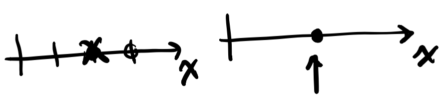
PROBABILITY Review | Appendix C | Chapter 2

Random Variables —

1D

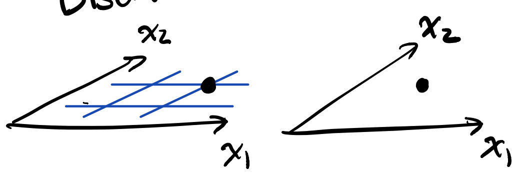
Regular Variable

Discrete Case   Continuous Case

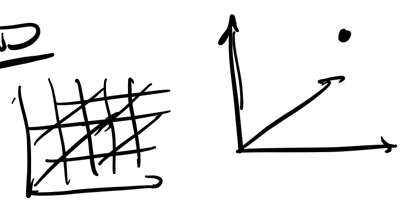


2D

Discrete



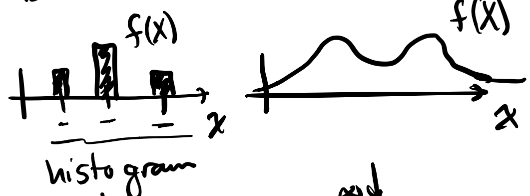
3D



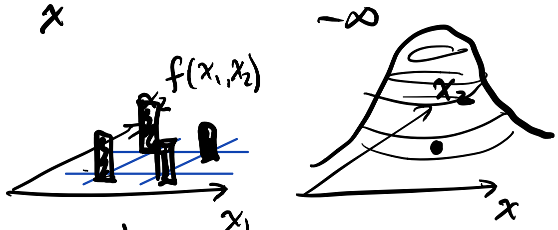
Blowing up points to density cloud

Random Variable $f(x) \geq 0$

Discrete   Continuous



$\sum_{x} f(x) = 1$ $\int_{-\infty}^{\infty} f(x)\, dx = 1$

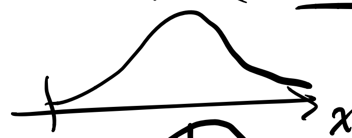


$\sum_{x_1}\sum_{x_2} f(x_1,x_2) = 1$ $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,x_2)\,dx_1 dx_2 = 1$

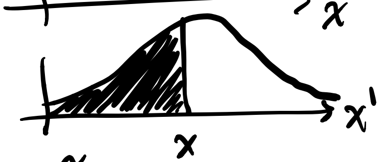
$\sum_{x_1}\sum_{x_2}\sum_{x_3} f(x_1,x_2,x_3) = 1$ $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,x_2,x_3)\,dx_1 dx_2 dx_3 = 1$

How do we define random variables? (continuous)

- density function: $f(x)$



- cummulative density function: $F(x)$



Note: measure theory - define $F(x)$ first...

$\frac{\partial F}{\partial x} = f(x)$

$\frac{\partial}{\partial x} \longrightarrow$ $F(x) = \int_{-\infty}^{x} f(x')\,dx'$

fund. thm of calculus / Leibnitz integration rule

$$\frac{\partial F}{\partial x} = f(x)$$

Expected Value:

"Expected value of $g(x)$ given $f(x)$"

$$E_{x \sim f}\big(g(x)\big) = \int_{-\infty}^{\infty} g(x) f(x)\,dx$$

Moments: $x \in \mathbb{R}$

"kth moment of f about c"

$$\mu_k = \int_{-\infty}^{\infty} (x - c)^k f(x)\,dx$$

1D:

- 1 moment of f about 0

$$\mu = \int_{-\infty}^{\infty} x\, f(x)\,dx = E_{x \sim f}[x] = E[x]$$

"mean of x"
"expected value of x"

- 2nd moment of f about μ

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx$$

"variance of $f(x)$"

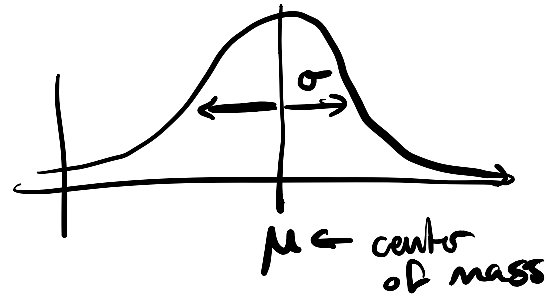
$$\sqrt{\sigma^2} = \sigma = \text{standard deviation}$$

- and so on w higher order terms...

