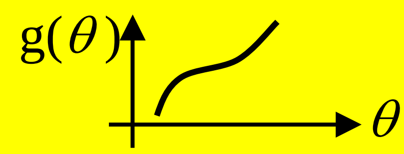
7.6 MLE for Transformed Parameters

Given PDF $p(\mathbf{x};\theta)$ but want an estimate of $\alpha = g(\theta)$

What is the MLE for α??

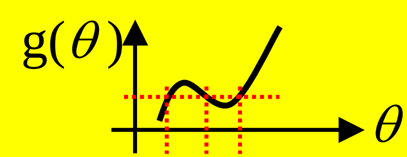
Two cases:

1. $\alpha = g(\theta)$ is a one-to-one function



$$\hat{\alpha}_{ML} \text{ maximizes } p(\mathbf{x};g^{-1}(\alpha))$$

2. $\alpha = g(\theta)$ is not a one-to-one function



Need to define modified likelihood function:

$$\overline{p}_T(\mathbf{x};\alpha) = \underbrace{\max_{\{\theta:\,\alpha=g(\theta)\}} p(\mathbf{x};\theta)}$$

- For each α, find all θ's that map to it
- Extract largest value of $p(\mathbf{x};\theta)$ over this set of θ's

$$\hat{\alpha}_{ML} \text{ maximizes } \overline{p}_T(\mathbf{x};\alpha)$$

<u>Invariance Property of MLE</u>

Another Big Advantage of MLE!

<u>Theorem 7.2: Invariance Property of MLE</u>

If parameter θ is mapped according to $\alpha = g(\theta)$ then the MLE of α is given by

$$\boxed{\hat{\alpha} = g(\hat{\theta})}$$

where $\hat{\theta}$ is the MLE for θ found by maximizing $p(\mathbf{x};\theta)$

Note: when $g(\theta)$ is not one-to-one the MLE for α maximizes the <u>modified</u> likelihood function

"Proof":

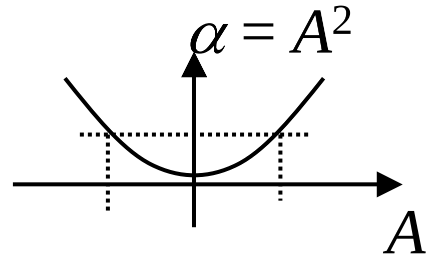
Easy to see when $g(\theta)$ is one-to-one

Otherwise… can "argue" that maximization over θ inside definition for modified LF ensures the result.

Ex. 7.9: Estimate Power of DC Level in AWGN

$x[n] = A + w[n]$ noise is $N(0,\sigma^2)$ & White

Want to Est. Power: $\alpha = A^2 \Rightarrow$



$\Rightarrow$ For each α value there are 2 PDF's to consider

$$p_{T_1}(\mathbf{x};\alpha) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_n (x[n] - \sqrt{\alpha})^2 \right]$$

$$p_{T_2}(\mathbf{x};\alpha) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_n (x[n] + \sqrt{\alpha})^2 \right]$$

Then:
$$\hat{\alpha}_{ML} = \left[\arg \max_{\sqrt{\alpha} \geq 0} \left\{ p(\mathrm{x};\sqrt{\alpha}),\, p(\underline{x};-\sqrt{\alpha}) \right\} \right]^2$$
$$= \left[\arg \max_{-\infty < A < \infty} p(\mathrm{x};A) \right]^2$$
$$= \left[\hat{A}_{ML} \right]^2$$

Demonstration that Invariance Result Holds for this Example

Ex. 7.10: Estimate Power of WGN in dB

$x[n] = w[n]$ WGN w/ var = σ^2 unknown

Recall: $P_{\text{noise}} = \sigma^2$

Can show that the MLE for variance is: $\hat{P}_{noise} = \frac{1}{N}\sum_{n=0}^{N-1} x^2[n]$

