^K \frac{p_{k,i}}{r_k} = \sum_{k=1}^K \frac{p_{k,0}}{r_k} = \sum_{k=1}^K \frac{1}{Kr_k} \triangleq \epsilon > 0. \quad (26)$$

For $\sum_{k=1}^K p_{k,i}/r_k$ to be greater than ϵ , there must exist at least one k' such that $p_{k',i}/r_{k'} \geq \epsilon$ and hence $p_{k',i} \geq \epsilon r_{k'} \geq \epsilon \min_k r_k$. We conclude that for all i we have:

$$\max_k p_{k,i} \geq \left(\min_k r_k \right) \left(\sum_{k=1}^K \frac{1}{Kr_k} \right). \quad (27)$$

Upon assuming that p_i approaches the fixed-point p_∞ and taking limits yields the desired result. $\square$

Corollary 1 (Convex objectives). *Assume that the objectives $J_k(w)$ are convex for all $k = 1, \dots, K$. Then, $w_\infty \in \mathcal{E}_r = \mathcal{P} \cap \mathcal{F}_r$ is exact Pareto optimal and solves the min-max problem (1).*

Proof. The result is immediate after recognizing that Pareto stationarity implies weak Pareto optimality [28, Lemma 2.2] for convex objectives. Hence, $w^\infty \in \mathcal{P}$. Theorem 1 already established that $w^\infty \in \mathcal{F}_r$. Under these conditions, Proposition 1 ensures that $w^\infty \in \mathcal{E}_r$ and also solves (1). $\square$

V. EMPIRICAL EVALUATION

We empirically evaluate our algorithm using a pair of synthetic experiments: when the objectives $\{J_k(w)\}_{k=1}^K$ are (i) all convex and are (ii) all non-convex¹. Specifically, for the convex scenario, we consider (i) $J_k(w) = \sqrt{1 + \|w - w_k\|^2} - 1$; for the non-convex scenario we take (ii) $J_k(w) = 1 - e^{-\|w - w_k\|^2}$ adapted from [9] to deal with more than two objectives. Specifically, the K anchor points $\{w_k\}_{k=1}^K$ are chosen uniformly at random on the unit $(d-1)$ -surface, and we also choose the preference vector r by sampling uniformly at random in the interior of the probability simplex Δ_+^K for which we define as $\Delta_+^K = \{y \in \Delta^K : y_k > 1/3K \forall k\}$. We impose such strict positivity to avoid extreme cases where some objectives are essentially ignored. We choose the initial model w_0 by randomly sampling from the unit $(d-1)$ -sphere. We account for these randomnesses by running 30 independent experiments, unless specified otherwise.

We compare the proposed algorithm to (i) the subgradient algorithm (2) where the active index k_i^{act} is chosen by breaking any tie at random; (ii) the smooth-max approach [5], [13], [15] that updates $w \leftarrow w - \mu \nabla \text{LSE}_\tau[r_1 J_1(w), \dots, r_K J_K(w)]$, where $\text{LSE}_\tau[v_1, \dots, v_K]$ is the smooth-max function defined as $\text{LSE}_\tau[v_1, \dots, v_K] = \log \sum_{k=1}^K e^{v_k/\tau}$; and EPO Search [8], which has the form of $w \leftarrow w - \mu G(w)\beta$ where $\beta \in \Delta^K$ is chosen by solving a K -dimensional linear program [8] at every iteration.

¹Code is available at https://github.com/sangwoo-p/EPO_AL

A. Visualization of the optimization trajectory

We first visualize the optimization trajectory of the schemes considered when all the objectives are non-convex in Fig. 1. We omit the smooth-max approach as it fails to converge to the optimal point [13], [14] for large enough τ that gives us a distinct optimization trajectory along with the subgradient algorithm (2). We set $d = 3$ and $K = 2$. An interesting observation here is that existing approaches prioritize converging to the fairness constraint before searching for the Pareto front, while the proposed EPO-AL algorithm rapidly converges to the Pareto front and then sweeps it for a solution that also satisfies the fairness condition. We set $\mu = 0.1$ for all the schemes along with $\eta = 10$ for EPO-AL, $r = [0.2, 0.8]^\top$, and choose the two anchors following [9].

B. Iteration/time complexity

We now consider the iteration complexity of the four algorithms considered by measuring the minimum number of iterations required to achieve a specified target accuracy. In order to fairly compare different algorithms, we set the step size μ for each algorithm separately by searching over $\mathcal{G}_\mu = [10^{-3}, 10^{-1}]$ in a log-scaled grid of size 10. As for the EPO-AL and smooth-max algorithm, we also set the penalty parameter η and temperature parameter τ by searching over $\mathcal{G}_\eta = [10^{-1}, 10^2]$ and $\mathcal{G}_\tau = [10^{-2}, 10]$ respectively, both in a log-scaled grid of size 10. We set the maximum number of iterations as 1000 and set the dimension of the model as $d = 100$, i.e., $w \in \mathbb{R}^{100}$.

