CSEP 590B

Spring 2011

4: MLE, EM

Outline

HW#2 Discussion

MLE: Maximum Likelihood Estimators

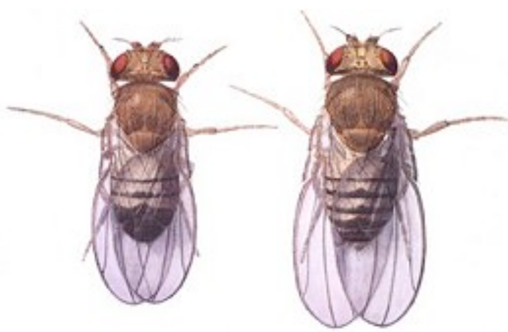
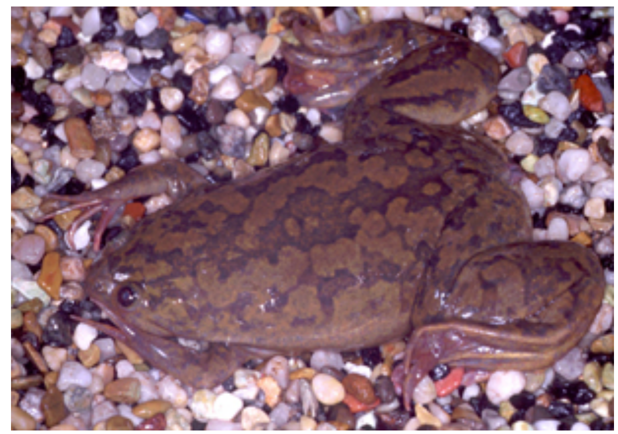
EM: the Expectation Maximization Algorithm

Next: Motif description & discovery

HW # 2 Discussion

	Species	Name	Description	Access -ion	score to #1
1	Homo sapiens (Human)	MYOD1_HUMAN	Myoblast determination protein 1	P15172	~1700 ?
2	Homo sapiens (Human)	TAL1_HUMAN	T-cell acute lymphocytic leukemia protein 1 (TAL-1)	P17542	143
3	Mus musculus (Mouse)	MYOD1_MOUSE	Myoblast determination protein 1	P10085	1494
4	Gallus gallus (Chicken)	MYOD1_CHICK	Myoblast determination protein 1 homolog (MYOD1 homolog)	P16075	1020
5	Xenopus laevis (African clawed frog)	MYODA_XENLA	Myoblast determination protein 1 homolog A (Myogenic factor 1)	P13904	978
6	Danio rerio (Zebrafish)	MYOD1_DANRE	Myoblast determination protein 1 homolog (Myogenic factor 1)	Q90477	893
7	Branchiostoma belcheri (Amphioxus)	Q8IU24_BRABE	MyoD-related	Q8IU24	426
8	Drosophila melanogaster (Fruit fly)	MYOD_DROME	Myogenic-determination protein (Protein nautilus) (dMyd)	P22816	368
9	Caenorhabditis elegans	LIN32_CAEEL	Protein lin-32 (Abnormal cell lineage protein 32)	Q10574	118
10	Homo sapiens (Human)	SYFM_HUMAN	Phenylalanyl-tRNA synthetase, mitochondrial	O95363	~55?



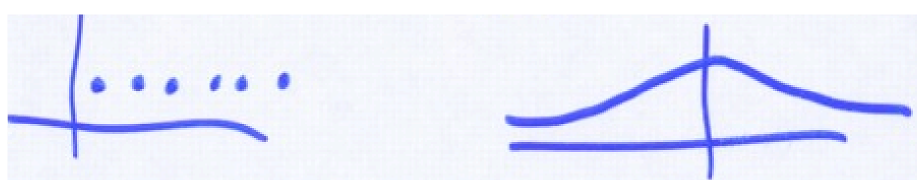

Courtesy of Ralf Sommer

MyoD

http://www.rcsb.org/pdb/explore/jmol.do?structureId=1MDY&bionumber=1

Probability Basics, I

	Ex.	Ex.
Sample Space		
	$\{1, 2, \ldots, 6\}$	$\mathbb{R}$
Distribution		
	$p_1, \ldots, p_6 \geq 0; \sum_{1 \leq i \leq 6} p_i = 1$	$f(x) >= 0; \int_{\mathbb{R}} f(x)dx = 1$
e.g.		
	$p_1 = \cdots = p_6 = 1/6$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$



pdf, not probability

Probability Basics, II

	Ex.	Ex.
Expectation	$E(g) = \sum_{1 \le i \le 6} g(i) p_i$	$E(g) = \int_{\mathbb{R}} g(x) f(x) dx$
Population		
mean	$\mu = \sum_{1 \le i \le 6} i p_i$	$\mu = \int_{\mathbb{R}} x f(x) dx$
variance	$\sigma^2 = \sum_{1 \le i \le 6} (i - \mu)^2 p_i$	$\sigma^2 = \int_{\mathbb{R}} (x - \mu)^2 f(x) dx$

Sample

mean
$$\bar{x} = \sum_{1 \le i \le n} x_i / n$$

variance
$$\bar{s}^2 = \sum_{1 \le i \le n} (x_i - \bar{x})^2 / n$$

Learning From Data: MLE

Maximum Likelihood Estimators

Parameter Estimation

Assuming sample $x_1, x_2, ..., x_n$ is from a parametric distribution $f(x|\theta)$, estimate θ.

E.g.: Given sample HHTTTTTHTHTTTHH of (possibly biased) coin flips, estimate

$$\theta = \text{probability of Heads}$$

$f(x|\theta)$ is the Bernoulli probability mass function with parameter θ

Likelihood

$P(x \mid \theta)$: Probability of event x given *model* θ

Viewed as a function of x (fixed θ), it's a *probability*

E.g., $\Sigma_x P(x \mid \theta) = 1$

Viewed as a function of θ (fixed x), it's a *likelihood*

E.g., $\Sigma_\theta P(x \mid \theta)$ can be anything; *relative* values of interest.

E.g., if θ = prob of heads in a sequence of coin flips then

$P(HHTHH \mid .6) > P(HHTHH \mid .5)$,

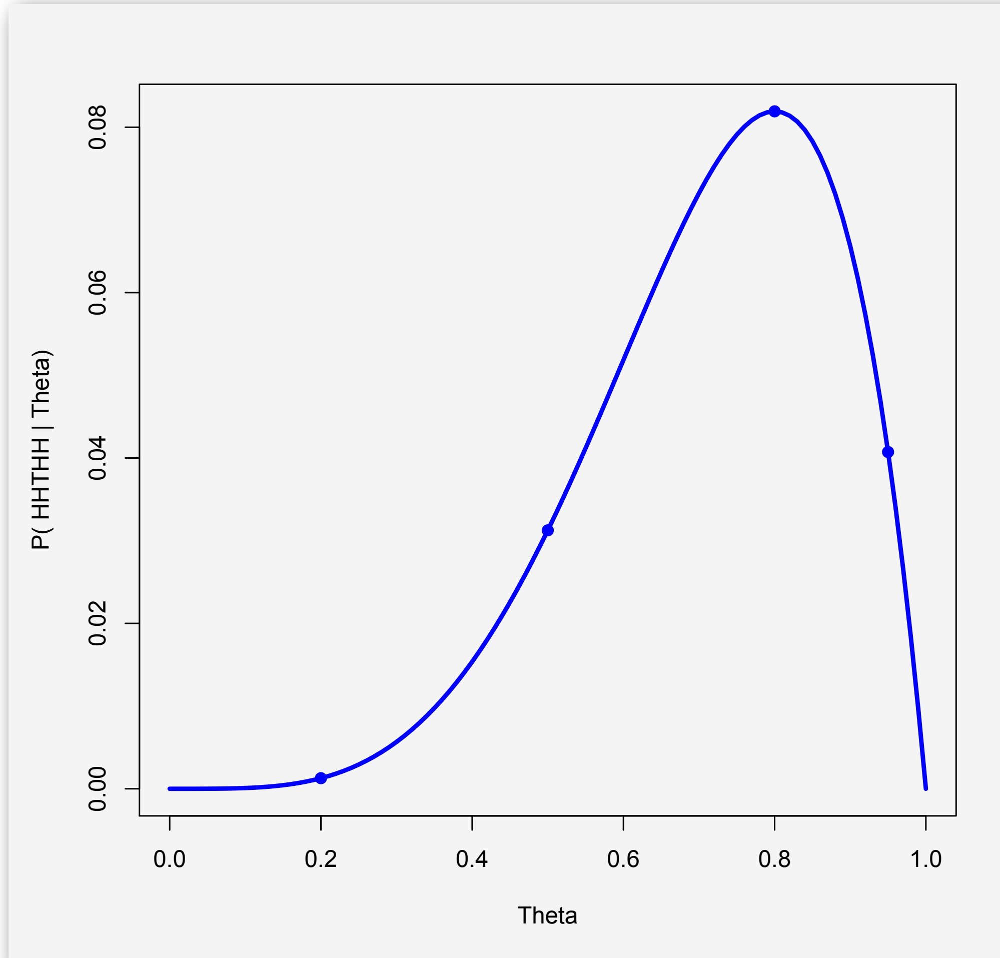
I.e., event HHTHH is *more likely* when $\theta = .6$ than $\theta = .5$

And what θ make HHTHH *most* likely?

