Direct-Search for a Class of Stochastic Min-Max Problems

Sotiris Anagnostidis
Institute for Machine Learning
ETH Zürich, Switzerland
sotirios.anagnostidis@inf.ethz.ch

Aurelien Lucchi
Institute for Machine Learning
ETH Zürich, Switzerland
aurelien.lucchi@inf.ethz.ch

Youssef Diouane
ISAE-SUPAERO
Université de Toulouse, France
youssef.diouane@isae-supaero.fr

Abstract

Recent applications in machine learning have renewed the interest of the community in min-max optimization problems. While gradient-based optimization methods are widely used to solve such problems, there are however many scenarios where these techniques are not well-suited, or even not applicable when the gradient is not accessible. We investigate the use of direct-search methods that belong to a class of derivative-free techniques that only access the objective function through an oracle. In this work, we design a novel algorithm in the context of min-max saddle point games where one sequentially updates the min and the max player. We prove convergence of this algorithm under mild assumptions, where the objective of the max-player satisfies the Polyak-Lojasiewicz (PL) condition, while the min-player is characterized by a nonconvex objective. Our method only assumes dynamically adjusted accurate estimates of the oracle with a fixed probability. To the best of our knowledge, our analysis is the first one to address the convergence of a direct-search method for min-max objectives in a stochastic setting.

1 INTRODUCTION

Recent applications in the field of machine learning, including generative models (Goodfellow et al., 2014) or robust optimization (Ben-Tal et al., 2009), have triggered significant interest for the optimization of min-max functions of the form

$$\min_{x \in X} \max_{y \in Y} f(x, y) = \mathbb{E}[f(x, y, \xi)],$$

where $\xi$ is a random variable characterized by some distribution. In machine learning, $\xi$ is for instance often drawn from a distribution that depends on the training data.

In practice, min-max problems are often solved using gradient-based algorithms, especially simultaneous gradient descent ascent (GDA) that simply alternates between a gradient descent step for $x$ and a gradient ascent step for $y$. While these algorithms are attractive due to their simplicity, there are however cases where the gradient of the objective function is not accessible, such as when modelling distributions with categorical variables (Jang et al., 2016), tuning hyper-parameters (Audet and Orban, 2006; Marzat et al., 2011) and multi-agent reinforcement learning with bandit feedback (Zhang et al., 2019). A resurgence of interest has recently emerged for applications in black-box optimization (Bogunovic et al., 2018; Liu et al., 2019) and black-box poisoning attack (Liu et al., 2020), where an attacker deliberately modifies the training data in order to tamper with the model’s predictions. This can be formulated as a min-max optimization problem, where only stochastic accesses to the objective function are available (Wang et al., 2020).

In this work, we investigate the use of direct-search methods to optimize min-max objective functions without requiring access to the gradients of the objective $f$. Direct-search methods have a long history in the field of optimization, dating back from the seminal paper of Hooke and Jeeves (1961). The appeal of these methods is due to their simplicity but also potential ability to deal with non-trivial objective functions. Although there are variations among random search techniques, most of them can be summarized conceptually as sampling random directions from a search space and moving towards directions that decrease the objective function value. We note that these

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).
techniques are sometimes named derivative-free methods, but it is important to distinguish them from other techniques that try to estimate derivatives based on finite difference (Spall, 2003) or smoothing (Nesterov and Spokoiny, 2017). We refer the reader to the surveys by Lewis et al. (2000); Rios and Sahinidis (2013) for a comprehensive review of direct-search methods.

Solving the saddle point problem (1) is equivalent to finding a saddle point\(^1\) \((x^*, y^*)\) such that

\[
 f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*) \quad \forall x \in X, \quad \forall y \in Y.
\]

There is a rich literature on saddle point optimization for the particular class of convex-concave functions (i.e. when \(f\) is convex in \(x\) and concave in \(y\)) that are differentiable. Although this type of objective function is commonly encountered in applications such as constrained convex minimization, many saddle point problems of interest do not satisfy the convex-concave assumption. This for instance includes applications such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), robust optimization (Ben-Tal et al., 2009; Bogunovic et al., 2018) and multi-agent reinforcement learning (Omidshafiei et al., 2017). For min-max problems without access to derivatives, the literature is in fact very scarce. Most existing techniques such as Hare and Macklem (2013); Hare and Nutini (2013); Custódio et al. (2021) consider finite-max functions, i.e., functions of the form \(f(x) = \max\{f_i(x) : i = 1, 2, \ldots, N\}\) where \(N > 0\) is finite and each \(f_i\) is continuously differentiable. Other techniques such as Bertsimas and Nohadani (2010); Bertsimas et al. (2010) are restricted to functions \(f\) that are convex with respect to \(x\) or only provide asymptotic convergence analysis (Menickelly and Wild, 2020). We refer the reader to Section 2 for a more detailed discussion of prior approaches.

