PNI Bootcamp

Basics of Statistics and Model Fitting

Stephen Keeley

Outline

- Basics of probability notation
- Models, parameters
- Sampling and inference
- Distributions
- Marginalization/conditionalization
    - Break
- Gaussian and Poisson neuron estimation problems
- MLE

Basics of Probability

- An **event (or sample)** is an outcome or set of outcomes from a random process
    - ex: tossing a coin three times
        - Event A = getting exactly two heads = {HTH, HHT, THH}
    - Rolling dice
        - Event A = result is even = {2,4,6}

- An **sample space**, ***S*,** is the set of all possible outcomes (events) of a process (random variable)
  - ex: 6-sided dice = {1, 2, 3, 4, 5, 6 }
- In it's most basic form, when all outcomes have the same probability, we can write

$$P(A) = \frac{\text{\# of outcomes in A}}{\text{\# of outcomes in S}}$$

Probability facts to remember

- $P(A)$ is always bounded between 0 and 1
- $P(A,B)$ is the probability of both event A and event B occurring
- $P(A,B) = P(A)*P(B)$ if the events are *independent*
- Otherwise , $P(A,B) = P(A|B)*P(B)$
  - That is, the probability of event A and B occurring is the probability of event A given event B occurs times the probability event B occurs
  - Can also be written the other way around!
- In other words $P(A|B) = P(A)$ when the events are independent

- In statistical modeling, we wish to describe the probability of ***all possible events***, to understand the process that generates the events

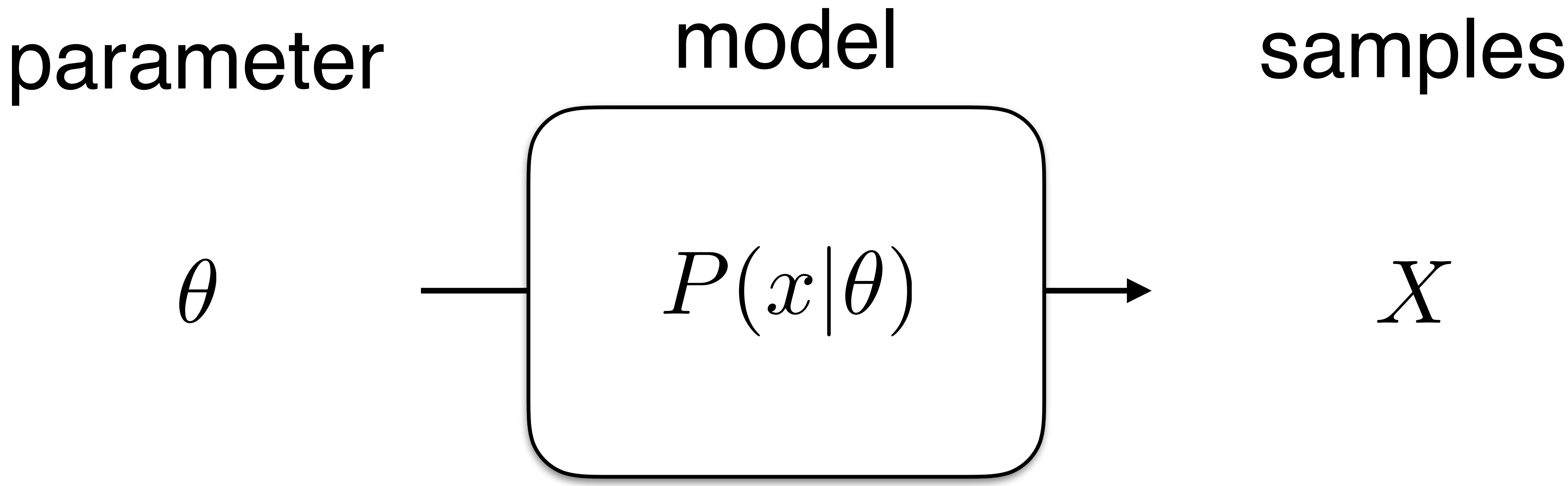
- We use a **model**

- Models are functions of a variable that represent all possible values of **events** in a given set, **S**

