

Mechanizing Soundness of Off-Policy Evaluation

Jared Yeager  

Manning College of Information and Computer Sciences,
University of Massachusetts Amherst, MA, USA

J. Eliot B. Moss  

Manning College of Information and Computer Sciences,
University of Massachusetts Amherst, MA, USA

Michael Norrish  

School of Computing, Australian National University, Canberra, Australia

Philip S. Thomas  

Manning College of Information and Computer Sciences,
University of Massachusetts Amherst, MA, USA

Abstract

There are reinforcement learning scenarios – e.g., in medicine – where we are compelled to be as confident as possible that a policy change will result in an improvement before implementing it. In such scenarios, we can employ *off-policy evaluation* (OPE). The basic idea of OPE is to record histories of behaviors under the current policy, and then develop an estimate of the quality of a proposed new policy, seeing what the behavior would have been under the new policy. As we are evaluating the policy without actually using it, we have the “off-policy” of OPE. Applying a concentration inequality to the estimate, we derive a confidence interval for the expected quality of the new policy. If the confidence interval lies above that of the current policy, we can change policies with high confidence that we will do no harm.

We focus here on the mathematics of this method, by mechanizing the soundness of off-policy evaluation. A natural side effect of the mechanization is both to clarify all the result’s mathematical assumptions and preconditions, and to further develop HOL4’s library of verified statistical mathematics, including concentration inequalities. Of more significance, the OPE method relies on importance sampling, whose soundness we prove using a measure-theoretic approach. In fact, we generalize the standard result, showing it for contexts comprising both discrete and continuous probability distributions.

2012 ACM Subject Classification Theory of computation → Interactive proof systems; Theory of computation → Logic and verification; Theory of computation → Sequential decision making; Mathematics of computing → Hypothesis testing and confidence interval computation

Keywords and phrases Formal Methods, HOL4, Reinforcement Learning, Off-Policy Evaluation, Concentration Inequality, Hoeffding

Digital Object Identifier 10.4230/LIPIcs.ITP.2022.32

Supplementary Material *Software (Source Code)*: https://github.com/jdyeager/itp_ope
archived at `swh:1:dir:a554d5232ce611cfaa8df6b8c4fa0c164e51b2bb`

Funding This material is based upon work supported by the National Science Foundation under Grant No. CCF-2018372. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

1 Introduction

Reinforcement learning (RL) algorithms are machine learning algorithms that learn to make sequences of optimal or nearly optimal decisions through interactions with their environment. Their use has been proposed for a variety of high-risk high-reward applications including improving sepsis treatment [21], insulin dosing for type 1 diabetes treatment [40], epilepsy



© Jared Yeager, J. Eliot B. Moss, Michael Norrish, and Philip S. Thomas;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Interactive Theorem Proving (ITP 2022).

Editors: June Andronick and Leonardo de Moura; Article No. 32; pp. 32:1–32:20

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

treatment [12], and for various applications related to autonomous vehicles [10]. However, none of these *proposed* applications have come to fruition, partially due to concerns about safety. If the RL algorithm proposed a new policy (mechanism for making decisions) that is worse than a currently used policy it could be dangerous or costly.

To ensure that the policies produced by RL algorithms are safe, RL researchers have recently increased their focus on *off-policy evaluation* (OPE) methods – methods that use historical data collected from the use of a current policy to estimate and bound the performance of a newly proposed policy without requiring the (possibly dangerous) newly proposed policy to actually be used [32, 18, 39]. These methods are based on *importance sampling*. At a high level, the idea of importance sampling for RL is to consider previous runs of the current policy (or policies) and to take a weighted average of the observed historical performance, where the weight corresponds to the likelihood of the historical observations under the newly proposed policy divided by their likelihood under the current policy. Because importance sampling provides unbiased estimates of the performance of the newly proposed policy [38], confidence intervals for the mean of a random variable (e.g., based on Hoeffding’s inequality [14] or Student’s *t*-test [34]) can be applied to obtain confidence intervals for the performance of the new policy if it were to be used.

Particularly in cases like medicine, we desire more than just confidence in a statistical result in the sense of confidence intervals, but also confidence that the overall *process* of OPE is sound. A step in that direction is mechanizing a proof of OPE’s soundness, much like how other researchers provided mechanized proofs for the soundness of supervised learning generalization guarantees (as opposed to the reinforcement learning guarantees we provide) [3]. A complete program of establishing confidence would further prove correctness of a software implementation of OPE down to machine code – a step we leave to future work.

Concerning the mathematics of OPE, previous work has offered hand proofs for the cases of discrete probability distributions and continuous ones, but not for hybrid distributions – ones with both discrete and continuous components. Yet hybrid distributions arise quite naturally in practice because many systems have both discrete controls, such as on-off switches, and continuous ones, such as throttles and torque actuators. Another example is when a robot component experiences contact with an external object, which constrains actions and movement, versus no-contact when actions and movement are less constrained. Thus there can be a discrete probability that a throttle cannot be used in the current state, and in other states a continuous distribution of how much throttle to apply.

In addition to rounding out the mathematics of OPE by handling hybrid distributions, mechanizing a proof of OPE clarifies the preconditions of its soundness, ensuring that hand proofs did not overlook anything of significance.

1.1 Contributions

Our starting point is the existing HOL4 libraries, which include a number of theorems about probability and measure theory, but not much in the way of *statistics*.

The main results we present here are:

- OPE gives an unbiased estimate for any property (statistic) of the new policy representable as an *integrable* function on histories.
- Thus, more specifically, OPE gives an unbiased estimate of the standard RL performance metric: the expected discounted sum of rewards [35], also known as the expected *return*.
- Applying Hoeffding’s inequality, we obtain a confidence interval for the expected return for any desired confidence level.

Other similar results we do not explicitly present here:

- OPE gives an unbiased estimate for any property (statistic) of the new policy that can be represented as a positive measurable function on histories.
- From that result we further show that OPE gives an unbiased estimate for any point of the cumulative distribution function of any property (statistic) of the new policy representable as a measurable function on histories.

Other contributions include:

- Proof of Hoeffding’s inequality (and Hoeffding’s Lemma) in HOL4.
- Machinery in HOL4 to represent trajectories and histories, and measure spaces over them.
- Machinery in HOL4 to represent arbitrary n -way product measure spaces.

Our HOL4 sources are available from https://github.com/jdyeager/itp_ope.

2 Background

First, let us further clarify the starting point for our work, namely theories already present in the HOL4 libraries. Hurd [16, 17] developed the original formulation of measure theory and probability. Coble [8] added Lebesgue integration, Radon-Nikodym derivative, and random variable theories. Mhamdi [26] added formalization of “almost everywhere” and proved Markov’s inequality, and later refined integration to allow extended real results [27]. Tian [41] applied that extension to measure theory and probability. There are similar developments of much of this background theory in Isabelle/HOL (by Hölzl and Heller [15]), and in Coq (by Boldo et al. [5]).

While much of what OPE builds on is standard mathematics, such as measure theory, there are a few mathematical topics worthy of mention here. One is how Radon-Nikodym derivatives (already in HOL4) allow us to fold the discrete, continuous, and hybrid probability distribution cases into one general case. Another is the expression of OPE itself (not in HOL4). Among more standard results, we review briefly below Markov’s inequality (in HOL4), Hoeffding’s Lemma, and Hoeffding’s inequality (both new to HOL4 – Hoeffding’s inequality has been proven in Coq [3]). Before we review these mathematical topics, we first mention some standard definitions we use in our proofs:

Extended reals: The extended reals ($\bar{\mathbb{R}}$) consist of the real numbers augmented with $+\infty$ and $-\infty$, and are useful in capturing unbounded measures and integrals.

Sigma algebras: Let X be a set of points and Σ be a set of subsets of X , $\Sigma \subseteq 2^X$.¹ We say (X, Σ) is a *sigma algebra* if $X \in \Sigma$ and Σ is closed under complementation and countable unions. This appears in HOL4 as (note the indexing/extracting functions `space` and `subsets`):

```
subset-class X Σ  $\stackrel{\text{def}}{=} \forall s. s \in \Sigma \Rightarrow s \subseteq X,$ 
algebra a  $\stackrel{\text{def}}{=} \text{subset-class (space a) (subsets a) } \wedge \emptyset \in \text{subsets a} \wedge$ 
 $(\forall s. s \in \text{subsets a} \Rightarrow \text{space a} - s \in \text{subsets a}) \wedge$ 
 $\forall s t. s \in \text{subsets a} \wedge t \in \text{subsets a} \Rightarrow s \cup t \in \text{subsets a},$  and
 $\sigma\text{-algebra a} \stackrel{\text{def}}{=} \text{algebra a} \wedge \forall c. \text{countable c} \wedge c \subseteq \text{subsets a} \Rightarrow \bigcup c \in \text{subsets a}.$ 
```

There is a standard sigma algebra over the extended reals, called `Borel` in HOL4. This is useful when integrating measurable functions.

¹ Σ is a conventional name for the measurable sets, not to be confused with summation.

