

Constructive law of large numbers with application to countable Markov chains

Péter Gács
Boston University

AMS meeting, New Orleans, 2011

Convergence speed in the law of large numbers

- Typical laws of large numbers: **higher moment** assumptions. Convergence speed is an important part of the statement and the proof.
- In case of identically distributed, (pairwise) independent variables, the law of large numbers follows already from the **existence of the expected value**. In this case, the question of convergence speed is more subtle.
- One weird-sounding sentence summarizes the main result: the convergence is effective if and only if **the expected value is a computable number**.

Convergence speed in the law of large numbers

- Typical laws of large numbers: **higher moment** assumptions. Convergence speed is an important part of the statement and the proof.
- In case of identically distributed, (pairwise) independent variables, the law of large numbers follows already from the **existence** of the **expected value**. In this case, the question of convergence speed is more subtle.
- One weird-sounding sentence summarizes the main result: the convergence is effective if and only if **the expected value is a computable number**.

Convergence speed in the law of large numbers

- Typical laws of large numbers: **higher moment** assumptions. Convergence speed is an important part of the statement and the proof.
- In case of identically distributed, (pairwise) independent variables, the law of large numbers follows already from the **existence** of the **expected value**. In this case, the question of convergence speed is more subtle.
- One weird-sounding sentence summarizes the main result: the convergence is effective if and only if **the expected value is a computable number**.

- [Vyugin](#) in 1998: example of a countable Markov chain with a computable distribution, for which the **convergence** of relative frequencies to their limit, (guaranteed for example by Birkoff's ergodic theorem), is provably **non-effective**. His Markov chain is **not ergodic**.
- My original goal and result was that for ergodic Markov chains, the convergence is effective. By now, this has been proved for general ergodic dynamical systems by [Avigad, Gerhardy, Towsner](#) and later simpler by [Galatolo, Hoyrup, Rojas](#).
- My proof for Markov chains constructivizes a standard proof via the renewal theorem and the law of large numbers. I hope that the findings on the convergence speed in the law of large numbers still hold some interest, and also help illustrating the main issues in a more elementary setting.

- [Vyugin](#) in 1998: example of a countable Markov chain with a computable distribution, for which the **convergence** of relative frequencies to their limit, (guaranteed for example by Birkoff's ergodic theorem), is provably **non-effective**. His Markov chain is **not ergodic**.
- My original goal and result was that for ergodic Markov chains, the convergence is effective. By now, this has been proved for **general ergodic dynamical systems** by [Avigad, Gerhardy, Towsner](#) and later simpler by [Galatolo, Hoyrup, Rojas](#).
- My proof for Markov chains constructivizes a standard proof via the renewal theorem and the law of large numbers. I hope that the findings on the convergence speed in the law of large numbers still hold some interest, and also help illustrating the main issues in a more elementary setting.

- [Vyugin](#) in 1998: example of a countable Markov chain with a computable distribution, for which the **convergence** of relative frequencies to their limit, (guaranteed for example by Birkoff's ergodic theorem), is provably **non-effective**. His Markov chain is **not ergodic**.
- My original goal and result was that for ergodic Markov chains, the convergence is effective. By now, this has been proved for **general ergodic dynamical systems** by [Avigad, Gerhardy, Towsner](#) and later simpler by [Galatolo, Hoyrup, Rojas](#).
- My proof for Markov chains constructivizes a standard proof via the renewal theorem and the law of large numbers. I hope that the findings on the convergence speed in the law of large numbers still hold some interest, and also help illustrating the main issues in a more elementary setting.

Effective convergence

- We will work in **constructive metric spaces** (which have a natural definition).
- A sequence of functions $x_1(t), x_2(t), \dots$ between constructive metric spaces **converges effectively** (pointwise) to $y(t)$ if there is an **upper semicomputable** $m(\varepsilon, t)$ (with real values) such that for all t , all $\varepsilon > 0$ and every $n \geq m(\varepsilon, t)$ we have $d(x_n(t), y(t)) \leq \varepsilon$. We call $m(\varepsilon, t)$ the **threshold function**.
- We could have required $m(\varepsilon, t)$ to be computable as well as to take integer values (but not both).
- If $m(\varepsilon, t)$ does not depend on t then the convergence is **effectively uniform** in t .

