

Lecture 3

Wednesday, January 7, 2026

12:39 PM

modes of convergence:

Definition 1.5.1. Let X be a random variable and $\{X_n\}$ be a sequence of random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. $X_n \rightarrow X$ a.s. if $\mathbb{P}(\{\omega \in \Omega : X_n(\omega) \rightarrow X(\omega)\}) = 1$ (why is this event in $\mathcal{F}$?).
2. $X_n \rightarrow X$ in probability if $\mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$ for every $\varepsilon > 0$.
3. $X_n \rightarrow X$ in $\mathcal{L}^p$ if $\|X_n - X\|_p \rightarrow 0$ as $n \rightarrow \infty$.
4. $X_n \rightarrow X$ in law (or in distribution, or weakly) if $\mathbb{E}(f(X_n)) \rightarrow \mathbb{E}(f(X))$ for every bounded continuous function f .

Proposition 1.5.2. Let X be a random variable and $\{X_n\}$ be a sequence of random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The following implications hold:

1. $X_n \rightarrow X$ a.s. $\implies X_n \rightarrow X$ in probability.
 2. $X_n \rightarrow X$ in $\mathcal{L}^p \implies X_n \rightarrow X$ in probability.
 3. $X_n \rightarrow X$ in $\mathcal{L}^p \implies X_n \rightarrow X$ in $\mathcal{L}^q$ ($q \leq p$).
 4. $X_n \rightarrow X$ in probability $\implies X_n \rightarrow X$ in law.
- in general!

Proof: 1) $\limsup \mathbb{P}(|X_n - X| > \varepsilon) \leq \mathbb{P}(\limsup \{|X_n - X| > \varepsilon\})$

$$= \mathbb{P}(|X_n - X| > \varepsilon, \text{ i.o.}) = 0.$$

2) By Chebyshev's ineq. $\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p} \rightarrow 0.$

3) $x \mapsto x^{p/q}$ is convex, so by Jensen's inequality:

$$(\mathbb{E}[|X_n - X|^q])^{p/q} \leq \mathbb{E}[|X_n - X|^p] \rightarrow 0.$$

$$\begin{aligned} 4) \quad |\mathbb{E}[f(X_n) - f(X)]| &\leq \mathbb{E}|f(X_n) - f(X)| \\ &\leq \mathbb{E}\left[\mathbb{1}_{\{|f(X_n) - f(X)| > \varepsilon\}} |f(X_n) - f(X)|\right] \\ &\quad + \mathbb{E}\left[\mathbb{1}_{\{|f(X_n) - f(X)| < \varepsilon\}} |f(X_n) - f(X)|\right] \end{aligned}$$

(by cont. of f , $\exists \delta > 0$) $\Rightarrow \leq 2\|f\|_\infty \cdot \mathbb{P}(|X_n - X| > \delta) + \varepsilon, \quad \forall \varepsilon > 0.$

$$\rightarrow 0 \text{ as } n \rightarrow \infty.$$

so $\limsup |\mathbb{E}f(X_n) - \mathbb{E}f(X)| \leq \varepsilon, \quad \forall \varepsilon \Rightarrow X_n \rightarrow X$ in law



Important non-examples:

(E.g.) convergence in law $\not\Rightarrow$ convergence in probability
 $\mathbb{P}(\{a_1\} = \{a_2\}) \rightarrow \text{fair coin}$

$$\Omega = \{a_1, a_2, \dots\}$$

$$X: \Omega \rightarrow \{\pm 1\}: X(a_1) = -X(a_2) = 1$$

$$X_n \triangleq (-1)^n X$$

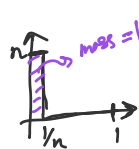
very weak convergence!

then $X_n \rightarrow X$ in law (distribution) but not in any other form! in particular, the sample paths $n \mapsto X_n(\omega)$ for a fixed ω is just fluctuating!

$$\begin{aligned} \text{why?} \\ E(f(X_n)) &= E(f((-1)^n X)) = \int f((-1)^n x) P(d\omega) \\ &= f((-1)^n) \cdot \frac{1}{2} + f((-1)^n (-1)) \cdot \frac{1}{2} \\ &= \frac{1}{2} (f((-1)^n) + f((-1)^{n+1})) \\ &\stackrel{\text{why?}}{=} \frac{1}{2} (f(1) + f(-1)) \end{aligned}$$

consider $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}([0, 1])$, P is uniform meas. on $[0, 1]$.