$\underbrace{\qquad\qquad}$

↓

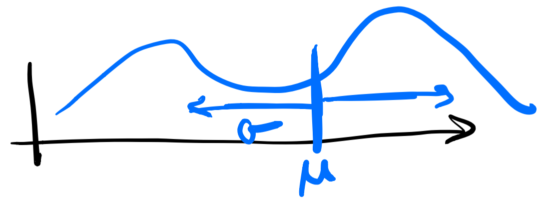
constructing a taylor expansion of $f(x)$

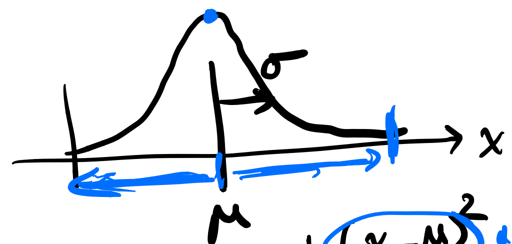


Gaussian distribution:

the mean & variance define $f(x)$

μ, σ^2 are enough to define $f(x)$ completely





also called Normal distribution: $x \sim \mathcal{N}(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

$\underbrace{\qquad}_{\text{shape}}$

2D: $\boxed{x \in \mathbb{R}^2}$ $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ $f: x \mapsto \mathbb{R}_+$ $\int f(x)\,dx = 1$

function $g(x)$

Expected of $g(x)$ $\quad E_{x\sim f}(g(x)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f(x_1, x_2)\, dx_1\, dx_2$

Moments

- 1st moment about 0 $\quad E(x) = \mu = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \underline{x} f(x_1, x_2)\, dx_1\, dx_2$

  ↑ vector $\qquad$ vector

- 2nd moment about μ

  outer prod.

  $E[\underbrace{(x-\mu)(x-\mu)^T}_{\text{matrix}}] = \Sigma = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu)(x-\mu)^T f(x_1, x_2)\, dx_1\, dx_2$

  covariance matrix

• 3rd ^{4} moment... (not as important for us...)

3 tensor $M_{ijk} = E\left((x-\mu)_i (x-\mu)_j (x-\mu)_k\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu)_i (x-\mu)_j (x-\mu)_k \, f(x_1,x_2)\, dx_1 dx_2$

N-Dimensions $x \in \mathbb{R}^n$ $\quad f: x \to \mathbb{R}_+$

$$E(x) = \mu = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \underset{\uparrow \text{vector}}{x} \, f(x_1 \dots x_n)\, dx_1 \dots dx_n \quad \leftarrow \text{mean vector}$$

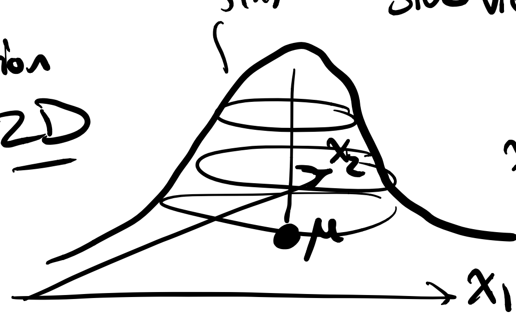
$$\underbrace{E\left[(x-\mu)(x-\mu)^T\right]}_{\text{matrix}} = \Sigma = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x-\mu)(x-\mu)^T f(x_1,\dots,x_n)\, dx_1 \dots dx_n$$

$\uparrow$ covariance matrix

Gaussian (Normal) Distribution

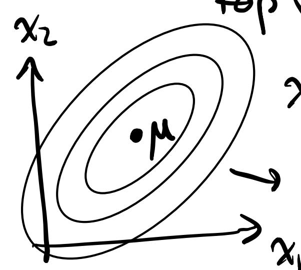
$x \sim N(\mu, \Sigma)$

2D



$$\boxed{f(x) = \left((2\pi)^n \det(\Sigma)\right)^{-1/2} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}}$$

top view...