Effective convergence

- We will work in **constructive metric spaces** (which have a natural definition).
- A sequence of functions $x_1(t), x_2(t), \dots$ between constructive metric spaces **converges effectively** (pointwise) to $y(t)$ if there is an **upper semicomputable** $m(\varepsilon, t)$ (with real values) such that for all t , all $\varepsilon > 0$ and every $n \geq m(\varepsilon, t)$ we have $d(x_n(t), y(t)) \leq \varepsilon$. We call $m(\varepsilon, t)$ the **threshold function**.
- We could have required $m(\varepsilon, t)$ to be computable as well as to take integer values (but not both).
- If $m(\varepsilon, t)$ does not depend on t then the convergence is **effectively uniform** in t .

Effective convergence

- We will work in **constructive metric spaces** (which have a natural definition).
- A sequence of functions $x_1(t), x_2(t), \dots$ between constructive metric spaces **converges effectively** (pointwise) to $y(t)$ if there is an **upper semicomputable** $m(\varepsilon, t)$ (with real values) such that for all t , all $\varepsilon > 0$ and every $n \geq m(\varepsilon, t)$ we have $d(x_n(t), y(t)) \leq \varepsilon$. We call $m(\varepsilon, t)$ the **threshold function**.
- We could have required $m(\varepsilon, t)$ to be computable as well as to take integer values (but not both).
- If $m(\varepsilon, t)$ does not depend on t then the convergence is **effectively uniform** in t .

- We will work in **constructive metric spaces** (which have a natural definition).
- A sequence of functions $x_1(t), x_2(t), \dots$ between constructive metric spaces **converges effectively** (pointwise) to $y(t)$ if there is an **upper semicomputable** $m(\varepsilon, t)$ (with real values) such that for all t , all $\varepsilon > 0$ and every $n \geq m(\varepsilon, t)$ we have $d(x_n(t), y(t)) \leq \varepsilon$. We call $m(\varepsilon, t)$ the **threshold function**.
- We could have required $m(\varepsilon, t)$ to be computable as well as to take integer values (but not both).
- If $m(\varepsilon, t)$ does not depend on t then the convergence is **effectively uniform** in t .

It is sometimes more convenient to work with the inverse, inspired by the notion of “order function” in Schnorr:

A function $b(n, t)$ is a **shrinkorder function** if: it is upper semicomputable, with $b(n, t) \searrow 0$ for all t .

Proposition

The sequence $x_1(t), x_2(t), \dots$ converges effectively to $y(t)$ if and only if there is a shrinkorder function $b(n, t)$ with $d(x_n(t), y(t)) \leq b(n, t)$ for all n, t .

For the case of a sequence of numbers x_n , we can even require $b(n)$ to be **computable**.

Proposition

Let $z_n \geq 0$, $y = \sum_n z_n$. If z_n, y are (uniformly) computable then the convergence is effective.

More generally, if the convergence is effective relative to $(z_n), y$.

Indeed, $b(n) = y - \sum_{i=1}^n z_n$ is a shrinkorder function.

Detachment: Faster computability of z_n and y does not imply faster convergence to y : these two convergences are **detached** from each other. Indeed, given an **arbitrary** computable shrinkorder function $b(n)$ with $b(0) = 1$, the sum of the series $z_n = b(n-1) - b(n)$ converges as slowly as $b(n)$.

We can make $b(n)$ computable even in linear time without making it decrease faster.

This remark applies to all main effectiveness results in the talk.

Computable distributions

There are many, equivalent, ways to define computability of probability distributions. Our results are sufficiently interesting even if one confines interest to integer-valued random variables. In this case, the distribution P is computable if and only if $P(x)$ is computable for all integers x .

Proposition

There exists a computable distribution over the nonnegative integers, with a non-computable expected value.

Proof. Let $\alpha_i = 0$ or 2^{-k} for some k , with noncomputable $\sum_i \alpha_i < 1$. Construct $p(n)$ gradually: If $\alpha_i = 0$ then add 2^{-i} to $p(0)$. If $\alpha_i = 2^{-k}$ then add 2^{-i} to $p(2^{i-k})$. □

Effective stochastic convergence

A sequence of random variables X_1, X_2, \dots with joint probability distribution P **effectively converges to Y in probability, or stochastically**, if there is an upper semicomputable function $m(\delta, \varepsilon)$ such that $n \geq m(\delta, \varepsilon)$ implies $P[|X_n - Y| > \delta] < \varepsilon$.

