ECE 275A: Parameter Estimation I
The Expectation Maximization Algorithm

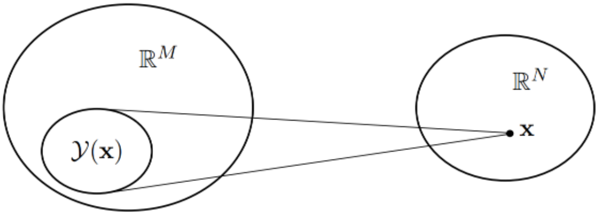
Florian Meyer

Electrical and Computer Engineering Department
University of California San Diego

UC San Diego
JACOBS SCHOOL OF ENGINEERING

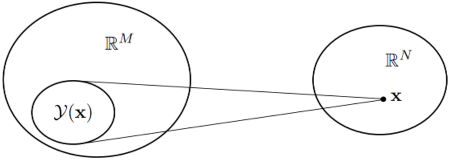
Complete and Incomplete Data

- The *complete* data $\mathbf{y} = \left(y_1, y_2, \cdots, y_M\right)^\mathsf{T} \in \mathbb{R}^M$ are what we would like to have since it leads to an effective ML estimator; unfortunately, it is not available
- The *incomplete* data $\mathbf{x} = \left(x_1, x_2, \cdots, x_N\right)^\mathsf{T} \in \mathbb{R}^N$ are what we observe
- Complete and incomplete data are related via a many-to-one mapping; the dimension of $\mathbf{x}$ is smaller than the dimension of $\mathbf{y}$, i.e., $N < M$



Many-to-one mapping from **y** to **x**

Many-to-one Mapping: $\mathbf{y} \rightarrow \mathbf{x}$



- An complete data point $\boldsymbol{y}$, uniquely determines an associated incomplete data point $\boldsymbol{x}$
- An incomplete data point $\boldsymbol{x}$, however, does not uniquely determine an associated complete data point $\boldsymbol{y}$; any value of $\boldsymbol{x}$ corresponds to a set of complete datapoints, $\mathcal{Y}(\boldsymbol{x})$
- The pdf of $\mathbf{x}$ can be expressed as the integral of the pdf of $\mathbf{y}$ over the set $\mathcal{Y}(\mathbf{x})$, i.e.,

$$f(\mathbf{x};\boldsymbol{\theta}) = \int_{\mathcal{Y}(\mathbf{x})} f(\mathbf{y};\boldsymbol{\theta})\, d\mathbf{y}$$

Example: Nonlinear Signal-in-Noise Model

- Consider the nonlinear signal-in-noise model

$$\mathbf{x} = \sum_{k=1}^{p} \mathbf{s}_k(\theta_k) + \mathbf{w} \tag{1}$$

where $\mathbf{s}_k(\theta_k) \in \mathbb{R}^N$ is the deterministic signal that depends on the k-th parameter θ_k in some nonlinear manner, $\mathbf{w} \in \mathbb{R}^N$ is a white Gaussian noise with zero mean and variance σ^2 for each element, and $\mathbf{x} \in \mathbb{R}^N$ is the observed data

- Calculation of the ML estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} \ln f(\mathbf{x}; \boldsymbol{\theta})$ amounts to the solution of the following linear inverse problem, i.e.,

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\mathbf{x}) = \arg\min_{\boldsymbol{\theta}} \left\| \mathbf{x} - \sum_{k=1}^{p} \mathbf{s}_k(\theta_k) \right\|^2 \tag{2}$$

which requires a p-D numerical minimization and is considered too difficult

Example: Nonlinear Signal in Noisel

- A much simpler ML estimation problem would be given by the following p nonlinear signal-in-noise models:

$$\mathbf{y}_k = \mathbf{s}_k(\theta_k) + \mathbf{w}_k, \quad k = 1, 2, \cdots, p \tag{3}$$

  where all vectors are N-D and $\mathbf{w}_k$'s are *mutually independent* white Gaussian noise with zero mean and variance σ_k^2 for each element

