APPROXIMATING NASH EQUILIBRIA IN NORMAL-FORM GAMES VIA STOCHASTIC OPTIMIZATION

Ian Gemp
DeepMind
London, UK
imgemp@deepmind.com

Luke Marris
DeepMind
London, UK
marris@deepmind.com

Georgios Piliouras
DeepMind
London, UK
gpil@deepmind.com

ABSTRACT

We propose the first loss function for approximate Nash equilibria of normal-form games that is amenable to unbiased Monte Carlo estimation. This construction allows us to deploy standard non-convex stochastic optimization techniques for approximating Nash equilibria, resulting in novel algorithms with provable guarantees. We complement our theoretical analysis with experiments demonstrating that stochastic gradient descent can outperform previous state-of-the-art approaches.

1 INTRODUCTION

Nash equilibrium (NE) famously encodes stable behavioral outcomes in multi-agent systems and is arguably the most influential solution concept in game theory. Formally speaking, if n players independently choose n, possibly mixed, strategies (x_i for $i \in [n]$) and their joint strategy ($\boldsymbol{x} = \prod_i x_i$) constitutes a *Nash equilibrium*, then no player has any incentive to unilaterally deviate from their strategy. This concept has sparked extensive research in various fields, ranging from economics (Milgrom and Weber, 1982) to machine learning (Goodfellow et al., 2014), and has even inspired behavioral theory generalizations such as quantal response equilibria which allow for more realistic models of boundedly rational agents (McKelvey and Palfrey, 1995).

Unfortunately, when considering Nash equilibria beyond the 2-player, zero-sum scenario, two significant challenges arise. First, it becomes unclear how n independent players would collectively identify a Nash equilibrium when multiple equilibria are possible, giving rise to the *equilibrium selection* problem (Harsanyi et al., 1988). Secondly, even approximating a single Nash equilibrium is known to be computationally intractable and specifically PPAD-complete (Daskalakis et al., 2009). Combining both problems together, e.g., testing for the existence of equilibria with welfare greater than some fixed threshold is NP-hard and it is in fact even hard to approximate (Austrin et al., 2011).

From a machine learning practitioner's perspective, such computational complexity results hardly give pause for thought as collectively we have become all too familiar with the unreasonable effectiveness of heuristics in circumventing such obstacles. Famously, non-convex optimization is NP-hard, even if the goal is to compute a local minimizer (Murty and Kabadi, 1985), however, stochastic gradient descent (SGD) and variants succeed in training billion parameter models (Brown et al., 2020).

Unfortunately, computational techniques for Nash equilibrium have so far not achieved anywhere near the same level of success. In contrast, most modern NE solvers for n-player, m-action, general-sum, normal-form games (NFGs) are practically restricted to a handful of players and/or actions per player except in special cases, e.g., symmetric (Wiedenbeck and Brinkman, 2023) or mean-field games (Pérolat et al., 2022). For example, when running the suite of all 7 applicable methods from the hallmark `gambit` library (McKelvey et al., 2016) on a 4-player Blotto game, we find only brute-force pure-NE enumeration is able to return any NE within a 1 hour time limit. Scaling solvers to large games is difficult partially due to the fact that an NFG is represented by a tensor with an exponential nm^n entries; even *reading* this description into memory can be computationally prohibitive. More to the point, any computational technique that presumes *exact* computation of the *expectation* of any function sampled according to $\boldsymbol{x}$ similarly does not have any hope of scaling beyond small instances.

This inefficiency arguably lies at the core of the differential success between ML optimization and equilibrium computation. For example, numerous techniques exist that reduce the problem of Nash computation to the minimization of the expectation of a random variable (Section 3). Unfortunately, unlike the source of randomness in ML applications where batch learning suffices to easily produce unbiased estimators, these techniques do not extend easily to game theory which incorporates non-linear functions such as maximum and best-response. This raises our motivating goal:

Can we solve for Nash equilibria via unbiased stochastic optimization?

Our results. Following in the successful steps of the interplay between ML and stochastic optimization, we reformulate the approximation of Nash equilibria in an NFG as a stochastic non-convex optimization problem admitting unbiased Monte-Carlo estimation. This enables the use of powerful solvers and advances in parallel computing to efficiently enumerate Nash equilibria for n-player, general-sum games. Furthermore, this re-casting allows practitioners to incorporate other desirable objectives into the problem such as "find an approximate Nash equilibrium with welfare above ω" or "find an approximate Nash equilibrium nearest the current observed joint strategy" resolving the equilibrium selection problem in an effectively ad-hoc and application tailored manner. Concretely, we make the following contributions by producing:

- A loss $\mathcal{L}^{\tau}(\boldsymbol{x})$ 1) whose global minima well approximate Nash equilibria in normal form games, 2) admits unbiased Monte-Carlo estimation, and 3) is Lipschitz and bounded.
- An efficient randomized algorithm for approximating Nash equilibria in a novel class of games. The algorithm emerges by employing the family of $\mathcal{X}$-armed bandit approaches to $\mathcal{L}^{\tau}(\boldsymbol{x})$ and connecting their global stochastic optimization guarantees to global approximate Nash guarantees.
- An empirical comparison of SGD against state-of-the-art baselines for approximating NEs in large games. In some games, vanilla SGD actually improves upon previous state-of-the-art; in others, SGD is slowed by saddle points, a familiar challenge in deep learning (Dauphin et al., 2014).

Overall, this perspective showcases a promising new route to approximating equilibria at scale in practice. We conclude the paper with discussion for future work.

2 PRELIMINARIES

In an n-player, normal-form game, each player $i \in \{1, \ldots, n\}$ has a strategy set $\mathcal{A}_i = \{a_{i1}, \ldots, a_{im_i}\}$ consisting of m_i pure strategies. These strategies can be naturally indexed, so we redefine $\mathcal{A}_i = \{1, \ldots, m_i\}$ as an abuse of notation. Each player i also has a utility function, $u_i : \mathcal{A} = \prod_i \mathcal{A}_i \rightarrow [0, 1]$, (equiv. "payoff tensor") that maps joint actions to payoffs in the unit-interval . We denote the average cardinality of the players' action sets by $\bar{m} = \frac{1}{n} \sum_k m_k$ and maximum by $m^* = \max_k m_k$. Player i may play a mixed strategy by sampling from a distribution over their pure strategies. Let player i's mixed strategy be represented by a vector $x_i \in \Delta^{m_i - 1}$ where $\Delta^{m_i - 1}$ is the $(m_i - 1)$-dimensional probability simplex embedded in $\mathbb{R}^{m_i}$. Each function u_i is then extended to this domain so that $u_i(\boldsymbol{x}) = \sum_{\boldsymbol{a} \in \mathcal{A}} u_i(\boldsymbol{a}) \prod_j x_{ja_j}$ where $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $a_j \in \mathcal{A}_j$ denotes player j's component of the joint action $\boldsymbol{a} \in \mathcal{A}$. For convenience, let x_{-i} denote all components of $\boldsymbol{x}$ belonging to players other than player i.

