Formalization of a Stochastic Approximation Theorem

Koundinya Vajjha
University of Pittsburgh, United States

Barry Trager
IBM Research, United States

Avraham Shinnar
IBM Research, United States

Vasily Pestun
IBM Research, United States
IHÉS, France

Abstract

Stochastic approximation algorithms are iterative procedures which are used to approximate a target value in an environment where the target is unknown and direct observations are corrupted by noise. These algorithms are useful, for instance, for root-finding and function minimization when the target function or model is not directly known. Originally introduced in a 1951 paper by Robbins and Monro, the field of Stochastic approximation has grown enormously and has come to influence application domains from adaptive signal processing to artificial intelligence. As an example, the Stochastic Gradient Descent algorithm which is ubiquitous in various subdomains of Machine Learning is based on stochastic approximation theory. In this paper, we give a formal proof (in the Coq proof assistant) of a general convergence theorem due to Aryeh Dvoretzky [16], which implies the convergence of important classical methods such as the Robbins-Monro and the Kiefer-Wolfowitz algorithms. In the process, we build a comprehensive Coq library of measure-theoretic probability theory and stochastic processes.

2012 ACM Subject Classification Mathematics of computing → Stochastic processes; Mathematics of computing → Nonlinear equations; Computing methodologies → Optimization algorithms

Keywords and phrases Formal Verification, Stochastic Approximation, Stochastic Processes, Probability Theory, Optimization Algorithms

Supplementary Material `https://github.com/IBM/FormalML/releases/tag/ITP2022`

Funding *Koundinya Vajjha*: Vajjha acknowledges support from the Alfred P. Sloan Foundation under grant number G-2018-10067 and the Andrew W. Mellon Foundation.

1 Introduction

A very frequent problem occuring in various contexts of statistical learning is the following: Let Y be a real-valued random variable that depends on a parameter x. We may say that $P(Y|x)$ is the probability distribution of Y conditioned or dependent on a parameter x. Next, suppose we are given a function $f(y, x)$ and we want to find x that solves the equation

$$\mathbb{E}_P f(Y, x) = 0 \tag{1}$$

Moreover, assume that $P(Y|x)$ is not available to us explicitly, but only in an implicit or sampling form, that is, we are provided a sampling oracle which takes a parameter x and returns a sample of Y drawn from x-dependent probability distribution $P(Y|x)$.

▶ **Example 1** (Kolmogorov's Strong Law of Large Numbers). Let $f(y, x) = y - x$, and let Y be independent of x. To solve equation (1) in this situation means to solve $\mathbb{E}[Y] = x$, that is to find the expected value x of a random variable Y given an oracle from which we can sample Y, in other words to construct a statistical estimator of $\mathbb{E}[Y]$ given a series of samples $y_0, y_1, \ldots$. The following iterative algorithm does the job

$$x_{n+1} := x_n + a_n(y_n - x_n) \tag{2}$$

for $n = 0, 1, 2, \ldots$ where $x_0 := 0$ and $a_n = \frac{1}{n+1}$. Indeed, the iterations (2) are equivalent to the standard sample mean estimator $x_n = \frac{1}{n}\sum_{k=0}^{n-1} y_k$. Notice that the iterations (2) have the form $x_{n+1} = x_n + a_n f(y_n, x_n)$. The theorem that the estimator x_n converges almost surely (with probability one) to the true expectation value is famously known as the "Kolmogorov's Strong Law of Large Numbers" (SLLN).

▶ **Example 2** (Banach's fixed point and optimal control). Now consider the opposite example, where Y that depends on x in a deterministic way, say $Y = g(x)$ where $g : \mathbb{R} \to \mathbb{R}$ is a certain function (and event space is a single point). In this case, when we pass x to the oracle, the oracle deterministically returns to us the value of the function g evaluated at x. In this case, solving the equation (1) for the function $f(y, x) = y - x = g(x) - x$ means solving the equation

$$g(x) = x \tag{3}$$

If g is a γ-contraction map[1] the standard proof of the Banach fixed point theorem tells us that the iterations

$$x_{n+1} := x_n + a_n(g(x_n) - x_n) \tag{4}$$

for a suitable choice of a_n, for example $a_n = \frac{1}{n+1}$, form a Cauchy sequence $x_1, x_2, \ldots$ that converges to the fixed point of the map $g : \mathbb{R} \to \mathbb{R}$. A variation of this process is applied to solve Bellman's equation for optimal control of Markov Decision Process (MDP) where γ-contraction map g comes from Bellman's optimality operator for MDPs with discount parameter $0 < \gamma < 1$.

▶ **Example 3** (Stochastic gradient descent). Now, as a variation of (1), suppose that we want to find x that minimizes the expectation value $\mathbb{E}[L(Y, x)]$ of a certain loss function $L(Y, x)$ in a context where Y is sampled by an oracle from an x-dependent probability distribution. Assuming that $\mathbb{E}[L(Y, x)]$ is a locally convex analytic function, finding a local minimum is equivalent to solving the stationary point equation

$$\nabla_x \mathbb{E}[L(Y, x)] = 0 \tag{5}$$

Since ∇_x is a linear operator, the above equation is equivalent to $\mathbb{E}[\nabla_x L(Y, x)] = 0$ and therefore is again an example of the equation (1) with $f(y, x) := -\nabla_x L(y, x)$. The iterative sequence

$$x_{n+1} := x_n - a_n \nabla_x L(y_n, x_n) \tag{6}$$

is known as stochastic gradient descent. This algorithm is a typical component of most of machine learning algorithms that search for a parameter x that minimizes the expected value

[1] this means that in some norm $||\bullet||$ it holds that $||g(x) - g(x')|| < \gamma||x - x'||$ for all x, x' with $\gamma < 1$

of the loss function $L(Y, x)$ given samples of Y.[2] Under suitable conditions on $f(y, x) = -\nabla_x L(y, x)$ and the parameters a_n (for example, $a_n = \frac{1}{n+1}$ would satisfy all required assumptions) one can prove convergence of (6) to the critical point of the loss function $L(Y, x)$.

These three examples demonstrate the ubiquity of the problem (1), and many more applications could be mentioned in a longer report.

In all these cases the solution of the problem (1) has the form

$$x_{n+1} := x_n + a_n f(y_n, x_n) \tag{7}$$

and is called *a stochastic approximation algorithm*.

