Neural Interactive Proofs

Lewis Hammond [*] [1] **Sam Adam-Day** [*] [1]

Abstract

We consider the problem of how a trusted, but computationally bounded agent (a 'verifier') can *learn* to interact with one or more powerful but untrusted agents ('provers') in order to solve a given task without being misled. More specifically, we study the case in which agents are represented using neural networks and refer to solutions of this problem as *neural interactive proofs*. First we introduce a unifying framework based on prover-verifier games (Anil et al., 2021), which generalises previously proposed interaction 'protocols'. We then describe several new protocols for generating neural interactive proofs, and provide a (theoretical) comparison of both new and existing approaches. In so doing, we aim to create a foundation for future work on neural interactive proofs and their application in building safer AI systems.

1. Introduction

Recent years have witnessed the proliferation of large machine learning (ML) systems (Villalobos et al., 2022), useful for solving an increasingly wide range of problems. Often, however, it can be difficult to trust the output of these systems, raising concerns about their safety and limiting their applicability to high-stakes situations (Amodei et al., 2016; Bengio et al., 2023; Hendrycks et al., 2023). At the same time, traditional approaches in verification do not scale to today's most powerful systems (Seshia et al., 2022). There is thus a pressing need to identify new angles via which to gain such assurances.

In response to this need, we take inspiration from *interactive proofs* (IPs) (Goldwasser et al., 1985), one of the most important developments in computational complexity theory and cryptography. In an IP, a computationally bounded but trustworthy *verifier* agent interacts with a more powerful but untrustworthy *prover* agent in order to solve a given problem (Figure 1). Given reasonable assumptions, it can be shown that such interactions allow the verifier to solve many more kinds of problem than it could alone, all while limiting the chance of being misled by the prover.

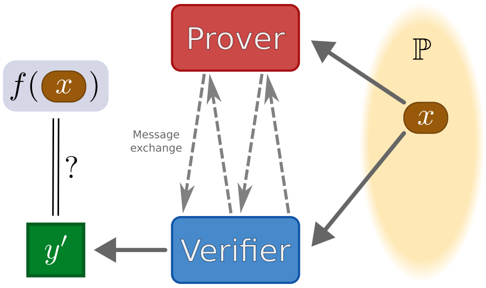

Figure 1. On receiving input x from distribution $\mathbb{P}$ the prover and verifier exchange messages and the verifier eventually decides on an output y', which is compared to $f(x)$.

In this work, we investigate *neural* interactive proofs, in which the prover and verifier agents are represented using neural networks. While a small handful of similar proposals have been made in recent years (Irving et al., 2018; Anil et al., 2021; Wäldchen et al., 2022), these existing approaches are limited in the strength of their guarantees.

One plausible assumption about the future of advanced AI is that we will have access to trusted weaker models and untrusted stronger models (Shlegeris, 2023). This insight is core to many proposals for scalable oversight, which at present is one of the main agendas in ensuring the safety of advanced AI (Bowman et al., 2022; Burns et al., 2023). The present paper follows in this vein.

1.1. Contributions

In this (ongoing) work, we seek to provide the first comprehensive treatment of neural interactive proofs. In particular, we provide the following contributions: (i) a unifying framework that generalises existing neural IP models; (ii) several new neural IP models, including those that allow for zero–knowledge proofs; (iii) a (theoretical) comparison of both new and existing models. In so doing, we hope to create a

[*] Equal contribution [1] Department of Computer Science, University of Oxford, Oxford, United Kingdom. Correspondence to: Lewis Hammond <lewis.hammond@cs.ox.ac.uk>, Sam Adam-Day <me@samadamday.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

foundation for future work on neural interactive proofs and their application in building safer ML systems.

1.2. Related Work

The most closely related work to ours is that of Anil et al. (2021), who introduce *prover-verifier games* played between neural networks, which we generalise and build on. While an important first step, this work is limited in the strength of the protocols that result from their model (as we show further below), which is only applied to very small problem instances. In this paper, we overcome the first limitation (our forthcoming work overcomes the second). Very closely related are the works of Irving et al. (2018); Brown-Cohen et al. (2023) and Wäldchen et al. (2022), whose protocols make use of two provers in competition with one another. We compare such *multi-prover* protocols (Ben-Or et al., 1988) against one another (in the context of learning agents), and against single-prover protocols.

Another departure from these works is that we explicitly study the difficulty of the learning problem faced by the verifier. We can thus be seen as building on earlier models of computationally bounded agents in game-theoretic settings (Papadimitriou & Yannakakis, 1994; Chang, 2006; Halpern & Pass, 2008; Orton, 2021), though these do not consider learning. Relatedly, Goldwasser et al. (2020) recently introduced interactive proofs for *PAC verification*, which is similar in spirit to our work, but the verifier protocols they consider are hand-crafted and only applied to simple ML models. In contrast, we take inspiration from Gowal et al. (2019) and hypothesise that such ideas can best be scaled to real-world ML systems if the verifier can *learn* the protocol.

Other work on the verification of neural networks faces similar scalability problems, with today's techniques typically only applying to networks with hundreds of thousands of parameters and relatively simple properties (Albarghouthi, 2021), as opposed to the hundreds of billions present in state-of-the-art models (Villalobos et al., 2022). Most of these techniques aim to provide proofs of properties such as the robustness of models to small perturbations of their inputs. Alternative directions such as *proof of learning* (Jia et al., 2021) and *proof of inference* (Ghodsi et al., 2017) aim to verify that a given model is the result of a given training process, or that a given output is the result of running a given model on a given input, respectively. In contrast, we aim to verify, not assume, that the prover implements a function that solves the given problem (i.e. the verifier might be misled if it blindly copies the prover).

2. Preliminaries

We begin with some brief technical background on interactive proofs and games.

2.1. Proof Protocols

Definition 2.1 (Goldwasser et al., 1985; Goldreich, 2001). Given $S \subseteq X$, an *interactive proof protocol* for S is a pair $\langle p, v \rangle$ where p is a *prover*, defined as a probability distribution on messages $m_{t+1} \sim p(M^p \mid x, \boldsymbol{m}_{1:t})$ and v is a *verifier*, defined as $m_{t+1} \sim v(M^v \mid x, \boldsymbol{m}_{1:t})$, from message spaces M^v and M^p respectively, and where $\boldsymbol{m}_{i:j} \coloneqq (m_i, \dots, m_j)$. The sequence length T is determined by the verifier, whose eventual output is $m_T \in \{0, 1\} \subseteq M^v$, denoting 'reject' or 'accept'. We denote the (stochastic) sequence of messages $\boldsymbol{m}$ produced by $\langle p, v \rangle$ on input x as $\langle p, v \rangle(x)$. We say that $\langle p, v \rangle$ is *valid* if it satisfies, for every $x \in S$, where $\epsilon_{\mathrm{c}} + \epsilon_{\mathrm{s}} < 1$:

- **Completeness**: If $x \in S$, then $\langle p, v \rangle(x)_T = 1$ with probability at least $1 - \epsilon_{\mathrm{c}}$,
- **Soundness**: If $x \notin S$, then $\langle p', v \rangle(x)_T = 0$ with probability at least $1 - \epsilon_{\mathrm{s}}$ for any prover p'.

The classes of decision problems S for which there exists a valid interactive proof protocol depends on the power of the prover and verifier. For example, in the the original formulation due to Goldwasser et al. (1985), the prover is unbounded and the verifier is a probabilistic polynomial time Turing machine, which gives rise to the class IP. If we instead only require the protocol to be sound in the face of efficiently implementable provers (i.e. computational soundness), this gives rise to *arguments* (as opposed to proofs) (Brassard et al., 1988).

Definition 2.2 (Goldwasser et al., 1985; Goldreich, 2001). We say that $\langle p, v \rangle$ is *(ϵ_k-statistically) zero-knowledge* if for every verifier v' there is some verifier z such that $\max_{x \in S} \frac{1}{2} \sum_{\boldsymbol{m}} \Big| \mathbb{P}\left(\langle p, v' \rangle(x) = \boldsymbol{m}\right) - \mathbb{P}\left(z(x) = \boldsymbol{m}\right) \Big| \leqslant \epsilon_{\mathrm{k}}$. We call z a *simulator*.

While *validity* can be viewed as a property of the verifier, being *zero-knowledge* can be viewed as a property of the prover. Intuitively, $\langle p, v \rangle$ is zero-knowledge if the verifier learns only whether $x \in S$ and nothing else, i.e. v' does not gain any additional power through their interaction with p.

2.2. Games

In this work, we study n-player games $\mathcal{G} = (\Sigma, \mathcal{L})$ where $\Sigma \coloneqq \Sigma^1 \times \cdots \times \Sigma^n$ is a product strategy space and $\mathcal{L}$ consists of loss functions $\mathcal{L}^i : \Sigma \to \mathbb{R}$. Each player i selects a strategy $\sigma^i \in \Sigma^i$ in an attempt to minimise their loss $\mathcal{L}^i(\sigma)$. We use $\mathcal{G}(\sigma^i)$ to denote the $(n-1)$-player game induced when player i plays strategy σ^i in $\mathcal{G}$, but where the remaining $n-1$ players have not yet chosen their strategies. In practice, we assume that each player's strategy space Σ^i is defined by some finite number of parameters $\boldsymbol{\Theta}^i$, and will often refer to $\boldsymbol{\theta}^i \in \boldsymbol{\Theta}^i$ instead of σ^i. Within these games, we make use of two standard equilibrium concepts.