32:4 Mechanizing Soundness of Off-Policy Evaluation

Measurable functions: For (X, Σ) and (Y, T) sigma algebras, $f : X \rightarrow Y$ is a *measurable function* (from (X, Σ) to (Y, T)) if $\forall E \in T : f^{-1}(E) \in \Sigma$. In HOL4:

$$\begin{aligned} \vdash f \in \text{measurable } a \ b &\iff \\ \sigma\text{-algebra } a \ \wedge \ \sigma\text{-algebra } b \ \wedge \ f \in (\text{space } a \ \rightarrow \ \text{space } b) \ \wedge \\ \forall s. s \in \text{subsets } b \ \Rightarrow \ f^{-1} s \cap \text{space } a \in \text{subsets } a. \end{aligned}$$

In HOL4 these are called *Borel-measurable* functions if (Y, T) is the Borel sigma algebra.

Measure spaces: For $\mu : \Sigma \rightarrow \overline{\mathbb{R}}$, we say μ is a *measure* and (X, Σ, μ) a *measure space* if: (X, Σ) is a sigma algebra; $\mu(\emptyset) = 0$; and μ is positive and countably additive. Expressed in detail in HOL4 this is:

$$\begin{aligned} \text{measure-space } m &\stackrel{\text{def}}{=} \\ \sigma\text{-algebra } (\text{m-space } m, \text{measurable-sets } m) \ \wedge \ \text{positive } m \ \wedge \ \text{countably-additive } m, \text{ where} \\ \text{positive } m &\stackrel{\text{def}}{=} \text{measure } m \ \emptyset = 0 \ \wedge \ \forall s. s \in \text{measurable-sets } m \ \Rightarrow \ 0 \leq \text{measure } m \ s, \text{ and} \\ \text{countably-additive } m &\stackrel{\text{def}}{=} \\ \forall f. f \in (\mathbb{N} \rightarrow \text{measurable-sets } m) \ \wedge \ (\forall i \ j. i \neq j \ \Rightarrow \ \text{DISJOINT } (f \ i) \ (f \ j)) \ \wedge \\ \bigcup (\text{IMAGE } f \ \mathbb{N}) \in \text{measurable-sets } m \ \Rightarrow \\ \text{measure } m \ (\bigcup (\text{IMAGE } f \ \mathbb{N})) &= \text{suminf } (\text{measure } m \circ f). \end{aligned}$$

Sigma finite measure spaces: A measure space is *sigma finite* if it can be partitioned into countably many measurable sets of finite measure, or equivalently (as represented in HOL4), the space is the limit of measurable sets of finite measure:

$$\begin{aligned} \text{sigma-finite-measure-space } m &\stackrel{\text{def}}{=} \text{measure-space } m \ \wedge \ \text{sigma-finite } m, \text{ where} \\ \text{sigma-finite } m &\stackrel{\text{def}}{=} \\ \exists f. f \in (\mathbb{N} \rightarrow \text{measurable-sets } m) \ \wedge \ (\forall n. f \ n \subseteq f \ (\text{SUC } n)) \ \wedge \\ \bigcup (\text{IMAGE } f \ \mathbb{N}) &= \text{m-space } m \ \wedge \ \forall n. \text{measure } m \ (f \ n) < +\infty. \end{aligned}$$

Almost Everywhere: A property is said to be true *almost everywhere* (a.e.) if the set of points where the property is false has measure zero. In HOL4 this appears as a quantifier:

$$\begin{aligned} (\text{AE } x :: m. P \ x) &\stackrel{\text{def}}{=} \exists N. \text{null-set } m \ N \ \wedge \ \{x \mid x \in \text{m-space } m \ \wedge \ \neg P \ x\} \subseteq N, \text{ where} \\ \text{null-set } m \ s &\stackrel{\text{def}}{=} s \in \text{measurable-sets } m \ \wedge \ \text{measure } m \ s = 0. \end{aligned}$$

Density: We will find it useful to take a measure space (X, Σ, μ) and positive function f , and re-weight μ by f , thus incorporating f into μ . This is done by integrating over f :

$$\begin{aligned} \text{density } m \ f &\stackrel{\text{def}}{=} (\text{m-space } m, \text{measurable-sets } m, f * m), \text{ where} \\ f * m &\stackrel{\text{def}}{=} (\lambda s. \int^+ m (\lambda x. f \ x \cdot \mathbb{1} \ s \ x)). \end{aligned}$$

Probability: Probability, both mathematically and in HOL4, is a renaming of measure theory concepts:

$$\begin{aligned} \text{prob-space } p &\stackrel{\text{def}}{=} \text{measure-space } p \ \wedge \ \text{measure } p \ (\text{m-space } p) = 1, \\ \text{p-space } &\stackrel{\text{def}}{=} \text{m-space}, \\ \text{events } &\stackrel{\text{def}}{=} \text{measurable-sets}, \\ \text{prob } &\stackrel{\text{def}}{=} \text{measure}, \\ \text{real-random-variable } X \ p &\stackrel{\text{def}}{=} \\ \text{random-variable } X \ p \ \text{Borel} &\ \wedge \ \forall x. x \in \text{p-space } p \ \Rightarrow \ X \ x \neq -\infty \ \wedge \ X \ x \neq +\infty, \\ \text{random-variable } X \ p \ s &\stackrel{\text{def}}{=} X \in \text{measurable } (\text{p-space } p, \text{events } p) \ s, \text{ and} \\ \text{expectation } &\stackrel{\text{def}}{=} \int. \end{aligned}$$

For convenience of notation, we use $P_p[s]$ and $\mathbb{E}_p[f]$ in lieu of `prob` and `expectation` respectively, where the subscript p is the probability space and the set s is intersected with `p-space` p .

Markov's inequality: This, perhaps less basic, result gives an upper bound to how much of a positive function is above a given threshold:

$$\vdash \text{prob-space } p \wedge \text{integrable } p \ X \wedge 0 < c \Rightarrow \\ P_p[\{x \mid c \leq |X \ x|\}] \leq c^{-1} \cdot \mathbb{E}_p[(\lambda \ x. |X \ x|)].$$

We use this in proving Hoeffding's inequality.

Hoeffding's lemma: This result gives an upper bound to a moment generating function of a real-valued random variable with expectation 0, bounded almost everywhere between a and b :

$$\mathbb{E}[e^{cX}] \leq \exp\left(\frac{c^2(b-a)^2}{8}\right).$$

This is used in proving Hoeffding's inequality.

Hoeffding's inequality: This result bounds how much a sum of random variables deviates from its expectation. For n variables X_i , respectively bounded almost everywhere between a_i and b_i , for $t > 0$, where $S_n = \sum_{i=1}^n X_i$:

$$P(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

2.1 Off-Policy Evaluation

Here we present a standard mathematical formulation of an RL problem, known as a *partially observable Markov decision process* (POMDP) with a *state-free* policy [19, Section 7.1], using the motivating example of optimizing dosing for type 1 diabetes treatment [4, 40]. In this motivating application an RL algorithm is used to determine how much insulin should be injected into a person's blood prior to their eating each meal, in order to keep their blood glucose (blood sugar) near ideal levels.

Let \mathcal{S} , \mathcal{O} , and \mathcal{A} be sets of states, observations, and actions respectively. For simplicity when first presenting these methods, we assume that these sets are finite, resulting in subsequent distributions being discrete. However, our formalization is for the general setting where these can be arbitrary measurable spaces and distributions over these sets can be discrete, continuous, or hybrid. In the diabetes treatment example, each state $s \in \mathcal{S}$ is a complete characterization of the patient and the meal they will eat. The RL agent does not know or observe the state, but rather makes an observation $o \in \mathcal{O}$. For example, in prior work [4, 40], the observation corresponded to a vector of three real numbers indicating the patients' current blood glucose level (from a sample of the person's blood), target blood glucose level (as specified by a doctor), and the size of the meal they are about to eat (in grams of carbohydrates, roughly estimated by the patient). Each action $a \in \mathcal{A}$ is a positive real number corresponding to an amount of insulin that could be injected.

The environment that an RL agent interacts with has state $S_t \in \mathcal{S}$ at time $t \in \{-1, 0, 1, \dots, T-1\}$, where S_{-1} denotes an initial state. Here T , the number of time steps before the process terminates, is called the *horizon*. The initial state is sampled from an *initial state distribution* d_0 . The RL agent does not necessarily observe the state itself, and instead makes some observation $O_t \in \mathcal{O}$ where $O_t \sim \Omega(\cdot | S_{t-1})$ – where the notation $X \sim Z(\cdot | Y)$ indicates drawing X from the joint distribution $Z(x, y)$ with y fixed with value

Y , or equivalently, drawing X from the distribution $\lambda x.Z(x, Y)$. Next the RL agent selects an action $A_t \in \mathcal{A}$ according to a policy π , where $A_t \sim \pi(\cdot|O_t)$. This action causes the state of the environment to change to $S_t \sim P(\cdot|S_{t-1}, A_t)$, where P is called the *transition function*. This transition of the environment results in the agent’s receiving a real-valued reward, $R_t \sim d_R(\cdot|S_{t-1}, A_t, S_t)$.