Motivated by a wide range of applications, we therefore focus on a nonconvex and nonconcave stochastic setting where the max player satisfies the PL condition (see Definition 5 in Section 4), which is known to be a weaker assumption compared to convexity (Karimi and Schmidt, 2015). In summary, our main contributions are:

- We design a novel direct-search algorithm for such min-max problems and provide non-asymptotic convergence guarantees in terms of first-order Nash equilibrium. Concretely, we prove convergence to an \(\varepsilon\)-first-order Nash Equilibrium (for a definition see Section 3) in \(O(\varepsilon^{-2} \log(\varepsilon^{-1}))\) iterations, which is comparable to the rate achieved by gradient-based techniques (Nouiehed et al., 2019).
- We derive theoretical convergence guarantees in a stochastic setting where one only has access to accurate estimates of the objective function, with some fixed probability. We prove our results for the case where the min player optimizes a non-convex function while the max player optimizes a PL function.
- We validate empirically our theoretical findings, including settings where derivatives are not available.

2 RELATED WORK

Direct-search methods for minimization problems The general principle behind direct-search methods is to optimize a function \(f(x)\) without having access to its gradient \(\nabla f(x)\). There is a large number of algorithms that are part of this broad family including golden-section search techniques or random search (Rastrigin, 1963). Among the most popular algorithms in machine learning are evolution strategies and population-based algorithms that have demonstrated promising results in reinforcement learning (Salimans et al., 2017; Maheswaranathan et al., 2018) and bandit optimization (Flaxman et al., 2004).

At a high-level, these techniques work by maintaining a distribution over parameters and duplicate the individuals in the population with higher fitness. Often these algorithms are initialized at a random point and then adapt their search space, depending on which area contains the best samples (i.e. the lowest function value when minimizing \(f(x)\)). New samples are then generated from the best regions in a process repeated until convergence. The most well-known algorithms that belong to this class are evolutionary-like algorithms, including for instance CMA-ES (Hansen et al., 2003). Evolutionary strategies have recently been shown to be able to solve various complex tasks in reinforcement learning such as Atari games or robotic control problems, see e.g. Salimans et al. (2017). Their advantages in the context of reinforcement learning are their reduced sensitivity to noisy or uninformative gradients (potentially increasing their ability to avoid local minima (Conti et al., 2017)) and the ease with which one can implement a distributed or parallel version.

Convergence guarantees for direct-search methods Proofs of convergence for direct-search methods are based on a specific construction for the sampling directions, often that they positively span the whole search space (Conn et al., 2009), or that they are dense in certain types of directions (known as refining directions) at the limit point (Audet and Dennis Jr, 2006). In addition, they also typically

---

\(^1\)In the game theory literature, such point is commonly referred to as (global) Nash equilibrium, see e.g. Liang and Stokes (2018).
rely on the use of a forcing function that imposes each new selected iterate to decrease the function value adequately. This technique has been analyzed in Vicente (2013) who proved convergence under mild assumptions in $O(\epsilon^{-2})$ iterations for the goal $\|\nabla f(x)\| < \epsilon$. The number of required steps is reduced to $O(\epsilon^{-1})$ for convex functions $f$, and to $O(\log(\epsilon^{-1}))$ for strongly-convex functions (Konevčný and Richtárik, 2014). This is on par with the steepest descent method for unconstrained optimization (Nesterov, 2013) apart from some constants that depend on the dimensionality of the problem.

**Stochastic estimates of the function** In our analysis, we only assume access to stochastic estimates of the objective function.

$$f(x, y) = \mathbb{E}[\tilde{f}(x, y, \xi)],$$

where $\xi$ is a random variable that captures the randomness of the objective function. The origin of the noise could be privacy related, or caused by a noisy adversary. Most commonly, it might arise from online streaming data, distributed and batch-sized updates due to the sheer size of the problem. Stochastic gradient descent is often used to optimize Eq. (2), where one often assumes access to accurate estimates of $f$ and consider updates only in expectation (Johnson and Zhang, 2013). To establish similar convergence rates to the deterministic case, an alternative solution consists of adapting the accuracy of these estimates dynamically, which can be ensured by averaging multiple samples together. This approach has for instance been analyzed in the context of trust-region methods (Blanchet et al., 2019) and line-search methods (Paquette and Scheinberg, 2018; Bergou et al., 2018), including direct-search for the minimization of nonconvex functions (Dzahini, 2020).

**Algorithms for finding equilibria in games** Since the pioneering work of von Neumann (1928), equilibria in games have received great attention. Most past results focus on convex-concave settings (Chen et al., 2014; Hien et al., 2017). Notably, Cherukuri et al. (2017) studied convergence of the GDA algorithm under strictly convex-concave assumptions. For problems where the function does not satisfy this condition however, convergence to a saddle point is not guaranteed. More recent results focus on relaxing these conditions. The work of Nouiehed et al. (2019) analyzed gradient descent-ascent under a similar scenario, where the objective of the max player satisfies the PL condition and where the min player optimizes a nonconvex objective. Ostrovskii et al. (2020); Wang et al. (2020) analyze a nonconvex-concave class of problems, while Lin et al. (2020) present a two-scale variant of the GDA algorithm for a similar scenario, providing a replacement for the alternating updates scheme.