Definition 2.3. A *local Nash equilibrium* (LNE) on $\hat{\boldsymbol{\Theta}} \subseteq \boldsymbol{\Theta}$ is a strategy profile $\boldsymbol{\theta}_\star \in \hat{\boldsymbol{\Theta}}$ such that:

$$\boldsymbol{\theta}^i_\star \in \operatorname*{argmin}_{\boldsymbol{\theta}^i \in \hat{\boldsymbol{\Theta}}^i} \mathcal{L}^i(\boldsymbol{\theta}^i, \boldsymbol{\theta}^{-i}_\star),$$

for all $i \in [n]$. If $\hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}$ then $\boldsymbol{\theta}_\star$ is a (global) Nash equilibrium (NE). We denote the local and global NEs of G by $\mathrm{LNE}(G)$ and $\mathrm{NE}(G)$ respectively.

Definition 2.4. A *local Stackelberg equilibrium* led by player i (LSE_i) on $\hat{\boldsymbol{\Theta}} \subseteq \boldsymbol{\Theta}$ is a strategy profile $\boldsymbol{\theta}_\star \in \hat{\boldsymbol{\Theta}}$ such that:

$$\boldsymbol{\theta}^i_\star \in \operatorname*{argmin}_{\boldsymbol{\theta}^i \in \hat{\boldsymbol{\Theta}}^i} \max_{\boldsymbol{\theta}^{-i}_\star \in \mathrm{LNE}(G(\boldsymbol{\theta}^i))} \mathcal{L}^i\big(\boldsymbol{\theta}^i, \boldsymbol{\theta}^{-i}_\star\big).$$

If $\hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}$ then $\boldsymbol{\theta}_\star$ is a (global) Nash equilibrium (NE). We denote the local and global i-led SEs of G by $\mathrm{LSE}_i(G)$ and $\mathrm{SE}_i(G)$ respectively.

More specifically, we consider *approximate* versions of these concepts, where the argmin for each agent i has some tolerance $e^i \in \mathbb{R}_{\geqslant 0}$. Given a vector $\boldsymbol{e} = (e^i, \ldots, e^n)$, we denote the approximate equilibria as $\boldsymbol{e}$-NE and $\boldsymbol{e}$-SE, respectively.

3. Prover-Verifier Games

We consider the problem of how a trusted but computationally bounded verifier can learn to interact with one or more powerful but untrusted provers in order to solve a given task. We represent this task as a function $f : X \to Y$ and a distribution $\mathbb{P}(X)$, whereupon receiving and input $x \in X$, the verifier interacts with the prover in order generate a 'proof' that $y' = f(x)$, for some $y' \in Y$. Importantly, we assume that: (a) the verifier is not capable of solving the task alone; (b) the prover *is* capable of solving the task alone; but (c) the prover may not be perfectly cooperative or aligned with the verifier, and so cannot be trusted to provide the correct answer without some form of proof.

In the remainder of this section, we make the above setting more concrete by introducing a generalisation of the prover-verifier game model originally introduced by Anil et al. (2021). This generalised definition is sufficiently broad so as to capture several other models (Irving et al., 2018; Wäldchen et al., 2022), as well as the new models that we introduce in this paper.

Definition 3.1. A (generalised) *prover-verifier game* (PVG) for a decision problem $(f, \mathbb{P})$ is a game $\mathcal{G} = (\Sigma, \mathcal{L})$ comprising strategy sets and loss functions for n_p provers and n_v verifiers, where $n_p + n_v = n$. Given $x \sim \mathbb{P}$, play proceeds according to a (possibly stochastic) *messaging specification* μ that determines the set of players whose turn it is at each time t. When $i \in \mu_t$, player i samples a message $m^i_t \sim \sigma^i(M^i \mid M^*, X) \in \Sigma^i$ given x and the previous messages $\boldsymbol{m}_{1:t-1} \in M^*$, where $M = \bigcup_{i \in [n]} M^i$. Play terminates at timestep T if and only if $m^i_T \in Y$ where $i \in \mu_T$ is a verifier.

Different PVGs represent different messaging specifications between the prover(s) and verifier(s), with the basic idea being that we wish to construct a game such that its equilibria correspond to valid proof protocols. For example, Anil et al. (2021) introduce the following model, which they refer to as an 'Abstract Decision Problem' (`adp`). In this model, the prover sends a single message to the verifier, and the verifier must make its decision in response. Note that when there is just one prover or verifier, we denote their index $i \in [n]$ as p or v, respectively. For games with multiple provers or verifiers, we let $\{p_1, \ldots, p_{n_p}, v_1, \ldots, v_{n_v}\} = [n]$.

Definition 3.2 (Anil et al., 2021)**.** The `adp` model is a PVG where $Y = \{0, 1\}$ and $n_p = n_v = 1$. Σ^p is given by a set of deterministic distributions $\sigma^p(m^p \mid x)$ – i.e. functions $\delta^p : X \to M^p$ – and Σ^v contains the convex combinations of functions $\delta^v : X \times M^p \to Y$. The specification μ is such that $\mu_0 = \{p\}$ and $\mu_1 = \{v\}$ and the loss functions are given by:

$$\mathcal{L}^p(\boldsymbol{\sigma}) = -\mathbb{E}\big[\log \sigma^v(1 \mid x, m^p)\big],$$
$$\mathcal{L}^v(\boldsymbol{\sigma}) = -\mathbb{E}\big[\log \sigma^v(y \mid x, m^p)\big],$$

where the expectation is taken with respect to $x \sim \mathbb{P}$ and $m^p \sim \sigma^p(\cdot \mid x)$.

Other forms of protocol can be characterised, for example, as a competition between two provers, which is judged by the verifier. Formal definitions of the `debate` model (Irving et al., 2018; Brown-Cohen et al., 2023) and the 'Merlin-Arthur Classifier' (`mac`) model (Wäldchen et al., 2022) can be found in Appendix A. Limitations of the `adp` model are also discussed in Appendix A.

4. Neural Interactive Proofs

Unlike in the case of standard interactive proofs, we assume that the verifier has some access to a small dataset of labelled examples and aims to *learn* a strategy for interacting with the prover that generalises to any input. More formally, we assume that the verifier has access to a dataset $\mathcal{D} = \{(x_j, y_j)\}_{1 \leqslant j \leqslant m}$ sampled $x_j \sim \mathbb{P}$ and $y_j = f(x_j)$. Letting $\ell(\boldsymbol{\sigma}, x) := \mathbb{E}[\boldsymbol{\sigma}(x)_T \neq y]$, where $\boldsymbol{\sigma}(x)_T$ is the output of the strategy profile $\boldsymbol{\sigma}$ in a PVG $\mathcal{G}$, the verifier's aim is to minimise ℓ across inputs $x \in X$.

In this work, we consider two ways of operationalising this aim: the expected loss (or *risk*), and the worst-case loss. We denote the empirical risk with respect to $\mathcal{D}$ as $\mathcal{L}^{\mathrm{ER}}_{\mathcal{D}}(\boldsymbol{\sigma}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} \ell(\boldsymbol{\sigma}, x)$, and the empirical worst-case loss by $\mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}) = \max_{(x,y)\in\mathcal{D}} \ell(\boldsymbol{\sigma}, x)$. Minimisers of these losses are denoted $\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}$ and $\boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}$ respectively. We drop $\mathcal{D}$ from our notation when considering the full space inputs X.

While the expected loss is by far the most common objective in contemporary ML, from the perspective of safety-critical systems, we would ideally like to minimise the worst-case loss. In other words, we wish to find some $\boldsymbol{\sigma}$ such that $\mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}) - \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{WC}})$ is small. Moreover, this is despite the fact that $\min_{\sigma^v} \mathcal{L}^{\text{WC}}(\sigma^v)$ is large, denoting the case in which the verifier acts alone as best it can.

Taking inspiration from the adp model above, let us consider a PVG with $Y = \{0, 1\}$, $n_p = n_v = 1$, $\mu_t = \{t \mod 2\}$, and loss functions given by:

$$\mathcal{L}^p(\boldsymbol{\sigma}) = \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma} \mid f(x) = 1) - \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma} \mid f(x) = 0),$$
$$\mathcal{L}^v(\boldsymbol{\sigma}) = \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma} \mid f(x) = 1) + \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma} \mid f(x) = 0).$$

We refer to this model as nip, and it is straightforward to show that this corresponds closely to the notion of an interactive proof.

Theorem 4.1. *Let $(f, \mathbb{P})$ be a probabilistic decision problem which has a valid proof protocol and $\mathcal{G}$ a* nip *game. Then $\boldsymbol{\sigma}$ is a valid IP protocol if and only if it is a (strict) approximate verifier-leading Stackelberg equilibrium of $\mathcal{G}$.*

While this proposition essentially reduces the problem of finding a valid proof protocol to the task of finding an equilibrium in a given game, this task is far from straightforward. In particular, there are two key difficulties. Firstly, there is the challenge of learning to minimise the *worst-case* (as opposed to the expected) loss. Secondly, there is the challenge of finding a *Stackelberg* equilibrium. In what follows, we address each of these difficulties in turn.

