Read Chapter 4

III Expectation

- $(\Omega, \mathcal{F}, P)$ a probability model
- $A \in \mathcal{F}$, $I_A(\omega) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise} \end{cases}$ indicator fn of A
- note $I_A^{-1} B = ?$
- note $I_{\bigcap_{i=1}^n A_i} = \prod_{i=1}^n I_{A_i}$, $I_{A^c} = 1 - I_A$

$$I_{\bigcup_{i=1}^n A_i} = I_{(\bigcap_{i=1}^n A_i^c)^c} = 1 - I_{\bigcap_{i=1}^n A_i^c} = 1 - \prod_{i=1}^n (1 - I_{A_i})$$

- $\omega_1, \ldots, \omega_n$ n independent performance from random system
- then $\frac{1}{n} \sum_{i=1}^n I_A(\omega_i) = r_n(A) \xrightarrow{\text{SLLN}} P(A)$ as $n \to \infty$

- if $A_1, \ldots, A_k \in \mathcal{F}$ $a_1, \ldots, a_k \in \mathbb{R}$ and
  $Y = a_1 I_{A_1} + \cdots + a_k I_{A_k}$ (a simple function) then

$$\frac{1}{n} \sum_{i=1}^n Y(\omega_i) \to a_1 P(A_1) + \cdots + a_k P(A_k) = E[Y]$$

- note - there are many choices of k, $a_1, \ldots, a_n$, $A_1, \ldots, A_k$ which we can use to represent a simple fn Y

  - each such choice leads to the same value of $E[Y]$ (all give same partition of Ω and same value on each partition element)

Def For simple function $Y = \sum_{i=1}^{k} a_i A_i : (\Omega, \mathcal{F}) \to (\mathbb{R}^1, \mathcal{B}^1)$ the <u>expectation</u> of Y is given by

$$\mathbb{E}[Y] = \sum_{i=1}^{k} a_i \, P(A_i)$$

Prop ① If Y_1, Y_2 are simple fns then
(i) $\mathbb{E}[1] = 1$
(ii) $\mathbb{E}[aY_1 + bY_2] = a\mathbb{E}[Y_1] + b\mathbb{E}[Y_2]$
(iii) $Y_1 \le Y_2$ implies $\mathbb{E}[Y_1] \le \mathbb{E}[Y_2]$
(iv) $Y_1 = Y_2$ a.s. then $\mathbb{E}[Y_1] = \mathbb{E}[Y_2]$.

Proof: (iii) $Y_2 - Y_1$ is a simple fn and it is nonnegative. Therefore $0 \le \mathbb{E}[Y_2 - Y_1] = \mathbb{E}[Y_2] - \mathbb{E}[Y_1]$.

Prop ② If $Y : (\Omega, \mathcal{F}) \to (\mathbb{R}^1, \mathcal{B}^1)$ is nonnegative then $\exists$ a sequence of nonnegative simple fns $Y_n \uparrow Y$. Further if W_n is another sequence of simple fns and $W_n \uparrow Y$ then $\lim_{n\to\infty} \mathbb{E}[Y_n] = \lim_{n\to\infty} \mathbb{E}[W_n]$

Proof: accept

- <u>note</u> - if $Y : \Omega \to \mathbb{R}$ then define

$$Y^+(\omega) = \begin{cases} Y(\omega), & Y(\omega) \ge 0 \\ 0, & \text{otherwise} \end{cases} \quad \underline{\text{positive part of } Y}$$

$$Y^-(\omega) = \begin{cases} -Y(\omega), & Y(\omega) \le 0 \\ 0, & \text{otherwise} \end{cases} \quad \underline{\text{negative part of } Y}$$

- if $Y : (\Omega, \mathcal{F}) \to (\mathbb{R}^1, \mathcal{B}^1)$ and $B \in \mathcal{B}^1$ then

$(Y^+)^{-1}B = Y^{-1}B \cap Y^{-1}[0,\infty) \in \mathfrak{F}$ and thus $Y^+ : (\Omega, \mathfrak{F}) \to (\mathbb{R}^1, \mathcal{B}^1)$, similarly $Y^- : (\Omega, \mathfrak{F}) \to (\mathbb{R}^1, \mathcal{B}^1)$

Def For nonnegative $Y : (\Omega, \mathfrak{F}) \to (\mathbb{R}^1, \mathcal{B}^1)$ the expectation of Y is $E[Y] = \lim_{n\to\infty} E[Y_n]$ where Y_n is a sequence of simple fn's s.t. $Y_n \uparrow Y$.

For general $Y : (\Omega, \mathfrak{F}) \to (\mathbb{R}^1, \mathcal{B}^1)$ the expectation of Y is $E[Y] = E[Y^+] - E[Y^-]$ provided at least one of $E[Y^+]$, $E[Y^-]$ is finite.

note - sometimes we use the following notation

$$E[Y] = \int_\Omega Y(\omega)\, P(d\omega)$$

Prop 3 If $Y, Y_n : (\Omega, \mathfrak{F}) \to (\mathbb{R}^1, \mathcal{B}^1)$ then

(i) $E[aY_1 + bY_2] = aE[Y_1] + bE[Y_2]$ provided $aE[Y_1] + bE[Y_2]$ is defined

(ii) $E[|Y|] = E[Y^+] + E[Y^-]$

(iii) $Y_1 \leq Y_2$ implies $E[Y_1] \leq E[Y_2]$, $E[Y_1]$ defined.

