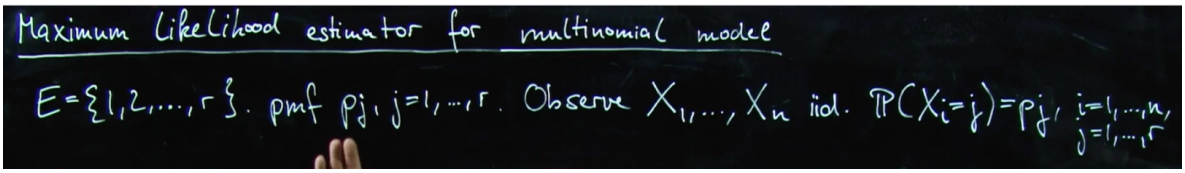
Recitation: maximum likelihood estimator for multinomial

Recitation problem statement

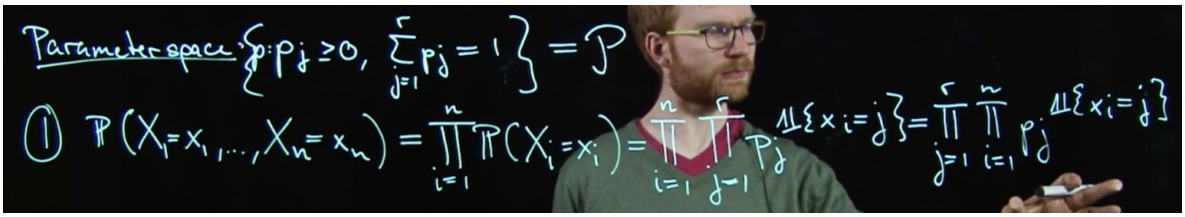
Consider a finite space $E = \{a_1, a_2, \dots, a_r\}$ of size $r \geq 2$ and let X be a random variable taking values in E. For $j = 1, \dots, r$, let $p_j^* = \mathbf{P}[X = a_j]$. Consider a sample of n i.i.d. copies $X_1, \dots, X_n$ of X. Based on this sample, we would like to estimate the multivariate parameter $p^* = (p_1^*, \dots, p_r^*)$.

1. What is the parameter space ?
2. Write the likelihood associated with the model described above.
3. Compute the maximum likelihood estimator $\hat{p}$ of p^*.
4. Using the central limit theorem, show that $\hat{p}$ is asymptotically normal. Compute the asymptotic covariance matrix. Denote it by Σ.
5. Prove that Σ is not invertible. Conclude that the theorem for the MLE could not have been applied here. What condition is not satisfied ?

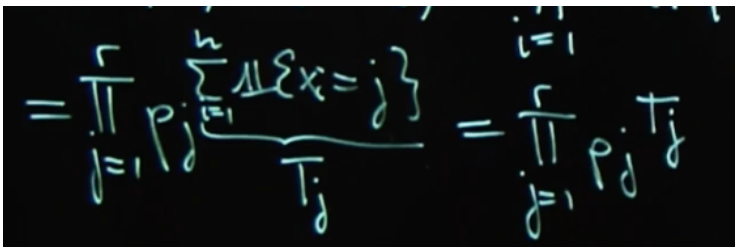


- binomial is a case of multinomial where r=2

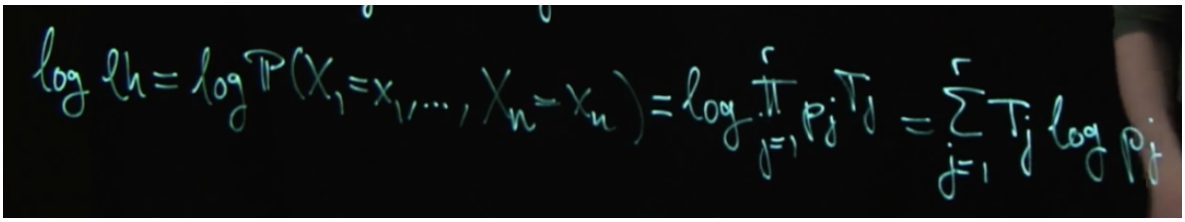
X_i must all equal one of the values of E (1 to r)