4.1. Worst-Case Loss

There are several techniques for minimising the worst-case loss using finitely many data $\mathcal{D}$ generated from $\mathbb{P}$. The best approach will depend on the nature of the decision problem, and the affordances of those training the relevant models.

4.1.1. Empirical Risk Minimisation

The simplest approach is to simply return some $\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}$. The question then becomes: when is minimising the empirical risk with respect to $\mathcal{D}$ sufficient for minimising the worst-case risk with respect to X? The following result shows that we can break this down into two properties: (a) the empirical worst-case loss being similar to the actual worst-case loss; and (b) for a given $\mathcal{D}$, the empirical worst-case loss of $\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}$ being similar to that of $\boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}$. Worst-case uniform convergence and robustness do not always hold, but can do when the decision problem is sufficiently 'regular'.

Definition 4.2. Σ has the *worst-case uniform convergence* property if there is $m^{\text{WCUC}} : (0, 1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0, 1)$, if $|\mathcal{D}| \geqslant m^{\text{WCUC}}(\epsilon, \delta)$ then $\mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}) - \mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}) \leqslant \epsilon$ for all $\boldsymbol{\sigma}$, with probability $1 - \delta$.

Definition 4.3. Σ has the *worst-case robustness* property if there is $m^{\text{WCR}} : (0, 1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0, 1)$, if $|\mathcal{D}| \geqslant m \geqslant m^{\text{WCR}}(\epsilon, \delta)$ then $\mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}) - \mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}) \leqslant \epsilon$ with probability $1 - \delta$.

Theorem 4.4. *If Σ has the worst-case uniform convergence property and the worst-case robustness property then there is some $m^{WCUC} : (0, 1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0, 1)$, if $|\mathcal{D}| \geqslant m^{WC}(\epsilon, \delta)$ then $\mathcal{L}^{WC}(\boldsymbol{\sigma}^{ER}_{\mathcal{D}}) - \mathcal{L}^{WC}(\boldsymbol{\sigma}^{WC}) \leqslant \epsilon$ with probability $1 - \delta$.*

4.1.2. Adversarial Training

One of the most natural ways to optimise the worst-case loss is to introduce an adversary. This can be done using a third agent, a, whose strategy space is $X_0 \times X_1$ (where $X_j = \{x \mid f(x) = j\}$) and whose loss function is $\mathcal{L}^a(\boldsymbol{\sigma}, (x_0, x_1)) = -\ell(\boldsymbol{\sigma}, x_0) - \ell(\boldsymbol{\sigma}, x_1)$. We then replace the terms $\mathcal{L}^{\text{WC}}(\boldsymbol{\sigma} \mid f(x) = i)$ in the original loss functions for the prover and verifier with with $\ell(\boldsymbol{\sigma}, x_1) - \ell(\boldsymbol{\sigma}, x_0)$ and $\ell(\boldsymbol{\sigma}, x_1) + \ell(\boldsymbol{\sigma}, x_0)$ respectively. The verifier-leading Stackelberg equilibria of the original nip game are then identical to the verifier-prover-leading Stackelberg equilibria in this new three-player game, denoted G^a.

Theorem 4.5. *Let $(f, \mathbb{P})$ be a probabilistic decision problem and $\mathcal{G}$ a* nip *game. Then (σ^p, σ^v) is an approximate verifier-leading SE ($\boldsymbol{e}$-SE$_v$) of $\mathcal{G}$ if and only if there is some σ^a such that $(\sigma^p, \sigma^v, \sigma^a)$ is an approximate verifier-prover SE ($\boldsymbol{e}$-SE$_{v,p}$) of G^a (the adversarial version of $\mathcal{G}$).*

Unlike the classical learning-theoretic approach above in which we assumed we were given a fixed dataset $\mathcal{D}$ of (x, y) pairs, we are here assuming access to an adversary capable of outputting any $x \in X$. This stronger assumption may not always hold, but when it does, learning can be more efficient (Goldman & Kearns, 1995).

4.2. Solving Stackelberg Games

In this subsection we consider two ways solving for Stackelberg equilibria using gradient-based methods.

4.2.1. Bi-Level Optimisation

Computing Stackelberg equilibria can be naturally modelled as a bi-level optimisation problem. A standard solution to such problems using gradient-based methods is to employ a timescale separation, meaning that some of the variables are updated faster than the others (Borkar, 2008). In particular, we take the sequential nature of the problem setting into account by explicitly modelling the dependence of $\boldsymbol{\theta}^p$ on $\boldsymbol{\theta}^v$ and updating $\boldsymbol{\theta}^p$ more quickly as part of an 'inner loop'.

Fiez et al. (2020) show that if $\alpha^v = o(\alpha^p)$ then with high probability the following dynamics will converge locally to the neighbourhood of a LSE_v:

$$\boldsymbol{\theta}^p_{t+1} = \boldsymbol{\theta}^p_t - \alpha^p(t) \cdot \nabla_p \mathcal{L}^p,$$

$$\boldsymbol{\theta}^v_{t+1} = \boldsymbol{\theta}^v_t - \alpha^v(t) \cdot \nabla_v \mathcal{L}^v - \nabla_p \mathcal{L}^v \left(\nabla^2_p \mathcal{L}^p\right)^{-1} \nabla_{pv} \mathcal{L}^p,$$

where we drop the dependence on $\boldsymbol{\theta}$ from our notation and write ∇_v and ∇_p for $\nabla_{\boldsymbol{\theta}^v}$ and $\nabla_{\boldsymbol{\theta}^p}$, respectively.

4.2.2. Opponent Shaping

The updates above require computing an inverse Hessian vector product, which is intractable when $\boldsymbol{\theta}^p$ is large. Replacing the term $\left(\nabla^2_p \mathcal{L}^p\right)^{-1}$ with $\alpha^p(t+1)$ leads to the LOLA (Learning with Opponent Learning Awareness) update (Foerster et al., 2018), which aims to actively influence the future policy updates of its opponents. While LOLA may fail to converge, interpolating between the LOLA update and LookAhead (Zhang & Lesser, 2010) – an algorithm known as *Stable Opponent Shaping (SOS)* – leads to local convergence to stable fixed points in differentiable games under self-play (Letcher et al., 2019).

5. Extensions

Finally, we generalise the `nip` model along two natural dimensions in order to strengthen the properties of the resulting protocols.

5.1. Multiple Provers

Multi-prover interactive proofs (MIPs) are a natural generalisation of classical IPs (Ben-Or et al., 1988), whose additional power results from the fact that while the two provers may correlate their strategies, they are prevented from communicating with one another during their interactions with the verifier (Babai et al., 1991). This allows the verifier to 'cross-examine' the provers.

We define the `nmip` model identically to the `nip` model, but now with two provers, p_1 and p_2, each of which has the same loss. Valid MIP protocols are defined as in Definition 2.1, with the soundness condition altered such that v must be robust to any choice of p'_1, p'_2. Using a similar proof to that of Theorem 4.1, it can be shown that the equilibria of the `nmip` PVG correspond to valid MIP protocols. The only subtlety is that due to the provers' ability to coordinate on a joint strategy and shared random signal beforehand, we must consider *correlated* equilibria (defined in Appendix B).

Theorem 5.1. *Let $(f, \mathbb{P})$ be a probabilistic decision problem which has a valid proof protocol and $\mathcal{G}$ a* `nmip` *game. Then $\boldsymbol{\sigma}$ is a valid MIP protocol if and only if it is a (strict) approximate verifier-leading* correlated *Stackelberg equilibrium of $\mathcal{G}$.*

5.2. Zero Knowledge

We now consider whether the prover can prevent the verifier from learning any *more* than to solve the task at hand. We formalise this question as whether the resulting proof protocol is *zero-knowledge* (ZK). Far from a theoretical curiosity, this may have important practical implications, such as for the prevention of model cloning/stealing via black-box access (Orekondy et al., 2019).

Recall from Definition 2.2 that $\langle p, v\rangle$ is ZK if for any v' and any x such that $f(x) = 1$, there is another weak agent $z \in V$ capable of generating approximately the same distribution over message sequences as $\langle v', p\rangle(x)$, *without* interacting with p. While it is possible to introduce a precise analogue of this condition, we wish to investigate whether the verifier can gain knowledge regardless of the value of $f(x)$.

We therefore provide a slightly stronger condition that requires statistical closeness between $\langle v', p\rangle(x, y)$ and $z(x, y)$ for *any* x. In other words, short of the answer y, there is nothing else that v' can learn by interacting with p. We can implement this model using a PVG by introducing an additional verifier agent $v_2 = v'$ and a simulator $v_3 = z$, whose objectives $\mathcal{L}^{v_2}$ and $\mathcal{L}^{v_3} := -\mathcal{L}^{v_2}$ are defined via:

$$\mathcal{L}^{v_3}(\boldsymbol{\sigma}) := -\max_x \frac{1}{2} \sum_{\boldsymbol{m}} \left| \begin{array}{l} \mathbb{P}\left(\langle \sigma^p, \sigma^{v_2}\rangle(x, y) = \boldsymbol{m}\right) \\ -\,\mathbb{P}\left(\sigma^{v_3}(x, y) = \boldsymbol{m}\right) \end{array} \right|,$$

and by adding the term $\mathcal{L}^{v_3}$ to the prover's loss $\mathcal{L}^p$. We refer to the resulting model as `zk-nip`, and show that the equilibria of this game correspond to ZK proof protocols.