A large body of literature explored different versions of assumptions on the domain of variables, on the function $f(y, x)$ and on the step-sizes (learning rates) a_n, under which the convergence of x_n could be proven in various senses: as convergence in L^2, as convergence in probability, as convergence with probability 1.[3]

Robbins and Monro introduced in [24] the field of Stochastic Approximation by proving the L^2 convergence of the process (7) for $f(y, x) = b - y$ to the value x that solves equation $\mathbb{E}[y](x) = b$. For this theorem, Robbins and Monro assumed

$$a_n \to 0, \qquad \sum_{n=1}^{\infty} a_n = \infty, \qquad \sum_{n=1}^{\infty} a_n^2 < \infty, \tag{8}$$

that Y is bounded with probability 1, and that the function $M(x) := E[Y](x)$ is (i) non-decreasing, (ii) the solution x_* of $M(x) = b$ exists, and (iii) the derivative at the solution is positive $M'(x)|_{x=x_*} > 0$.

Kiefer and Wolfowitz [20] took a similar approach but considered the problem of estimating the parameter x where the function $M(x)$ has a maximum, and proved convergence in probability.

Wolfowitz [36] weakened the assumption of Robbins-Monro about boundness of Y: instead his version assumes only that the variance of Y is bounded uniformly over x, and $M(x)$ is bounded, and with those assumptions Wolfowitz proves convergence in probability.

Blum [10] weakened further the assumptions of Robbins-Monro and Wolfowitz and proved substantially stronger result, namely that the iterative sequence (7) converges with probability 1. Blum requires the variance of Y be uniformly bounded over x, but he allows the expectation value $M(x) = \mathbb{E}[Y](x)$ to be bounded by a linear function of x

$$|M(x)| \leq A|x| + B \quad A, B \geq 0 \tag{9}$$

instead of a constant. Blum's proof is based on a version of Kolmogorov's inequality adopted in a suitable way by Loève [22] where instead of series of independent random variables, a certain dependence was allowed but constrained by a conditional expectation value. This extension of Kolomogorov's inequality to the conditional situation was related to earlier works of Borel, Lévy and Doob about convergence with probability 1 of certain stochastic processes.

---

[2] In the context of supervised learning, y will stand for (y_{in}, y_{out}) tuples sampled from training data, and x stands for the model parameters, e.g. neural network weights. If $N_x : y_{in} \mapsto y_{out}$ is a neural network, then with a quadratic supervised loss one normally takes $L(y, x) := (y_{out} - N_x(y_{in}))^2$ where $y = (y_{in}, y_{out})$

[3] The notion of "convergence with probability 1" is the same as the notion of "convergence almost surely", but different from the notion of "convergence in probability", which is much weaker.

Finally, the most general form of stochastic approximation was formulated by Dvoretzky [16]. In the original Robbins-Monro stochastic approximation (7), the next value x_{n+1} is determined through the previous value x_n and the sample y_n. Dvoretzky allowed more general estimator algorithms in which x_{n+1} is determined through a certain function that can take as arguments complete history of all previous values $x_1, \dots, x_{n+1}$ and the current sample y_n.

Concretely, let $T_n : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function of n-variables. Consider the stochastic process

$$x_{n+1} := T_n(x_1, \dots, x_n) + W_n \tag{10}$$

where $W_1, W_2, \dots$ are random variables, with W_n dependent on the previous history $X_1, \dots, X_n$ such that

$$\mathbb{E}(W_n | x_1, \dots, x_n) = 0 \tag{11}$$

Another way to formulate Dvoretzky's setup is to say that for any sequence of random variables $X_1, X_2, \dots$ where we have conditional probability distribution of X_{n+1} dependent on the complete history $x_1, \dots, x_n$, and then *define*

$$\begin{aligned} T(x_1, \dots, x_n) &\overset{def}{=} \mathbb{E}[X_{n+1} | x_1, \dots, x_n] \\ W_n &\overset{def}{=} X_{n+1} - \mathbb{E}[X_{n+1} | x_1, \dots, x_n] \end{aligned} \tag{12}$$

in this way we automatically get the relation (10) with noise terms W_n that satisfy (11).

For example, in the Robbins-Monro version we take (7) with $f(y, x) = b - y$ which gives $X_{n+1} := X_n + a_n(b - Y_n)$ and hence the Robbins-Monro process is a specialization of Dvoretzky's process with

$$\begin{aligned} &T(x_1, \dots, x_n) := x_n + a_n(b - M(x_n)) \\ &W_n := a_n(M(x_n) - Y_n) \end{aligned} \tag{13}$$

whereas before in the context of Robbins-Monro we had $M(x_n) := \mathbb{E}(Y_n | x_n)$.

To prove his result, Dvoretzky assumed that:

1. there exists a point x_* such that

$$|T_n(x_1, \dots, x_n) - x_*| \leq \max(\alpha_n, (1 + \beta_n)|x_n - x_*| - \gamma_n) \tag{14}$$

   where $\alpha_n, \beta_n, \gamma_n$ are sequences of non-negative real numbers with

$$\alpha_n \to 0 \tag{15}$$

$$\sum_n \beta_n < \infty \tag{16}$$

$$\sum_n \gamma_n = \infty \tag{17}$$

2. The cumulative variance of the noise terms W_n is bounded

$$\sum_{n=1}^{\infty} \mathbb{E}[W_n^2] < \infty, \qquad \mathbb{E}[W_n | X_1, \dots, X_n] = 0 \tag{18}$$

and proved that the iterative sequence (10) converges with probability 1 to the fixed point x_*.

The Robbins-Monro theorem in its strongest form (that is, under the weakest assumptions of Blum (9)) becomes easy consequence of Dvoretzky theorem. We only have to check that given assumptions of Blum we can apply Dvoretzky. First, Blum's assumption that the variance of Y_n is bounded by, say, σ^2 for all x and n, given the relation (13), implies $\sum_{n=1}^{\infty} \mathbb{E}[W_n^2] = \sum_{n=1}^{\infty} a_n^2 \sigma^2 < \infty$, and therefore the Dvoretzky's assumption (18) about limited cumulative variance of his noise terms holds. Second, given Blum's $M(x)$ in equation (9), we will construct $\alpha_n, \beta_n, \gamma_n$ that satisfy (15) and such that the bound on the operator T in (14) holds. To do that, first choose a real-valued series $\{\rho_n\}$ with $\rho_n > 0$ and $\rho_n \to 0$ such that[4] $\sum_n \rho_n a_n = \infty$. For simplicity assume, by a change of coordinates, that the fixed point $x_* = 0$. Assuming that $M(x)$ is regular and monotonic in a neighborhood of x_* there is an inverse map M^{-1} and then we define the sequence $\{\eta_n\} = \{M^{-1}(\rho_n)\}$ for sufficiently small ρ_n. Next, define (for sufficiently large n)

$$\begin{aligned} \alpha_n &:= \max(\eta_n, B a_n) \\ \beta_n &:= 0 \\ \gamma_n &:= a_n \rho_n \end{aligned} \tag{19}$$

A case-by-case check for $|x| \leq \eta_n$ and for $|x| > \eta_n$ shows that Dvoretzky's bound on (14) holds given the relation (13).