Note: You may recall a result for estimating variance that divides by N–1 rather than by N … that estimator is unbiased, this estimate is biased (but asymptotically unbiased)

To get the dB version of the power estimate:

$$\hat{P}_{dB} = \underbrace{10\log_{10}\left[\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]\right]}_{\text{Using Invariance Property !}}$$

7.7: Numerical Determination of MLE

Note: In all previous examples we ended up with a closed-form expression for the MLE: $\hat{\theta}_{ML} = f(\mathbf{x})$

Ex. 7.11: $x[n] = r^n + w[n]$ noise is $N(0,\sigma^2)$ & white

Estimate *r*

If $-1 < r < 0$ then this signal is a decaying oscillation that might be used to model:
- A Ship's "Hull Ping"
- A Vibrating String, Etc.

To find MLE:

$$\frac{\partial \ln p(\mathrm{x};\theta)}{\partial \theta} = 0$$

$$\Rightarrow \sum_{n=0}^{N-1}(x[n] - r^n)nr^{n-1} = 0$$

No closed-form solution for the MLE

So…we can't <u>always</u> find a closed-form MLE!

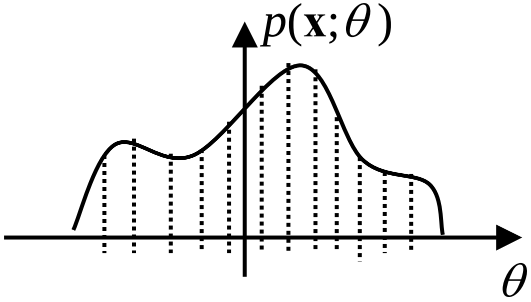
But a main advantage of MLE is:

We can <u>always</u> find it numerically!!!

(<u>Not always *computationally* efficiently, though</u>)

<u>Brute Force Method</u>

Compute $p(\mathbf{x};\theta)$ on a fine grid of θ values



<u>Advantage</u>: Sure to Find maximum

(if grid is fine enough)

<u>Disadvantage</u>: <u>Lots</u> of Computation

(especially w/ a fine grid)

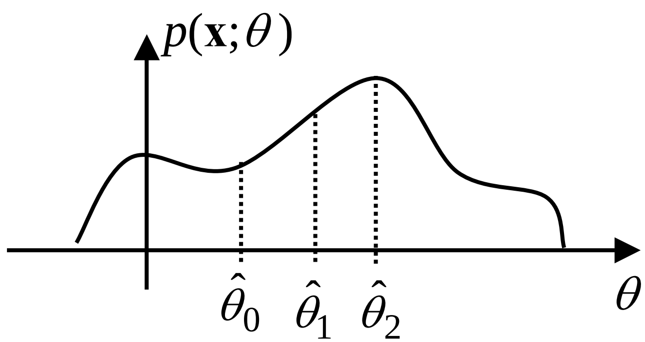
Iterative Methods for Numerical MLE

Step #1: Pick some "initial estimate" $\hat{\theta}_0$

Step #2: Iteratively improve it using

$$\hat{\theta}_{i+1} = f(\hat{\theta}_i, \mathbf{x}) \qquad \text{such that} \quad \lim_{i \to \infty} p(\mathrm{x}\,;\theta_i) = \max_{\theta} p(\mathrm{x};\theta)$$

"Hill Climbing in the Fog"



Note: A so-called "Greedy" maximization algorithm will always move up even though taking an occasional step downward may be the better global strategy!