Theorem 5.2. *Let $(f, \mathbb{P})$ be a probabilistic decision problem which has a valid ϵ_k-statistically zero-knowledge protocol and $\mathcal{G}$ a* `zk-nip` *game. Then $\boldsymbol{\sigma}$ is a valid ϵ_k-zero-knowledge IP protocol if it is a (strict) approximate verifier-prover-leading Stackelberg equilibrium ($\boldsymbol{e}$-$\mathrm{SE}_{v_1,p}$) of $\mathcal{G}$.*

We note that in this model the equilibria contain a Stackelberg hierarchy with more than two levels. Namely, the prover p commits to their strategy before agents v_2 and v_3. Technically, v_2 should also be viewed as committing before v_3, but $\mathcal{G}(\sigma^{v_1}, \sigma^p)$ is a zero-sum game and so reaching an NE is sufficient.

6. Discussion

In this work, we provide the first unifying framework for neural interactive proof protocols and conduct a theoretical analysis, including introducing several new models. A natural next step is to conduct a complementary empirical investigation and a comparison of the real-world efficacy of these models. This challenge is the subject of our forthcoming work.

Impact Statement

The aim of this work is to advance efforts towards building safer and more trustworthy systems. We therefore expect (and hope) that it will contribute towards positive societal benefits, as elaborated in the main body.

References

Albarghouthi, A. Introduction to neural network verification. *Foundations and Trends in Programming Languages*, 7 (1–2):1–157, 2021.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv:2108.12099*, 2016.

Anil, C., Zhang, G., Wu, Y., and Grosse, R. Learning to give checkable answers with prover-verifier games. *arXiv:2108.12099*, 2021.

Babai, L., Fortnow, L., and Lund, C. Non-deterministic exponential time has two-prover interactive protocols. *Computational Complexity*, 1(1):3–40, 1991.

Ben-Or, M., Goldwasser, S., Kilian, J., and Wigderson, A. Multi-prover interactive proofs: How to remove intractability assumptions. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, pp. 113–131, New York, NY, USA, 1988. Association for Computing Machinery.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahneman, D., Brauner, J., and Mindermann, S. Managing AI risks in an era of rapid progress. *arXiv:2310.17688*, 2023.

Borkar, V. S. *Stochastic Approximation*. Hindustan Book Agency, 2008.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Kernion, J., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lovitt, L., Elhage, N., Schiefer, N., Joseph, N., Mercado, N., DasSarma, N., Larson, R., McCandlish, S., Kundu, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Mann, B., and Kaplan, J. Measuring progress on scalable oversight for large language models. *arXiv:2211.03540*, 2022.

Brassard, G., Chaum, D., and Crépeau, C. Minimum disclosure proofs of knowledge. *Journal of Computer and System Sciences*, 37(2):156–189, 1988.

Brown-Cohen, J., Irving, G., and Piliouras, G. Scalable AI safety via doubly-efficient debate. *arXiv:2311.14125*, 2023.

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., and Wu, J. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv:2312.09390*, 2023.

Chang, C.-L. On the computational power of players in two-person strategic games. Master's thesis, National Taiwan University, 2006.

Fiez, T., Chasnov, B., and Ratliff, L. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 3133–3144, 2020.

Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 122–130, 2018.

Fürer, M., Goldreich, O., Mansour, Y., Sipser, M., and Zachos, S. On completeness and soundness in interactive proof systems. *Advances in Compututing Research*, 5: 429–442, 1989.

Ghodsi, Z., Gu, T., and Garg, S. Safetynets: Verifiable execution of deep neural networks on an untrusted cloud. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pp. 4675–4684, 2017.

Goldman, S. and Kearns, M. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995.

Goldreich, O. *Foundations of Cryptography: Basic Tools*. Cambridge University Press, 2001.

Goldwasser, S., Micali, S., and Rackoff, C. The knowledge complexity of interactive proof-systems. In *Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing*, 1985.

Goldwasser, S., Rothblum, G. N., Shafer, J., and Yehudayoff, A. Interactive proofs for verifying machine learning. Technical Report 58, Electronic Colloquium Computational Complexity, 2020.

Gowal, S., Dvijotham, K., Stanforth, R., Mann, T., and Kohli, P. A dual approach to verify and train deep networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

Halpern, J. Y. and Pass, R. Game theory with costly computation. *arXiv:0809.0024*, 2008.

Hendrycks, D., Mazeika, M., and Woodside, T. An overview of catastrophic AI risks. *arXiv:2306.12001*, 2023.

Irving, G., Christiano, P., and Amodei, D. AI safety via debate. *arXiv:1805.00899*, 2018.

Jia, H., Yaghini, M., Choquette-Choo, C. A., Dullerud, N., Thudi, A., Chandrasekaran, V., and Papernot, N. Proof-of-learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy (SP)*, 2021.

Letcher, A., Balduzzi, D., Racanière, S., Martens, J., Foerster, J. N., Tuyls, K., and Graepel, T. Differentiable game mechanics. *Journal of Machine Learning Research*, 20 (84):1–40, 2019.

Orekondy, T., Schiele, B., and Fritz, M. Knockoff nets: Stealing functionality of black-box models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Orton, T. Modeling precomputation in games played under computational constraints. In Zhou, Z.-H. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2005–2011, 8 2021.

Papadimitriou, C. H. and Yannakakis, M. On complexity as bounded rationality. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing (STOC)*, 1994.

Seshia, S. A., Sadigh, D., and Sastry, S. S. Toward verified artificial intelligence. *Communications of the ACM*, 65 (7):46–55, 2022.

Shamir, A. IP = PSPACE. *Journal of the ACM*, 39(4): 869–877, 1992.

Shlegeris, B. Untrusted smart models and trusted dumb models. Alignment Forum, 2023. URL https://www.alignmentforum.org/posts/LhxHcASQwpNa3mRNk/untrusted-smart-models-and-trusted-dumb-models. Date accessed: 8 July 2024.

Skalse, J., Hammond, L., Griffin, C., and Abate, A. Lexicographic multi-objective reinforcement learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3430–3436, 2022.

Villalobos, P., Sevilla, J., Besiroglu, T., Heim, L., Ho, A., and Hobbhahn, M. Machine learning model sizes and the parameter gap. *arXiv:2207.02852*, 2022.

Wäldchen, S., Sharma, K., Zimmer, M., and Pokutta, S. Merlin-arthur classifiers: Formal interpretability with interactive black boxes. *arXiv:2206.00759*, 2022.

Zhang, C. and Lesser, V. Multi-agent learning with policy prediction. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pp. 927–934, 2010.

A. Additional Proof Protocols

While a comprehensive comparison is beyond the scope of the present work, we include a brief description and high-level comparison of additional models for neural interactive proofs from the literature, with a summary given in Table 1.

Model	Provers	Verifiers	Steps	Complexity	Zero Knowledge	Reference
adp	1	1	2	NP	✗	(Anil et al., 2021)
debate	2	1	T	PSPACE	✗	(Irving et al., 2018)
mac	2	1	2	?	✗	(Wäldchen et al., 2022)
nip	1	1	T	PSPACE	✗	Ours
nmip	2	1	T	NEXP	✗	Ours
zk-nip	1	3	T	PSPACE	✓	Ours
zk-nmip	2	3	T	NEXP	✓	Ours

Table 1. A comparison between various PVG models.

Under the debate protocol, two provers declare their answers as $m_0 \in Y \subseteq M^1$ and $m_1 \in Y \subseteq M^2$ respectively, and the verifier chooses to accept the answer given by prover y' after witnessing the ensuing debate $\boldsymbol{m}_{2:T-1}$.

Definition A.1 (Irving et al., 2018)**.** The debate model is a PVG where $n_p = 2$ and $n_v = 1$, with $\{1, 2\} \subseteq Y$. The protocol is $\mu_t = \{t \mod 3 + 1\}$ and we let $M^v = \{0, 1, 2\}$. The loss functions are given by:

$$\begin{aligned}
\mathcal{L}^1(\boldsymbol{\sigma}) &= -\,\mathbb{P}(y' = 1), \\
\mathcal{L}^2(\boldsymbol{\sigma}) &= -\,\mathbb{P}(y' = 2), \\
\mathcal{L}^3(\boldsymbol{\sigma}) &= -\,\mathbb{P}(m_{y'-1} = y),
\end{aligned}$$

for the two provers and the verifier, respectively.

The Merlin-Arthur Classifier (mac) protocol is closely related to debate. In this setting, the verifier must output the correct class y for the input (or output ‘not sure’) based on a feature $m_0 = \phi \in \Phi$ provided by either a helpful (player 1) or unhelpful (player 2) prover, not knowing which provided the feature.

