Introduction to Graduate Probability - MA946 - Appendices

Stefan Adams
2021

A. Modes of Convergence

We shall review in this chapter the basic modes of convergence of random variables. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables taking values in some metric space (E,d), that is, each $X_n\colon \Omega\to E$ is a measurable map between a given probability space $(\Omega,\mathcal{F},\mathbb{P})$ and the range or target space (E,d) where one equips the metric space E with its Borel-σ-field (algebra) $\mathcal{B}(E)$. Let X be a random variable taking values in (E,d).

Definition A.1 (always surely or almost everywhere or with probability 1 or strongly). *The sequence $(X_n)_{n\in\mathbb{N}}$ converges almost surely or almost everywhere or with probability 1 or strongly towards X if*

$$\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) = \mathbb{P}\big(\{\omega\in\Omega\colon \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\big) = 1.$$

This means that the values of X_n approach the value of X, in the sense that events for which X_n does not converge to X have probability 0. We write $X_n \xrightarrow{\text{a.s.}} X$ for almost sure convergence.

Definition A.2 (Convergence in probability). *The sequence $(X_n)_{n\in\mathbb{N}}$ converges in probability to X if*

$$\lim_{n\to\infty}\mathbb{P}\big(d(X_n,X)>\varepsilon\big)=0, \qquad \text{for all } \varepsilon>0.$$

We write $X_n \xrightarrow{\text{P}} X$ for convergence in probability.

Proposition A.3 (Markov's inequality). *Let Y be a real-valued random variable and $f\colon[0,\infty)\to[0,\infty)$ an increasing function. Then, for all $\varepsilon>0$ with $f(\varepsilon)>0$,*

$$\mathbb{P}\big(|Y|\geq\varepsilon\big)\leq\frac{\mathbb{E}\big[f\circ|Y|\big]}{f(\varepsilon)}.$$

Corollary A.4 (Chebyshev's inequality, 1867). *For all $Y\in\mathcal{L}^2$ and $\varepsilon>0$,*

$$\mathbb{P}\big(|Y-\mathbb{E}[Y]|\geq\varepsilon\big)\leq\frac{\mathsf{var}(Y)}{\varepsilon^2}.$$

By Chebyshev's inequality the convergence in probability is equivalent to $\mathbb{E}[d(X_n,X)\wedge 1]\to 0$ as $n\to\infty$. This is related to the almost sure convergence as follows.

Lemma A.5 (Subsequence criterion). *Let $X, X_1, X_2, \dots$ be random variables in (E,d). Then $(X_n)_{n\in\mathbb{N}}$ converges to X in probability if and only if every subsequence $N'\subset\mathbb{N}$ has a further subsequence $N''\subset\mathbb{N}'$ such that $X_n\to X$ almost surely along N''. In particular, $X_n\xrightarrow{\text{a.s.}}X$ implies that $(X_n)_{n\in\mathbb{N}}$ converges to X in probability.*

Definition A.6 (Convergence in distribution). *We say that X_n converges in distribution to X, if, for every bounded continuous function $f\colon E\to\mathbb{R}$,*

$$\lim_{n\to\infty}\mathbb{E}[f(X_n)]=\mathbb{E}[f].$$

We write $X_n\xrightarrow{\text{d}}X$ for convergence in distribution.

Remark A.7. *(a) $X_n\xrightarrow{\text{d}}X$ is equivalent to weak convergence of the distributions.*

(b) if $X_n \xrightarrow{\mathrm{d}} X$ and $g\colon E \to \mathbb{R}$ continuous, then $g(X_n) \xrightarrow{\mathrm{d}} g(X)$. But note that, if $E = \mathbb{R}$ and $X_n \xrightarrow{\mathrm{d}} X$, this does not imply that $\mathbb{E}[X_n]$ converges to $\mathbb{E}[X]$, as $g(x) = x$ is not a bounded function on $\mathbb{R}$.

(c) Suppose $E = \{1, \dots, m\}$ is finite and $d(x, y) = 1 - \mathbb{1}_{x=y}$. Then $X_n \xrightarrow{\mathrm{d}} X$ if and only if $\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$ for all $k \in E$.

(d) Let $E = [0, 1]$ and $X_n = 1/n$ almost surely. Then $X_n \xrightarrow{\mathrm{d}} X$, where $X = 0$ almost surely. However, note that $\lim_{n\to\infty} \mathbb{P}(X_n = 0) = 0 \neq \mathbb{P}(X = 0)$.

