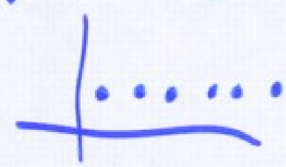
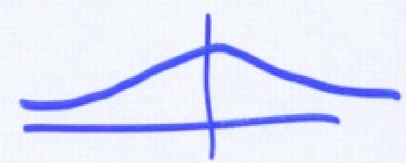
Maximum Likelihood and Expectation Maximization

Probability Basics

	Ex	Ex
Sample Space	$\{1, \dots, 6\}$	$\mathbb{R}$
Distribution	$p_1 \dots p_6 \geq 0, \sum p_i = 1$	p.d.f $f(x) \geq 0, \int_{\mathbb{R}} f(x)dx = 1$
	eg $p_1 = \dots = p_6 = \frac{1}{6}$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$





Population vs Sample

population mean	$\mu = \sum i p_i$	$\mu = \int x f(x) dx$
population Variance	$\sigma^2 = \sum (i-\mu)^2 p_i$	$\sigma^2 = \int (x-\mu)^2 f(x) dx$
sample mean	$\sum x_i / n$	
sample Variance	$\sum (x_i - \bar{x})^2 / n$	

Parameter Estimation

- Assuming sample $x_1, x_2, ..., x_n$ is from a parametric distribution f(x| theta), estimate theta.
- E.g.:

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

$$\theta = (\mu, \sigma^2)$$

Maximimum Likelihood Estimation

One (of many) approaches to parameter est.

Likelihood of $x_1 \cdots x_n$ =

$$L(x_1, .., x_n | \Theta) = \underbrace{\prod_{i=1}^{n}}_{\text{assume indp.}} f(x_i | \Theta)$$

View this as a function of Θ

what Θ maximizes the likelihood

Typical approach: $\frac{\partial}{\partial\Theta} L(x | \Theta) = 0$

or $\frac{\partial}{\partial\Theta} \ln L(x|\Theta) = 0$

Example

$X_1 \cdots X_n$ coin flips; Θ = prob of heads

n_0 tails, n_1 heads, $n_0 + n_1 = n$

$$L(x_1 \cdots x_n \mid \Theta) = (1-\Theta)^{n_0} (\Theta)^{n_1}$$

$$\ln L = n_0 \ln(1-\Theta) + n_1 \ln \Theta$$

$$\frac{d}{d\Theta} \ln L = \frac{-n_0}{1-\Theta} + \frac{n_1}{\Theta} = 0$$

$$n_0 \Theta = n_1 (1-\Theta)$$

$$(n_0 + n_1)\Theta = n_1$$

$$\boxed{\Theta = \frac{n_1}{n}}$$

And verify it's max, not min & not better on boundary

Example $x_i \sim N(\mu, \sigma^2)$, $\sigma^2 = 1$, μ unknown

$$L(x_1 \cdots x_n \mid \Theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \Theta)^2}{2}}$$

$$\ln L(x_1 \cdots x_n \mid \Theta) = \sum -\tfrac{1}{2}\ln 2\pi - \frac{(x_i - \Theta)^2}{2}$$

$$\frac{dL}{d\Theta} = -\sum (x_i - \Theta) = 0$$

$$\left(\sum x_i\right) - n\Theta = 0$$

$$\boxed{\Theta = \sum x_i / n}$$

And verify it's max, not min
& not better on boundary

Example $x_i \sim N(\mu, \sigma^2)$, both unknown

...