- indicator function 1(x_i = j) means that if x_i = j then pj will remain, otherwise pj = 1 due to exponent



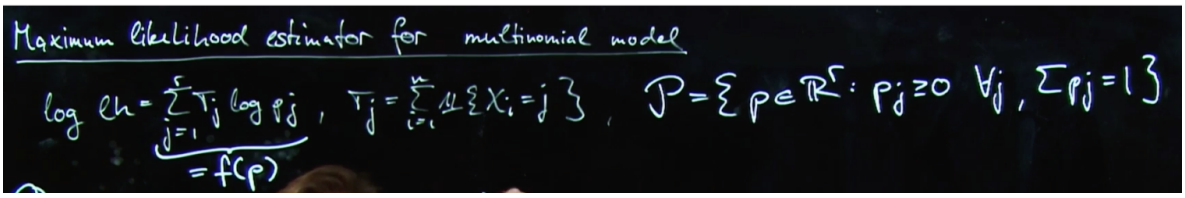
- T_j means how many times have we observed each element j
- in binomial this would be number of heads and tails in coin flip example

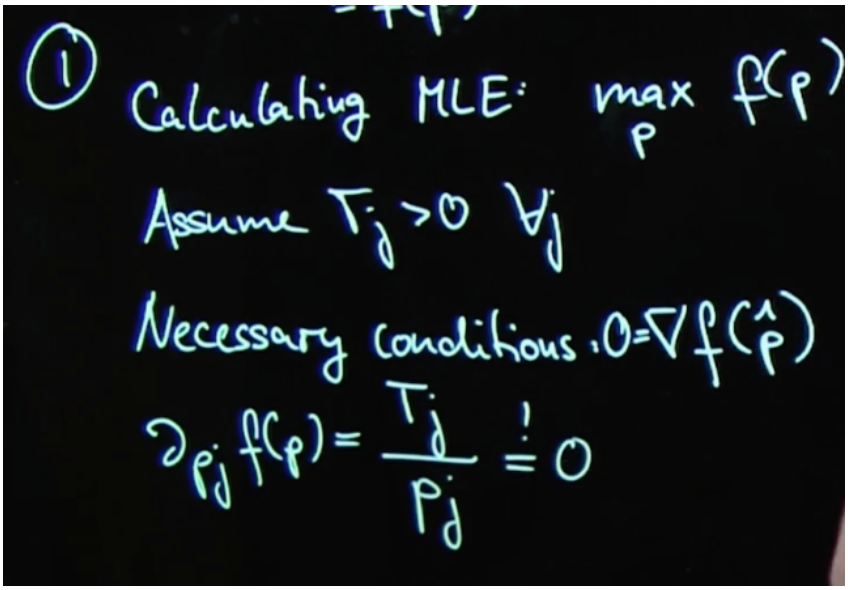


- log of product becomes a sum and the T_j can be pulled outside the log
-

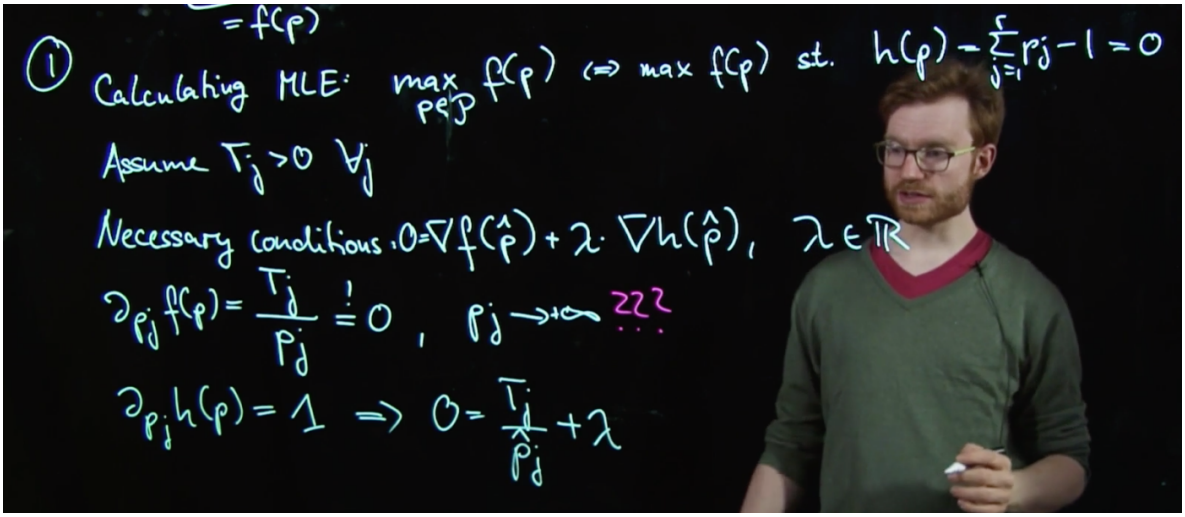
Compute MLE

- take the likelihood given the parameters,
- plug in observations
- maximise wrt the parameters (pj in this case)