◇

B. Law of large numbers and the central limit theorem

Definition B.1 (Variance and covariance). *Let $X, Y \in \mathcal{L}^2$ be real-valued random variables.*

(a)

$$\mathsf{var}(X) := \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)^2\big] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

*is called the **variance**, and $\sqrt{\mathsf{var}(X)}$ the **standard deviation** of X with respect to $\mathbb{P}$.*

(b)

$$\mathsf{Cov}(X, Y) := \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

*is called the **covariance** of X and Y. It exists since $|XY| \leq X^2 + Y^2$.*

*(c) If $\mathsf{Cov}(X, Y) = 0$, then X and Y are called **uncorrelated**.*

Theorem B.2 (Weak law of large numbers, $\mathcal{L}^2$-version). *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of uncorrelated (e.g. independent) real-valued random variables in $\mathcal{L}^2$ with bounded variance, in that $v := \sup_{n\in\mathbb{N}} \mathsf{var}(X_n) < \infty$. Then for all $\varepsilon > 0$*

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathbb{E}[X_i]\big)\Big| \geq \varepsilon\Big) \leq \frac{v}{n\varepsilon^2} \underset{n\to\infty}{\longrightarrow} 0,$$

and thus $1/n \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \xrightarrow{\mathrm{P}} 0$. In particular, if $\mathbb{E}[X_i] = \mathbb{E}[X_1]$ for all $i \in \mathbb{N}$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{P}} \mathbb{E}[X_1].$$

We now present a second version of the weak law of large numbers, which does not require the existence of the variance. To compensate we must assume that the random variables, instead of being pairwise uncorrelated, are even pairwise independent and identically distributed.

Theorem B.3 (Weak law of large numbers, $\mathcal{L}^1$-version). *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of pairwise independent, identically distributed real-valued random variables in $\mathcal{L}^1$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{P}} \mathbb{E}[X_1].$$

Theorem B.4 (Strong law of large numbers). *If $(X_n)_{n\in\mathbb{N}}$ is a sequence of pairwise uncorrelated real-valued random variables in $\mathcal{L}^2$ with $v := \sup_{n\in\mathbb{N}} \mathsf{var}(X_n) < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathbb{E}\big[X_i\big]\big) \to 0 \text{ almost surely as } n \to \infty.$$

Theorem B.5 (Central limit theorem; A.M. Lyapunov 1901, J.W. Lindeberg 1922,P. Leévy 1922). *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed real-valued random variables in $\mathcal{L}^2$ with $\mathbb{E}[X_i] = m$ and $\mathsf{var}(X_i) = v > 0$. Then,*

$$S_n^* := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - m}{\sqrt{v}} \xrightarrow{\mathrm{d}} \mathsf{N}(0,1).$$

The normal distribution is defined in the following section.

C. Normal distribution

A real-valued random variable X is **normally** distributed with mean μ and variance $\sigma^2 > 0$ if

$$\mathbb{P}(X > x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_x^\infty \mathrm{e}^{-\frac{(u-\mu)^2}{2\sigma^2}} \,\mathrm{d}u, \qquad \text{for all } x \in \mathbb{R}.$$

We write $X \sim \mathsf{N}(\mu, \sigma^2)$. We say that X is standard normal distributed if $X \sim \mathsf{N}(0,1)$.

A random vector $X = (X_1, \ldots, X_n)$ is called a **Gaussian random vector** if there exits an $n \times m$ matrix A, and an n-dimensional vector $b \in \mathbb{R}^n$ such that $X^T = AY + b$, where Y is an m-dimensional vector with independent standard normal entries, i.e. $Y_i \sim \mathsf{N}(0,1)$ for $i = 1, \ldots, m$. Likewise, a random variable $Y = (Y_1, \ldots, Y_m)$ with values in $\mathbb{R}^m$ has the m-dimensional standard Gaussian distribution if the m coordinates are standard normally distributed and independent. The covariance matrix of $X = AY + b$ is then given by

$$\mathsf{Cov}(Y) = \mathbb{E}\big[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^T\big] = AA^T.$$

Lemma C.1. *If A is an orthogonal $n \times n$ matrix, i.e. $AA^T = \mathbb{1}$, and X is a n-dimensional standard Gaussian vector, then AX is also a n-dimensional standard Gaussian vector.*