We take inspiration from the work of Liu et al. (2019); Nouiehed et al. (2019); Sanjabi et al. (2018) to design a novel alternating direct-search algorithm, where the inner maximization problem is solved almost exactly before performing a single step towards improving the strategy of the minimization player. We are able to prove convergence of our direct-search algorithm under this procedure, which has been proven to be more stable than the analogous simultaneous one, as rigorously shown in Gidel et al. (2018) and Zhang and Yu (2019) for a variety of algorithms.

### 3 PRELIMINARIES

Throughout, we use $\|\cdot\|$ to denote the Euclidean norm; that is, for $x \in \mathbb{R}^n$ we have $\|x\| = \sqrt{x^\top x}$.

#### 3.1 Min-Max Games

We consider the optimization problem defined in Eq. (1) for which a common notion of optimality is the concept of Nash equilibrium as mentioned previously, which is formally defined as follows.

**Definition 1.** We say that a point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is a Nash equilibrium of the game if

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*) \quad \forall x \in \mathcal{X}, \quad \forall y \in \mathcal{Y}.$$  

A Nash equilibrium is a point where the change of strategy of each player individually does not lead to an improvement from her viewpoint. Such a Nash equilibrium point always exists for convex-concave games (Jin et al., 2019), but not necessarily for nonconvex-nonconcave games. Even when they exist, finding Nash equilibria is known to be an NP-hard problem, which has led to the introduction of local characterizations as discussed in Jin et al. (2019); Adolphe et al. (2018). Here we use the notion of a first-order Nash equilibrium (FNE) (for a definition we refer to Pang and Razaviyayn (2016)). We focus on the problem of converging to such a FNE point, or an approximate FNE defined as follows (adapted from Nouiehed et al. (2019) in the absence of constraints).

**Definition 2.** For a function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, a point $(x^*, y^*) \in \mathbb{R}^n \times \mathbb{R}^m$ is said to be an $\epsilon$-first-order Nash Equilibrium ($\epsilon$-FNE) if: $\|\nabla_x f(x^*, y^*)\| \leq \epsilon$ and $\|\nabla_y f(x^*, y^*)\| \leq \epsilon$.

#### 3.2 Direct-Search Methods

**Spanning set** Direct-search methods typically rely on the smoothness of the objective function, which we
denote by \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) in this section, and on appropriate choice of sampling points to prove convergence. The key idea to guarantee convergence is that one of the sampled directions will form an acute angle with the negative gradient. This can be ensured by sampling from a Positive Spanning Set (PSS). The quality of a spanning set \( \mathcal{D} \) is typically measured using a notion of cosine measure defined as

\[
\kappa(\mathcal{D}) = \min_{\theta \neq u \in \mathbb{R}^n} \max_{d \in \mathcal{D}} \frac{u^T d}{\|u\| \|d\|}.
\]

In the following, we will consider positive spanning sets such that \( \kappa(\mathcal{D}) \geq \kappa_{\text{min}} > 0 \) and \( d_{\text{min}} \leq \|d\| \leq d_{\text{max}}, \forall d \in \mathcal{D} \). These assumptions require \( |\mathcal{D}| \geq n + 1 \). Common choices are i) the positive and negative orthonormal bases \( \mathcal{D} = [I_n - I_n] = \{e_1, \ldots, e_n, -e_1, \ldots, -e_n\} \) of size \( |\mathcal{D}| = 2n \), ii) a minimal positive basis with uniform angles of size \( |\mathcal{D}| = n + 1 \) (see Corollary 2.6 of Conn et al. (2009) and Kolda et al. (2003)) or iii) even rotations of these matrices (Gratton et al., 2016).

### Algorithm 1: Direct-search \((f, x_0, c, T)\)

**Input:** \( f \): objective function, with \( f_k \) it’s estimate at step \( k \)
- \( c \): forcing function constant
- \( T \): number of steps

Initialize step size value \( \sigma_0 \). Choose \( \gamma > 1 \).

Create the Positive Spanning Set \( \mathcal{D} \).

for \( k = 0, \ldots, T - 1 \) do

1. **Offspring generation:**
   Generate the points
   \[
   x'_{k+i} = x_k + \sigma_k d^i, \quad \forall d^i \in \mathcal{D}.
   \]

2. **Parent Selection:**
   Choose \( x' = \arg \min_i f_k(x') \).