Likelihood Function

P(HHTHH | θ):
Probability of HHTHH,
given P(H) = θ:

θ	$\theta^4(1-\theta)$
0.2	0.0013
0.5	0.0313
0.8	0.0819
0.95	0.0407



Maximum Likelihood Parameter Estimation

One (of many) approaches to param. est.

Likelihood of (indp) observations $x_1, x_2, ..., x_n$

$$L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

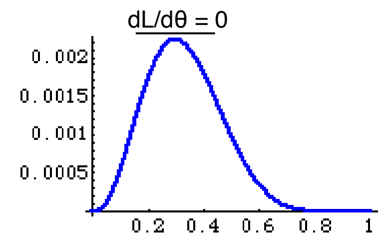
As a function of θ, what θ maximizes the likelihood of the data actually observed

Typical approach: $\frac{\partial}{\partial\theta} L(\vec{x} \mid \theta) = 0$ or $\frac{\partial}{\partial\theta} \log L(\vec{x} \mid \theta) = 0$

Example 1

n coin flips, $x_1, x_2, \ldots, x_n$; n_0 tails, n_1 heads, $n_0 + n_1 = n$;

θ = probability of heads



$$L(x_1, x_2, \ldots, x_n \mid \theta) = (1-\theta)^{n_0}\theta^{n_1}$$

$$\log L(x_1, x_2, \ldots, x_n \mid \theta) = n_0 \log(1-\theta) + n_1 \log\theta$$

$$\frac{\partial}{\partial\theta} \log L(x_1, x_2, \ldots, x_n \mid \theta) = \frac{-n_0}{1-\theta} + \frac{n_1}{\theta}$$

Setting to zero and solving:

$$\boxed{\hat{\theta} = \frac{n_1}{n}}$$

Observed fraction of successes in sample is MLE of success probability in population

(Also verify it's max, not min, & not better on boundary)

Bias

A desirable property: An estimator Y of a parameter θ is an *unbiased* estimator if

$$E[Y] = \theta$$

For coin ex. above, MLE is unbiased:

Y = fraction of heads = $(\Sigma_{1 \leq i \leq n} X_i)/n$,

(X_i = indicator for heads in i^{th} trial) so

$$E[Y] = (\Sigma_{1 \leq i \leq n} E[X_i])/n = n\,\theta/n = \theta$$

by linearity of expectation

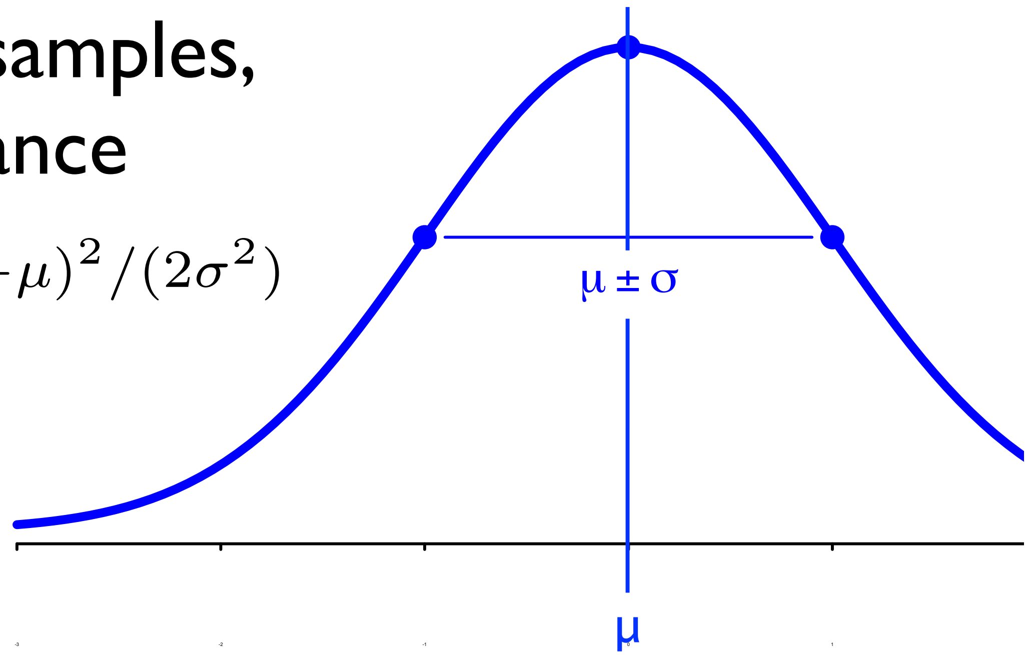
Parameter Estimation

Assuming sample $x_1, x_2, ..., x_n$ is from a parametric distribution $f(x|\theta)$, estimate θ.

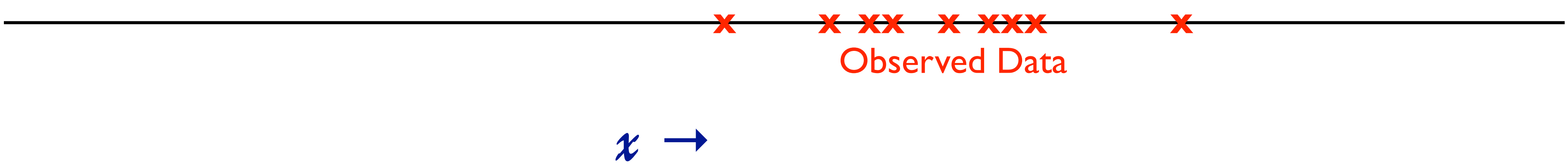
E.g.: Given n normal samples, estimate mean & variance

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

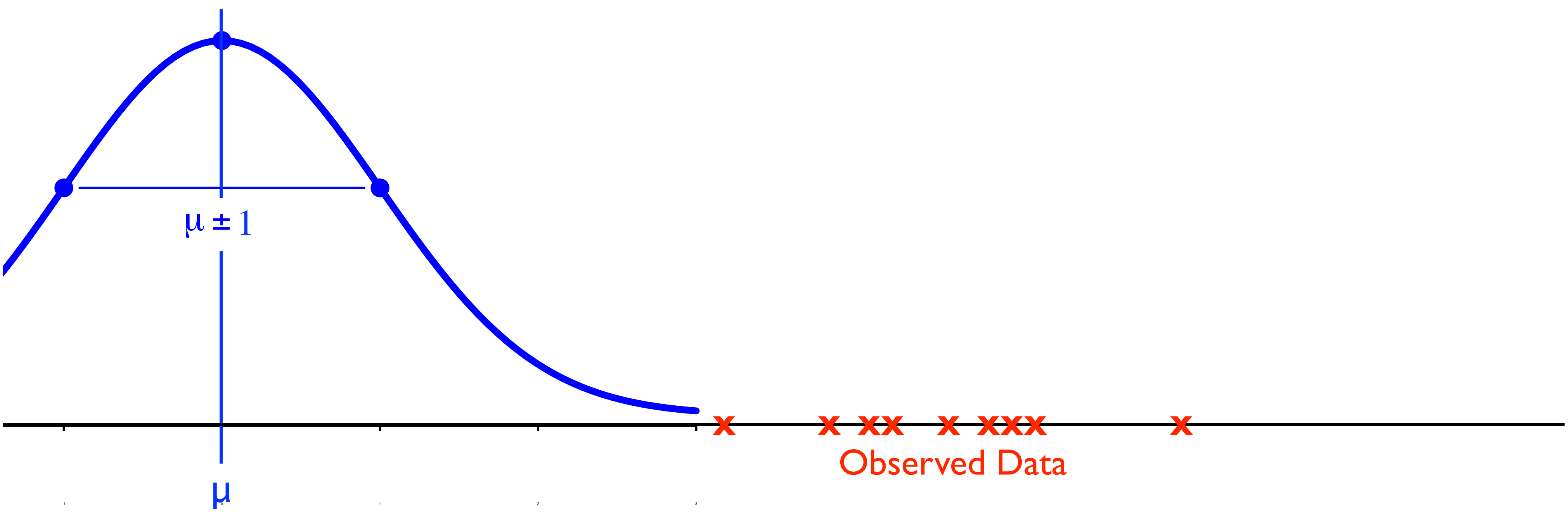
$$\theta = (\mu, \sigma^2)$$



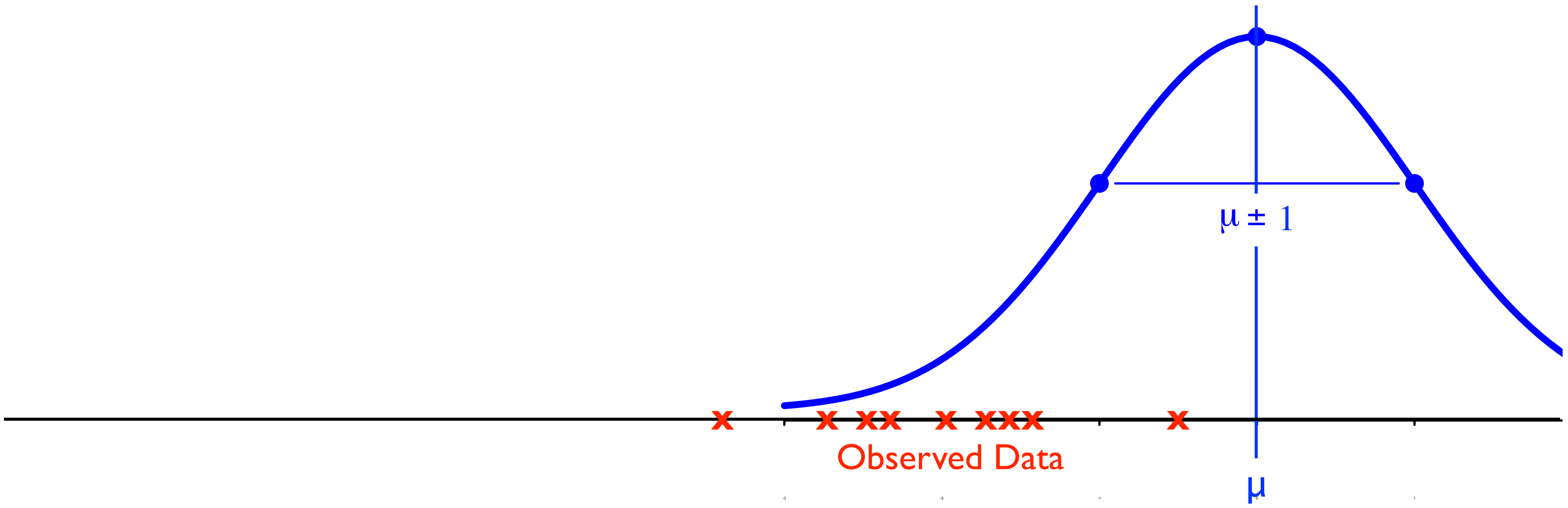
Ex2: I got data; a little birdie tells me it's normal, and promises $\sigma^2 = 1$