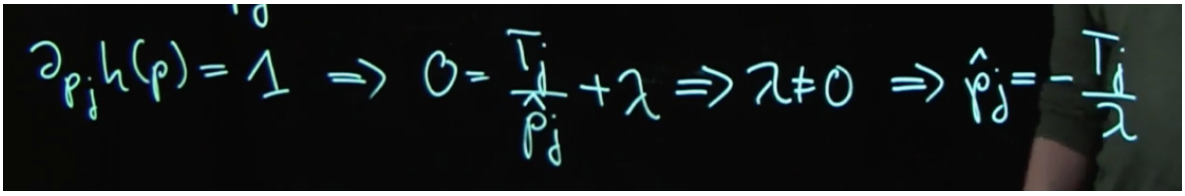


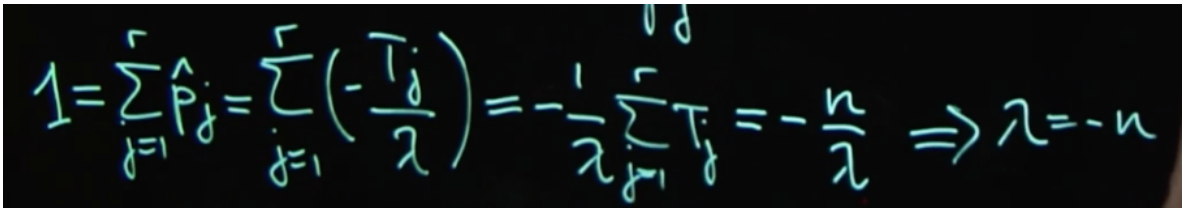


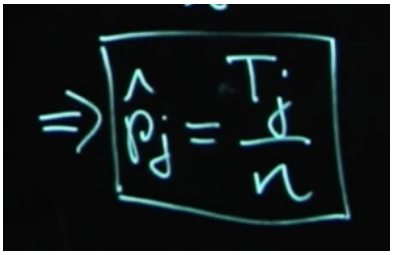
- p_hat is the optimum we want to find
- if p_j is strictly positive then this derivative will never become 0, only approach it as pj -> infinity
- but they have to lie between 0 and 1, so it is a constrained optimisation problem, not unconstrained



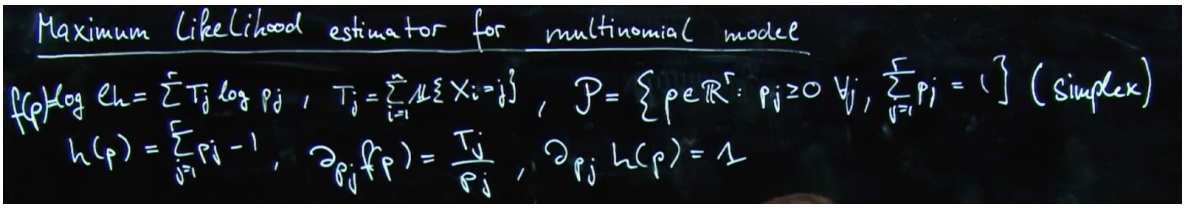
- add in additional h() function, which the derivative gives us lambda as an additional term
- using Lagrange multiplier
- there is only one lambda for all dimensions

