Chapter 1 : Review

$$\underset{\sim}{f} : \mathbb{R}^p \longrightarrow \mathbb{R}^k \qquad \underset{\sim}{f}(\underset{\sim}{x}) = \begin{bmatrix} f_1(\underset{\sim}{x}) \\ \vdots \\ f_k(\underset{\sim}{x}) \end{bmatrix}$$

$$\underset{\sim}{x} \longmapsto \underset{\sim}{y}$$

Jacobian $\underset{\sim}{f}'(\underset{\sim}{x}) = \left[\dfrac{\partial f_i}{\partial x_j}(\underset{\sim}{x}) \right]$

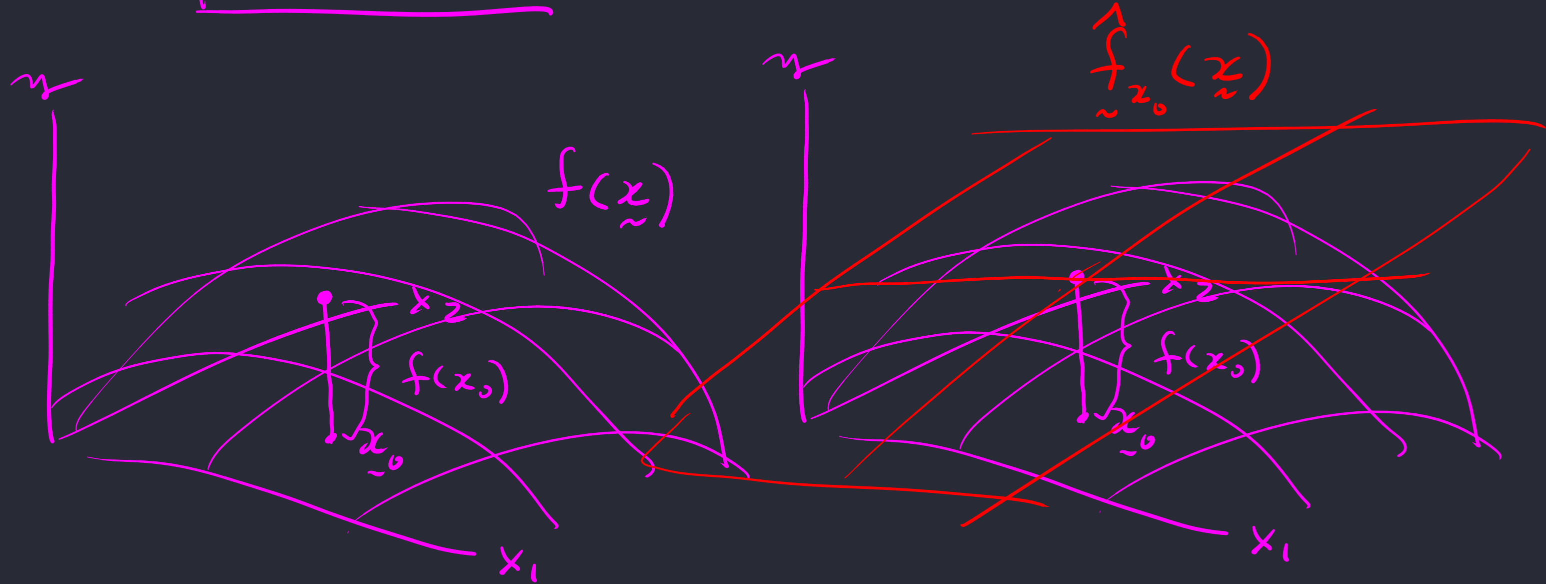
$k \times p$

Also denoted $J(\underset{\sim}{x}) = J_f(\underset{\sim}{x})$

Why? Linear approximation to diff. fcn. at $\underset{\sim}{x}_0$

$$\underbrace{\hat{\underset{\sim}{f}}_{x_0}(\underbrace{\underset{\sim}{x}}_{p\times 1})}_{k\times 1} = \underbrace{\underset{\sim}{f}(\underset{\sim}{x}_0)}_{k\times 1} + \underbrace{\underset{\sim}{f}'(\underset{\sim}{x}_0)}_{k\times p}\underbrace{(\underset{\sim}{x} - \underset{\sim}{x}_0)}_{p\times 1}$$

$p=2 \quad k=1$



Special case: $k = p$

Then J is square and

$|J(\underset{\sim}{x})|$ is Jacobian determinant.

If $\underset{\sim}{f}$ is 1-1, differentiable $\mathbb{R}^p \to \mathbb{R}^p$

Under conditions: $\Omega \subseteq \mathbb{R}^p$:

$$\int_\Omega h(\underset{\sim}{x})\, d\underset{\sim}{x} = \int_{f(\Omega)} \frac{h(f^{-1}(\underset{\sim}{y}))}{\Big|\underbrace{|J(\underset{\sim}{f}^{-1}(\underset{\sim}{y}))|}_{\text{determinant}}\Big|}\, d\underset{\sim}{y}$$

← absolute value

$$\int_\Omega d\underset{\sim}{x}$$

Area (Volume) of Ω

$$\int_{f(\Omega)} d\underset{\sim}{y} = \int_\Omega \|J_f(\underset{\sim}{x})\|\, d\underset{\sim}{x}$$

Area (Volume) of $f(\Omega)$

Jacobian of a linear (affine) transformation

Let A be a square matrix

$$\underset{\sim}{f}(\underset{\sim}{x}) = \underset{\sim}{a} + A\underset{\sim}{x}$$

Then $\underset{\sim}{f}'(\underset{\sim}{x}) = A$ (constant for all $\underset{\sim}{x}$)