Definition A.2 (Wäldchen et al., 2022)**.** The mac model is a PVG where $n_p = 2$ and $n_v = 1$, with $M^v = Y = [k] \cup \{0\}$ representing k classes and a ‘not sure’ answer (0), and $M^1 = M^2 = \Phi$ are a set of features. Unlike debate and adp, the protocol is *stochastic*, with $\mu_0 = i$ for $i \sim \mathsf{unif}(1, 2)$ and $\mu_1 = \{3\}$. The loss functions are given by:

$$\begin{aligned}
\mathcal{L}^1(\boldsymbol{\sigma}) &= -\mathbb{E}\big[\log \sigma^v(y \mid x, m_0)\big], \\
\mathcal{L}^2(\boldsymbol{\sigma}) &= \mathbb{E}\big[\log\big(\sigma^v(y \mid x, m_0) + \sigma^v(0 \mid x, m_0)\big)\big], \\
\mathcal{L}^3(\boldsymbol{\sigma}) &= -(1-\gamma) \cdot \mathbb{E}\big[\log \sigma^v(y \mid x, m_0) \mid \mu_0 = \{1\}\big] \\
&\quad - \gamma \cdot \mathbb{E}\big[\log\big(\sigma^v(y \mid x, m_0) + \sigma^v(0 \mid x, m_0)\big) \mid \mu_0 = \{2\}\big],
\end{aligned}$$

where $\gamma \in [0, 1]$ is a hyperparameter.

For a given PVG model model and solution concept E, we denote the class of decision problems solved by $\langle p, v \rangle \in E(\mathcal{G})$, with provers and verifiers in classes of Turing machines C_p and C_v, respectively, as $\texttt{model}_E(\mathsf{C}_p, \mathsf{C}_v)$. Typically, we consider the case when $\mathsf{C}_p = \mathsf{ALL}$ and $\mathsf{C}_v = \mathsf{PPT}$, where where ALL is the set of all Turing machines and PPT is the class of all probabilistic polynomial time machines. In this setting, we can draw analogies between the PVG models we discuss and the complexity classes they correspond to.

For example, by employing well-known results about the complexity class IP (Shamir, 1992), it follows immediately from Theorem 4.1 that nip corresponds to PSPACE. Irving et al. (2018) similarly prove that debate corresponds to PSPACE. On the other hand, while Anil et al. (2021) show that the SE$_v$s of adp correspond exactly to valid interactive proof protocols (when the verifier is deterministic), the theoretical strength of this result is severely limited due to its stipulation of zero soundness error.

Proposition A.3 (Anil et al., 2021)**.** *Let $(f, \mathbb{P})$ be a (probabilistic) decision problem and $\mathcal{G}$ a* `adp` *game. Suppose that there exists some deterministic $\delta^v_\star$ such that $\exists \delta^p \forall x \big(\langle \delta^p, \delta^v_\star \rangle(x)_T = y \big)$ and $\forall \delta^p \forall x \big(\langle \delta^p, \delta^v_\star \rangle(x)_T = 1 \implies y = 1 \big)$. Then $\langle \delta^p, \sigma^v \rangle$ is a valid interactive proof protocol (with $\epsilon_c = \epsilon_s = 0$) for $\{x : f(x) = 1\}$:*

- *If and only if $\langle \delta^p, \sigma^v \rangle \in \mathrm{SE}_v(G)$,*
- *Only if $\langle \delta^p, \sigma^v \rangle \in \mathrm{NE}(G)$.*

Allowing for a soundness error is widely held to be critical to the power of interactive proofs. Indeed, if a set S has a valid interactive proof protocol with $\epsilon_s = 0$, then $S \in \mathsf{NP}$.[1] Similarly, the restriction to deterministic verifiers is also theoretically significant: if a set S has a valid interactive proof protocol where v is deterministic, then we must also have $\epsilon_s = 0$. Unfortunately, if we relax these assumptions then the correspondence between the SE_vs of an `adp` PVG and valid proof protocols no longer holds.

Proposition A.4. *There is a probabilistic decision problem $(f, \mathbb{P})$ and an* `adp` *game $\mathcal{G}$ such that – even though there exists some valid interactive proof protocol $\langle \delta^p, \sigma^v_\star \rangle$ with $\epsilon_c = 0$ – the fact that $\langle \delta^p, \sigma^v \rangle \in \mathrm{SE}_v(G)$ is neither necessary nor sufficient for $\langle \delta^p, \sigma^v \rangle$ to be valid.*

Proof. Let use consider the specific PVG with $X = \{0, 1, 2, 3\}$ and $f(x) = x \mod 2$, with the following deterministic strategies for the prover (who has message space $M^p = X$):

$$\delta^p_1(x) = x \mod 2 \qquad \delta^p_2(x) = 2 - |x - 2| \qquad \delta^p_3(x) = x,$$

and with the verifier choosing a strategy σ^v that forms a convex combination over:

$$\delta^v_1(x, m^p) = [0 < m^p < 3] \qquad \delta^v_2(x, m^p) = [m^p < 2] \qquad \delta^v_3(x, m^p) = 1,$$

where $[\cdot]$ are Iverson brackets (i.e. an indicator function), and thus the codomain of each δ^v is $y = \{0, 1\}$. We write σ^v explicitly as $(p\delta^v_1, q\delta^v_2, r\delta^v_3)$, where $p + q + r = 1$. Writing these strategies out explicitly we have:

x	0	1	2	3
$\delta^p_1(x)$	1	0	1	0
$\delta^p_2(x)$	0	1	2	1
$\delta^p_3(x)$	0	1	2	3

m^p	0	1	2	3
$\delta^v_1(x, m^p)$	0	1	1	0
$\delta^v_2(x, m^p)$	1	1	0	0
$\delta^v_3(x, m^p)$	1	1	1	1

Let $\sigma^v_\star = (\frac{5}{8}\delta^v_1, \frac{3}{8}\delta^v_2, 0\delta^v_3)$. Then if $y = 1$ (i.e., either $x = 1$ or $x = 3$) we have $\langle \delta^p_1, \sigma^v_\star \rangle(x)_T = y$ w.p. 1, and hence $\epsilon_c = 0$. Similarly, note that that for any δ^p, we have that $\langle \delta^p, \sigma^v_\star \rangle(x)_T = 1$ w.p. $\leqslant \frac{5}{8} =: \epsilon_s$ for any x such that $f(x) = 0$ (i.e., either $x = 0$ or $x = 2$). As $\epsilon_c + \epsilon_c = \frac{5}{8} < 1$, then $\langle \delta^p_1, \sigma^v_\star \rangle$ is a valid proof protocol.

Suppose, for the remainder of the proof, that $\mathbb{P}(0) = \mathbb{P}(1) = \mathbb{P}(2) = a < \frac{1}{3}$ and $\mathbb{P}(3) = 1 - 3a$. First, we show lack of necessity. By the reasoning above, we have that $(\delta^p_1, \sigma^v_\star)$ is a sound and complete interactive proof protocol for $\{x : f(x) = 1\}$. But under the loss functions for `adp` $\mathcal{L}^p(\delta^p_1, \sigma^v_\star) = a \log \frac{64}{9}$ while $\mathcal{L}^p(\delta^p_2, \sigma^v_\star) = a \log \frac{64}{15}$, and so the prover can beneficially deviate by switching to δ^p_2. Thus, $(\delta^p_1, \sigma^v_\star)$ is not an SE_v.

Second, we show lack of sufficiency. As we explain further below, the unique SE_v of the resulting PVG is given by $(\delta^p_3, \sigma^v_\dagger)$, where $\sigma^v_\dagger = (b\delta^v_1, b\delta^v_2, (1 - 2b)\delta^v_3)$ and $b = \frac{3a}{2}$. Under this equilibrium, however, we have that $\langle \delta^p_3, \sigma^v_\dagger \rangle(1)_T = f(1) = 1$ w.p. $2b$ (hence $\epsilon_c = 1 - 2b$) and $\langle \delta^p_3, \sigma^v_\dagger \rangle(0)_T = 1 \neq f(0)$ w.p. $1 - b$ (hence $\epsilon_s = 1 - b$). Therefore, we have $\epsilon_c + \epsilon_c = 2 - 3b$, and so $\langle \delta^p_3, \sigma^v_\dagger \rangle$ is valid if and only if $b > \frac{1}{3}$. But because $b = \frac{3a}{2}$, this is false for any $a \leqslant \frac{2}{9}$. In such cases, being an SE_v is insufficient for validity, completing the proof.

The intuition behind the equilibrium $(\delta^p_3, \sigma^v_\dagger)$ is that the larger the probability mass on the setting when $x = 3$ (i.e. the smaller a is) the more the verifier (and also the prover, as $f(3) = 1$) has an overriding incentive to make sure that it outputs the correct answer in this particular case. Because $\langle \delta^p, \delta^v \rangle(3)_T = 0$ if $\delta^p = \delta^p_1$ or $\delta^p = \delta^p_2$ (for any δ^v), the verifier is thus incentivised to encourage the prover to play δ^p_3. The only way the prover can lower its loss by playing δ^p_3 is if the verifier plays δ^v_3 with high probability.