One universal theme passing through the various versions of stochastic approximation convergence theorems is the choice of the scheduling of the step-sizes (or learning rates) a_n.

In the Robbins-Monro scheduling assumption (8), it is clear that the step-sizes have to converge to zero (otherwise the model would fluctuate and never converge to the exact solution). The second assumption $\sum_{n=1}^{\infty} a_n = \infty$ that says that the rates should not converge to zero too fast is also sensible, as otherwise it is easy to imagine a learning schedule with a_n dropping to zero so fast that the iterative process does not reach the fixed point x_* from an initial point x_0 (for a concrete example, see [14, pp. 5]). The third assumption, $\sum_{n=1}^{\infty} a_n^2 < \infty$, is more subtle and technical, it primarily ensures that even in situations when the noise-terms have self-correlation they would not move the iterative process out of its track of converging with probability 1 to the exact fixed point. In certain situations, a slower decreasing of learning rate schedule still leads to convergence, and is faster in practice.

As the above discussion has shown, the Robbins-Monro paper spawned a huge literature on the analysis and applications of such stochastic algorithms. This is because the problem of estimating unknown parameters of a model from observed data is quite a fundamental one, with variants of this problem appearing in one form or another in control theory, learning theory and other fields of engineering.

Because of the pervasive reach of stochastic approximation methods, any serious formalization effort of an algorithm involving parameter estimation when the underlying model is unknown will eventually have to contend with formalizing tricky stochastic convergence proofs. We chose to formalize Dvoretzky's theorem as it implies the convergence of both the Robbins-Monro and Kiefer-Wolfowitz algorithms, various stochastic gradient descent agorithms and various reinforcement learning algorithms such as Q-learning based on Bellman's optimality operator.

---

[4] for example, if we start with $a_n = \frac{1}{n+1}$ take $\rho_n = \frac{1}{\log(n+1)}$, in general take $\rho_n^{-1} = \sum_{k=1}^{n} a_n$, (see [1]).

▶ Remark. Throughout the text which follows, hyperlinks to theorems, definitions and lemmas which have formal equivalents in the Coq development are indicated by a ✿.

2 Dvoretzky's Theorem

After Dvoretzky's original publication [16] of his theorem and several very useful extensions, several shorter proofs have been proposed. A simplified proof was published by Wolfowitz [37] who like Blum relied on the conditional version of Kolmogorov's law exposed by Loève [22]. A third, more simplified proof was published by Derman and Sacks [15], who again relied on the conditional version of Kolomogorov's law, streamlined the chain of inequality manipulations with Dvoretzky's bounding series parameters $(\alpha_n, \beta_n, \gamma_n)$ and used Chebyshev's inequality and the Borel-Cantelli lemma to arrive at a very short proof. Robbins and Siegmund generalized the theorem to the context where the variables take value in generic Hilbert spaces using the methods of supermartingale theory [23], as did Venter [34]. For a survey see Lai [21]. Dvoretzky himself published a revisited version in [17].

We have chosen to formalise the proof following Derman and Sacks [15] as this version appeared to us as being the shortest and most suitable to formalize using constructions from our library of formalized probability theory.

In this paper we present complete formalization of the scalar version of Dvoretzky's theorem, with random variables taking value in one-dimensional real vector space $\mathcal{H} = \mathbb{R}$. A formalization of the version of the theorem dealing with vector-valued random variables is in progress.

Here is a full statement of Dvoretzky's theorem:

▶ **Theorem 4** (Regular Dvoretzky's Theorem ✿). *Assuming the following:*

H$_1$ **:** *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space*

H$_2$ **:** *For* $n = 1, 2, \dots$

H$_3$ **:** *Let* $\mathcal{F}_n$ *be an increasing sequence of sub* σ*-fields of* $\mathcal{F}$

H$_4$ **:** *Let* X_n *be* $\mathcal{F}_n$*-measurable random variables taking values in a Hilbert space* $\mathcal{H}$ *(here* $\mathcal{H} \equiv \mathbb{R}$*)*

H$_5$ **:** *Let* $T_n : \mathcal{H}^n \to \mathcal{H}$ *be a measurable function*

H$_6$ **:** *Let* W_n *be* $\mathcal{F}_{n+1}$*-measurable random variables taking values in* $\mathcal{H}$ *such that*

$$X_{n+1} = T(x_1, \dots, x_n) + W_n$$

H$_7$ **:** $\mathbb{E}(W_n|\mathcal{F}_n) = 0$

H$_8$ **:** $\sum_{n=1}^{\infty} \mathbb{E}W_n^2 < \infty$

H$_9$ **:** *Let* $\alpha_n, \beta_n, \gamma_n$ *be a series of real numbers such that*

H$_{10}$ **:** $\alpha_n \geq 0$

H$_{11}$ **:** $\beta_n \geq 0$

H$_{12}$ **:** $\gamma_n \geq 0$

H$_{13}$ **:** $\lim_{n=\infty} \alpha_n = 0$

H$_{14}$ **:** $\lim_{n=\infty} \sum_{k=1}^{n} \beta_k < \infty$

H$_{15}$ **:** $\lim_{n=\infty} \sum_{k=1}^{n} \gamma_k = \infty$

H$_{16}$ **:** *Let* x_* *be a point of* $\mathcal{H}$ *such that for all* $n = 1, 2, \dots$ *and for all* $x_1, \dots, x_n \in \mathcal{H}$*,*

$$|T_n(x_1, \dots, x_n) - x_*| \leq \max(\alpha_n, (1 + \beta_n)|x_n - x_*| - \gamma_n)$$

Then the sequence of random variables $X_1, X_2, \dots$ *converges with probability 1 to* x_**:*

$$P\{\lim_{n=\infty} X_n = x_*\} = 1$$

An increasing sequence $\mathcal{F}_n$ of sub-σ-fields of $\mathcal{F}$ (a filtration) formalizes a notion of a discrete stochastic process moving forward in time steps n, where $\mathcal{F}_n$ formalizes the history of the process up to the time step n. Assuming an $\mathcal{F}_n$-measurable random variable X_n means assuming a stochastic variable X_n that is included into the history up to the time step n.