(E.2) Convergence a.s. but not in L^p :



$$X_n(\omega) = n I_{(0, 1/n]}(\omega), \quad \omega \in [0, 1]$$

$X_n \rightarrow 0$ a.s. but $\|X_n\|_1 = 1 \Rightarrow X_n \not\rightarrow 0$ in L^1
 $\Rightarrow X_n \not\rightarrow 0$ in any $L^p, p \geq 1$.

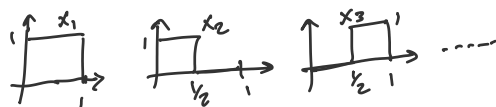
(E.3) ($q < p$) Convergence in L^q but not in L^p :

$$X_n(\omega) = n^{1/p} I_{(0, 1/n]}(\omega)$$

$X_n \rightarrow 0$ in $L^q, q < p$ but not for any $q \geq p$!

(E.4) Convergence in L^p but not a.s.: "Dancing sequence" or "Typewriter seq."

$\{X_n\}_{n=1}^\infty \stackrel{\text{A}}{=} \left\{ \begin{array}{l} \text{For each } n=1, 2, 3, \dots \\ \rightarrow \text{divide } [0, 1] \text{ to } 2^{n-1} \text{ equal intervals, } U_k^n, k=1, 2, \dots, 2^{n-1} \\ \rightarrow \text{sequentially from left to right, set} \\ X_{nk}(\omega) = I_{U_k^n}(\omega) = \begin{cases} +1 & \text{if } \omega \in U_k^n \\ 0 & \text{o.w.} \end{cases} \end{array} \right.$



$$X_n \rightarrow 0 \text{ in } L^p \text{ for any } p \geq 1. \quad \left(E(|X_n|^p) = \frac{1}{2^{n-1}} \rightarrow 0 \text{ as } n \rightarrow \infty \right)$$

But, fix any $\omega \in [0, 1]$, the sequence $X_n(\omega)$ always oscillates between 0, 1 $\Rightarrow X_n(\omega) \not\rightarrow 0$ for any $\omega \in [0, 1]$.

random -

$$\Rightarrow X_n(\omega) \not\rightarrow 0 \text{ a.s.}$$

In (E.4), we learned $X_n \rightarrow X$ in prob. $\not\Rightarrow X_n \rightarrow X$ a.s.
because possible irregularity like the "dancing seq."

Question: Is there an additional regularity assumption s.t.
convergence in prob $\Rightarrow$ conv. a.s. ?!

Lemma: if $X_n \rightarrow X$ in prob. $P(|X_n - X| > \varepsilon) \rightarrow 0, \forall \varepsilon > 0.$
and
 $\sum_n P(|X_n - X| > \varepsilon) < \infty, \forall \varepsilon > 0.$
 $\Rightarrow X_n \rightarrow X$ a.s.

Proof: By Borell-Cantelli, $P(|X_n - X| > \varepsilon, \text{i.o.}) = 0, \forall \varepsilon > 0.$
Thus, $P(\text{fix any } k, |X_n - X| > 1/k \text{ for finite \# of } n\text{'s}) = 0.$
So, for any fixed $k, \exists N$ s.t. $\forall n \geq N, |X_n(\omega) - X(\omega)| \leq 1/k$
for a.e. $\omega \in \Omega.$

Corollary: if $X_n \rightarrow X$ in prob., then there exist a
subsequence $X_{n_k} \rightarrow X$ a.s.

Proof: $P(|X_n - X| > \varepsilon) \rightarrow 0$; choose the subsequence $n_k \uparrow \infty$
such that $P(|X_{n_k} - X| > \varepsilon) \leq \frac{1}{2^k}$. Apply Lemma above. $\square$

In (E.2), we learned that (by $\prod_{n=1}^{\infty} \frac{X_n}{Y_n}$)
 $X_n \rightarrow X$ a.s. $\not\Rightarrow X_n \rightarrow X$ in L^p .

this also automatically means

$X_n \rightarrow X$ in prob $\not\Rightarrow X_n \rightarrow X$ in L^p .

Question: Is there an additional regularity that allows this ?!

(DET) **Theorem 1.5.12 (Dominated convergence)**. Let $X_n \rightarrow X$ in probability, and suppose there exists a nonnegative $Y \in L^p$, with $p \geq 1$, such that $|X_n| \leq Y$ a.s. for all n . Then X and X_n are in L^p for all n , $X_n \rightarrow X$ in L^p , and $E(X_n) \rightarrow E(X)$.