$$\ln L(x_1 \ldots x_n | \theta_1 \theta_2) = \sum -\frac{1}{2} \ln 2\pi\theta_2 - \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial \theta_1} \ln L(x_1 \ldots x_n | \theta_1 \theta_2) = \sum \frac{x_i - \theta_1}{\theta_2} = 0$$

$$\Rightarrow \quad \boxed{\theta_1 = \sum x_i / n}$$

$$\frac{\partial}{\partial \theta_2} \ln L(x_1 \ldots x_n | \theta_1 \theta_2) = \sum -\frac{2\pi}{2 \cdot 2\pi\theta_2} + \frac{(x_i - \theta_1)^2}{2\theta_2^2}$$

$$\sum -\frac{1}{2\theta_2} + \frac{(x_i-\theta_1)^2}{2\theta_2^2} = 0$$

$$\sum (x_i - \theta_1)^2 = n\theta_2$$

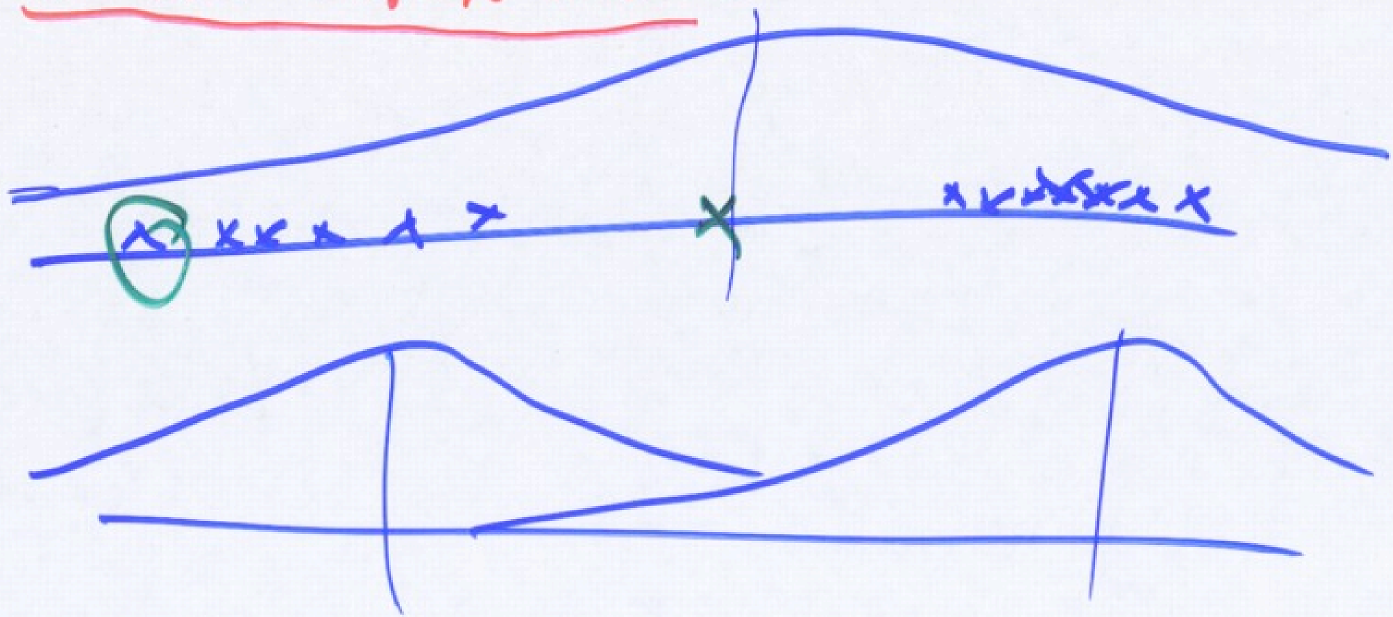
$$\boxed{\theta_2 = \sum (x_i - \theta_1)^2 / n}$$

A Biased (but consistent) estimate of population variance

An Example of Overfitting

Unbiased estimate: $\sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{n-1}$