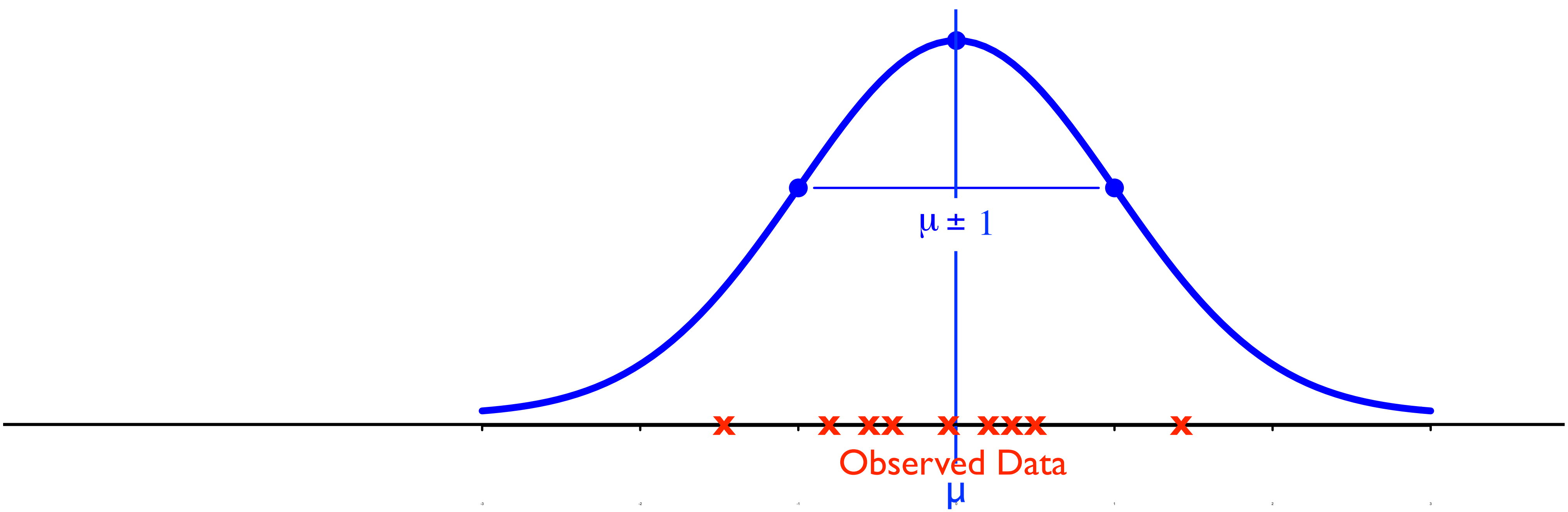
Which is more likely: (a) this?



Which is more likely: (b) or this?

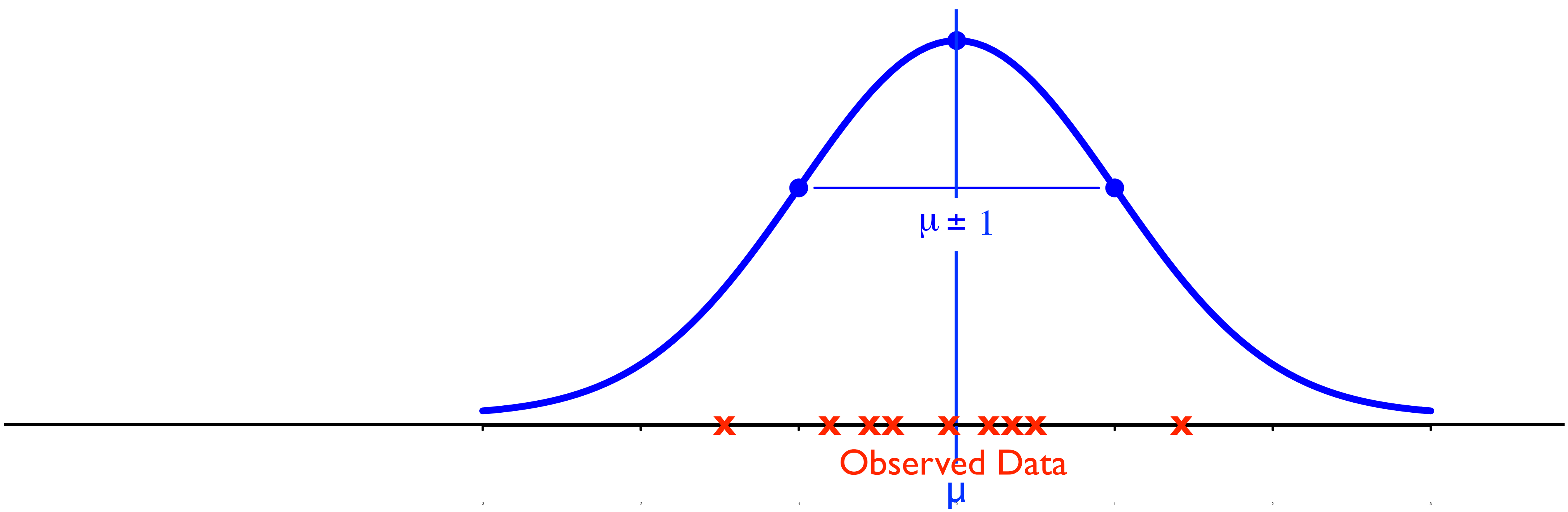


Which is more likely: (c) or *this*?



Which is more likely: (c) or *this*?

Looks good by eye, but how do I optimize my estimate of μ ?



Ex. 2: $x_i \sim N(\mu, \sigma^2),\ \sigma^2 = 1,\ \mu$ unknown

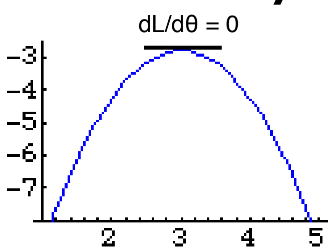
$$L(x_1, x_2, \ldots, x_n|\theta) = \prod_{1\le i\le n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2}$$

$$\ln L(x_1, x_2, \ldots, x_n|\theta) = \sum_{1\le i\le n} -\frac{1}{2}\ln 2\pi - \frac{(x_i-\theta)^2}{2}$$

$$\frac{d}{d\theta}\ln L(x_1, x_2, \ldots, x_n|\theta) = \sum_{1\le i\le n} (x_i - \theta)$$

$$= \left(\sum_{1\le i\le n} x_i\right) - n\theta = 0$$

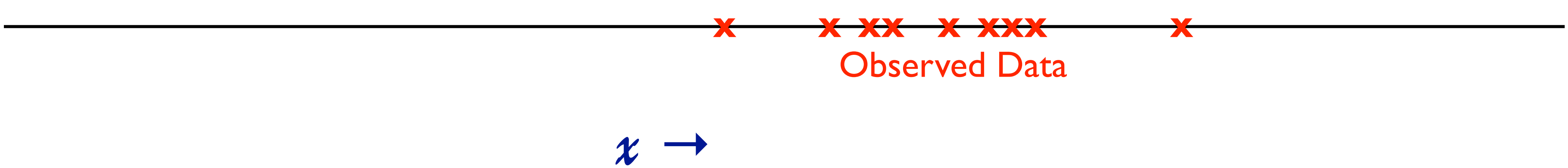
And verify it's max, not min & not better on boundary



