Re-evaluating Evaluation

David Balduzzi[*]   Karl Tuyls[*]   Julien Perolat[*]   Thore Graepel[*]

Abstract

"What we observe is not nature itself, but nature exposed to our method of questioning." – Werner Heisenberg

Progress in machine learning is measured by careful evaluation on problems of outstanding common interest. However, the proliferation of benchmark suites and environments, adversarial attacks, and other complications has diluted the basic evaluation model by overwhelming researchers with choices. Deliberate or accidental cherry picking is increasingly likely, and designing well-balanced evaluation suites requires increasing effort. In this paper we take a step back and propose *Nash averaging*. The approach builds on a detailed analysis of the *algebraic structure of evaluation* in two basic scenarios: agent-vs-agent and agent-vs-task. The key strength of Nash averaging is that it automatically adapts to redundancies in evaluation data, so that results are not biased by the incorporation of easy tasks or weak agents. Nash averaging thus encourages maximally inclusive evaluation – since there is no harm (computational cost aside) from including all available tasks and agents.

1 Introduction

Evaluation is a key driver of progress in machine learning, with e.g. ImageNet [1] and the Arcade Learning Environment [2] enabling subsequent breakthroughs in supervised and reinforcement learning [3, 4]. However, developing evaluations has received little *systematic* attention compared to developing algorithms. Immense amounts of compute is continually expended smashing algorithms and tasks together – *but the results are almost never used to **evaluate and optimize evaluations***. In a striking asymmetry, results are almost exclusively applied to evaluate and optimize algorithms.

The classic train-and-test paradigm on common datasets, which has served the community well [5], is reaching its limits. Three examples suffice. Adversarial attacks have complicated evaluation, raising questions about which attacks to test against [6–9]. Training agents far beyond human performance with self-play means they can only really be evaluated against each other [10, 11]. The desire to build increasingly general-purpose agents has led to a proliferation of environments: Mujoco, DM Lab, Open AI Gym, Psychlab and others [12–15].

In this paper we pause to ask, and partially answer, some basic questions about evaluation: **Q1.** Do tasks test what we think they test? **Q2.** When is a task redundant? **Q3.** Which tasks (and agents) matter the most? **Q4.** How should evaluations be evaluated?

We consider two scenarios: *agent vs task* (AvT), where algorithms are evaluated on suites of datasets or environments; and *agent vs agent* (AvA), where agents compete directly as in Go and Starcraft. Our goal is to treat tasks and agents symmetrically – with a view towards, ultimately, co-optimizing agents and evaluations. From this perspective AvA, where the task is (beating) another agent, is especially interesting. Performance in AvA is often quantified using Elo ratings [16] or the closely related TrueSkill [17]. There are two main problems with Elo. Firstly, Elo bakes-in the assumption that relative skill is transitive; but Elo is meaningless – it has no predictive power – in cyclic games

---

[*]DeepMind. Email: { dbalduzzi | karltuyls | perolat | thore }@google.com

like rock-paper-scissors. Intransitivity has been linked to biodiversity in ecology, and may be useful when evolving populations of agents [18–21]. Secondly, an agent's Elo rating can be inflated by instantiating many copies of an agent it beats (or conversely). This can cause problems when Elo guides hyper-optimization methods like population-based training [22]. Similarly, the most important decision when constructing a task-suite is which tasks to include. It is easy, and all too common, to bias task-suites in favor of particular agents or algorithms.

1.1 Overview

Section 2 presents background information on Elo and tools for working with antisymmetric matrices, such as the Schur decomposition and combinatorial Hodge theory. A major theme underlying the paper is that the fundamental algebraic structure of tournaments and evaluation is antisymmetric [23]. Techniques specific to antisymmetric matrices are less familiar to the machine learning community than approaches like PCA that apply to symmetric matrices and are typically correlation-based.

Section 3 presents a unified approach to representing evaluation data, where agents and tasks are treated symmetrically. A basic application of the approach results in our first contribution: a **multidimensional Elo rating** (mElo) that handles cyclic interactions. We also sketch how the Schur decomposition can uncover latent skills and tasks, providing a partial answer to **Q1**. We illustrate mElo on the domain of training an AlphaGo agent [24].

The second contribution of the paper is **Nash averaging**, an evaluation method that is **invariant** to redundant tasks and agents, see section 4. The basic idea is to play a meta-game on evaluation data [25]. The meta-game has a unique maximum entropy Nash equilibrium. The key insight of the paper is that the maxent Nash adapts automatically to the presence of redundant tasks and agents. The maxent Nash distribution thus provides a principled answer to **Q2** and **Q3**: which tasks and agents do and do not matter is determined by a meta-game. Finally, expected difficulty of tasks under the Nash distribution on agents yields a partial answer to **Q4**. The paper concludes by taking a second look at the performance of agents on Atari. We find that, under Nash averaging, human performance ties with the best agents, suggesting better-than-human performance has not yet been achieved.

1.2 Related work

Legg and Hutter developed a definition of intelligence which, informally, states "intelligence measures an agent's ability to achieve goals in a wide range of environments" [26, 27]. Formally, they consider all computable tasks weighted by algorithmic complexity [28–30]. Besides being incomputable, the distribution places (perhaps overly) heavy weight on the simplest tasks.

A comprehensive study of performance metrics for machine learning and AI can be found in [31–35]. There is a long history of psychometric evaluation in humans, some of which has been applied in artificial intelligence [36–38]. Bradley-Terry models provide a general framework for pairwise comparison [39]. Researchers have recently taken a second look at the arcade learning environment [2] and introduced new performance metrics [40]. However, the approach is quite particular. Recent work using agents to evaluate games has somewhat overlapping motivation with this paper [41–45]. Item response theory is an alternative, and likely complementary, approach to using agents to evaluate tasks [46] that has recently been applied to study the performance of agents on the Arcade Learning Environment [47].

Our approach draws heavily on work applying combinatorial Hodge theory to statistical ranking [48] and game theory [49–51]. We also draw on empirical game theory [52, 53], by using a meta-game to "evaluate evaluations", see section 4. Empirical game theory has been applied to domains like poker and continuous double auctions, and has recently been extended to asymmetric games [54–58]. *von Neumann winners* in the dueling bandit setting and NE-regret are related to Nash averaging [59–62].

2 Preliminaries

Notation. Vectors are column vectors. $\mathbf{0}$ and $\mathbf{1}$ denote the constant vectors of zeros and ones respectively. We sometimes use subscripts to indicate dimensions of vectors and matrices, e.g. $\mathbf{r}_{n \times 1}$ or $\mathbf{S}_{m \times n}$ and sometimes their entries, e.g. $\mathbf{r}_i$ or $\mathbf{S}_{ij}$; no confusion should result. The unit vector with a 1 in coordinate i is $\mathbf{e}_i$. Proofs and code are in the appendix.

2.1 The Elo rating system

Suppose n agents play a series of pairwise matches against each other. Elo assigns a rating r_i to each player $i \in [n]$ based on their wins and losses, which we represent as an n-vector $\mathbf{r}$. The predicted probability of i beating j given their Elo ratings is

$$\hat{p}_{ij} := \frac{10^{r_i/400}}{10^{r_i/400} + 10^{r_j/400}} = \sigma(\alpha r_i - \alpha r_j), \quad \text{where} \quad \sigma(x) = \frac{1}{1 + e^{-x}} \quad \text{and} \quad \alpha = \frac{\log(10)}{400}.$$

The constant α is not important in what follows, so we pretend $\alpha = 1$. Observe that only the *difference* between Elo ratings affects win-loss predictions. We therefore impose that Elo ratings sum to zero, $\mathbf{r}^\mathsf{T}\mathbf{1} = 0$, without loss of generality. Define the loss,

$$\ell_{\text{Elo}}(p_{ij}, \hat{p}_{ij}) = -p_{ij} \log \hat{p}_{ij} - (1 - p_{ij}) \log(1 - \hat{p}_{ij}), \quad \text{where} \quad \hat{p}_{ij} = \sigma(r_i - r_j)$$

and p_{ij} is the true probability of i beating j. Suppose the tth match pits player i against j, with outcome $S_{ij}^t = 1$ if i wins and $S_{ij}^t = 0$ if i loses. Online gradient descent on ℓ_{Elo} obtains

$$r_i^{t+1} \leftarrow r_i^t - \eta \cdot \nabla_{r_i} \ell_{\text{Elo}}(S_{ij}^t, \hat{p}_{ij}^t) = r_i^t + \eta \cdot (S_{ij}^t - \hat{p}_{ij}^t).$$

Choosing learning rate $\eta = 16$ or 32 recovers the updates introduced by Arpad Elo in [16].

The win-loss probabilities predicted by Elo ratings can fail in simple cases. For example, rock, paper and scissors will all receive the same Elo ratings. Elo's predictions are $\hat{p}_{ij} = \frac{1}{2}$ for all i, j – and so Elo has no predictive power for any given pair of players (e.g. paper beats rock with $p = 1$).

What are the Elo update's fixed points? Suppose we batch matches to obtain empirical estimates of the probability of player i beating j: $\bar{p}_{ij} = \sum_n \frac{S_{ij}^n}{N_{ij}}$. As the number of matches approaches infinity, the empirical estimates approach the true probabilities p_{ij}.

Proposition 1. *Elo ratings are at a stationary point under batch updates iff the matrices of empirical probabilities and predicted probabilities have the same row-sums (or, equivalently the same column-sums):*

$$\nabla_{r_i} \left[\sum_j \ell_{\text{Elo}}(\bar{p}_{ij}, \hat{p}_{ij}) \right] = 0 \ \forall i \quad \textit{iff} \quad \sum_j \bar{p}_{ij} = \sum_j \hat{p}_{ij} \ \forall i.$$

Many different win-loss probability matrices result in identical Elo ratings. The situation is analogous to how many different joint probability distributions can have the same marginals. We return to this topic in section 3.1.

2.2 Antisymmetric matrices

We recall some basic facts about antisymmetric matrices. Matrix $\mathbf{A}$ is antisymmetric if $\mathbf{A} + \mathbf{A}^\mathsf{T} = 0$. Antisymmetric matrices have even rank and *imaginary* eigenvalues $\{\pm i\lambda_j\}_{j=1}^{\text{rank}(\mathbf{A})/2}$. Any antisymmetric matrix $\mathbf{A}$ admits a real **Schur decomposition**:

$$\mathbf{A}_{n \times n} = \mathbf{Q}_{n \times n} \cdot \mathbf{\Lambda}_{n \times n} \cdot \mathbf{Q}_{n \times n}^\mathsf{T},$$

where $\mathbf{Q}$ is orthogonal and $\mathbf{\Lambda}$ consists of zeros except for (2×2) diagonal-blocks of the form:

$$\mathbf{\Lambda} = \begin{pmatrix} 0 & \lambda_j \\ -\lambda_j & 0 \end{pmatrix}.$$

The entries of $\mathbf{\Lambda}$ are *real* numbers, found by multiplying the eigenvalues of $\mathbf{A}$ by $i = \sqrt{-1}$.