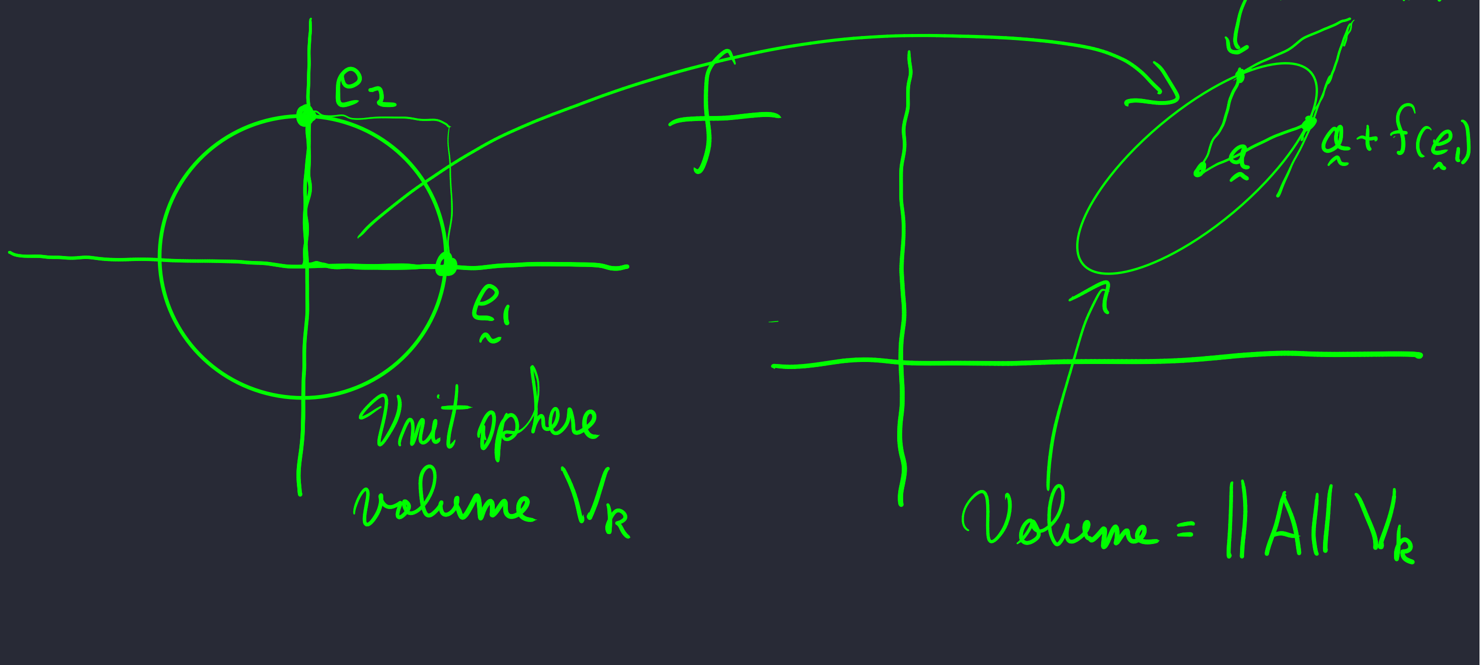
Is $\underset{\sim}{f}'(\underset{\sim}{x}) = A\underset{\sim}{x}$?

☐ Yes
☐ No

Why?

Let A be a square matrix $\ni |A| \neq 0$

$$\underset{\sim}{f}(\underset{\sim}{x}) = \underset{\sim}{a} + A\underset{\sim}{x}$$



Transforming a random vector:

Random vector: $\underset{\sim}{X}$ in $\mathbb{R}^p$, density h_X

$$\underset{\sim}{Y} = \underset{\sim}{f}(\underset{\sim}{X}), \quad f \text{ 1-1 + differentiable}$$

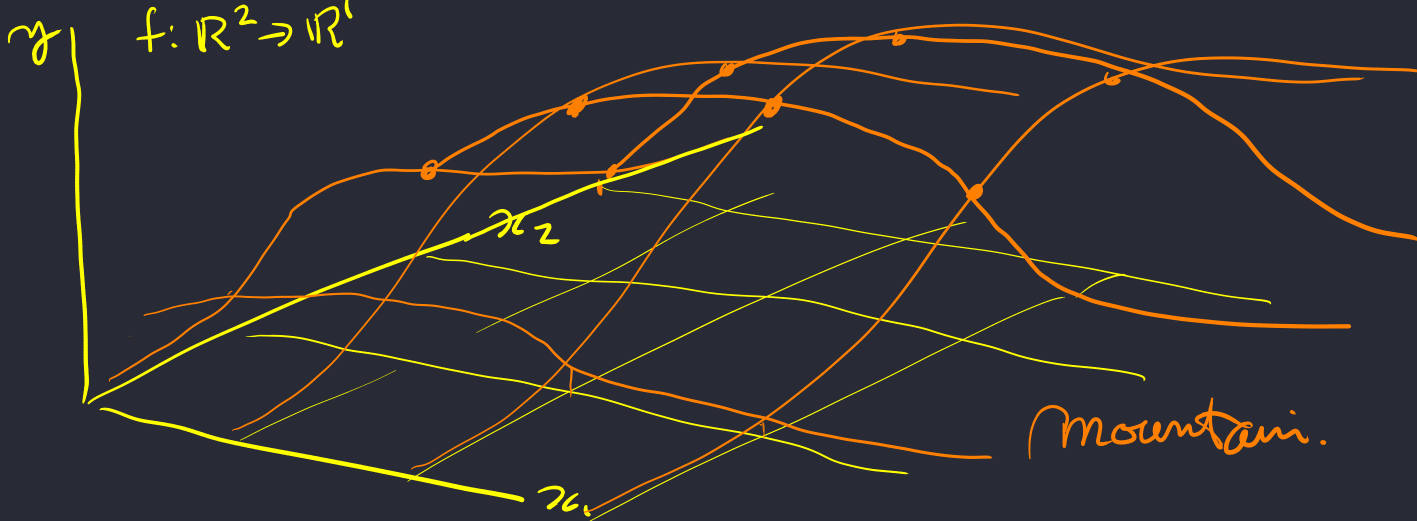
then density for $\underset{\sim}{Y}$ is

$$h_{\underset{\sim}{Y}}(\underset{\sim}{y}) = \frac{h_{\underset{\sim}{X}}(\underset{\sim}{f}^{-1}(\underset{\sim}{y}))}{\|J(f^{-1}(\underline{y}))\|}$$