We have also formalized the extended version of Dvoretzky's theorem in which $\alpha_n, \beta_n, \gamma_n$ are promoted to real valued functions and T_n is promoted to be an $\mathcal{F}_n$-measurable random variable. The hypotheses that have been modified in the extended version are marked by the symbol $\star$ below:

▶ **Theorem 5** (Extended Dvoretzky's theorem ✿). *Assuming the following:*
H$_1$: *Let $(\Omega, \mathcal{F}, P)$ be a probability space*
H$_2$: *For $n = 1, 2, \ldots$*
H$_3$: *Let $\mathcal{F}_n$ be an increasing sequence of sub σ-fields of $\mathcal{F}$*
H$_4$: *Let X_n be $\mathcal{F}_n$-measurable random variables taking values in a Hilbert space $\mathcal{H}$ (in this paper we restrict to one-dimensional value space $\mathcal{H} \equiv \mathbb{R}$)*
$\star$**H$_5$:** *Let T_n be $\mathcal{F}_n$-measurable $\mathcal{H}$-valued random variable*
H$_6$: *Let W_n be $\mathcal{F}_{n+1}$-measurable $\mathcal{H}$-valued random variables such that:*

$$X_{n+1} = T(x_1, \ldots, x_n) + W_n$$

H$_7$: $\mathbb{E}(W_n|\mathcal{F}_n) = 0$
H$_8$: $\sum_{n=1}^{\infty} \mathbb{E}W_n^2 < \infty$
$\star$**H$_9$:** *Let $\alpha_n, \beta_n, \gamma_n : \Omega \to \mathbb{R}$ be functions*[5] *such that:*
H$_{10}$: $\alpha_n \geq 0$
H$_{11}$: $\beta_n \geq 0$
H$_{12}$: $\gamma_n \geq 0$
$\star$**H$_{13}$:** $\lim_{n\to\infty} \alpha_n = 0$ *with probability 1*
$\star$**H$_{14}$:** $\lim_{n\to\infty} \sum_{k=1}^{n} \beta_k < \infty$ *with probability 1*
$\star$**H$_{15}$:** $\lim_{n\to\infty} \sum_{k=1}^{n} \gamma_k = \infty$ *with probability 1*
H$_{16}$*: Let x_* be a point of $\mathcal{H}$ such that for all $n = 1, 2, \ldots$ and for all $x_1, \ldots, x_n \in \mathcal{H}$ we have:*

$$|T_n(x_1, \ldots, x_n) - x_*| \leq \max(\alpha_n, (1 + \beta_n)|x_n - x_*| - \gamma_n) \tag{20}$$

Then the sequence of random variables $X_1, X_2, \ldots$ converges with probability 1 to x_*:

$$P\{\lim_{n\to\infty} X_n = x_*\} = 1$$

We now turn to describing the formalization of the above theorems. First, we give a description of our comprehensive supporting Probability Theory library in Section 3 (which may be of independent interest), then we shall give an overview of the proof of Theorem 4 in Section 4.1, and finally detail the variants of this theorem we have formalized in Section 4.2.

3 Formalized Probability Library

Our formalization of both Dvoretzky theorems is built on top of our general library of formalized Probability Theory. In particular, we are not restricted to discrete probability measures.

---

[5] Technically, Dvoretzky in his revisited paper [17] requires $\alpha_n, \beta_n, \gamma_n$ to be $\mathcal{F}_n$-measurable, but we formalized the theorem without having to assume this.

3.1 σ-Algebras and Probability Spaces

We first introduce `pre_events` ✿ which are just subsets of a type `T` i.e., maps `T` → `Prop`. Then we define σ-algebras, `SigmaAlgebra(T)` ✿, as collections of `pre_events` which are closed under countable union and complement and include the full subset of all elements in `T`:

```
Class SigmaAlgebra (T : Type) :=
   {
   sa_sigma : pre_event T → Prop;
   sa_countable_union (collection: nat → pre_event T) :
     (forall n, sa_sigma (collection n)) →
     sa_sigma (pre_union_of_collection collection);
   sa_complement (A:pre_event T) :
     sa_sigma A → sa_sigma (pre_event_complement A) ;
   sa_all : sa_sigma pre_Ω
   }.
```

Then, we label `pre_events` which are members of a σ-algebra as `events` ✿. Special σ-algebras, like that generated by a set of `pre_events` ✿ and the Borel σ-algebra ✿, are constructed as usual.

One interesting feature of the formalization of both of these is that they are both provided with alternative characterizations, which is useful for using the definitions. For the borel σ-algebra, we define two variants: `borel_sa` ✿, defined as the σ-algebra generated by the half-open intervals, and `open_borel_sa` ✿, defined as the σ-algebra generated by the open sets. After proving that the definitions yield the same σ-algebra ✿, we can choose which definition is simpler to work with in a given context, simplifying some proofs.

For the definition of $\sigma(X)$, the σ-algebra generated by a set X, we start with the standard definition ✿: the intersection ✿ of the set of σ-algebras that contain X ✿. This is useful, but as it is non-constructive, it lacks a convenient induction principle. As an alternative, we define the explicit closure of a set of `events` ✿, built by starting with the set (augmented by Ω), and repeatedly adding in complements and countable unions. In Coq, this is naturally defined using an inductive data type. This closure is then shown to be (a σ-algebra ✿ and) equivalent to $\sigma(X)$ ✿.

While definitions generally use the standard definition of $\sigma(X)$, some theorems are more easily proven by switching to the equivalent closure-based characterization. This enables induction, providing an easy way to extend a property on the generating set to the generated σ-algebra, by showing that complements and countable unions preserve the property in question.

Next, we introduce probability spaces ✿ over a σ-algebra, equipped with a measure mapping each `event` to a real number r, such that $0 \leq r \leq 1$.

```
Class ProbSpace {T : Type} (σ : SigmaAlgebra T) :=
{
  ps_P : event σ → R;
  ps_proper :> Proper (event_equiv ==> eq) ps_P ;
  ps_countable_disjoint_union (collection: nat → event σ) :
  (* Assume: collection is a subset of Sigma and
     its elements are pairwise disjoint. *)
    collection_is_pairwise_disjoint collection →
    sum_of_probs_equals ps_P collection (ps_P (union_of_collection collection));
  ps_one : ps_P Ω= R1;
  ps_pos (A:event σ): (0 <= ps_P A)
}.
```

The usual properties of probability spaces, such as monotonicity ✿, complements ✿, and non-disjoint unions ✿, are verified.

3.2 Almost Everywhere

Having defined probability spaces, we can introduce a commonly used assertion in probabilistic proofs: that a certain property holds *almost everywhere* on a probability space. By this we mean the set of points where the property holds includes a measurable event of measure 1. We define a predicate `almost` ✿ to indicate propositions which hold almost everywhere. It is parameterized by a probability space and proposition on that space.

```
Definition almost {Ts:Type} {dom: SigmaAlgebra Ts}(prts: ProbSpace dom) (P:Ts → Prop)
  := exists E, ps_P E = 1 ∧ forall x, E x → P x.
```

We have introduced machinery to make it more convenient to reason about `almost` propositions. For example, if we want to show that `almost P → almost Q → almost R`, we reduce the proof to showing that `almost (P → Q → R)` ✿, which itself is implied by `P → Q → R` ✿. Usual theorem proving tools can then be used.