[1] On the other hand, having non-zero completeness error still results in IP (Fürer et al., 1989).

Given that δ_3^v is being played with some probability, then the loss from extra probability mass placed on δ_1^v or δ_2^v is symmetric, hence we only need to find the probability of the verifier playing δ_3^v. The negative loss for the verifier is given by:

$$a \log b + a \log(2b) + a \log b + (1-3a)\log(1-2b).$$

Given that we must have $0 < b < \frac{1}{2}$ to avoid infinite loss, the verifier-optimal choice of b can be found analytically by solving for the stationary points of the expression above with respect to b, resulting in the solution $b = \frac{3a}{2}$. □

B. Proofs

In this section, we provide proofs for the results presented in the main text.

B.1. Protocol Correspondences

We first prove the correspondence result for `nip` games, which we restate here for convenience.

Theorem 4.1. *Let $(f, \mathbb{P})$ be a probabilistic decision problem which has a valid proof protocol and $\mathcal{G}$ a* `nip` *game. Then $\boldsymbol{\sigma}$ is a valid IP protocol if and only if it is a (strict) approximate verifier-leading Stackelberg equilibrium of $\mathcal{G}$.*

Proof. Before beginning the proof notice that for any $\boldsymbol{\sigma}'$, the least upper bound of the resulting completeness and soundness errors are given by $\epsilon_{\mathrm{c}}(\boldsymbol{\sigma}') \coloneqq \sup_{x:f(x)=1} \ell(\boldsymbol{\sigma}', x)$ and $\epsilon_{\mathrm{s}}(\boldsymbol{\sigma}') \coloneqq \sup_{x:f(x)=0} \ell(\boldsymbol{\sigma}', x)$ respectively.

In an approximate equilibrium, the prover and verifier each select their strategies in a way that brings them within some e^p or e^v of the loss from their optimal strategy, respectively. We will show that $\boldsymbol{\sigma}$ is a valid IP protocol if and only if it is a strict verifier-leading $\boldsymbol{e}$-SE of $\mathcal{G}$. Concretely, we set: Concretely, we set:

$$\begin{aligned}
e^p &\coloneqq 1 - \min_{\boldsymbol{\sigma}^\star \in \boldsymbol{\Sigma}^\star} \mathcal{L}^v(\boldsymbol{\sigma}^\star),\\
e^v &\coloneqq 1 - \min_{\sigma^v \in \Sigma^v} \max_{\sigma^p \in e^p-\mathrm{LNE}(G(\sigma^v))} \mathcal{L}^v(\boldsymbol{\sigma}),
\end{aligned}$$

where (because G is a two-player game) e^p-LNE$(G(\sigma^v))$ contains the prover's approximate (local) best-responses to σ^v, denoted e^p-BR(σ^v). $\boldsymbol{\Sigma}^\star$ is the set of valid proof protocols for $(f, \mathbb{P})$, defined as:

$$\begin{aligned}
\boldsymbol{\Sigma}^\star \coloneqq \Big\{\sigma^\star \in \boldsymbol{\Sigma} : \; & \forall x \mathbb{E}[\boldsymbol{\sigma}_\star(x) \neq y \mid y = 1] \leqslant \epsilon_c^\star\\
& \wedge \forall \sigma^p \forall x \mathbb{E}[\langle \sigma^p, \sigma_\star^v \rangle (x) \neq y \mid y = 0] \leqslant \epsilon_s^\star\\
& \wedge \epsilon_c^\star + \epsilon_s^\star < 1 \Big\},
\end{aligned}$$

where we assume that $\boldsymbol{\Sigma}^\star \neq \varnothing$.

First, let us assume that $\boldsymbol{\sigma}$ is a valid IP protocol, with bounds $\epsilon_{\mathrm{c}}(\boldsymbol{\sigma})$ and $\epsilon_{\mathrm{s}}(\boldsymbol{\sigma})$. Let us show that $\boldsymbol{\sigma}$ is a Stackelberg equilibrium. The optimal strategy $\sigma_\dagger^p$ for the prover in response to σ^v has loss $\mathcal{L}^p(\sigma_\dagger^p, \sigma^v) = \epsilon_{\mathrm{c}}(\sigma_\dagger^p, \sigma^v) - \epsilon_{\mathrm{s}}(\sigma_\dagger^p, \sigma^v)$. Suppose for a contradiction that the prover can deviate from σ^p to reduce their loss by at least e^p:

$$\begin{aligned}
\epsilon_{\mathrm{c}}(\sigma_\dagger^p, \sigma^v) - \epsilon_{\mathrm{s}}(\sigma_\dagger^p, \sigma^v) + e^p &= \mathcal{L}^p(\sigma_\dagger^p, \sigma^v) + e^p\\
&\leqslant \mathcal{L}^p(\boldsymbol{\sigma})\\
&= \epsilon_{\mathrm{c}}(\boldsymbol{\sigma}) - \epsilon_{\mathrm{s}}(\boldsymbol{\sigma}).
\end{aligned}$$

Because $\boldsymbol{\sigma}$ is a valid IP protocol, then $\epsilon_{\mathrm{s}}(\sigma_\dagger^p, \sigma^v) \leqslant \epsilon_{\mathrm{s}}(\boldsymbol{\sigma})$. Combining these inequalities yields a contradiction:

$$\begin{aligned}
\epsilon_{\mathrm{c}}(\boldsymbol{\sigma}) + \epsilon_{\mathrm{s}}(\boldsymbol{\sigma}) &= \epsilon_{\mathrm{c}}(\boldsymbol{\sigma}) - \epsilon_{\mathrm{s}}(\boldsymbol{\sigma}) + 2\epsilon_{\mathrm{s}}(\boldsymbol{\sigma})\\
&\geqslant \epsilon_{\mathrm{c}}(\sigma_\dagger^p, \sigma^v) - \epsilon_{\mathrm{s}}(\sigma_\dagger^p, \sigma^v) + e^p + 2\epsilon_{\mathrm{s}}(\sigma_\dagger^p, \sigma^v)\\
&= \epsilon_{\mathrm{c}}(\sigma_\dagger^p, \sigma^v) + \epsilon_{\mathrm{s}}(\sigma_\dagger^p, \sigma^v) + e^p\\
&\geqslant \min_{\boldsymbol{\sigma}^\star \in \boldsymbol{\Sigma}^\star} \Big(\epsilon_{\mathrm{c}}(\boldsymbol{\sigma}^\star) + \epsilon_{\mathrm{s}}(\boldsymbol{\sigma}^\star)\Big) + e^p\\
&= 1.
\end{aligned}$$

Thus $\boldsymbol{\sigma}$ is a strict $\boldsymbol{e}$-SE$_v$ of $\mathcal{G}$.

Conversely, let us assume that $\boldsymbol{\sigma}$ is a strict $\boldsymbol{e}$-SE$_v$ of $\mathcal{G}$ with the aim of showing that $\boldsymbol{\sigma}$ a valid IP protocol. Define $\hat{\Sigma}^p$ as the set of approximate best responses to σ^v that maximise the soundness error, i.e. maximise the loss ℓ when $f(x) = 0$. Next, let us take some $\sigma^p_\dagger \in \operatorname{argmin}_{\hat{\sigma}^p \in \hat{\Sigma}^p} \epsilon_{\mathrm{c}}(\hat{\sigma}^p, \sigma^v)$. We claim that $\boldsymbol{\sigma}$ is a valid IP protocol with soundness and completeness errors $\epsilon_{\mathrm{s}}(\sigma^p_\dagger, \sigma^v)$ and $\epsilon_{\mathrm{c}}(\sigma^p_\dagger, \sigma^v)$ respectively. Clearly, by construction, we have that $\epsilon_{\mathrm{s}}(\tilde{\sigma}^p, \sigma^v) \leqslant \epsilon_{\mathrm{s}}(\sigma^p_\dagger, \sigma^v)$ for any prover strategy $\tilde{\sigma}^p$. Finally, for a contradiction, let us assume that:

$$\epsilon_{\mathrm{s}}(\sigma^p_\dagger, \sigma^v) + \epsilon_{\mathrm{c}}(\sigma^p_\dagger, \sigma^v) \geqslant 1 = e^v + \min_{\sigma^v \in \Sigma^v} \left(\max_{\sigma^p \in e^p\text{-BR}(\sigma^v)} \mathcal{L}^v(\boldsymbol{\sigma}) \right).$$

It follows immediately that $\boldsymbol{\sigma}$ is not a strict $\boldsymbol{e}$-SE$_v$ of $\mathcal{G}$ after all, as the verifier can deviate to lower their loss by at least e^v, and thus we must have $\epsilon_{\mathrm{s}}(\sigma^p_\dagger, \sigma^v) + \epsilon_{\mathrm{c}}(\sigma^p_\dagger, \sigma^v) < 1$, meaning that $\boldsymbol{\sigma}$ is valid. □

Here we establish a correspondence between the Stackelberg equilibria of a `nip` game G and its adversarial counterpart G^a.