A More Complex Problem



2 distributions $f_1(x)$, $f_2(x)$

$f_1(x|\theta_1)$, $f_2(x|\theta_2)$

Mixing parameter τ_1 τ_2

$\tau_1 + \tau_2 = 1$

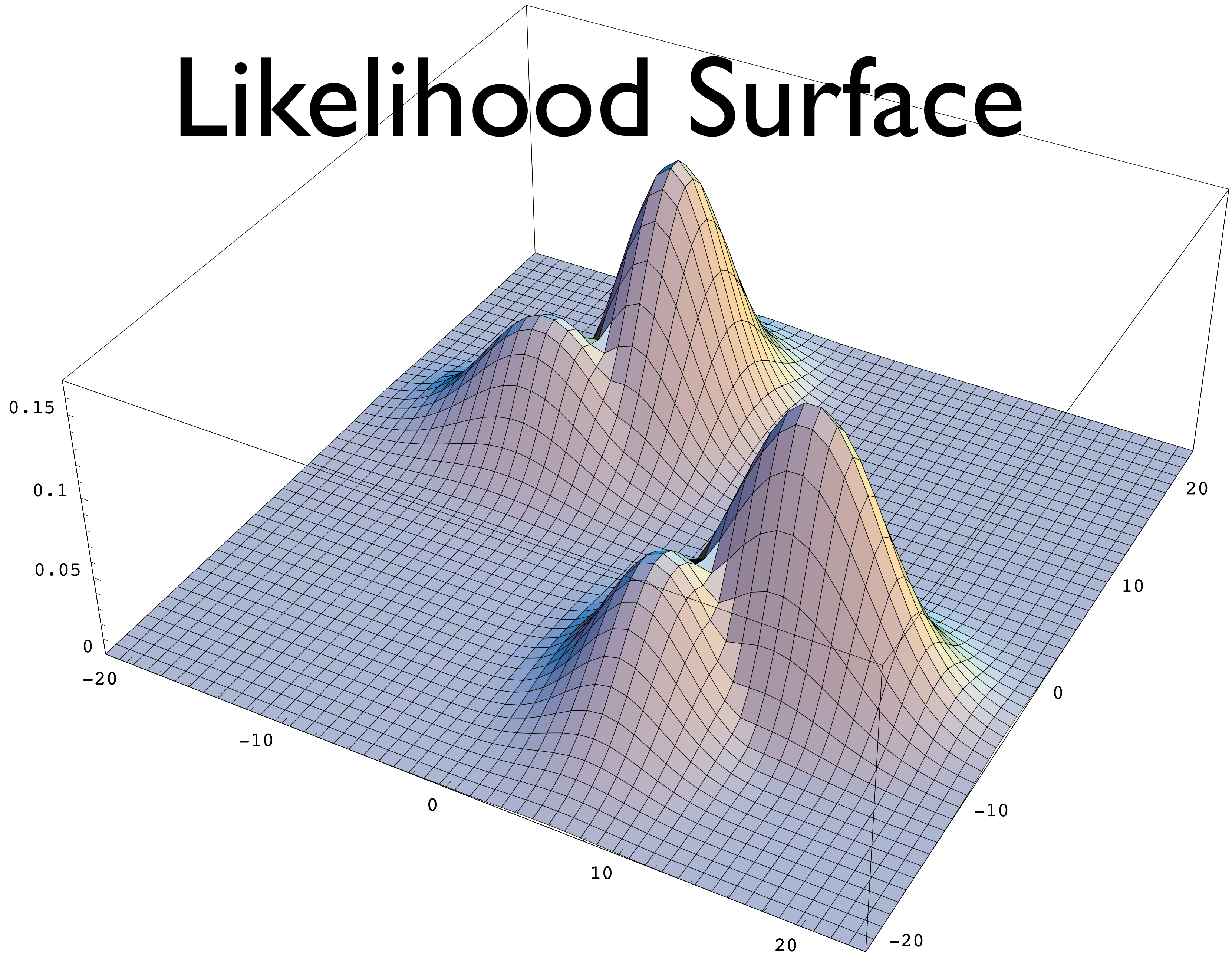