Proof: needs Monotone Convergence Theorem and Fatou's Lemma.
see page 36, in van Handel's note.

Theorem 1.5.10 (Monotone convergence). Let $\{X_n\}$ be a sequence of random variables such that $0 \leq X_1 \leq X_2 \leq \dots$ a.s. Then $\mathbb{E}(X_n) \nearrow \mathbb{E}(\lim_{n \rightarrow \infty} X_n)$.

↳ pointwise!

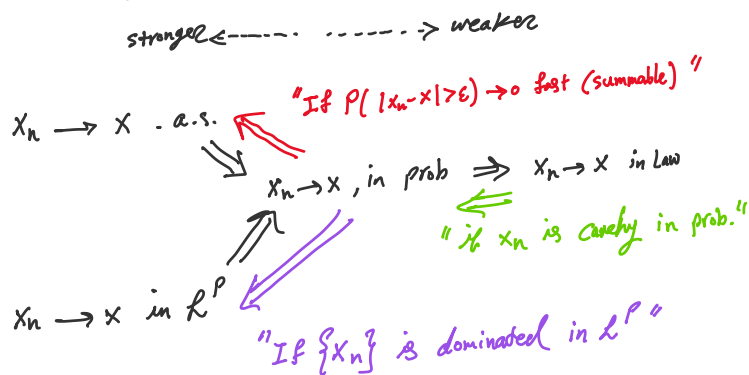
Lemma 1.5.11 (Fatou's lemma). Let $\{X_n\}$ be a sequence of a.s. nonnegative random variables. Then $\mathbb{E}(\liminf X_n) \leq \liminf \mathbb{E}(X_n)$. In there exists a Y such that $X_n \leq Y$ a.s. for all n and $\mathbb{E}(Y) < \infty$, then $\mathbb{E}(\limsup X_n) \geq \limsup \mathbb{E}(X_n)$.

But, the usefulness of DCT is that often $\{X_n\}$ is a seq. of bounded r.v.; that, $X_n \rightarrow X$ in prob $\Rightarrow \begin{cases} X_n \rightarrow X \text{ in any } L^p \\ \mathbb{E}[X_n] \rightarrow \mathbb{E}[X] \end{cases}$

we like L^p convergence because, it is strong enough and also is a "complete space"; i.e.
 $\{X_n\}$ is Cauchy in L^p ($\|X_n - X_m\|_{L^p} \rightarrow 0$ as $n, m \rightarrow \infty$)
 Then, $\exists X \in L^p$ s.t. $X_n \rightarrow X$, in L^p .

* the strong convergence enables us to argue about sample paths
 $t \mapsto X_t(\omega)$ for each fixed $\omega \in \Omega$.

Summary of convergence modes:



Induced measures, laws, and independenc:

Definition 1.6.1. If $h : \Omega \rightarrow S$ is measurable, then the map $\mu : S \rightarrow [0, 1]$ defined by $\mu(A) = \mathbb{P}(h^{-1}(A))$ is a probability measure (why?), called the *induced measure*.

If h is a r.v. then μ_h on $\mathcal{B}$ is called the law of h .

So, two different r.v.s can have the same law, for example

$$X : \Omega \rightarrow \mathbb{R} \quad X(\omega) = \omega$$

$X: (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{N}(\text{Lebesgue})) \rightarrow \mathbb{R}$, $Y(\omega) = \Phi^{-1}(\omega)$ where
 Φ is the CDF of $\mathcal{N}(0,1)$.

$\Rightarrow$ Law of X = Law of Y (why?).

- This is the reason when we say X is a r.v. with certain law we may not specify what is the domain probability space!

Definition 1.6.2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

1. A countable set of events $\{A_n\}_{n \in I} \subset \mathcal{F}$ are independent if

$$\mathbb{P}(A_{k_1} \cap A_{k_2} \cap \dots \cap A_{k_n}) = \mathbb{P}(A_{k_1}) \mathbb{P}(A_{k_2}) \dots \mathbb{P}(A_{k_n}),$$

for all finite ($n < \infty$) subsets $\{k_1, \dots, k_n\}$ of the index set I .

2. A countable set $\{\mathcal{G}_n\}_{n \in I}$ of σ -algebras $\mathcal{G}_n \subset \mathcal{F}$ are independent if any finite set of events $A_1, \dots, A_n$ from distinct $\mathcal{G}_k$ are independent.