Theorem 4.5. *Let $(f, \mathbb{P})$ be a probabilistic decision problem and $\mathcal{G}$ a* `nip` *game. Then (σ^p, σ^v) is an approximate verifier-leading SE ($\boldsymbol{e}$-SE$_v$) of $\mathcal{G}$ if and only if there is some σ^a such that $(\sigma^p, \sigma^v, \sigma^a)$ is an approximate verifier-prover SE ($\boldsymbol{e}$-SE$_{v,p}$) of G^a (the adversarial version of $\mathcal{G}$).*

Proof. First consider some $\boldsymbol{\sigma}_\star = (\sigma^p_\star, \sigma^v_\star, \sigma^a_\star) \in (e^p, e^v, 0)$-SE$_{v,p}(\mathcal{G}^a)$. By definition, the adversary best responds to $(\boldsymbol{\sigma}^p_\star, \boldsymbol{\sigma}^v_\star)$. Considering their loss:

$$\mathcal{L}^a(\boldsymbol{\sigma}) = -\ell((\sigma^p, \sigma^v), x_0) - \ell((\sigma^p, \sigma^v), x_1),$$

this is achieved by picking x_0 that maximises $\ell((\sigma^p, \sigma^v), x_0)$ and x_1 that maximises $\ell((\sigma^p, \sigma^v), x_1)$. Furthermore, the prover e^p-best responds to $\boldsymbol{\sigma}^v_\star$ given that (x_0, x_1) will be chosen in this way. This means that:

$$\mathcal{L}^p(\boldsymbol{\sigma}_\star) := \ell\left((\boldsymbol{\sigma}^p_\star, \boldsymbol{\sigma}^v_\star), \operatorname*{argmax}_{x_1 \in X_1} \ell((\boldsymbol{\sigma}^p_\star, \boldsymbol{\sigma}^v_\star), x_1) \right) - \ell\left((\boldsymbol{\sigma}^p_\star, \boldsymbol{\sigma}^v_\star), \operatorname*{argmax}_{x_0 \in X_0} \ell((\boldsymbol{\sigma}^p_\star, \boldsymbol{\sigma}^v_\star), x_0) \right)$$

is within e^p of the minimum. Now note that:

$$\ell\left((\boldsymbol{\sigma}^p, \boldsymbol{\sigma}^v), \operatorname*{argmax}_{x_i \in X_i} \ell((\boldsymbol{\sigma}^p, \boldsymbol{\sigma}^v), x_i) \right) = \mathcal{L}^{\mathrm{WC}}\big((\boldsymbol{\sigma}^p, \boldsymbol{\sigma}^v) \mid f(x) = i\big),$$

for $i \in \{0, 1\}$. Therefore, we have that:

$$\mathcal{L}^p(\boldsymbol{\sigma}^p_\star, \boldsymbol{\sigma}^v_\star) = \mathcal{L}^{\mathrm{WC}}\big((\boldsymbol{\sigma}^p_\star, \boldsymbol{\sigma}^v_\star) \mid f(x) = 1\big) - \mathcal{L}^{\mathrm{WC}}\big((\boldsymbol{\sigma}^p_\star, \boldsymbol{\sigma}^v_\star) \mid f(x) = 0\big)$$

is within e^p of the minimum. In other words, the prover e^p-best responds to $\boldsymbol{\sigma}^v_\star$ under the loss functions of $\mathcal{G}$. Using similar reasoning for the verifier, we see that $(\sigma^p_\star, \sigma^v_\star) \in \boldsymbol{e}$-SE$_v(\mathcal{G})$.

Conversely, let $(\sigma^p_\star, \sigma^v_\star)$ be a verifier-leading (e^p, e^v)-Stackelberg equilibrium. Let $\sigma^a_\star$ be the strategy for the adversary which selects (x_0, x_1) such that $\ell((\sigma^p, \sigma^v), x_0)$ and $\ell((\sigma^p, \sigma^v), x_1)$ are maximised. Then by repeating the above argument in reverse we see that $(\boldsymbol{\sigma}^p_\star, \boldsymbol{\sigma}^v_\star, \boldsymbol{\sigma}^a_\star)$ is a verifier-prover-leading $(e^p, e^v, 0)$-Stackelberg equilibrium, i.e. $\boldsymbol{\sigma}_\star = (\sigma^p_\star, \sigma^v_\star, \sigma^a_\star) \in (e^p, e^v, 0)$-SE$_{v,p}(\mathcal{G}^a)$. □

We now prove the correspondence result for `nmip` games. The proof is very similar to that of the correspondence for `nip` games, so we limit ourselves to noting the differences.

Theorem 5.1. *Let $(f, \mathbb{P})$ be a probabilistic decision problem which has a valid proof protocol and $\mathcal{G}$ a* `nmip` *game. Then $\boldsymbol{\sigma}$ is a valid MIP protocol if and only if it is a (strict) approximate verifier-leading* correlated *Stackelberg equilibrium of $\mathcal{G}$.*

Proof. We follow the proof of Theorem 4.1. This time we define the approximation bound $\boldsymbol{e}$ as follows.

$$\begin{aligned} e^{p_1} = e^{p_2} &:= 1 - \min_{\boldsymbol{\sigma}^\star \in \boldsymbol{\Sigma}^\star} \mathcal{L}^v(\boldsymbol{\sigma}^\star), \\ e^v &:= 1 - \min_{\sigma^v \in \Sigma^v} \max_{\sigma^{p_1} \in e^{p_1}\text{-BR}(\sigma^v),\ \sigma^{p_2} \in e^{p_2}\text{-BR}(\sigma^v)} \mathcal{L}^v(\boldsymbol{\sigma}). \end{aligned}$$

In the nmip model, the provers are assumed to be able to agree on a joint strategy $\boldsymbol{\sigma}^p = (\sigma^{p_1}, \sigma^{p_2})$ beforehand – including a commonly observed source of randomness – though their interactions with the verifier during the game are independent. The source of randomness then essentially forms a *correlation device* for the provers, allowing them to sample their actions using the agreed upon joint strategy $\boldsymbol{\sigma}^p$. If neither prover has an incentive to deviate from this agreement given their action (provided by this 'correlation device'), then we say that they are playing as in a *correlated equilibrium*.[2] Since p_1 and p_2 have the same loss, for a correlated Stackelberg equilibrium we can consider their combined strategy $\boldsymbol{\sigma}^p$ effectively as the strategy of a single player aiming to minimise the shared loss.

Now assume that $\boldsymbol{\sigma}$ is a valid MIP protocol, with bounds $\epsilon_{\mathrm{c}}(\boldsymbol{\sigma})$ and $\epsilon_{\mathrm{s}}(\boldsymbol{\sigma})$. The optimal strategy pair $(\sigma^{p_1}_{\dagger}, \sigma^{p_2}_{\dagger})$ for the provers in response to σ^v has (shared) loss:

$$\epsilon_{\mathrm{c}}\left(\sigma^{p_1}_{\dagger}, \sigma^{p_2}_{\dagger}, \sigma^v\right) - \epsilon_{\mathrm{s}}\left(\sigma^{p_1}_{\dagger}, \sigma^{p_2}_{\dagger}, \sigma^v\right).$$

Supposing for a contradiction that the provers can deviate from $(\sigma^{p_1}, \sigma^{p_2})$ to reduce their loss by at least e^{p_1} $(= e^{p_1})$, as before we obtain that:

$$\epsilon_{\mathrm{c}}\left(\sigma^{p_1}_{\dagger}, \sigma^{p_2}_{\dagger}, \sigma^v\right) - \epsilon_{\mathrm{s}}\left(\sigma^{p_1}_{\dagger}, \sigma^{p_2}_{\dagger}, \sigma^v\right) + e^{p_1} \leqslant \epsilon_{\mathrm{c}}(\boldsymbol{\sigma}) - \epsilon_{\mathrm{s}}(\boldsymbol{\sigma}).$$

and $\epsilon_{\mathrm{s}}\left(\sigma^{p_1}_{\dagger}, \sigma^{p_2}_{\dagger}, \sigma^v\right) \leqslant \epsilon_{\mathrm{s}}(\boldsymbol{\sigma})$. Combining these inequalities again yields the contradiction:

$$\epsilon_{\mathrm{c}}(\boldsymbol{\sigma}) + \epsilon_{\mathrm{s}}(\boldsymbol{\sigma}) \geqslant 1.$$

The converse direction proceeds as before, again noting that in correlated Stackelberg equilibrium the combined strategy $(\sigma^{p_1}, \sigma^{p_2})$ can reasoned about as the strategy of a single player. In particular, the fact that σ^{p_1} and σ^{p_2} can vary together when considering a best response to the verifier's strategy means that the verifier's strategy must be robust to both provers independently, which is what is required in the soundness condition. □

Finally, we prove the correspondence result for zk-nip games. As above, the proof is very similar to the previous cases, so we do not include the full proof.