Likelihood

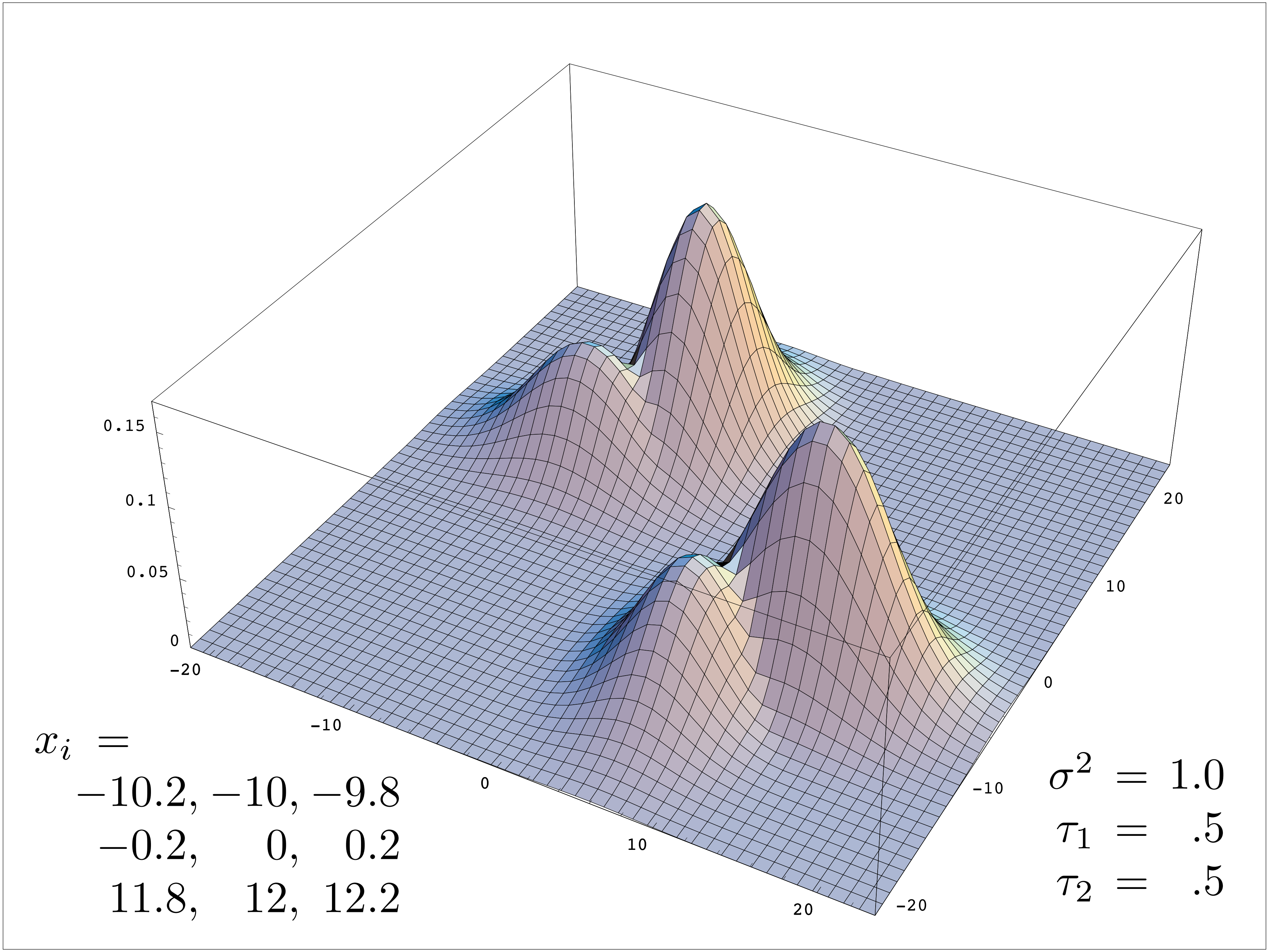
$$L(x_1 \dots x_n \mid \underbrace{\tau_1 \tau_2 \mu_1 \mu_2 \sigma_1^2 \sigma_2^2 \dots}_{\Theta})$$