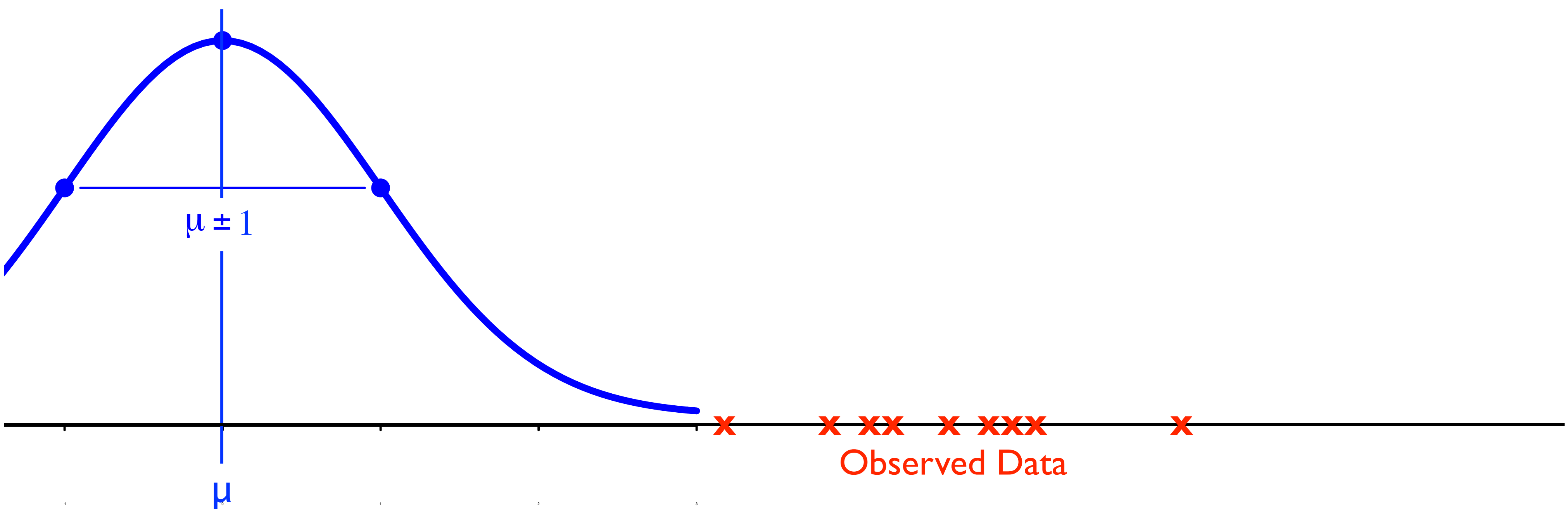
$$\boxed{\hat{\theta} = \left(\sum_{1\le i\le n} x_i\right)/n = \bar{x}}$$

Sample mean is MLE of population mean

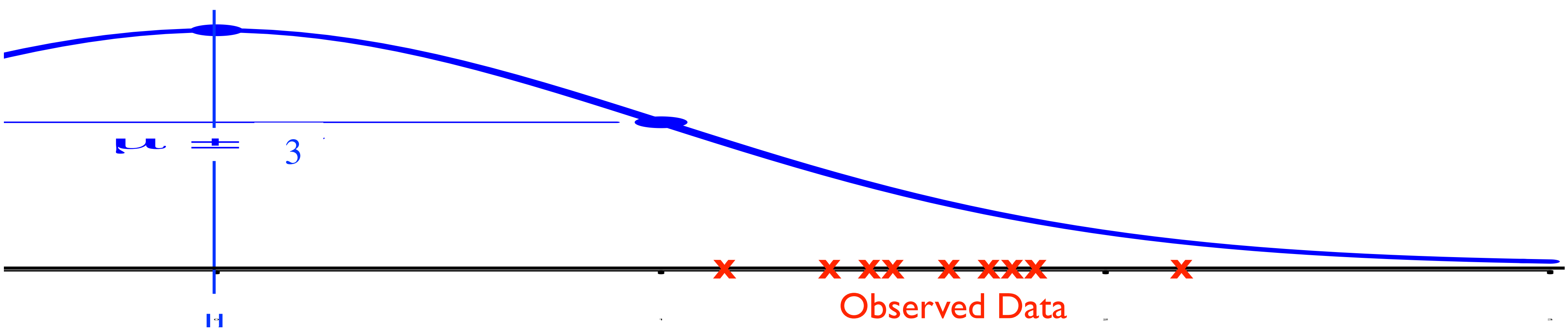
Ex3: I got data; a little birdie tells me it's normal (but does *not* tell me σ^2)



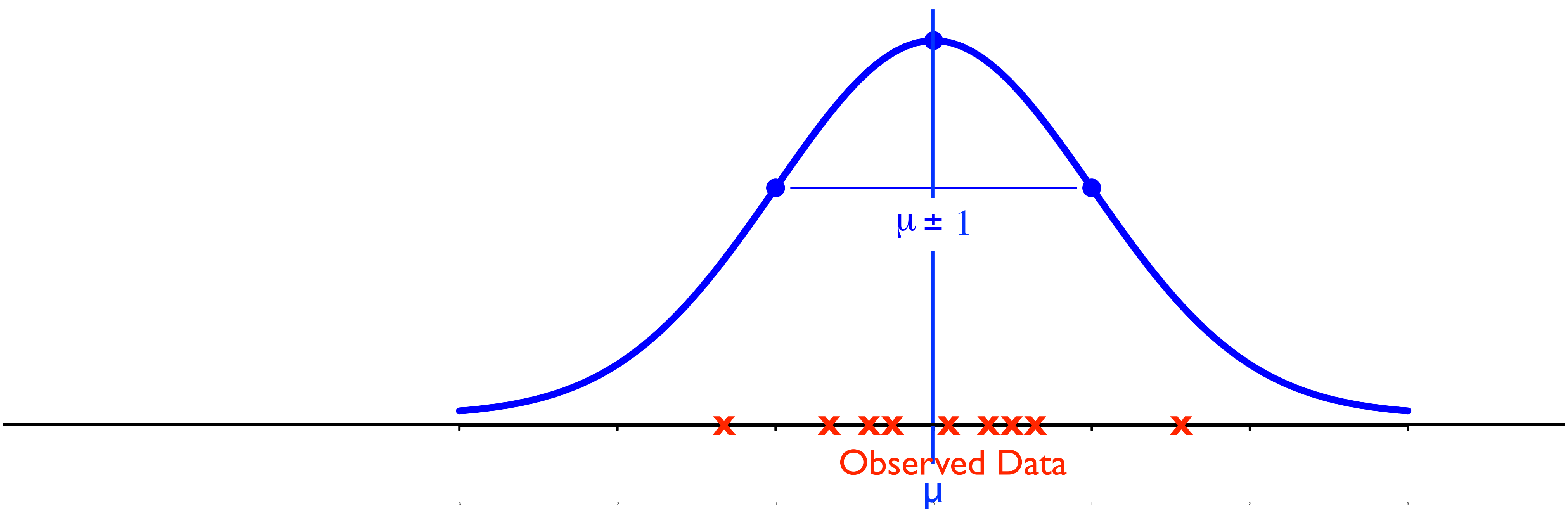
Which is more likely: (a) this?



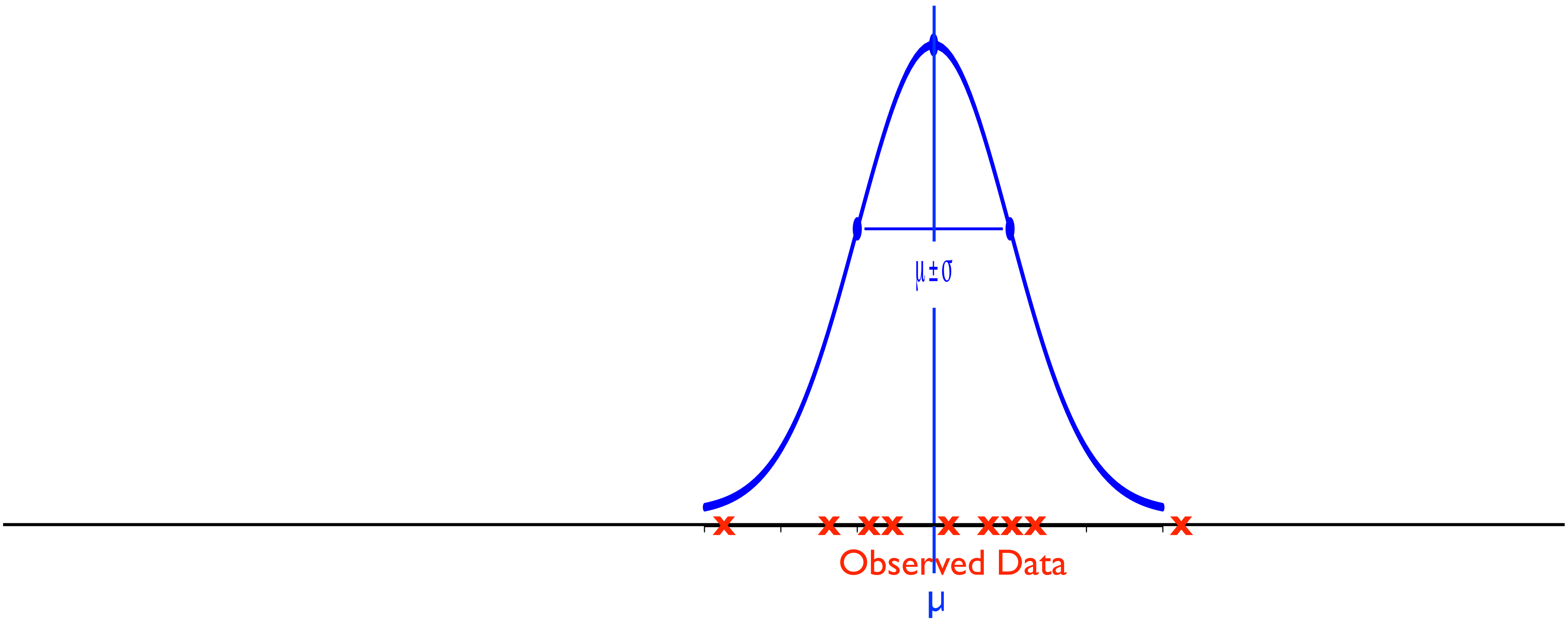
Which is more likely: (b) or this?