A *trajectory* $\mathcal{T} = (S_{-1}, O_0, A_0, S_0, R_0, O_1, A_1, S_1, R_1, \dots, O_{T-1}, A_{T-1}, S_{T-1}, R_{T-1})$ is the sequence of states, observations, actions, and rewards observed during one run from time $t = 0$ to time $t = T - 1$ (one such run is called an *episode*). A *history* H is similar, but contains only the terms that are known to the RL algorithm, namely $H = (O_0, A_0, R_0, O_1, A_1, R_1, \dots, O_{T-1}, A_{T-1}, R_{T-1})$ (keeping the observations but dropping the states).

The *return* of an episode is the discounted sum of rewards, and can be viewed as a function g of the trajectory or history: $g(\mathcal{T}) = g(H) = \sum_{t=0}^{T-1} \gamma^t R_t$, where $\gamma \in [0, 1]$ is a parameter that discounts rewards that occur later in the trajectory. The performance of a policy π is the expected discounted sum of rewards that result from using the policy to make decisions: $J(\pi) = \mathbb{E}[g(H)|\pi]$.² We assume that we have access to n histories $(H_i)_{i=1}^n$ generated by past policies $(\beta_i)_{i=1}^n$, and also to a newly proposed policy π . OPE methods [32] use the historical data $D = (H_i)_{i=1}^n$ to estimate $J(\pi)$.

To further ground this notation and terminology, consider again our diabetes treatment example. Here, insulin injections are given prior to each meal, and each day is considered a separate episode. Assuming three meals per day, this corresponds to a horizon of $T = 3$. S_{-1} corresponds to a complete description of the patient and meal just prior to eating breakfast, O_0 is the observation about the state of the person and meal (blood glucose, target blood glucose, and meal size) just prior to breakfast, A_0 is the amount of insulin (a real number) to be injected prior to eating breakfast, S_0 is the state of the patient just prior to eating lunch, etc. The reward R_t is designed to penalize deviation from optimal blood glucose levels, with larger penalties for dangerously low blood levels, called *hypoglycemia*. The precise specification of these rewards must be carefully designed by experts to ensure that an agent that maximizes the expected discounted sum of rewards will produce the desired behavior [4, Page 11]. Current basic insulin dosage calculators determine the injection size using equations similar to:

$$\text{injection size} = \frac{\text{blood glucose} - \text{target blood glucose}}{CF} + \frac{\text{meal size}}{CR}, \quad (1)$$

where CF and CR are parameters chosen by a doctor [40], and which should be fine-tuned over time. When viewed as a policy, the injection size is the action A_t ; and the blood glucose, target blood glucose, and meal size correspond to the observation O_t . If (1) were used to select actions, this would correspond to a *deterministic policy*. To make this policy stochastic (as required for off-policy evaluation and RL in general), one might add a Gaussian random variable with a mean of 0 (we will call this “noise”) to the injection size, and $\pi(A_t|O_t)$ would then correspond to the probability of A_t being the action given observation O_t when using (1) with noise added.

However, the addition of noise to the action might result in unsafe injection sizes that deviate significantly (but with low probability) from the injection size intended by the doctor when they specified values for CF and CR . Though there are more sophisticated techniques

² The definition of $J(\pi)$ is *not* a conditional expected value. The conditioning notation indicates that the history H was generated by selecting actions using the policy π .

for creating a stochastic policy for this application that could be trusted [40], here we present one intuitive motivating alternative: the magnitude of the noise might be limited to some maximum value, η_{\max} . Let $E[A_t|O_t]$ denote the expected injection size given observation O_t – that is, the injection prescribed by (1). Notice that the inclusion of clipped noise to the action results in the policy being a hybrid distribution, with probability density on the open interval $(E[A_t|O_t] - \eta_{\max}, E[A_t|O_t] + \eta_{\max})$ and probability masses at $E[A_t|O_t] - \eta_{\max}$ and $E[A_t|O_t] + \eta_{\max}$.

Lastly, for the diabetes treatment example the data D corresponds to data collected using initial values for CR and CF chosen by a doctor, with one history per day. The goal of an RL agent would be to use this data to find a new policy (new values for CR and CF) that results in an increased expected return. If the rewards, R_t , are defined appropriately, this would correspond to values for CR and CF that better regulate the patient’s blood glucose levels. However, if the new policy is worse, it could have devastating consequences. For example, a single instance of severe hypoglycemia triples the five year mortality rate for a person with type 1 diabetes [25] and can have other severe consequences [40].

The importance sampling [20] estimator for $J(\pi)$ is then $IS = \frac{1}{n} \sum_{i=1}^n \left(\frac{\Pr(H_i|\pi)}{\Pr(H_i|\beta_i)} g(H_i) \right)$. Denoting actions and observations at time t in the i^{th} trajectory or history as A_t^i and O_t^i respectively, some simplification shows IS is equivalent to [32, 38]:

$$IS = \frac{1}{n} \sum_{i=1}^n (\rho_i(H_i) g(H_i)), \quad \text{where } \rho_i(H_i) = \prod_{t=0}^{T-1} \frac{\pi(A_t^i|O_t^i)}{\beta_i(A_t^i|O_t^i)}.$$

Peer reviewed (but not machine verified) proofs have shown that, when for all i and H $\Pr(H|\beta_i) = 0 \rightarrow \Pr(H|\pi) = 0$ (a condition assumed from here on, for histories and trajectories), the importance sampling estimator is unbiased [32, 37]. That is, $E[IS] = J(\pi)$.

We now provide an overview of the proof that the importance sampling estimator is unbiased. Let \mathcal{T}_π and \mathcal{H}_π denote the sets of all possible trajectories and histories when using policy π . We write $\mathcal{T} \sim \pi$ or $H \sim \pi$ to denote that a trajectory or history is generated using the policy π . Similarly, when a policy is used as a subscript on a probability it indicates that the relevant random variables come from using the specified policy. For example, $\Pr_\pi(H)$ is the probability of history H when policy π is used. As with g , the ρ_i can be viewed as functions of histories or trajectories: $\rho_i(\mathcal{T}) = \rho_i(H) = \prod_{t=0}^{T-1} \frac{\pi(A_t^i|O_t^i)}{\beta_i(A_t^i|O_t^i)}$. We now begin with $J(\pi) = E_{\mathcal{T} \sim \pi} [g(\mathcal{T})]$ and derive an equality to $E[IS]$, following the four steps that we later use in our HOL4 proof. For the first three steps, it suffices to consider unindexed β and ρ .

1. We begin with a change of measure – changing from trajectories generated by π to trajectories generated by β .

$$E_{\mathcal{T} \sim \pi} [g(\mathcal{T})] = \sum_{\mathcal{T} \in \mathcal{T}_\pi} \Pr_\pi(\mathcal{T}) g(\mathcal{T}) = \sum_{\mathcal{T} \in \mathcal{T}_\pi} \Pr_\beta(\mathcal{T}) \frac{\Pr_\pi(\mathcal{T})}{\Pr_\beta(\mathcal{T})} g(\mathcal{T}) = E_{\mathcal{T} \sim \beta} \left[\frac{\Pr_\pi(\mathcal{T})}{\Pr_\beta(\mathcal{T})} g(\mathcal{T}) \right],$$

2. Next we show that the ratio of the probability of \mathcal{T} under π divided by the probability under β does not depend on functions that are not known in practice (e.g., the transition and observation functions P and Ω). That is:

$$\begin{aligned} \frac{\Pr_\pi(\mathcal{T})}{\Pr_\beta(\mathcal{T})} &= \frac{d_0(S_{-1}) \prod_{t=0}^{T-1} \Omega(O_t|S_{t-1}) \pi(A_t|O_t) P(S_t|S_{t-1}, A_t) d_R(R_t|S_{t-1}, A_t, S_t)}{d_0(S_{-1}) \prod_{t=0}^{T-1} \Omega(O_t|S_{t-1}) \beta(A_t|O_t) P(S_t|S_{t-1}, A_t) d_R(R_t|S_{t-1}, A_t, S_t)} \\ &= \prod_{t=0}^{T-1} \frac{\pi(A_t|O_t)}{\beta(A_t|O_t)} = \rho(\mathcal{T}). \end{aligned}$$

Combining with the previous result, we therefore have that $J(\pi) = E_{\mathcal{T} \sim \beta} [\rho(\mathcal{T}) g(\mathcal{T})]$.

3. Next, as neither $\rho(\mathcal{T})$ nor $g(\mathcal{T})$ depend on the states, we apply the law of total probability to sum out the states and get: $\mathbb{E}_{\mathcal{T} \sim \beta} [\rho(\mathcal{T})g(\mathcal{T})] = \mathbb{E}_{H \sim \beta} [\rho(H)g(H)]$. Combining with the previous result, we therefore have that $J(\pi) = \mathbb{E}_{H \sim \beta} [\rho(H)g(H)]$.
4. Finally, we bring these results – and another application of the law of total probability – together to show (writing $H^n \sim \beta^n$ for the more precise $(H_i)_{i=1}^n \sim (\beta_i)_{i=1}^n$):

$$\begin{aligned} \mathbb{E}_{H^n \sim \beta^n} [\text{IS}] &= \mathbb{E}_{H^n \sim \beta^n} \left[\frac{1}{n} \sum_{i=1}^n (\rho_i(H_i)g(H_i)) \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{H^n \sim \beta^n} [\rho_i(H_i)g(H_i)] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{H_i \sim \beta_i} [\rho_i(H_i)g(H_i)] = \frac{1}{n} \sum_{i=1}^n J_{\mathcal{T}}(\pi) = J_{\mathcal{T}}(\pi). \end{aligned}$$

Our goal is to mechanize this proof, establishing the result for hybrid (mixed discrete and continuous) distributions.