(iv) if $Y_1 = Y_2$ a.s. then $E[Y_1] = E[Y_2]$

Proof: accept

Def $Y_n : (\Omega, \mathfrak{F}) \to (\mathbb{R}^1, \mathcal{B}^1)$ converges almost surely to $Y : (\Omega, \mathfrak{F}) \to (\mathbb{R}^1, \mathcal{B}^1)$ if $P(\{\omega \mid \lim_{n\to\infty} Y_n(\omega) \neq Y(\omega)\}) = 0$.

Prop ④ Suppose that Y_n is a sequence of r.v.'s converging almost surely to r.v. Y. Then

(i) (Monotone Convergence) if $0 \le Y_1 \le Y_2 \le \ldots$ then $E[Y_n] \to E[Y]$

(ii) (Dominated Convergence) if $\exists$ W s.t. $|Y_n| \le |W|$ $\forall n$ and $E[|W|] < \infty$ then $E[Y_n] \to E[Y]$

Proof: accept.

note – a r.v. Y leads to prob. model $(\mathbb{R}^1, \mathcal{B}^1, P_Y)$

Prop ⑤ For r.v. Y and $h: (\mathbb{R}^1, \mathcal{B}^1) \to (\mathbb{R}^1, \mathcal{B}^1)$ then $E_P[h(Y)] = E_{P_Y}[h]$ provided $E_P[h(Y)]$ exists.

Proof: accept. (do for simple fns) | eg

Prop ⑥ (i) If Y is a discrete r.v. and $E|h(Y)| < \infty$ then
$$E[h(Y)] = \sum_{y \in \mathbb{R}} h(y)\, p_Y(y)$$

(ii) If Y is an abs. cont. r.v. and $E|h(Y)| < \infty$ then
$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) f_Y(y)\, dy.$$

Proof: accept.

Def: For r.v. Y we define:

(a) the <u>mean</u> of Y given by $\mu = E[Y]$ provided the expectation exists.

(b) the <u>variance</u> of Y given by $\sigma^2 = Var[Y] = E[(Y-\mu)^2]$ provided this expectation exists.

For stochastic process $\{Y_t : t \in T\}$ with $T \subseteq \mathbb{R}$ we define

(a) the <u>mean function</u> of Y given $\mu(t) = E[Y_t]$ provided the expectation exists.

(b) the <u>autocovariance function</u> of Y given by
$\sigma(t,u) = COV[Y_t, Y_u] = E[(Y_t - \mu_t)(Y_u - \mu_u)]$
provided this expectation exists

(c) the <u>autocorrelation function</u> of Y given by

$$\rho(t,u) = \frac{\sigma_{tu}}{\sqrt{\sigma_{tt}}\sqrt{\sigma_{uu}}}$$

provided all the expectations exist.

- <u>note</u> ① $COV[Y_t, Y_u] \overset{Ex}{=} E[Y_t Y_u] - \mu_t \mu_u$

- <u>note</u> ② if $EY_t^2 < \infty$, $EY_u^2 < \infty$ then $E|Y_t Y_u| < \infty$

Proof: $|Y_t Y_u| = |Y_t||Y_u| \le \max\{|Y_t|^2, |Y_u|^2\}$

$\le |Y_t|^2 + |Y_u|^2$

* independence: $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$

- a stochastic process $\{Y_t \mid t \in T\}$ for which $\mathbb{E}Y_t^2 < \infty \ \forall t \in T$ will be called an <u>L^2-process</u>

- <u>note</u> ③ - if $T = \{1, \dots, n\}$ then

$$\underset{\sim}{\mu} = \begin{pmatrix} \mathbb{E}[Y_1] \\ \vdots \\ \mathbb{E}[Y_n] \end{pmatrix} = \mathbb{E}[\underset{\sim}{Y}] \text{ is called the \underline{mean vector}}$$

$$\Sigma = \begin{pmatrix} \mathrm{cov}[Y_1, Y_1] & \mathrm{cov}[Y_1, Y_2] & \cdots & \mathrm{cov}[Y_1, Y_n] \\ \vdots & & & \\ \mathrm{cov}[Y_n, Y_1] & \cdots & & \mathrm{cov}[Y_n, Y_n] \end{pmatrix}$$

$$= \mathbb{E}[(\underset{\sim}{Y} - \underset{\sim}{\mu})(\underset{\sim}{Y} - \underset{\sim}{\mu})']$$

is called the <u>variance matrix</u>.

* e.g.

<u>Prop 7</u> (<u>Markov's Inequality</u>)
If $Y \geq 0$ is a r.v. then $P(Y \geq t) \leq \mathbb{E}[Y]/t$.