$x \in \mathbb{R}^2$

shape of these level set ellipses is given by Σ.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

$\int f(x)dx = 1$ $\quad$ shape

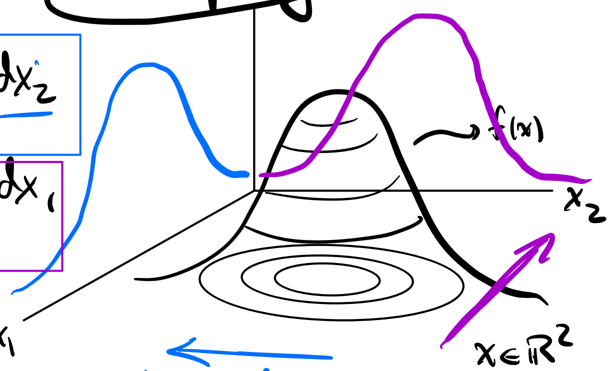
Note:

$f(x_1,\dots,x_n)$: joint distribution

Marginal Distribution → "collapsing" onto a lower dim

$$\boxed{f_{X_1}(x) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2}$$

$$\boxed{f_{X_2}(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1}$$

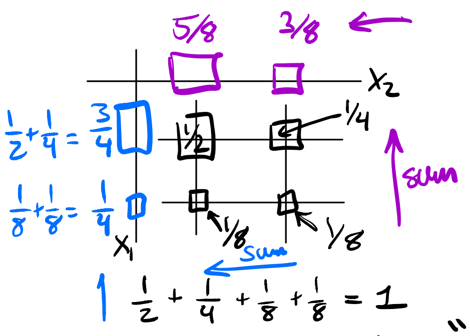


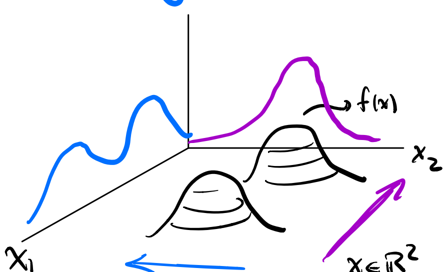
integrate
"collapse the distribution along x_2"

marginalize

Discrete case

looking down from top.





integrate
"collapse the distribution along x_2"

"written in the margins."

Conditional Distribution: slice along a dimension

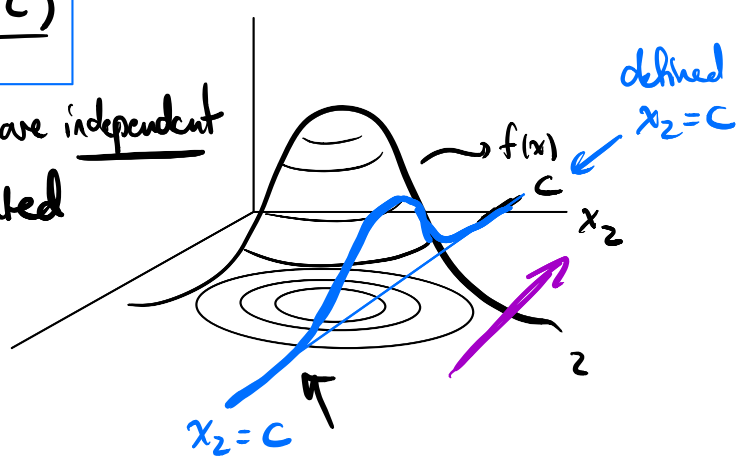
$$\boxed{f(x_1 \mid X_2 = c) = \frac{1}{\int_{-\infty}^{\infty} f(x_1, c)\, dx_1} f(x_1, c)}$$