3. **Sufficient Decrease:**
   if \( f_k(x') < f_k(x_k) - \rho(\sigma_k) \) then
   (Iteration is successful)
   Update and increase step size
   \[
   x_{k+1} = x', \quad \sigma_{k+1} = \min\{\sigma_{\text{max}}, \gamma \sigma_k\}.
   \]
   else
   (Iteration is unsuccessful)
   Decrease step size
   \[
   x_{k+1} = x_k, \quad \sigma_{k+1} = \gamma^{-1} \sigma_k.
   \]
   end

end

return \( x_T \)

---

**Forcing function** Another critical component to guarantee that the function value decreases at each step appropriately is a forcing function \( \rho \) that satisfies \( \rho(\sigma) \rightarrow 0 \) when \( \sigma \rightarrow 0 \). Given such \( \sigma \), direct-search methods sample new points according to the rule

\[
x' = x + \sigma d,
\]

and accept points for which

\[
f(x') < f(x) - \rho(\sigma),
\]

\( \sigma \in \mathcal{D} \). If the previous condition holds for some \( \sigma \in \mathcal{D} \), then the new point is accepted, the step is deemed successful and the \( \sigma \) parameter is increased, otherwise \( \sigma \) is decreased and the above process is repeated. We use a parameter \( \gamma \) to indicate these updates of the step size. For convenience and without loss of generality, we will only consider spanning sets with vectors of unitary length \( d_{\text{min}} = d_{\text{max}} = 1 \) and a forcing function

\[
\rho(\sigma) = c\sigma^2.
\]

The direct-search scheme is displayed in Algorithm 1.

### 4 STOCHASTIC DIRECT-SEARCH

The full algorithm we analyze to solve the min-max objective is presented in Algorithm 2. It consists of two steps: i) first solve the maximization problem w.r.t. the \( y \) variable using Algorithm 1, and ii) perform one update step for the \( x \) variable. In this section, we first analyze the convergence properties of Algorithm 1 in the setting where we only have access to estimates of the objective function \( f \),

\[
f(x) = \mathbb{E}[f(x, \xi)].
\]

Let \( (\Omega, \mathcal{F}, P) \) be a probability space with elementary events denoted with \( \omega \). We denote the random quantities for the iterate by \( x_k = X_k(\omega) \) and for the step size by \( \sigma_k = \Sigma_k(\omega) \). Similarly let \( \{F^0_k, F^\xi_k\} \) be the estimates of \( f(X_k) \) and \( f(X_k + \Sigma d_k) \), for each \( d_k \) in a set \( \mathcal{D} \), with their realizations \( f^0_k = F^0_k(\omega) \), \( f^\xi_k = F^\xi_k(\omega) \). At each iteration the influence of the noise on function evaluations is random. We will assume that, when conditioned on all the past iterates, these estimates are sufficiently accurate with a sufficiently high probability. We formalize this concept in the two definitions below.

**Definition 3.** \((\epsilon_f\text{-accurate})\) The estimates \( \{F^0_k, F^\xi_k\} \) are said to be \( \epsilon_f \)-accurate with respect to the corresponding sequence if

\[
|F^0_k - f(X_k)| \leq \epsilon_f \Sigma^2_k \quad \text{and} \quad |F^\xi_k - f(X_k + \Sigma d_k)| \leq \epsilon_f \Sigma^2_k.
\]

**Definition 4.** \((p_f\text{-probabilistically } \epsilon_f\text{-accurate})\) The estimates \( \{F^0_k, F^\xi_k\} \) are said to be \( p_f \)-probabilistically
\( \epsilon_f \)-accurate with respect to the corresponding sequence if the events

\[
J_k = \{ \text{The estimates } \{ F_k^0, F_k^g \} \text{ are } \epsilon_f \text{-accurate} \}
\]
satisfy the condition\(^3\)

\[
P(J_k | F_{k-1}) = \mathbb{E}[1_{J_k} | F_{k-1}] \geq p_f,
\]

where \( F_{k-1} \) is the sigma-algebra generated by the sequence \( \{ F_k^0, F_k^g, \ldots, F_{k-1}^0, F_{k-1}^g \} \).

As the step size \( \sigma \) gets smaller, meaning that we are getting closer to the optimum, we require the accuracy over the function values to increase. However, the probability to encounter a good estimation remains the same throughout. A significant challenge arises, as steps may satisfy our sufficient decrease condition specified in Eq. (5) falsely, leading to a potential increase in terms of the objective value. This increase can potentially be very large, leading to divergence, and we therefore need to require an additional assumption regarding the variance of the error.

**Assumption 1.** The sequence of estimates \( \{ F_k^0, F_k^g \} \) are said to satisfy a \( l_f \)-variance condition if for all \( k \geq 0 \)

\[
\mathbb{E}[|F_k^0 - f(X_k)|^2 | F_{k-1}] \leq \ell_f^2 \sigma_k^4,
\]

\[
\mathbb{E}[|F_k^g - f(X_k + \Sigma_k d_k)|^2 | F_{k-1}] \leq \ell_f^2 \sigma_k^4.
\]

Based on the above assumptions, we reach the following conclusion regarding inaccurate steps (similar to Lemma 2.5 in Paquette and Scheinberg (2018)).