Proof: $P(Y \geq t) = \mathbb{E}[I_{\{Y \geq t\}}]$

$\leq \mathbb{E}[\frac{Y}{t} I_{\{Y \geq t\}}] \leq \mathbb{E}[\frac{Y}{t}] = \mathbb{E}[Y]/t$

<u>Corollary</u> <u>Chebyshev's Inequality</u>
If Y is a r.v. with mean μ and variance σ^2 then $P(|Y-\mu| \geq k\sigma) \leq 1/k^2$

Proof: $P(|Y-\mu| \geq k\sigma) = P((Y-\mu)^2 \geq k^2\sigma^2)$ and apply Markov to r.v. $(Y-\mu)^2$.

eg $\underset{\sim}{X} \sim N_k(\underset{\sim}{\mu}, \Sigma)$

- recall $\underset{\sim}{Z} = \Sigma^{-1/2}(\underset{\sim}{X} - \underset{\sim}{\mu}) \sim N_k(\underset{\sim}{0}, I)$

$$\text{so } \underset{\sim}{X} = \underset{\sim}{\mu} + \Sigma^{1/2}\underset{\sim}{Z}$$

- also recall when $\underset{\sim}{Z} \sim N_k(0, I)$ then $Z_1, \ldots, Z_k$ are independent and $Z_i \sim N(0,1)$

- therefore $E(\underset{\sim}{Z}) = \begin{pmatrix} E(Z_1) \\ \vdots \\ E(Z_k) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \underset{\sim}{0}$

- also $Var(\underset{\sim}{Z}) = E\left((\underset{\sim}{Z} - \underset{\sim}{0})(\underset{\sim}{Z} - \underset{\sim}{0})'\right)$

$$= E(\underset{\sim}{Z}\underset{\sim}{Z}')$$

$$= \left(E(Z_i Z_j) \right)$$

$i \neq j$, $E(Z_i Z_j) \overset{\text{independence}}{=} E(Z_i)E(Z_j) = 0$

$i = j$ $E(Z_i^2) = Var(Z_i) = 1$

$$\therefore E(\underset{\sim}{Z}) = \underset{\sim}{0}, \quad Var(\underset{\sim}{Z}) = I$$

- then $E(\underset{\sim}{X}) = E(\underset{\sim}{\mu} + \Sigma^{1/2}\underset{\sim}{Z})$ so

$E(X_i) = E\left(\mu_i + \sum_{j=1}^{k} c_{ij} Z_j\right)$ $\quad \Sigma^{1/2} = (c_{ij})$

$$= \mu_i + \sum_{j=1}^{k} c_{ij} E(Z_j) = \mu_i$$

$\therefore E(\underset{\sim}{X}) = \underset{\sim}{\mu}$

- $Var(\underset{\sim}{X}) = (Cov(X_i, X_j))$

$$= (E((X_i - \mu_i)(X_j - \mu_j)))$$

- $X_i = \mu_i + \sum_{l=1}^{k} c_{il} Z_l \quad , \quad X_j = \mu_j + \sum_{m=1}^{k} c_{jm} Z_m$

- so $E((X_i - \mu_i)(X_j - \mu_j))$

$$= \sum_{l=1}^{k} \sum_{m=1}^{k} c_{il} c_{jm} E(Z_l Z_m)$$

$$= \sum_{l=1}^{k} c_{il} c_{jl}$$

$$= (\Sigma^{1/2} \Sigma^{1/2})_{ij} = (\Sigma)_{ij}$$

$\therefore Var(\underset{\sim}{X}) = \Sigma$

Markov's Inequality

Prop When $Y \geq 0$ and $t > 0$ then $P(Y \geq t) \leq \frac{E(Y)}{t}$

Proof: $P(Y \geq t) = E(I_{\{Y \geq t\}}) \leq E(I_{\{Y \geq t\}} \frac{Y}{t})$

$\leq E(\frac{Y}{t}) = E(Y)/t.$

- when do we have equality?

Corollary $P(Y \geq t) = \frac{E(Y)}{t}$ iff $P(Y = t) = 1 - P(Y = 0)$

Proof: $\Rightarrow$ $0 \leq E(Y) = E(I_{\{Y < t\}} Y) + E(I_{\{Y \geq t\}} Y)$

$= E(I_{\{Y < t\}} Y) + t P(Y \geq t)$ which implies $E(I_{\{Y < t\}} Y) = 0$

which implies $P(0 < Y < t) = 0$. Then $0 \leq E(Y)$

$= t P(Y = t) + E(I_{\{Y > t\}} Y) \geq t P(Y = t) + t P(Y > t)$

$= t P(Y \geq t)$ with the inequality strict whenever

$P(Y > t) > 0$. Therefore $P(Y > t) = 0$ and $E(Y)$

$= t P(Y = t) + 0 \cdot P(Y = 0)$

$\Leftarrow$ Obvious.

- **note** - if Y is positive the inequality is exact then Y is degenerate at t.

<u>Corollary</u> $E(Y)=0$ iff $P(Y=0)=1$

Proof: For every n, $P(Y \geq \frac{1}{n}) \leq nE(Y) = 0$ and

so $P(Y>0) = \lim_{n\to\infty} P(Y \geq \frac{1}{n}) = 0$.