The joint strategy $\boldsymbol{x} \in \prod_i \Delta^{m_i - 1}$ is a Nash equilibrium if and only if, for all $i \in \{1, \ldots, n\}$, $u_i(z_i, x_{-i}) \leq u_i(\boldsymbol{x})$ for all $z_i \in \Delta^{m_i - 1}$, i.e., no player has any incentive to unilaterally deviate from $\boldsymbol{x}$. Nash is typically relaxed with ϵ-Nash, our focus: $u_i(z_i, x_{-i}) \leq u_i(\boldsymbol{x}) + \epsilon$ for all $z_i \in \Delta^{m_i - 1}$.

As an abuse of notation, let the atomic action $a_i = e_i$ also denote the m_i-dimensional "one-hot" vector with all zeros aside from a 1 at index a_i; its use should be clear from the context. We also introduce $\nabla^i_{x_i}$ as player i's utility gradient. And for convenience, denote by $H^i_{il} = \mathbb{E}_{x_{-il}}[u_i(a_i, a_l, x_{-il})]$ the bimatrix game approximation (Janovskaja, 1968) between players i and l with all other players marginalized out; x_{-il} denotes all strategies belonging to players other than i and l and $u_i(a_i, a_l, x_{-il})$ separates out l's strategy x_l from the rest of the players x_{-i}. Similarly, denote by $T^i_{ilq} = \mathbb{E}_{x_{-ilq}}[u_i(a_i, a_l, a_q, x_{-ilq})]$ the 3-player tensor approximation to the game. Note player i's utility can now be written succinctly as $u_i(x_i, x_{-i}) = x_i^\top \nabla^i_{x_i} = x_i^\top H^i_{il} x_l = x_i T^i_{ilq} x_l x_q$ for any l, q where we use Einstein notation for tensor arithmetic. For convenience, define $\texttt{diag}(z)$ as the function that places a vector z on the diagonal of a square matrix, and $\texttt{diag3} : z \in \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d \times d}$ as a 3-tensor of shape (d, d, d) where $\texttt{diag3}(z)_{iii} = z_i$. Following convention from differential

Loss	Function	Obstacle
Exploitabilty	$\max_k \epsilon_k(\boldsymbol{x})$	max of r.v.
Nikaido-Isoda (NI)	$\sum_k \epsilon_k(\boldsymbol{x})$	max of r.v.
Fully-Diff. Exp	$\sum_k \sum_{a_k \in \mathcal{A}_k} [\max(0, u_k(a_k, x_{-i}) - u_k(\boldsymbol{x}))]^2$	max of r.v.
Gradient-based NI	NI w/ $\texttt{BR}_k \leftarrow \texttt{aBR}_k = \Pi_\Delta\Big(x_k + \eta \nabla_{x_k} u_k(\boldsymbol{x})\Big)$	Π_Δ of r.v.
Unconstrained	Loss + Simplex Deviation Penalty	sampling from $x_i \in \mathbb{R}^{m_k}$

Table 1: Previous loss functions for NFGs and their obstacles to unbiased estimation. Note that $\epsilon_k(\boldsymbol{x}) = \max_z u_k(z, x_{-k}) - u_k(\boldsymbol{x})$ contains a max operator (see equivalent definition in equation (1)).

geometry, let $T_v\mathcal{M}$ be the tangent space of a manifold $\mathcal{M}$ at v. For the interior of the d-action simplex Δ^{d-1}, the tangent space is the same at every point, so we drop the v subscript, i.e., $T\Delta^{d-1}$. We denote the projection of a vector $z \in \mathbb{R}^d$ onto this tangent space as $\Pi_{T\Delta^{d-1}}(z) = z - \frac{1}{d}\mathbf{1}\mathbf{1}^\top z$ and call $\Pi_{T\Delta^{d-1}}(\nabla^i_{x_i})$ a *projected-gradient*. We drop $d-1$ when the dimensionality is clear from the context. Finally, let $\mathcal{U}(S)$ denote a discrete uniform distribution over elements from set S.

3 RELATED WORK

Representing the problem of computing a Nash equilibrium as an optimization problem is not new. A variety of loss functions and pseudo-distance functions have been proposed. Most of them measure some function of how much each player can exploit the joint strategy by unilaterally deviating:

$$\epsilon_k(\boldsymbol{x}) \overset{\text{def}}{=} u_k(\texttt{BR}_k, x_{-k}) - u_k(\boldsymbol{x}) \text{ where } \texttt{BR}_k \in \arg\max_z u_k(z, x_{-k}). \tag{1}$$

As argued in the introduction, we believe it is important to be able to subsample payoff tensors of normal-form games in order to scale to large instances. As Nash equilibria can consist of mixed strategies, it is advantageous to be able to sample from an equilibrium to estimate its exploitability ϵ. However none of these losses is amenable to unbiased estimation under sampled play. Each of the functions currently explored in the literature is biased under sampled play either because 1) a random variable appears as the argument of a complex, nonlinear (non-polynomial) function or because 2) how to sample play is unclear. Exploitability, Nikaido-Isoda (NI) (Nikaidô and Isoda, 1955) (also known by `NashConv` (Lanctot et al., 2017) and ADI (Gemp et al., 2022)), as well as fully-differentiable options (Shoham and Leyton-Brown, 2008, p. 106, Eqn 4.31) introduce bias when a max over payoffs is estimated using samples from $\boldsymbol{x}$. Gradient-based NI (Raghunathan et al., 2019) requires projecting the result of a gradient-ascent step onto the simplex; for the same reason as the max, this is prohibitive because it is a nonlinear operation which introduces bias. Lastly, unconstrained optimization approaches (Shoham and Leyton-Brown, 2008, p. 106) that instead penalize deviation from the simplex lose the ability to sample from strategies when each iterate $\boldsymbol{x}$ is no longer a distribution (i.e., $x_k \notin \Delta^{m_k-1}$). Table 1 summarizes these complications.