We say that $X_n \rightarrow Y$ **almost surely, effectively**, if the last inequality is replaced with $P[\sup_{n \geq m(\delta, \varepsilon)} |X_n - Y| > \delta] < \varepsilon$.

Proposition

The sequence X_n converges to Y in probability effectively if and only if there are shrinkorder functions $b_1(n), b_2(n)$ with the property

$$P[|X_n - Y| > b_1(n)] \leq b_2(n)$$

for all n . Similar characterization holds for almost sure convergence.

Of course, we could make $b_1(n) = b_2(n)$.

It is not surprising that an effective law of large numbers allows to compute the expected value:

Proposition

Let X_1, X_2, \dots be a sequence of identically distributed random variables with distribution P and expected value μ . If $\frac{1}{n} \sum_{i=1}^n X_i$ converges to μ effectively in probability, then μ is computable from P .

Corollary

If μ is not computable then even if the distribution P is computable, the convergence to μ in the weak law of large numbers is not effective.

We have seen a computable distribution P over the nonnegative integers, with a non-computable expected value.

It is not surprising that an effective law of large numbers allows to compute the expected value:

Proposition

Let X_1, X_2, \dots be a sequence of identically distributed random variables with distribution P and expected value μ . If $\frac{1}{n} \sum_{i=1}^n X_i$ converges to μ effectively in probability, then μ is computable from P .

Corollary

If μ is not computable then even if the distribution P is computable, the convergence to μ in the weak law of large numbers is not effective.

We have seen a computable distribution P over the nonnegative integers, with a non-computable expected value.

Theorem (Constructive strong law)

Let X_1, X_2, \dots be a sequence of identically distributed, pairwise independent random variables with distribution P . Let

$\mathbf{E}|X| = \mu < \infty$. Let $S_n = \sum_{i=1}^n X_i$. The average S_n/n converges to $\mathbf{E}X$ effectively in (P, μ) .

The earlier remark on **detachment** applies to this result, too: the speed of computability of μ does not influence directly the speed of convergence in the law of large numbers.

Straightforward constructivization of a textbook proof. The truncated variables

$$Y_n = X_n 1_{[|X_n| < n]}.$$

have variances, and even $\sum_{n>0} \text{Var} Y_n / n^2 < \infty$ (due to identical distribution), effectively. Prove the strong law for Y_n first.

Bound the difference $X_n - Y_n$ using

$$P[(\exists k \geq n) Y_k \neq X_k] \leq \sum_{k \geq n} P[|X_k| \geq k] = \sum_{k \geq n} P[|X_1| \geq k].$$

The latter sum is a tail of $\sum_{k \geq 0} P[|X_1| \geq k]$, related to the tails of

$$\int_0^{\infty} P[|X_1| \geq t] dt = \mathbf{E}|X_1|.$$

Let T_0, T_1, T_2, \dots be integer random variables in $[0, \infty]$, where T_0 has distribution Q over $\mathbb{Z} \cap [0, \infty)$, and for $i > 0$ the values $T_i - T_{i-1} > 0$ are identically distributed with distribution R and $\mathbf{E}(T_i - T_{i-1}) = \mu < \infty$, and independent (also from T_0). Define

$$X_m = 1 \text{ if } (\exists i) m = T_i, \text{ and } 0 \text{ otherwise .}$$

X_0, X_1, \dots is called a (positive recurrent) delayed renewal sequence.

Theorem

$\frac{1}{n} \sum_{i=1}^n X_i$ converges to $1/\mu$ almost surely, effectively in (Q, R, μ) .

The proof is rather routine from the law of large numbers.

- Let $T^n(x, y) = \mathbf{P} [X_{k+n} = y \mid X_k = x]$ be the n -step transition matrix of a Markov chain X_0, X_1, \dots with a countable state space.
- Let P^x be the conditional distribution when starting from x (this is determined by $T(x, y)$).
- The chain is **irreducible** if all states are mutually **accessible**:

$$\pi_{x,y} = P^x [(\exists n) X_n = y \mid X_0 = x] > 0.$$

- The chain is **aperiodic** if the smallest period of the return time distribution, when starting from any state, is 1.
- If the expected return time m_x of state x is finite then x is called **positive recurrent**.