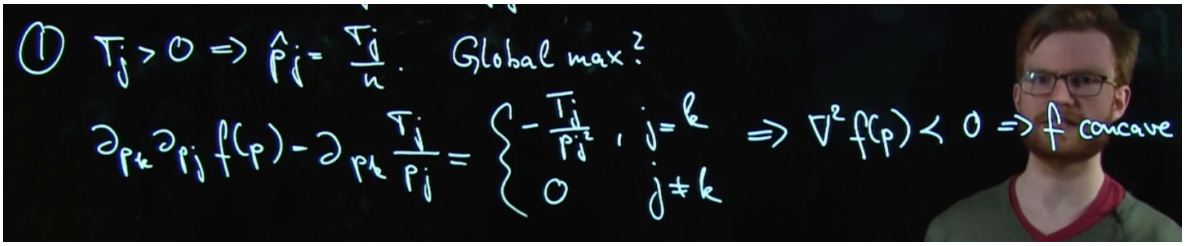




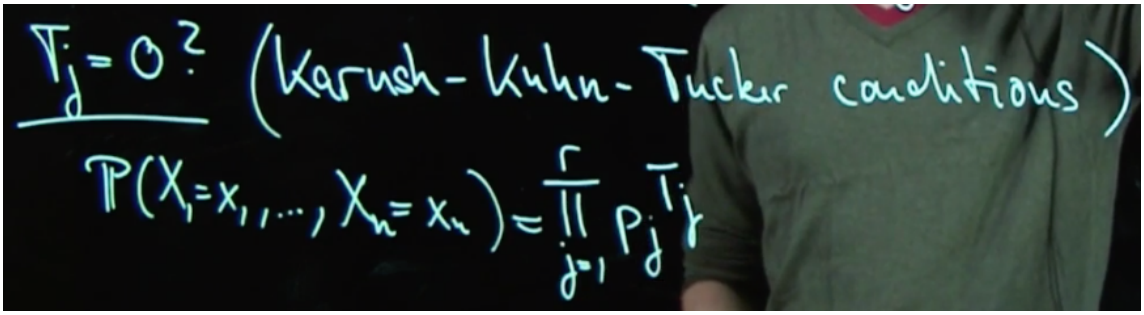


- this is just the frequency of each element
- this is assuming none of the T_j are 0



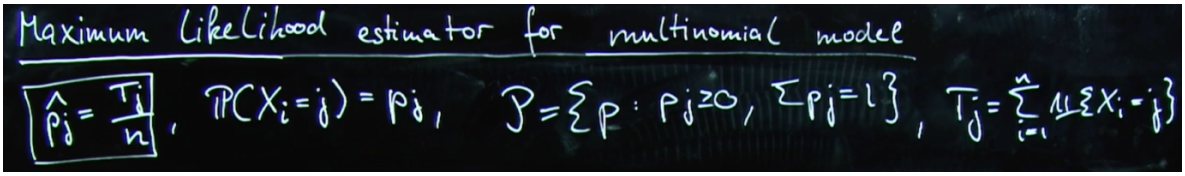


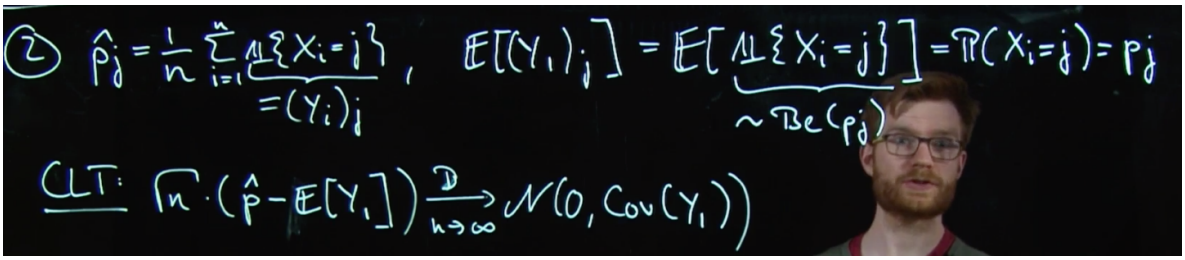
- the log likelihood is concave so we know that p_hat is a global maximum
- the Hessian is strictly concave