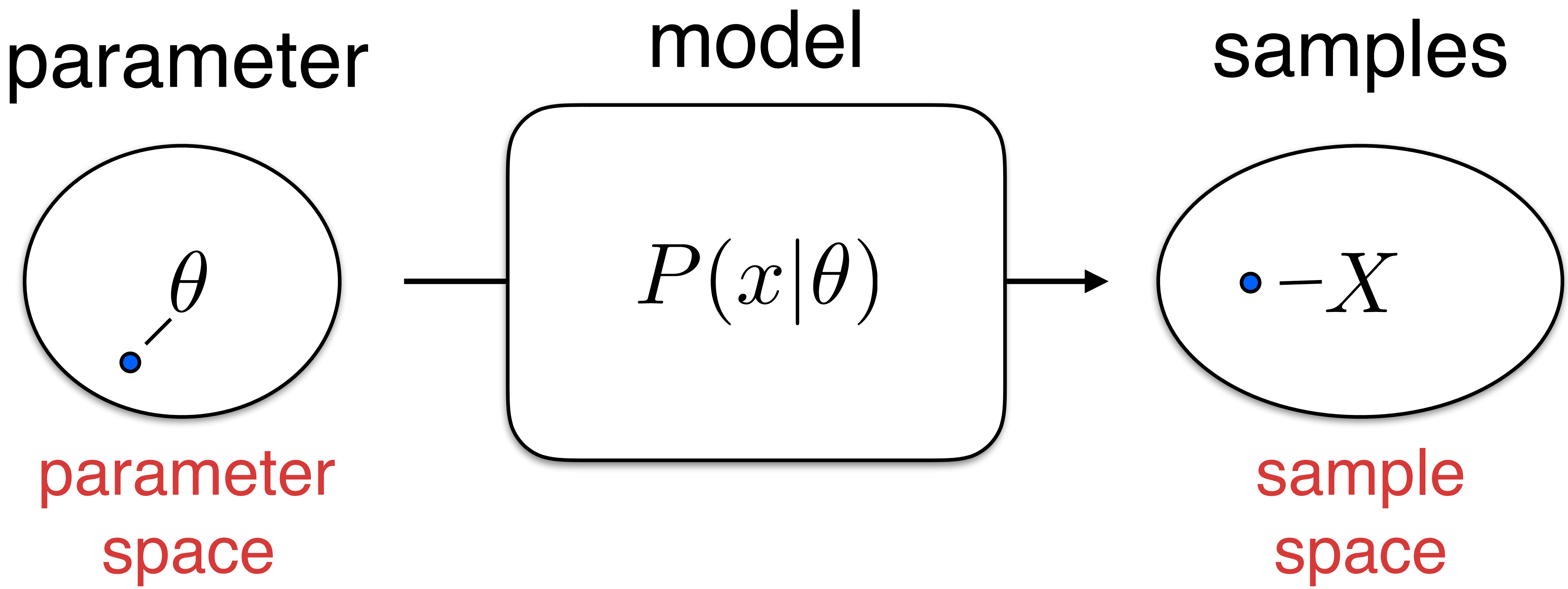
- Models include **parameters**, ***θ,*** that describe the shape of the mapping from events to probabilities