<u>Corollary</u> (<u>Chebyshev's Inequality</u>)
If Y is a r.v. with mean μ and variance σ^2, then $P(|Y-\mu| \geq k\sigma) \leq 1/k^2$ for any $k>0$, with equality iff $P(Y=\mu \pm k\sigma) = 1-P(Y=\mu)$

Proof: To get the inequality apply MI to

$(Y-\mu)^2$ and note $P((Y-\mu)^2 \geq k^2\sigma^2) = P(|Y-\mu| \geq k\sigma)$.

For equality note $|Y-\mu| = k\sigma$ iff $Y=\mu+k\sigma$

or $Y=\mu-k\sigma$ and $|Y-\mu|=0$ iff $Y=\mu$.

<u>Corollary</u> $P(Y=\mu)=1$ iff $\sigma^2=0$.

- <u>note</u> whenever $g(Y) \geq 0$ then MI gives

$$P(g(Y) \geq t) \leq E(g(Y))/t.$$

Prop (Cauchy-Schwartz Inequality)

If $E(X^2)<\infty$, $E(Y^2)<\infty$ then

$$|E(XY)| \le \sqrt{E(X^2)}\,\sqrt{E(Y^2)}$$

with equality iff $Y=cX$ or $X=cY$ a.s.
If $Y=cX$ and $EY^2=0$ then $c=0$ and if $EY^2>0$, $EX^2>0$ then $c=E(XY)/EX^2$.

Proof: Suppose $EY^2=0$. Then $P(Y^2=0)=P(Y=0)=1$ and so $P(XY=0)=1$ which implies $0=E(XY)$ $=\sqrt{E(X^2)}\,\sqrt{E(Y^2)}$ and $Y=0X$.

Suppose then $E(X^2)>0$, $E(Y^2)>0$. Then $0\le(Y-cX)^2=Y^2-2cXY+c^2X^2$ so $0\le E(Y-cX)^2$ $=E(Y^2)-2cE(XY)+c^2E(X^2)$ which is minimized by $c=E(XY)/E(X^2)$ and the minimized value is

$E(Y^2)-2(E(XY))^2/E(X^2)+(E(XY))^2/E(X^2)$

$=E(Y^2)-(E(XY))^2/E(X^2)$ which gives the result.

If $0=E(Y-cX)^2$ then $Y=\left(\frac{E(XY)}{E(X^2)}\right)X$

Corollary (Correlation Inequality)

If $Var(X) < \infty$, $Var(Y) < \infty$ then

$$|Corr(X,Y)| \le 1$$

with equality iff $Y = E(Y) + 0(X - E(X))$ when $Var(Y) = 0$, and $Y = E(Y) + \frac{Cov(X,Y)}{Var(X)}(X - E(X))$ when $Var(X), Var(Y) > 0$.

Proof: This follows from CS using r.v.s $(X - E(X))$ and $(Y - E(Y))$.

eg <u>best affine predictor</u>

- predict Y from X via an affine fn $a+bX$ assuming $E(Y^2)<\infty$, $E(X^2)<\infty$.

- what choice of a and b are best?

- find values minimizing $E((Y-a-bX)^2)$

$$= E(((Y-\mu_Y)-(a'-b(X-\mu_X)))^2 \qquad a' = a-\mu_Y-b\mu_X$$

$$= E((Y-\mu_Y)^2) - 2a'E((Y-\mu_Y))$$

$$-2bE((Y-\mu_Y)(X-\mu_X)) +$$

$$+(a')^2 - 2a'b\,E((X-\mu_X)) + b^2 E(X-\mu_X)^2$$

$$= Var(Y) - 2b\,Cov(X,Y) + b^2\,Var(X) + (a')^2$$

$\therefore$ choose $a = \mu_Y + b\mu_X$

$$b = \frac{Cov(X,Y)}{Var(X)}$$

and best affine predictor of Y from X is

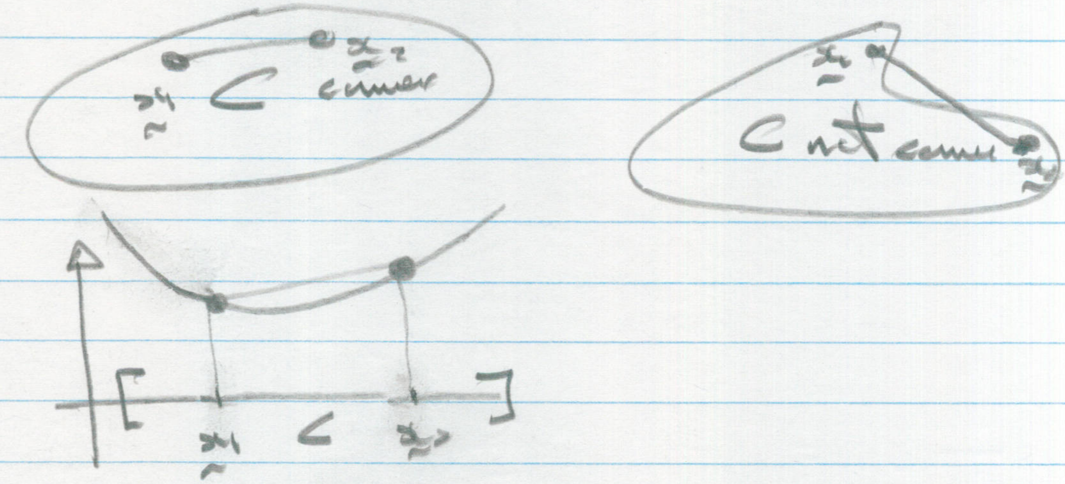
$$Y = \mu_Y + \frac{Cov(X,Y)}{Var(X)}(X-\mu_X)$$