On top of the basic `almost` definition, we defined `almostR2` ✿, which says that a binary relation holds `almost` everywhere.

```
Definition almostR2 (R:Td→ Td→ Prop) (r1 r2:Ts → Td) : Prop
  := almost (fun x ⇒ R (r1 x) (r2 x)).
```

This is useful, since it inherits many properties from the base relation (e.g. it is a preorder if the base relation is ✿), and simplifies definitions.

3.3 Measurability and Expectation

We next introduce the concept of measurable functions with respect to two σ-algebras. Since we are focusing on probability spaces, we call these measurable functions `RandomVariables` ✿.

```
(* A random variable is a mapping from a probability space to a sigma algebra. *)
Class RandomVariable {Ts:Type} {Td:Type}
      (dom: SigmaAlgebra Ts)
      (cod: SigmaAlgebra Td)
      (rv_X: Ts → Td)
  := (* for every element B in the sigma algebra, the preimage
     of rv_X on B is an event in the probability space *)
    rv_preimage_sa: forall (B: event cod), sa_sigma (event_preimage rv_X B).
```

In order to define the `Expectation` of a `RandomVariable`, we follow the usual technique of first treating the case of finite range functions ✿, then extending to nonnegative functions ✿ (resulting in an extended real) and then to general random variables. In the general case, the expectation is the difference of the expectation of the positive and negative parts of a random variable. ✿ Exceptions are handled using the Coq `option` type. For example, the difference of the expectations of the positive and negative parts of a random variable is not defined if they are both the same infinity. This exception is captured by allowing the difference to be `None` in that case.

```
Definition Expectation (rv_X : Ts → R) : option Rbar :=
  Rbar_minus' (NonnegExpectation (pos_fun_part rv_X))
              (NonnegExpectation (neg_fun_part rv_X)).
```

Originally our results about `Expectation` were for random variables taking images in the reals, but as we introduced limiting processes we needed to extend the concept to random variables taking values in the extended reals including plus and minus infinity for which we use and extend the implementation of `Rbar` in the Coquelicot library [13].

This requires extending the support for limits in Coquelicot, allowing for sequences of functions over the extended reals ✿. The approach we took was to copy over all the definitions and lemmas in Coquelicot's `Lim_seq` module, extending them as appropriate, and re-proving them. A few changes were made, such as defining the extended version of `is_lim_seq` ✿ to hold when the inf and sup sequence limits coincide. The original definition uses filters, and is problematic to extend to the extended reals. Pleasantly, however, almost all of the lemmas continue to hold with minor modification.

The above construction of Expectation and its properties (including linearity ✿✿, the monotone convergence theorem ✿, and other standard results) are then generalized and proven for this generalization to functions whose image is the extended reals ✿.

On top of our general definition of `Expectation`, we define the `IsFiniteExpectation` property, which asserts that a function has a well-defined, finite expectation ✿✿. For functions that satisfy this property, we can define their `FiniteExpectation` ✿✿, which returns their (real) expectation. This simplifies working with such functions, and avoids otherwise necessary side-conditions on properties such as linearity ✿✿.

3.4 L^p Spaces

Using these building blocks, we can define L^p spaces, which are the space of measurable functions where the p-th power of its absolute value has finite expectation ✿.

```
Definition IsLp {Ts} {dom: SigmaAlgebra Ts} (prts: ProbSpace dom) (n:R) (rv_X:Ts→ R)
    := IsFiniteExpectation prts (rvpower (rvabs rv_X) (const n)).
```

This space is then quotiented, identifying functions that are equal `almost` everywhere (see Section 3.2) ✿. We use a quotient construction ✿ that avoids needing axioms beyond those already proposed in Coq's standard libraries. This quotienting operation is required in order to define a norm on the space (defined as the p-th root of the `Expectation` of the absolute value of the p-th power of the function), as having a zero `Expectation` only implies that a non-negative function is zero `almost` everywhere.

For nonnegative p, L^p is shown to be a module space ✿. For $1 \leq p \leq \infty$, it is shown to be a Banach space (complete normed module space) ✿ ✿.

Furthermore, the important special case of L^2 is proven to be a Hilbert space ✿, where the inner product of x and y is defined as the `Expectation` of the product of x and y.

3.5 Conditional Expectation

Building on top of this work, we turn to the definition of conditional expectation, defining it with respect to a general σ-algebra `dom2` (the ambient σ-algebra being `dom`). We first postulate a relational definition ✿, characterized by the universal property of conditional expectations: for any event P that is in the sub σ-algebra `dom2`, if we multiply the original function and its conditional expectation by that event's associated indicator function, we get equal expectations.

```
Definition is_conditional_expectation {Ts:Type} {dom: SigmaAlgebra Ts}
    (prts: ProbSpace dom) (dom2 : SigmaAlgebra Ts)
    (f : Ts → R) (ce : Ts → Rbar)
```

```
    {rvf : RandomVariable dom borel_sa f}
    {rvce : RandomVariable dom2 Rbar_borel_sa ce}
  := forall P (dec:dec_pre_event P),
    sa_sigma (SigmaAlgebra := dom2) P →
    Expectation (rvmult f (EventIndicator dec)) =
    Rbar_Expectation (Rbar_rvmult ce (EventIndicator dec)).
```

Using this definition, we can show uniqueness (where equality is almost everywhere) ✿, and many standard properties of conditional expectation, such as linearity ✿✿, preservation of `Expectation` ✿, (`almost`) monotonicity ✿, and the tower law ✿. We also show the "factor out" property ✿, which enables factoring out of a conditional expectation a random variable that is measurable with respect to the sub σ-algebra. In addition, we verify its interactions with limits (e.g. the conditional version of the monotone convergence theorem ✿), and prove Jensen's lemma ✿, bounding how convex functions affect the conditional expectation.

After having proven these properties for the `is_conditional_expectation` relation, we still need to show that the conditional expectation generally exists (at least for functions that are non-negative or have finite expectation).

To do this, we build on our work on L^p spaces (Section 3.4), and in particular our proof that that L^2 is a Hilbert space. Given an L^2 function, this implies that the subset of functions which are measurable with respect to a smaller σ-algebra `dom2` forms a linear subspace.

The L^2 conditional expectation ✿ of an L^2 random variable X with respect to `dom2` is then defined as the orthogonal projection ✿ of X onto that subspace. For this construction and definitions of Hilbert spaces we use the library from the formal development of the Lax-Milgram theorem [12]. Note that this definition is for a function in the quotiented space (recall that L^2 is quotiented to identify functions that are equal almost everywhere).