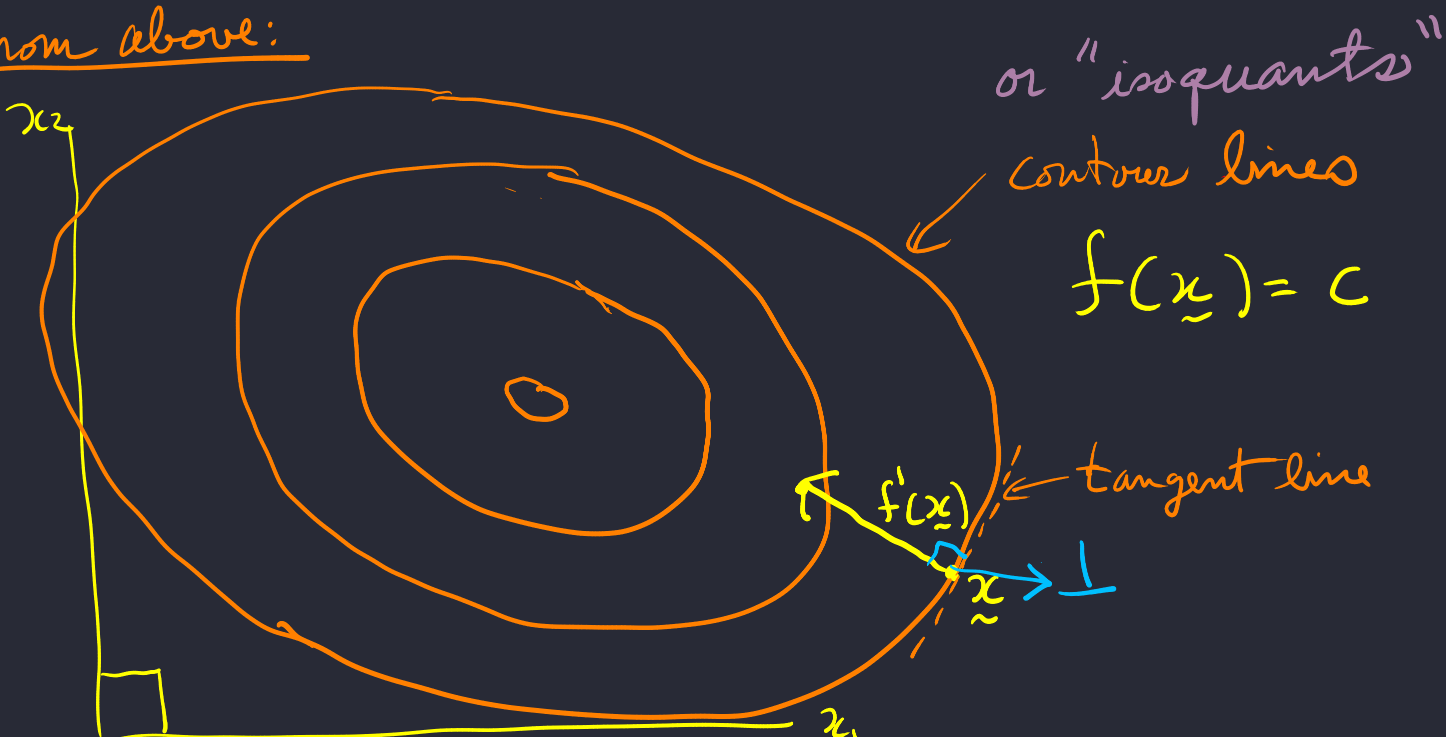
Special case: $k=1$

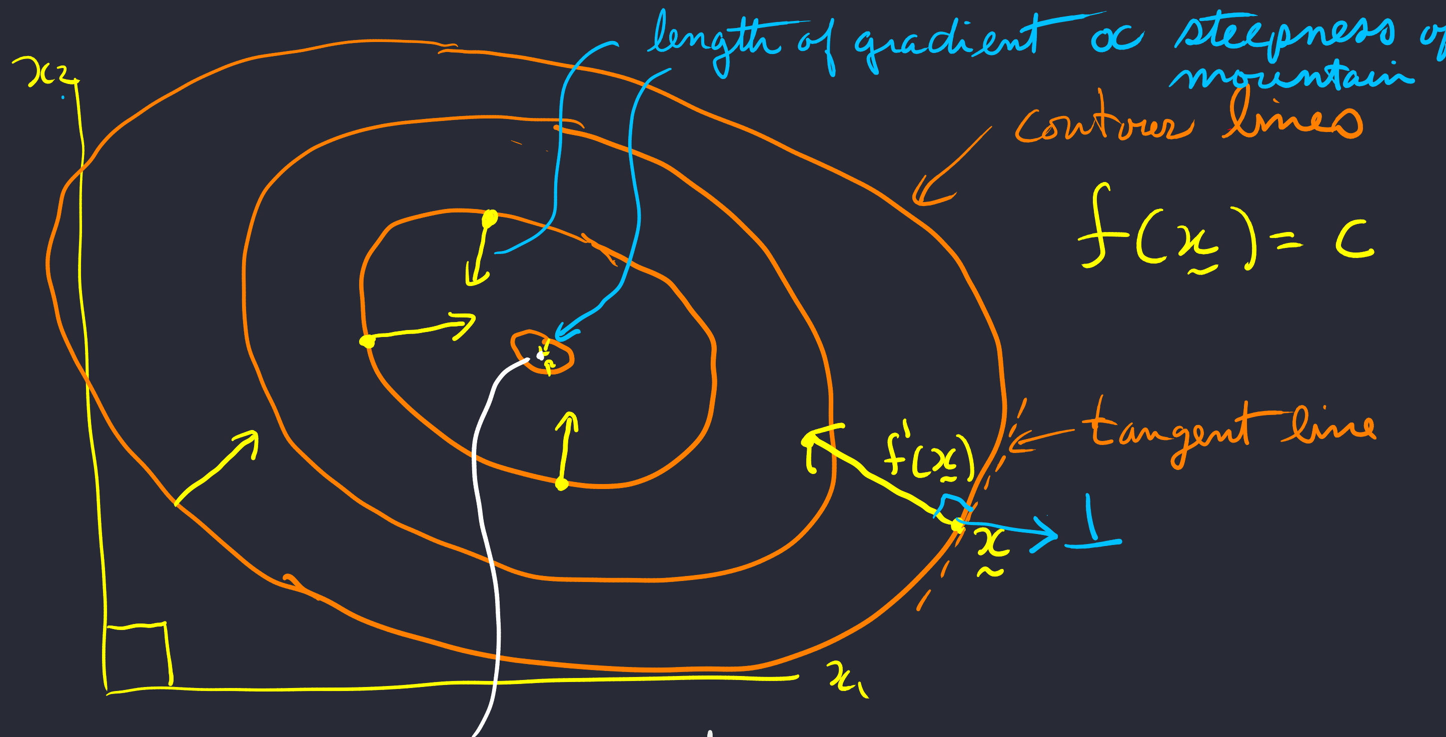
$\underset{\sim}{f}'(\underset{\sim}{x})$ is called the ______

of $p=1$: ______



From above:





At the summit $f'(\underset{\sim}{x}) = \underset{\sim}{0}$

Note: $f(\underset{\sim}{x}) \in \mathbb{R}^1$

$f'(\underset{\sim}{x}) \in \mathbb{R}^2$

<u>Hessian</u> 2nd derivative

$$f : \mathbb{R}^p \to \mathbb{R}$$

$$f''(\underset{\sim}{x}) = \left[\frac{\partial f_i}{\partial x_j}(\underset{\sim}{x}) \right]$$ is a $p \times p$ matrix

<u>Quadratic approximation at $\underset{\sim}{x}_0$</u>

$$\hat{f}_{\underset{\sim}{x}_0}(\underset{\sim}{x}) = f(\underset{\sim}{x}_0) + f'(\underset{\sim}{x}_0)(\underset{\sim}{x} - \underset{\sim}{x}_0)$$

$$+ \frac{1}{2} (\underset{\sim}{x} - \underset{\sim}{x}_0)' f''(\underset{\sim}{x}_0)(\underset{\sim}{x} - \underset{\sim}{x}_0)$$

... and beyond ... Taylor's theorem.