Convergence Issues:

1. May not converge
2. May converge, but to local maximum
   - good initial guess is needed !!
   - can use rough grid search to initialize
   - can use multiple initializations

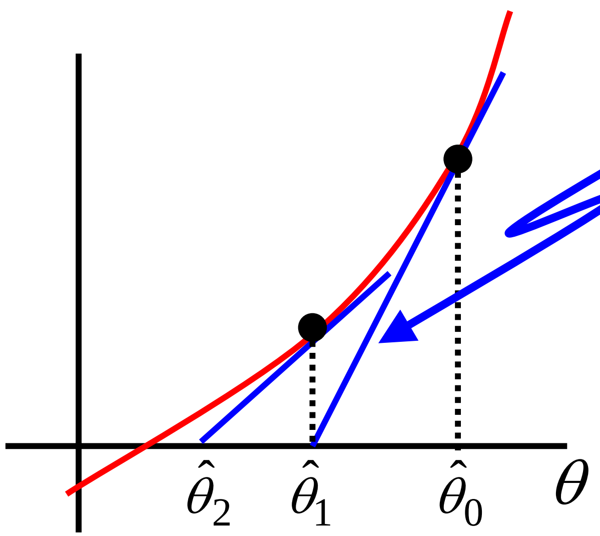
Iterative Method: Newton-Raphson MLE

The MLE is the maximum of the LF… so set derivative to 0:

$$\underbrace{\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}}_{\triangleq g(\theta)} = 0$$

So… MLE is a zero of $g(\theta)$

Newton-Raphson is a numerical method for finding the zero of a function… so it can be applied here… **Linearize $g(\theta)$**



$$g(\theta) \approx \underbrace{g(\theta_k) + \left[\left. \frac{dg(\theta)}{d\theta} \right|_{\theta=\hat{\theta}_k} \right] (\theta - \hat{\theta}_k)}_{\substack{set=0 \\ solve\ for\ \hat{\theta}_{k+1}}}$$

Truncated Taylor Series

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \frac{g(\hat{\theta}_k)}{\left[\left. \frac{dg(\theta)}{d\theta} \right|_{\theta=\hat{\theta}_k} \right]}$$

Now… using our "definition of convenience": $g(\theta) = \dfrac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}$

So then the Newton-Raphson MLE iteration is:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \left\{ \left[\underbrace{\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}}_{} \right]^{-1} \frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} \right\} \Bigg|_{\theta = \hat{\theta}_k}$$

Look Familiar???
Looks like $I(\theta)$, except: $I(\theta)$ is evaluated at the true θ, and has an expected value

Iterate until convergence criterion is met:

$$|\hat{\theta}_{k+1} - \hat{\theta}_k| < \varepsilon$$

You get to choose!

Generally:

For a given PDF model, compute derivatives analytically…

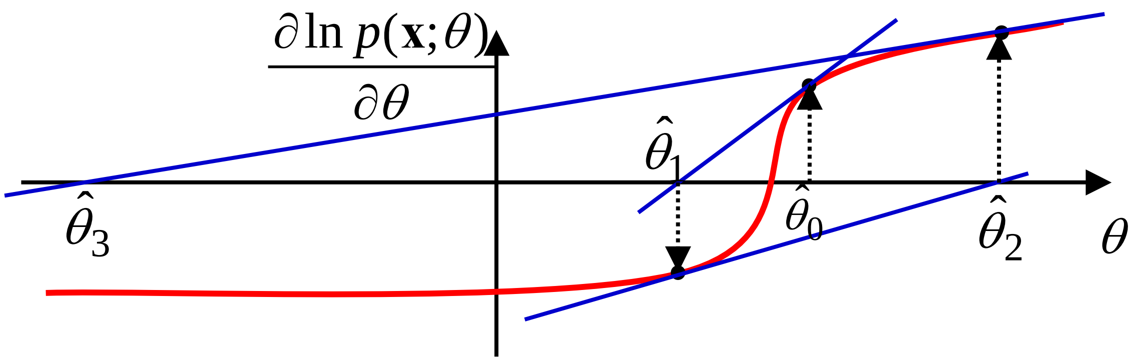
or… compute derivatives numerically:

$$\left. \frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} \right|_{\hat{\theta}_k} \approx \frac{\ln p(\mathbf{x};\hat{\theta}_k + \Delta\theta) - \ln p(\mathbf{x};\hat{\theta}_k)}{\Delta\theta}$$