4 NASH EQUILIBRIUM AS STOCHASTIC OPTIMIZATION

We will now develop our proposed loss function which is amenable to unbiased estimation. Subsections 4.1-4.4 provide a warm-up in which we assume an interior (fully-mixed) Nash equilibrium exists. Subsection 4.5 then shows how to relax that assumption allowing us to approximate partially mixed equilibria as well (including pure equilibria). Our key technical insight is to pay special attention to the geometry of the simplex. To our knowledge, prior works have failed to recognize the role of the tangent space $T\Delta$. Proofs are in the appendix.

4.1 STATIONARITY ON THE SIMPLEX INTERIOR

Lemma 1. *Assuming player i's utility, $u_i(x_i, x_{-i})$, is concave in its own strategy x_i, a strategy in the interior of the simplex is a best response $\texttt{BR}_i$ if and only if it has zero projected-gradient*[1] *norm.*

In NFGs, each player's utility is linear in x_i, thereby satisfying the concavity condition of Lemma 1.

---

[1] Not to be confused with the nonlinear (biased) projected gradient operator in (Hazan et al., 2017).

4.2 PROJECTED GRADIENT NORM AS LOSS

An equivalent description of a Nash equilibrium is a joint strategy $\boldsymbol{x}$ where every player's strategy is a best response to the equilibrium (i.e., $x_i = \texttt{BR}_i$ so that $\epsilon_i(\boldsymbol{x}) = 0$). Lemma 1 states that any interior best response has zero *projected-gradient* norm, which inspires the following loss function

$$\mathcal{L}(\boldsymbol{x}) = \sum_k \eta_k ||\Pi_{T\Delta}(\nabla^k_{x_k})||^2 \tag{2}$$

where $\eta_k > 0$ represent scalar weights, or equivalently, step sizes to be explained next.

Proposition 1. *The loss $\mathcal{L}$ is equivalent to* `NashConv`*, but where player k's best response is approximated by a single step of projected-gradient ascent with step size η_k:* $\texttt{aBR}_k = x_k + \eta_k \Pi_{T\Delta}(\nabla^k_{x_k})$.

This connection was already pointed out in prior work for unconstrained problems (Gemp et al., 2022; Raghunathan et al., 2019), but this result is the first for strategies constrained to the simplex.

4.3 CONNECTION TO TRUE EXPLOITABILITY

In general, we can bound exploitability in terms of the projected-gradient norm as long as each player's utility is concave (this result extends to subgradients of non-smooth functions).

Lemma 2. *The amount a player can gain by exploiting a joint strategy $\boldsymbol{x}$ is upper bounded by a quantity proportional to the norm of the projected-gradient:*

$$\epsilon_k(\boldsymbol{x}) \leq \sqrt{2}||\Pi_{T\Delta}(\nabla^k_{x_k})||. \tag{3}$$

This bound is not tight on the boundary of the simplex, which can be seen clearly by considering x_k to be part of a pure strategy equilibrium. In that case, this analysis assumes x_k can be improved upon by a projected-gradient ascent step (via the equivalence pointed out in Proposition 1). However, that is false because the probability of a pure strategy cannot be increased beyond 1. We mention this to provide further intuition for why our "warm-up" loss $\mathcal{L}(\boldsymbol{x})$ is only valid for interior equilibria.

Note that $||\Pi_{T\Delta}(\nabla^k_{x_k})|| \leq ||\nabla^k_{x_k}||$ because $\Pi_{T\Delta}$ is a projection. Therefore, this improves the naive bounds on exploitability and distance to best responses given using the "raw" gradient $\nabla^k_{x_k}$.

Lemma 3. *The exploitability of a joint strategy $\boldsymbol{x}$, is upper bounded by a function of $\mathcal{L}(\boldsymbol{x})$:*

$$\epsilon \leq \sqrt{\frac{2n}{\min_k \eta_k}} \sqrt{\mathcal{L}(\boldsymbol{x})} \overset{\text{def}}{=} f(\mathcal{L}). \tag{4}$$

4.4 UNBIASED ESTIMATION

As discussed in Section 3, a primary obstacle to unbiased estimation of $\mathcal{L}(\boldsymbol{x})$ is the presence of complex, nonlinear functions of random variables, with the projection of a point onto the simplex being one such example (see Π_Δ in Table 1). However, $\Pi_{T\Delta}$, the projection onto the tangent space of the simplex, *is linear*! This is the key that allows us to design an unbiased estimator (Lemma 5).

Our proposed loss requires computing the squared norm of the *expected value* of the gradient under the players' mixed strategies, i.e., the l-th entry of player k's gradient equals $\nabla^k_{x_{kl}} = \mathbb{E}_{a_{-k} \sim x_{-k}} u_k(a_{kl}, a_{-k})$. By analogy, consider a random variable Y. In general, $\mathbb{E}[Y]^2 \neq \mathbb{E}[Y^2]$. This means that we cannot just sample projected-gradients and then compute their average norm to estimate our loss. However, consider taking two independent samples from two corresponding identically distributed, independent random variables $Y^{(1)}$ and $Y^{(2)}$. Then $\mathbb{E}[Y^{(1)}]^2 = \mathbb{E}[Y^{(1)}]\mathbb{E}[Y^{(2)}] = \mathbb{E}[Y^{(1)}Y^{(2)}]$ by properties of expected value over products of independent random variables. This is a common technique to construct unbiased estimates of expectations over polynomial functions of random variables. Proceeding in this way, define $\nabla^{k(1)}_{x_k}$ as a random variable distributed according to the distribution induced by all other players' mixed strategies ($j \neq k$). Let $\nabla^{k(2)}_{x_k}$ be independent and distributed identically to $\nabla^{k(1)}_{x_k}$. Then

$$\mathcal{L}(\boldsymbol{x}) = \mathbb{E}[\sum_k \eta_k \underbrace{(\hat{\nabla}^{k(1)}_{x_k} - \frac{\mathbf{1}}{m_k}(\mathbf{1}^\top \hat{\nabla}^{k(1)}_{x_k})\mathbf{1})}_{\text{projected-gradient 1}}{}^\top \underbrace{(\hat{\nabla}^{k(2)}_{x_k} - \frac{\mathbf{1}}{m_k}(\mathbf{1}^\top \hat{\nabla}^{k(2)}_{x_k})\mathbf{1})}_{\text{projected-gradient 2}}] \tag{5}$$