Jensen's Inequality

Def A $C \subseteq \mathbb{R}^k$ is convex if whenever $\underset{\sim}{x}_1, \underset{\sim}{x}_2 \in C$, $p \in [0,1]$ then $p\underset{\sim}{x}_1 + (1-p)\underset{\sim}{x}_2 \in C$.

A function $h: C \to \mathbb{R}$ is convex if $h(p\underset{\sim}{x}_1 + (1-p)\underset{\sim}{x}_2) \leq p\, h(\underset{\sim}{x}_1) + (1-p)\, h(\underset{\sim}{x}_2)$

eg



Prop (Jensen's Inequality)

If Y is a r.v. taking values in a convex set $C \subseteq \mathbb{R}$ with mean $E(Y)$ then $E(h(Y)) \geq h(E(Y))$ with equality iff $h(x) = a + b$ for some a, b.

Proof: Note a convex set $C \subseteq \mathbb{R}$ is an interval say $C = (a,b)$ so and $a < Y < b$ and so $E(Y) \in (a,b)$.

Now for $x \in (a,b)$ and ε small enough

$$h(x) = h\left(\tfrac{1}{2}(x+\varepsilon) + \tfrac{1}{2}(x-\varepsilon)\right) \leq \tfrac{1}{2}h(x+\varepsilon) + \tfrac{1}{2}(h(x-\varepsilon))$$

and so
$$\frac{h(x) - h(x-\varepsilon)}{\varepsilon} \leq \frac{h(x+\varepsilon) - h(x-\varepsilon)}{2\varepsilon}$$
$$\leq \frac{h(x+\varepsilon) - (2h(x) - h(x+\varepsilon))}{2\varepsilon}$$
$$= \frac{h(x+\varepsilon) - h(x)}{\varepsilon}$$

For $0 < \varepsilon' < \varepsilon$ and p s.t. $(1-p)\varepsilon = \varepsilon'$

$$\frac{h(x) - h(x-\varepsilon')}{\varepsilon'} = \frac{h(x) - h(px + (1-p)(x-\varepsilon))}{\varepsilon'}$$
$$\geq \frac{h(x) - p\,h(x) - (1-p)\,h(x-\varepsilon)}{\varepsilon'}$$
$$= \frac{(1-p)(h(x) - h(x-\varepsilon))}{\varepsilon'} = \frac{h(x) - h(x-\varepsilon)}{\varepsilon}$$

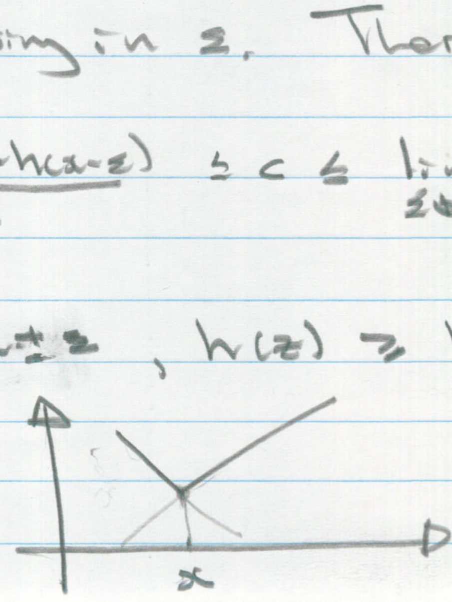
Therefore $\dfrac{h(x) - h(x-\varepsilon)}{\varepsilon}$ is nonincreasing in ε

while using a similar argument $\dfrac{h(x+\varepsilon) - h(x)}{\varepsilon}$

is nondecreasing in ε. Therefore, if $c \in \mathbb{R}$ satisfies

$$\lim_{\varepsilon \downarrow 0} \frac{h(x) - h(x-\varepsilon)}{\varepsilon} \leq c \leq \lim_{\varepsilon \downarrow 0} \frac{h(x+\varepsilon) - h(x)}{\varepsilon}$$ then