$f(x_1, x_2)$: say that x_1 & x_2 are independent

or independently distributed

if

$$\boxed{f(x_1, x_2) = f_1(x_1) f_2(x_2)}$$



Independent Variables $x_1 \dots x_n$

$$f(x_1 \dots x_n) = f_1(x_1)\, f_2(x_2) \dots f_n(x_n)$$

→ individual distributions

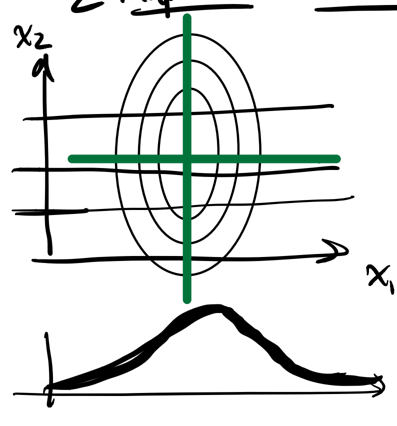
this gives → joint distribution

2D:

joint dist of 2 independent variables

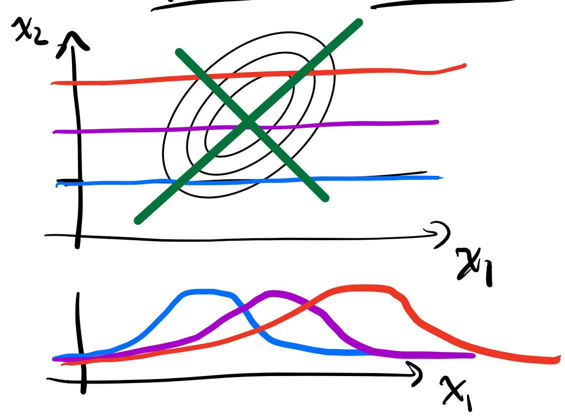
$$f(x_1 | x_2) = f_1(x_1)$$

No matter what x_2 is the conditional prob of x_1 doesn't change



Gaussian level sets

2 dependent variables.



$$f(x_1, x_2)$$

$$\boxed{f(x_1 | X_2 = c) = \frac{1}{\int_{-\infty}^{\infty} f(x_1, c)\, dx_1} f(x_1, c)}$$

$$f(x_1 | x_2 = c) = f(x_1, x_2 = c) \times \frac{1}{f_2(c) \underbrace{\int_{-\infty}^{\infty} f_1(x_1)\, \cancel{f_2(c)}\, dx_1}_{1}} f_1(x_1) f_2(c) = \frac{1}{\cancel{f_2(c)}} f_1(x_1) \cancel{f_2(c)}$$

$$f_{x_1}(x_1) = \int_{-\infty}^{\infty} \underbrace{f(x_1,x_2)}_{f_1(x_1)f_2(x_2)} dx_2 = f_1(x_1)\underbrace{\int_{-\infty}^{\infty} f_2(x_2)\,dx_2}_{1} = f_1(x_1)$$

Covariance Matrices

$x \in \mathbb{R}^n$

$$\Sigma = E\left((x-\mu)(x-\mu)^T\right)$$

$$= E\left[\begin{bmatrix} x_1-\mu_1 \\ \vdots \\ x_n-\mu_n \end{bmatrix}\begin{bmatrix} x_1-\mu_1 & \cdots & x_n-\mu_n \end{bmatrix}\right]$$

$$= E\begin{bmatrix} (x_1-\mu_1)^2 = \sigma_1^2 & & (x_1-\mu_1)(x_n-\mu_n) \\ & \ddots & \\ (x_n-\mu_n)(x_1-\mu_1) & & (x_n-\mu_n)^2 = \sigma_n^2 \end{bmatrix}$$

Covariance

σ_{1n}

σ_{n1}

variance terms

$\rho_{ij} = \dfrac{\sigma_{ij}}{\sigma_i \sigma_j}$

correlation of x_i & x_j

$$\sigma_i^2 = E\left((x_i-\mu_i)(x_i-\mu_i)\right)$$

$$\sigma_{ij} = E\left((x_i-\mu_i)(x_j-\mu_j)\right)$$

$\rho_{ij} = \dfrac{\sigma_{ij}}{\sigma_i \sigma_j}$

correlation of x_i & x_j

$$\sigma_{ij} = \Sigma_{ij} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(x_i-\mu_i)(x_j-\mu_j)\, f(x_1 \ldots x_n)\,dx_1 \cdots dx_n$$

what happens if $x_1 \dots x_n$ are independent?