2.2 Radon-Nikodym Derivatives

Radon-Nikodym derivatives generalize the concept of *probability density functions* (PDFs) and *probability mass functions* (PMFs) and allow us to transform one measure space into another via the density operator above. They capture the idea of a point-wise density ratio between two measure spaces, allowing a change of measure by integration.

Consider measure spaces (X, Σ, μ) and (X, Σ, ν) , and function f such that $\nu = f * \mu$, where f is called a *Radon-Nikodym derivative*. If we take (X, Σ, μ) to be the canonical uniform measure space on the real line and f to be the PDF of a normal distribution, then $\nu(s) = (f * \mu)(s)$ would be the probability that a random number drawn from a normal distribution is in s , and $\nu((-\infty, x))$ would be the normal *cumulative distribution function* (CDF) at x .

In this context discrete, continuous, and hybrid measures (e.g., probability measures, though this discussion applies to all measures) refer to characteristics of ν , such as whether it characterizes a distribution that has point masses (like a Bernoulli distribution), density (like a normal distribution), or both. In cases where we do not have access to ν (such as a π -weighted measure space for trajectories), we might use a different measure μ together with a correction term that “re-weights” the samples from μ . This re-weighting term is the Radon-Nikodym derivative.

We want our proofs to capture discrete, continuous, and hybrid ν . PDFs cannot capture discrete distributions, PMFs cannot capture continuous distributions, and neither can capture hybrid distributions. Taking a measure-theoretic approach using Radon-Nikodym derivatives allows us to capture all of these cases. Though this could also be achieved using PDFs and Dirac delta functions, Dirac delta functions are *generalized functions* [29] not functions, are not currently in the HOL4 libraries, and present challenges for completing a formal proof [33].

3 Proofs

We now present salient aspects of the mechanized proofs, covering Hoeffding’s inequality, product spaces and isomorphisms, trajectories and histories, importance sampling, OPE, and lastly confidence intervals on OPE.

3.1 Hoeffding’s Inequality

Much of what is needed to prove Hoeffding’s inequality is already in HOL4’s library. We do need a theorem about the expectation of the product of independent random variables:

► **Lemma 1** (Product of Independent Variables).

$$\begin{aligned} &\vdash \text{prob-space } p \wedge (\forall i. i < n \Rightarrow \text{real-random-variable } X_i p) \wedge \\ &\quad \text{independent } p X \text{ (count } n) \wedge (\forall i. i < n \Rightarrow \text{integrable } p X_i) \Rightarrow \\ &\quad \mathbb{E}_p[(\lambda x. \prod_{i < n} (X_i x))] = \prod_{i < n} \mathbb{E}_p[X_i]. \end{aligned}$$

In addition to the standard presentation of Hoeffding’s lemma, we prove a generalized version (not centered at 0):

► **Lemma 2** (Hoeffding’s lemma).

$$\begin{aligned} &\vdash \text{prob-space } p \wedge \text{real-random-variable } X p \wedge \mathbb{E}_p[X] = 0 \wedge a \leq 0 \wedge 0 \leq b \wedge \\ &\quad (\text{AE } x :: p. a \leq X x \wedge X x \leq b) \Rightarrow \\ &\quad \mathbb{E}_p[(\lambda x. \exp(c \cdot X x))] \leq \exp(c^2 \cdot (b - a)^2 / 8); \text{ and} \\ &\vdash \text{prob-space } p \wedge \text{real-random-variable } X p \wedge (\text{AE } x :: p. a \leq X x \wedge X x \leq b) \Rightarrow \\ &\quad \mathbb{E}_p[(\lambda x. \exp(c \cdot (X x - \mathbb{E}_p[X])))] \leq \exp(c^2 \cdot (b - a)^2 / 8). \end{aligned}$$

Hoeffding’s inequality then follows with some algebra:

► **Lemma 3** (Hoeffding’s inequality [14]). *Given p a probability space, n a positive natural number, and $X_{i < n}$, n independent random variables drawn from p that almost surely lie in their respective intervals $[a_i, b_i]$, let $S_n x = \sum_{i < n} (X_i x)$. We then have that the probability that the sum of the X_i minus their expectation exceeds some fixed t is bounded above by an expression in t and the a_i and b_i :*

$$\vdash P_p[\{ x \mid t \leq S_n x - \mathbb{E}_p[(\lambda x. S_n x)] \}] \leq \exp(-2 \cdot t^2 / \sum_{i < n} (b_i - a_i)^2)$$

We also prove a corollary, for positive δ , that is more helpful in constructing confidence intervals:

$$\begin{aligned} &\vdash 1 - \delta \leq \\ &\quad P_p[\{ x \mid \\ &\quad \quad n^{-1} \cdot S_n x - \text{sqrt}(\ln \delta^{-1} \cdot \sum_{i < n} (b_i - a_i)^2 / (2 \cdot n^2)) \leq \\ &\quad \quad \mathbb{E}_p[(\lambda y. n^{-1} \cdot S_n y)] \}]. \end{aligned}$$

The proof is straightforward, directly paralleling one found in Wikipedia:³

$$\begin{aligned} P(S_n - \mathbb{E}[S_n] \geq t) & \qquad \qquad \qquad \text{where } S_n = \sum_{i=1}^n X_i \\ &= P(\exp(s(S_n - \mathbb{E}[S_n])) \geq \exp(st)) && \text{for } s > 0, \text{ by monotonicity of exp} \\ &\leq \exp(-st) \mathbb{E}[\exp(s(S_n - \mathbb{E}[S_n]))] && \text{by Markov’s inequality} \\ &= \exp(-st) \prod_{i=1}^n \mathbb{E}[\exp(s(X_i - \mathbb{E}[X_i]))] && \text{by algebraic manipulation} \\ &\leq \exp(-st) \prod_{i=1}^n \exp\left(\frac{s^2(b_i - a_i)^2}{8}\right) && \text{by Hoeffding’s lemma} \end{aligned}$$

³ See https://en.wikipedia.org/wiki/Hoeffding%27s_inequality, as of February, 2022. The Wikipedia formulation assumes the probability space for the X_i without naming it. In HOL4 it is an explicit argument p , and x is a point in that space.

32:10 Mechanizing Soundness of Off-Policy Evaluation

$$\begin{aligned}
&= \exp\left(-st + \frac{1}{8}s^2 \sum_{i=1}^n (b_i - a_i)^2\right) && \text{by algebraic manipulation} \\
&= \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) && \text{by setting } s = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}
\end{aligned}$$

The corollary can be obtained by setting $t = \sqrt{\ln \delta^{-1} \sum_{i=1}^n (b_i - a_i)^2 / 2}$ and dividing everything by n .

3.2 Product Spaces

Our proofs use measure spaces that are products of arbitrary (finite) numbers of measure spaces. A technical limitation of HOL4's type system is that it does not support arbitrary n -tuples of types. To model such things requires using an indexing function of type $\text{num} \rightarrow \alpha$ where α subsumes the types of all the dimensions. The existing Martingale theory package provides theorems about pairwise products of measure spaces. We recapitulate that approach in the inductive step of building n -fold products.

► **Definition 4** (pi-m-space). *The space of a product measure space is the product of the spaces of the component measure spaces. We use a subsidiary definition `updated-at`, which defines n -fold product sets. The term $f(n \mapsto e)$ denotes a function that is everywhere the same as f except at n , where it maps to e .*

$$\begin{aligned}
\text{pi-m-space } 0 \text{ } mn &\stackrel{\text{def}}{=} \{ (\lambda i. \text{ARB}) \} \\
\text{pi-m-space (SUC } n) \text{ } mn &\stackrel{\text{def}}{=} \text{updated-at } n \text{ (pi-m-space } n \text{ } mn) \text{ (m-space (mn } n)), \text{ where} \\
\text{updated-at } n \text{ } fs \text{ } s &\stackrel{\text{def}}{=} \{ f(n \mapsto e) \mid f \in fs \wedge e \in s \}.
\end{aligned}$$