3. A countable set $\{X_n\}_{n \in I}$ of random variables are independent if the σ -algebras generated by these random variables $\mathcal{X}_n = \sigma(X_n)$ are independent.

Conditional Expectation: $(\Omega, \mathcal{F}, \mathbb{P})$
 Recall Cond. Probability $A, B \in \mathcal{F}$

$$\mathbb{P}(A|B) := \begin{cases} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} & \text{if } \mathbb{P}(B) \neq 0 \\ 0 & \text{o.w.} \end{cases}$$

- if X is a r.v. then Cond. Expect. is

$$\mathbb{E}(X|B) := \mathbb{E}(X \mathbb{I}_B) / \mathbb{P}(B)$$

now, if Y is another r.v. and $B = \{Y = y_1\}$ then.

$$\mathbb{E}(X|B) = \mathbb{E}(X|Y=y_1) = f(y_1) \text{ a function of } y_1!$$

So, it make sense to define $\mathbb{E}(X|Y)$ as another r.v.

$$\text{s.t. } \mathbb{E}(X|Y) = f(Y) \text{ where } f(y) \triangleq \mathbb{E}(X|Y=y)!$$

Intuitively, $\mathbb{E}(X|Y)$ is a "good estimate" of X if Y is observed.

Definition 2.1.2. Let $X \in \mathcal{L}^1$ be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subset \mathcal{F}$ be a finite σ -algebra generated by the partition $\{A_k\}_{k=1, \dots, n}$. Then

$$\mathbb{E}(X|\mathcal{G}) \equiv \sum_{k=1}^n \mathbb{E}(X|A_k) I_{A_k},$$

where $\mathbb{E}(X|A) = \mathbb{E}(X I_A) / \mathbb{P}(A)$ if $\mathbb{P}(A) > 0$ and we may define $\mathbb{E}(X|A)$ arbitrarily if $\mathbb{P}(A) = 0$. The conditional expectation thus defined is unique up to a.s. equivalence, i.e., any two random variables $Y, \tilde{Y}$ that satisfy the definition obey $Y = \tilde{Y}$ a.s. (why?).

Remark 2.1.3. $X \in \mathcal{L}^1$ ensures that $\mathbb{E}(X|A_k)$ is well defined and finite.

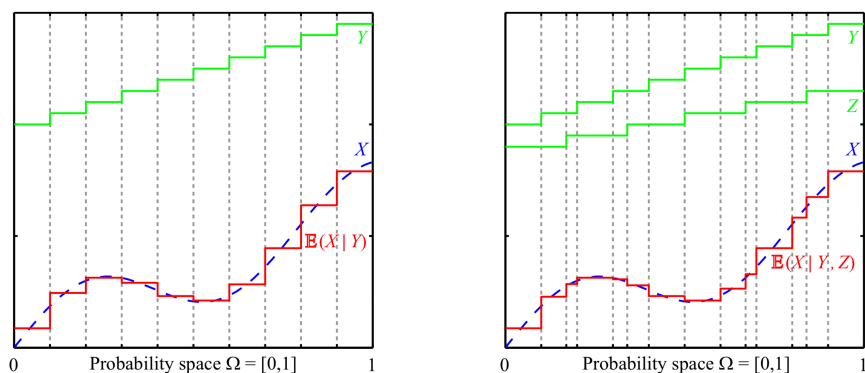


Figure 2.1. Illustration of the discrete conditional expectation on $([0, 1], \mathcal{B}([0, 1]), \mathbb{P})$, where $\mathbb{P}$ is the uniform measure. A random variable X is conditioned with respect to a discrete random variable Y (left) and with respect to two discrete random variables Y, Z (right). This amounts to averaging X (w.r.t. $\mathbb{P}$) over each bin in the partition generated by Y and Y, Z , respectively.