$$\sigma_{ij} = \Sigma_{ij} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) f_i(x_i) f_j(x_j) f_1(x_1) \cdots f_n(x_n)\, dx_i\, dx_j\, dx\_$$

$$\underbrace{\int_{-\infty}^{\infty} (x_i - \mu_i) f_i(x_i)\, dx_i}_{0} \underbrace{\int_{-\infty}^{\infty} (x_j - \mu_j) f_j(x_j)\, dx_j}_{0} \underbrace{\int_{-\infty}^{\infty} f_1\, dx_1}_{1} \underbrace{\int_{-\infty}^{\infty} f_2\, dx_2}_{1} \cdots$$

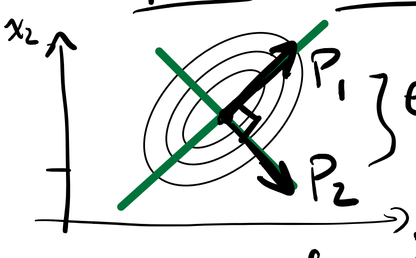
if x_i & x_j are independent

$$\Rightarrow \boxed{\sigma_{ij} = \Sigma_{ij} = 0}$$

if all $x_1 \dots x_n$ are independent $\Rightarrow \Sigma$ diagonal

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}$$

2 dependent variables.

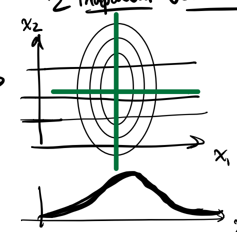


} eigenvectors of Σ

Gaussian

diagonal Σ

2D. joint dist of 2 independent variables



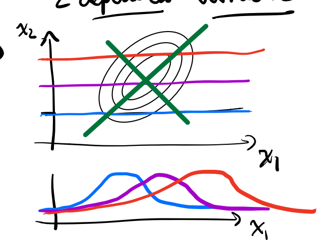
coord. transform on space P

random vector

$$x = \overbrace{\begin{bmatrix} P_1 & P_2 \end{bmatrix}}^{P} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

non diagonal Σ

2 dependent variables.



in the exponential $\rightarrow$ orthonormal

$$f(x) = Ce^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$\bar{\mu} = P^{-1}\mu$$

$$\underbrace{f(Pz)}_{\bar{f}(z)} \approx e^{-\frac{1}{2}(Pz - P\bar{\mu})^T \Sigma^{-1} (Pz - P\bar{\mu})}$$

$$\bar{f}(z) \approx e^{-\frac{1}{2}(z-\bar{\mu})^T \underbrace{P^T \Sigma^{-1} P}_{\text{diagonalizing } \Sigma} (z - \bar{\mu})}$$

$$\underbrace{\bar{\Sigma}^{-1}}_{\text{diagonal}} = P^T \Sigma^{-1} P \Rightarrow \underline{\Sigma^{-1}} = \underline{P}\,\underline{\bar{\Sigma}^{-1}}\,\underline{P^T}$$

$P = [p_1\ p_2]$ right eigenvectors of Σ

$P^T = \begin{bmatrix} p_1^T \\ p_2^T \end{bmatrix}$ left eigen vectors

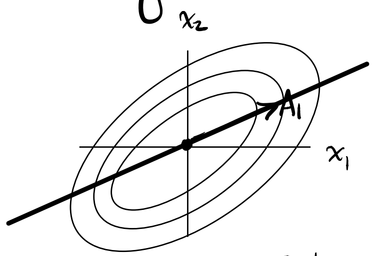
Σ : sym.