- We can combine all data vectors $\mathbf{y}_k$ into a single vector $\mathbf{y} = \left(\mathbf{y}_1^{\mathsf{T}}, \mathbf{y}_2^{\mathsf{T}}, \cdots, \mathbf{y}_p^{\mathsf{T}}\right)^{\mathsf{T}} \in \mathbb{R}^M$ with $M = pN$ that is described by a single nonlinear signal model, i.e.,

$$\mathbf{y} = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{w}'$$

  where $\mathbf{s}(\boldsymbol{\theta}) = \left(\mathbf{s}_1^{\mathsf{T}}(\theta_1), \mathbf{s}_2^{\mathsf{T}}(\theta_2), \cdots, \mathbf{s}_p^{\mathsf{T}}(\theta_p)\right)^{\mathsf{T}}$ and $\mathbf{w}' = \left(\mathbf{w}_1^{\mathsf{T}}, \mathbf{w}_2^{\mathsf{T}}, \cdots, \mathbf{w}_p^{\mathsf{T}}\right)^{\mathsf{T}}$; note that the complete data $\mathbf{y}$ has p times as many elements as the incomplete data $\mathbf{x}$

Example: Nonlinear Signal-in-Noise Model

- As for (3), since each data vector $\mathbf{y}_k$ contains only a single scalar parameter θ_k and the noise are statistically independent for each k, the ML estimate of $\boldsymbol{\theta}$ from the complete data $\mathbf{y}$, i.e., $\hat{\boldsymbol{\theta}}_{\text{ML}}(\mathbf{y}) = \arg\max_{\boldsymbol{\theta}} \ln f(\mathbf{y}; \boldsymbol{\theta})$, amounts to the solution of the following p nonlinear LS problems, i.e.,

$$\hat{\theta}_{\text{ML},k}(\mathbf{y}_k) = \arg\min_{\theta_k} \left\| \mathbf{y}_k - \mathbf{s}_k(\theta_k) \right\|^2, \quad k = 1, 2, \cdots, p \qquad (4)$$

- These are p 1-D minimizations, which are much simpler than the original p-D minimization problem

Example: Nonlinear Signal-in-Noise Model

- We can formally establish a relation between the given signal model (1) and the hypothetical signal model (3) as

$$\mathbf{x} = \sum_{k=1}^{p} \mathbf{y}_k = \sum_{k=1}^{p} \mathbf{s}_k(\theta_k) + \mathbf{w} \tag{5}$$

  where $\mathbf{w} = \sum_{k=1}^{p} \mathbf{w}_k$. Since $\mathbf{w}_k$'s are independent, we have $\sigma^2 = \sum_{k=1}^{p} \sigma_k^2$

- Note that (5) is a *many-to-one mapping* $\mathbf{y} \to \mathbf{x}$; while $\mathbf{y} = \left(\mathbf{y}_1^{\mathsf{T}}, \mathbf{y}_2^{\mathsf{T}}, \cdots, \mathbf{y}_p^{\mathsf{T}}\right)^{\mathsf{T}}$ uniquely specifies $\mathbf{x}$, $\mathbf{x}$ does not specify $\mathbf{y}$

- The hypothetical ML estimation problem (4) is much simpler than the original ML estimation problem (2); thus, it would be desirable to solve the hypothetical problem instead of the original problem

- Unfortunately, the hypothetical problem cannot be solved directly since the complete data $\mathbf{y}$ is unavailable

The Expectation Maximization (EM) Algorithm

- Maximization of $\ln f(\mathbf{x};\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is difficult, so we want to maximize $\ln f(\mathbf{y};\boldsymbol{\theta})$ instead; however, $\mathbf{y}$, is unavailable
- The idea of EM algorithm is the following: instead of maximizing $\ln f(\mathbf{y};\boldsymbol{\theta})$ directly, we *estimate* $\ln f(\mathbf{y};\boldsymbol{\theta})$ from $\mathbf{x}$ and then *maximize* this estimate of $\ln f(\mathbf{y};\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$
- To estimate $\ln f(\mathbf{y};\boldsymbol{\theta})$ from $\mathbf{x}$, we compute the **conditional expectation** of $\ln f(\mathbf{y};\boldsymbol{\theta})$ given $\mathbf{x}$, i.e., $E\left[\ln f(\mathbf{y};\boldsymbol{\theta})|\mathbf{x};\boldsymbol{\theta}\right]$
- Since $\boldsymbol{\theta}$ is unknown for calculation of the conditional expectation, we initialize the parameter with value $\boldsymbol{\theta}^{(0)}$ and then iteratively update the parameter value by maximizing $\ln f(\mathbf{y};\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$

The i-th Iteration of EM Algorithm

E-M Iteration

Given the current parameter estimate $\boldsymbol{\theta}^{(i)}$, the i-th iteration of the EM algorithm consists of the following two successive steps:

(E) Calculate an estimate of $\ln f(\mathbf{y};\boldsymbol{\theta})$, i.e., the expectation of $\ln f(\mathbf{y};\boldsymbol{\theta})$ using conditional pdf $f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$ with the current parameter estimate $\boldsymbol{\theta}^{(i)}$

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}) \triangleq E\left[\ln f(\mathbf{y};\boldsymbol{\theta})|\mathbf{x};\boldsymbol{\theta}^{(i)}\right] = \int_{\mathcal{Y}(\mathbf{x})} \big(\ln f(\mathbf{y};\boldsymbol{\theta})\big) f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}^{(i)}) d\mathbf{y}$$

(M) Find a new parameter value $\boldsymbol{\theta}^{(i+1)}$ that maximizes $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)})$ with respect to $\boldsymbol{\theta}$

$$\boldsymbol{\theta}^{(i+1)} \triangleq \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}) = \arg\max_{\boldsymbol{\theta}} E\left[\ln f(\mathbf{y};\boldsymbol{\theta})|\mathbf{x};\boldsymbol{\theta}^{(i)}\right]$$

- This iterative procedure is terminated when $\|\boldsymbol{\theta}^{(i+1)} - \boldsymbol{\theta}^{(i)}\| < \epsilon$ with some small ϵ.

Comments on EM Algorithm

- As shown in class, the EM algorithm is guaranteed to converge to a (possibly only local) maximum of $\ln f(\mathbf{y}; \boldsymbol{\theta})$
- Since the ML estimate is the *global* maximum of $\ln f(\mathbf{y}; \boldsymbol{\theta})$, the EM algorithm does not necessarily produce the ML estimate
- However, the EM algorithm does have the desirable property of increasing the likelihood $f(\mathbf{x}; \boldsymbol{\theta}^{(i)})$ at each iteration
- There are two potential challenges related to developing an EM algorithm
    1. First, the formulation of a good "hypothetical ML problem" for a given ML problem is not straightforward and requires some intuition and creativity
    2. Second, determining the conditional expectation $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)})$ in closed form may be difficult or even impossible

Example: Nonlinear Signal-in-Noise Model (cont.)

- For simplicity, we assume that $\|\boldsymbol{s}_k(\theta_k)\|^2 = \mathrm{E_s}$, i.e., all "signals" have the same energy $\mathrm{E_s}$ that is independent of the value of θ_k

- **E-step:** As derived in class, we obtain

$$Q\big(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}\big) = \sum_{k=1}^{p} \frac{1}{\sigma_k^2} \boldsymbol{s}_k^{\mathrm{T}}(\theta_k) \hat{\boldsymbol{y}}_k(\boldsymbol{x}, \boldsymbol{\theta}^{(i)}) + h\big(\boldsymbol{x}, \boldsymbol{\theta}^{(i)}\big) \tag{6}$$

  where $\hat{\mathbf{y}}_k\big(\boldsymbol{x}, \boldsymbol{\theta}^{(i)}\big) = E\big[\mathbf{y}_k | \boldsymbol{x}, \boldsymbol{\theta}^{(i)}\big]$ is an estimate of the complete data, $\mathbf{y}_k$, from the observed incomplete data, $\boldsymbol{x}$, and $h\big(\boldsymbol{x}, \boldsymbol{\theta}^{(i)}\big)$ is a term that does not depent on $\boldsymbol{\theta}$

- The estimate of the complete data is obtained as

$$\hat{\boldsymbol{y}}_k\big(\boldsymbol{x}, \boldsymbol{\theta}^{(i)}\big) = \boldsymbol{s}_k\big(\theta_k^{(i)}\big) + \frac{\sigma_k^2}{\sigma^2}\left[\boldsymbol{x} - \sum_{l=1}^{p} \boldsymbol{s}_l\big(\theta_l^{(i)}\big)\right] \tag{7}$$

Example: Nonlinear Signal-in-Noise Model (cont.)

- **M-step:** Based on (7), we directly obtain

$$\begin{aligned}\boldsymbol{\theta}^{(i+1)} &= \arg\max_{\boldsymbol{\theta}} Q\big(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}\big) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{p} \frac{1}{\sigma_k^2} \boldsymbol{s}_k^{\mathrm{T}}(\theta_k)\, \hat{\boldsymbol{y}}_k\big(\boldsymbol{x}, \boldsymbol{\theta}^{(i)}\big)\end{aligned}$$