**Lemma 1.** Let Assumption 1 hold for \( p_f \)-probabilistically \( \epsilon_f \)-accurate estimates of a function. Then for \( k \geq 0 \) we have

\[
\mathbb{E}[1_{J_k} | F_k^0 - f(X_k)| | F_{k-1}] \leq (1 - p_f)^{1/2} l_f \sigma_k^2,
\]

\[
\mathbb{E}[1_{J_k} | F_k^g - f(X_k + \Sigma_k d_k)| | F_{k-1}] \leq (1 - p_f)^{1/2} l_f \sigma_k^2.
\]

**Computing the estimates** In order to satisfy Assumption 1 we can perform multiple function evaluations and average them out (see for instance Tropp (2015)). We therefore get an estimate \( F_k^0 = \frac{1}{|S_k^0|} \sum_{\xi_i \in S_k^0} \hat{f}(X_k, \xi_i) \), where \( S_k^0, S_k^g \) correspond to independent samples for \( F_k^0 \) and \( F_k^g \) respectively. Assuming bounded variance, i.e. \( \mathbb{E}[|\hat{f}(x, \xi) - f(x)|^2] \leq \sigma_f^2 \), known concentration results (see e.g. Tripuraneni et al. (2018); Chen et al. (2018)) guarantee that we can obtain \( p_f \)-probabilistically \( \epsilon_f \)-accurate estimates for

\[
|S_k^0| \geq \mathcal{O}(1) \left( \frac{\sigma_f^2}{\epsilon_f 2 \sigma_k^2} \log \left( \frac{1}{1 - p_f} \right) \right)
\]

number of evaluations (the same result holds for \( S_k^g \)). To also satisfy Assumption 1, we additionally require

\[
|S_k^0| \geq \frac{\sigma_f^2}{\ell_f 2 \sigma_k^2}.
\]

### 4.1 Convergence of Stochastic Direct-Search

In order to study the convergence properties of Algorithm 1, we introduce the following (random) Lyapunov function:

\[
\Phi_k = v(f(X_k) - f^\star) + (1 - v) \Sigma_k^2,
\]

where \( v \in (0, 1) \) is a constant. We denote by \( f^\star \) the minimum of the function \( f \), assumed to exist and potentially achieved at multiple positions. The Lyapunov function \( \Phi_k \) will be used to track the progress of the gradient norm \( \|\nabla f(X_k)\| \), which will serve as a measure of convergence.

Theorem 2 presented below ensures that the Lyapunov function decreases over iterations. Using this result, one can guarantee that the sequence of step-sizes decreases and then exploit the fact that for sufficiently small step sizes (and accurate estimates), the steps are successful, i.e. they decrease the objective function. The proof of the next Theorem is mainly inspired by Dzahini (2020); Audet et al. (2021).

**Theorem 2.** Let a function \( f \) with a minimum value \( f^\star \), with Lipschitz continuous gradients with a constant \( L \). Let also \( f \) be \( p_f \)-probabilistically \( \epsilon_f \)-accurate, while also having bounded noise variance according to Assumption 1 with constant \( l_f \). Then:

\[
\mathbb{E}[(\Phi_{k+1} - \Phi_k) | F_{k-1}] \leq -p_f(1 - v)(1 - \frac{1}{2}) \Sigma_k^2.
\]

The constants \( c, v \) and \( p_f \) should satisfy

\[
c - 2 \epsilon_f > 0, \quad \frac{p_f}{\sqrt{1 - p_f}} \geq \frac{4 v l_f}{(1 - v)(1 - \gamma^{-2})},
\]

\[
\frac{v}{1 - v} \geq \frac{1}{c - 2 \epsilon_f} (\gamma^2 - \frac{1}{\gamma^2}).
\]

Next, we characterize the number of steps required to converge by using a renewal-reward process adapted from Blanchet et al. (2019). Let us define the random process \( \{ \Phi_k, \Sigma_k \} \), with \( \Phi_k \geq 0 \) and \( \Sigma_k \geq 0 \). Let us also denote with \( W_k \) a random walk process and \( F_k \) the
σ-algebra generated by \( \{\Phi_0, \Sigma_0, W_0, \ldots, \Phi_k, \Sigma_k, W_k\} \) with \( W_0 = 1 \),

\[
P(W_{k+1} = 1 \mid \mathcal{F}_k) = p, \\
P(W_{k+1} = -1 \mid \mathcal{F}_k) = 1 - p. \tag{7}
\]

We also define a family of stopping times \( \{T_\epsilon\}_{\epsilon > 0} \) with respect to \( \{\mathcal{F}_k\}_{k \geq 0} \) for \( \epsilon > 0 \).