$\downarrow$

diagonalizable by an orthonormal P (rotation)

Σ & Σ^{-1} have the same eigenvectors

Slicing a Gaussian Distribution



$x \sim N(0, \underbrace{\Sigma}_{2\times 2})$

$\boxed{x = A_1 z} \rightarrow$

$z \in \mathbb{R}^2$

$z \sim N(0, \Sigma_z)$, Σ_z is 1×1

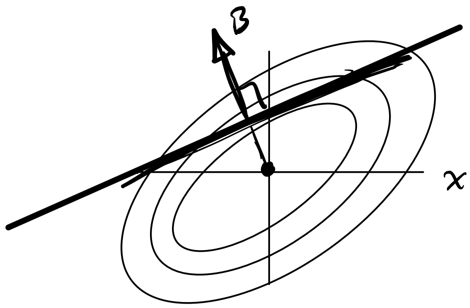
$f(x) \approx e^{-\frac{1}{2}(x-0)^T \Sigma^{-1} (x-0)}$

$$\underbrace{e^{-1/2\, z \overbrace{A_1^T \Sigma^{-1} A_1}^{\Sigma_z^{-1}} z}}_{\text{still Gaussian in } z}$$

$\Sigma_z^{-1} = A_1^T \Sigma^{-1} A_1$

$\Sigma_z = (A_1^T \Sigma^{-1} A_1)^{-1}$

Constraints on x.



$B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \in \mathbb{R}^2$

$B^T x = 1$

$\begin{bmatrix} B_1 & B_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1$

$B_1 x_1 + B_2 x_2 = 1$

$x_1 = \frac{1}{B_1} - \frac{B_2}{B_1} x_2$

Multivariate Gaussians are "jointly gaussian"

"Any slice of a jointly Gaussian distribution is still jointly Gaussian"

$\Rightarrow$ lin combs of joint Gauss. rand. variables are joint. Gauss

$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/B_1 \\ 0 \end{bmatrix} + \begin{bmatrix} -B_2/B_1 \\ 1 \end{bmatrix} x_2$

$\approx e^{-\frac{1}{2}(x^T \Sigma^{-1} x)}$

$$\underbrace{e^{-1/2 \left(\begin{bmatrix} 1/B_1 \\ 0 \end{bmatrix} + \begin{bmatrix} -B_2/B_1 \\ 1 \end{bmatrix} x_2 \right)^T \Sigma^{-1} \left(\quad \right)}}_{x_2}$$

Matrix Derivatives:

before $\frac{\partial f}{\partial x}$ ← vector

$f(X) \quad X \in \mathbb{R}^{m \times n} \qquad \frac{\partial f}{\partial X} = ?$

1. Vectorize X. ⇒ "stack X into a vector" (stack up the cols of X)

$$\text{vec}(X) \in \mathbb{R}^{mn}$$

$$\boxed{\text{vec}(ABC) = (C^T \otimes A)\,\text{vec}(B)}$$ ← wikipedia

kronecker product

$$\frac{d\,\text{vec} f}{\partial\, \text{vec} X}$$

Ex. $\underset{\text{scalar}}{f(Q)} = x^T Q x \Rightarrow \boxed{\frac{\partial\, \text{vec} f}{\partial\, \text{vec} Q} = x^T \otimes x^T}$

$$\text{vec}(f) = \text{vec}(\underset{A}{x^T} Q \underset{C}{x}) = (x^T \otimes x^T)\,\text{vec}\, Q$$

Ex. $\underset{\text{matrix}}{f(X)} = AX$ $\qquad \boxed{\frac{\partial\, \text{vec} f}{\partial\, \text{vec} X} = I \otimes A}$

$$\text{vec}(f) = \text{vec}(AXI) = (I \otimes A)\,\text{vec}(X)$$

2. think about derivative differently.
doing derivative elementwise...

Recall

$f(x) = c^T x$ $\quad \frac{\partial f}{\partial x} = c^T \quad \Delta f = \frac{\partial f}{\partial x} \cancel{\Delta x}$

row col vector

General version:

what is $\langle \cdot, \cdot \rangle$ for matrices?

$$\boxed{\Delta f = \left\langle \frac{\partial f}{\partial x}, \Delta x \right\rangle}$$

define derivative to do this...

$$\langle A, B \rangle = \sum_{ij} A_{ij} B_{ij}$$

$$\langle A, B \rangle = \sum_{ij} A_{ij} B_{ij} = \mathrm{Tr}(A^T B)$$

Many times!