Theorem 5.2. *Let $(f, \mathbb{P})$ be a probabilistic decision problem which has a valid ϵ_k-statistically zero-knowledge protocol and $\mathcal{G}$ a* zk-nip *game. Then $\boldsymbol{\sigma}$ is a valid ϵ_k-zero-knowledge IP protocol if it is a (strict) approximate verifier-prover-leading Stackelberg equilibrium ($\boldsymbol{e}$-$\mathrm{SE}_{v_1,p}$) of $\mathcal{G}$.*

Proof. Recall that in zk-nip we introduce two new verifiers, v_2 and v_3, where v_2 is a 'dishonest' verifier that attempts to extract knowledge from the prover p, in the sense of being able to generate a distribution over transcripts $\boldsymbol{m}$ that cannot be generated (by the simulator v_3) simply by learning y. If this is possible, then p must be providing more knowledge than y to v_2. In this new game, $\mathcal{L}^{v_3} := -\mathcal{L}^{v_2}$ are defined via:

$$\mathcal{L}^{v_2}(\boldsymbol{\sigma}) := -\max_x \frac{1}{2} \sum_{\boldsymbol{m}} \Big| \mathbb{P}\left(\langle \sigma^p, \sigma^{v_2} \rangle(x, y) = \boldsymbol{m}\right) - \mathbb{P}\left(\sigma^{v_3}(x, y) = \boldsymbol{m}\right) \Big|.$$

In addition, we add the term $\mathcal{L}^{v_3}$ to the prover's standard loss. There is a slight subtlety here, however, in that we would like the prover to minimise ϵ_{k} conditional on the protocol being valid, as opposed to reducing ϵ_{k} while potentially violating the validity of the protocol. (This would be true, for example, if the prover adopted a strategy that always outputs a single message m^p at every turn.) This leads to a *lexicographic* objective in which p first minimises $\mathcal{L}^{\mathrm{WC}}(\boldsymbol{\sigma} \mid f(x) = 1) - \mathcal{L}^{\mathrm{WC}}(\boldsymbol{\sigma} \mid f(x) = 0)$ and, from the set of such minimisers, minimises $\mathcal{L}^{v_3}(\boldsymbol{\sigma})$. While this may seem somewhat convoluted, such objectives can be easily optimised in practice via a simple modification of any standard (MA)RL algorithm (Skalse et al., 2022).

In the remainder of the proof, we assume that the ϵ_{k} bound is tight – i.e. it is the minimal ZK bound achievable for any valid IP protocol for $(f, \mathbb{P})$. We also note that if:

$$\max_x \frac{1}{2} \sum_{\boldsymbol{m}} \Big| \mathbb{P}\left(\langle \sigma^p, \sigma^{v_2} \rangle(x, y) = \boldsymbol{m}\right) - \mathbb{P}\left(\sigma^{v_3}(x, y) = \boldsymbol{m}\right) \Big| \leqslant \epsilon_{\mathrm{k}},$$

[2] We note that there is a slight discrepancy in our usage of this term from the classic definition, as we consider equilibria in which only the provers (not the verifier) correlate their strategies. In our formulation, the provers correlate their behavioural strategies and can randomise each time they send a message. However, because each prover is assumed to have perfect recall, then there is an equivalent joint mixed strategy in which the provers use their correlation device to each sample a deterministic strategy, which they then follow.

then:

$$\max_{x : f(x)=1} \frac{1}{2} \sum_{\boldsymbol{m}} \Big| \mathbb{P}\left(\langle p, v' \rangle(x) = \boldsymbol{m}\right) - \mathbb{P}\left(z(x) = \boldsymbol{m}\right) \Big| \leqslant \epsilon_{\text{k}},$$

where:

$$\begin{aligned} p(m^p \mid \boldsymbol{m}_{1:t}, x) &\coloneqq \sigma^p(m^p \mid \boldsymbol{m}_{1:t}, x, 1), \\ v'(m^v \mid \boldsymbol{m}_{1:t}, x) &\coloneqq \sigma^{v_2}(m^v \mid \boldsymbol{m}_{1:t}, x, 1), \\ z(m^v \mid \boldsymbol{m}_{1:t}, x) &\coloneqq \sigma^{v_3}(m^v \mid \boldsymbol{m}_{1:t}, x, 1). \end{aligned}$$

Conditioning on the value of $y = f(x)$ does not introduce any new information for either p, v', or z, because the prover is assumed to be able to compute $f(x)$, and the ZK error for the 'dishonest' verifier v' and simulator z is based only on the case where $f(x) = 1$, meaning that their behaviour when $f(x) = 0$ is irrelevant for the purposes of the ZK bound in Definition 2.2. As such, the bound introduced by our ZK variant is at least as strong as this definition, and so it suffices to consider the former.

Let us first suppose that $\boldsymbol{\sigma}_\star$ is an $\boldsymbol{e}$-$\text{SE}_{v_1,p}$ of $\mathcal{G}$, where $\boldsymbol{e}$ is defined as in the proof of Theorem 4.1 and the approximation errors for v_2 and v_3 are 0. First note that for any σ^{v_1} and σ^p then $\mathcal{G}(\sigma^{v_1}, \sigma^p)$ is zero-sum game between v_2 and v_3. Hence, under $\boldsymbol{\sigma}_\star$ – which is an NE in $\mathcal{G}(\sigma^{v_1}_\star, \sigma^p_\star)$ – we have a unique value $\epsilon_{\text{k}}(\sigma^{v_1}_\star, \sigma^p_\star) \coloneqq \mathcal{L}^{v_3}(\boldsymbol{\sigma}_\star) = -\mathcal{L}^{v_2}(\boldsymbol{\sigma}_\star)$.

In particular, because the prover p seeks to minimise $\mathcal{L}^{v_3}$ given that it is best-responding to $\sigma^{v_1}_\star$, we must have that $\epsilon_{\text{k}} \coloneqq \min_{(\sigma^{v_1}, \sigma^p) \in \boldsymbol{e}\text{-SE}_v(\mathcal{G}')} \epsilon_{\text{k}}(\sigma^{v_1}, \sigma^p)$, where $\mathcal{G}'$ is the `nip` game underlying the `zk-nip` game in question. In other words, we end up with a valid proof protocol for $\mathcal{G}'$ (as per the reasoning in the proof of Theorem 4.1) that minimises the ZK error.[3] Thus, we have that $\boldsymbol{\sigma}_\star$ is a valid ϵ_{k}-statistically zero-knowledge protocol for $(f, \mathbb{P})$. □

B.2. Worst-Case Loss

The next result establishes that, under certain conditions, minimising the empirical risk is sufficient to minimise the worst-case loss.

Theorem 4.4. *If Σ has the worst-case uniform convergence property and the worst-case robustness property then there is some $m^{WCUC} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$, if $|\mathcal{D}| \geqslant m^{\text{WC}}(\epsilon, \delta)$ then $\mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}) - \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{WC}}) \leqslant \epsilon$ with probability $1 - \delta$.*

Proof. Let us begin by defining $m^{\text{WC}}(\epsilon, \delta) \coloneqq \max\left[m^{\text{WCUC}}(\frac{\epsilon}{3}, \sqrt{\delta}), m^{\text{WCR}}(\frac{\epsilon}{3}, \sqrt{\delta})\right]$. Next, we expand $\mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}) - \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{WC}})$ into four expressions, which we denote by E_1 to E_4, respectively:

$$\begin{aligned} \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}) - \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{WC}}) &= \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}) - \mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}) \\ &\quad + \mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}) - \mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}) \\ &\quad + \mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}) - \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}) \\ &\quad + \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}) - \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{WC}}). \end{aligned}$$

Fix some $\epsilon, \delta \in (0,1)$ and let $m = m^{\text{WC}}(\epsilon, \delta)$. Consider some $\mathcal{D}$ drawn iid from $\mathbb{P}$ such that $|\mathcal{D}| \geqslant m$. Then by worst-case uniform convergence we have that, with probability $1 - \sqrt{\delta}$, $\mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}) - \mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}) \leqslant \frac{\epsilon}{3}$ for $\boldsymbol{\sigma} \in \{\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}, \boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}, \boldsymbol{\sigma}^{\text{WC}}\}$. Thus, we have directly that $E_1 \leqslant \frac{\epsilon}{3}$, and furthermore that:

$$\begin{aligned} E_4 &= \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}) - \mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}) \\ &\quad + \mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}) - \mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}^{\text{WC}}) \\ &\quad + \mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}^{\text{WC}}) - \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{WC}}) \\ &\leqslant \frac{\epsilon}{3} + 0 + 0. \end{aligned}$$

The second two terms are non-positive as $\boldsymbol{\sigma}^{\text{WC}}_{\mathcal{D}}$ minimises $\mathcal{L}^{\text{WC}}_{\mathcal{D}}$, and $\mathcal{L}^{\text{WC}}_{\mathcal{D}}(\boldsymbol{\sigma}) - \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}) \leqslant 0$ for any $\boldsymbol{\sigma}$. This second observation also implies that $E_3 \leqslant 0$. Finally, by worst-case robustness, we have that, with probability $1 - \sqrt{\delta}$, $E_2 \leqslant \frac{\epsilon}{3}$. Hence, $\mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{ER}}_{\mathcal{D}}) - \mathcal{L}^{\text{WC}}(\boldsymbol{\sigma}^{\text{WC}}) \leqslant \epsilon$ with probability $1 - \delta$, as required. □

[3] Here we assume a *strong* Stackelberg equilibrium in which v_1 is assumed to break any ties in favour of p, hence our minimisation over $(\sigma^{v_1}, \sigma^p) \in \boldsymbol{e}\text{-SE}_v(\mathcal{G}')$.