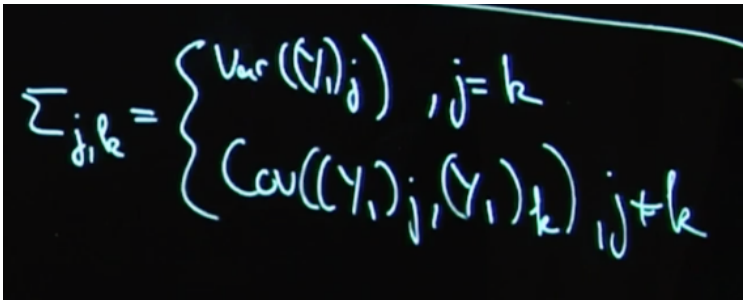
- if T_j is 0 then it will not enter the product
- gives same result T_j / n

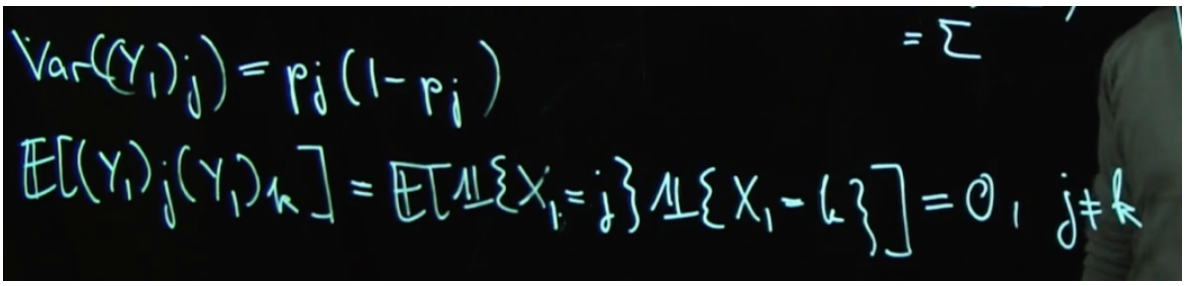
Asymptotic variance



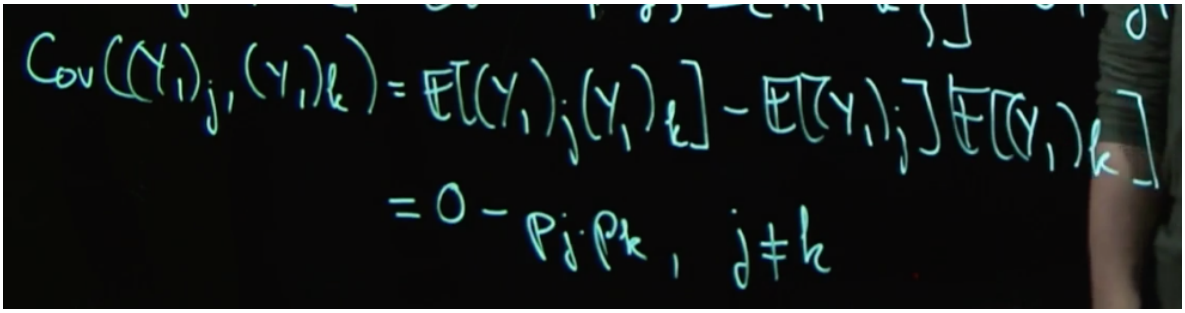


- can apply CLT as Y is an average of iid random variable and becomes Bernoulli with parameter pj