Proposition 2. *Given matrix $\mathbf{S}_{m \times n}$ with rank r and singular value decomposition $\mathbf{U}_{m \times r} \mathbf{D}_{r \times r} \mathbf{V}_{r \times n}^\mathsf{T}$. Construct antisymmetric matrix*

$$\mathbf{A}_{(m+n) \times (m+n)} = \begin{pmatrix} \mathbf{0}_{m \times m} & \mathbf{S}_{m \times n} \\ -\mathbf{S}_{n \times m}^\mathsf{T} & \mathbf{0}_{n \times n} \end{pmatrix}.$$

Then the thin Schur decomposition of $\mathbf{A}$ is $\mathbf{Q}_{(m+n) \times 2r} \mathbf{\Lambda}_{2r \times 2r} \mathbf{Q}_{2r \times (m+n)}^\mathsf{T}$ where the nonzero pairs in $\mathbf{\Lambda}_{2r \times 2r}$ are $\pm$ the singular values in $\mathbf{D}_{r \times r}$ and

$$\mathbf{Q}_{(m+n) \times 2r} = \begin{pmatrix} -\mathbf{u}_1 & \mathbf{0}_{m \times 1} & \cdots & -\mathbf{u}_r & \mathbf{0}_{m \times 1} \\ \mathbf{0}_{n \times 1} & \mathbf{v}_1 & \cdots & \mathbf{0}_{n \times 1} & \mathbf{v}_r \end{pmatrix}.$$

3

Combinatorial Hodge theory is developed by analogy with differential geometry, see [48–51]. Consider a fully connected graph with vertex set $[n] = \{1, \ldots, n\}$. Assign a **flow** $\mathbf{A}_{ij} \in \mathbb{R}$ to each edge of the graph. The flow in the opposite direction ji is $\mathbf{A}_{ji} = -\mathbf{A}_{ij}$, so flows are just $(n \times n)$ antisymmetric matrices. The flow on a graph is analogous to a vector field on a manifold.

The combinatorial **gradient** of an n-vector $\mathbf{r}$ is the flow: $\mathrm{grad}(\mathbf{r}) := \mathbf{r}\mathbf{1}^\mathsf{T} - \mathbf{1}\mathbf{r}^\mathsf{T}$. Flow $\mathbf{A}$ is a **gradient flow** if $\mathbf{A} = \mathrm{grad}(\mathbf{r})$ for some $\mathbf{r}$, or equivalently if $\mathbf{A}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ for all i, j. The **divergence** of a flow is the n-vector $\mathrm{div}(\mathbf{A}) := \frac{1}{n}\mathbf{A} \cdot \mathbf{1}$. The divergence measures the contribution to the flow of each vertex, considered as a source. The **curl** of a flow is the three-tensor $\mathrm{curl}(\mathbf{A})_{ijk} = \mathbf{A}_{ij} + \mathbf{A}_{jk} - \mathbf{A}_{ik}$. Finally, the **rotation** is $\mathrm{rot}(\mathbf{A})_{ij} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{curl}(\mathbf{A})_{ijk}$.

Theorem (Hodge decomposition, [48]). *(i)* $\mathrm{div} \circ \mathrm{grad}(\mathbf{r}) = \mathbf{r}$ *for any* $\mathbf{r}$ *satisfying* $\mathbf{r}^\mathsf{T}\mathbf{1} = 0$.
(ii) $\mathrm{div} \circ \mathrm{rot}(\mathbf{A}) = \mathbf{0}_{n \times 1}$ *for any flow* $\mathbf{A}$. *(iii)* $\mathrm{rot} \circ \mathrm{grad}(\mathbf{r}) = \mathbf{0}_{n \times n}$ *for any vector* $\mathbf{r}$.
(iv) The vector space of antisymmetric matrices admits an orthogonal decomposition

$$\{\textit{flows}\} = \{\textit{antisymmetric matrices}\} = \mathrm{im}(\mathrm{grad}) \oplus \mathrm{im}(\mathrm{rot})$$

with respect to the standard inner product $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{ij} \mathbf{A}_{ij}\mathbf{B}_{ij}$. *Concretely, any antisymmetric matrix decomposes as*

$$\mathbf{A} = \{\textit{transitive component}\} + \{\textit{cyclic component}\} = \mathrm{grad}(\mathbf{r}) + \mathrm{rot}(\mathbf{A}) \quad \textit{where} \quad \mathbf{r} = \mathrm{div}(\mathbf{A}).$$

Sneak peak. The divergence recovers Elo ratings or just plain average performance depending on the scenario. The Hodge decomposition separates transitive (captured by averages or Elo) from cyclic interactions (rock-paper-scissors), and explains when Elo ratings make sense. The Schur decomposition is a window into the latent skills and tasks not accounted for by Elo and averages.

3 On the algebraic structure of evaluation

The Schur decomposition and combinatorial Hodge theory provide a unified framework for analyzing evaluation data in the AvA and AvT scenarios. In this section we provide some basic tools and present a multidimensional extension of Elo that handles cyclic interactions.

3.1 Agents vs agents (AvA)

In AvA, results are collated into a matrix of win-loss probabilities based on relative frequencies. Construct $\mathbf{A} = \mathrm{logit}\,\mathbf{P}$ with $\mathbf{A}_{ij} := \log \frac{p_{ij}}{1 - p_{ij}}$. Matrix $\mathbf{A}$ is antisymmetric since $p_{ij} + p_{ji} = 1$.

When can Elo correctly predict win-loss probabilities? The answer is simple in logit space:

Proposition 3. *(i) If probabilities* $\mathbf{P}$ *are generated by Elo ratings* $\mathbf{r}$ *then the divergence of its logit is* $\mathbf{r}$. *That is,*

$$\textit{if } p_{ij} = \sigma(r_i - r_j) \; \forall i, j \textit{ then } \mathrm{div}(\mathrm{logit}\,\mathbf{P}) = \left(\frac{1}{n}\sum_{j=1}^{n}(r_i - r_j)\right)_{i=1}^{n} = \mathbf{r}.$$

(ii) There is an Elo rating that generates probabilities $\mathbf{P}$ *iff* $\mathrm{curl}(\mathrm{logit}\,\mathbf{P}) = 0$. *Equivalently, iff* $\log \frac{p_{ij}}{p_{ji}} + \log \frac{p_{jk}}{p_{kj}} + \log \frac{p_{ki}}{p_{ik}} = 0$ *for all* i, j, k.

Elo is, essentially, a uniform average in logit space. Elo's predictive failures are due to the cyclic component $\tilde{\mathbf{A}} := \mathrm{rot}(\mathrm{logit}\,\mathbf{P})$ that uniform averaging ignores.

Multidimensional Elo (mElo$_{2k}$). Elo ratings bake-in the assumption that relative skill is transitive. However, there is no single dominant strategy in games like rock-paper-scissors or (arguably) StarCraft. Rating systems that can handle intransitive abilities are therefore necessary. An obvious approach is to learn a feature vector $\mathbf{w}$ and a rating vector $\mathbf{r}_i$ per player, and predict $\hat{p}_{ij} = \sigma(\mathbf{r}_i^\mathsf{T}\mathbf{w} - \mathbf{r}_j^\mathsf{T}\mathbf{w})$. Unfortunately, this reduces to the standard Elo rating since $\mathbf{r}_i^\mathsf{T}\mathbf{w}$ is a scalar.

Handling intransitive abilities requires learning an approximation to the cyclic component $\tilde{\mathbf{A}}$. Combining the Schur and Hodge decompositions allows to construct low-rank approximations that extend Elo. Note, antisymmetric matrices have *even* rank. Consider

$$\mathbf{A}_{n \times n} = \mathrm{grad}(\mathbf{r}) + \tilde{\mathbf{A}} \approx \mathrm{grad}(\mathbf{r}) + \mathbf{C}^\mathsf{T}\begin{pmatrix} 0 & 1 & \\ -1 & 0 & \\ & & \ddots \end{pmatrix}\mathbf{C} =: \mathrm{grad}(\mathbf{r}) + \mathbf{C}_{n \times 2k}^\mathsf{T}\mathbf{\Omega}_{2k \times 2k}\mathbf{C}_{2k \times n}$$

4

where the rows of $\mathbf{C}$ are orthogonal to each other, to $\mathbf{r}$, and to $\mathbf{1}$. The larger $2k$, the better the approximation. Let $\mathbf{mElo}_{2k}$ assign each player Elo rating r_i and $2k$-dimensional vector $\mathbf{c}_i$. Vanilla Elo uses $2k = 0$. The mElo_{2k} win-loss prediction is

$$\mathbf{mElo}_{2k}\text{:} \ \hat{p}_{ij} = \sigma\left(r_i - r_j + \mathbf{c}_i^\intercal \cdot \mathbf{\Omega}_{2k \times 2k} \cdot \mathbf{c}_j\right) \text{ where } \mathbf{\Omega}_{2k \times 2k} = \sum_{i=1}^{k}(\mathbf{e}_{2i-1}\mathbf{e}_{2i}^\intercal - \mathbf{e}_{2i}\mathbf{e}_{2i-1}^\intercal).$$

Online updates can be computed by gradient descent, see section E, with orthogonality enforced.

3.2 Application: predicting win-loss probabilities in Go

Elo ratings are widely used in Chess and Go. We compared the predictive capabilities of Elo and the simplest extension mElo_2 on eight Go algorithms taken from extended data table 9 in [24]: seven variants of AlphaGo, and Zen. The Frobenius norms and logistic losses are $\|\mathbf{P} - \hat{\mathbf{P}}\|_F = 0.85$ and $\ell_{\log} = 1.41$ for Elo vs the empirical probabilities and $\|\mathbf{P} - \hat{\mathbf{P}}_2\|_F = 0.35$ and $\ell_{\log} = 1.27$ for mElo_2.

To better understand the difference, we zoom in on three algorithms that were observed to interact non-transitively in [58]: α_v with value net, α_p with policy net, and Zen. Elo's win-loss predictions are poor (Table **Elo**: Elo incorrectly predicts both that α_p likely beats α_v and α_v likely beats Zen), whereas mElo_2 (Table **mElo$_2$**) correctly predicts likely winners in all cases (Table **empirical**), with more accurate probabilities:

Elo	α_v	α_p	Zen
α_v	-	0.41	0.58
α_p	0.59	-	0.67
Zen	0.42	0.33	-

empirical	α_v	α_p	Zen
α_v	-	0.7	0.4
α_p	0.3	-	1.0
Zen	0.6	0.0	-

mElo$_2$	α_v	α_p	Zen
α_v	-	0.72	0.46
α_p	0.28	-	0.98
Zen	0.55	0.02	-

3.3 Agents vs tasks (AvT)

In AvT, results are represented as an $(m \times n)$ matrix $\mathbf{S}$: rows are agents, columns are tasks, entries are scores (e.g. accuracy or total reward). Subtract the total mean, so the sum of all entries of $\mathbf{S}$ is zero. We recast both agents and tasks as *players* and construct an antisymmetric $(m+n) \times (m+n)$-matrix. Let $\mathbf{s} = \frac{1}{m}\mathbf{S} \cdot \mathbf{1}_{m \times 1}$ and $\mathbf{d} = -\frac{1}{n}\mathbf{S}^\intercal \cdot \mathbf{1}_{n \times 1}$ be the **average skill of each agent** and the **average difficulty of each task**. Define $\tilde{\mathbf{S}} = \mathbf{S} - (\mathbf{s} \cdot \mathbf{1}^\intercal - \mathbf{1} \cdot \mathbf{d}^\intercal)$. Let $\mathbf{r}$ be the concatenation of $\mathbf{s}$ and $\mathbf{d}$. We construct the antisymmetric matrix

$$\mathbf{A}_{(m+n) \times (m+n)} = \mathrm{grad}(\mathbf{r}) + \underbrace{\begin{pmatrix} \mathbf{0}_{m \times m} & \tilde{\mathbf{S}}_{m \times n} \\ -\tilde{\mathbf{S}}_{n \times m}^\intercal & \mathbf{0}_{n \times n} \end{pmatrix}}_{\tilde{\mathbf{A}}} = \begin{pmatrix} \mathrm{grad}(\mathbf{s}) & \mathbf{S}_{m \times n} \\ -\mathbf{S}_{n \times m}^\intercal & \mathrm{grad}(\mathbf{d}) \end{pmatrix}.$$

The top-right block of $\mathbf{A}$ is agent performance on tasks; the bottom-left is task difficulty for agents. The top-left block compares agents by their average skill on tasks; the bottom-right compares tasks by their average difficulty for agents. Average skill and difficulty explain the data if the score of agent i on task j is $\mathbf{S}_{ij} = \mathbf{s}_i - \mathbf{d}_j$, the agent's skill minus the task's difficulty, for all i, j. Paralleling proposition 3, averages explain the data, $\mathbf{S} = \mathbf{s}\mathbf{1}^\intercal - \mathbf{1}\mathbf{d}^\intercal$, iff $\mathrm{curl}(\mathbf{A}) = \mathbf{0}$.