Specifically, we define the target performance J^* as the minimum value attained by the subgradient algorithm (2) throughout all of the possible step size choices. We then evaluate the iteration complexity for a fixed choice of μ (for all the algorithms) as well as of η (for EPO-AL) and of τ (for smooth-max) by $i^o(\mu, \eta, \tau) = \min\{i : |\max_k r_k J_k(w_i) - J^*| \leq \epsilon\}$ with the tolerance level set to $\epsilon = 0.01$. We then finally evaluate the iteration complexity i^o for each scheme by choosing the minimum $i^o(\mu, \eta, \tau)$ among the possible choices of μ , η , and τ , i.e., $i^o = \min_{\mu \in \mathcal{G}_\mu} i^o(\mu, \eta, \tau)$ for the subgradient algorithm (2) and EPO Search; $i^o = \min_{(\mu, \tau) \in \mathcal{G}_\mu \times \mathcal{G}_\tau} i^o(\mu, \eta, \tau)$ for smooth-max; and $i^o = \min_{(\mu, \eta) \in \mathcal{G}_\mu \times \mathcal{G}_\eta} i^o(\mu, \eta, \tau)$ for EPO-AL.

Fig. 2 (left) shows the iteration complexity as a function of the number of objectives K for both convex and non-convex functions $J_k(w)$. It is observed that both the classical EPO Search [8] and the proposed EPO-AL algorithm scale well with the number of objectives, unlike the subgradient algorithm (2) that scales poorly upon increasing the number of objectives. Since the iteration count does not capture the computational complexity associated with each iteration (see Fig. 1), we next investigate the minimum *total* complexity required to reach the target performance J^* .

Fig. 2 (right) shows the wall-clock time complexity t^o , defined as the actual total time required to process the number of iterations i^o . The wall-clock time is evaluated on Apple M1 hardware. The fact that EPO Search [8] involves the solution

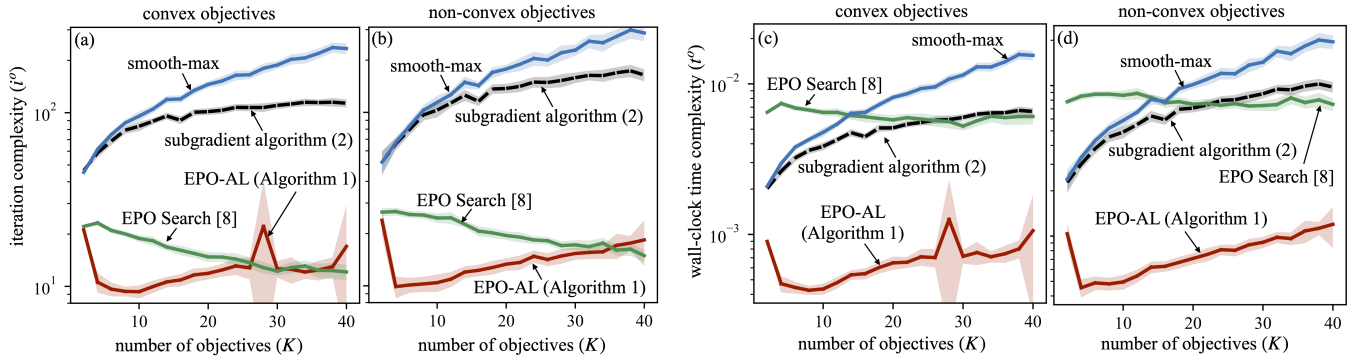


Fig. 2. Iteration complexity i^o (a,b) and wall-clock time complexity t^o (c,d) as a function of number of objectives K . The results are averaged over 30 independent experiments after removing the minimum and maximum, where each experiment assumes different preference vector r and different initial model w_0 . Shaded area corresponds to 99% confidence interval.

of a linear program at every iteration results in higher per-iteration complexity and hence higher total runtime compared to the proposed EPO-AL strategy, which only involves a single timescale.

VI. CONCLUSION

A new algorithm was proposed for min-max optimization via exact Pareto optimization. To derive the strategy we made use of primal-dual consensus techniques via the augmented Lagrangian, resulting in an algorithm which scales better with the number of objectives than a subgradient-based approach, while maintaining a lower per-iteration complexity than other smoothing-based algorithms. Experimental results showed that the proposed algorithm achieves the target performance by imposing lower total complexity as compared to the other benchmarks, demonstrating its scalability with the number of objectives.