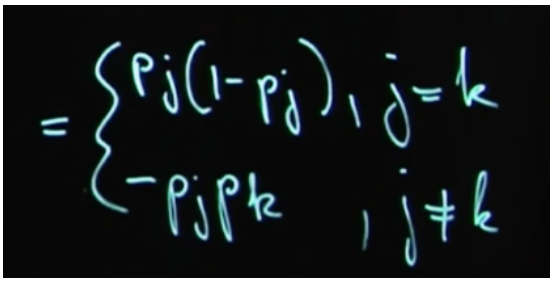




- 0 because one of the identity functions will always be 0 since j does not equal k

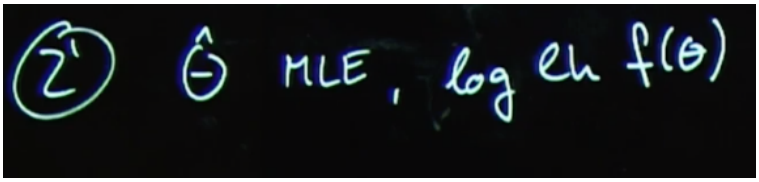


covariance matrix

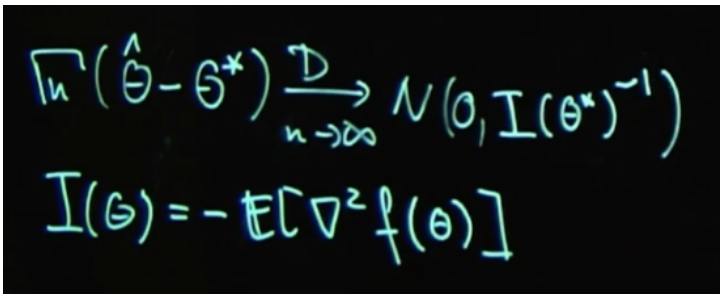


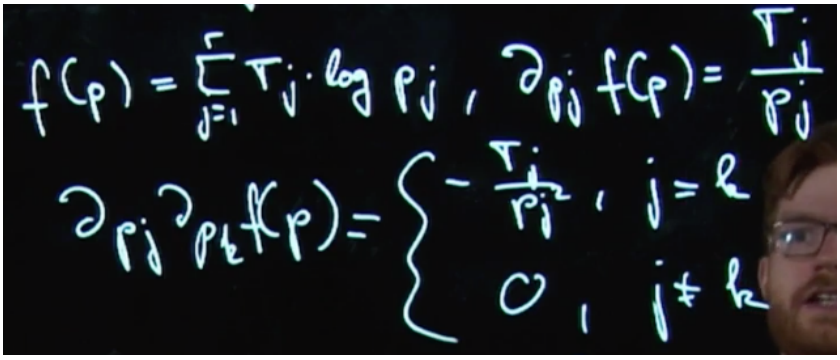
- pj(1 - pj) on the diagonal where j=k
- negative on the off-diagonals
- negative shows an anti-correlation since when j is big k is small (similar to the identity functions)