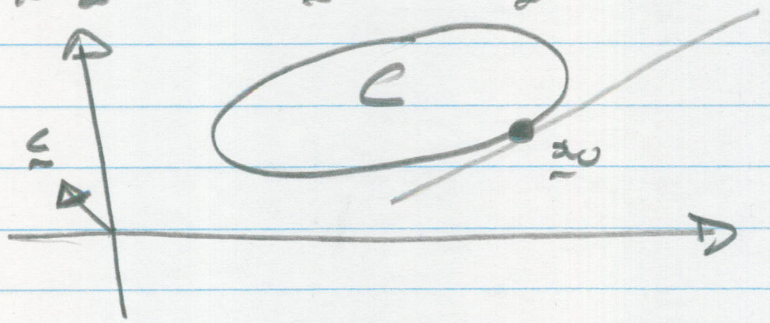
putting $z = x \pm \varepsilon$, $h(z) \geq h(x) + c(z-x)$ $\forall z \in C$



Applying this with $x = E(Y)$ gives $h(Y) \geq h(E(Y)) + c\,(Y - E(Y))$ and so $E(h(Y)) \geq h(E(Y))$. If equality holds then $h(Y) - h(E(Y)) - c(Y - E(Y)) \geq 0$ with expectation 0 which implies $h(Y) = a + bY$ where $a = h(E(Y)) + cE(Y)$ and $b = c$.

- we want to generalize this to $\mathbb{R}^k$ and for this we need a few results
- we quote without proof two results we should all know and then use them.

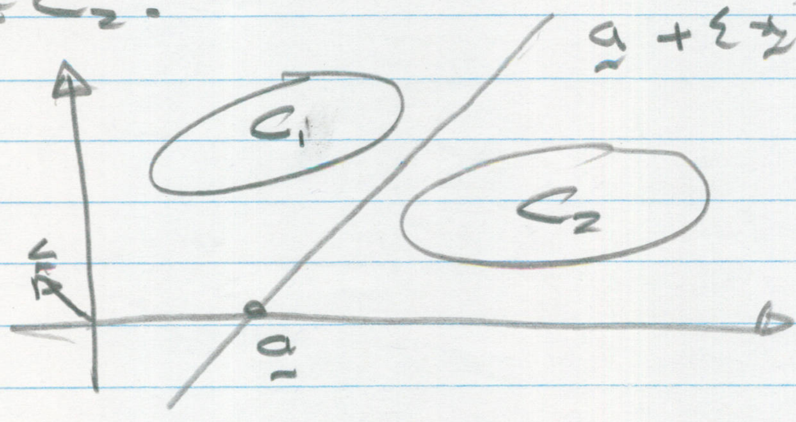
Theorem (Supporting Hyperplane Theorem)
If $C \subseteq \mathbb{R}^k$ is convex and $\underset{\sim}{x}_0$ is not an interior point of C (there isn't a ball about x_0 contained in C) then $\exists\, \underset{\sim}{c} \in \mathbb{R}^k \setminus \{\underset{\sim}{0}\}$ s.t. $\underset{\sim}{c}'\underset{\sim}{x} \geq \underset{\sim}{c}'\underset{\sim}{x}_0 \;\forall\, \underset{\sim}{x} \in C$.



$\underset{\sim}{x}_0 + \{\underset{\sim}{x} : \underset{\sim}{c}'\underset{\sim}{x} = 0\}$

supporting hyperplane.

Theorem (Separating Hyperplane Theorem)
If $C_1, C_2 \subseteq \mathbb{R}^k$ are disjoint and convex then $\exists\, \underset{\sim}{c} \in \mathbb{R}^k \setminus \{\underset{\sim}{0}\}$ s.t. $\underset{\sim}{c}'\underset{\sim}{x} \leq \underset{\sim}{c}'\underset{\sim}{y} \;\forall\, \underset{\sim}{x} \in C_1$, $\underset{\sim}{y} \in C_2$.



$\underset{\sim}{a} + \{\underset{\sim}{x} : \underset{\sim}{c}'\underset{\sim}{x} = 0\}$

separating hyperplane

<u>Theorem</u> If $P_{\underset{\sim}{X}}(C) = 1$ where $C \subseteq \mathbb{R}^k$ is convex and if $E(\underset{\sim}{X})$ exists then $E(\underset{\sim}{X}) \in C$.

Proof: Consider the set $C^* = C - E(\underset{\sim}{X})$ which is convex and $E(\underset{\sim}{X}) \in C$ iff $\underset{\sim}{0} \in C'$.

Suppose $\underset{\sim}{0} \notin C^*$. Now $\{\underset{\sim}{0}\}$ and C are disjoint convex sets so by SHT $\exists\ \underset{\sim}{c} \neq 0$ s.t. $0 = \underset{\sim}{c}'\underset{\sim}{0} \leq \underset{\sim}{c}'y \quad \forall y \in C^*$. Therefore if $\underset{\sim}{Y} = \underset{\sim}{X} - E(\underset{\sim}{X})$ we have $P(\underset{\sim}{Y} \in C^*) = 1$ and so $\underset{\sim}{c}'\underset{\sim}{Y} \geq 0$ a.s. and $E(\underset{\sim}{c}'\underset{\sim}{Y})$ $= \underset{\sim}{c}'E(\underset{\sim}{Y}) = 0$ and so $P(\underset{\sim}{c}'\underset{\sim}{Y} = 0) = 1$

This implies $P(\underset{\sim}{Y} \in \{y : \underset{\sim}{c}'y = 0\}) = 1$ and $\{y : \underset{\sim}{c}'y = 0\} \cap C^*$ is of dimension $k-1$.