Inference:

- Get a sample from an unknown distribution (population).
- What does the sample tell you about the distribution

Example: Distribution: $N(\underbrace{\mu, \sigma^2}_{\text{unknown}})$

Sample $n = 10$

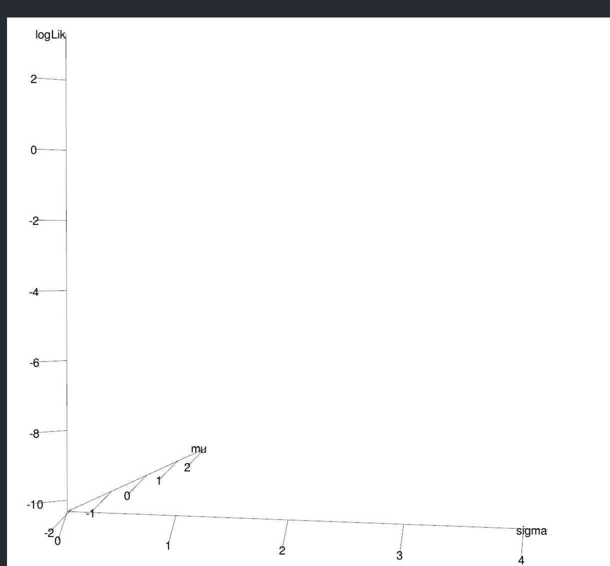
0.85887292 0.05047837 1.18787131 1.18794773 -0.15461645 -0.75116726 0.10407667 -0.26744312 0.10795396 1.44423848

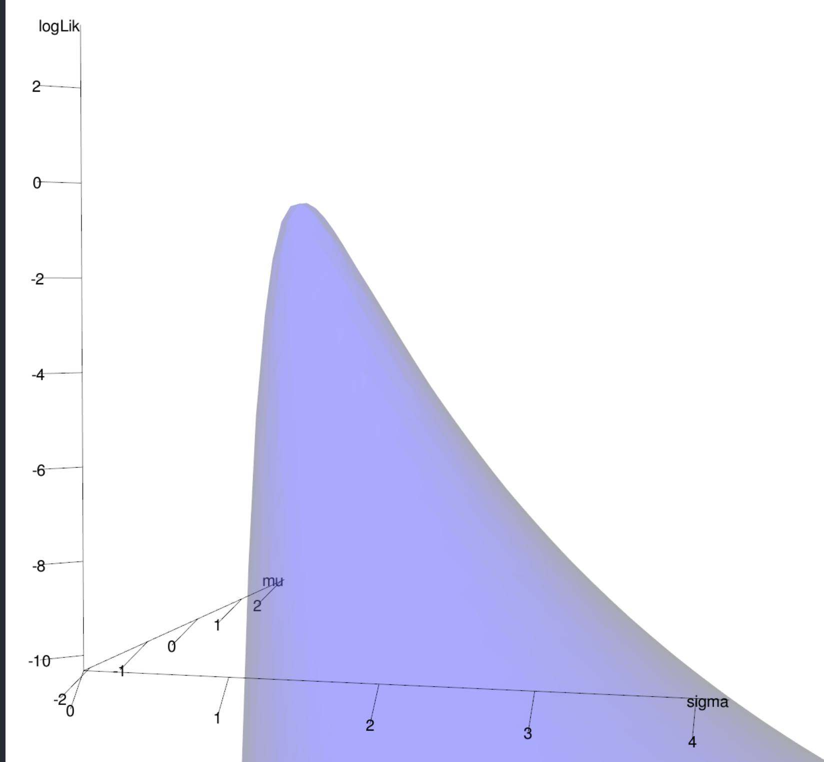
mean $= 0.3768$ sd $= 0.7386$

So...?

(log) likelihood function:

$\ln f(\underset{\sim}{x}; \mu, \sigma)$ as a function of μ + σ





What do we do with this???

- Find the summit?
- Slice it?
- Do a quadratic approximation?
- Consider a gradient at some "H_0"?
- Multiply it by a prior density?
  - and use Bayes Formula to get a posterior.
- Multiply it by a prior and use
  - MCMC
  - HMC

  } to simulate from the posterior.

Answer: $\boxed{}$

The great divide

Everyone uses the likelihood.

Bayesian

Combine it with a prior to create a posterior probability for parameters

Frequentist

Priors are "subjective"

- Posit one or more hypothetical "true" value.
- Consider the sampling variability of the likelihood under these hypothetical values.

Bayesian: Bayes formula

$$\pi(\underset{\sim}{\theta} \mid \underset{\sim}{x}_{obs}) = \frac{\overbrace{f(\underset{\sim}{x}_{obs} \mid \underset{\sim}{\theta})}^{\text{likelihood}} \; \overbrace{\pi(\underset{\sim}{\theta})}^{\text{prior}}}{\underbrace{\underbrace{\int \boxed{f(x_{obs} \mid \underset{\sim}{\theta}) \, \pi(\underset{\sim}{\theta})} \, d\underset{\sim}{\theta}}_{\text{norming "constant"}}}_{\text{FINDING THIS CAN BE HARD}}}$$

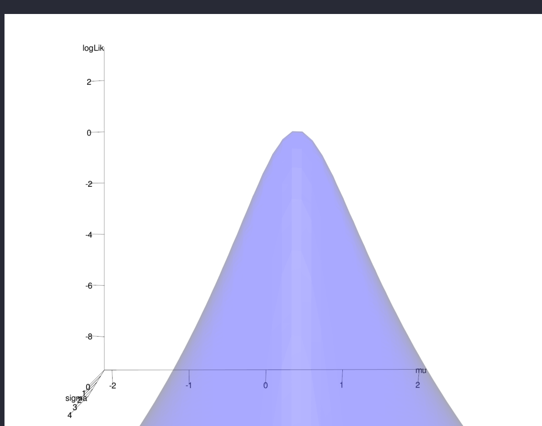
Some modern solutions:

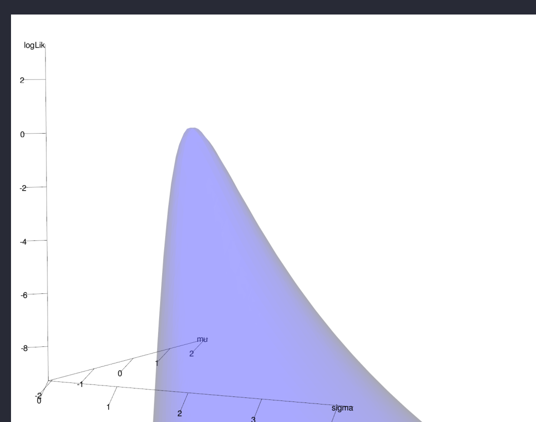
Use only $\boxed{f(x_{obs} \mid \underset{\sim}{\theta})\, \pi(\underset{\sim}{\theta})}$

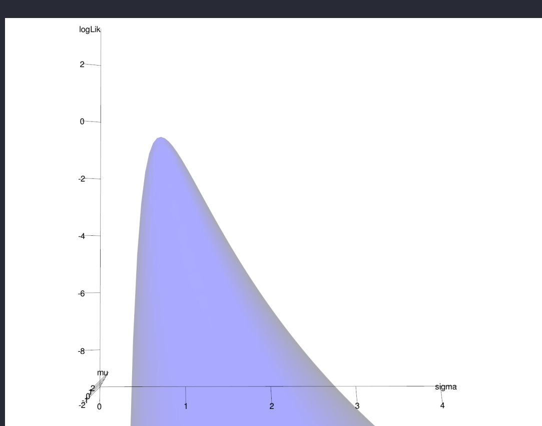
and MCMC or HMC

Frequentist methods

All based on sampling variation of likelihood.
i.e. What would happen if we were to resample over and over.