► **Definition 5** (pi-measurable-sets). *The measurable sets of a product measure space are those of the smallest sigma algebra formed by the rectangular sets of the component measure spaces. We use a subsidiary definition `pi-prod-sets`, which defines the n -fold rectangular sets. `pi-sig-alg` is a packaging of the space and measurable sets.*

$$\begin{aligned}
\text{pi-measurable-sets } 0 \text{ } mn &\stackrel{\text{def}}{=} \text{POW } \{ (\lambda i. \text{ARB}) \} \\
\text{pi-measurable-sets (SUC } n) \text{ } mn &\stackrel{\text{def}}{=} \\
&\text{subsets} \\
&\text{(sigma (pi-m-space (SUC } n) \text{ } mn)} \\
&\quad \{ \{ f(n \mapsto e) \mid f \in fs \wedge e \in s \} \mid \\
&\quad fs \in \text{pi-measurable-sets } n \text{ } mn \wedge s \in \text{measurable-sets (mn } n) \}); \\
\text{pi-prod-sets } n \text{ } fsts \text{ } sts &\stackrel{\text{def}}{=} \{ \text{updated-at } n \text{ } fs \text{ } s \mid fs \in fsts \wedge s \in sts \}; \text{ and} \\
\text{pi-sig-alg } n \text{ } mn &\stackrel{\text{def}}{=} (\text{pi-m-space } n \text{ } mn, \text{pi-measurable-sets } n \text{ } mn).
\end{aligned}$$

► **Definition 6** (pi-measure). *The measure of a set in a product measure space is obtained by integrating over an indicator function of that set. pi-measure-space is a packaging of all components of the measure space.*

$$\begin{aligned} \text{pi-measure } 0 \text{ } mn &\stackrel{\text{def}}{=} (\lambda fs. \mathbb{1} fs (\lambda i. \text{ARB})) \\ \text{pi-measure (SUC } n) \text{ } mn &\stackrel{\text{def}}{=} \\ &(\lambda fs. \int^+ (mn \ n) (\lambda e. \int^+ (\text{pi-measure-space } n \ mn) (\lambda f. \mathbb{1} fs f (n \mapsto e))))); \text{ and} \\ \text{pi-measure-space } n \ mn &\stackrel{\text{def}}{=} \\ &(\text{pi-m-space } n \ mn, \text{pi-measurable-sets } n \ mn, \text{pi-measure } n \ mn). \end{aligned}$$

3.3 Isomorphisms

We show that these product spaces are indeed measure spaces. The proof uses recursion to show an $(n + 1)$ -way product space of sigma finite measure spaces is a sigma finite measure space because it is “equivalent to” a 2-way product of an n -way product and the remaining space. Firstly, we address the 2-way product being a sigma finite measure space:

► **Theorem 7** (2-Way Product is a Sigma Finite Measure Space). *A product of 2 sigma finite measure spaces is a sigma finite measure space.*

$$\vdash \text{sigma-finite-measure-space } m_1 \wedge \text{sigma-finite-measure-space } m_2 \Rightarrow \text{sigma-finite-measure-space } (m_1 \times m_2).$$

We also need results formalizing what it means for a measure space to be “equivalent to” another one. Specifically, two measure spaces are isomorphic if there is a “measure preserving” function between them, along the lines of Halmos [13, p. 164]:

► **Definition 8** (Isomorphic Measure Spaces).

$$m_0 \cong m_1 \stackrel{\text{def}}{=} \exists f. f \in \text{measure-preserving } m_0 \ m_1$$

where

$$\begin{aligned} \text{measure-preserving } m_0 \ m_1 &\stackrel{\text{def}}{=} \\ &\{ f \mid \\ &f \in \text{measurability-preserving } (\text{sig-alg } m_0) (\text{sig-alg } m_1) \wedge \\ &\forall s. s \in \text{measurable-sets } m_0 \Rightarrow \text{measure } m_0 \ s = \text{measure } m_1 (\text{IMAGE } f \ s) \}; \text{ and} \\ \text{measurability-preserving } a \ b &\stackrel{\text{def}}{=} \\ &\{ f \mid \\ &\sigma\text{-algebra } a \wedge \sigma\text{-algebra } b \wedge \text{BIJ } f \ (\text{space } a) \ (\text{space } b) \wedge \\ &(\forall s. s \in \text{subsets } a \Rightarrow \text{IMAGE } f \ s \in \text{subsets } b) \wedge \\ &\forall s. s \in \text{subsets } b \Rightarrow f^{-1} \ s \cap \text{space } a \in \text{subsets } a \}. \end{aligned}$$

These definitions allow us to formalize the utility of isomorphisms:

► **Theorem 9** (Isomorphisms Preserve Sigma Finiteness). *Isomorphism to a sigma finite measure space implies being a sigma finite measure space*

$$\vdash \text{sigma-finite-measure-space } m_0 \wedge m_0 \cong m_1 \Rightarrow \text{sigma-finite-measure-space } m_1.$$

Finally, we combine Theorems 7 and 9 with the existence of a measure-preserving function:

32:12 Mechanizing Soundness of Off-Policy Evaluation

► **Theorem 10** (*n*-Way Product is a Sigma Finite Measure Space). *A product of n sigma finite measure spaces is a sigma finite measure space.*

$$\begin{aligned} &\vdash (\forall i. i < n \Rightarrow \text{sigma-finite-measure-space } (mn\ i)) \Rightarrow \\ &\quad \text{sigma-finite-measure-space } (\text{pi-measure-space } n\ mn); \text{ because} \\ &\vdash \text{sigma-finite-measure-space } (\text{pi-measure-space } n\ mn) \wedge \\ &\quad \text{sigma-finite-measure-space } (mn\ n) \Rightarrow \\ &\quad (\lambda (f, e). f(n \mapsto e)) \in \\ &\quad \text{measure-preserving } (\text{pi-measure-space } n\ mn \times mn\ n) \\ &\quad (\text{pi-measure-space } (\text{SUC } n)\ mn). \end{aligned}$$

This approach appears again when working with trajectory and history measure spaces.

3.4 Variable Name Conventions and Standard Assumptions

In our model of OPE, we model sets of states, observations, actions, and rewards. At the very least, we need sigma algebras of these, and as the later distributions will need base measure spaces, we need full measure spaces for each of these. Below, these are respectively denoted ms , mo , ma , mr . The underlying type variables we use for these are σ , ω , α , ρ respectively (for less general results, ρ , the reward type, is usually of type `extreal`). For convenience, these standard measure parameters are omitted from argument/parameter lists henceforth, as are assumptions that these spaces are either measure spaces or sigma finite measure spaces.

Individual states are typically s and s' (s' being the later state); observations are o , actions are a , r is reward, and h is history or trajectory (they seldom appear simultaneously). The length of trajectories or histories is n or T , while the number of rows in the database is N .

3.5 Trajectories and Histories as Types

We now show our development of trajectories and histories (described in Section 2.1) as types in HOL4. The respective types are defined:

► **Definition 11** (Histories and Trajectories as HOL Types).

$$\begin{aligned} (\alpha, \rho, \omega) \text{ hist} &= \text{hnil} \mid \text{hcons } ((\alpha, \rho, \omega) \text{ hist}) \omega \alpha \rho \\ (\alpha, \rho, \sigma, \omega) \text{ traj} &= \text{init } \sigma \mid \text{tcons } ((\alpha, \rho, \sigma, \omega) \text{ traj}) \omega \alpha \sigma \rho \end{aligned}$$

A history is isomorphic to a list of observation, action, and reward triples; these are drawn from the type variables ω , α , and ρ respectively. The order of arguments in our definition is deliberate: we consider the non-recursive arguments to the “cons” functions to be at the end (rather than head) of the history. A trajectory adds a required initial state of type σ , and then adds a state between each action and reward of a history. The function `t-hist` extracts the state-less triples from a trajectory, giving a history. We write $|t|$ to record the number of steps in a trajectory (i.e., $|\text{init } s| = 0$), and `t-st` t is the function that returns the final state of trajectory t .

The definition and machinery of history measure spaces directly parallels, and was developed after, that of trajectories. It suffices to speak only of trajectories here.

Originally, we considered measure spaces of all trajectories of all lengths. This was abandoned for two reasons. First, it was unnecessary. For any database of histories (which contains a finite number of histories), there is a longest trajectory among those from which

the histories are derived. All other histories/trajectories can be extended to the same length via “do nothing” actions. This is why a maximum number of steps is taken in the literature as described earlier. Moreover, only fixed-length PDF-weighted trajectory measure spaces can form probability spaces, which is necessary later on.

Even though we abandoned the use of different trajectory lengths, we pursued it long enough that it inspired trajectory spaces and measurable sets to be defined in a not-directly-recursive manner (unlike the n -way product space). The development parallels that of product spaces directly, as opposed to only within an inductive step.