Which is more likely: (c) or this?

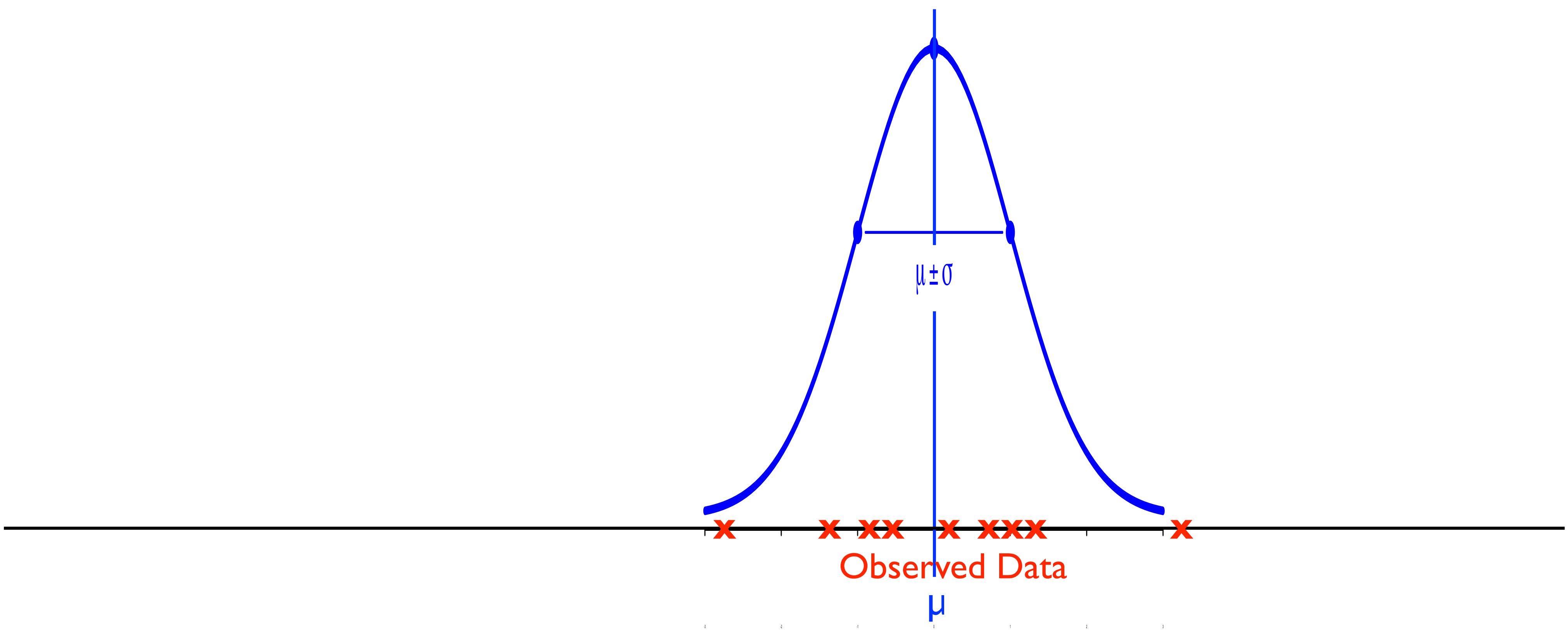


Which is more likely: (d) or *this*?



Which is more likely: (d) or *this*?

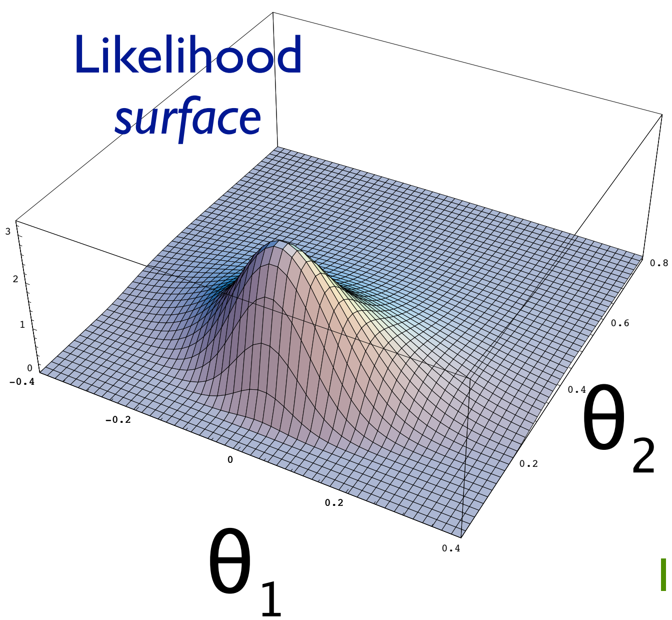
Looks good by eye, but how do I optimize my estimates of μ & σ ?



Ex 3: $x_i \sim N(\mu, \sigma^2)$, μ, σ^2 both unknown

$$\ln L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2) = \sum_{1 \le i \le n} -\frac{1}{2} \ln 2\pi\theta_2 - \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial \theta_1} \ln L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2) = \sum_{1 \le i \le n} \frac{(x_i - \theta_1)}{\theta_2} = 0$$



$$\boxed{\hat{\theta}_1 = \left(\sum_{1 \le i \le n} x_i \right) / n = \bar{x}}$$

Sample mean is MLE of population mean, again

In general, a problem like this results in 2 equations in 2 unknowns. Easy in this case, since θ_2 drops out of the $\partial/\partial\theta_1 = 0$ equation

Ex. 3, (cont.)

$$\ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) = \sum_{1 \le i \le n} -\frac{1}{2} \ln 2\pi\theta_2 - \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial\theta_2} \ln L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2) = \sum_{1 \le i \le n} -\frac{1}{2}\frac{2\pi}{2\pi\theta_2} + \frac{(x_i - \theta_1)^2}{2\theta_2^2} = 0$$

$$\boxed{\hat{\theta}_2 = \left(\sum_{1 \le i \le n}(x_i - \hat{\theta}_1)^2\right)/n = \bar{s}^2}$$

Sample variance is MLE of *population* variance

Ex. 3, (cont.)

Bias? if Y is sample mean

$Y = (\Sigma_{1 \le i \le n} X_i)/n$

then

$E[Y] = (\Sigma_{1 \le i \le n} E[X_i])/n = n\ \mu/n = \mu$

so the MLE is an *unbiased* estimator of population mean

Similarly, $(\Sigma_{1 \le i \le n} (X_i-\mu)^2)/n$ is an unbiased estimator of σ^2. Unfortunately, if μ is unknown, estimated *from the same data*, as above, $\hat{\theta}_2 = \sum_{1 \le i \le n} \frac{(x_i - \hat{\theta}_1)^2}{n}$ is a consistent, but *biased* estimate of population variance. (An example of *overfitting.*) Unbiased estimate is:

I.e., $\lim_{n \to \infty}$ = correct

$$\hat{\theta}'_2 = \sum_{1 \le i \le n} \frac{(x_i - \hat{\theta}_1)^2}{n-1}$$

Moral: MLE is a great idea, but not a magic bullet

More on Bias of $\hat{\theta}_2$

Biased? Yes. Why? As an extreme, think about n = 1. Then $\hat{\theta}_2 = 0$; probably an underestimate!

Also, consider n = 2. Then $\hat{\theta}_1$ is exactly between the two sample points, the position that exactly minimizes the expression for θ_2. Any other choices for θ_1, θ_2 make the likelihood of the observed data slightly *lower*. But it's actually pretty unlikely that two sample points would be chosen exactly equidistant from, and on opposite sides of the mean, so the MLE $\hat{\theta}_2$ systematically underestimates θ_2.

(But not by much, & bias shrinks with sample size.)

Summary

MLE is *one* way to estimate *parameters* from *data*

You choose the *form* of the model (normal, binomial, ...)