$A = [A_1 \cdots A_n]$ $\quad \begin{bmatrix} -A_1^T- \\ -A_n^T- \end{bmatrix} [B_1 \cdots B_n]$

$f(X) = \mathrm{Tr}(\underline{\quad X \quad})$ $\quad B = [B_1 \cdots B_n]$ $\quad A_1^T B_1 \ldots A_n^T B_n$

$$\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB) = \mathrm{Tr}(BCA)$$

Ex. $f(X) = \mathrm{Tr}(AXB) = \mathrm{Tr}(BAX)$

$= \langle A^T B^T, X \rangle$ $\quad \boxed{\frac{\partial f}{\partial X} = A^T B^T}$

Useful Relationships:

element wise: $\frac{\partial f}{\partial X_{ij}} = [A^T B^T]_{ij}$

$$\boxed{\frac{\partial}{\partial A} Tr(BAC) = B^T C^T} \Leftarrow$$

$Tr(BAC) = Tr(CBA)$

$= \langle B^T C^T, A \rangle$

$$\boxed{\frac{\partial}{\partial A} Tr(ABA^T) = A(B+B^T)} \Leftarrow$$

page 68 of book

$f(A) = Tr(ABA^T)$ $\quad \Delta f = \langle \_\_, \Delta A \rangle$

$\Delta f = Tr(\Delta A B A^T) + \underbrace{Tr(A B \Delta A^T)}$

$Tr(M) = Tr(M^T)$

$\swarrow \; Tr(\Delta A B^T A^T)$

$= Tr(\Delta A (B+B^T) A^T)$

$= Tr((B+B^T) A^T \Delta A) = \langle \underbrace{A(B+B^T)}_{\frac{\partial f}{\partial A}}, \Delta A \rangle$

BACK TO ESTIMATION Chapter 2

PROBABILITY PERSPECTIVE ON LS.

LINEAR MEAS. $\tilde{y} = Hx + v$ $v \sim N(0, R)$

FIND LINEAR ESTIMATOR

$\hat{x} = M\tilde{y} + n$

- unbiased ⊛
- minimum variance ⊛

unbiased

$E[\hat{x}] = E[x]$

est true

if not

$E[\hat{x} - x] = \text{BIAS.}$

$$
\begin{aligned}
E[\hat{x}] &= E[MHx + Mv + n] \\
&= E[MHx] + E[Mv] + E[n] \\
&= MHE[x] + M\cancel{E[v]}^{0} + n
\end{aligned}
$$

linearity of $E[\cdot]$ (from integral)

$E[\hat{x}] = MHE[x] + n$

$\Rightarrow MH = I, \; n = \emptyset$

$\hat{x} = M\tilde{y}$ ...

sum of variance terms

$$\min_M J = \frac{1}{2} E\left[\underbrace{(\hat{x}-x)^T(\hat{x}-x)}_{\sum_i (\hat{x}_i - x_i)^2}\right] = \frac{1}{2} \underline{\mathrm{Tr}}\, E\left[\underbrace{(\hat{x}-x)(\hat{x}-x)^T}_{\text{covariance of } \hat{x}-x}\right]$$

s.t. $MH = I$ $\quad (\hat{x} = M\tilde{y})$

↓

$$\min_M J = \frac{1}{2} \mathrm{Tr}\, E\left(\underline{(\hat{x}-x)(\hat{x}-x)^T}\right)$$

s.t. $MH = I$

Parallel Axis Thm: (for unbiased estimator)