	Exact	Sample Others	Sample All
Estimator of $\nabla^{k(p)}_{x_k}$	$[u_k(a_{kl}, x_{-k})]_l$	$[u_k(a_{kl}, a_{-k} \sim x_{-k})]_l$	$m_k u_k(a_{kl} \sim \mathcal{U}(\mathcal{A}_k), a_{-k} \sim x_{-k}) e_l$
$\hat{\nabla}^{k(p)}_{x_k}$ Bounds	$[0, 1]$	$[0, 1]$	$[0, m_k]$
$\hat{\nabla}^{k(p)}_{x_k}$ Query Cost	$\prod_{i=1}^n m_i$	m_k	1
$\hat{\mathcal{L}}$ Bounds	$\pm{}^1/_4 \sum_k \eta_k m_k$	$\pm{}^1/_4 \sum_k \eta_k m_k$	$\pm{}^1/_4 \sum_k \eta_k m_k^3$
$\hat{\mathcal{L}}$ Query Cost	$n \prod_{i=1}^n m_i$	$2n\bar{m}$	$2n$

Table 2: Examples and Properties of Unbiased Estimators of Loss and Player Gradients ($\hat{\nabla}^{k(p)}_{x_k}$).

where $\hat{\nabla}^{k(p)}_{x_k}$ is an unbiased estimator of player k's gradient. This estimator can be constructed in several ways. The most expensive, an exact estimator, is constructed by marginalizing player k's payoff tensor over all other players' strategies. However, a cheaper estimate can be obtained at the expense of higher variance by approximating this marginalization with a Monte Carlo estimate of the expectation. Specifically, if we sample a single action for each of the remaining players, we can construct an unbiased estimate of player k's gradient by considering the payoff of each of its actions against the sampled background strategy. Lastly, we can consider constructing an estimate of player k's gradient by sampling only a single action from player k to represent their entire gradient. Each of these approaches is outlined in Table 2 along with the query complexity (Babichenko, 2016) of computing the estimator and bounds on the values it can take (Lemma 9).

We can extend Lemma 3 to one that holds under T samples with probability $1 - \delta$ by applying, for example, a Hoeffding bound: $\epsilon \leq f\big(\hat{\mathcal{L}}(\boldsymbol{x}) + \mathcal{O}(\sqrt{\frac{1}{T}\ln(1/\delta)}\big)$.

4.5 INTERIOR EQUILIBRIA

We discussed earlier that $\mathcal{L}(\boldsymbol{x})$ captures interior equilibria. But some games may only have *partially mixed* equilibria, i.e., equilibria that lie on the boundary of the simplex. We show how to circumvent this shortcoming by considering quantal response equilibria (QREs), specifically, logit equilibria. By adding an entropy bonus to each player's utility, we can

- guarantee **all** equilibria are interior,
- still obtain unbiased estimates of our loss,
- maintain an upper bound on the exploitability ϵ of any approximate Nash equilibrium in the original game (i.e., the game without an entropy bonus).

Define $u^\tau_k(\boldsymbol{x}) = u_k(\boldsymbol{x}) + \tau S(x_k)$ where Shannon entropy $S(x_k) = -\sum_l x_{kl}\ln(x_{kl})$ is 1-strongly concave with respect to the 1-norm (Beck and Teboulle, 2003). It is known that Nash equilibria of entropy-regularized games satisfy the conditions for logit equilibria (Leonardos et al., 2021), which are solutions to the fixed point equation $x_k = \texttt{softmax}(\frac{1}{\tau}\nabla^k_{x_k})$. The `softmax` makes clear that all probabilities have positive mass at positive temperature.

Recall that in order to construct an unbiased estimate of our loss, we simply needed to construct unbiased estimates of player gradients. The introduction of the entropy term to player k's utility is special in that it depends entirely on known quantities, i.e., the player's own mixed strategy. We can directly and deterministically compute $\tau\frac{dS}{dx_k} = -\tau(\ln(x_k) + \mathbf{1})$ and add this to our estimator of $\nabla^{k(p)}_{x_k}$: $\hat{\nabla}^{k\tau(p)}_{x_k} = \hat{\nabla}^{k(p)}_{x_k} + \tau\frac{dS}{dx_k}$. Consider our loss function refined from (2) with changes in blue:

$$\mathcal{L}^\tau(\boldsymbol{x}) = \sum_k \eta_k ||\Pi_{T\Delta}(\nabla^{k\tau}_{x_k})||^2. \tag{6}$$

As mentioned above, the utilities with entropy bonuses are still concave, therefore, a similar bound to Lemma 2 applies. We use this to prove the QRE counterpart to Lemma 3 where ϵ_{QRE} is the exploitability of an approximate equilibrium in a game with entropy bonuses.

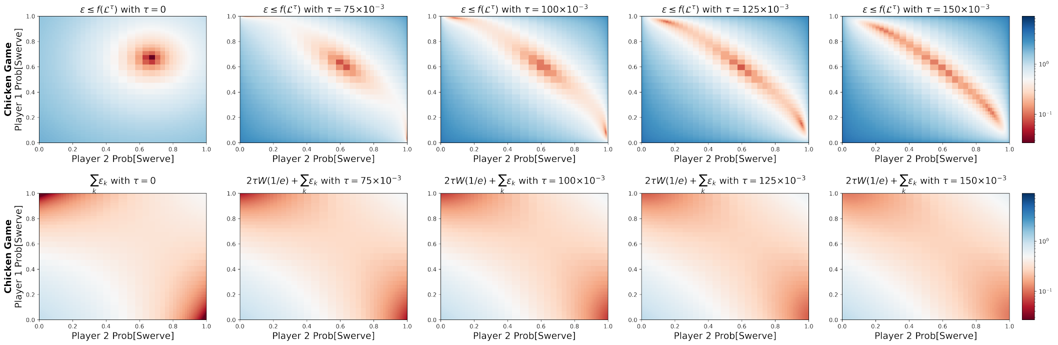

Figure 1: Effect of Sampled Play on a Biased Loss. The first row displays the expectation of the upper bound guaranteed by our proposed loss $\mathcal{L}^\tau$. The second row displays the expectation of NashConv under sampled play, i.e., $\sum_k \epsilon_k$ where $\epsilon_k = \mathbb{E}_{a_{-k} \sim x_{-k}}[\max_{a_k} u_k^\tau(\boldsymbol{a})] - \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{x}}[u_k^\tau(\boldsymbol{a})]$. To be consistent, we add the offset $n\tau W(1/e)$ to NashConv per Lemma 16, which relates the exploitability at positive temperature to that at zero temperature. The resulting loss surface clearly shows NashConv fails to recognize any interior Nash equilibrium due to its inherent bias.

Lemma 4. *The entropy regularized exploitability, ϵ_{QRE}, of a joint strategy $\boldsymbol{x}$, is upper bounded as:*

$$\epsilon_{QRE} \leq \sqrt{\frac{2n}{\min_k \eta_k}} \sqrt{\mathcal{L}^\tau(\boldsymbol{x})} \overset{\text{def}}{=} f(\mathcal{L}^\tau). \tag{7}$$

Lastly, we establish a connection between quantal response equilibria and Nash equilibria that allows us to approximate Nash equilibria in the original game via minimizing our modified loss $\mathcal{L}^\tau(\boldsymbol{x})$.