Math chooses the *value(s)* of parameter(s)

Has the intuitively appealing property that the parameters maximize the *likelihood* of the observed data; basically just assumes your sample is “representative”

- Of course, unusual samples will give bad estimates (estimate normal human heights from a sample of NBA stars?) but that is an unlikely event

Often, but not always, MLE has other desirable properties like being *unbiased,* or at least *consistent*

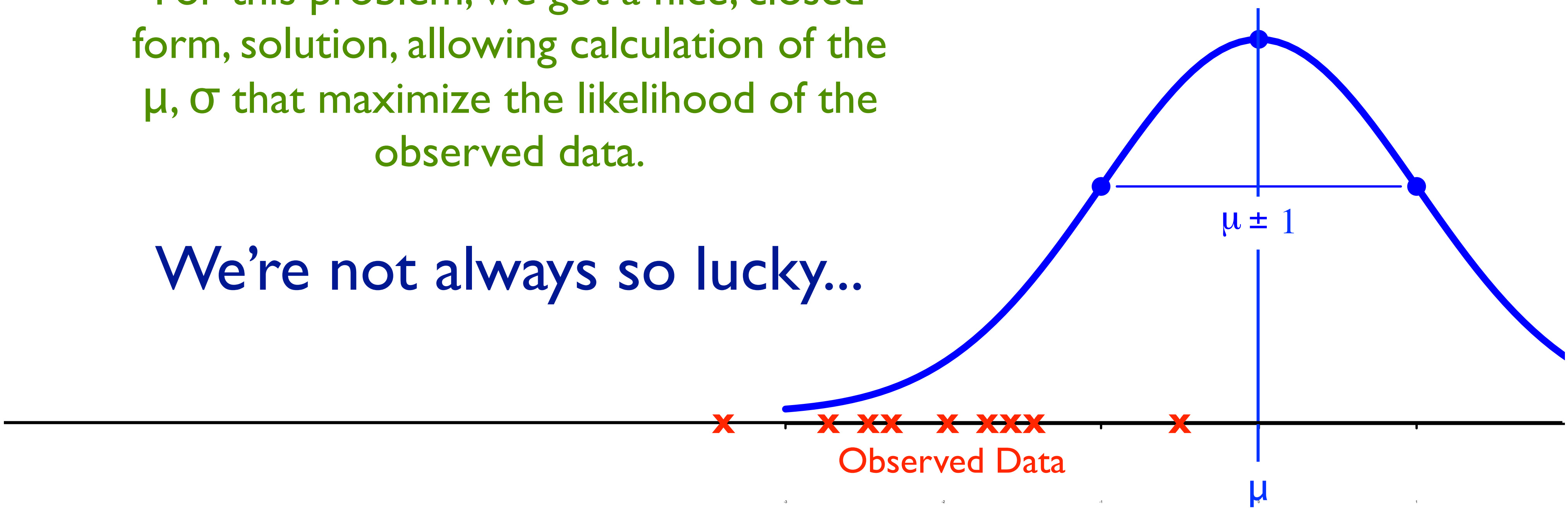
EM

The Expectation-Maximization
Algorithm

Above:
How to estimate μ given data

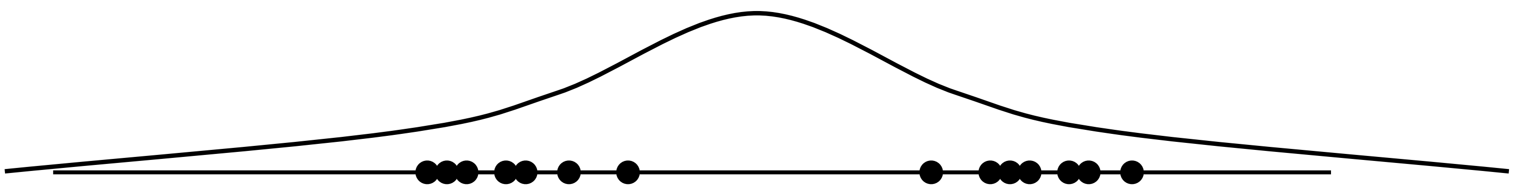
For this problem, we got a nice, closed form, solution, allowing calculation of the μ, σ that maximize the likelihood of the observed data.

We're not always so lucky...

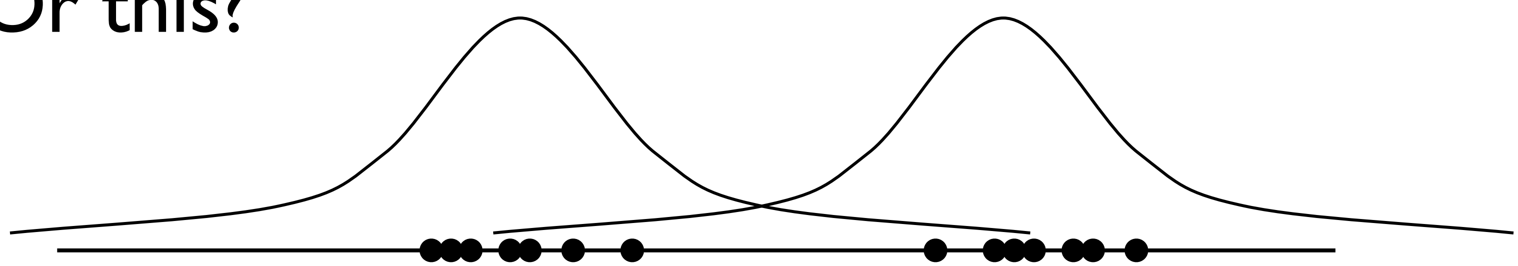


More Complex Example

This?



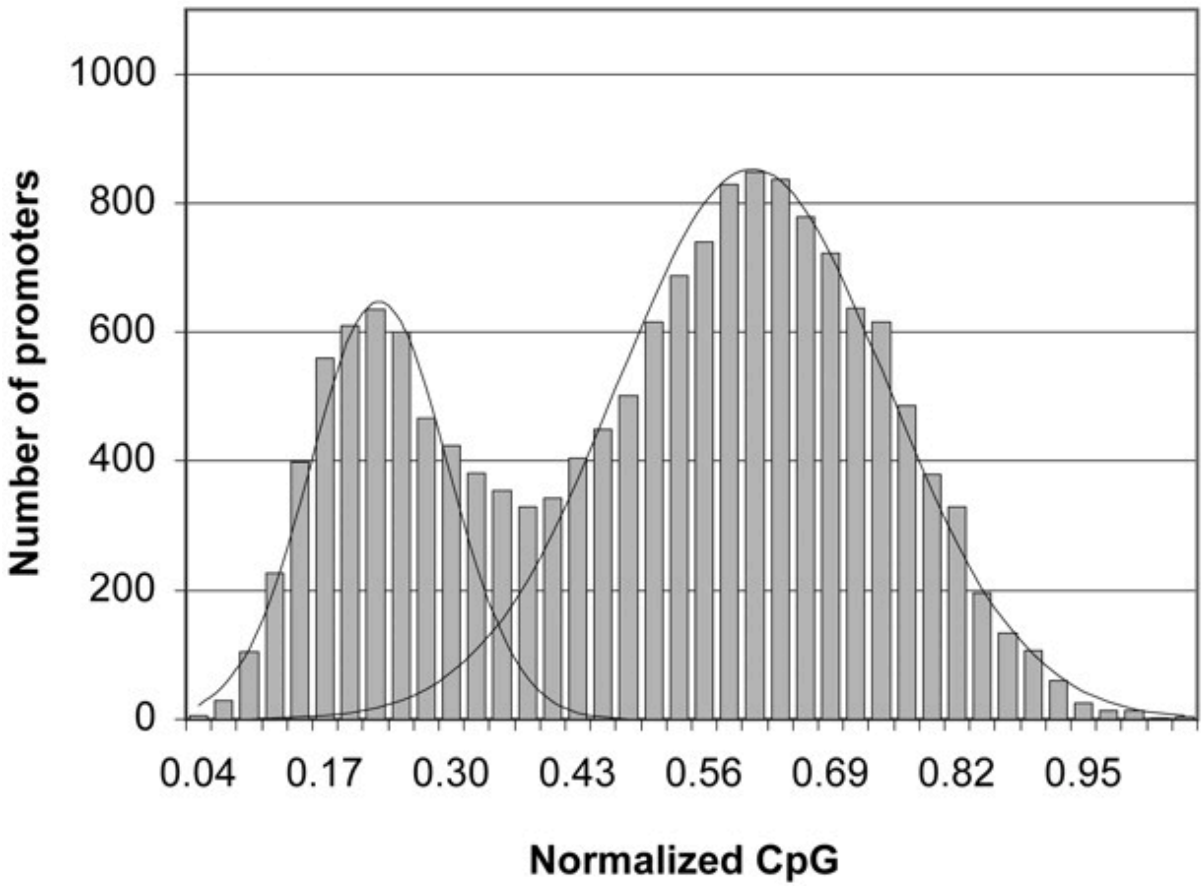
Or this?



(A modeling decision, not a math problem...,
but if later, what math?)

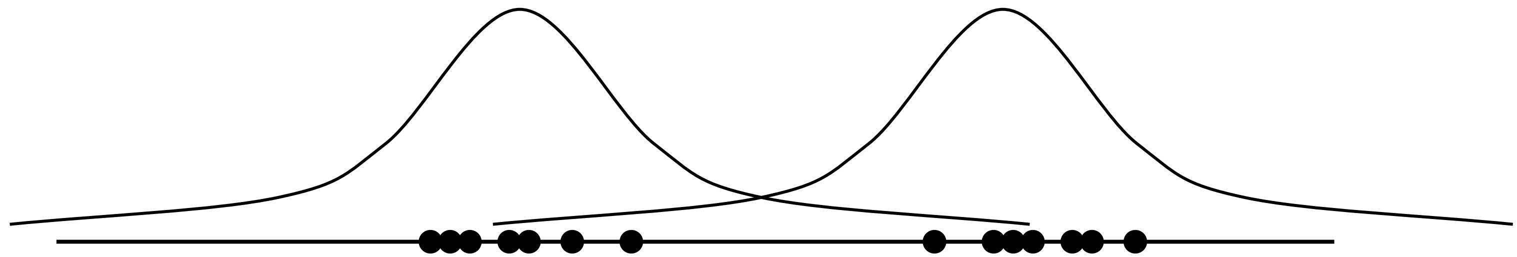
A Real Example:

CpG content of human gene promoters



"A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters" Saxonov, Berg, and Brutlag, PNAS 2006;103:1412-1417