So by induction (Theorem trivially true when $k=0$ since $C = \mathbb{R}^0 = \{\underset{\sim}{0}\}$)

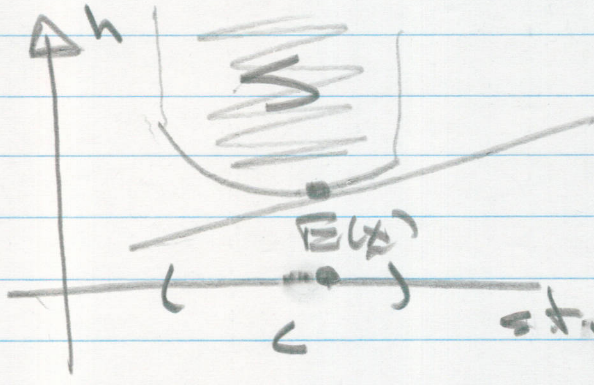
$\underset{\sim}{0} \in \{y : \underset{\sim}{c}'y = 0\} \cap C^*$ ($\Rightarrow\Leftarrow$)

Theorem (Jensen's Inequality)

If $C \subseteq \mathbb{R}^k$ is convex, $P(\underset{\sim}{X} \in C) = 1$, $E(\underset{\sim}{X})$ exists and $h : C \to \mathbb{R}$ is convex then $E(h(\underset{\sim}{X})) \geq h(E(\underset{\sim}{X}))$.

Proof: Again we proceed by induction on k

We have $E(\underset{\sim}{X}) \in C$ and so $(E(\underset{\sim}{X}), h(E(\underset{\sim}{X})))$ is a boundary point of $S = \{(\underset{\sim}{x}, \underset{\sim}{y}) : \underset{\sim}{x} \in C, \underset{\sim}{y} \geq h(\underset{\sim}{x})\}$ and note, for $p \in [0,1]$ and $(\underset{\sim}{x}_1, \underset{\sim}{y}_1), (\underset{\sim}{x}_2, \underset{\sim}{y}_2) \in S$ then $p(\underset{\sim}{x}_1, \underset{\sim}{y}_1) + (1-p)(\underset{\sim}{x}_2, \underset{\sim}{y}_2)$
$= (p\underset{\sim}{x}_1 + (1-p)\underset{\sim}{x}_2, p\underset{\sim}{y}_1 + (1-p)\underset{\sim}{y}_2) \in S$ since $p\underset{\sim}{x}_1 + (1-p)\underset{\sim}{x}_2 \in C$ and $h(p\underset{\sim}{x}_1 + (1-p)\underset{\sim}{x}_2)$
$\leq p\,h(\underset{\sim}{x}_1) + (1-p)\,h(\underset{\sim}{x}_2) \leq p\underset{\sim}{y}_1 + (1-p)\underset{\sim}{y}_2$.



Then by the Supporting Hyperplane Theorem $\exists\, \underset{\sim}{c} \in \mathbb{R}^{k+1}$

s.t. $\sum_{i=1}^{k+1} c_i z_i \geq \sum_{i=1}^{k} c_i E(X_i) + c_{k+1} h(E(\underset{\sim}{X}))$

for every $\underset{\sim}{z} \in S$. Note that $c_{k+1} \geq 0$

otherwise large values of z_{k+1} would violate the inequality. Now put $Z_{k+1} = h(\underset{\sim}{X})$ $z_i = X_i$ for $i = 1, \ldots, k$ gives

$$c_{k+1} h(E(\underset{\sim}{X})) \leq c_{k+1} h(\underset{\sim}{X}) + \sum_{i=1}^{k} c_i (X_i - E(X_i)).$$

and taking expectations (using $E(\sum_{i=1}^{k} c_i (X_i - E(X_i))) = 0$)

$$c_{k+1} h(E(\underset{\sim}{X})) \leq c_{k+1} E(h(\underset{\sim}{X}))$$

which gives the result if $c_{k+1} > 0$. If $c_{k+1} = 0$ then $0 \leq \sum_{i=1}^{k} c_i (X_i - E(X_i))$ and since ~~—~~ has expectation 0 then $P(\sum_{i=1}^{k} c_i X_i = \sum_{i=1}^{k} c_i E(X_i)) = 1$ and so the probability dist. of $\underset{\sim}{X}$ is concentrated on the hyperplane and by induction the result follows

<u>Corollary</u> Equality iff $h(\underset{\sim}{x}) = a + \underset{\sim}{c}'\underset{\sim}{x}$ for some $a \in \mathbb{R}$

Proof: Exercise.

Conditional Expectation

Def Suppose Y is a r.v. with $E|Y| < \infty$ and $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{G})$. The conditional expectation of Y given X is defined as a function $E(Y|X) : (\mathbb{R}^1, \mathcal{B}^1) \to (\mathbb{R}^1, \mathcal{B}^1)$ satisfying

$$E(Y h(X)) = E(E(Y|X) h(X)) \qquad (*)$$

for every $h : (\mathcal{X}, \mathcal{G}) \to (\mathbb{R}^1, \mathcal{B}^1)$ for which both expectations exist

- note ① $(*)$ is really the Theorem of Total Expectation.