Here, we know from normal theory that max occurs at

$\mu = \bar{x}$, $\sigma = \hat{\sigma}$

$= 0.3768$ $= 0.7387$

How close to "true" value?

Three standard approaches

1) Likelihood Ratio Tests (Wilk's)
   Use random variability in height of log lik.
2) Wald Tests: Use quadratic approx. at MLE.
3) Fisher score (aka "Rao") test
   Use gradient at hypothetical value

$\boxed{\text{Wouldn't it be better to use the gradient at the MLE?}}$ Why? or Why not?

Likelihood Ratio = log Likelihood difference

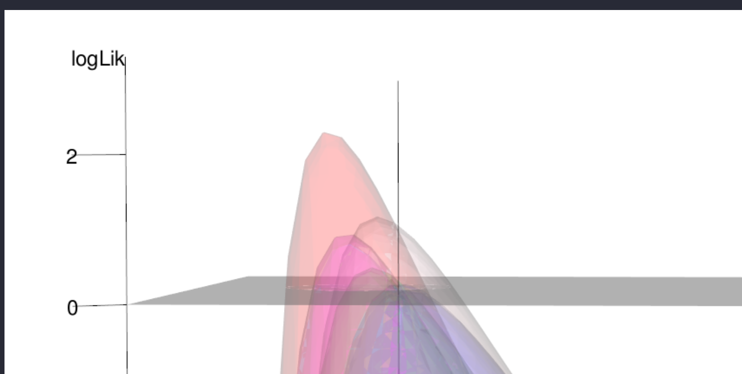
Take a hypothetical value, e.g. $\mu=0, \sigma=1$ and consider random variability in height of logLik. under repeated sampling.

Issue: height is arbitrary for a continuous distribution. So need to normalize by specifying height (doesn't matter which) at a choice of μ, σ.

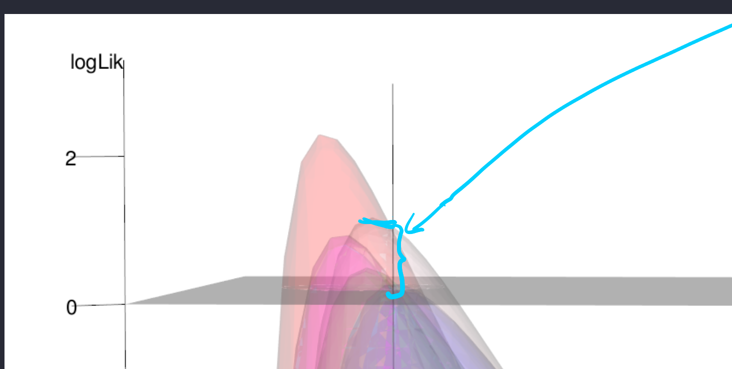
Common choice: height $= 0$ when μ, σ at hypothetical values.

Let's take a number of samples from N(0,1)

```
> set.seed(4567)
> sams <- lapply(rep(10,5), rnorm, mean = 0, sd = 1)
> sams
[[1]]
 [1] -0.7358250 -0.9025461  0.2524151  0.6150259  1.3545344  1.6023185  0.4432036  0.2114059 -0.3523372  1.5262018

[[2]]
 [1]  1.5186540  1.2311758 -0.3413057  0.3480193  0.1033780  0.3460563  0.1758544  0.9994781 -1.6544060 -1.4433098

[[3]]
 [1]  0.18435328 -0.18692255 -0.05300318  0.67699652  1.72128013  1.15726733 -0.02079913  0.30526505  0.01921551 -0.34847418

[[4]]
 [1]  1.72528724 -0.11417979 -0.47668452  1.17492289  0.87798073 -1.10289564 -0.94173811 -0.01198302  0.04414184  0.19493947

[[5]]
 [1] -0.7108700  0.3049154  0.6592707 -1.0953988  0.6362406  0.5217352  0.4086028 -0.5748284  1.3277637 -0.5658254
```



5 random log Likelihood mountains

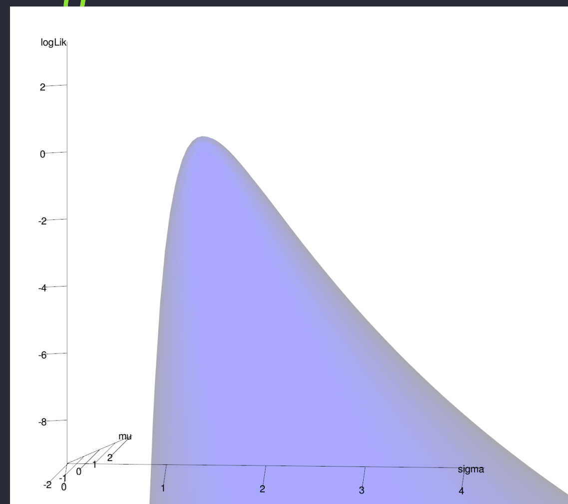


Summit altitude above 0 has asymptotic $\frac{1}{2}\chi^2_d$ distribution

d = # of parameters

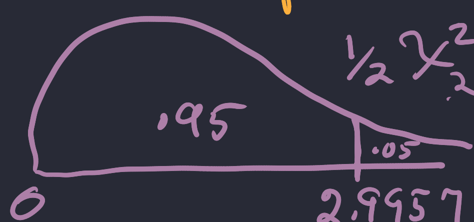
So back to original data.

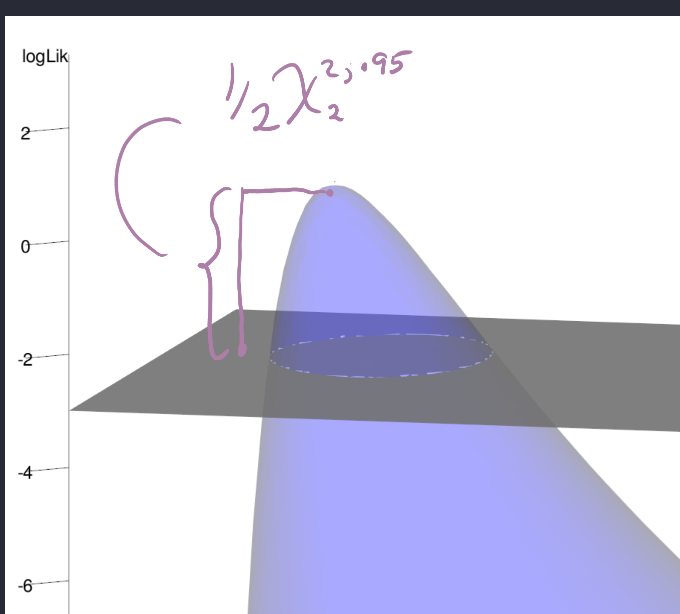
what to do with:

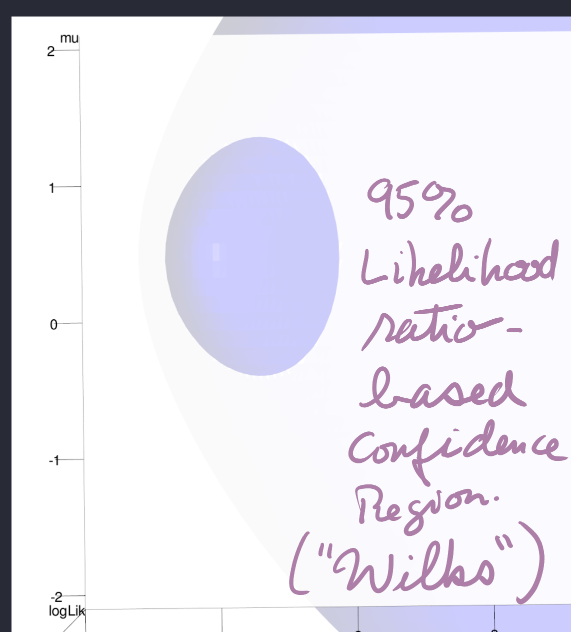


Answer: Invert sampling distribution

Take all points (μ, σ) such that the height of the summit is within $\frac{1}{2}\chi^{2,.95}_{2} = 2.9957$ of height of logLikelihood.



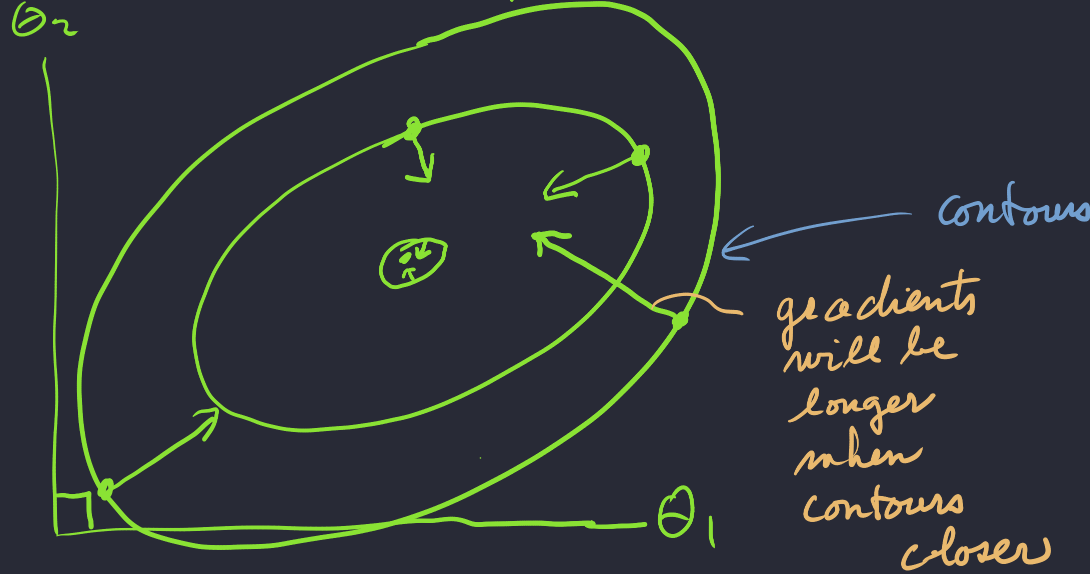


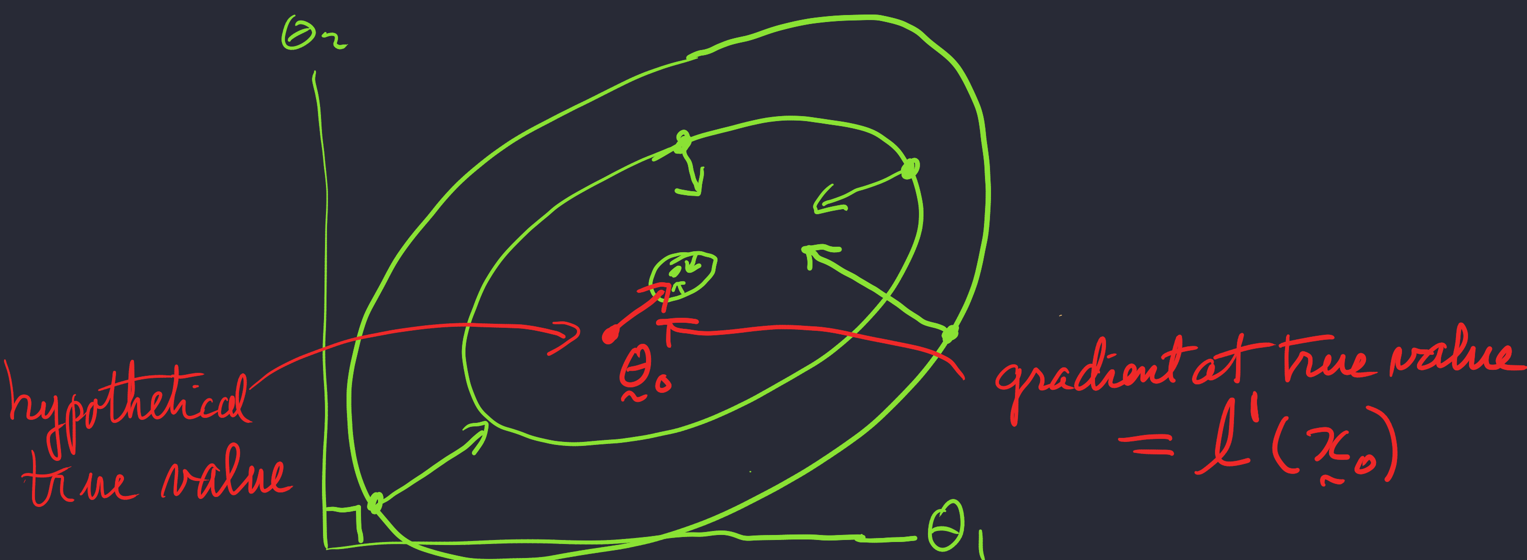


Wald: Same idea with quadratic appx at MLE [shape?]

Gradient = Score :

Looking at likelihood from above





Score Moments :

① $E_{\underset{\sim}{\theta}_0}\left(l'(\underset{\sim}{\theta}_0)\right) = \underset{\sim}{0}$

② $Var_{\underset{\sim}{\theta}_0}\left(l'(\underset{\sim}{\theta}_0)\right) = E_{\underset{\sim}{\theta}_0}\left(-l''(\underset{\sim}{\theta}_0)\right)$

$= I(\theta_0)$ "Fisher information"

Proof: Start with $\int f(y \mid \underset{\sim}{\theta})\, dy = 1$

Regularity assumptions:

- Fixed support
- $\frac{d}{d\underset{\sim}{\theta}}$ & $\int$ commute many times.