The failure of averages to explain performance is encapsulated in $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{A}}$. By proposition 2, the SVD of $\tilde{\mathbf{S}}$ and Schur decomposition of $\tilde{\mathbf{A}}$ are equivalent. If the SVD is $\tilde{\mathbf{S}}_{m \times n} = \mathbf{U}_{m \times r}\mathbf{D}_{r \times r}\mathbf{V}_{r \times n}^\intercal$ then the rows of $\mathbf{U}$ represent the latent abilities exhibited by agents and the rows of $\mathbf{V}$ represent the latent problems posed by tasks.

4 Invariant evaluation

Evaluation is often based on metrics like average performance or Elo ratings. Unfortunately, two (or two hundred) tasks or agents that look different may test/exhibit identical skills. Overrepresenting particular tasks or agents introduces biases into averages and Elo – biases that can only be detected *post hoc*. Humans must therefore decide which tasks or agents to retain, to prevent redundant agents or tasks from skewing results. At present, *evaluation is not automatic and does not scale.* To be scalable and automatic, an evaluation method should *always benefit* from including additional agents and tasks. Moreover, it should *adjust automatically and gracefully* to redundant data.

Definition 1. *An **evaluation method** maps data to a real-valued function on players (that is, agents or agents and tasks):*

$$\mathcal{E} : \big\{ \textit{evaluation data} \big\} = \big\{ \textit{antisymmetric matrices} \big\} \to \Big[\big\{ \textit{players} \big\} \to \mathbb{R} \Big].$$

Desired properties. An evaluation method should be:

P1. *Invariant:* adding redundant copies of an agent or task to the data should make no difference.

P2. *Continuous:* the evaluation method should be robust to small changes in the data.

P3. *Interpretable:* hard to formalize, but the procedure should agree with intuition in basic cases.

Elo and uniform averaging over tasks are examples of evaluation methods that invariance excludes.

4.1 Nash averaging

This section presents an evaluation method satisfying properties $P1, P2, P3$. We discuss AvA here, see section D for AvT. Given antisymmetric logit matrix $\mathbf{A}$, define a two-player meta-game with *payoffs* $\mu_1(\mathbf{p}, \mathbf{q}) = \mathbf{p}^\mathsf{T}\mathbf{A}\mathbf{q}$ and $\mu_2(\mathbf{p}, \mathbf{q}) = \mathbf{p}^\mathsf{T}\mathbf{B}\mathbf{q}$ for the row and column meta-players, where $\mathbf{B} = \mathbf{A}^\mathsf{T}$. The game is symmetric because $\mathbf{B} = \mathbf{A}^\mathsf{T}$ and zero-sum because $\mathbf{B} = -\mathbf{A}$.

The row and column meta-players pick 'teams' of agents. Their payoff is the expected log-odds of their respective team winning under the joint distribution. If there is a dominant agent that has better than even odds of beating the rest, both players will pick it. In rock-paper-scissors, the only unbeatable-on-average team is the uniform distribution. In general, the value of the game is zero and the Nash equilibria are teams that are unbeatable in expectation.

A problem with Nash equilibria (NE) is that they are not unique, which forces the user to make choices and undermines interpretability [63, 64]. Fortunately, for zero-sum games there is a natural choice of Nash:

Proposition 4 (maxent NE). *For antisymmetric $\mathbf{A}$ there is a* unique symmetric *Nash equilibrium* $(\mathbf{p}^*, \mathbf{p}^*)$ *solving* $\max_{\mathbf{p} \in \Delta_n} \min_{\mathbf{q} \in \Delta_n} \mathbf{p}^\mathsf{T}\mathbf{A}\mathbf{q}$ *with greater entropy than any other Nash equilibrium.*

Maxent Nash is maximally indifferent between players with the same empirical performance.

Definition 2. *The **maxent Nash evaluation method** for AvA is*

$$\mathcal{E}_m : \big\{ \textit{evaluation data} \big\} = \big\{ \textit{antisymmetric matrices} \big\} \xrightarrow{\textit{maxent NE}} \Big[\big\{ \textit{players} \big\} \xrightarrow{\textit{Nash average}} \mathbb{R} \Big],$$

*where $\mathbf{p}_\mathbf{A}^*$ is the **maxent Nash equilibrium** and $\mathbf{n}_\mathbf{A} := \mathbf{A} \cdot \mathbf{p}_\mathbf{A}^*$ is the **Nash average**.*

Invariance to redundancy is best understood by looking at an example; for details see section C.

Example 1 (invariance). Consider two logit matrices, where the second adds a redundant copy of agent C to the first:

$\mathbf{A}$	A	B	C
A	0.0	4.6	-4.6
B	-4.6	0.0	4.6
C	4.6	-4.6	0.0

and

$\mathbf{A}'$	A	B	C_1	C_2
A	0.0	4.6	-4.6	-4.6
B	-4.6	0.0	4.6	4.6
C_1	4.6	-4.6	0.0	0.0
C_2	4.6	-4.6	0.0	0.0

The maxent Nash for $\mathbf{A}$ is $\mathbf{p}_\mathbf{A}^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. It is easy to check that $(\frac{1}{3}, \frac{1}{3}, \frac{\alpha}{3}, \frac{1-\alpha}{3})$ is Nash for $\mathbf{A}'$ for any $\alpha \in [0, 1]$ and thus the maxent Nash for $\mathbf{A}'$ is $\mathbf{p}_{\mathbf{A}'}^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6})$. Maxent Nash automatically detects the redundant agents C_1, C_2 and distributes C's mass over them equally.

Uniform averaging is not invariant to adding redundant agents; concretely $\operatorname{div}(\mathbf{A}) = \mathbf{0}$ whereas $\operatorname{div}(\mathbf{A}') = (-1.15, 1.15, 0, 0)$, falsely suggesting agent B is superior. In contrast, $\mathbf{n}_\mathbf{A} = \mathbf{0}_{3 \times 1}$ and $\mathbf{n}_{\mathbf{A}'} = \mathbf{0}_{4 \times 1}$ (the zero-vectors have different sizes because there are different numbers of agents). Nash averaging correctly reports no agent is better than the rest in both cases.

Theorem 1 (main result for AvA[2]). *The maxent NE has the following properties:*

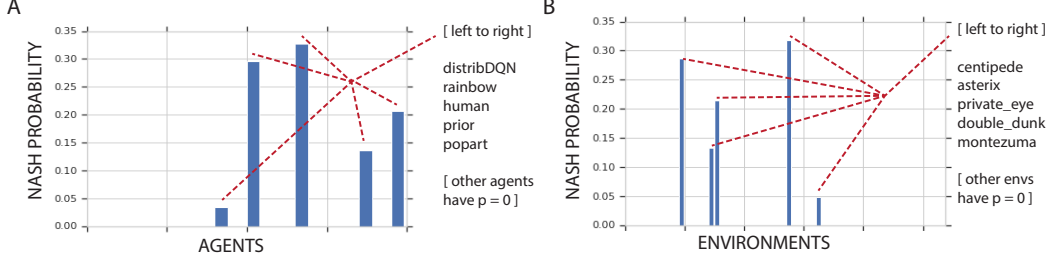[2]The main result for AvT is analogous, see section D.

Figure 1: (A) The Nash $\mathbf{p}_a^*$ assigned to agents; (B) the Nash $\mathbf{p}_e^*$ assigned to environments.

P1. **Invariant:** *Nash averaging, with respect to the maxent NE, is invariant to redundancies in* $\mathbf{A}$.

P2. **Continuous:** *If* $\mathbf{p}^*$ *is a Nash for* $\hat{\mathbf{A}}$ *and* $\epsilon = \|\mathbf{A} - \hat{\mathbf{A}}\|_{max}$ *then* $\mathbf{p}^*$ *is an* ϵ-*Nash for* $\mathbf{A}$.

P3. **Interpretable: (i)** *The maxent NE on* $\mathbf{A}$ *is the uniform distribution,* $\mathbf{p}^* = \frac{1}{n}\mathbf{1}$, *iff the meta-game is* cyclic, *i.e.* $\mathrm{div}(\mathbf{A}) = \mathbf{0}$. **(ii)** *If the meta-game is* transitive, *i.e.* $\mathbf{A} = \mathrm{grad}(\mathbf{r})$, *then the maxent NE is the uniform distribution on the player(s) with highest rating(s) – there could be a tie.*

See section C for proof and formal definitions. For interpretability, if $\mathbf{A} = \mathrm{grad}(\mathbf{r})$ then the transitive rating is all that matters: Nash averaging measures performance against the best player(s). If $\mathrm{div}(\mathbf{A}) = \mathbf{0}$ then no player is better than any other. Mixed cases cannot be described in closed form.

The continuity property is quite weak: theorem 1.2 shows the *payoff* is continuous: *a team that is unbeatable for* $\hat{\mathbf{A}}$ *is* ϵ-*beatable for nearby* $\mathbf{A}$. Unfortunately, Nash equilibria themselves can jump discontinuously when $\mathbf{A}$ is modified slightly. Perturbed best response converges to a more stable approximation to Nash [65, 66] that unfortunately is not invariant.

Example 2 (continuity). Consider the cyclic and transitive logit matrices

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{T} = \begin{pmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{pmatrix}.$$

The maxent Nash equilibria and Nash averages of $\mathbf{C} + \epsilon\mathbf{T}$ are

$$\mathbf{P}_{\mathbf{C}+\epsilon\mathbf{T}}^* = \begin{cases} \left(\frac{1+\epsilon}{3}, \frac{1-2\epsilon}{3}, \frac{1+\epsilon}{3}\right) & \text{if } 0 \leq \epsilon \leq \frac{1}{2} \\ (1,0,0) & \text{if } \frac{1}{2} < \epsilon \end{cases} \quad \text{and} \quad \mathbf{n}_{\mathbf{C}+\epsilon\mathbf{T}} = \begin{cases} (0,0,0) & 0 \leq \epsilon \leq \frac{1}{2} \\ (0, -1-\epsilon, 1-2\epsilon) & \frac{1}{2} < \epsilon \end{cases}$$

The maxent Nash is the uniform distribution over agents in the cyclic case ($\epsilon = 0$), and is concentrated on the first player when it dominates the others ($\epsilon > \frac{1}{2}$). When $0 < \epsilon < \frac{1}{2}$ the optimal team has most mass on the first and last players. Nash jumps discontinuously at $\epsilon = \frac{1}{2}$.