"how does covariance shift when you shift the mean / of a distribution"

axis of rotation — moment of inertia

unbiased $\hat{x} = \underbrace{MH}_{I} x$

$$E\left((\hat{x}-x)(\hat{x}-x)^T\right) = E[\hat{x}\hat{x}^T] - E[x\hat{x}^T] - E[\hat{x}x^T] + E[xx^T]$$

with respect to distribution of $\hat{x}$

$$= E[\hat{x}\hat{x}^T] - E[xx^T]\underbrace{H^TM^T}_{I} - \underbrace{MH}_{I}E[xx^T] + E[xx^T]$$

$$= E[\hat{x}\hat{x}^T] - E[xx^T]$$

Minimize variance

$$\min_M J = \frac{1}{2}\mathrm{Tr}\left(\underline{E[\hat{x}\hat{x}^T]} - \underline{E[xx^T]}\right)$$

$$\text{s.t. } \underline{MH = I} \Rightarrow (I - MH) = 0$$

NOW OPTIMIZE . . .

$$\mathcal{L}(M, \Lambda) = \frac{1}{2}\mathrm{Tr}\left(\underline{E[\hat{x}\hat{x}^T]} - \cancel{E[xx^T]}\right) + \langle \underline{\Lambda}^T, I - MH \rangle$$

primal dual

$$\mathrm{Tr}(\Lambda(I - MH))$$

$$\frac{\partial \mathcal{L}}{\partial M} = 0 \qquad \frac{\partial \mathcal{L}}{\partial \Lambda} = 0$$

$$\hat{x} = My = M(Hx + \underline{v}) = \underline{MHx} + \underline{Mv}$$

$$E[\hat{x}\hat{x}^T] = E(MHxx^TH^TM^T) + E(MHxv^TM^T) + E(Mvx^TH^TM^T) + E(Mvv^TM^T)$$

$$MH = I$$

$$= E(\cancel{xx^T}) + E(\cancel{xv^T})M^T + ME(\cancel{vx^T}) + \underline{ME[vv^T]M^T}$$

0 0

x, v are independent

noise to be independent from parameters (x)

MRM^T

$$E(xv^T) = E(x)\cancel{E(v^T)}^{0} \quad \star \qquad v \sim N(0,R)$$

$$\mathcal{L}(M,\Lambda) = \frac{1}{2}\mathrm{Tr}(MRM^T) + \mathrm{Tr}(\Lambda(I - MH))$$

$$\frac{\partial \mathcal{L}}{\partial M} = \frac{1}{2}\frac{\partial}{\partial M}\left(\mathrm{Tr}(MRM^T) - \mathrm{Tr}(\Lambda M H)\right)$$

$$\boxed{\frac{\partial}{\partial A}\mathrm{Tr}(BAC) = B^T C^T}$$

$$\boxed{\frac{\partial}{\partial A}\mathrm{Tr}(ABA^T) = A(B+B^T)}$$

$$= \frac{1}{2}M(\underbrace{R+R^T}_{2R}) - \Lambda^T H^T$$

$$-\mathrm{Tr}(\Lambda M H)$$

$$= MR - \Lambda^T H^T = 0$$

$$\Longrightarrow M = \Lambda^T H^T R^{-1} \Longleftarrow \text{LQR term...}$$

$$\frac{\partial \mathcal{L}}{\partial \Lambda} = \frac{\partial}{\partial \Lambda}\mathrm{Tr}(\Lambda(I - MH))$$

$$= (\underline{I - MH})^T = 0 \Longrightarrow MH = I$$

Combining ...

$$(\Lambda^T H^T R^{-1})H = I \Longrightarrow \Lambda^T = (H^T R^{-1} H)^{-1}$$

$$M$$

$$M = (H^T R^{-1} H)^{-1} H^T R^{-1}$$

$$\boxed{\hat{x} = M\tilde{y} = (H^T R^{-1} H)^{-1} H^T R^{-1} \tilde{y}}$$ ← minimum variance linear estimator

Note: has the same form as weighted least squares where $W = R^{-1}$

weight matrix = inverse of the covariance of noise

if you have noise $v \sim N(0, R)$

$$\tilde{y} = Hx + v$$

optimal way to weight LS is to invert covariance R ... use R^{-1} as a weighting matrix.

Next time: what if you have a prior estimate on x

$$x = \hat{x}_a + w \qquad w \sim N(0, Q)$$