Lemma 16 **($\mathcal{L}^\tau$ Scores Nash Equilibria).** *Let $\mathcal{L}^\tau(\boldsymbol{x})$ be our proposed entropy regularized loss function with payoffs bounded in $[0, 1]$ and $\boldsymbol{x}$ be an approximate QRE. Then it holds that*

$$\epsilon \leq n\tau(W(1/e) + \frac{\bar{m} - 2}{e}) + 2\sqrt{\frac{n \max_k m_k}{\min_k \eta_k}} \sqrt{\mathcal{L}^\tau(\boldsymbol{x})} \tag{8}$$

where W is the Lambert function: $W(1/e) = W(\exp(-1)) \approx 0.278$.

This upper bound is plotted as a heatmap for a familiar Chicken game in the top row of Figure 1. First, notice how pure equilibria are not visible as minima for zero temperature, but appear for slightly warmer temperatures. Secondly, notice that NashConv in the bottom row is unable to capture the interior Nash equilibrium because of its high bias under sampled play. In contrast, our proposed loss $\mathcal{L}^\tau$ is guaranteed to capture all equilibria at low temperature τ.

5 ANALYSIS

In the preceding section we established a loss function that upper bounds the exploitability of an approximate equilibrium. In addition, the zeros of this loss function have a one-to-one correspondence with quantal response equilibria (which approximate Nash equilibria at low temperature).

Here, we derive properties that suggest it is "easy" to optimize. While this function is generally non-convex and may suffer from a proliferation of saddle points (Figure 2) , it is Lipschitz continuous (over the relevant subset of the interior) and bounded. These are two commonly made assumptions in the literature on non-convex optimization, which we leverage in Section 6. In addition, we can derive its gradient, its Hessian, and characterize its behavior around global minima.

Lemma 17. *The gradient of $\mathcal{L}^\tau(\boldsymbol{x})$ with respect to player l's strategy x_l is*

$$\nabla_{x_l} \mathcal{L}^\tau(\boldsymbol{x}) = 2 \sum_k \eta_k B_{kl}^\top \Pi_{T\Delta}(\nabla_{x_k}^{k\tau}) \tag{9}$$

where $B_{ll} = -\tau[I - \frac{1}{m_l}\mathbf{1}\mathbf{1}^\top]diag(\frac{1}{x_l})$ and $B_{kl} = [I - \frac{1}{m_k}\mathbf{1}\mathbf{1}^\top]H_{kl}^k$ for $k \neq l$.

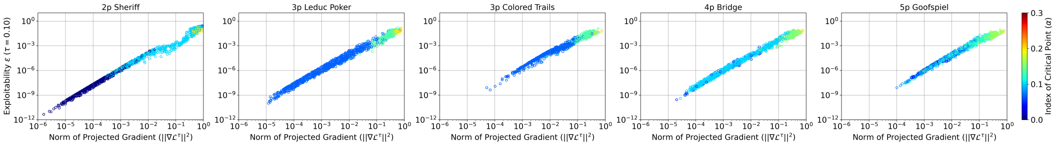

Figure 2: Analysis of Loss Landscape. We reapply the analysis of (Dauphin et al., 2014), originally designed to understand the success of SGD in deep learning, to "slices" of several popular extensive form games. To construct a slice (or *meta-game*), we randomly sample 6 deterministic policies and then consider the corresponding n-player, 6-action normal-form game at $\tau = 0.1$ (with payoffs normalized to $[0, 1]$). The index of a critical point $\boldsymbol{x}_c$ ($\nabla_{\boldsymbol{x}} \mathcal{L}^{\tau}(\boldsymbol{x}_c) = \mathbf{0}$) indicates the fraction of negative eigenvalues in the Hessian of $\mathcal{L}^{\tau}$ at $\boldsymbol{x}_c$; $\alpha = 0$ indicates a local minimum, 1 a maximum, else a saddle point. We see a positive correlation between exploitability (y-axis), *projected*-gradient norm (x-axis), and α (color) indicating a lower prevalence of local minima at high exploitability.

Lemma 19. *The Hessian of $\mathcal{L}^{\tau}(\boldsymbol{x})$ can be written*

$$\texttt{Hess}(\mathcal{L}^{\tau}) = 2\big[\tilde{B}^{\top}\tilde{B} + T\Pi_{T\Delta}(\tilde{\nabla}^{\tau})\big] \tag{10}$$

where $\tilde{B}_{kl} = \sqrt{\eta_k} B_{kl}$, $\Pi_{T\Delta}(\tilde{\nabla}^{\tau}) = [\eta_1 \Pi_{T\Delta}(\nabla^{1\tau}_{x_1}), \ldots, \eta_n \Pi_{T\Delta}(\nabla^{n\tau}_{x_n})]$, and we augment T (the 3-player approximation to the game, T^{k}_{lqk}) so that $T^{l}_{lll} = \tau \texttt{diag3}(\frac{1}{x_l^2})$.

At an equilibrium, the latter term disappears because $\Pi_{T\Delta}(\nabla^{k\tau}_{x_k}) = \mathbf{0}$ for all k (Lemma 1). If $\mathcal{X}$ was $\mathbb{R}^{n\bar{m}}$, then we could simply check if $\tilde{B}$ is full-rank to determine if $Hess \succ 0$. However, $\mathcal{X}$ is a simplex product, and we only care about curvature in directions toward which we can update our strategy profile $\boldsymbol{x}$. Toward that end, define M to be the $n(\bar{m}+1) \times n\bar{m}$ matrix that stacks $\tilde{B}$ on top of a repeated identity matrix that encodes orthogonality to the simplex:

$$M(\boldsymbol{x}) = \begin{bmatrix} -\tau\sqrt{\eta_1}\Pi_{T\Delta}(\frac{1}{x_1}) & \sqrt{\eta_1}\Pi_{T\Delta}(H^{1}_{12}) & \ldots & \sqrt{\eta_1}\Pi_{T\Delta}(H^{1}_{1n}) \\ \vdots & \vdots & \vdots & \vdots \\ \sqrt{\eta_n}\Pi_{T\Delta}(H^{n}_{n1}) & \ldots & \sqrt{\eta_n}\Pi_{T\Delta}(H^{n}_{n,n-1}) & -\tau\sqrt{\eta_n}\Pi_{T\Delta}(\frac{1}{x_n}) \\ \mathbf{1}_1^{\top} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & \mathbf{1}_n^{\top} \end{bmatrix} \tag{11}$$