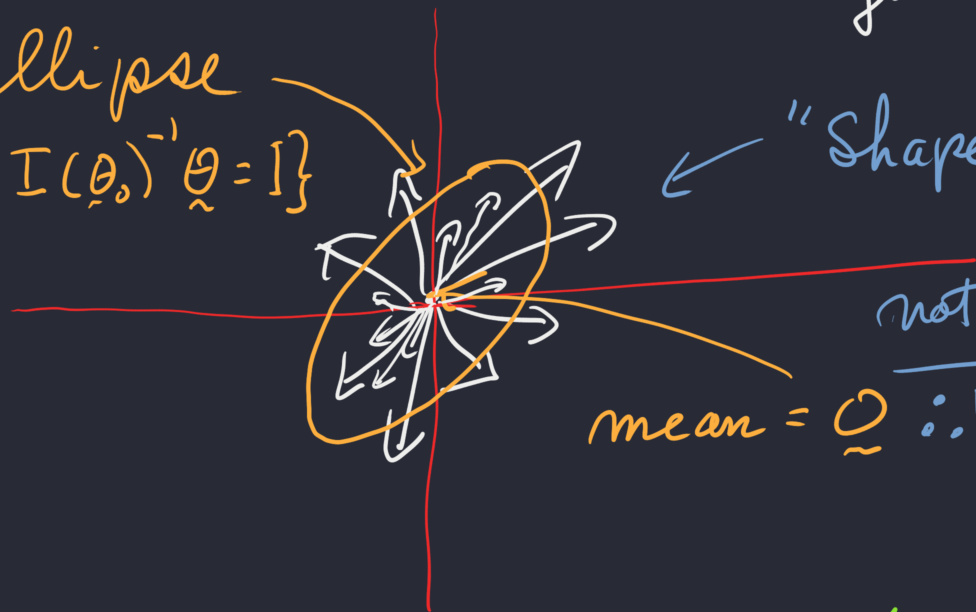
Distribution of gradients around $\underset{\sim}{\theta}_0$

given $\underset{\sim}{\theta}_0$ true value

SD ellipse

$\{\underset{\sim}{\theta} : \underset{\sim}{\theta}^T I(\underset{\sim}{\theta}_0)^{-1} \underset{\sim}{\theta} = 1\}$

"Shape" $\sqrt{I(\theta_0)}$



not asymptotic

mean = $\underset{\sim}{0}$ ∴ Very Important

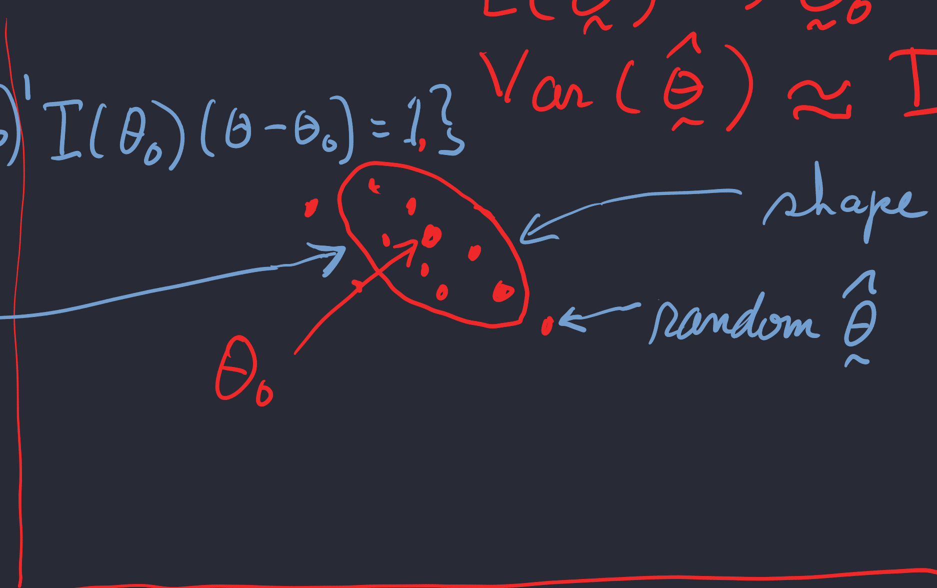
Asymptotic distribution of MLE $\hat{\underset{\sim}{\theta}}$

$E(\hat{\underset{\sim}{\theta}}) \longrightarrow \underset{\sim}{\theta}_0$

$Var(\hat{\underset{\sim}{\theta}}) \approx I^{-1}(\underset{\sim}{\theta}_0)$

$\{\underset{\sim}{\theta} : (\underset{\sim}{\theta} - \underset{\sim}{\theta}_0)' I(\theta_0)(\theta - \theta_0) = 1\}$

shape $\sqrt{I^{-1}(\theta_0)}$



random $\hat{\underset{\sim}{\theta}}$

θ_0

What if you are only interested in μ?

<u>Bayes</u>: "easy" use marginal posterior for μ

<u>Frequentist</u>: <u>Not always clear</u>

- e.g Behrens-Fisher problem
  Inference for $\mu_1 - \mu_2$ with 2 normal samples when you can't assume $\sigma_1 = \sigma_2$.
- Use additional principles to get e.g. t-test, etc.
- Use profile likelihood

Exponential families:

Special families that
— make things easy for
- Bayesian : conjugate priors
- Frequentist : finite-dimensional sufficient statistics as $n \uparrow$

Note: Likelihood is always sufficient.
(whole function, not necessarily MLE)

$n=1$ $\quad f(\underset{\sim}{x} \mid \underset{\sim}{\psi}) = h(\underset{\sim}{x})\, g(\underset{\sim}{\psi}) \exp\left(\sum_{i=1}^{k} \underbrace{\theta_i(\psi)}_{\text{canonical parameter}} T_i(\underset{\sim}{x})\right)$

n i.i.d

$$f(x_1, \dots, x_n \mid \underset{\sim}{\psi}) = \prod_{i=1}^{n} h(x_i)\, g^n(\psi)$$

$$\times \exp\left\{\sum_{i=1}^{k} \theta_i(\underset{\sim}{\psi}) \underbrace{\left(\sum_{j=1}^{n} T_j(\underset{\sim}{x})\right)}_{\text{suff. stat.}}\right\}$$

Redo using exponential form

Example: $N(\mu, \sigma)$

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2} \cdot x^2 + \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} \mu^2\right\}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi}}}_{h} \underbrace{\frac{1}{\sigma}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}}_{g} \times \exp\left\{\underbrace{\frac{1}{\sigma^2}\times\left(-\frac{x^2}{2}\right)}_{\theta_1 \times T_1} + \underbrace{\frac{\mu}{\sigma^2}x}_{\theta_2 \times T_2}\right\}$$

Express g as a function of $\theta_1 = \frac{1}{\sigma^2}$ $\theta_2 = \frac{\mu}{\sigma^2}$

Now: $\sigma^2 = \frac{1}{\theta_1}$ $\mu = \sigma^2\theta_2 = \theta_2/\theta_1$

So
$$k(\underset{\sim}{\theta}) = \frac{1}{\sigma}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}$$
$$= \sqrt{\theta_1}\exp\left\{-\frac{1}{2}\frac{\theta_2^2}{\theta_1^2}\times\theta_1\right\}$$
$$= \sqrt{\theta_1}\exp\left\{-\frac{1}{2}\theta_2^2/\theta_1\right\}$$

Then
$$E\left(\begin{pmatrix}T_1\\T_2\end{pmatrix}\middle|\,\underset{\sim}{\theta}\right) = K'(\theta)$$
$$\mathrm{Var}\left(\begin{pmatrix}T_1\\T_2\end{pmatrix}\middle|\,\underset{\sim}{\theta}\right) = K''(\theta)$$

$K^{(n)}(\theta)$ is cumulant generating function.