(*) As you see, the $\mathbb{E}(X|Y)$ doesn't depend on the realized value of Y , rather depend on "the partitions" generated by this realization of Y .

now, a more complicated case as a limit of simple functions:

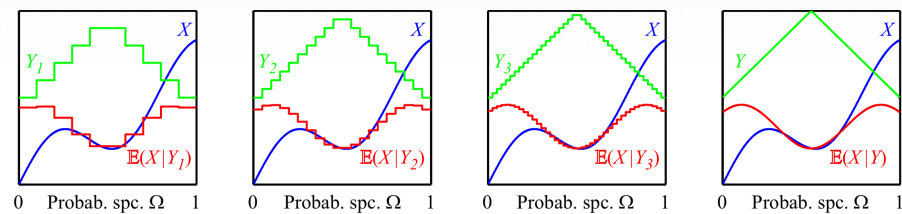


Figure 2.2. We would like to calculate $\mathbb{E}(X|Y)$, where $\mathbb{P}$ is the uniform measure on the interval, $X : [0, 1] \rightarrow \mathbb{R}$ is some random variable and $Y : [0, 1] \rightarrow \mathbb{R}$ is given by $Y(x) = x$ for $x \leq \frac{1}{2}$, $Y(x) = 1 - x$ for $x \geq \frac{1}{2}$. Intuitively we should have $\mathbb{E}(X|Y)(x) = \frac{1}{2}X(x) + \frac{1}{2}X(1 - x)$, but definition 2.1.2 does not allow us to conclude this. However, a sequence of approximations covered by definition 2.1.2 appears to give rise to the expected result. But how to prove it?

A natural way to define this limit is as follows:
 Find a seq. $Y_n \rightarrow Y$ a.s. and define $M_n = \mathbb{E}[X|Y_n]$
 then show that $M_n \rightarrow M := \mathbb{E}[X|Y]$ a.s.
 → showing this convergence requires "martingale convergence theorem"
 that we skip for now in favor of synching with lectures.
 → Instead, we provide the "general definition of $\mathbb{E}[X|Y]$ ":

Definition 2.3.1 (Kolmogorov). Let $X \in \mathcal{L}^1$ and let $\mathcal{F}$ be any σ -algebra. Then $\mathbb{E}(X|\mathcal{F})$ is, by definition, the unique $\mathcal{F}$ -measurable random variable that satisfies the relation $\mathbb{E}(I_A X) = \mathbb{E}(I_A \mathbb{E}(X|\mathcal{F}))$ for all events $A \in \mathcal{F}$.

This is usually taken as the starting point in developing conditional expectations, but with your knowledge of martingales it should now be evident that this is just the limiting case of the familiar discrete conditional expectation, but in disguise. On the other hand, this definition is often much easier to deal with: the definition itself does not involve taking any limits (the limits are only involved in proving *existence* of an object that satisfies the definition!)

Lemma 2.1.4. Let $X, Y \in \mathcal{L}^1$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, let $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ be finite σ -algebras, and let $\alpha, \beta \in \mathbb{R}$.

1. $\mathbb{E}(\alpha X + \beta Y | \mathcal{G}) = \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y)$ a.s.
2. If $X \geq 0$ a.s., then $\mathbb{E}(X | \mathcal{G}) \geq 0$ a.s.

2.1. Conditional expectations and martingales: a trial run

49

3. $\mathbb{E}(\mathbb{E}(X | \mathcal{G})) = \mathbb{E}(X)$.
4. Tower property: if $\mathcal{H} \subset \mathcal{G}$, then $\mathbb{E}(\mathbb{E}(X | \mathcal{G}) | \mathcal{H}) = \mathbb{E}(X | \mathcal{H})$ a.s.
5. If X is $\mathcal{G}$ -measurable, then $\mathbb{E}(X | \mathcal{G}) = X$ a.s.
6. If X is $\mathcal{G}$ -measurable and $XY \in \mathcal{L}^1$, then $\mathbb{E}(XY | \mathcal{G}) = X \mathbb{E}(Y | \mathcal{G})$ a.s.
7. If $\mathcal{H}$ and $\sigma\{X, \mathcal{G}\}$ are independent, then $\mathbb{E}(X | \sigma\{\mathcal{G}, \mathcal{H}\}) = \mathbb{E}(X | \mathcal{G})$ a.s.
8. If $\mathcal{H}$ and X are independent, then $\mathbb{E}(X | \mathcal{H}) = \mathbb{E}(X)$ a.s.
9. Monotone and dominated convergence, Fatou's lemma, and Jensen's inequality all hold for conditional expectations also; e.g., if $0 \leq X_1 \leq X_2 \leq \dots$ a.s., then $\mathbb{E}(X_n | \mathcal{G}) \nearrow \mathbb{E}(\lim_n X_n | \mathcal{G})$ a.s. (monotone convergence).