where $\Pi_{T\Delta}(z \in \mathbb{R}^{a \times b}) = [I_a - \frac{1}{a}\mathbf{1}_a\mathbf{1}_a^{\top}]z$ subtracts the mean from each column of z and $\frac{1}{x_i}$ is shorthand for $\texttt{diag}(\frac{1}{x_i})$. If $M(x)z = \mathbf{0}$ for a nonzero vector $z \in \mathbb{R}^{n\bar{m}}$, this implies there exists a z that 1) is orthogonal to the ones vectors of each simplex (i.e., is a valid equilibrium update direction) and 2) achieves zero curvature in the direction z, i.e., $z^{\top}(\tilde{B}^{\top}\tilde{B})z = z^{\top}(Hess)z = 0$, and so $Hess$ is not positive definite. Conversely, if $M(\boldsymbol{x})$ is of rank $n\bar{m}$ for a quantal response equilibrium $\boldsymbol{x}$, then the Hessian of $\mathcal{L}^{\tau}$ at $\boldsymbol{x}$ in the tangent space of the simplex product ($\mathcal{X} = \prod_i \mathcal{X}_i$) is positive definite. In this case, we call $\boldsymbol{x}$ *polymatrix*-isolated: **polymatrix** because we only require information of the local polymatrix approximation of the game (i.e., the H^{i}_{ij} matrices) to construct M and **isolated** because it implies $\boldsymbol{x}$ is not connected to any other equilibria.

Definition 1 (*Polymatrix*-Isolated Equilibrium)**.** *A Nash equilibrium $\boldsymbol{x}^*$ is polymatrix-isolated iff $\boldsymbol{x}^*$ is isolated according to its local polymatrix game approximation.*

By analyzing the rank of M, we can confirm that many classical matrix games including Rock-Paper-Scissors, Chicken, Matching Pennies, and Shapley's game all induce strongly convex $\mathcal{L}^{\tau}$'s at zero temperature (i.e., they have unique mixed Nash equilibria). In contrast, a game like Prisoner's Dilemma has a unique pure strategy that will not be captured by our loss at zero temperature.

6 ALGORITHMS

We have formally transformed the approximation of Nash equilibria in NFGs into a **stochastic** optimization problem. To our knowledge, this is the first such formulation that allows one-shot

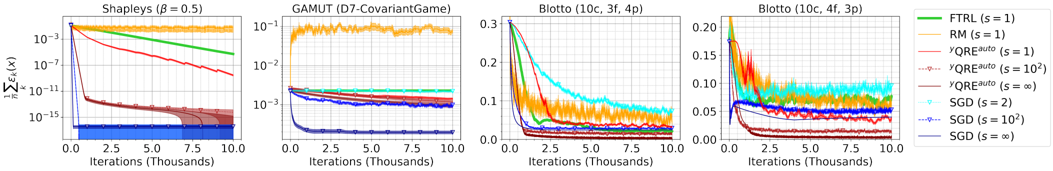

Figure 3: Comparison of SGD on $\mathcal{L}^{\tau=0}$ against baselines on four games evaluated in (Gemp et al., 2022). The number of samples used to estimate each update iteration (i.e., minibatch size) is indicated by s. From left to right: 2-player, 3-action, nonsymmetric; 6-player, 5-action, nonsymmetric; 4-player, 66-action, symmetric; 3-player, 286-action, symmetric. SGD struggles at saddle points in Blotto.

unbiased Monte-Carlo estimation which is critical to introduce the use of powerful algorithms capable of solving high dimensional optimization problems. We explore two off-the-shelf approaches.

6.1 STOCHASTIC GRADIENT DESCENT

Stochastic gradient descent is the workhorse of high-dimensional stochastic optimization. It comes with guaranteed convergence to stationary points (Cutkosky et al., 2023), however, it may converge to local, rather than global minima. It also enjoys implicit gradient regularization (Barrett and Dherin, 2020), seeking "flat" minima and performs approximate Bayesian inference (Mandt et al., 2017). Despite the lack of global convergence guarantee, we find it performs well empirically in games previously examined by the literature: modified Shapley's (Ostrovski and van Strien, 2013), GAMUT D7 (Nudelman et al., 2004), Blotto (Arad and Rubinstein, 2012). Figure 3 shows SGD is competitive with scalable techniques to approximating NEs: FTRL (Shalev-Shwartz and Singer, 2006; Shalev-Shwartz et al., 2012), Regret Matching (Hart and Mas-Colell, 2000), ADIDAS (Gemp et al., 2022). Shapley's game induces a strongly convex $\mathcal{L}$ (see Section 5) leading to SGD's strong performance. Blotto reaches low, but nonzero ϵ, demonstrating the challenges of saddle points.

6.2 HIGH PROBABILITY, GLOBAL POLYNOMIAL CONVERGENCE RATES VIA BANDITS

We explore one other algorithmic approach to non-convex optimization based on minimizing regret, which enjoys finite time **global** convergence rates. $\mathcal{X}$-armed bandits (Bubeck et al., 2011) systematically explore the space of solutions by refining a mesh over the joint strategy space, trading off exploration versus exploitation of promising regions. Several approaches exist (Bartlett et al., 2019; Valko et al., 2013) with open source implementations, e.g., (Li et al., 2023). Applying $\mathcal{X}$-armed bandits to our $\mathcal{L}^\tau$ can be thought of as a stochastic generalization of the *exclusion method* and other bandit approaches for Nash equilibria (Berg and Sandholm, 2017; Zhou et al., 2017).

Equipped with these techniques, we can establish a high probability polynomial-time **global** convergence rate to Nash equilibria in n-player, general-sum games under mild assumptions. The quality of this approximation improves as $\tau \to 0$, at the same time increasing the constant on the convergence rate via the Lipschitz constant $\sqrt{\hat{L}}$ defined below. For clarity, we assume users provide a temperature in the form $\tau = \frac{1}{\ln(1/p)}$ with $p \in (0, 1)$ which ensures all equilibria have probability mass greater than $\frac{p}{m^*}$ for all actions (Lemma 11). Lower p corresponds with lower temperature.