Convergence Issues of Newton-Raphson:

1. May not converge
2. May converge, but to local maximum

   - good initial guess is needed !!
   - can use rough grid search to initialize
   - can use multiple initializations



Some Other Iterative MLE Methods

1. Scoring Method
   - Replaces second-partial term by $I(\theta)$
2. Expectation-Maximization (EM) Method
   - Guarantees convergence to at least a local maximum
   - Good for complicated multi-parameter cases

7.8 MLE for Vector Parameter

Another nice property of MLE is how easily it carries over to the vector parameter case.

The vector parameter is: $\boldsymbol{\theta} = \begin{bmatrix}\theta_1 & \theta_2 & \cdots & \theta_p\end{bmatrix}^T$

$\hat{\boldsymbol{\theta}}_{ML}$ is the vector that satisfies: $\underbrace{\dfrac{\partial \ln p(\mathrm{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}} = 0$

Derivative w.r.t. a vector

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \dfrac{\partial f(\boldsymbol{\theta})}{\partial \theta_1} \\ \dfrac{\partial f(\boldsymbol{\theta})}{\partial \theta_2} \\ \vdots \\ \dfrac{\partial f(\boldsymbol{\theta})}{\partial \theta_p} \end{bmatrix}$$

Ex. 7.12: Estimate DC Level and Variance

$x[n] = A + w[n]$ noise is $N(0,\sigma^2)$ and white

Estimate: DC level A and Noise Variance σ^2 $\Rightarrow$ $\boldsymbol{\theta} = \begin{bmatrix} A \\ \sigma^2 \end{bmatrix}$

LF is: $$p(\mathbf{x}; A, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} [x[n] - A]^2 \right\}$$

Solve: $$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \stackrel{set}{=} \mathbf{0}$$

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} \big(x[n] - A\big) = \frac{N}{\sigma^2}\big(\bar{x} - A\big) = 0$$

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=0}^{N-1} \big(x[n] - A\big)^2 = 0$$

$$\boxed{\hat{\theta}_{ML} = \begin{bmatrix} \bar{x} \\ \frac{1}{N} \sum_n (x[n] - \bar{x})^2 \end{bmatrix}}$$

Interesting: For this problem…

First estimate A just like scalar case

The subtract it off and then estimate variance like scalar case

Properties of Vector ML

The **asymptotic properties** are captured in **Theorem 7.3**:

If $p(\mathbf{x};\theta)$ satisfies some "regularity" conditions, then the MLE is asymptotically distributed according to

$$\boxed{\hat{\boldsymbol{\theta}}_{ML} \overset{a}{\sim} N(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta}))}$$

where $\mathbf{I}(\theta)$ = Fisher Information Matrix

So the vector ML is asymptotically:

- unbiased
- efficient

Invariance Property Holds for Vector Case

If $\alpha = g(\theta)$, then $\hat{\boldsymbol{\alpha}}_{ML} = g(\hat{\boldsymbol{\theta}}_{ML})$

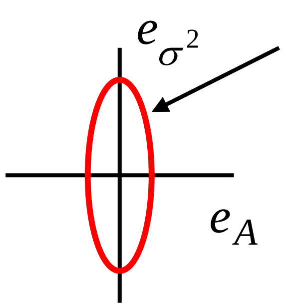
Ex. 7.12 Revisited

It can be shown that:
$$E\{\hat{\boldsymbol{\theta}}\} = \begin{bmatrix} A \\ \dfrac{(N-1)}{N}\sigma^2 \end{bmatrix} \qquad \operatorname{cov}\{\hat{\boldsymbol{\theta}}\} = \begin{bmatrix} \dfrac{\sigma^2}{N} & 0 \\ 0 & \dfrac{2(N-1)}{N^2}\sigma^4 \end{bmatrix}$$

For large N then :
$$E\{\hat{\boldsymbol{\theta}}\} \approx \begin{bmatrix} A \\ \sigma^2 \end{bmatrix} = \boldsymbol{\theta} \qquad \operatorname{cov}\{\hat{\boldsymbol{\theta}}\} \approx \begin{bmatrix} \dfrac{\sigma^2}{N} & 0 \\ 0 & \dfrac{2}{N}\sigma^4 \end{bmatrix} = \mathbf{I}^{-1}(\boldsymbol{\theta})$$

which we see satisfies the asymptotic property.