4.2 Application: re-evaluation of agents on the Arcade Learning Environment

To illustrate the method, we re-evaluate the performance of agents on Atari [2]. Data is taken from results published in [67–70]. Agents include `rainbow`, dueling networks, `prioritized` replay, `pop-art`, `DQN`, count-based exploration and baselines like `human`, random-action and `no-action`. The 20 agents evaluated on 54 environments are represented by matrix $\mathbf{S}_{20\times54}$. It is necessary to standardize units across environments with quite different reward structures: for each column we subtract the $\min$ and divide by the $\max$ so scores lie in $[0, 1]$.

We introduce a meta-game where row meta-player picks aims to pick the best distribution $\mathbf{p}_a^*$ on agents and column meta-player aims to pick the hardest distribution $\mathbf{p}_e^*$ on environments, see section D for details. We find a Nash equilibrium using an LP-solver; it should be possible to find the maxent Nash using the algorithm in [71, 72]. The Nash distributions are shown in figure 1. The supports of the distributions are the 'core agents' and the 'core environments' that form unexploitable teams. See appendix for tables containing all skills and difficulties. panel B.

Figure 2A shows the skill of agents under uniform $\frac{1}{n}\mathbf{S}\cdot\mathbf{1}$ and Nash $\mathbf{S}\cdot\mathbf{p}_e^*$ averaging over environments; panel B shows the difficulty of environments under uniform $-\frac{1}{m}\mathbf{S}^\intercal\cdot\mathbf{1}$ and Nash $-\mathbf{S}^\intercal\cdot\mathbf{p}_a^*$ averaging over agents. There is a tie for top between the agents with non-zero mass – including human. This follows by the indifference principle for Nash equilibria: strategies with support have equal payoff.
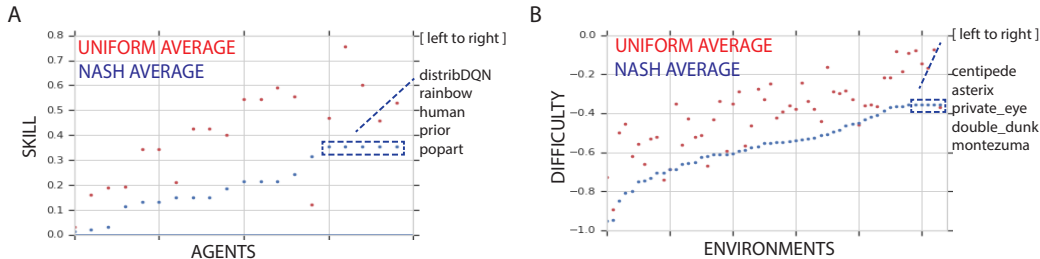
Figure 2: **Comparison of uniform and Nash averages.** (A) Skill of agents by uniform $\frac{1}{n}\mathbf{S} \cdot \mathbf{1}$ and Nash $\mathbf{S} \cdot \mathbf{p}_e^*$ averaging over environments. (B) Difficulty of environments under uniform $-\frac{1}{m}\mathbf{S}^\mathsf{T} \cdot \mathbf{1}$ and Nash $-\mathbf{S}^\mathsf{T} \cdot \mathbf{p}_a^*$ averaging over agents. Agents and environments are sorted by Nash-averages.

Our results suggest that the better-than-human performance observed on the Arcade Learning Environment is because ALE is skewed towards environments that (current) agents do well on, and contains fewer environments testing skills specific to humans. Solving the meta-game automatically finds a distribution on environments that evens out the playing field and, simultaneously, identifies the most important agents and environments.

5 Conclusion

A powerful guiding principle when deciding what to measure is to find quantities that are *invariant* to naturally occurring transformations. The determinant is computed over a basis – however, the determinant is *invariant to the choice of basis* since $\det(G^{-1}AG) = \det(A)$ for any invertible matrix G. Noether's theorem implies the dynamics of a physical system with symmetries obeys a conservation law. The speed of light is fundamental because it is invariant to the choice of inertial reference frame.

One must have symmetries in mind to talk about invariance. **What are the naturally occurring symmetries in machine learning?** The question admits many answers depending on the context, see e.g. [73–79]. In the context of evaluating agents, that are typically built from neural networks, it is unclear *a priori* whether two seemingly different agents – based on their parameters or hyperparameters – are *actually* different. Further, it is increasingly common that environments and tasks are parameterized – or are learning agents in their own right, see self-play [10, 11], adversarial attacks [6–9], and automated curricula [80]. The overwhelming source of symmetry when evaluating learning algorithms is therefore **redundancy**: different agents, networks, algorithms, environments and tasks that do basically the same job.

Nash evaluation computes a distribution on players (agents, or agents and tasks) that automatically adjusts to redundant data. It thus provides an invariant approach to measuring agent-agent and agent-environment interactions. In particular, Nash averaging encourages a maximally inclusive approach to evaluation: computational cost aside, the method should only benefit from including as many tasks and agents as possible. Easy tasks or poorly performing agents will not bias the results. As such Nash averaging is a significant step towards more *objective* evaluation.

Nash averaging is not always the right tool. Firstly, it is only as good as the data: *garbage in, garbage out*. Nash decides which environments are important based on the agents provided to it, and conversely. As a result, the method is blind to differences between environments that do not make a difference to agents and vice versa. Nash-based evaluation is likely to be most effective when applied to a diverse array of agents and environments. Secondly, for good or ill, Nash averaging removes control from the user. One may have good reason to disagree with the distribution chosen by Nash. Finally, Nash is a harsh master. It takes an adversarial perspective and may not be the best approach to, say, constructing automated curricula – although boosting is a related approach that works well [81, 82]. It is an open question whether alternate invariant evaluations can be constructed, game-theoretically or otherwise.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[2] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: An evaluation platform for general agents," *J. Artif. Intell. Res.*, vol. 47, pp. 253–279, 2013.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 02 2015.

[5] D. Donoho, "50 years of Data Science," in *Based on a presentation at the Tukey Centennial workshop*, 2015.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *arXiv:1312.6199*, 2013.

[7] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble Adversarial Training: Attacks and Defenses," in *ICLR*, 2018.

[8] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie, J. Wang, Z. Zhang, Z. Ren, A. Yuille, S. Huang, Y. Zhao, Y. Zhao, Z. Han, J. Long, Y. Berdibekov, T. Akiba, S. Tokui, and M. Abe, "Adversarial Attacks and Defences Competition," in *arXiv:1804.00097*, 2018.

[9] J. Uesato, B. O'Donoghue, A. van den Oord, and P. Kohli, "Adversarial Risk and the Dangers of Evaluating Against Weak Attacks," in *ICML*, 2018.

[10] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, pp. 354–359, 2017.

[11] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm," in *arXiv:1712.01815*, 2017.

[12] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *IROS*, 2012.

[13] C. Beattie, J. Z. Leibo, D. Teplyashin, T. Ward, M. Wainwright, H. Küttler, A. Lefrancq, S. Green, V. Valdés, A. Sadik, J. Schrittwieser, K. Anderson, S. York, M. Cant, A. Cain, A. Bolton, S. Gaffney, H. King, D. Hassabis, S. Legg, and S. Petersen, "DeepMind Lab," in *arXiv:1612.03801*, 2016.

[14] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI Gym," 2016.

[15] J. Z. Leibo, C. de Masson d'Autume, D. Zoran, D. Amos, C. Beattie, K. Anderson, A. G. Castañeda, M. Sanchez, S. Green, A. Gruslys, S. Legg, D. Hassabis, and M. M. Botvinick, "Psychlab: A Psychology Laboratory for Deep Reinforcement Learning Agents," in *arXiv:1801.08116*, 2018.

[16] A. E. Elo, *The Rating of Chess players, Past and Present*. Ishi Press International, 1978.

[17] R. Herbrich, T. Minka, and T. Graepel, "TrueSkill: a Bayesian skill rating system," in *NIPS*, 2007.

[18] M. Frean and E. R. Abraham, "Rock-scissors-paper and the survival of the weakest," *Proc. R. Soc. Lond. B*, no. 268, pp. 1323–1327, 2001.

[19] B. Kerr, M. A. Riley, M. W. Feldman, and B. J. M. Bohannan, "Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors," *Nature*, no. 418, pp. 171–174, 2002.

[20] R. A. Laird and B. S. Schamp, "Competitive Intransitivity Promotes Species Coexistence," *The American Naturalist*, vol. 168, no. 2, 2006.

[21] A. Szolnoki, M. Mobilia, L.-L. Jiang, B. Szczesny, A. M. Rucklidge, and M. Perc, "Cyclic dominance in evolutionary games: a review," *J R Soc Interface*, vol. 11, no. 100, 2014.

[22] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, C. Fernando, and K. Kavukcuoglu, "Population based training of neural networks," *CoRR*, vol. abs/1711.09846, 2017.

[23] D. Balduzzi, S. Racanière, J. Martens, J. Foerster, K. Tuyls, and T. Graepel, "The mechanics of n-player differentiable games," in *ICML*, 2018.

[24] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. P. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[25] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Perolat, D. Silver, and T. Graepel, "A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning," in *NIPS*, 2017.

[26] S. Legg and M. Hutter, "A universal measure of intelligence for artificial agents," in *IJCAI*, 2005.

[27] S. Legg and J. Veness, "An Approximation of the Universal Intelligence Measure," in *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, 2013.

[28] R. J. Solomonoff, "A formal theory of inductive inference I, II," *Inform. Control*, vol. 7, no. 1-22, 224-254, 1964.

[29] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems Inform. Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[30] G. J. Chaitin, "On the length of computer programs for computing finite binary sequences," *J Assoc. Comput. Mach.*, vol. 13, pp. 547–569, 1966.

[31] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, no. 30, pp. 27–38, 2009.

[32] J. Hernández-Orallo, P. Flach, and C. Ferri, "A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss," *JMLR*, no. 13, pp. 2813–2869, 2012.

[33] J. Hernández-Orallo, *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press, 2017.

[34] J. Hernández-Orallo, "Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement," *Artificial Intelligence Review*, vol. 48, no. 3, pp. 397–447, 2017.

[35] R. S. Olson, W. La Cava, P. Orzechowski, R. J. Urbanowicz, and J. H. Moore, "PMLB: a large benchmark suite for machine learning evaluation and comparison," *BioData Mining*, vol. 10, p. 36, Dec 2017.

[36] C. Spearman, "'General Intelligence,' objectively determined and measured," *Am. J. Psychol.*, vol. 15, no. 201, 1904.

[37] A. Woolley, C. Fabris, A. Pentland, N. Hashmi, and T. Malone, "Evidence for a Collective Intelligence Factor in the Performance of Human Groups," *Science*, no. 330, pp. 686–688, 2010.

[38] S. Bringsjord, "Psychometric artificial intelligence," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 23, no. 3, pp. 271–277, 2011.