Theorem 4 **(BLiN PAC Rate).** *Assume* $\eta_k = \eta = 2/\hat{L}$, $\tau = \frac{1}{\ln(1/p)}$, *and a previously pulled arm is returned uniformly at random (i.e.,* $t \sim U([T])$*). Then for any* $w > 0$

$$\epsilon_t \leq w\Big[\frac{n}{\ln(1/p)}\big(W(1/e) + \frac{\bar{m}-2}{e}\big) + 4(1 + (4c^2 C_z)^{1/3})\sqrt{nm^*\hat{L}}\Big(\frac{\ln T}{T}\Big)^{\frac{1}{2(d_z+2)}}\Big] \tag{12}$$

with probability $(1 - w^{-1})(1 - 2T^{-2})$ *where* $W(1/e) \approx 0.278$, $m^* = \max_k m_k$, $2|\mathcal{L}^\tau| \leq c \leq \frac{1}{4}\Big(\frac{\ln(m^*)}{\ln(1/p)} + 2\Big)$ *(Lemma 10),* $\hat{L} = \Big(\frac{\ln(m^*)}{\ln(1/p)} + 2\Big)\Big(\frac{m^{*2}}{p\ln(1/p)} + n\bar{m}\Big)$ *(Corollary 1), the zooming dimension* $d_z = \frac{1}{2}n\bar{m}$, *and the zooming constant* $C_z = |\mathcal{X}^*|^{-1}\Big(\frac{4}{r_\eta^2\sigma_{-\infty}}\Big)^{n\bar{m}}$ *(Corollary 33).*

The convergence rate for BLiN (Feng et al., 2022) depends on bounds on the exploitability in terms of the loss (Lemma 16), bounds on estimates of the loss (Lemma 10), Lipschitz bounds on the

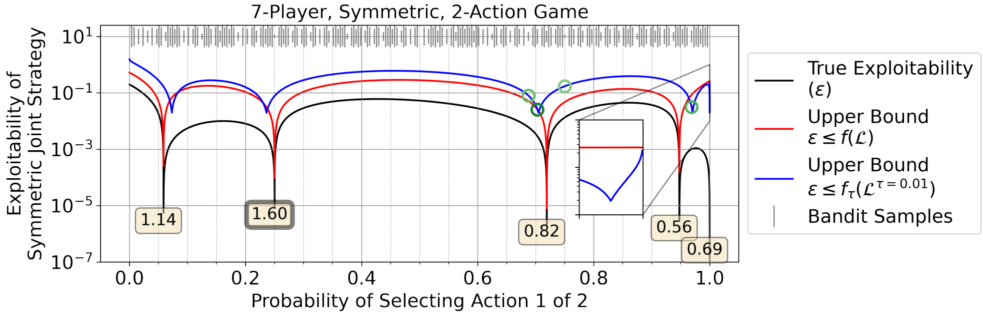

Figure 4: Bandit-based (BLiN) Nash solver applied to an artificial 7-player, symmetric, 2-action game. We search for a symmetric equilibrium, which is represented succinctly as the probability of selecting action 1. The plot shows the true exploitability ϵ of all symmetric strategies in black and indicates there exist potentially 5 NEs (the dips in the curve). Upper bounds on our unregularized loss $\mathcal{L}$ capture 4 of these equilibria, missing only the pure NE on the right. By considering our regularized loss, $\mathcal{L}^\tau$, we are able to capture this pure NE (see zoomed inset). The bandit algorithm selects strategies to evaluate, using 10 Monte-Carlo samples for each evaluation (arm pull) of $\mathcal{L}^\tau$. These samples are displayed as vertical bars above with the height of the vertical bar representing additional arm pulls. The best arms throughout search are denoted by green circles (darker indicates later in the search). The boxed numbers near equilibria display the welfare of the strategy.

infinity norm of the gradient (Corollary 1), and the number of distinct strategies ($n\bar{m} = \sum_k m_k$). This result also depends on the *near-optimality* or *zooming*-dimension d_z and zooming constant C_z which quantify the number of near optimal states. In particular, we assume $\mathcal{L}(s(z))$ is locally $(\sigma_{-\infty})$-strongly convex with respect to $||\cdot||_\infty$ about each global optimum within a ball of radius r_η. Here, $s : [0,1]^{n(\bar{m}-1)} \rightarrow \prod_i \Delta^{m_i - 1}$ is any function that maps from the unit hypercube to a product of simplices; we analyze two such maps in the appendix. Next, we present an additional convergence rate result using an alternative $\mathcal{X}$-bandit approach, StoSOO (Valko et al., 2013).

Theorem 5 (StoSOO Rate)**.** *Corollary* 1 *of Valko et al. (2013) implies that with probability* $(1 - w^{-1})(1-\delta)$ *for any* $w > 0$*, a uniformly randomly drawn arm (i.e.,* $t \sim U([T])$*) achieves*

$$\epsilon_t \leq w\Big[\frac{n}{\ln(1/p)}(W(1/e) + \frac{\bar{m}-2}{e}) + \sqrt{2nm^*\hat{L}}\sqrt{\xi_1\sqrt{\frac{\log_b(Tk/\delta)}{2\log_b(e)k}} + \xi_2 b^{-\frac{1}{dC}\sqrt{T/k}}}\Big] \tag{13}$$

where $d = n(\bar{m}-1)$, $\xi_1 = (2+2^{2/d})$, $\xi_2 = \frac{1}{4}db^{2(1+2/d)}$, $k = T\log_b(T)^{-3}$, *b is the branching factor for partitioning cells, and the near-optimality constant* $C = |\mathcal{X}^*|^{-1}\sqrt{2\pi d}\Big(\frac{b^2d^2}{5r_\eta^2\sigma_{-2}}\Big)^{d/2}$ *(Lemma 39).*

Here we assume $\mathcal{L}(s(z))$ is locally (σ_{-2})-strongly convex with respect to $||\cdot||_2$ about each global optimum within a ball of radius r_η. Theorem 5 implies a $\tilde{\mathcal{O}}(T^{-1/4})$ global convergence rate (Proposition 2), however this is achieved only after an exponential-length burn in time.

7 CONCLUSION

In this work, we proposed a stochastic loss for approximate Nash equilibria in normal-form games. An unbiased loss estimator of Nash equilibria is the "key" to the stochastic optimization "door" which holds a wealth of research innovations uncovered over several decades. Thus, it allows the development of new algorithmic techniques for computing equilibria. We consider bandit and vanilla SGD methods in this work, but theses are only two of the many options now at our disposal (e.g, adaptive methods (Antonakopoulos et al., 2022), Gaussian processes (Calandriello et al., 2022), evolutionary algorithms (Hansen et al., 2003), etc.). Such approaches as well as generalizations of these techniques to extensive-form, imperfect-information games are promising directions for future work. Similarly to how deep learning research first balked at and then marched on to train neural networks via NP-hard non-convex optimization, we hope computational game theory can march ahead to make useful equilibrium predictions of large multiplayer systems.