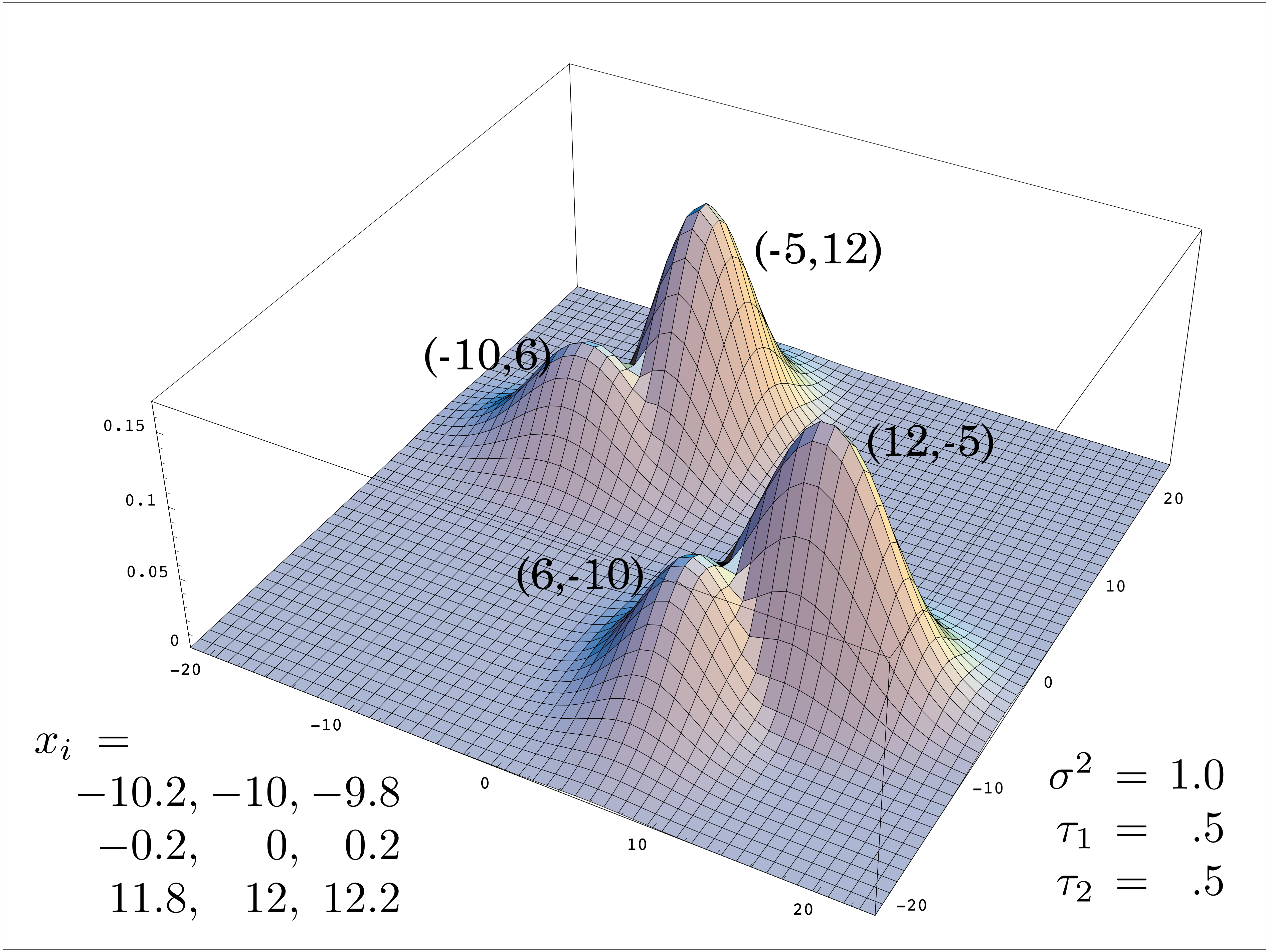
$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f_j(x_i \mid \theta)$$