Gaussian Mixture Models / Model-based Clustering



Parameters θ

means	μ_1	μ_2
variances	σ_1^2	σ_2^2
mixing parameters	τ_1	$\tau_2 = 1 - \tau_1$

P.D.F. $f(x|\mu_1, \sigma_1^2)$ $f(x|\mu_2, \sigma_2^2)$

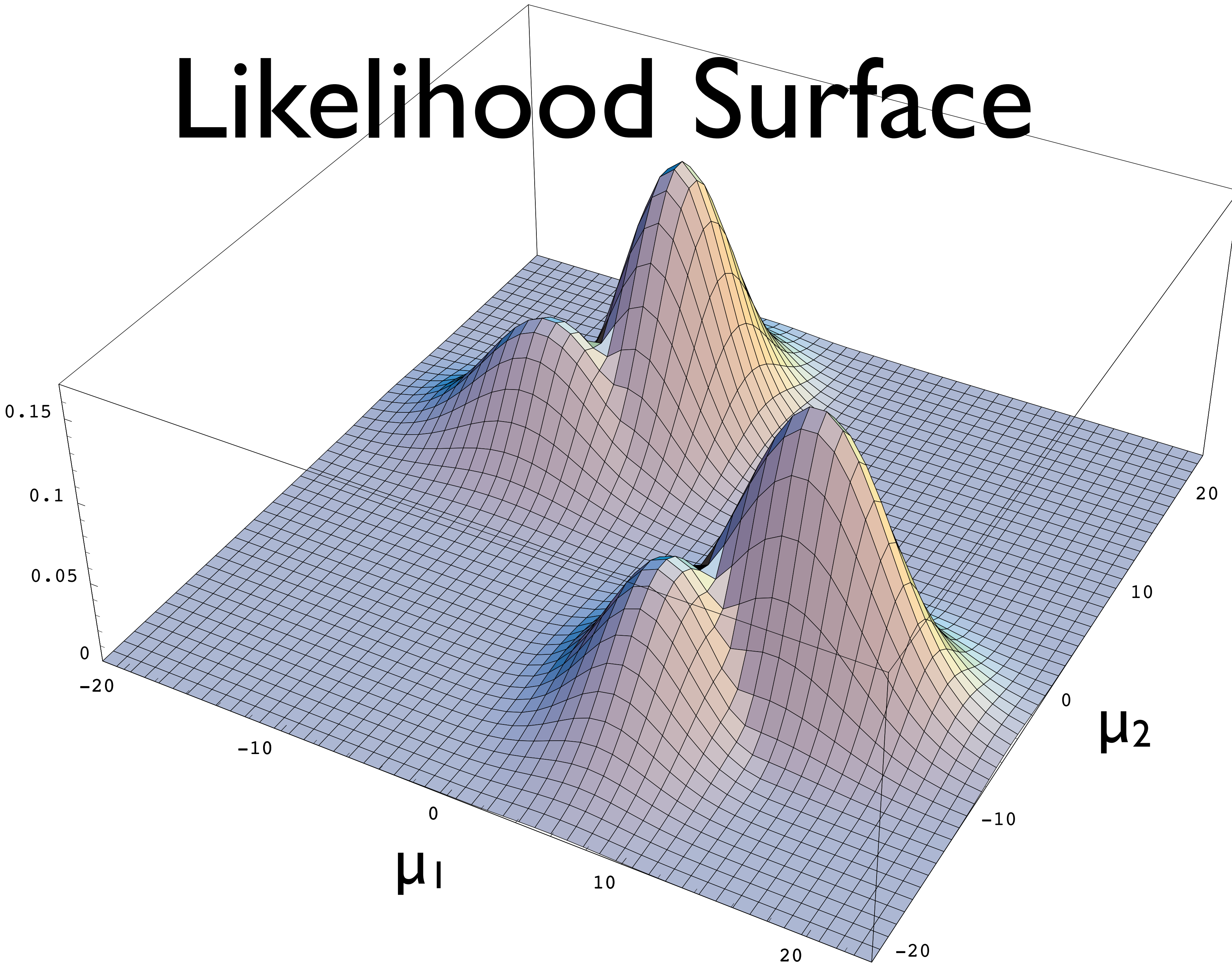
Likelihood

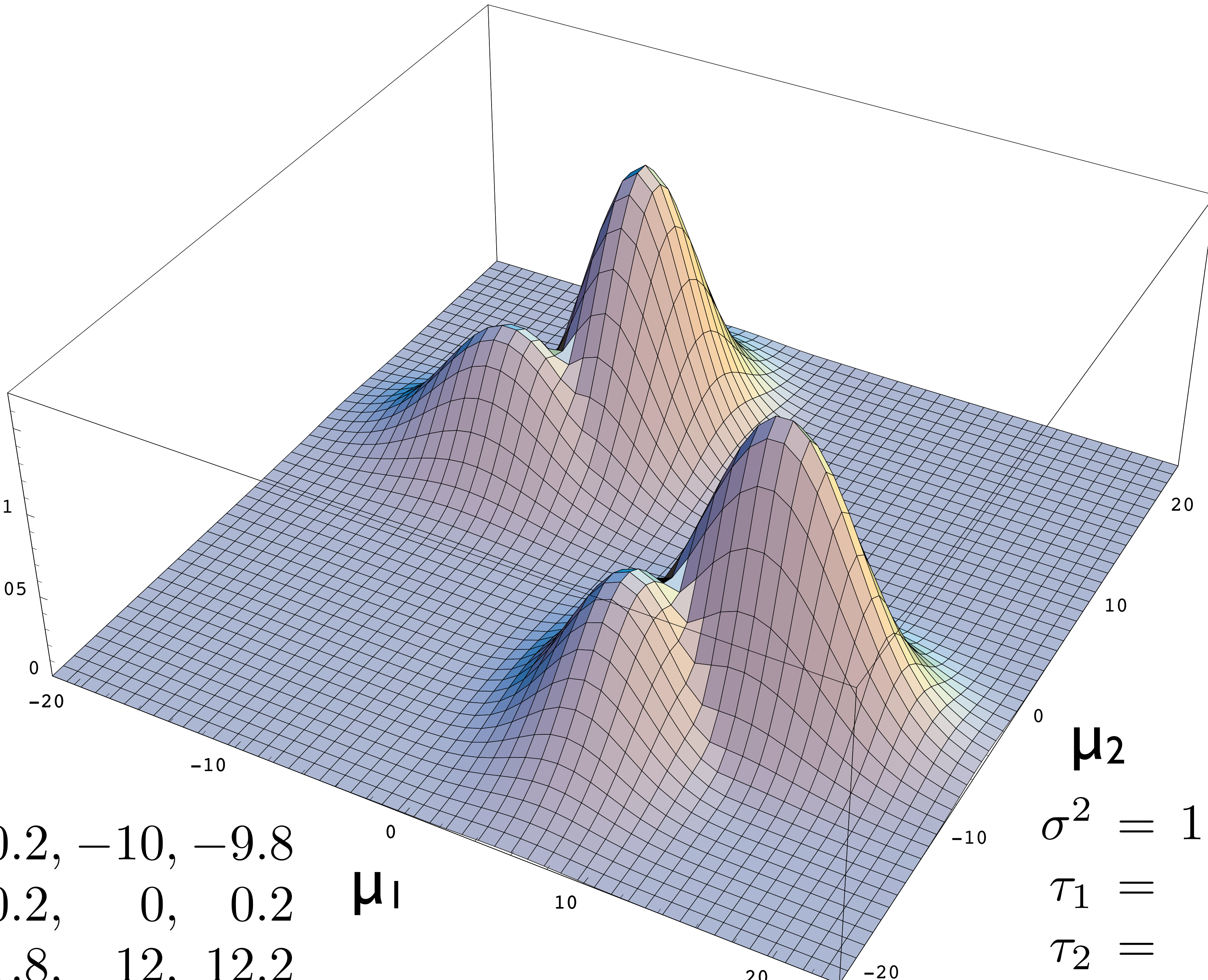
$$L(x_1, x_2, \ldots, x_n|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2)$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i|\mu_j, \sigma_j^2)$$