[39] D. R. Hunter, "MM algorithms for generalized Bradley-Terry models," *Annals of Statistics*, vol. 32, no. 1, pp. 384–406, 2004.

[40] M. C. Machado, M. G. Bellemare, E. Talvitie, J. Veness, M. Hausknecht, and M. Bowling, "Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents," *Journal of Artificial Intelligence Research (JAIR)*, vol. 61, pp. 523–562, 2018.

[41] A. Liapis, G. N. Yannakakis, and J. Togelius, "Towards a Generic Method of Evaluating Game Levels," in *Artificial Intelligence in Digital Interactive Entertainment (AIIDE)*, 2013.

[42] B. Horn, S. Dahlskog, N. Shaker, G. Smith, and J. Togelius, "A Comparative Evaluation of Procedural Level Generators in the Mario AI Framework," in *Foundations of Digital Games*, 2014.

[43] T. S. Nielsen, G. Barros, J. Togelius, and M. J. Nelson, "General video game evaluation using relative algorithm performance profiles," in *EvoApplications*, 2015.

[44] F. de Mesentier Silva, S. Lee, J. Togelius, and A. Nealen, "AI-based Playtesting of Contemporary Board Games," in *Foundations of Digital Games (FDG)*, 2017.

[45] V. Volz, J. Schrum, J. Liu, S. M. Lucas, A. M. Smith, and S. Risi, "Evolving Mario Levels in the Latent Space of a Deep Convolutional Generative Adversarial Network," in *GECCO*, 2018.

[46] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of item response theory*. Sage Publications, 1991.

[47] F. Martínez-Plumed and J. Hernández-Orallo, " AI results for the Atari 2600 games: difficulty and discrimination using IRT," in *Workshop on Evaluating General-Purpose AI (EGPAI at IJCAI)*, 2017.

[48] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, "Statistical ranking and combinatorial Hodge theory," *Math. Program., Ser. B*, vol. 127, pp. 203–244, 2011.

[49] O. Candogan, I. Menache, A. Ozdaglar, and P. A. Parrilo, "Flows and Decompositions of Games: Harmonic and Potential Games," *Mathematics of Operations Research*, vol. 36, no. 3, pp. 474–503, 2011.

[50] O. Candogan, A. Ozdaglar, and P. A. Parrilo, "Near-Potential Games: Geometry and Dynamics," *ACM Trans Econ Comp*, vol. 1, no. 2, 2013.

[51] O. Candogan, A. Ozdaglar, and P. A. Parrilo, "Dynamics in near-potential games," *Games and Economic Behavior*, vol. 82, pp. 66–90, 2013.

[52] W. E. Walsh, D. C. Parkes, and R. Das, "Choosing samples to compute heuristic-strategy nash equilibrium," in *Proceedings of the Fifth Workshop on Agent-Mediated Electronic Commerce*, 2003.

[53] M. P. Wellman, "Methods for empirical game-theoretic analysis," in *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, pp. 1552–1556, 2006.

[54] S. Phelps, S. Parsons, and P. McBurney, "An Evolutionary Game-Theoretic Comparison of Two Double-Auction Market Designs," in *Agent-Mediated Electronic Commerce VI, Theories for and Engineering of Distributed Mechanisms and Systems, AAMAS Workshop*, pp. 101–114, 2004.

[55] S. Phelps, K. Cai, P. McBurney, J. Niu, S. Parsons, and E. Sklar, "Auctions, Evolution, and Multi-agent Learning," in *AAMAS and 7th European Symposium on Adaptive and Learning Agents and Multi-Agent Systems (ALAMAS)*, pp. 188–210, 2007.

[56] M. Ponsen, K. Tuyls, M. Kaisers, and J. Ramon, "An evolutionary game-theoretic analysis of poker strategies," *Entertainment Computing*, vol. 1, no. 1, pp. 39–45, 2009.

[57] D. Bloembergen, K. Tuyls, D. Hennes, and M. Kaisers, "Evolutionary dynamics of multi-agent learning: A survey," *J. Artif. Intell. Res. (JAIR)*, vol. 53, pp. 659–697, 2015.

[58] K. Tuyls, J. Perolat, M. Lanctot, J. Z. Leibo, and T. Graepel, "A Generalised Method for Empirical Game Theoretic Analysis ," in *AAMAS*, 2018.

[59] M. Dudik, K. Hofmann, R. E. Schapire, A. Slivkins, and M. Zoghi, "Contextual Dueling Bandits," in *COLT*, 2015.

[60] A. Balsubramani, Z. Karnin, R. E. Schapire, and M. Zoghi, "Instance-dependent Regret Bounds for Dueling Bandits," in *COLT*, 2016.

[61] P. R. Jordan, C. Kiekintveld, and M. P. Wellman, "Empirical game-theoretic analysis of the TAC supply chain game," in *6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2007), Honolulu, Hawaii, USA, May 14-18, 2007*, p. 193, 2007.

[62] P. R. Jordan, *Practical Strategic Reasoning with Applications in Market Games.* PhD thesis, 2010.

[63] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior.* Princeton University Press, Princeton NJ, 1944.

[64] J. F. Nash, "Equilibrium Points in n-Person Games," *Proc Natl Acad Sci U S A*, vol. 36, no. 1, pp. 48–49, 1950.

[65] J. Hofbauer and W. H. Sandholm, "On the global convergence of stochastic fictitious play," *Econometrica*, vol. 70, no. 6, pp. 2265–2294, 2002.

[66] W. H. Sandholm, *Population Games and Evolutionary Dynamics.* MIT Press, 2010.

[67] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas, "Dueling Network Architectures for Deep Reinforcement Learning," in *ICML*, 2016.

[68] H. van Hasselt, A. Guez, M. Hessel, V. Mnih, and D. Silver, "Learning values across many orders of magnitude," in *NIPS*, 2016.

[69] G. Ostrovski, M. G. Bellemare, A. van den Oord, and R. Munos, "Count-Based Exploration with Neural Density Models," in *ICML*, 2017.

[70] M. Hessel, J. Modayil, H. van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining Improvements in Deep Reinforcement Learning," in *AAAI*, 2018.

[71] L. E. Ortiz, R. E, Schapire, and S. M. Kakade, "Maximum entropy correlated equilibrium," in *Technical Report TR-2006-21, CSAIL MIT*, 2006.

[72] L. E. Ortiz, R. E. Schapire, and S. M. Kakade, "Maximum entropy correlated equilibria," in *AISTATS*, 2007.

[73] P. Diaconis, *Group Representations in Probability and Statistics.* Institute of Mathematical Statistics, 1988.

[74] Y. LeCun, , L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[75] R. Kondor and T. Jebara, "A kernel between sets of vectors," in *ICML*, 2003.

[76] R. Kondor, "Group theoretical methods in machine learning," in *PhD dissertation*, 2008.

[77] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep Sets," in *NIPS*, 2017.

[78] J. Hartford, D. R. Graham, K. Leyton-Brown, and S. Ravanbakhsh, "Deep Models of Interactions Across Sets," in *ICML*, 2018.

[79] R. Kondor, Z. Lin, and S. Trivedi, "Clebsch–Gordan Nets: a Fully Fourier Space Spherical Convolutional Neural Network," in *NIPS*, 2018.

[80] S. Sukhbaatar, Z. Lin, I. Kostrikov, G. Synnaeve, A. Szlam, and R. Fergus, "Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play," in *ICLR*, 2017.

[81] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, 1996.

[82] R. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*. MIT Press, 2012.

[83] R. J. Vandenberg and C. E. Lance, "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research," *Organizational Research Methods*, vol. 3, no. 1, pp. 4–70, 2000.

APPENDIX

A Invariance: Further motivation

Consider agents A, B, C evaluated on a benchmark suite comprising three tasks:

	task 1	task 2	task 3	**average**	**rank**
agent A	89	93	76	86	1st
agent B	85	85	85	85	2nd
agent C	79	74	99	84	3rd

On average, agent A performs best and agent C performs worst. Consider a second benchmark suite containing an additional fourth task. On the second benchmark suite, agent A performs worst and agent C performs best on average. However, a closer look at the second suite reveals that the additional task is a minor variant of one of the original three tasks.

	task 1	task 2	task 3a	task 3b	**average**	**rank**
agent A	89	93	76	77	84	3rd
agent B	85	85	85	84	85	2nd
agent C	79	74	99	98	88	1st

Measuring performance by uniformly averaging over tasks in a benchmark suite is sensitive to the set of tasks that are included in the suite. Including redundant tasks, whether consciously or not, can easily skew average performance in favor of or against particular agents.

The problem becomes more serious when agent performance is measured against other agents, as in games such as Go, Chess or StarCraft. It is easy to manipulate the measured performance of agents by tweaking the composition of the population used to evaluate them. Consider the following example:

	agent A	agent B	agent C	**Elo**
agent A	0.5	0.9	0.1	0
agent B	0.1	0.5	0.9	0
agent C	0.9	0.1	0.5	0

The three agents exhibit rock-paper-scissors dynamics; their Elo ratings (normalized to sum to zero) are all zero. However, adding a second copy of agent C decreases the Elo rating of agent A and increases the Elo rating of agent B:

	agent A	agent B	agent C_1	agent C_2	**Elo**
agent A	0.5	0.9	0.1	0.1	-63
agent B	0.1	0.5	0.9	0.9	63
agent C_1	0.9	0.1	0.5	0.5	0
agent C_2	0.9	0.1	0.5	0.5	0

That is, the Elo ratings of agents A and B are easily manipulated by changing the structure of the population.

The examples above suggest it is important to find evaluation metrics that are *invariant* to redundant changes in the population of agents or suite of tasks.

Related work. A different notion of measurement invariance has been proposed in the psychometric and consumer research literatures [83]. There, measurement invariance refers to the statistical property that a measurement measures the same construct across a predefined set of groups. For example, whether or not a question in an IQ test is measurement invariant has to do with whether or not the question is interpreted in the same way by individuals with different cultural backgrounds.

B   Proofs of propositions

Proof of proposition 1.

Proposition 1. *Batch Elo updates are stable if the matrices of empirical probabilities and predicted probabilities have the same row-sums (or, equivalently the same column-sums):*

$$\nabla_i \sum_j \ell_{Elo}(\bar{p}_{ij}, \hat{p}_{ij}) = 0 \; \forall i \quad \textit{iff} \quad \sum_j \bar{p}_{ij} = \sum_j \hat{p}_{ij} \; \forall i.$$

Proof. Player i's weights are updated in one batch after observing win-loss probabilities $\bar{p}_{ij}$ for each player $j = 1, \ldots, n$. Observe that

$$\nabla_i \left[\sum_{j=1}^n \ell_{\text{Elo}}(\bar{p}_{ij}, \hat{p}_{ij}) \right] = \sum_{j=1}^n (\bar{p}_{ij} - \hat{p}_{ij}).$$

The result follows. □

Proof of proposition 2.