Lemma C.2. *Let X_1 and X_2 be independent and normally distributed with zero mean and variance $\sigma^2 > 0$. Then $X_1 + X_2$ and $X_1 - X_2$ are independent and normally distributed with mean 0 and variance $2\sigma^2$.*

Proposition C.3. *If X and Y are n-dimensional Gaussian vectors with $\mathbb{E}\big[X\big] = \mathbb{E}\big[Y\big]$ and $\mathsf{Cov}(X) = \mathsf{Cov}(Y)$, then X and Y have the same distribution.*

Corollary C.4. *A Gaussian random vector X has independent entries if and only if its covariance matrix is diagonal. In other words, the entries in a Gaussian vector are uncorrelated if and only if they are independent.*

Lemma C.5 (Inequalities). *Let $X \sim \mathsf{N}(0,1)$. Then for all $x > 0$,*

$$\frac{x}{x^2+1} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \leq \mathbb{P}(X > x) \leq \frac{1}{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\,.$$

D. Gaussian integration formulae

For any $a > 0$,

$$\int_{-\infty}^{\infty} \mathrm{e}^{-ax^2} \,\mathrm{d}x = \sqrt{\pi/a}.$$

For $b \in \mathbb{C}$ and $a > 0$,

$$I(b) = \int_{-\infty}^{\infty} \mathrm{e}^{-a/2x^2 + bx} \,\mathrm{d}x = \mathrm{e}^{b^2/2a} \sqrt{2\pi/a}\,.$$

Let $A \in \mathbb{R}^{n\times n}$, $A = A^T > 0$ (i.e. all eigenvalues of A are positive), and define $C = A^{-1}$ and write $\langle \phi, \psi \rangle$ for the scalar product of $\phi, \psi \in \mathbb{R}^n$.

$$\int_{\mathbb{R}^n} \mathrm{e}^{-\frac{1}{2}\langle \phi, A\phi\rangle} \prod_{i=1}^n \mathrm{d}\phi_i = (2\pi)^{n/2} \det(A^{-\frac{1}{2}}) = \det(2\pi C)^{\frac{1}{2}}.$$

For any $J \in \mathbb{C}^n$ we obtain

$$\int_{\mathbb{R}^n} \mathrm{e}^{-\frac{1}{2}\langle \phi, A\phi\rangle + \langle J, \phi\rangle} \prod_{i=1}^n \mathrm{d}\phi_i = \det(2\pi C)^{\frac{1}{2}} \mathrm{e}^{\frac{1}{2}\langle J, CJ\rangle}\,.$$

Let $C \in \mathbb{R}^{n\times n}$ be invertible matrix and $C > 0$. The probability measure $\mu_C \in \mathcal{M}_1(\mathbb{R}^n)$ defined by

$$\mu_C(\mathrm{d}\phi) = \frac{1}{\sqrt{\det(2\pi C)}} \mathrm{e}^{-1/2\langle \phi, C^{-1}\phi\rangle} \prod_{i=1}^n \mathrm{d}\phi_i,$$

is called the **Gaussian measure** on $\mathbb{R}^n$ with mean zero and covariance matrix C.

The covariance splitting formula. Let $C_i = C_i^T, i = 1, 2$, be positive invertible matrices. Define $C = C_1 + C_2$. Then for all $F \in \mathcal{L}(\mu_C)$,

$$\begin{aligned}\int_{\mathbb{R}^n} F(\phi)\mu_C(\mathrm{d}\phi) &= \int_{\mathbb{R}^n} \mu_{C_1}(\mathrm{d}\phi_1) \int_{\mathbb{R}^n} \mu_{C_2}(\mathrm{d}\phi_2) F(\phi_1 + \phi_2) \\ &= \int_{\mathbb{R}^n} \mu_{C_1}(\mathrm{d}\phi) \int_{\mathbb{R}^n} \mu_{C_2}(\mathrm{d}(\phi - \phi_1)) F(\phi).\end{aligned}$$

In other words, if $C = C_1 + C_2$, the Gaussian random variable ϕ is the sum of two independent (see above) Gaussian random variables, $\phi = \phi_1 + \phi_2$, and the Gaussian measure factors, i.e. $\mu_C = \mu_{C_1} \otimes \mu_{C_2}$.