**Assumption 2.** Given the random quantities \( \{\Phi_k, \Sigma_k, W_k\} \), we make the following assumptions.

i. There exists \( \lambda > 0 \) such that \( \Sigma_{\max} = \Sigma_0 e^{\lambda_{\max}} \) for \( j_{\max} \in \mathbb{Z} \), and \( \Sigma_k \leq \Sigma_{\max} \) for all \( k \).

ii. There exists \( \Sigma_\epsilon = \Sigma_0 e^{\lambda_\epsilon} \), with \( j_\epsilon \in \mathbb{Z} \), such that

\[
1_{T_\epsilon > k} \Sigma_{k+1} \geq 1_{T_\epsilon > k} \min \{\Sigma_k e^{AW_{k+1}}, \Sigma_\epsilon\}
\]

where \( W_{k+1} \) satisfies Equation (7) with probability \( p > \frac{1}{2} \).

iii. There exists a nondecreasing function \( h(\cdot) : [0, \infty) \rightarrow (0, \infty) \) and a constant \( \Theta > 0 \) such that

\[
1_{T_\epsilon > k} \mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] \leq 1_{T_\epsilon > k} (\Phi_k - \Theta h(\Sigma_k)).
\]

Assumption 2 (ii) requires that step sizes tend to increase when below a specific threshold, while Assumption 2 (iii) requires that the random function \( \Phi \) decreases in expectation (already proved in Theorem 2). Under this assumption, the following results hold for the stopping time \( T_\epsilon \) (Blanchet et al., 2019).

**Theorem 3.** Under Assumption 2, we have

\[
\mathbb{E}[T_\epsilon] \leq \frac{p}{2p - 1} \frac{\Phi_0}{\Theta h(\Sigma_\epsilon)} + 1.
\]

In this analysis, our goal is to show that the norm of the gradient decreases below a threshold

\[
T_\epsilon = \inf\{k \geq 0 : \|\nabla f(X_k)\| \leq \epsilon\}.
\]

We assume that Assumption 2 (i) holds by the choice of \( \Sigma_{\max} \). We also know from Lemma 4 that for \( \|\nabla f(X)\| > \epsilon \) and \( \Sigma \leq C\epsilon \) then a successful step occurs, provided that estimates are accurate. Then following Lemma 4.10 from Paquette and Scheinberg (2018) we get that Assumption 2 (ii) also holds, for \( \Sigma_\epsilon = C\epsilon \). Based on the results of Theorem 3 and Lemma 4, we can now prove convergence for a non-convex bounded function.

**Theorem 5.** Assume that the Assumptions of Theorem 2 hold additionally \( p_f > \frac{1}{2} \). Then to get \( \|\nabla f(X_k)\| \leq \epsilon \), the expected stopping time of Algorithm 1 is

\[
\mathbb{E}[T_\epsilon] \leq O\left(1 + \frac{K_{\min}^2}{2p_f - 1} (f(\mathcal{X}_0) - f^* + \Sigma^2_0) (L + c + \epsilon_f)^2 \right) + 1.
\]

Note that for the deterministic scenario where \( \epsilon_f = 0 \), the above bound matches known results of direct-search in the nonconvex case (Vicente, 2013; Konvex and Richtárik, 2014). We now establish faster convergence for a function \( f \), additionally satisfying the PL condition, defined below.

**Definition 5.** (Polyak-Lojasiewicz Condition). A differentiable function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) with the minimum value \( f^* = \min_{x \in \mathbb{R}^n} f(x) \) is said to be \( \mu \)-Polyak-Lojasiewicz (\( \mu \)-PL) if:

\[
\frac{1}{2} \|\nabla f(x)\|^2 \geq \mu(f(x) - f^*).
\]

The PL condition is the weakest among a large family of function classes that include convex functions and other nonconvex ones (Karimi and Schmidt, 2015). Again we can guarantee convergence that closely matches results for deterministic direct-search under strong convexity, by proving that the number of iterations required to halve the distance to the optimum objective value is constant in terms of the accuracy \( \epsilon \).

**Theorem 6.** Let a function \( f \) with a minimum value \( f^* \) and satisfying the PL condition with a constant \( \mu \) and Lipschitz continuous gradients with a constant \( L \). Let also \( f \) be \( p_f \)-probabilistically \( \epsilon_f \)-accurate, while also having bounded noise variance according to Assumption 1 with constant \( l_f \). Then to get \( \|\nabla f(X_k)\| \leq \epsilon \), the expected stopping time of
Algorithm 1 is

\[ E[T_f] \leq O(1) \frac{\kappa_{\min}^2 (c + L)^2}{(2p_f - 1)\mu} \left( 1 + \frac{1}{c} \right) \log \left( \frac{L(f(X_0) - f^*)}{\epsilon} \right). \]  

(9)

The constants \(c, v\) and \(p_f > \frac{1}{2}\) should satisfy

\[ c > \max\{4\kappa_f, 2\sqrt{2}f\}, \quad \frac{p_f}{\sqrt{1 - p_f}} \geq \frac{4v_f}{(1 - v)(1 - \gamma^2)} \]

and \(\frac{v}{1 - v} \geq \max\left\{ \frac{1}{c - 2\kappa_f} (\gamma^2 - \frac{1}{\gamma^2}), \frac{72\gamma^2}{c} \right\} \).