Proposition 2. *Given matrix* $\mathbf{S}_{m \times n}$ *with rank* r *and singular value decomposition* $\mathbf{UDV}^\mathsf{T}$. *Construct antisymmetric matrix*

$$\mathbf{A}_{(m+n) \times (m+n)} = \begin{pmatrix} \mathbf{0}_{m \times m} & \mathbf{S}_{m \times n} \\ -\mathbf{S}_{n \times m}^\mathsf{T} & \mathbf{0}_{n \times n} \end{pmatrix}.$$

Then the thin Schur decomposition of $\mathbf{A}$ *is* $\mathbf{Q \Lambda Q}^\mathsf{T}$ *where the eigenpairs in the* $\mathbf{\Lambda}_{2r \times 2r}$ *are* $\pm$ *the singular values in* $\mathbf{D}$ *and*

$$\mathbf{Q}_{(m+n) \times r} = \begin{pmatrix} -\mathbf{u}_1 & \mathbf{0}_{m \times 1} & \cdots & -\mathbf{u}_r & \mathbf{0}_{m \times 1} \\ \mathbf{0}_{n \times 1} & \mathbf{v}_1 & \cdots & \mathbf{0}_{n \times 1} & \mathbf{v}_r \end{pmatrix}$$

Proof. Direct computation; multiply out the matrices. □

Proof of proposition 3.

Proposition 3. *(i) If probabilities* $\mathbf{P}$ *are generated by Elo ratings* $\mathbf{r}$ *then the divergence of its logit is* $\mathbf{r}$. *That is,*

$$\textit{if } p_{ij} = \sigma(r_i - r_j) \; \forall i, j \; \textit{ then } \; \text{div}(\text{logit}\,\mathbf{P}) = \left(\frac{1}{n} \sum_{j=1}^n (r_i - r_j) \right)_{i=1}^n = \mathbf{r}.$$

(ii) There is an Elo rating that generates probabilities $\mathbf{P}$ *iff* $\text{curl}(\text{logit}\,\mathbf{P}) = 0$. *Alternatively, iff* $\log \frac{p_{ij}}{p_{ji}} + \log \frac{p_{jk}}{p_{kj}} + \log \frac{p_{ki}}{p_{ik}} = 0$ *for all* i, j, k.

Proof. For the first claim, apply definitions and recall $\mathbf{r}^\mathsf{T} \mathbf{1} = 0$. For the second, apply the Hodge decomposition from section 2.2. □

Proof of proposition 4.

Proposition 4 (maxent NE). *The game* $\max_{\mathbf{p} \in \Delta_n} \min_{\mathbf{q} \in \Delta_n} \mathbf{p}^\mathsf{T} \mathbf{A} \mathbf{q}$, *where* $\mathbf{A}$ *is antisymmetric, has a unique symmetric Nash equilibrium* $(\mathbf{p}^*, \mathbf{p}^*)$, *with greater entropy than any other Nash equilibrium.*

Proof. The Nash equilibria in a two-player zero-sum are rectangular: if $(\mathbf{p}, \mathbf{q})$ and $(\mathbf{u}, \mathbf{v})$ are Nash equilibria then so are $(\mathbf{p}, \mathbf{v})$ and $(\mathbf{u}, \mathbf{q})$. Further, they form a convex polytope. Since $\mathbf{A}^\mathsf{T} = -\mathbf{A}$, the set of Nash equilibria is also symmetric: if $(\mathbf{p}, \mathbf{q})$ is a Nash equilibrium then so is $(\mathbf{q}, \mathbf{p})$. The entropy $H(\mathbf{p}) := -\sum_{i=1}^n p_i \log p_i$ is strictly concave and therefore achieves a unique maximum on the compact, convex, symmetric set of Nash equilibria. □

C   Proof of theorem 1

Theorem (main result for AvA). *The maxent NE has the following properties:*

P1. Invariant: *Nash averaging, with respect to the maxent NE, is invariant to redundancies in* $\mathbf{A}$.

P2. Continuous: *If* $\mathbf{p}^*$ *is a Nash for* $\hat{\mathbf{A}}$ *and* $\epsilon = \|\mathbf{A} - \hat{\mathbf{A}}\|_{max}$ *then* $\mathbf{p}^*$ *is an* ϵ-*Nash for* $\mathbf{A}$.

P3. Interpretable: *(i) The maxent NE on* $\mathbf{A}$ *is the uniform distribution,* $\mathbf{p}^* = \frac{1}{n}\mathbf{1}$, *iff the meta-game is* cyclic, *i.e.* $\mathrm{div}(\mathbf{A}) = \mathbf{0}$. *(ii) If the meta-game is* transitive, *i.e.* $\mathbf{A} = \mathrm{grad}(\mathbf{r})$, *then the maxent NE is the uniform distribution on the player(s) with highest rating(s) – there could be a tie.*

C.1   Proof of theorem 1.1

First, we more precisely formalize invariance to redundancy.

Definition 3 (invariance to copying the last row and column). *Given antisymmetric matrix* $\mathbf{A}_{n \times n}$, *denote the right-most column by* $\mathbf{a}_n$. *Assume the right-most column (and bottom row) of* $\mathbf{A}$ *differs from all other columns. Construct antisymmetric matrix*

$$\mathbf{A}'_{(n+1) \times (n+1)} = \begin{pmatrix} \mathbf{A} & \mathbf{a}_n \\ -\mathbf{a}_n^{\mathsf{T}} & 0 \end{pmatrix} \tag{1}$$

by adding an additional copy of the right-most column (and bottom row) to $\mathbf{A}$. *A family of functions*

$$\left\{ \mathbf{p}_k : \{\text{antisymmetric } k \times k \text{ matrices}\} \to \mathbb{R}^k \right\}_{k=1}^{\infty}$$

is invariant to adding a row and column according to (1) *if*

$$\mathbf{p}_{n+1}(\mathbf{A}')^{\mathsf{T}} = \left(\mathbf{p}_n(\mathbf{A})[1], \dots, \mathbf{p}_n(\mathbf{A})[n-1], \frac{\mathbf{p}_n(\mathbf{A})[n]}{2}, \frac{\mathbf{p}_n(\mathbf{A})[n]}{2} \right).$$

If the copied row is not unique and receives positive mass under maxent Nash, then maxent Nash will already be spreading mass across the copies. In that case, adding *yet another* copy will result in the maxent Nash on the larger mass spreading mass evenly across all copies.

Lemma 1. *Suppose* $\mathbf{A}_{n \times n}$ *is antisymmetric with Nash equilibrium* $\mathbf{p}$. *Construct* $\mathbf{A}'$ *from* $\mathbf{A}$ *by adding a redundant copy of the right-most column and bottom row according* (1). *Then*

$$\mathbf{p}'_{\alpha} = \left(p_1, \dots, p_{n-1}, \alpha \cdot p_n, (1 - \alpha) \cdot p_n \right)$$

is a Nash equilibrium for $\mathbf{A}'$ *for all* $\alpha \in [0, 1]$. *Conversely, if* $\mathbf{p}'$ *is a Nash equilibrium for* $\mathbf{A}'$ *then*

$$(p'_1, \dots, p'_{n-1}, p'_n + p'_{n+1})$$

is a Nash equilibrium for $\mathbf{A}$.

Proof. Since the value of the game on $\mathbf{A}$ is zero, it follows that $\mathbf{p}$ is a Nash equilibrium iff all the coordinates of $\mathbf{Ap}$ are nonnegative, i.e. $\mathbf{Ap} \succeq \mathbf{0}_{n \times 1}$. Direct computation shows that $\mathbf{A}'\mathbf{p}'_{\alpha} \succeq \mathbf{0}_{(n+1) \times 1}$, which implies $\mathbf{p}'_{\alpha}$ is a Nash equilibrium for $\mathbf{A}'$. The converse follows similarly. $\square$

Finally, we prove theorem 1.2.

Proof. The definition and lemma are stated in the particular case where the last column and row are copied. This is simply for notational convenience. They generalize trivially to copying arbitrary row/columns into arbitrary positions, and can be applied inductively to the cases where $\mathbf{A}'$ is constructed from $\mathbf{A}$ by inserting multiple redundant copies. $\square$

C.2 Proof of theorem 1.2

Recall the max-norm on matrices is $\|\mathbf{S}\|_{\max} := \max_{ij} |\mathbf{S}_{ij}|$.

Definition 4 (ϵ-Nash equilibrium). *A joint strategy* $(\mathbf{p}^*, \mathbf{q}^*)$ *is an ϵ-Nash equilibrium for* $\mathbf{A}$ *if the benefit from either player deviating, separately, is at most ϵ:*

$$\max_{\mathbf{p}'}(\mathbf{p}' - \mathbf{p}^*)^{\mathsf{T}}\mathbf{A}\mathbf{q}^* \leq \epsilon \quad \text{and} \quad \max_{\mathbf{q}'}(\mathbf{p}^*)^{\mathsf{T}}\mathbf{A}(\mathbf{q}^* - \mathbf{q}') \leq \epsilon.$$

We are now ready to prove

P2 Continuous: If $\mathbf{p}^*$ is a Nash for $\hat{\mathbf{A}}$ and $\epsilon = \|\mathbf{A} - \hat{\mathbf{A}}\|_{\max}$ then $\mathbf{p}^*$ is an ϵ-Nash for $\mathbf{A}$.

Proof. Suppose $(\mathbf{p}^*, \mathbf{p}^*)$ is a Nash equilibrium for the antisymmetric matrix $\hat{\mathbf{A}}$. Observe that

$$(\mathbf{p}' - \mathbf{p}^*)^{\mathsf{T}}\mathbf{A}\mathbf{p}^* = (\mathbf{p}')^{\mathsf{T}}\hat{\mathbf{A}}\mathbf{p}^* + (\mathbf{p}')^{\mathsf{T}}\mathbf{A}\mathbf{p}^* - (\mathbf{p}')^{\mathsf{T}}\hat{\mathbf{A}}\mathbf{p}^*$$

for any distribution $\mathbf{p}'$ because $(\mathbf{p}^*)^{\mathsf{T}}\mathbf{A}\mathbf{p}^* = 0$ since $\mathbf{A}$ is antisymmetric. It follows that

$$\max_{\mathbf{p}'}\left\{(\mathbf{p}' - \mathbf{p}^*)^{\mathsf{T}}\mathbf{A}\mathbf{p}^*\right\} \leq \max_{\mathbf{p}'}\left\{(\mathbf{p}')^{\mathsf{T}}\hat{\mathbf{A}}\mathbf{p}^*\right\} + \max_{\mathbf{p}'}\left\{(\mathbf{p}')^{\mathsf{T}}(\mathbf{A} - \hat{\mathbf{A}})\mathbf{p}^*\right\}$$

The first term on the right-hand-side is ≤ 0 since $\mathbf{p}^*$ is a Nash equilibrium for $\hat{\mathbf{A}}$ and the value of the game is zero. The second term on the right-hand-side is $\leq \epsilon$ because $\|\mathbf{A} - \hat{\mathbf{A}}\|_{\max} \leq \epsilon$ and $\mathbf{p}'$ and $\mathbf{p}^*$ are probability distributions. $\square$

Note that since $\|\mathbf{S}\|_{\max} \leq \|\mathbf{S}\|_2 \leq \|\mathbf{S}\|_F$ for any $\mathbf{S}$, the divergence from Nash is also controlled by how well $\hat{\mathbf{A}}$ approximates $\mathbf{A}$ in the operator or Frobenius norms.

The proof is adapted from the proof of the following lemma in [58].

Lemma 2 (Nash on approximate games). *Suppose* $(\mathbf{p}^*, \mathbf{q}^*)$ *is a Nash equilibrium for* $\hat{\mathbf{A}}$ *and that* $\epsilon = \|\mathbf{A} - \hat{\mathbf{A}}\|_{max}$. *Then* $(\mathbf{p}^*, \mathbf{q}^*)$ *is a 2ϵ-Nash for* $\mathbf{A}$.