► **Definition 12** (traj-m-space-n). *A product measure space’s space is a cross product of the component measure spaces’ spaces. We mirror that with traj-cross, which takes a trajectory of sets and returns a set of trajectories, each element of which falls component-wise within the input. In order to get a trajectory of spaces of the desired length, we make a helper function named traj-n-gen.*

$$\begin{aligned} \text{traj-m-space-n } n \text{ } ms \text{ } mo \text{ } ma \text{ } mr &\stackrel{\text{def}}{=} \\ &\text{traj-cross (traj-n-gen } n \text{ (m-space } ms) \text{ (m-space } mo) \text{ (m-space } ma) \text{ (m-space } mr));} \\ \text{traj-cross (init } ss) \text{ (init } s) &\stackrel{\text{def}}{=} s \in ss \\ \text{traj-cross (init } ss) \text{ (tcons } h \text{ } w \text{ } a \text{ } s \text{ } r) &\stackrel{\text{def}}{=} F \\ \text{traj-cross (tcons } hs \text{ } ws \text{ } as \text{ } ss \text{ } rs) \text{ (init } s) &\stackrel{\text{def}}{=} F \\ \text{traj-cross (tcons } hs \text{ } ws \text{ } as \text{ } ss \text{ } rs) \text{ (tcons } h \text{ } w \text{ } a \text{ } s \text{ } r) &\stackrel{\text{def}}{=} \\ &w \in ws \wedge a \in as \wedge s \in ss \wedge r \in rs \wedge \text{traj-cross } hs \text{ } h; \text{ and} \\ \text{traj-n-gen } 0 \text{ } sg \text{ } og \text{ } ag \text{ } rg &\stackrel{\text{def}}{=} \text{init } sg \\ \text{traj-n-gen (SUC } n) \text{ } sg \text{ } og \text{ } ag \text{ } rg &\stackrel{\text{def}}{=} \text{tcons (traj-n-gen } n \text{ } sg \text{ } og \text{ } ag \text{ } rg) \text{ } og \text{ } ag \text{ } sg \text{ } rg. \end{aligned}$$

► **Definition 13** (traj-measurable-sets-n). *A product measure space’s measurable sets form a sigma algebra containing all cross products (the “rectangular” sets) of the component measure spaces’ measurable sets. We mirror that with traj-rect-sets-n, which takes a trajectory of sets of measurable sets, to a set of trajectories of measurable sets, to a set of rectangular sets of trajectories. traj-sig-alg-n is a pairing of the space and measurable sets.*

$$\begin{aligned} \text{traj-measurable-sets-n } n &\stackrel{\text{def}}{=} \\ &\text{subsets (sigma (traj-m-space-n } n) \text{ (traj-rect-sets-n } n));} \\ \text{traj-rect-sets-n } n &\stackrel{\text{def}}{=} \\ &\text{IMAGE traj-cross} \\ &\text{(traj-cross} \\ &\text{(traj-n-gen } n \text{ (measurable-sets } ms) \text{ (measurable-sets } mo) \text{ (measurable-sets } ma) \text{ (measurable-sets } mr));} \text{ and} \\ \text{traj-sig-alg-n } n &\stackrel{\text{def}}{=} \text{(traj-m-space-n } n, \text{ traj-measurable-sets-n } n). \end{aligned}$$

► **Definition 14** (traj-measure-n). *The measure of a set is obtained by integrating over an indicator function of that set. traj-measure-space-n is a pairing of all components of the measure space:*

32:14 Mechanizing Soundness of Off-Policy Evaluation

$$\begin{aligned}
\text{traj-measure-n } 0 &\stackrel{\text{def}}{=} (\lambda hs. \int^+ ms (\lambda s. \mathbb{1} hs (\text{init } s))) \\
\text{traj-measure-n (SUC } n) &\stackrel{\text{def}}{=} \\
&(\lambda hs. \\
&\quad \int^+ (\text{traj-measure-space-n } n) \\
&\quad (\lambda h. \\
&\quad \quad \int^+ mo \\
&\quad \quad (\lambda o. \\
&\quad \quad \quad \int^+ ma \\
&\quad \quad \quad (\lambda a. \int^+ ms (\lambda s. \int^+ mr (\lambda r. \mathbb{1} hs (\text{tcons } h o a s r)))))); \text{ and} \\
\text{traj-measure-space-n } n &\stackrel{\text{def}}{=} (\text{traj-m-space-n } n, \text{traj-measurable-sets-n } n, \text{traj-measure-n } n).
\end{aligned}$$

Many steps in later proofs require these to be measure spaces, which we prove inductively by showing they form sigma finite measure spaces:

► **Theorem 15** (History and Trajectory Measure Spaces).

- ⊢ sigma-finite-measure-space (traj-measure-space-n n); *because*
- ⊢ traj-measure-space-n 0 \cong ms; *and*
- ⊢ traj-measure-space-n (SUC n) \cong traj-measure-space-n n \times mo \times ma \times ms \times mr.
- ⊢ sigma-finite-measure-space (hist-space n); *because*
- ⊢ hist-space (SUC n) \cong hist-space n \times mo \times ma \times mr.

3.6 Preconditions and Useful Functions

Before developing the main OPE results, the mechanization identifies our standard preconditions on d_0 (initial state distribution), Ω (the observation distribution), β (current policy), π (new policy), P (transition distribution), and d_R (reward distribution), grouping them together in a predicate `valid-dist-gen-funs`. The conditions are about as minimal as one could hope: the distribution functions need to be positive, non-infinite, measurable, and integrate to 1 (form a probability space when used for making a density space). We omit the full definition, but assume it in all contexts mentioning our various distribution functions hereafter.

► **Definition 16** (Trajectory PDF, Importance Ratio, and Return). *Using those preconditions, we define a number of functions, specifically the return, importance ratio, and PDF:*

$$\begin{aligned}
\text{traj-pdf } d_0 P \Omega d_R \beta (\text{init } s) &\stackrel{\text{def}}{=} d_0 s \\
\text{traj-pdf } d_0 P \Omega d_R \beta (\text{tcons } h w a s r) &\stackrel{\text{def}}{=} \\
&\text{traj-pdf } d_0 P \Omega d_R \beta h \cdot \Omega (\text{t-st } h) w \cdot \beta w a \cdot P (\text{t-st } h) a s \cdot d_R (\text{t-st } h) a s r; \\
\text{importance-ratio } \pi \beta (\text{init } s) &\stackrel{\text{def}}{=} 1 \\
\text{importance-ratio } \pi \beta (\text{tcons } h w a s r) &\stackrel{\text{def}}{=} \text{importance-ratio } \pi \beta h \cdot \pi w a \cdot (\beta w a)^{-1}; \text{ and} \\
\text{traj-return } \gamma (\text{init } s) &\stackrel{\text{def}}{=} 0 \\
\text{traj-return } \gamma (\text{tcons } h w a s r) &\stackrel{\text{def}}{=} \text{traj-return } \gamma h + \gamma^{|h|} \cdot r.
\end{aligned}$$

There are similarly defined analogous functions for histories: `hist-pdf`, `h-importance-ratio`, and `hist-return`.

► **Definition 17** (History PDF). *hist-pdf is particularly interesting in that it requires a helper function that serves as a PDF of history and final state, where the non-final states are integrated away recursively.*

$$\begin{aligned} \text{hist-pdf } ms \ d_0 \ P \ \Omega \ d_R \ \beta \ h &\stackrel{\text{def}}{=} \int^+ ms \ (\text{hist-1st-pdf } ms \ d_0 \ P \ \Omega \ d_R \ \beta \ h); \text{ and} \\ \text{hist-1st-pdf } ms \ d_0 \ P \ \Omega \ d_R \ \beta \ hnil \ s' &\stackrel{\text{def}}{=} d_0 \ s' \\ \text{hist-1st-pdf } ms \ d_0 \ P \ \Omega \ d_R \ \beta \ (hcons \ h \ o \ a \ r) \ s' &\stackrel{\text{def}}{=} \\ \beta \ o \ a \cdot \int^+ ms \ (\lambda s. \text{hist-1st-pdf } ms \ d_0 \ P \ \Omega \ d_R \ \beta \ h \ s \cdot \Omega \ s \ o \cdot P \ s \ a \ s' \cdot d_R \ s \ a \ s' \ r). \end{aligned}$$

Many steps in later proofs require these (and their history analogs) to be measurable:

$$\begin{aligned} &\vdash \text{traj-pdf } d_0 \ P \ \Omega \ d_R \ \beta \in \text{Borel-measurable } (\text{traj-sig-alg-n } n); \\ &\vdash (\forall w \ a. w \in \text{m-space } mo \wedge a \in \text{m-space } ma \wedge \beta \ w \ a = 0 \Rightarrow \pi \ w \ a = 0) \Rightarrow \\ &\quad \text{importance-ratio } \pi \ \beta \in \text{Borel-measurable } (\text{traj-sig-alg-n } n); \text{ and} \\ &\vdash \text{sig-alg } mr = \text{Borel} \Rightarrow \text{traj-return } \gamma \in \text{Borel-measurable } (\text{traj-sig-alg-n } n). \end{aligned}$$

3.7 Importance Sampling

We now start to address OPE as described in Section 2.1. The first step is to shift the density measure space from the π PDF to the β PDF, by showing that the PDF ratio is a Radon-Nikodym derivative from one space to the other:

$$\begin{aligned} &\vdash (\forall w \ a. w \in \text{m-space } mo \wedge a \in \text{m-space } ma \wedge \beta \ w \ a = 0 \Rightarrow \pi \ w \ a = 0) \wedge \\ &\quad f \in \text{Borel-measurable } (\text{traj-sig-alg-n } n) \Rightarrow \\ &\quad \int (\text{density } (\text{traj-measure-space-n } n) (\text{traj-pdf } d_0 \ P \ \Omega \ d_R \ \pi)) f = \\ &\quad \int (\text{density } (\text{traj-measure-space-n } n) (\text{traj-pdf } d_0 \ P \ \Omega \ d_R \ \beta)) \\ &\quad (\lambda h. \text{traj-pdf } d_0 \ P \ \Omega \ d_R \ \pi \ h \cdot (\text{traj-pdf } d_0 \ P \ \Omega \ d_R \ \beta \ h)^{-1} \cdot f \ h). \end{aligned}$$