We can then define conditional expectation on the unquotiented space by injecting the inputs into the quotiented space, using the conditional expectation operator just defined on L^2 functions, and then choosing a representative from the equivalence class of functions it returns ✿. This unquotienting gives insight into why most theorems about conditional expectations only `almost` hold, as it is defined on equivalence classes of `almost` equal functions.

Next, we extend our notion of conditional expectation to nonnegative functions whose usual expectation is finite using the property that L^2 functions are dense in L^1. In particular, given a nonnegative L^1 function f, we can define an L^2 sequence $g_n = \min(f, n)$. The conditional expectation of f is defined as the limit of the conditional expectation of the g_n ✿.

Using this definition directly has some disadvantages: it forces essentially all theorems, including simple ones such as the result being non-negative, or that the conditional expectation is the identity operation on functions that are already measurable with respect to the sub σ-algebra `dom2`, be only `almost` valid. To address this, we wrap this definition in a wrapper ✿ that takes the function returned by the original (limit based) definition and tweaks it slightly, producing a "fixed" function `almost` equivalent to the original, but where such simple properties hold unconditionally.

Finally, we extend this to all measurable functions by taking the difference of the nonnegative conditional expectation of its positive and negative parts ✿. While this function is defined for all measurable functions, it can only be shown to be a conditional expectation (the `is_conditional_expectation` relation defined above) for functions that are either non-negative ✿ or have finite expectation ✿. Using this property, we now lift all of the properties proven above for the relational version to our explicitly defined version, verifying that it satisfies all the expected properties. For convenience, we also provide a wrapper definition `FiniteConditionalExpectation` ✿, which assumes that the function has finite expectation, and returns a function whose image is in $\mathbb{R}$ (insted of the extended reals), and lift all the expected properties to it.

Connecting back to L^p spaces, we can use Jensen's lemma about convex functions to show that if a function is in L^p then its conditional expectation is as well ✿, allowing us to view conditional expectation as a (contractive ✿) operation on L^p spaces. Furthermore, we show that it minimizes the L^2-loss for an L^2 function ✿.

We chose this approach to defining conditional expectation (via an orthonormal projection on L^2) since we could rely on an existing library of Hilbert space theory [12], thus avoiding other tedious constructions involving Radon-Nikodym derivatives etc.

3.6 Filtrations and Martingales

We next introduce a notion of σ-algebra filtrations ✿, which are an increasing sequence of σ-algebras. We say that a sequence of random variables X_n `IsAdapted` ✿ to a filtration F_n, if each random variable of the sequence is measurable with respect the corresponding σ-algebra.

Building on these definitions and our development of conditional expectation, we started developing the basics of martingale theory ✿.

Additionally, the language of filtrations and adapted processes enables us to represent the history of a stochastic process, which is critical for stating and verifying properties of stochastic approximation methods.

3.7 Additional results

There are many other results proven in the library; here we highlight two that are used in the Derman-Sacks proof: Chebyshev's inequality and the Borel-Cantelli lemma.

Chebyshev's inequality ✿ which states that given a random variable X and a positive constant a, the probability of $\|X\| \geq a$ is less that or equal to the expectation of X^2/a^2.

```
Lemma Chebyshev_ineq_div_mean0
(X : Ts → R) (rv : RandomVariable dom borel_sa X) (a : posreal) :
Rbar_le (ps_P (event_ge dom (rvabs X) a))
        (Rbar_div_pos
        (NonnegExpectation (rvsqr X))
        (mkposreal _ (rsqr_pos a))).
```

Another is the Borel-Cantelli lemma ✿ which states that if the sum of probabilities of a sequence of events is finite, then the probability of all but finitely many of them occuring is 0.

```
Theorem Borel_Cantelli (E : nat → event dom) :
(forall (n:nat), sa_sigma (E n)) →
ex_series (fun n ⇒ ps_P (E n)) →
ps_P (inter_of_collection
        (fun k ⇒ union_of_collection
                 (fun n ⇒ E (n + k)))) = 0.
```

In this theorem statement, `ex_series f`, defined in Coquelicot, assert that the infinite series of partial sums $\lim_{n\to\infty} \sum_{0\leq i\leq n} f(i)$ converges to a finite limit.

4 Formalization Challenges/Overview

We will now sketch the key pieces which go into the formalization of the Derman-Sacks proof.

4.1 Overview of the proof

The Derman-Sacks proof relies on a number of prerequisites in Probability Theory and Real Analysis. For example, the proof begins by stating that we may replace the series $\sum_n \mathbb{E}W_n^2 < \infty$ by the series $\sum_n \frac{\mathbb{E}W_n^2}{\alpha_n^2} < \infty$ where $\alpha_n \to 0$. This statement invokes a classical theorem of du Bois-Reymond [11] which states:

▶ **Theorem 6.** ✿ *Let (a_n) be a sequence of nonnegative real numbers. The series $\sum_n a_n$ converges if and only if there is another sequence of positive real numbers (b_n) such that $b_n \to \infty$ and $\sum_n a_n b_n < \infty$.*

In other words, this theorem states that *no worst convergent series exists* (see [4]). This elementary theorem did require some effort to formalize, in part because existing proofs such as the one in [4] require the sequence (a_n) to consist only of positive terms, while our application (Dvoretzky's theorem) needed them to be non-negative. Additionally, we had to prove convergence of the product series without using the integral test (as used in [4]), because it was unavailable in our library. Our final proof of Theorem 6 involved a case analysis in which we case on whether the sequence (a_n) was eventually positive or not ✿, and we bypassed the need to use the integral test by using an exercise from Rudin's *Principles of Mathematical Analysis* [25].

The main workhorse of the Derman-Sacks proof is the sequence $Z_n := W_n \operatorname{sgn} T_n$. First, they apply the following theorem[6] to the sequence of random variables (Z_n):

▶ **Theorem 7** (Loève [22] ✿). *Let $X_1, X_2, \ldots$ be a sequence of random variables adapted to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Assume that $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = 0$ almost surely for all n and also that $\sum_{n=1}^{\infty} \mathbb{E}X_n^2$ converges. Then we have that $\sum_{n=1}^{\infty} X_n$ converges almost surely.*

to conclude that the series $\sum_n Z_n$ converges almost surely. To apply this theorem we need to prove that (Z_n) is adapted to the filtration $\mathcal{F}$, which critically uses the fact that $T_n : \mathcal{H}^n \to \mathcal{H}$ is a measurable function. (Here we take $\mathcal{H} = \mathbb{R}$.) The proof of the theorem uses $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = 0$ to show that since the sequence is adapted, we have $\mathbb{E}[X_i X_j] = 0$ for all $i \neq j$. This depends on the "factor out" property of conditional expectation ✿ (see Section 3.5).