Our result is slightly sharper because we specialize to antisymmetric matrices.

C.3 Proof of theorem 1.3

Proof. **(i)** If $\operatorname{div}(\mathbf{A}) = \mathbf{0}$ then $\frac{1}{n}\mathbf{1}^{\mathsf{T}}\mathbf{A} = \mathbf{0}^{\mathsf{T}}$ and $\mathbf{A}\frac{1}{n}\mathbf{1} = \mathbf{0}$ implying the uniform distribution is a Nash equilibrium because there is no incentive to deviate from $(\frac{1}{n}\mathbf{1}, \frac{1}{n}\mathbf{1})$. The uniform distribution also has maximum entropy. Conversely, suppose $\operatorname{div}(\mathbf{A}) \neq \mathbf{0}$. Then $\mathbf{1}^{\mathsf{T}}\mathbf{A}$ has at least one positive and one negative coordinate because we know $\mathbf{1}^{\mathsf{T}}\mathbf{A}\mathbf{1} = 0$ by antisymmetry. It follows that if the row player chooses $\frac{1}{n}\mathbf{1}^{\mathsf{T}}$ then the column player is incentivized to choose a distribution with more mass on the positive coordinate and less on the negative. In other words, the column player will not play the uniform distribution, and the uniform distribution is therefore not a Nash equilibrium.

(ii) By assumption $\mathbf{A} = \operatorname{grad}(\mathbf{r}) = \mathbf{r}\mathbf{1}^{\mathsf{T}} - \mathbf{1}\mathbf{r}^{\mathsf{T}}$, so $\mathbf{p}^{\mathsf{T}}\mathbf{A}\mathbf{q} = \mathbf{p}^{\mathsf{T}}\mathbf{r} - \mathbf{r}^{\mathsf{T}}\mathbf{q}$ decouples into an independent maximization problem with respect to $\mathbf{p}$ and minimization problem with respect to $\mathbf{q}$. It follows that the optimal distribution $\mathbf{p}^*$ concentrates mass on the maximal coordinate(s) of $\mathbf{r}$ and so does $\mathbf{q}^*$. That is, to be a Nash equilibrium, $\mathbf{p}^*$ and $\mathbf{q}^*$ must place their mass on the maximal coordinate if it is unique and can distribute it arbitrarily over the set of maximal coordinates if there is a tie. Adding the condition that the Nash equilibrium has maximum entropy entails placing the uniform distribution over the maximal coordinate(s). $\square$

D Nash averaging for agent-vs-task

Given score matrix $\mathbf{S}_{m \times n}$, construct antisymmetric matrix

$$\mathbf{A}_{(m+n) \times (m+n)} = \begin{pmatrix} \mathbf{0}_{m \times m} & \mathbf{S}_{m \times n} \\ -\mathbf{S}_{n \times m}^{\mathsf{T}} & \mathbf{0}_{n \times n} \end{pmatrix}.$$

Note this differs from the antisymmetrization used in section 3.3, see next remark.

Remark 1. The graph structure underlying AvT is *bipartite*: agents interact with tasks and tasks with agents, but there are no direct agent-agent or task-task interactions. When done in full generality, the definitions of div, grad and curl take into account the graph structure, see [48]. In particular, div, grad and curl are computed differently on bipartite graphs than fully connected graphs – note the definitions in section 2.2 are specific to fully connected graphs. Working in full generality is overkill for our purposes. It suffices to introduce slightly *ad hoc* notation to handle the specific case of AvT.

Introduce the notation

$$\mathrm{div}_a(\mathbf{S}) = \frac{1}{m}\mathbf{S} \cdot \mathbf{1}_{m \times 1} \quad \text{and} \quad \mathrm{div}_e(\mathbf{S}) = -\frac{1}{n}\mathbf{S}^\mathsf{T} \cdot \mathbf{1}_{n \times 1},$$

where $\mathrm{div}_a(\mathbf{S})$ measures uniform **average skill of agents** on tasks and $\mathrm{div}_e(\mathbf{S})$ measures uniform **average difficulty of tasks** for agents. Let

$$\mathrm{grad}(\mathbf{s}, \mathbf{d}) = \mathbf{s} \cdot \mathbf{1}^\mathsf{T} - \mathbf{1} \cdot \mathbf{d}^\mathsf{T}.$$

Define a two-player zero-sum meta-game

$$\max_{(\mathbf{p}_a, \mathbf{p}_e) \in \Delta_m \times \Delta_n} \min_{(\mathbf{q}_a, \mathbf{q}_e) \in \Delta_m \times \Delta_n} (\mathbf{p}_a; \mathbf{p}_e)^\mathsf{T} \mathbf{A} (\mathbf{q}_a; \mathbf{q}_e).$$

The setup is the same as for AvA in the main text except the row and column meta-players each play two distributions: one on agents and one on tasks/environments. The same argument as in proposition 4 shows there is a unique symmetric maxent Nash equilibrium $\big((\mathbf{p}_a^*, \mathbf{p}_e^*), (\mathbf{p}_a^*, \mathbf{p}_e^*)\big)$.

Definition 5. *The **maxent Nash evaluation method** for AvT is*

$$\mathcal{E}_m : \big\{\text{evaluation data}\big\} = \big\{\text{antisymmetric matrices}\big\} \xrightarrow{\text{maxent NEs}} \Big[\big\{\text{players}\big\} \xrightarrow{\text{Nash averages}} \mathbb{R}\Big],$$

*where $\mathbf{p}_a^*$ and $\mathbf{p}_e^*$ are the **maxent Nash equilibria** over agents and environments and $\mathbf{n}_e := -\mathbf{S}^\mathsf{T} \cdot \mathbf{p}_a^*$ and $\mathbf{n}_a := \mathbf{S} \cdot \mathbf{p}_e^*$ are the **Nash averages** quantifying difficulty of environments (Nash averaged over agents) and skill of agents (Nash averaged over environments) respectively.*

Suppose $\mathbf{S}$ decomposes as

$$\mathbf{S} = \mathrm{grad}(\mathbf{s}, \mathbf{d}) + \tilde{\mathbf{S}},$$

where $\mathrm{div}_a(\tilde{\mathbf{S}}) = \mathbf{0}$ and $\mathrm{div}_e(\tilde{\mathbf{S}}) = \mathbf{0}$. Observe that

$$(\mathbf{p}_a; \mathbf{p}_e)^\mathsf{T} \mathbf{A} (\mathbf{q}_a; \mathbf{q}_e) = \mathbf{s}^\mathsf{T}(\mathbf{p}_a - \mathbf{q}_a) + \mathbf{d}^\mathsf{T}(\mathbf{p}_e - \mathbf{q}_e) + \mathbf{p}_a^\mathsf{T} \tilde{\mathbf{S}} \mathbf{q}_e + \mathbf{q}_a^\mathsf{T} \tilde{\mathbf{S}} \mathbf{p}_e$$

If $\tilde{\mathbf{S}} \equiv \mathbf{0}_{m \times n}$ then the game reduces to

$$(\mathbf{p}_a; \mathbf{p}_e)^\mathsf{T} \mathbf{A} (\mathbf{q}_a; \mathbf{q}_e) = \mathbf{s}^\mathsf{T}(\mathbf{p}_a - \mathbf{q}_a) + \mathbf{d}^\mathsf{T}(\mathbf{p}_e - \mathbf{q}_e)$$

and so the row player maximizes its payoff by putting all its agent-mass on the most skillful agent(s) and all of its environment-mass on the most difficult task(s) – and similarly for the column player.

It is easy to check that the maxent Nash $(\mathbf{p}_a^*, \mathbf{p}_e^*)$ has $\mathbf{p}_a$ a uniform distribution on all agents iff $\mathrm{div}_a(\mathbf{S}) = \mathbf{0}$, and has $\mathbf{p}_e$ a uniform distribution on tasks iff $\mathrm{div}_e(\mathbf{S}) = \mathbf{0}$. We thus obtain

Theorem 2 (main result for AvT). *The maxent NE has the following properties:*

P1. **Invariant:** *Nash averaging is invariant to redundancies in $\mathbf{A}$.*

P2. **Continuous:** *If $(\mathbf{p}_a^*, \mathbf{p}_e^*)$ is a Nash for $\hat{\mathbf{A}}$ and $\frac{\epsilon}{2} = \|\mathbf{A} - \hat{\mathbf{A}}\|_{max}$ then $(\mathbf{p}_a^*, \mathbf{p}_e^*)$ is an ϵ-Nash for $\mathbf{A}$.*

P3. **Interpretable:** *(i) The agent component of the maxent NE on $\mathbf{A}$ is the uniform distribution on agents, $\mathbf{p}_a^* = \frac{1}{n}\mathbf{1}$, iff $\mathrm{div}_a(\mathbf{A}) = \mathbf{0}$.*
(ii) The task component of the maxent NE on $\mathbf{A}$ is the uniform distribution on tasks, $\mathbf{p}_e^* = \frac{1}{m}\mathbf{1}$, iff $\mathrm{div}_e(\mathbf{A}) = \mathbf{0}$.
(iii) If the meta-game is transitive, *i.e. $\mathbf{A} = \mathrm{grad}(\mathbf{s}, \mathbf{d})$, then the maxent NE is the uniform distribution on the most skillful agent(s) and the uniform distribution on the most difficult task(s) – there could be ties.*
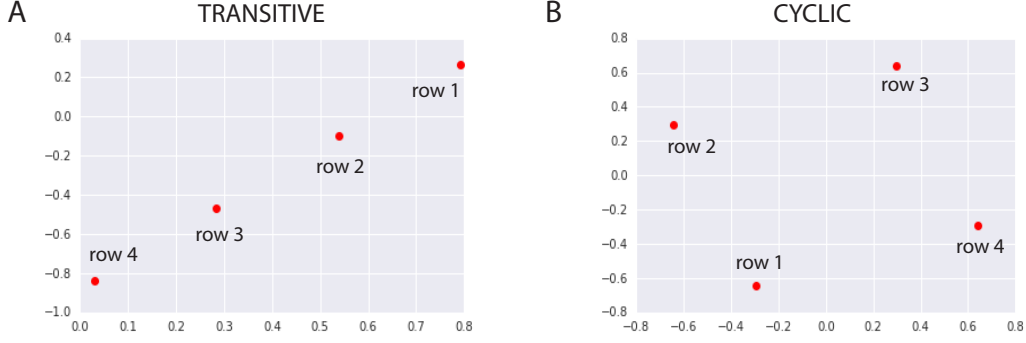
Figure 3: **Visualizing Schur decompositions.** (A) Rows of $\mathbf{Q}^{\mathbf{T}}_{4\times2}$ form a straight line, reflecting the transitive structure of $\mathbf{T}$. (B): Rows of $\mathbf{Q}^{\mathbf{C}}_{4\times2}$ lie on a circle centered at the origin.