<u>EXER:</u> <u>Check my math</u>

Easier formulation of Exponential families

Density: Single observation $\underset{\sim}{x}$.

$$f_X(\underset{\sim}{x} \mid \underset{\sim}{\theta}) = \exp\left\{ \underset{\sim}{\eta}(\underset{\sim}{\theta}) \cdot \underset{\sim}{T}(\underset{\sim}{x}) - A(\underset{\sim}{\theta}) + B(x) \right\}$$

Support of f must not depend on $\underset{\sim}{\theta}$

e.g. $U(\theta, \theta+1)$ is not an exponential family

IID Sample $\underset{\sim}{x}_1, \ldots, x_n$

$$\prod_{i=1}^{n} f_X(\underset{\sim}{x}_i \mid \underset{\sim}{\theta}) = \exp\left\{ \underset{\sim}{\eta}(\underset{\sim}{\theta}) \cdot \sum_{i=1}^{n} \underset{\sim}{T}(\underset{\sim}{x}_i) - nA(\underset{\sim}{\theta}) + \sum_{i=1}^{n} B(\underset{\sim}{x}_i) \right\}$$

Same form as above

Sufficiency: $\sum_{i=1}^{n} \underset{\sim}{T}(\underset{\sim}{x}_i)$ is sufficient and the sufficient statistic has dimension equal to the dimension of $\underset{\sim}{T}$.

- η is canonical parameter
- $K(\underset{\sim}{\eta}) = A(\underbrace{\underset{\sim}{\theta}(\underset{\sim}{\eta})}_{\text{inverse of } \underset{\sim}{\eta}(\underset{\sim}{\theta})})$ is the cumulant fcn

$$K'(\underset{\sim}{\eta}) = E_{\underset{\sim}{\eta}}(\underset{\sim}{T}) \qquad K''(\underset{\sim}{\eta}) = Var_{\underset{\sim}{\eta}}(\underset{\sim}{T})$$

$N(\mu, \sigma)$

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2) - \frac{1}{2}\log(2\pi\sigma^2)\right\}$$

$$= \exp\left\{\underbrace{-\frac{1}{2\sigma^2}\cdot x^2 + \frac{\mu}{\sigma^2}\cdot x}_{\eta(\underset{-}{\theta})\cdot \underset{\sim}{T}(x)} \underbrace{- \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2}_{-A(\underset{\sim}{\theta})} \underbrace{- \frac{1}{2}\log(2\pi)}_{\substack{B(x) \\ = \text{constant}}}\right\}$$

$$\eta_1 = -\frac{1}{2\sigma^2} \quad \eta_2 = \frac{\mu}{\sigma^2}$$

$$T_1(x) = x^2 \quad T_2(x) = x$$

To get K, express $A(\theta)$ as a function of η

$$\sigma^2 = \frac{-1}{2\eta_1}$$

(note: $\eta_1 < 0$)

$$\mu = \eta_2 \sigma^2 = -\frac{\eta_2}{2\eta_1}$$

So $A = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2$

$$= \frac{\eta_2^2 / 4\eta_1^2}{-1/\eta_1} + \frac{1}{2}\log\left(\frac{-1}{2\eta_1}\right)$$

$$= -\frac{\eta_2^2}{4\eta_1} - \frac{1}{2}\log(-2\eta_1)$$

$$A'(\underset{\sim}{\eta}) = \begin{pmatrix} \frac{\eta_2^2}{4\eta_1^2} - \frac{1}{2} \times \frac{1}{-2\eta_1} \times -2 \\ -\frac{2\eta_2}{4\eta_1} \end{pmatrix} = \begin{pmatrix} \mu^2 + \sigma^2 \\ \mu \end{pmatrix}$$

Pheew...

<u>The proof is much easier than the example</u>

Suppose A is expressed in terms of η

$$f(\underset{\sim}{x} \mid \underset{\sim}{\eta}) = \exp\{\underset{\sim}{\eta} \cdot T(x) - A(\underset{\sim}{\eta}) + B(x)\}$$

<u>Moment-generating function:</u>

$$m_{\underset{\sim}{\eta}}(\underset{\sim}{t}) = E_{\underset{\sim}{\eta}}(e^{\underset{\sim}{t} \cdot \underset{\sim}{T}})$$

$$= \int e^{\underset{\sim}{t} \cdot \underset{\sim}{T}} \, e^{\underset{\sim}{\eta} \cdot \underset{\sim}{T} - A(\underset{\sim}{\eta}) + B(x)} \, d\mu(\underset{\sim}{x})$$

density for $\underset{\sim}{\eta} + \underset{\sim}{t}$

$$= \int \exp\{(\underset{\sim}{t} + \underset{\sim}{\eta}) \cdot T - A(\underset{\sim}{\eta} + \underset{\sim}{t}) + A(\underset{\sim}{\eta} + \underset{\sim}{t}) - A(\underset{\sim}{\eta}) + B(\underset{\sim}{x})\} \, d\mu$$

$$= \exp\{A(\underset{\sim}{\eta}+\underset{\sim}{t}) - A(\underset{\sim}{\eta})\}$$

<u>cumulant generating function</u>

$$K_{\underset{\sim}{\eta}}(\underset{\sim}{t}) = \log m_{\underset{\sim}{\eta}}(\underset{\sim}{t})$$

$$= A(\underset{\sim}{\eta}+\underset{\sim}{t}) - A(\underset{\sim}{\eta})$$

$$E_{\underset{\sim}{\eta}}(\underset{\sim}{T}) = K'_{\underset{\sim}{\eta}}(\underset{\sim}{t})\Big|_{\underset{\sim}{t}=\underset{\sim}{0}} = \frac{\partial}{\partial \underset{\sim}{t}}\left(A(\eta+t) - A(\eta)\right)\Big|_{\underset{\sim}{t}=\underset{\sim}{0}}$$

$$= A'(\underset{\sim}{\eta})$$

$$Var_{\underset{\sim}{\eta}}(T) = K''(\underset{\sim}{\eta})$$