5 ALGORITHM & CONVERGENCE GUARANTEES

We now focus on the min-max problem presented in Eq. (1). To proceed, we make the following standard assumptions regarding the smoothness of \(f\).

**Assumption 3.** The function \(f\) is continuously differentiable in both \(x\) and \(y\) and there exist constants \(L_{11}, L_{12}, L_{21}\) and \(L_{22}\) such that for every \(x, x_1, x_2 \in \mathcal{X}\) and \(y, y_1, y_2 \in \mathcal{Y}\)

\[
\|\nabla_x f(x, y) - \nabla_x f(x', y)\| \leq L_{11}\|x - x'\|, \\
\|\nabla_x f(x, y_1) - \nabla_x f(x, y_2)\| \leq L_{21}\|y_1 - y_2\|, \\
\|\nabla_y f(x, y_1) - \nabla_y f(x, y_2)\| \leq L_{22}\|y_1 - y_2\|.
\]

We require that the objective of the max-player satisfies the PL condition.

**Assumption 4.** There exists a constant \(\mu > 0\) such that the function \(-f(x, y)\) in problem (1) is \(\mu\)-PL for any \(x \in \mathcal{X}\).

Following prior works on PL games, e.g., Nouiehed et al. (2019), we propose a sequential scheme for the updates of the two players presented in Algorithm 2 (for simplicity some of the algorithm’s constants are not depicted). This multi-step algorithm solves the maximization problem up to some accuracy, and it then performs a single (successful) Direct-Search (DR) step for the minimization problem (see Algorithm 3).

We formalize our Assumptions and our final result.

**Assumption 5.** The function \(f\) is defined on the whole domain \(\mathcal{X} \times \mathcal{Y} = \mathbb{R}^{|\mathcal{X}|} \times \mathbb{R}^{|\mathcal{Y}|}\). We also require \(f\) to be bounded below for every \(y \in \mathcal{Y}\) and bounded above for every \(x \in \mathcal{X}\).

Algorithm 2: Min-Max-Direct-search

**Input:** \(f\): objective function 
\((x_0, y_0)\): initial point
\(\sigma_0\): initial step for the min problem

for \(t = 1, \ldots, T\) do

\[ y_t = \text{Direct-search}(-f(x_{t-1}, y_{t-1}), y_{t-1}) \]

\[ x_t, \sigma_t = \text{One-Step-Direct-search} \]

end

return \((x_T, x_T)\).

**Theorem 7.** Suppose that the objective function \(f(x, y)\) satisfies Assumptions 3, 4 and 5. If the estimates are deterministic, then Algorithm 2 converges to an \(\epsilon\)-FNE within \(O(\epsilon^{-2} \log(\epsilon^{-1}))\) steps. When \(f(x, y)\) is \(\epsilon_x\)-accurate with probability \(p_x\) for every \(x\) satisfying assumptions of Theorem 5 and \(\epsilon_y\)-accurate with probability \(p_y\) for every \(y\), satisfying assumptions of Theorem 6, then with a probability at least \(\delta\), Algorithm 2 converges and the expected number of steps to converge to reach an \(\epsilon\)-FNE is

\[
O\left( \frac{1}{(2p_x - 1)(2p_y - 1)} \epsilon^{-2} \left( \log(\epsilon^{-1}) + \left[ \log \left( \frac{1 - p_x}{p_x} \right) \right]^{-1} \log \left( 1 - \epsilon^{-\frac{2}{(2p_x - 1)(2p_y - 1) \log \delta}} \right) \right) \right).
\]

Algorithm 2 performs in total \(O(\epsilon^{-2})\) updates for the minimization problem, and each minimization update requires \(O(\log(\epsilon^{-1}))\) updates for the maximization problem. The proof of Theorem 7 consists in showing that the maximization problem is solved with sufficient accuracy, for which we invoke the result of Theorem 6. We then proceed by showing that iteratively solving the minimization problem allows us to converge in terms of the min-max objective, which is done using the result of Theorem 5. We note that the sufficient decrease condition allows us to prove convergence for the last iterate instead of relying on the existence of an iterate \(k\) in the whole sequence that satisfies the required inequalities (as proven in the corresponding gradient based method by Nouiehed et al. (2019)).