No closed-form max

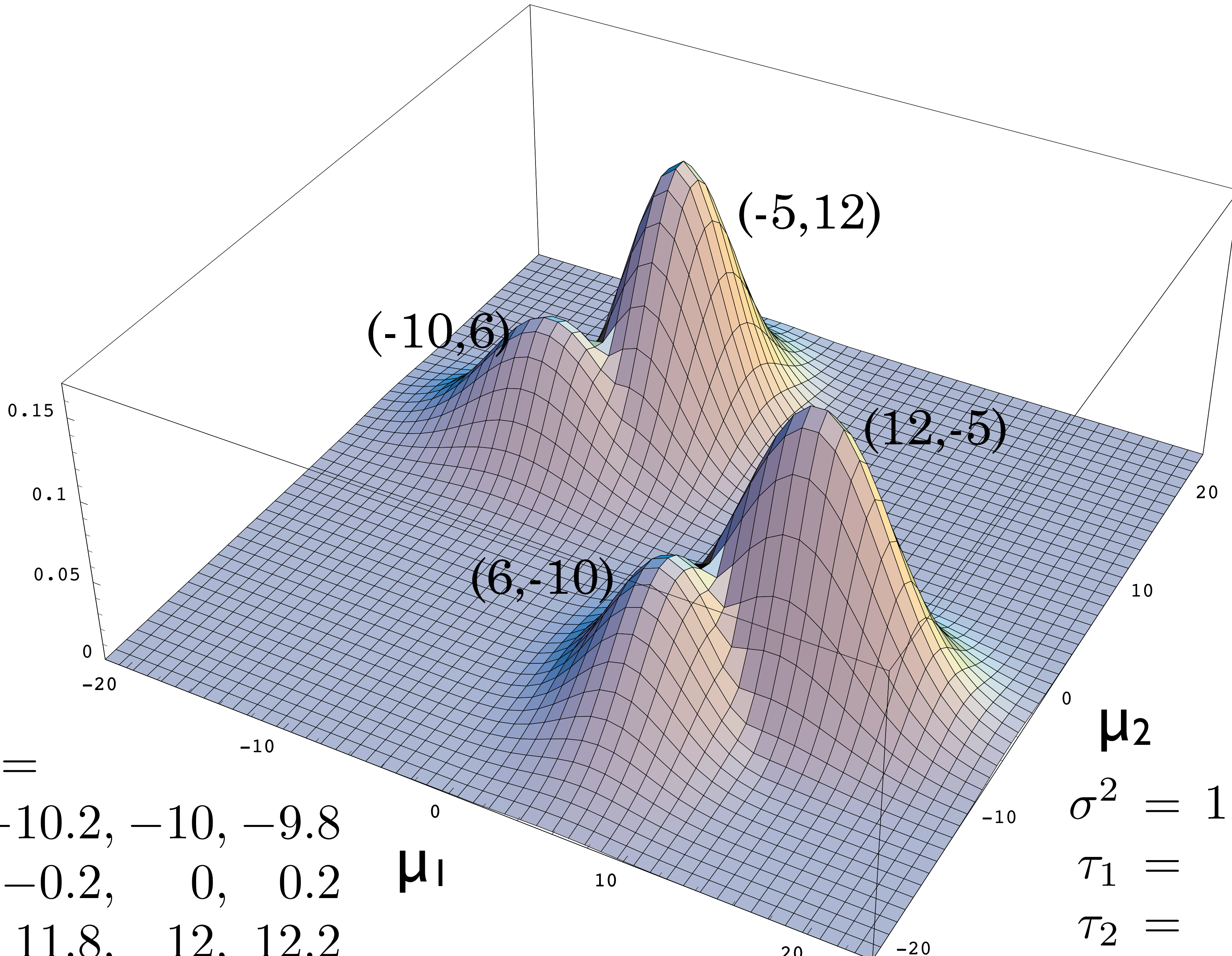
Likelihood Surface





$$x_i = \begin{array}{rrr} -10.2, & -10, & -9.8 \\ -0.2, & 0, & 0.2 \\ 11.8, & 12, & 12.2 \end{array}$$

$$\sigma^2 = 1.0$$
$$\tau_1 = .5$$
$$\tau_2 = .5$$



$$x_i = \begin{matrix} -10.2, & -10, & -9.8 \\ -0.2, & 0, & 0.2 \\ 11.8, & 12, & 12.2 \end{matrix}$$

$$\sigma^2 = 1.0$$
$$\tau_1 = .5$$
$$\tau_2 = .5$$

A What-If Puzzle

Likelihood

$$L(x_1, x_2, \ldots, x_n | \overbrace{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2}^{\theta})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)$$

Messy: no closed form solution known for finding θ maximizing L

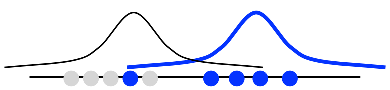
But *what if* we knew the *hidden data*?

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

EM as Egg vs Chicken

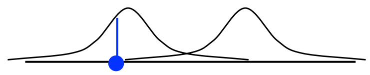
IF z_{ij} known, could estimate parameters θ

- E.g., only points in cluster 2 influence μ_2, σ_2



IF parameters θ known, could estimate z_{ij}

- E.g., if $|x_i - \mu_1|/\sigma_1 << |x_i - \mu_2|/\sigma_2$, then $z_{i1} >> z_{i2}$



But we know neither; (optimistically) iterate:

- E: calculate expected z_{ij}, given parameters
- M: calc "MLE" of parameters, given $E(z_{ij})$

Overall, a clever "hill-climbing" strategy

Simple Version: "Classification EM"

If z_{ij} < .5, pretend it's 0; z_{ij} > .5, pretend it's 1

I.e., *classify* points as component 0 or 1

Now recalc θ, assuming that partition

Then recalc z_{ij} , assuming that θ

Then re-recalc θ, assuming new z_{ij}, etc., etc.

"Full EM" is a bit more involved, but this is the crux.

Full EM

x_i's are known; θ unknown. Goal is to find MLE θ of:

$$L(x_1, \ldots, x_n \mid \theta) \qquad \text{(hidden data likelihood)}$$

Would be easy *if* z_{ij}'s were known, i.e., consider:

$$L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta) \qquad \text{(complete data likelihood)}$$

But z_{ij}'s aren't known.

Instead, maximize *expected* likelihood of visible data

$$E(L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta)),$$

where expectation is over distribution of hidden data (z_{ij}'s)

The E-step:

Find $E(Z_{ij})$, i.e. $P(Z_{ij}=1)$

Assume θ known & fixed

A (B): the event that x_i was drawn from f_1 (f_2)

D: the observed datum x_i

Expected value of z_{i1} is P(A|D) — $E = 0 \cdot P(0) + 1 \cdot P(1)$

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)}$$

$$\begin{aligned} P(D) &= P(D|A)P(A) + P(D|B)P(B) \\ &= f_1(x_i|\theta_1)\,\tau_1 + f_2(x_i|\theta_2)\,\tau_2 \end{aligned}$$

Repeat for each x_i

Complete Data Likelihood

Recall:

$$z_{1j} \;=\; \begin{cases} 1 & \text{if } x_1 \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

so, correspondingly,

$$L(x_1, z_{1j} \mid \theta) \;=\; \begin{cases} \tau_1 f_1(x_1 \mid \theta) & \text{if } z_{11} = 1 \\ \tau_2 f_2(x_1 \mid \theta) & \text{otherwise} \end{cases}$$

Formulas with "if's" are messy; can we blend more smoothly?
Yes, many possibilities. Idea 1:

$$L(x_1, z_{1j} \mid \theta) \;=\; z_{11} \cdot \tau_1 f_1(x_1 \mid \theta) + z_{12} \cdot \tau_2 f_2(x_1 \mid \theta)$$

Idea 2 (Better):

$$L(x_1, z_{1j} \mid \theta) \;=\; (\tau_1 f_1(x_1 \mid \theta))^{z_{11}} \cdot (\tau_2 f_2(x_1 \mid \theta))^{z_{12}}$$

M-step:

Find θ maximizing E(log(Likelihood))

(For simplicity, assume $\sigma_1 = \sigma_2 = \sigma; \tau_1 = \tau_2 = .5 = \tau$)

$$L(\vec{x}, \vec{z} \mid \theta) = \prod_{1 \le i \le n} \frac{\tau}{\sqrt{2\pi\sigma^2}} \exp\left(- \sum_{1 \le j \le 2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right)$$

$$E[\log L(\vec{x}, \vec{z} \mid \theta)] = E\left[\sum_{1 \le i \le n} \left(\log \tau - \frac{1}{2} \log 2\pi\sigma^2 - \sum_{1 \le j \le 2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right) \right]$$

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

Find θ maximizing this as before, using $E[z_{ij}]$ found in E-step. Result:

$$\boxed{\mu_j = \sum_{i=1}^{n} E[z_{ij}] x_i / \sum_{i=1}^{n} E[z_{ij}]}$$

(intuit: avg, weighted by subpop prob)

2 Component Mixture

$$\sigma_1 = \sigma_2 = 1; \ \tau = 0.5$$

			mu1	-20.00		-6.00		-5.00		-4.99
			mu2	6.00		0.00		3.75		3.75
x1	**-6**		**z11**		5.11E-12		1.00E+00		1.00E+00	
x2	**-5**		**z21**		2.61E-23		1.00E+00		1.00E+00	
x3	**-4**		**z31**		1.33E-34		9.98E-01		1.00E+00	
x4	**0**		**z41**		9.09E-80		1.52E-08		4.11E-03	
x5	**4**		**z51**		6.19E-125		5.75E-19		2.64E-18	
x6	**5**		**z61**		3.16E-136		1.43E-21		4.20E-22	
x7	**6**		**z71**		1.62E-147		3.53E-24		6.69E-26	

Essentially converged in 2 iterations

(Excel spreadsheet on course web)

Applications

Clustering is a remarkably successful exploratory data analysis tool

- Web-search, information retrieval, gene-expression, ...
- Model-based approach above is one of the leading ways to do it

Gaussian mixture models widely used

- With many components, empirically match arbitrary distribution
- Often well-justified, due to "hidden parameters" driving the visible data

EM is extremely widely used for "hidden-data" problems

- Hidden Markov Models

EM Summary

Fundamentally a maximum likelihood parameter estimation problem

Useful if hidden data, and if analysis is more tractable when 0/1 hidden data z known

Iterate:

E-step: estimate E(z) for each z, given θ
M-step: estimate θ maximizing E(log likelihood) given E(z) [where "E(logL)" is wrt random z ~ E(z) = p(z=1)]

EM Issues

Under mild assumptions (sect 11.6), EM is guaranteed to increase likelihood with every E-M iteration, hence will *converge*.

But it may converge to a *local*, not global, max. (Recall the 4-bump surface...)

Issue is intrinsic (probably), since EM is often applied to *NP-hard* problems (including clustering, above and motif-discovery, soon)

Nevertheless, widely used, often effective