Diagonal covariance matrix shows estimates are uncorrelated:



Error Ellipse is aligned with axes

This is why we could "decouple" the estimates

<u>MLE for the General Gaussian Case</u>

Let the data be general Gaussian: $\mathbf{x} \sim N\,(\boldsymbol{\mu}(\theta), \mathbf{C}(\theta))$

Thus $\partial \ln p(\mathbf{x};\theta)/\, \partial\theta$ will depend in general on $\dfrac{\partial \mathrm{u}(\theta)}{\partial\theta}$ <u>*and*</u> $\dfrac{\partial \mathrm{C}(\theta)}{\partial\theta}$

For each $k = 1, 2, \ldots, p$ set: $\dfrac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_k} = 0$

This gives p simultaneous equations, the k^{th} one being:

$$\underbrace{-\frac{1}{2} tr\left(\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial\theta_k}\right)}_{\text{Term \#1}} + \underbrace{\left[\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_k}\right]^T \mathbf{C}^{-1}(\boldsymbol{\theta})\,[\mathbf{x}-\boldsymbol{\mu}(\boldsymbol{\theta})]}_{\text{Term \#2}} - \underbrace{\frac{1}{2}[\mathbf{x}-\boldsymbol{\mu}(\boldsymbol{\theta})]^T \left[\frac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial\theta_k}\right][\mathbf{x}-\boldsymbol{\mu}(\boldsymbol{\theta})]}_{\text{Term \#3}} = 0$$

<u>Note</u>: for the "deterministic signal + noise" case: Terms #1 & #3 are zero

This gives general conditions to find the MLE…

but can't always solve it!!!

MLE for Linear Model Case

For this case we can solve these equations!

The signal model is: $\mathbf{x} = \mathbf{H}\theta + \mathbf{w}$ with the noise $\mathbf{w} \sim N(\mathbf{0},\mathbf{C})$

So terms #1 & #3 are zero and term #2 gives: $\underbrace{\left[\frac{\partial(\mathbf{H}\boldsymbol{\theta})^T}{\partial\boldsymbol{\theta}}\right]}_{=\mathbf{H}} \mathbf{C}^{-1}\left[\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\right] = \mathbf{0}$

Solving this gives: $\boxed{\hat{\theta}_{ML} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}}$

Hey! Same as chapter 4's MVU for linear model

For Linear Model: ML = MVU

Recall: the Linear Model is specified to have Gaussian noise

$\Rightarrow \boxed{\hat{\boldsymbol{\theta}}_{ML} \sim N(\boldsymbol{\theta},(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1})}$

EXACT…
Not Asymptotic!!

Numerical Solutions for Vector Case

Obvious generalizations… see p. 187

There is one issue to be aware of, though:

The numerical implementation needs $\partial \ln p(\mathbf{x};\theta)/\partial\theta$

For the general Gaussian case this requires: $\dfrac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$

…often hard to analytically: get $\mathbf{C}^{-1}(\theta)$ & then differentiate!

So… we use (3C.2):

$$\frac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial \theta_k} = -\underbrace{\mathbf{C}^{-1}(\boldsymbol{\theta})}\,\underbrace{\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \theta_k}}\,\underbrace{\mathbf{C}^{-1}(\boldsymbol{\theta})}$$

Get Analytically

Get Numerically

7.9 Asymptotic MLE

Useful when data samples $x[n]$ come from a WSS process

Reading Assignment Only