Next, it is shown that $|Z_n| \leq \alpha_n$ almost surely for sufficiently large n. This argument uses the Borel-Cantelli lemma ✿ and the Chebyshev inequality ✿, both of which needed a significant amount of probability theory to be set up (see Section 3.7). Using this bound for Z_n and the bound for $|T_n|$ in the hypothesis, an elementary argument shows that

$$|X_{n+1}| \leq \max(2\alpha_n, |T_n| + Z_n) \leq \max(2\alpha_n, (1+\beta_n)|X_n| + Z_n - \gamma_n)$$

almost surely for sufficiently large n.

Now, the conclusion $X_{n+1} \to 0$ almost surely follows by applying the following lemma:

▶ **Lemma 8.** ✿ *Let $\{a_n\}, \{b_n\}, \{c_n\}, \{\delta_n\}$ and $\{\xi_n\}$ be sequences of real numbers such that*
1. *$\{a_n\}, \{b_n\}, \{c_n\}, \{\xi_n\}$ are non-negative*
2. *$\lim_{n\to\infty} a_n = 0$, $\sum_n b_n < \infty$, $\sum_n c_n = \infty$, $\sum_n \delta_n$ converges.*
3. *For all n larger than some N_0, $\xi_{n+1} \leq \max(a_n, (1+b_n)\xi_n + \delta_n - c_n)$*

then, $\lim_{n\to\infty} \xi_n = 0$.

---

[6] the proof of this theorem is a modification of Theorem 6.2.1 in Ash's *Probability and Measure Theory* [5]

The proof of the lemma is somewhat unusual since it involves running an iteration backwards: the property (3) is applied repeatedly to derive an inequality between ξ_{n+1} and ξ_N for $n > N > N_0$ ✿. Besides using several properties of infinite products and list maximums, the final convergence result is an application of Abel's descending convergence criterion ✿ which says if the series $\sum_n b_n$ converges, and a_n is a bounded descending sequence, then the series $\sum_n a_n b_n$ also converges.

4.2 Variants of Dvoretzky's Theorem.

While Dvoretzky's theorem admits generalizations in many different ways, we chose to focus on formalizing the ones most suited for applications.

1. As already mentioned, we prove Theorem 5 which is a generalization of Theorem 4 in which the sequences of numbers $\alpha_n,\ \beta_n,\ \gamma_n$ are replaced by sequences of functions on the probability space. This generalization is called the *extended* Dvoretzky theorem ✿. All conditions on the sequences $\alpha_n,\ \beta_n,\ \gamma_n$ now hold pointwise, almost everywhere.
2. To apply Theorem 7 in the proof of Theorem 4 we needed to prove that (Z_n) is adapted to the filtration $\mathcal{F}$, which needed us to make assumptions on the functions T_n. These assumptions on T_n can be modified and generalized as:
   a. in the regular (non-extended) case, $T_n : \mathbb{R}^n \to \mathbb{R}$ are deterministic and measurable. ✿
   b. in the extended case, $T_n : \mathbb{R}^n \times \Omega \to \mathbb{R}$ are stochastic and $\mathcal{F}_n$-adapted. ✿

   Since Derman-Sacks do not explicitly state either assumption, we formalized Dvoretzky's theorem under both assumptions. It should be noted that Dvoretzky's original paper [16] and his revisited paper [17] treat both the above cases.
3. We have also formalized a corollary of the extended Dvoretzky's theorem ✿ which proves that the theorem holds in the context where the bound on T in (14) is assumed as follows with all other assumptions intact:

   $$|T_n(x_1, \ldots, x_n) - x_*| \leq \max(\alpha_n, (1 + \beta_n - \gamma_n)|x_n - x_*|)$$

   While this formulation is weaker compared to the original, it is convenient to have it for several applications of stochastic approximation theorems. A proof of this corollary used a classical analysis result of Abel [1] on the fact that the terms in a divergent sum-series could be multiplied by infinitesimally small series and the sum-series would still diverge ✿. This was addressed in Dvoretzky's paper [16, (5.1)].

5 Related work

While our results are general, our intended application was formalizing machine learning theory, on which there is a growing body of work [28, 29, 31, 18, 26, 8, 9]. Our work is a step in this direction, providing future developers of secure machine learning systems a library of formalized stochastic approximation results. Keeping this in mind, we have formalized different versions of our main result (Dvoretzky's theorem) to facilitate ease of use (see Section 4.2).

For the formalization itself, we make extensive use of the Coquelicot library of Boldo et al. [13] and the library which proved the Lax-Milgram theorem [12]. There have also been quite a few formalizations of probability theory in Coq: see Polaris [27], Infotheo [3], and Alea [6].

Alea is an early work and to the best of our knowledge incompatible with latest versions of Coq while Infotheo and Polaris either fundamentally focus on discrete probability theory (see [2]) or do not have the results we needed to prove Dvoretzky's theorem.

Formal proofs about convergence of random variables (the Central Limit Theorem) have been given in Avigad et al [7] using the Isabelle/HOL system. Parts of Martingale theory and stochastic processes have also recently made their way into the Lean math library [30].

To the best of our knowledge, our work presents the first formal proof of correctness of any theorem in Stochastic Approximation.

6 Applications & Future Work

Our own interest in stochastic approximation began with an attempt to extend our work on convergence proofs of (model-based) Reinforcement Learning (RL) algorithms [33] to include the model-free case. Model-based RL algorithms converge to an *optimal policy* (a sequence of actions which an agent should probabilistically perform so as to maximize its expected long-term reward) by making full use of the given transition probability structure of the agent. The term *model-free* refers to the fact that we have no information on how the agent performs it's transitions but can only *observe* its transitions. As we have emphasized above, this situation is perfectly suited for stochastic approximation techniques. Indeed, convergence proofs of Q-Learning (a prominent model-free RL algorithm) appeal to standard results of stochastic approximation (see Watkins & Dayan [35], Jaakkola et al. [19]), Tsitsiklis [32]. We plan to use our formalization of Dvoretzky's theorem to complete a convergence proof of the Q-learning algorithm.

Additionally, as part of this process, we have built up a large library for basic results on (general) probability spaces in Coq, including a general definition of conditional expectation. This library is publically available at `https://github.com/IBM/FormalML` and open source. We invite others to use our library and collaborate with us on extending and enhancing it.

References

1 N. H. Abel. Note sur le mémoire de Mr. L. Olivier No. 4 · du second tome de ce Journal, ajant pour titre „remarques sur les series infinies et leur convergence.". 1828(3):79–82, 1828. URL: `https://doi.org/10.1515/crll.1828.3.79`, `doi:doi:10.1515/crll.1828.3.79`.