We next replace the ratio of PDFs with the importance ratio, *via* reasonably straightforward algebra:

$$\begin{aligned} &\vdash h \in \text{traj-m-space-n } n \wedge \text{traj-pdf } d_0 \ P \ \Omega \ d_R \ \beta \ h \neq 0 \Rightarrow \\ &\quad \text{traj-pdf } d_0 \ P \ \Omega \ d_R \ \pi \ h \cdot (\text{traj-pdf } d_0 \ P \ \Omega \ d_R \ \beta \ h)^{-1} = \\ &\quad \text{importance-ratio } \pi \ \beta \ h. \end{aligned}$$

► **Theorem 18** (Importance Ratio). *We combine our last two results to allow us to calculate an expectation (integral, here) in π -weighted trajectory space given trajectories in β -weighted trajectory space:*

$$\begin{aligned} &\vdash (\forall w \ a. w \in \text{m-space } mo \wedge a \in \text{m-space } ma \wedge \beta \ w \ a = 0 \Rightarrow \pi \ w \ a = 0) \wedge \\ &\quad f \in \text{Borel-measurable } (\text{traj-sig-alg-n } n) \Rightarrow \\ &\quad \int (\text{density } (\text{traj-measure-space-n } n) (\text{traj-pdf } d_0 \ P \ \Omega \ d_R \ \pi)) f = \\ &\quad \int (\text{density } (\text{traj-measure-space-n } n) (\text{traj-pdf } d_0 \ P \ \Omega \ d_R \ \beta)) \\ &\quad (\lambda h. \text{importance-ratio } \pi \ \beta \ h \cdot f \ h). \end{aligned}$$

3.8 Off-Policy Evaluation

The next step is to shift from the β trajectory PDF to the β history PDF, justifying the use of the empirical estimator. To that end, after some involved changes in order of integration, we are able to show:

32:16 Mechanizing Soundness of Off-Policy Evaluation

► **Theorem 19** (Trajectories to Histories). *Integrals in the trajectory measure space (assumed sigma finite), can be recast as integrals in the history space:*

$$\begin{aligned} & \vdash (\forall x. x \in \text{hist-m-space-n } n \Rightarrow 0 \leq f x) \wedge \\ & f \in \text{Borel-measurable (hist-sig-alg-n } n) \Rightarrow \\ & \int^+ (\text{density (traj-measure-space-n } n) (\text{traj-pdf } d_0 P \Omega d_R \beta)) (f \circ \text{t-hist}) = \\ & \int^+ (\text{density (hist-space } n) (\text{hist-pdf } ms d_0 P \Omega d_R \beta)) f. \end{aligned}$$

A useful corollary – used in later proofs – follows:

► **Corollary 20.** *Since the PDF-weighted trajectory space is a probability space, setting the positive function in the earlier result to be the constant 1, we derive that the PDF-weighted history space is also a probability space:*

$$\begin{aligned} & \vdash \text{prob-space (density (traj-measure-space-n } n) (\text{traj-pdf } d_0 P \Omega d_R \beta)); \text{ and thus} \\ & \vdash \text{prob-space (density (hist-space } n) (\text{hist-pdf } ms d_0 P \Omega d_R \beta)). \end{aligned}$$

Finally, we combine Theorems 18 and 19:

► **Theorem 21.** *It is possible to estimate the expected return in π -weighted trajectory space by appeal to the empirical return of β -weighted history space:*

$$\begin{aligned} & \vdash (\forall w a. w \in \text{m-space } mo \wedge a \in \text{m-space } ma \wedge \beta w a = 0 \Rightarrow \pi w a = 0) \wedge \\ & f \in \text{Borel-measurable (hist-sig-alg-n } n) \Rightarrow \\ & \int (\text{density (traj-measure-space-n } n) (\text{traj-pdf } d_0 P \Omega d_R \pi)) (f \circ \text{t-hist}) = \\ & \int (\text{density (hist-space } n) (\text{hist-pdf } ms d_0 P \Omega d_R \beta)) \\ & (\lambda h. \text{h-importance-ratio } \pi \beta h \cdot f h). \end{aligned}$$

3.9 Database Estimate

To complete OPE, we go from an empirical estimate based on a single history, to one based on an average over histories. We first prove a lemma to help go from product space back to the individual spaces:

► **Theorem 22.** *An integral (expectation) over a sum (thus average) in product space can be reduced to a sum (or average) of integrals over individual spaces:*

$$\begin{aligned} & \vdash (\forall i. i < n \Rightarrow \text{prob-space } mn_i) \wedge (\forall i. i < n \Rightarrow \text{integrable } mn_i f_i) \Rightarrow \\ & \int (\text{pi-measure-space}_n mn) (\lambda x. \sum_{i < n} (f_i x_i)) = \sum_{i < n} (\int mn_i f_i). \end{aligned}$$

We use that to generalize Theorem 21:

► **Theorem 23.** *It is possible to estimate the expected return in π -weighted trajectory space by appeal to an average of empirical returns in β_i -weighted history spaces. Let*

$$\begin{aligned} p & = \text{pi-measure-space}_n (\lambda i. \text{density (hist-space } T) (\text{hist-pdf } ms d_0 P Z d_R \beta_i)) \\ \tau & = \text{density (traj-measure-space-n } T) (\text{traj-pdf } d_0 P Z d_R phi) \end{aligned}$$

and assume the various implicit measure spaces (ma , mo , etc.) are sigma finite measure spaces; that n is strictly positive, and that π is zero whenever any of the β_i are, i.e.:

$$\forall i o a. i < n \wedge o \in \text{m-space } mo \wedge a \in \text{m-space } ma \wedge \beta_i o a = 0 \Rightarrow \pi o a = 0$$

then:

$$\begin{aligned} & \vdash \text{integrable (density (hist-space } nT) (\text{hist-pdf } ms d_0 P \Omega d_R \pi)) f \Rightarrow \\ & \int p (\lambda D. n^{-1} \cdot \sum_{i < n} (\text{h-importance-ratio } \pi \beta_i D_i \cdot f D_i)) = \int \tau (f \circ \text{t-hist}). \end{aligned}$$

3.10 Confidence Interval

For brevity, we define the database-based estimate of the return as

$$\text{data-return}_n \pi \beta \gamma D \stackrel{\text{def}}{=} n^{-1} \cdot \sum_{i < n} (\text{h-importance-ratio } \pi \beta_i D_i \cdot \text{hist-return } \gamma D_i)$$

At this point we have shown that OPE gives an unbiased estimate of the expected return of the new policy. However, it is even more useful to have a confidence interval around this expectation. We develop a confidence interval using Hoeffding's inequality (other concentration inequalities could be used here). Using Hoeffding's inequality requires a further precondition: bounding almost everywhere the importance ratio times the return. The final result is:

► **Theorem 24** (Unbiasedness of Off-Policy Evaluation on Return). *Let*

$$\begin{aligned} S &= \sum_{i < n} (UB_i - LB_i)^2 / (2 \cdot n^2) \\ p &= \text{pi-measure-space}_n (\lambda i. \text{density (hist-space } T) (\text{hist-pdf } ms \ d_0 \ P \ \Omega \ d_R \ \beta_i)) \\ \tau &= \text{density (traj-measure-space-n } T) (\text{traj-pdf } d_0 \ P \ \Omega \ d_R \ \pi) \end{aligned}$$

and assume the various implicit measure spaces (ma , mo , etc.) are sigma finite measure spaces; that n and δ are both greater than 0, and that π is zero whenever any of the β_i are, i.e.:

$$\forall i \ o \ a. i < n \wedge o \in \text{m-space } mo \wedge a \in \text{m-space } ma \wedge \beta_i \ o \ a = 0 \Rightarrow \pi \ o \ a = 0$$

then:

$$\begin{aligned} \vdash \text{sig-alg } mr = \text{Borel} \wedge \\ (\forall h. h \in \text{hist-m-space-n } T \Rightarrow Gmin \leq \text{hist-return } \gamma h \wedge \text{hist-return } \gamma h \leq Gmax) \wedge \\ (\forall i. i < n \Rightarrow \\ \text{AE} h :: \text{density (hist-space } T) (\text{hist-pdf } ms \ d_0 \ P \ \Omega \ d_R \ \beta_i). \\ LB_i \leq \text{h-importance-ratio } \pi \beta_i h \cdot \text{hist-return } \gamma h \wedge \\ \text{h-importance-ratio } \pi \beta_i h \cdot \text{hist-return } \gamma h \leq UB_i) \Rightarrow \\ 1 - \delta \leq P_p[\{ D \mid \text{data-return}_n \pi \beta \gamma D - \text{sqrt}(\ln \delta^{-1} \cdot S) \leq \mathbb{E}_\tau[\text{traj-return } \gamma] \}] \end{aligned}$$

In other words, the one-sided confidence interval whose lower bound is $\sqrt{\ln \delta^{-1} \cdot S}$ less than our estimator captures the expected return of the new policy with probability at least $1 - \delta$. As such, if we have reason to believe – such as an upper-bounded confidence interval – that the expected return under the current policy is below the estimator confidence interval with probability at least $1 - \epsilon$, then the probability that the new policy's expected return is better is at least $(1 - \delta)(1 - \epsilon)$

3.11 The HOL4 Mechanization

The proof of Hoeffding's inequality took on the order of two person-months of effort and OPE took about three-person months. In terms of lines of HOL4 code, our running library of important and useful general-purpose results is currently around 4800 lines, Hoeffding's inequality and variants take another 500 lines, the development of product measure spaces and isomorphisms requires about 1000 lines, and the rest of the OPE development is approximately 3500 lines long, for a total of about 10,000 lines. As may often be the case when adding new theories, it is helpful to refine and enhance the theorem prover's formula simplification capabilities to reduce the number of small, tedious, steps – though many remain.