Probably too messy for closed-form solution

Likelihood Surface




0.15
0.1
0.05
0
-20
-10
0
10
20
20
10
0
-10
-20
$x_i =$
$-10.2, -10, -9.8$
$-0.2, 0, 0.2$
$11.8, 12, 12.2$
$\sigma^2 = 1.0$
$\tau_1 = .5$
$\tau_2 = .5$


(-5,12)
(-10,6)
(12,-5)
(6,-10)
0.15
0.1
0.05
0
-20
-10
0
10
20
20
10
0
-10
-20
$x_i =$
$-10.2, -10, -9.8$
$-0.2, 0, 0.2$
$11.8, 12, 12.2$
$\sigma^2 = 1.0$
$\tau_1 = .5$
$\tau_2 = .5$

Likelihood

$$L(x_1 \dots x_n \mid \underbrace{\tau_1 \tau_2 \mu_1 \mu_2 \sigma_1^2 \sigma_2^2 \dots}_{\Theta})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f_j(x_i \mid \theta)$$

Probably too messy for closed-form solution

Full data

$$\begin{array}{lll} x_1 & z_{11} & z_{12} \\ x_2 & z_{21} & z_{22} \\ x_3 & z_{31} & z_{32} \end{array}$$

$\underbrace{z_{11} \dots z_{32}}_{\text{Hidden Variables}}$

$z_{ij} \begin{cases} 0 \\ 1 & \text{iff } x_i \text{ comes from distribution } j \end{cases}$

EM as Egg vs Chicken

- *IF* parameters known, could estimate z_{ij}
- *IF* z_{ij} known, could estimate parameters
- But we know neither, so iterate:
  - E: calculate expected z_{ij}, given parameters
  - M: calc "MLE" of parameters, given $E(z_{ij})$

The E-Step

assume τ_j θ_j fixed

A event that x_i drawn from f_1

B . . . . f_2

D "data" x_i is observed

$P(A \mid D)$

$P(D \mid A)$

$$P(A \mid D) = \frac{P(D \mid A)\, P(A)}{P(D)}$$

Bayes rule

$$P(D) = \underbrace{P(D \mid A)}_{f_1(x_i \mid \theta_1)} \cdot \underbrace{P(A)}_{\tau_1} + \underbrace{P(D \mid B)}_{f_2(x_i \mid \theta_2)} \cdot \underbrace{P(B)}_{\tau_2}$$

Expected value of $Z_{i,1}$

The M-Step

$L(X_1 Z_{11} Z_{12} X_2 Z_{21} Z_{22} \ldots | \Theta T)$

X_i's known

if Z_{ij} known, then MLE ΘT easy

But we don't.

Instead maximize expected likelihood of visible data

$E(L(X_1, X_2, \ldots X_n | \Theta T))$

where Expectation is over distribution of hidden values (Z_{ij}'s)

$$L(\vec{X}, \vec{Z} \mid \Theta^T)$$

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{2} z_{ij} \cdot (x_i - \mu_j)^2}$$

$$E(\ln L(\vec{X}\vec{Z} \mid \Theta^T)) =$$

$$E\left[\sum_{i=1}^{n} -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{2} z_{ij} (x_i - \mu_j)^2\right]$$

$$= \sum_{i=1}^{n} -\frac{1}{2}\ln 2\pi\sigma^2 \underbrace{- \frac{1}{2\sigma^2}\sum_{j=1}^{2} E(z_{ij})(x_i - \mu_j)^2}$$

Find μ_j maximizing ↑ using $E(z_{ij})$ from E-step.