2 Reynald Affeldt, Jacques Garrigue, and Takafumi Saikawa. Reasoning with conditional probabilities and joint distributions in coq. *Computer Software*, 37(3):79–95, 2020. `doi:10.11309/jssst.37.3_79`.

3 Reynald Affeldt and Manabu Hagiwara. Formalization of Shannon's theorems in SSReflect-Coq. In *3rd Conference on Interactive Theorem Proving (ITP 2012), Princeton, New Jersey, USA, August 13–15, 2012*, volume 7406 of *Lecture Notes in Computer Science*, pages 233–249. Springer, Aug 2012.

4 J Marshall Ash. Neither a worst convergent series nor a best divergent series exists. *The College Mathematics Journal*, 28(4):296–297, 1997.

5 Robert B Ash, B Robert, Catherine A Doleans-Dade, and A Catherine. *Probability and measure theory*. Academic press, 2000.

6 Philippe Audebaud and Christine Paulin-Mohring. Proofs of randomized algorithms in Coq. *Science of Computer Programming*, 74(8):568–589, 2009.

7 Jeremy Avigad, Johannes Hölzl, and Luke Serafin. A formally verified proof of the central limit theorem. *Journal of Automated Reasoning*, 59(4):389–423, 2017.

8 Alexander Bagnall and Gordon Stewart. Certifying the True Error: Machine Learning in Coq with Verified Generalization Guarantees. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 2662–2669. AAAI Press, 2019. `doi:10.1609/aaai.v33i01.33012662`.

9 Alexander Bentkamp, Jasmin Christian Blanchette, and Dietrich Klakow. A Formal Proof of the Expressiveness of Deep Learning. *J. Autom. Reason.*, 63(2):347–368, 2019. doi:10.1007/s10817-018-9481-5.

10 Julius R. Blum. Approximation Methods which Converge with Probability one. *The Annals of Mathematical Statistics*, 25(2):382 – 386, 1954. doi:10.1214/aoms/1177728794.

11 Paul du Bois-Reymond. Eine neue Theorie der Convergenz und Divergenz von Reihen mit positiven Gliedern. 1873.

12 Sylvie Boldo, François Clément, Florian Faissole, Vincent Martin, and Micaela Mayero. A Coq formal proof of the Lax–Milgram theorem. In *6th ACM SIGPLAN Conference on Certified Programs and Proofs*, Paris, France, January 2017. URL: https://hal.inria.fr/hal-01391578, doi:10.1145/3018610.3018625.

13 Sylvie Boldo, Catherine Lelay, and Guillaume Melquiond. Coquelicot: A user-friendly library of real analysis for Coq. *Mathematics in Computer Science*, 9, 03 2014. doi:10.1007/s11786-014-0181-1.

14 Han-Fu Chen. *Stochastic approximation and its applications*, volume 64. Springer Science & Business Media, 2006.

15 C Derman and J Sacks. On Dvoretzky's stochastic approximation theorem. *The Annals of Mathematical Statistics*, 30(2):601–606, 1959.

16 A. Dvoretzky. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press, 1956.

17 Aryeh Dvoretzky. Stochastic approximation revisited. *Advances in Applied Mathematics*, 7(2):220–227, 1986. URL: https://www.sciencedirect.com/science/article/pii/0196885886900333, doi:https://doi.org/10.1016/0196-8858(86)90033-3.

18 Johannes Hoelzl. Markov processes in Isabelle/HOL. In *Proceedings of the 6th ACM SIGPLAN Conference on Certified Programs and Proofs*, CPP 2017, page 100–111, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3018610.3018628.

19 Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6(6):1185–1201, 1994.

20 J. Kiefer and J. Wolfowitz. Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics*, 23(3):462 – 466, 1952. doi:10.1214/aoms/1177729392.

21 Tze Leung Lai. Stochastic Approximation. *The Annals of Statistics*, 31(2):391–406, 2003. URL: http://www.jstor.org/stable/3448398.

22 Michel Loève. On almost sure convergence. Proc. Berkeley Sympos. math. Statist. Probability, California July 31 - August 12, 1950, 279-303 (1951)., 1951.

23 H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Jagdish S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971. URL: https://www.sciencedirect.com/science/article/pii/B9780126045505500158, doi:https://doi.org/10.1016/B978-0-12-604550-5.50015-8.

24 Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

25 Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

26 Daniel Selsam, Percy Liang, and David L. Dill. Developing Bug-Free Machine Learning Systems With Formal Mathematics. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3047–3056. PMLR, 2017. URL: http://proceedings.mlr.press/v70/selsam17a.html.

27 Joseph Tassarotti and Robert Harper. A separation logic for concurrent randomized programs. *Proceedings of the ACM on Programming Languages*, 3(POPL):1–30, 2019.

28 Joseph Tassarotti, Jean-Baptiste Tristan, and Koundinya Vajjha. A Formal Proof of PAC Learnability for Decision Stumps. *CoRR*, abs/1911.00385, 2019. URL: `http://arxiv.org/abs/1911.00385`, `arXiv:1911.00385`.

29 Joseph Tassarotti, Koundinya Vajjha, Anindya Banerjee, and Jean-Baptiste Tristan. A formal proof of PAC learnability for decision stumps. In *Proceedings of the 10th ACM SIGPLAN International Conference on Certified Programs and Proofs*, pages 5–17, 2021.

30 The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020)*, pages 367–381, 2020.

31 Jean-Baptiste Tristan, Joseph Tassarotti, Koundinya Vajjha, Michael L. Wick, and Anindya Banerjee. Verification of ML Systems via Reparameterization. *CoRR*, abs/2007.06776, 2020. URL: `https://arxiv.org/abs/2007.06776`, `arXiv:2007.06776`.

32 John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3):185–202, 1994.

33 Koundinya Vajjha, Avraham Shinnar, Barry Trager, Vasily Pestun, and Nathan Fulton. Certrl: formalizing convergence proofs for value and policy iteration in coq. In *Proceedings of the 10th ACM SIGPLAN International Conference on Certified Programs and Proofs*, pages 18–31, 2021.

34 J. H. Venter. On Dvoretzky Stochastic Approximation Theorems. *The Annals of Mathematical Statistics*, 37(6):1534 – 1544, 1966. `doi:10.1214/aoms/1177699145`.

35 Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

36 J. Wolfowitz. On the Stochastic Approximation Method of Robbins and Monro. *The Annals of Mathematical Statistics*, 23(3):457–461, 1952. URL: `http://www.jstor.org/stable/2236689`.

37 J. Wolfowitz. On Stochastic Approximation Methods. *The Annals of Mathematical Statistics*, 27(4):1151 – 1156, 1956. `doi:10.1214/aoms/1177728082`.