② $E(Y|X)(x)$ = the average value of $Y(\omega)$ given that $X(\omega) = x$ (given $\omega \in X^{-1}\{x\}$)

Prop If $E|Y| < \infty$ then $E(Y|X)$ exists and is unique a.s. P_X.

Proof: accept

- note ① - if we can find g s.t. $E(Y h(X)) = E(g(X) h(X))$ for every h then $g = E(Y|X)$

② - notation $E(Y|X)(x) = E(Y|X=x)$ and sometimes $E(Y|\mathcal{F}_X)$

Corollary $E(Yg(X)|X) = E(Y|X)g(X)$

Proof: $E(Yg(X)h(X)) = E(E(Y|X)g(X)h(X))$ by definition $\forall h$. Therefore by the Proposition $E(Yg(X)|X) = E(Y|X)g(X)$.

Prop Theorem of Total Expectation
If $E|Y| < \infty$ then $E(Y) = E(E(Y|X))$

Proof: Let $h \equiv 1$ in the definition.

- **note** If $Y = I_A$ then we write $P(A|X) = E(I_A|X)$ as the conditional probability of A given X

eg discrete X, Y

- suppose (X, Y) is a discrete random vector with probability function $p_{(X,Y)}$
- then $E(Y|X)$ is given by

$$E(Y|X)(x) = \sum_y y\, p_{Y|X}(y|x)$$

where $p_{Y|X}(y|x) = p_{(X,Y)}(x,y)/p_X(x)$

since $E(Yh(X)) = \sum_{x,y} h(x) y\, p_{(X,Y)}(x,y)$

$$= \sum_x h(x) \left(\sum_y y \, p_{Y|X}(y|x) \right) p_X(x)$$

holds for every h.

eg absolutely continuous (X, Y)

$$E(Y|X)(x) = \int_{-\infty}^{\infty} y \, f_{Y|X}(y|x) \, dy$$

where $f_{Y|X}(y|x) = f_{(X,Y)}(x,y) / f_X(x)$

eg $\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X}_1 \\ \underset{\sim}{X}_2 \end{pmatrix} \sim N_k\left(\begin{pmatrix} \underset{\sim}{\mu}_1 \\ \underset{\sim}{\mu}_2 \end{pmatrix}_{k-l}^{l}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix} \right)$

$$\underset{\sim}{X}_1 | \underset{\sim}{X}_2 = \underset{\sim}{x}_2 \sim N_l\left(\underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}' \right)$$

then $E(\underset{\sim}{X}_1 | \underset{\sim}{X}_2)(\underset{\sim}{x}_2) = \underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)$

eg <u>best predictor of Y from X</u>

- find $h(X)$ which minimizes

$$E((Y - h(X))^2) = E\Big((Y - E(Y|X) + E(Y|X) - h(X))(\quad) \Big)$$

$$= E((Y - E(Y|X))^2 + 2E\big((Y - E(Y|X))(E(Y|X) - h(X))\big)$$

$$+ E((E(Y|X) - h(X))^2)$$

and note ≥ 0 and $= 0$ when $h(X) = E(Y|X)$

and middle term equals by TTE

$$E\left(E\left((Y-E(Y|X))(E(Y|X)-h(X))\,|\,X\right)\right)$$

by Corollary

$$= E\left(E(Y-E(Y|X)\,|\,X)\,(E(Y|X)-h(X))\right)$$

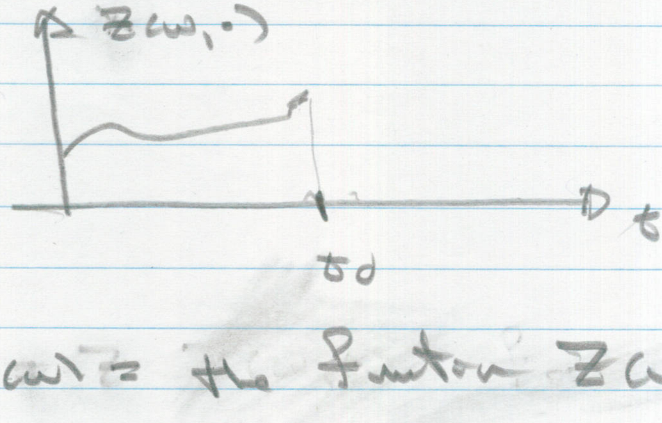
linearity

$$= E\left(\left(E(Y|X)-E(Y|X)E(1|X)\right)(E(Y|X)-h(X))\right)$$

$= 0$ since $E(Y_1+Y_2|X) \overset{Ex}{=} E(Y_1|X)+E(Y_2|X)$ and $E(1|X) \overset{Ex}{=} 1$.

- <u>note</u> - why do we need this general definition of conditional expectation?

  - because X could be observing a stoch. process at infinitely many times and so X doesn't have a density



$X(\omega) =$ the function $Z(\omega,\cdot):[0,t_d]\to\mathbb{R}^1$