E Code for computing mElo$_2$ updates

The routine `mElo2_update` takes as input: a pair of players i, j, the probability `p_ij` of player `i` beating player `j` (which could be 0 or 1 if only a single match is observed on the given round), the rating vector `r` and the $n \times 2$ matrix `c` quantifying non-transitive interactions. It returns updates to the `i`[th] and `j`[th] entries of `r` and `c`.

```
def mElo2_update(i, j, p_ij, r, c):
  p_hat_ij = sigmoid(r[i] − r[j] + c[i, 0] ∗ c[j, 1] − c[j, 0] ∗ c[i, 1])
  delta = p_ij − p_hat_ij
  r_update = [16 ∗ delta, −16 ∗ delta]
    # r has higher learning rate than c
  c_update = [
    [+delta ∗ c[j, 1], −delta ∗ c[i, 1]],
    [−delta ∗ c[j, 0], +delta ∗ c[i, 0]]
  ]
  return r_update, c_update
```

F On the geometry of antisymmetric matrices

This section provides some intuition for multidimensional Elo ratings by describing some of the underling geometry.

F.1 Visualizing the Schur decomposition

Consider the following logit matrices:

$\mathbf{T}$	A	B	C	D
A	0	1	2	3
B	-1	0	1	2
C	-2	-1	0	1
D	-3	-2	-1	0

and

$\mathbf{C}$	A	B	C	D
A	0	1	0	-1
B	-1	0	1	0
C	0	-1	0	1
D	1	0	-1	0

Note that $\mathbf{T}$ is transitive since $\mathbf{T} = \mathrm{grad} \circ \mathrm{div}(\mathbf{T})$ and $\mathbf{C}$ is cyclic since $\mathrm{div}(\mathbf{C}) = \mathbf{0}$. Both matrices have rank two, so the thin Schur decomposition can be written

$$\mathbf{Q}_{4\times2}\mathbf{\Lambda}_{2\times2}\mathbf{Q}^{\mathsf{T}}_{2\times4}$$

in either case. The matrices $\mathbf{Q}^{\mathbf{T}}_{4\times2}$ and $\mathbf{Q}^{\mathbf{C}}_{4\times2}$ arising from the respective Schur decompositions

$$\mathbf{Q}^{\mathbf{T}}_{4\times2} = \begin{pmatrix} 0.793 & 0.267 \\ 0.538 & -0.101 \\ 0.284 & -0.469 \\ 0.029 & -0.836 \end{pmatrix} \quad \text{and} \quad \mathbf{Q}^{\mathbf{C}}_{4\times2} = \begin{pmatrix} -0.296 & -0.642 \\ -0.642 & 0.296 \\ 0.296 & 0.642 \\ 0.642 & -0.296 \end{pmatrix}$$
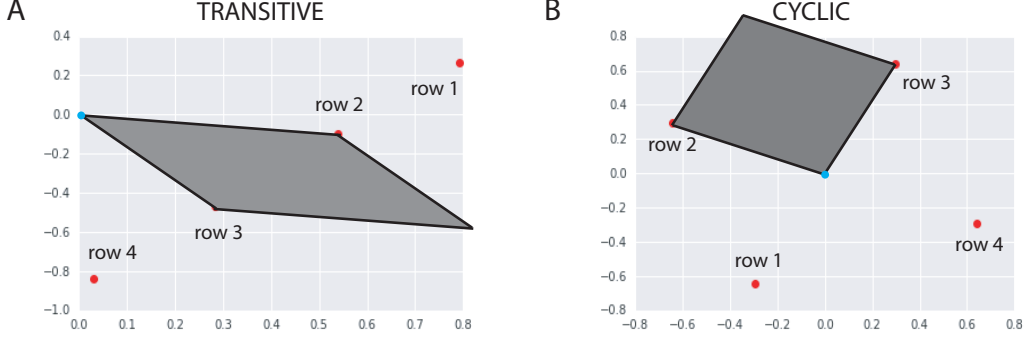
Figure 4: **Visualizing logits.** The entry $\mathbf{A}_{ij}$ of $\mathbf{A}$ is λ times the *signed area* of the parallelogram covered by the origin $(0,0)$, $\mathbf{Q}_{i,\bullet}$, $\mathbf{Q}_{i,\bullet} + \mathbf{Q}_{j,\bullet}$ and $\mathbf{Q}_{j,\bullet}$, where $\mathbf{Q}_{i,\bullet}$ and $\mathbf{Q}_{j,\bullet}$ are vectors corresponding to row i and row j of $\mathbf{Q}_{4\times 2}$.

can separately be thought of (and plotted) as four two-vectors, see figure 3. The rows of $\mathbf{Q}_{4\times 2}^{\mathbf{T}}$ form a straight line, see panel A. More generally, if $\mathbf{A} = \mathrm{grad}(\mathbf{r})$ and $\mathbf{r}$ is not identically zero then $\mathbf{A}$ is rank two with thin Schur decomposition $\mathbf{A} = \mathbf{Q}_{n\times 2}\mathbf{\Lambda}_{2\times 2}\mathbf{Q}_{2\times n}^{\mathsf{T}}$, where the rows of $\mathbf{Q}$ are points along a line. More precisely, the rows are of the form

$$\mathbf{q}_k = \mathbf{a} + \alpha_k \mathbf{b} \quad \text{for } k \in \{1, \ldots, n\}$$

where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$ and $\alpha_k \in \mathbb{R}$.

The rows of $\mathbf{Q}_{4\times 2}^{\mathbf{C}}$ lie on a circle centered at the origin, see panel B. This does not generalize to arbitrary cyclic matrices. However, the geometry of antisymmetric matrices has interesting connections with areas of parallelograms and the complex plane, see next subsection.

F.2 Areas and phases

Observe that

$$(u_1 \quad u_2)\begin{pmatrix} 0 & \lambda \\ -\lambda & 0 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \lambda \cdot (u_1 v_2 - u_2 v_1), \tag{2}$$

which is λ times the signed area of the parallelogram covered by $(0,0)$, $\mathbf{u}$, $\mathbf{u} + \mathbf{v}$, and $\mathbf{v}$. It follows that the Schur decomposition breaks $\mathbb{R}^n$ into an orthogonal collection of two-dimensional spaces – one for each $\pm\lambda_i$ block – and that the entries $\mathbf{A}_{ij}$ of $\mathbf{A}$ are sums of signed areas of parallelograms, one parallelogram per two-dimensional subspace. The case where there is a single two-dimensional subspace (since the rank of $\mathbf{A}$ is two) is illustrated in figure 4.

Alternatively, introduce complex numbers $w = u_1 + iu_2$ and $z = v_1 + iv_2$, where $w = |w| \cdot e^{i\cdot\phi_w}$ and $z = |z| \cdot e^{i\cdot\phi_z}$ in polar coordinates. Then, (2) can be rewritten

$$(u_1 \quad u_2)\begin{pmatrix} 0 & \lambda \\ -\lambda & 0 \end{pmatrix} = \mathrm{Im}(w \cdot \bar{z}) = |w| \cdot |z| \sin(\phi_w - \phi_z).$$

F.3 Comment on Schur and Hodge

The Schur decomposition is not guaranteed to be compatible with the Hodge decomposition. That is, $\mathrm{grad}(\mathbf{A})$ is not necessarily the span of two rows of the matrix $\mathbf{Q}_{n\times r}$ arising in the Schur decomposition. Roughly, this happens when $\tilde{\mathbf{A}} := \mathbf{A} - \mathrm{grad} \circ \mathrm{div}(\mathbf{A})$ is not a good $(\mathrm{rank}_{\mathbf{A}} - 2)$-approximation to $\mathbf{A}$ in the $\|\cdot\|_2$-norm.

We recommend to first extract the transitive component and then perform the Schur decomposition on $\tilde{\mathbf{A}} = \mathrm{rot}(\mathbf{A})$. Although this may not always be optimal with respect to the $\|\cdot\|_2$-norm, it has the important advantage that $\mathrm{div}(\mathbf{A})$ is readily understood by humans as a measure of average performance.

	Nash probability	Nash average	Uniform average
breakout	0.000	−0.450	−0.457
seaquest	0.000	−0.487	−0.298
tutankham	0.000	−0.798	−0.619
name_this_game	0.000	−0.751	−0.557
robotank	0.000	−0.689	−0.687
gopher	0.000	−0.370	−0.216
space_invaders	0.000	−0.387	−0.218
ms_pacman	0.000	−0.850	−0.500
surround	0.000	−0.807	−0.453
qbert	0.000	−0.541	−0.376
bowling	0.000	−0.553	−0.328
kangaroo	0.000	−0.659	−0.561
krull	0.000	−0.622	−0.669
yars_revenge	0.000	−0.534	−0.245
solaris	0.000	−0.552	−0.249
double_dunk	0.318	−0.354	−0.370
skiing	0.000	−0.576	−0.461
centipede	0.287	−0.354	−0.146
time_pilot	0.000	−0.624	−0.514
demon_attack	0.000	−0.542	−0.359
asteroids	0.000	−0.366	−0.084
berzerk	0.000	−0.611	−0.337
crazy_climber	0.000	−0.610	−0.592
freeway	0.000	−0.952	−0.729
wizard_of_wor	0.000	−0.456	−0.327
zaxxon	0.000	−0.657	−0.429
ice_hockey	0.000	−0.420	−0.357
road_runner	0.000	−0.587	−0.567
fishing_derby	0.000	−0.748	−0.662
alien	0.000	−0.686	−0.353
defender	0.000	−0.549	−0.425
private_eye	0.215	−0.354	−0.079
gravitar	0.000	−0.509	−0.165
beam_rider	0.000	−0.524	−0.377
phoenix	0.000	−0.365	−0.185
assault	0.000	−0.532	−0.336
kung_fu_master	0.000	−0.706	−0.521
enduro	0.000	−0.731	−0.531
montezuma_revenge	0.048	−0.354	−0.074
video_pinball	0.000	−0.431	−0.360
chopper_command	0.000	−0.548	−0.391
boxing	0.000	−0.705	−0.740
battle_zone	0.000	−0.612	−0.431
bank_heist	0.000	−0.652	−0.521
pitfall	0.000	−0.356	−0.090
pong	0.000	−0.946	−0.894
asterix	0.133	−0.354	−0.166
atlantis	0.000	−0.463	−0.283
star_gunner	0.000	−0.410	−0.363
amidar	0.000	−0.494	−0.288
frostbite	0.000	−0.594	−0.288
hero	0.000	−0.606	−0.350
venture	0.000	−0.571	−0.276
tennis	0.000	−0.514	−0.440

Figure 5: Evaluation of environments.

	Nash probability	Nash average	Uniform average
DQN_(w/o_MC)	0.000	0.030	0.189
DQN_(with_MC)	0.000	0.115	0.191
DQN-PixelCNN_(w/o_MC)	0.000	0.022	0.161
DQN-PixelCNN_(with_MC)	0.000	0.148	0.212
DQN	0.000	0.132	0.343
A3C	0.000	0.149	0.426
DDQN	0.000	0.244	0.556
PriorDDQN	0.000	0.213	0.543
DuelDDQN	0.034	0.354	0.600
DistribDQN	0.000	0.185	0.400
NoisyDQN	0.297	0.354	0.755
Rainbow	0.000	0.314	0.122
RANDOM	0.000	0.012	0.032
HUMAN	0.328	0.354	0.470
DQN_	0.000	0.132	0.343
DDQN_	0.000	0.149	0.426
DUEL	0.000	0.213	0.543
PRIOR	0.135	0.354	0.529
PRIOR_DUEL	0.000	0.214	0.590
PopArt	0.206	0.354	0.457

Figure 6: Evaluation of agents. Note, there are redundancies since agents are taken from multiple papers; these are ignored by Nash averaging.