REFERENCES

K. Antonakopoulos, P. Mertikopoulos, G. Piliouras, and X. Wang. AdaGrad avoids saddle points. In *International Conference on Machine Learning*, pages 731–771. PMLR, 2022.

A. Arad and A. Rubinstein. Multi-dimensional iterative reasoning in action: The case of the Colonel Blotto game. *Journal of Economic Behavior & Organization*, 84(2):571–585, 2012.

P. Austrin, M. Braverman, and E. Chlamtáč. Inapproximability of NP-complete variants of Nash equilibrium. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques: 14th International Workshop, APPROX 2011, and 15th International Workshop, RANDOM 2011, Princeton, NJ, USA, August 17-19, 2011. Proceedings*, pages 13–25. Springer, 2011.

Y. Babichenko. Query complexity of approximate Nash equilibria. *Journal of the ACM (JACM)*, 63 (4):36:1–36:24, 2016.

D. Barrett and B. Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2020.

P. L. Bartlett, V. Gabillon, and M. Valko. A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption. In *Algorithmic Learning Theory*, pages 184–206. PMLR, 2019.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

K. Berg and T. Sandholm. Exclusion method for finding Nash equilibrium in multiplayer games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. $\mathcal{X}$-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

D. Calandriello, L. Carratino, A. Lazaric, M. Valko, and L. Rosasco. Scaling Gaussian process optimization by evaluating a few unique candidates multiple times. In *International Conference on Machine Learning*, pages 2523–2541. PMLR, 2022.

A. Cutkosky, H. Mehta, and F. Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. *arXiv preprint arXiv:2302.03775*, 2023.

C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou. The complexity of computing a Nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009.

Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014.

A. Deligkas, J. Fearnley, A. Hollender, and T. Melissourgos. Pure-circuit: Strong inapproximability for PPAD. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 159–170. IEEE, 2022.

Y. Feng, Z. Huang, and T. Wang. Lipschitz bandits with batched feedback. *Advances in Neural Information Processing Systems*, 35:19836–19848, 2022.

B. Gao and L. Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

I. Gemp, R. Savani, M. Lanctot, Y. Bachrach, T. Anthony, R. Everett, A. Tacchetti, T. Eccles, and J. Kramár. Sample-based approximation of Nash in large many-player games via gradient descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 507–515, 2022.

B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley. KKT conditions, first-order and second-order optimization, and distributed optimization: tutorial and survey. *arXiv preprint arXiv:2110.01858*, 2021.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

N. Hansen, S. D. Müller, and P. Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary computation*, 11(1): 1–18, 2003.

J. C. Harsanyi, R. Selten, et al. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.

S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

E. Hazan, K. Singh, and C. Zhang. Efficient regret minimization in non-convex games. In *International Conference on Machine Learning*, pages 1433–1441. PMLR, 2017.

E. Janovskaja. Equilibrium points in polymatrix games. *Lithuanian Mathematical Journal*, 8(2): 381–384, 1968.

M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and T. Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4190–4203, 2017.

M. Lanctot, E. Lockhart, J.-B. Lespiau, V. Zambaldi, S. Upadhyay, J. Pérolat, S. Srinivasan, F. Timbers, K. Tuyls, S. Omidshafiei, D. Hennes, D. Morrill, P. Muller, T. Ewalds, R. Faulkner, J. Kramár, B. D. Vylder, B. Saeta, J. Bradbury, D. Ding, S. Borgeaud, M. Lai, J. Schrittwieser, T. Anthony, E. Hughes, I. Danihelka, and J. Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *CoRR*, abs/1908.09453, 2019. URL `http://arxiv.org/abs/1908.09453`.

S. Leonardos, G. Piliouras, and K. Spendlove. Exploration-exploitation in multi-agent competition: convergence with bounded rationality. *Advances in Neural Information Processing Systems*, 34: 26318–26331, 2021.

W. Li, H. Li, J. Honorio, and Q. Song. Pyxab – a python library for $\mathcal{X}$-armed bandit and online blackbox optimization algorithms, 2023. URL `https://arxiv.org/abs/2303.04030`.

C. K. Ling, F. Fang, and J. Z. Kolter. What game are we playing? end-to-end learning in normal and extensive form games. *arXiv preprint arXiv:1805.02777*, 2018.

J. Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791. PMLR, 2013.

S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18:1–35, 2017.

L. Marris, I. Gemp, and G. Piliouras. Equilibrium-invariant embedding, metric space, and fundamental set of 2x2 normal-form games. *arXiv preprint arXiv:2304.09978*, 2023.

R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

R. D. McKelvey, A. M. McLennan, and T. L. Turocy. Gambit: Software tools for game theory, version 16.0.1, 2016.

D. Milec, J. Černỳ, V. Lisỳ, and B. An. Complexity and algorithms for exploiting quantal opponents in large two-player games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6): 5575–5583, 2021.

P. R. Milgrom and R. J. Weber. A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society*, pages 1089–1122, 1982.

K. G. Murty and S. N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. Technical report, 1985.

H. Nikaidô and K. Isoda. Note on non-cooperative convex games. *Pacific Journal of Mathematics*, 5 (1):807815, 1955.

E. Nudelman, J. Wortman, Y. Shoham, and K. Leyton-Brown. Run the GAMUT: A comprehensive approach to evaluating game-theoretic algorithms. In *AAMAS*, volume 4, pages 880–887, 2004.

G. Ostrovski and S. van Strien. Payoff performance of fictitious play. *arXiv preprint arXiv:1308.4049*, 2013.

J. Pérolat, S. Perrin, R. Elie, M. Laurière, G. Piliouras, M. Geist, K. Tuyls, and O. Pietquin. Scaling mean field games by online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 2022.

T. Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9 (129-145):20, 1935.

A. Raghunathan, A. Cherian, and D. Jha. Game theoretic optimization via gradient-based Nikaido-Isoda function. In *International Conference on Machine Learning*, pages 5291–5300. PMLR, 2019.

S. Shalev-Shwartz and Y. Singer. Convex repeated games and Fenchel duality. *Advances in neural information processing systems*, 19, 2006.

S. Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Y. Shoham and K. Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

M. Valko, A. Carpentier, and R. Munos. Stochastic simultaneous optimistic optimization. In *International Conference on Machine Learning*, pages 19–27. PMLR, 2013.

B. Wiedenbeck and E. Brinkman. Data structures for deviation payoffs. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, 2023.

Y. Zhou, J. Li, and J. Zhu. Identify the Nash equilibrium in static games with random payoffs. In *International Conference on Machine Learning*, pages 4160–4169. PMLR, 2017.