The characteristic function of a Gaussian vector $X = (X_1, \ldots, X_n)$ with mean $\mu \in \mathbb{R}^n$ and covariance matrix C reads as

$$\varphi_X(t) = \mathbb{E}\Big[\mathrm{e}^{i\langle t, \mu\rangle - \frac{1}{2}\langle t, Ct\rangle}\Big], \quad t \in \mathbb{R}^n.$$

An $\mathbb{R}^n$-valued stochastic process $X = \{X_t \colon t \geq 0\}$ is called **Gaussian** if, for any integer $k \geq 1$ and real numbers $0 \leq t_1 < t_2 < \cdots < t_k < \infty$, the random vector $(X_{t_1}, \ldots, X_{t_k})$ has a joint normal distribution. If the distribution of $(X_{t+t_1}, \ldots, X_{t+t_n})$ does not depend on t, we say that the process is stationary. The finite-dimensional distributions of a Gaussian process X are determined by its expectation vector $m(t) := \mathbb{E}\big[X(t)\big], t \geq 0$, and its covariance matrix

$$\rho(s,t) := \mathbb{E}\big[(X_s - m(s))(X_t - m(t))^T\big], \qquad s, t \geq 0.$$

If $m(t) = 0$ for all $t \geq 0$, we say that X is a zero-mean Gaussian process.

Corollary D.1. *One-dimensional BM is a zero-mean Gaussian process with covariance formula*

$$\rho(s,t) = s \wedge t, \qquad s, t \geq 0.$$

E. Some useful properties of the weak topology of probability measures

A probability measure $\mu \in \mathcal{M}_1(E)$ on a metric space (E,d) is *tight* if for each $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset E$ such that $\mu(K_\varepsilon^{\mathrm{c}}) < \varepsilon$. A family $(\mu_n)_{n\in I}$ of probability measures on the metric space (E,d) is called a tight family if the set K_ε may be chosen independently of $n \in I$, that is, for all $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset E$ and $n_0 \in I$ such that $\mu_n(K_\varepsilon^{\mathrm{c}}) < \varepsilon$ for all $n \geq n_0$.

Definition E.1 (Weak convergence of probability measures). *A sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability measures on a metric space (E,d) converges weakly to $\mu \in \mathcal{M}_1(E)$ as $n \to \infty$ if*

$$\int_E f(x)\,\mu_n(\mathrm{d}x) \to \int_E f(x)\,\mu(\mathrm{d}x) \quad \textit{for all } f \in \mathcal{C}_{\mathrm{b}}(E) \textit{ as } n \to \infty.$$

Lemma E.2. *A sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability measures on a metric space (E,d) converges weakly to $\mu \in \mathcal{M}_1(E)$ as $n \to \infty$ if*

$$\begin{aligned} &\limsup_{n\to\infty} \mu_n(C) \leq \mu(C) \quad \textit{for all closed } C \subset E, \\ &\liminf_{n\to\infty} \mu_n(O) \geq \mu(O) \quad \textit{for all open } O \subset E. \end{aligned} \tag{E.1}$$

The set of probability measures $\mathcal{M}_1(E)$ on a Polish space (E,d) is itself a Polish space. Note that $\mathcal{M}_1(E) \subset \mathcal{M}(E)$ is a closed convex subset of the (vector) - space of all finite signed measures on E. We equip $\mathcal{M}(E)$ with the topology generated by sets

$$\Big\{\beta \in \mathcal{M}(E)\colon \Big|\int_E f(x)\,\mathrm{d}(\beta(x)-\alpha(x))\Big| < r\Big\},$$

where $\alpha \in \mathcal{M}(E)$, $f \in \mathcal{C}_{\mathrm{b}}(E)$, and $r > 0$. The norm on $\mathcal{M}(E)$ is the total variation norm

$$\|\alpha\|_{\mathrm{var}} := \sup\Big\{\int_E f(x)\alpha(\mathrm{d}x)\colon f \in \mathcal{C}_{\mathrm{b}}(E) \text{ with } \|f\|_\infty \leq 1\Big\}, \quad \alpha \in \mathcal{M}(E).$$