Another method of calculating asymptotic variance

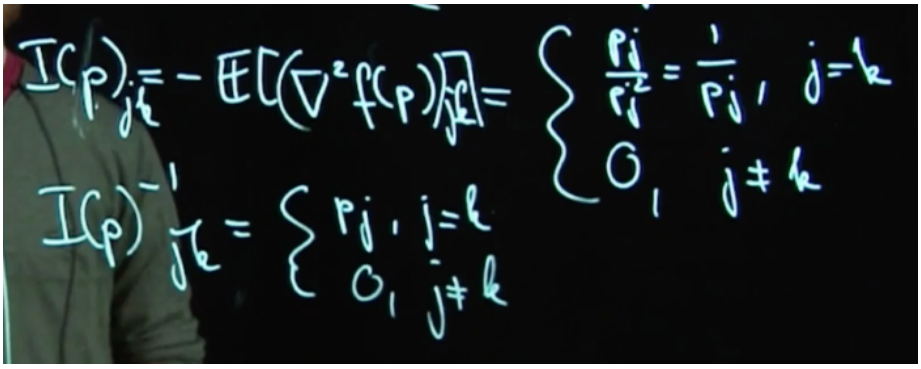


- think of theta as p here

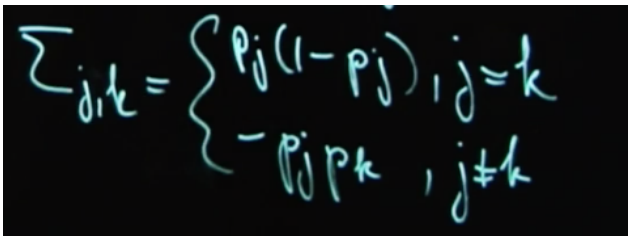




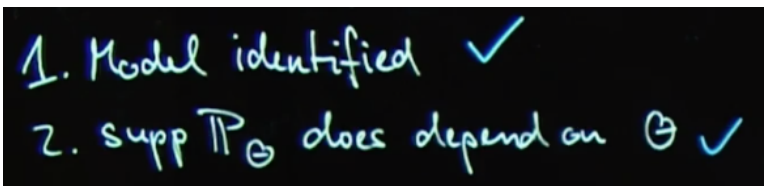
- the Hessian is a diagonal matrix as shown here



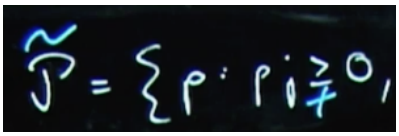
- the inverse Fisher Information matrix does not look like the covariance matrix as calculated before (below)

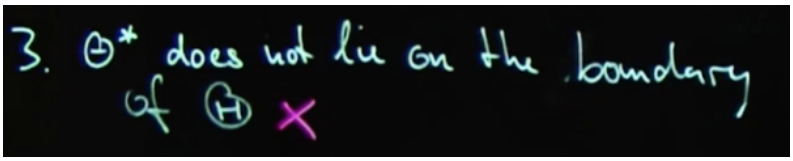


Explanation
some assumptions were not satisfied

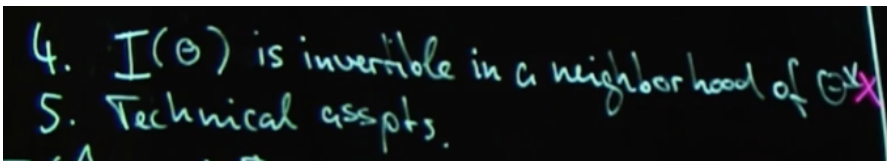


get second assumption by assuming new p_tilda (below) that is strictly above 0

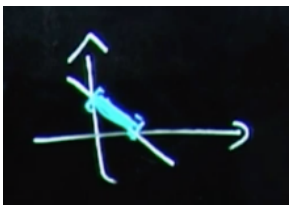




- bounded by linearity, this is a problem?

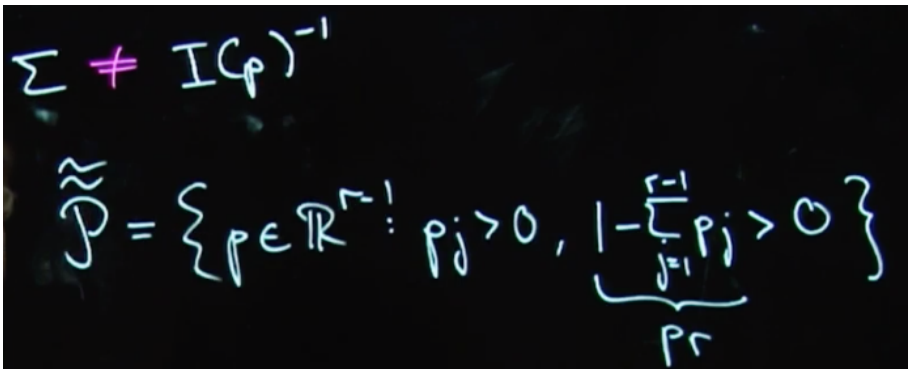


- technical assumptions is continuously differentiable



- since these assumptions aren't fulfilled then covariance is not equal to Fisher Information

Can get around this



- parameterise the set differently with 1 less parameter