- Let $T^n(x, y) = \mathbf{P} [X_{k+n} = y \mid X_k = x]$ be the n -step transition matrix of a Markov chain X_0, X_1, \dots with a countable state space.
- Let P^x be the conditional distribution when starting from x (this is determined by $T(x, y)$).
- The chain is **irreducible** if all states are mutually **accessible**:

$$\pi_{x,y} = P^x [(\exists n) X_n = y \mid X_0 = x] > 0.$$

- The chain is **aperiodic** if the smallest period of the return time distribution, when starting from any state, is 1.
- If the expected return time m_x of state x is finite then x is called **positive recurrent**.

- Let $T^n(x, y) = \mathbf{P} [X_{k+n} = y \mid X_k = x]$ be the n -step transition matrix of a Markov chain X_0, X_1, \dots with a countable state space.
- Let P^x be the conditional distribution when starting from x (this is determined by $T(x, y)$).
- The chain is **irreducible** if all states are mutually **accessible**:

$$\pi_{x,y} = P^x [(\exists n) X_n = y \mid X_0 = x] > 0.$$

- The chain is **aperiodic** if the smallest period of the return time distribution, when starting from any state, is 1.
- If the expected return time m_x of state x is finite then x is called **positive recurrent**.

- Let $T^n(x, y) = \mathbf{P} [X_{k+n} = y \mid X_k = x]$ be the n -step transition matrix of a Markov chain X_0, X_1, \dots with a countable state space.
- Let P^x be the conditional distribution when starting from x (this is determined by $T(x, y)$).
- The chain is **irreducible** if all states are mutually **accessible**:

$$\pi_{x,y} = P^x [(\exists n) X_n = y \mid X_0 = x] > 0.$$

- The chain is **aperiodic** if the smallest period of the return time distribution, when starting from any state, is 1.
- If the expected return time m_x of state x is finite then x is called **positive recurrent**.

- Let $T^n(x, y) = \mathbf{P} [X_{k+n} = y \mid X_k = x]$ be the n -step transition matrix of a Markov chain X_0, X_1, \dots with a countable state space.
- Let P^x be the conditional distribution when starting from x (this is determined by $T(x, y)$).
- The chain is **irreducible** if all states are mutually **accessible**:

$$\pi_{x,y} = P^x [(\exists n) X_n = y \mid X_0 = x] > 0.$$

- The chain is **aperiodic** if the smallest period of the return time distribution, when starting from any state, is 1.
- If the expected return time m_x of state x is finite then x is called **positive recurrent**.

Theorem

Assume that the chain with transition matrix T is irreducible, aperiodic and positive recurrent. Then $(p_x = 1/m_x : x \in \mathcal{X})$ is a probability distribution, and

$$\lim_{k \rightarrow \infty} T^k(x, y) = p_y.$$

Computability of $T(x, y)$ does not imply the computability of the equilibrium distribution p_x , as the following example shows. Let $T(0, i) = 2^{-i}$. For $i > 0$ let $T(i, 0) = \alpha_i$, $T(i, i) = 1 - \alpha_i$ for some $0 < \alpha_i < 1$ to be determined. We have the expected return time

$$m_0 = \sum_{i>0} 2^{-i} / \alpha_i.$$

Choose computable α_i while still making m_0 uncomputable, similarly to the example of non-computable expected value.

The sequence Y_n that is 1 if $X_n = y$ and 0 otherwise, is a delayed renewal sequence. By the law of large numbers for these, $\sum_{i=1}^n Y_n/n$ converges to $p_y = 1/m_y$ effectively almost surely relative to $(T(\cdot, \cdot), p_y)$. From this, it is routine to conclude:

Theorem (Computable ergodic theorem)

Let the stationary countable Markov chain X_0, X_1, X_2, \dots be irreducible, aperiodic and positive recurrent, with distribution P (this includes both $T(x, y)$ and p_x). Then for an arbitrary bounded computable function f ,

$$\frac{1}{n} \sum_{i=1}^n f(X_n) \rightarrow \mathbf{E}f(X_0)$$

almost surely, effectively in P .