The norm $\|\cdot\|_{\mathrm{var}}$ is lower semi-continuous on $\mathcal{M}(E)$ and therefore certainly measurable on $\mathcal{M}(E)$; and clearly, $\|\cdot\|_{\mathrm{var}}$ is bounded on $\mathcal{M}_1(E)$. The Lévy metric on $\mathcal{M}_1(E)$ is a complte separable metric, which is consistent (inherited from) with the restriction of the topology on $\mathcal{M}(E)$ to the closed and convex subset $\mathcal{M}_1(E)$. Following Lévy and Prohorov, define the Lévy metric as

$$\mathrm{d}(\alpha,\nu) := \inf\Big\{\delta > 0\colon \alpha(F) \leq \nu(F^{(\delta)}) + \delta \text{ and } \nu(F) \leq \alpha(F^{(\delta)}) + \delta \text{ for all closed } F \subset E\Big\},$$

$\alpha,\nu \in \mathcal{M}_1(E)$, where $F^{(\delta)}$ is defined relative to a complete metric on E, that is, $F^{(\delta)}$ is the open δ-hull of F. Since it is clear that $\mathrm{d}(\alpha,\nu) \leq \|\alpha-\nu\|_{\mathrm{var}}$, all that remains to show is that the Lévy metric d is compatible with the weak topology in Definition E.1 and Lemma E.2 and that $(\mathcal{M}_1(E),\mathrm{d})$ is a Polish space. To show this one uses the tightness criterion, Lemma E.2 (the upper bound), and the following: Suppose that $\mathcal{F} \subset \mathcal{C}_{\mathrm{b}}(E)$ is a set of uniformly bounded test functions which is equicontinuous on every compact subset of E. Then the weak convergence $\alpha_n \Rightarrow \nu$ implies that

$$\sup\Big\{\Big|\int f(x)\alpha_n(\mathrm{d}x) - \int f(x)\,\nu(\mathrm{d}x)\Big|\colon f \in \mathcal{F}\Big\} \to 0 \text{ as } n \to \infty.$$

This is the content of the following lemma which is proved in the book by Billingsley on Convergence of probability measures.

Lemma E.3 (Lévy & Prohorov). *The Lévy metric d (defined above) is compatible with the weak topology on $\mathcal{M}_1(E)$, and $(\mathcal{M}_1(E),\mathrm{d})$ is a Polish space.*

We will frequently use the following dual space for $\mathcal{M}(E)$ (note that $\mathcal{M}_1(E)$ is not a vector space).

Lemma E.4. *The duality relation*

$$(f,\nu) \in \mathcal{C}_{\mathrm{b}}(E) \times \mathcal{M}(E) \mapsto \int_E f(x)\,\nu(\mathrm{d}x)$$

determines a representation of $\mathcal{M}^*$ as $\mathcal{C}_{\mathrm{b}}(E)$.

Theorem E.5 (Prohorov). *Let (E,d) be a Polish space, and let $\Gamma \subset \mathcal{M}_1(E)$. The $\overline{\Gamma}$ is compact iff Γ is tight.*

We shall need some version for the path space $E := \mathcal{C}([0,1];\mathbb{R})$.

Proposition E.6. *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of probability measures on $\mathcal{C}([0,1];\mathbb{R})$ which converges weakly to μ. Let A be a Borel set in $\mathcal{C}([0,1];\mathbb{R})$ with $\mu(\partial A) = 0$. Then $\mu_n(A) \to \mu(A)$ as $\nu \to \infty$.*

We need an adaptation of Prohorov's theorem suited to the path space $\mathcal{C}([0,1];\mathbb{R})$,

Theorem E.7. *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of Borel probability measures on $\mathcal{C}([0,1];\mathbb{R})$ with the following two properties:*

(a) The finite dimensional distributions converge. That is, for any $0 \le t_1 < t_2 < \cdots < t_m \le 1$, $m \in \mathbb{N}$, there is a measure $\mu_{t^{(m)}} \in \mathcal{M}_1(\mathbb{R}^m)$ so that, as $n \to \infty$,

$$\int f(\omega(t_1),\ldots,\omega(t_m))\,\mu_n(\mathrm{d}\omega) \to \int f(x_1,\ldots,x_m)\,\mu_{t^{(m)}}(\mathrm{d}x) \text{ for all } f \in \mathcal{C}_{\mathrm{b}}(\mathbb{R}^m).$$

(b) $(\mu_n)_{n\in\mathbb{N}}$ is tight.

Then, there is a probability measure μ on $\mathcal{C}([0,1];\mathbb{R})$ so that $\mu_n \to \mu$ weakly as $n \to \infty$, and the finite dimensional distributions of μ are the $\mu_{t^{(m)}}$.