The following property makes precise the idea that $\mathbb{E}(X | \mathcal{F})$ can be interpreted as the *best estimate* of X given the information $\mathcal{F}$. In other words, we give the conditional expectation (a probabilistic concept) a *statistical* interpretation.

Proposition 2.3.3 (Least squares property). *Let $X \in \mathcal{L}^2$. Then $\mathbb{E}(X | \mathcal{F})$ is the least-mean-square estimate of X given $\mathcal{F}$, i.e., $\mathbb{E}(X | \mathcal{F})$ is the unique $\mathcal{F}$ -measurable random variable that satisfies $\mathbb{E}((X - \mathbb{E}(X | \mathcal{F}))^2) = \min_{Y \in \mathcal{L}^2(\mathcal{F})} \mathbb{E}((X - Y)^2)$, where $\mathcal{L}^2(\mathcal{F}) = \{Y \in \mathcal{L}^2 : Y \text{ is } \mathcal{F}\text{-measurable}\}$.*

Beside its statistical interpretation, you can also interpret this result *geometrically*: the conditional expectation $\mathbb{E}(X | \mathcal{F})$ is the *orthogonal projection* of $X \in \mathcal{L}^2$ onto the linear subspace $\mathcal{L}^2(\mathcal{F}) \subset \mathcal{L}^2$ with respect to the inner product $\langle X, Y \rangle = \mathbb{E}(XY)$.

Proof. First, note that $\mathbb{E}((X - Y)^2)$ is finite for any $Y \in \mathcal{L}^2(\mathcal{F})$ by Hölder's inequality. Clearly $\mathbb{E}((X - Y)^2) = \mathbb{E}((X - \mathbb{E}(X | \mathcal{F}) + \mathbb{E}(X | \mathcal{F}) - Y)^2)$. But $\Delta = \mathbb{E}(X | \mathcal{F}) - Y$ is $\mathcal{F}$ -measurable by construction, so using the elementary properties of the conditional expectation (and Hölder's inequality to show that $X\Delta \in \mathcal{L}^1$) we see that $\mathbb{E}(\mathbb{E}(X | \mathcal{F})\Delta) = \mathbb{E}(X\Delta)$. Thus $\mathbb{E}((X - Y)^2) = \mathbb{E}((X - \mathbb{E}(X | \mathcal{F}))^2) + \mathbb{E}(\Delta^2)$. [Recall the geometric intuition: $\Delta \in \mathcal{L}^2(\mathcal{F})$ and $\mathbb{E}(X | \mathcal{F})$ is the orthogonal projection of X onto $\mathcal{L}^2(\mathcal{F})$, so $H = X - \mathbb{E}(X | \mathcal{F}) \perp \mathcal{L}^2(\mathcal{F})$, and thus $\langle H, \Delta \rangle = \mathbb{E}(H\Delta) = 0$.] The least squares property follows, as $\mathbb{E}(\Delta^2) \geq 0$.

To prove that the minimum is unique, suppose that Y_* is another $\mathcal{F}$ -measurable random variable that minimizes $\mathbb{E}((X - Y)^2)$. Then $\mathbb{E}((X - Y_*)^2) = \mathbb{E}((X - \mathbb{E}(X | \mathcal{F}))^2)$. But by

2.3. Conditional expectations and martingales for real

60

the general formula above, $\mathbb{E}((X - Y_*)^2) = \mathbb{E}((X - \mathbb{E}(X | \mathcal{F}))^2) + \mathbb{E}((\mathbb{E}(X | \mathcal{F}) - Y_*)^2)$. It follows that we must have $\mathbb{E}((\mathbb{E}(X | \mathcal{F}) - Y_*)^2) = 0$; this implies that $\mathbb{E}(X | \mathcal{F}) = Y_*$ a.s. $\square$

Final remark about "Conditional Expectation":

we have three ways to think about Cond. Expec.

I) (Kolmogorov): most useful to use in theory, hard to interpret
showing uniqueness (a.s.) is easy, Existence is hard (R.N. Thm)

II) (mean squared estimator): follows directly by properties of L_2
but only applies to r.v. in L_2 ($E(|X|^2) < \infty$) and
not trivial to generalize it to $X \in L_1$ (only $E(|X|) < \infty$).

III) (limit of discrete cond. expectation):
with "martingale conv. thm" we can easily show
the existence but not the uniqueness; harder to work with
[only natural for separable σ -algebras (generated by countable
collection of events (or discrete σ -algebras)).]