4 Conclusion and Future Work

While the expected return, $J(\pi)$, is the most common performance metric in RL, for some high-risk applications parameters of the return distribution other than the expected value can better characterize the risk of applying the policy π . For example, RL researchers have studied using coherent risk measures [31, 7, 36, 30, 28] like the *conditional value at risk* (CVaR) [1] of the return distribution. The extension of our results to this setting will require an additional step: though we have shown off-policy pointwise convergence to the CDF of returns under π , we must show *uniform* convergence following the hand-checked proof of Chandak et al. [6]. Uniform convergence of off-policy estimates to the CDF of returns under π would allow for estimates and confidence intervals for all parameters of the return distribution [6], including CVaR, variance, and quantiles.

Our main result uses Hoeffding’s inequality to obtain a confidence interval. As concentration inequalities go, Hoeffding’s inequality is easy to prove, but also notoriously loose. Thus another fruitful direction for future work is mechanizing proofs of other, tighter, concentration inequalities, such as Maurer and Pontil’s empirical Bernstein bound [24], Anderson’s inequality [2] using Massart’s tight constants [23] for the Dvoretzky-Kiefer-Wolfowitz inequality [9], or an extremely tight confidence interval conjectured independently by multiple researchers [11, 22].

There are places where our development might be tightened by relaxing preconditions. For example, the constraint that the distribution generation functions form probability spaces isn’t necessary for showing that any of the PDF functions are measurable, but we still use the precondition `valid-dist-gen-funs` for brevity of proof statement in lemmas that are ultimately used when `valid-dist-gen-funs` in its entirety is necessary. There are also cases where a definition takes in extraneous information. For example, `pi-m-space` and `traj-m-space-n` take entire measure spaces, but use only the spaces of those measure spaces.

Finally, our histories are isomorphic to lists, and trajectories are in turn lists with extra information between adjacent elements, making them what the HOL4 library would call finite *paths*. It would be appealing to develop more generalized theories about measure spaces based on these library notions, allowing our trajectory and history results to fall out as special cases. Moreover, the product measure space results could be expressed in terms of list measure spaces.

More generally, given our development of measure spaces over histories and trajectories, a variety of other results in RL might be mechanized.

It is reassuring that our proof of the soundness of OPE brought no big surprises. In particular, it was encouraging that the ultimate pre-conditions were just the measure spaces being sigma finite and the distribution generating functions being non-negative, finite, measurable, and integrating to 1. At the same time, it is satisfying to record OPE’s generalization to hybrid distributions and to lay groundwork for several future directions.

References

- 1 Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.
- 2 B. D. Anderson and J. B. Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.
- 3 Alexander Bagnall and Gordon Stewart. Certifying the true error: Machine learning in Coq with verified generalization guarantees. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2662–2669, July 2019. doi:10.1609/aaai.v33i01.33012662.

- 4 M. Bastani. Model-free intelligent diabetes management using machine learning. Master's thesis, Department of Computing Science, University of Alberta, 2014.
- 5 Sylvie Boldo, François Clément, Florian Faissole, Vincent Martin, and Micaela Mayero. A Coq formalization of Lebesgue integration of nonnegative functions. *Journal of Automated Reasoning*, November 2021. doi:10.1007/s10817-021-09612-0.
- 6 Yash Chandak, Scott Niekum, Bruno Castro da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S. Thomas. Universal off-policy evaluation. In *Advances in Neural Information Processing Systems*, 2021.
- 7 Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems*, pages 3509–3517, 2014.
- 8 Aaron R. Coble. *Anonymity, information, and machine-assisted proof*. PhD thesis, Computer Lab., University of Cambridge, July 2010. doi:10.48456/tr-785.
- 9 A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27(3):642–669, 1956. doi:10.1214/aoms/1177728174.
- 10 Ahmad El Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.
- 11 Norbert Gaffke. Nonparametric one-sided testing for the mean and related extremum problems. *Mathematical Methods of Statistics*, 13(4):369–391, 2004.
- 12 Arthur Guez, Robert D. Vincent, Massimo Avoli, and Joelle Pineau. Adaptive treatment of epilepsy via batch-mode reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pages 1671–1678, 2008.
- 13 Paul R. Halmos. *Measure Theory*. Springer-Verlag, New York, NY, 1950.
- 14 W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
- 15 Johannes Hölzl and Armin Heller. Three chapters of measure theory in Isabelle/HOL. In Marko van Eekelen, Herman Geuvers, Julien Schmaltz, and Freek Wiedijk, editors, *Proceedings of Interactive Theorem Proving*, pages 135–151. Springer, 2011.
- 16 Joe Hurd. Formal verification of probabilistic algorithms. Technical Report UCAM-CL-TR-566, University of Cambridge, Computer Laboratory, May 2003. doi:10.48456/tr-566.
- 17 Joe Hurd. Verification of the Miller–Rabin probabilistic primality test. *The Journal of Logic and Algebraic Programming*, 56(1):3–21, 2003. doi:10.1016/S1567-8326(02)00065-6.
- 18 Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.
- 19 Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.
- 20 T. Kloek and H. K. van Dijk. Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. *Econometrica*, 46(1):1–19, 1978. doi:10.2307/1913641.
- 21 Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.
- 22 Erik Learned-Miller and Philip S. Thomas. A new confidence interval for the mean of a bounded random variable, 2020. arXiv:1905.06208.
- 23 P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability*, 18(3):1269–1283, July 1990. doi:10.1214/aop/1176990746.
- 24 A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the Twenty-Second Annual Conference on Learning Theory*, pages 115–124, 2009.
- 25 Rozalina G McCoy, Holly K Van Houten, Jeanette Y Ziegenfuss, Nilay D Shah, Robert A Wermers, and Steven A Smith. Increased mortality of patients with diabetes reporting severe hypoglycemia. *Diabetes Care*, 35(9):1897–1901, 2012.

- 26 Tarek Mhamdi, Osman Hasan, and Sofiène Tahar. On the formalization of the Lebesgue integration theory in HOL. In Matt Kaufmann and Lawrence C. Paulson, editors, *Proceedings of Interactive Theorem Proving*, volume 6172, pages 387–402. Springer, July 2010. doi: 10.1007/978-3-642-14052-5_27.
- 27 Tarek Mhamdi, Osman Hasan, and Sofiène Tahar. Formalization of entropy measures in HOL. In Marko van Eekelen, Herman Geuvers, Julien Schmaltz, and Freek Wiedijk, editors, *Proceedings of Interactive Theorem Proving*, volume 6898, pages 233–248. Springer, August 2011. doi:10.1007/978-3-642-22863-6_18.
- 28 Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 799–806. Citeseer, 2010.
- 29 A. Papoulis. *The Fourier Integral and its Implications*. McGraw-Hill Book Company, Inc., New York, NY, 1962.
- 30 Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. *arXiv preprint*, 2017. arXiv:1703.02702.
- 31 LA Prashanth and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. In *Advances in Neural Information Processing Systems*, pages 252–260, 2013.
- 32 D. Precup, R. S. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pages 759–766, 2000.
- 33 A. Shamilov, A. F. Yuzer, E. Agaoglu, and Y. Mert. A method of obtaining distributions of transformed random variables by using the Heaviside and the Dirac generalized functions. *Journal of Statistical Research*, 40(1):23–34, 2006.
- 34 Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.
- 35 R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.
- 36 Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the CVaR via sampling. In *AAAI*, pages 2993–2999, 2015.
- 37 P. S. Thomas, G. Theodorou, and M. Ghavamzadeh. High confidence off-policy evaluation. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence*, 2015.
- 38 Philip S. Thomas. *Safe Reinforcement Learning*. PhD thesis, University of Massachusetts Libraries, 2015.
- 39 Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.
- 40 Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.
- 41 Chun Tian. The new HOL-Probability based on $[0, +\text{Inf}]$ -measure theory, September 2019. URL: <https://github.com/HOL-Theorem-Prover/HOL/commit/c4db120ba392910141cc83672cc9bd6435e17c9a>.