REFERENCES

- [1] K. Miettinen, *Nonlinear multiobjective optimization*, vol. 12. Springer Science & Business Media, 1999.
- [2] Z. Fei, B. Li, S. Yang, C. Xing, H. Chen, and L. Hanzo, "A survey of multi-objective optimization in wireless sensor networks: Metrics, algorithms, and open problems," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 550–586, 2016.
- [3] O. Simeone, "A very brief introduction to machine learning with applications to communication systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 648–664, 2018.
- [4] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *ICML*, pp. 4615–4625, PMLR, 2019.
- [5] Q. Hou, Y. Cai, Q. Hu, M. Lee, and G. Yu, "Joint resource allocation and trajectory design for multi-UAV systems with moving users: Pointer network and unfolding," *IEEE Transactions on Wireless Communications*, vol. 22, no. 5, pp. 3310–3323, 2022.
- [6] Y. Wang, C. Yang, and M. Peng, "Hybrid precoding with low-resolution pss for wideband Terahertz communication systems in the face of beam squint," *arXiv preprint arXiv:2406.16303*, 2024.
- [7] S. M. Hamidi, A. Bereyhi, S. Asaad, and H. V. Poor, "Over-the-air fair federated learning via multi-objective optimization," *arXiv preprint arXiv:2501.03392*, 2025.
- [8] D. Mahapatra and V. Rajan, "Multi-task learning with user preferences: Gradient descent with controlled ascent in Pareto optimization," in *ICML*, pp. 6597–6607, PMLR, 2020.
- [9] X. Lin, H.-L. Zhen, Z. Li, Q.-F. Zhang, and S. Kwong, "Pareto multi-task learning," *NeurIPS*, vol. 32, 2019.
- [10] H. H. Bauschke, P. L. Combettes, H. H. Bauschke, and P. L. Combettes, *Correction to: convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2017.
- [11] C. Ras, M. Tam, and D. Ueda, "Identification of active subfunctions in finite-max minimisation via a smooth reformulation," *arXiv preprint arXiv:2404.10326*, 2024.
- [12] I. Zang, "A smoothing-out technique for min—max optimization," *Mathematical Programming*, vol. 19, pp. 61–77, 1980.
- [13] H. Gokcesu, K. Gokcesu, and S. S. Kozat, "Accelerating min-max optimization with application to minimal bounding sphere," *arXiv preprint arXiv:1905.12733*, 2019.
- [14] A. Epasto, M. Mahdian, V. Mirrokni, and E. Zampetakis, "Optimal approximation-smoothness tradeoffs for soft-max functions," *NeurIPS*, vol. 33, pp. 2651–2660, 2020.
- [15] X. Lin, X. Zhang, Z. Yang, F. Liu, Z. Wang, and Q. Zhang, "Smooth Tchebycheff scalarization for multi-objective optimization," *arXiv preprint arXiv:2402.19078*, 2024.
- [16] J. G. Lin, "On min-norm and min-max methods of multi-objective optimization," *Mathematical programming*, vol. 103, no. 1, pp. 1–33, 2005.
- [17] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," *arXiv preprint arXiv:1905.10497*, 2019.
- [18] D. Mahapatra and V. Rajan, "Exact Pareto optimal search for multi-task learning and multi-criteria decision-making," *arXiv preprint arXiv:2108.00597*, 2021.
- [19] M. Momma, C. Dong, and J. Liu, "A multi-objective/multi-task learning framework induced by Pareto stationarity," in *ICML*, pp. 15895–15907, PMLR, 2022.
- [20] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in ad hoc wns with noisy links —Part I: Distributed estimation of deterministic signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350–364, 2008.
- [21] Z. J. Towfic and A. H. Sayed, "Stability and performance limits of adaptive primal-dual networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2888–2903, 2015.
- [22] S. Vlaski, S. Kar, A. H. Sayed, and J. M. Moura, "Networked signal and information processing: Learning by multiagent systems," *IEEE Signal Processing Magazine*, vol. 40, no. 5, pp. 92–105, 2023.
- [23] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.
- [24] K. J. Arrow, L. Hurwicz, H. Uzawa, H. B. Chenery, S. Johnson, and S. Karlin, *Studies in linear and non-linear programming*, vol. 2. Stanford University Press Stanford, 1958.
- [25] S. A. Alghunaim and A. H. Sayed, "Linear convergence of primal-dual gradient methods and their performance in distributed optimization," *Automatica*, vol. 117, p. 109003, 2020.
- [26] L. Chen, H. Fernando, Y. Ying, and T. Chen, "Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance," *NeurIPS*, vol. 36, pp. 70045–70093, 2023.
- [27] J.-A. Désidéri, "Multiple-gradient descent algorithm (MGDA) for multi-objective optimization," *Comptes Rendus Mathématique*, vol. 350, no. 5–6, pp. 313–318, 2012.
- [28] H. Tanabe, E. H. Fukuda, and N. Yamashita, "Proximal gradient methods for multiobjective optimization and their applications," *Computational Optimization and Applications*, vol. 72, pp. 339–361, 2019.