6 EXPERIMENTS

One advantage of direct-search methods is their abilities to explore the space of parameters. This however comes at the price of a high dependency to the size of the parameter space (Vicente, 2013). For nonconvex optimization problems in \(\mathbb{R}^n\), the complexity of DS methods is of the order \(O(n^2)\) (Dodangeh et al.,...
2016). However, recent works by Gratton et al. (2015); Bergou et al. (2018) have shown that replacing the sampling procedure from a PSS by one that correlates with the gradient direction probabilistically, it is possible to achieve a dependence of the order $O(n)$. The sequential aspect of our method allows us to adopt this probabilistic perspective for the experiments to follow, thus lowering the computation cost.

6.1 Robust Optimization

Robustly-regularized estimators have been successfully used in prior work (Namkoong and Duchi, 2017) to deal with situations in which the empirical risk minimizer is susceptible to high amounts of noise. Formally, the problem of empirical risk minimization can be formulated as follows,

$$
\min_{\theta} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}_P[l(X; \theta)] : D(P \parallel \hat{P}_n) \leq \frac{\rho}{n} \right\},
$$

(10)

where $l(X; \theta)$ denotes the loss function, $X$ the data and $D(P \parallel \hat{P}_n)$ a distance function that measures the divergence between the true data distribution $P$ and the empirical data distribution $\hat{P}_n$. For the specific case of a binary classification problem, as for instance considered in Adolphs et al. (2018), Eq. (10) can be reformulated as

$$
\min_{\theta} \max_{p} \left\{ -\sum_{i=1}^{n} p_i [y_i \log(\hat{y}(X_i; \theta)) + (1 - y_i) \log(1 - \hat{y}(X_i; \theta))] - \lambda \sum_{i=1}^{n} \left( p_i - \frac{1}{n} \right)^2 \right\},
$$

where $y_i$ and $\hat{y}(X_i; \theta)$ correspond to the true and the predicted class of data point $X_i$ and $\lambda > 0$ controls the amount of regularization. Note that the aforementioned function is strongly-concave w.r.t $p$ (i.e. it satisfies our PL assumption) and can thus be solved efficiently. We consider this optimization problem on the Wisconsin breast cancer data set\(^4\), comparing the performance between our proposed direct-search method and GDA, using the same neural network as classifier. The zero-one loss is shown in Fig. 1 which clearly shows that our algorithm can consistently outperform GDA for different choices of regularization parameters.

6.2 Categorical Data

Generative Adversarial Networks (Goodfellow et al., 2014) are formulated as the saddle point problem:

$$
\min_{x} \max_{y} f(x, y) = \mathbb{E}_{y \sim P_{data}} [\log D_y(\theta)] + \mathbb{E}_{z \sim P_z} [\log(1 - D_y(G_x(z))],
$$

where $D_y : \mathbb{R}^n \rightarrow [0, 1]$ and $G_x : \mathbb{R}^m \rightarrow \mathbb{R}^n$ are the discriminator and generator networks. Although GANs have been used in a wide variety of applications (Goodfellow, 2016), very few approaches can deal with discrete data. The most severe impeding factor in such settings is the non existence of the gradient due to the non-smooth nature of the objective function. One advantage of direct-search techniques over gradient-based methods is that they can be used in such a context where gradients are not accessible. In some cases, we note that $\ell_2$ regularization can be used to increase the smoothness constant of the objective function.

We illustrate the performance of our direct-search algorithm on a simple example consisting of correlated categorical data, in Figure 2. For a more detailed discussion and more experimental results we refer the reader to the Appendix.

Scaling direct-search to higher dimensions still remains an active area of research, where recent developments

\(^4\)https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
include guided search (Maheswaranathan et al., 2018) and projection-based approaches (Wang et al., 2016).

In this work, we focus on the theoretical guarantees of our algorithm in the stochastic min-max setting. While we demonstrate a good empirical behavior on relatively small-scale problems, scaling our algorithm to large-scale problems will require further modifications to improve its scalability.

7 CONCLUSION

We presented and proved convergence results for a direct-search method in a stochastic minimization setting for both nonconvex and PL objective functions. We then extended these results to prove convergence for min-max objective functions, where the objective of the max-player satisfies the (PL) condition, while the min-player objective is nonconvex. Our experimental results establish that direct-search can outperform traditionally adopted optimization schemes, while also presenting a promising alternative for categorical settings. A potential direction for future work is to improve the scalability of our algorithm in order to run it on large-scale problems, such as adversarial poisoning attacks on benchmark computer vision datasets. Additional extensions of our work include the use of momentum to accelerate convergence as in Gidel et al. (2018) or developing an optimistic variant of our algorithm as in Daskalakis et al. (2017); Daskalakis and Panageas (2018).

8 Acknowledgements

Sotiris Anagnostidis is supported by the Onassis Foundation - Scholarship ID: F ZP 002-1/2019-2020.

References