- These functions generate a probability distribution over possible event values

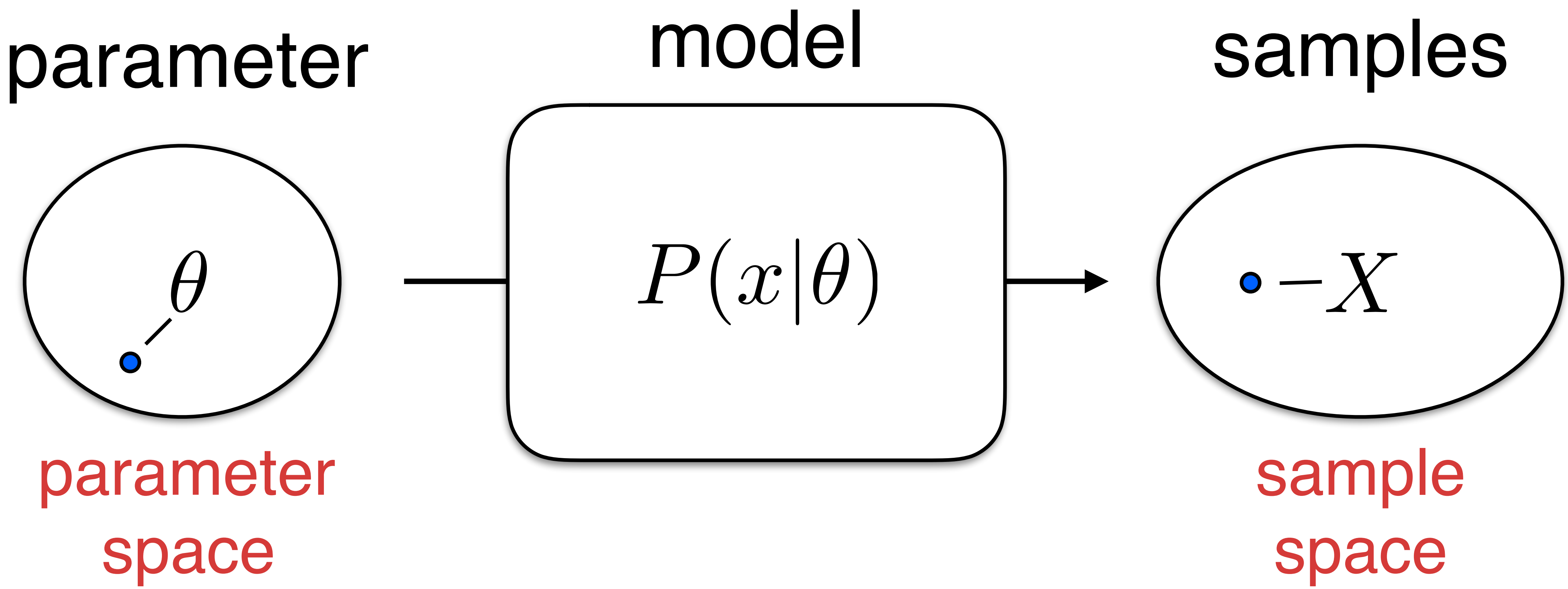
Big Picture



- “probability distribution”
- “events”
- “random variables”



- “probability distribution”
- “events”
- “random variables”



<u>examples</u>

1. coin flipping

$\theta = p(\text{“heads”})$

X = “H” or “T”

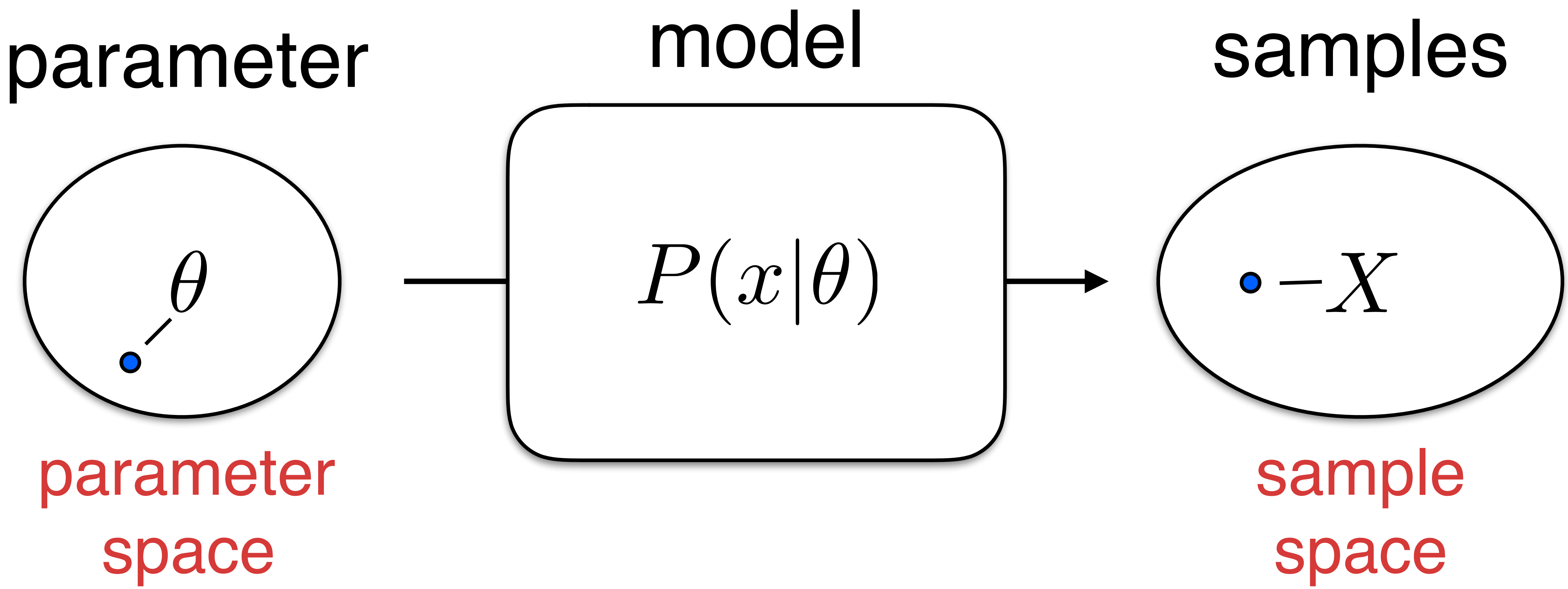
2. spike counts

$\theta =$ mean spike rate

$X \in \{0, 1, \ldots \ldots\}$

- “probability distribution”

- “events”
- “random variables”



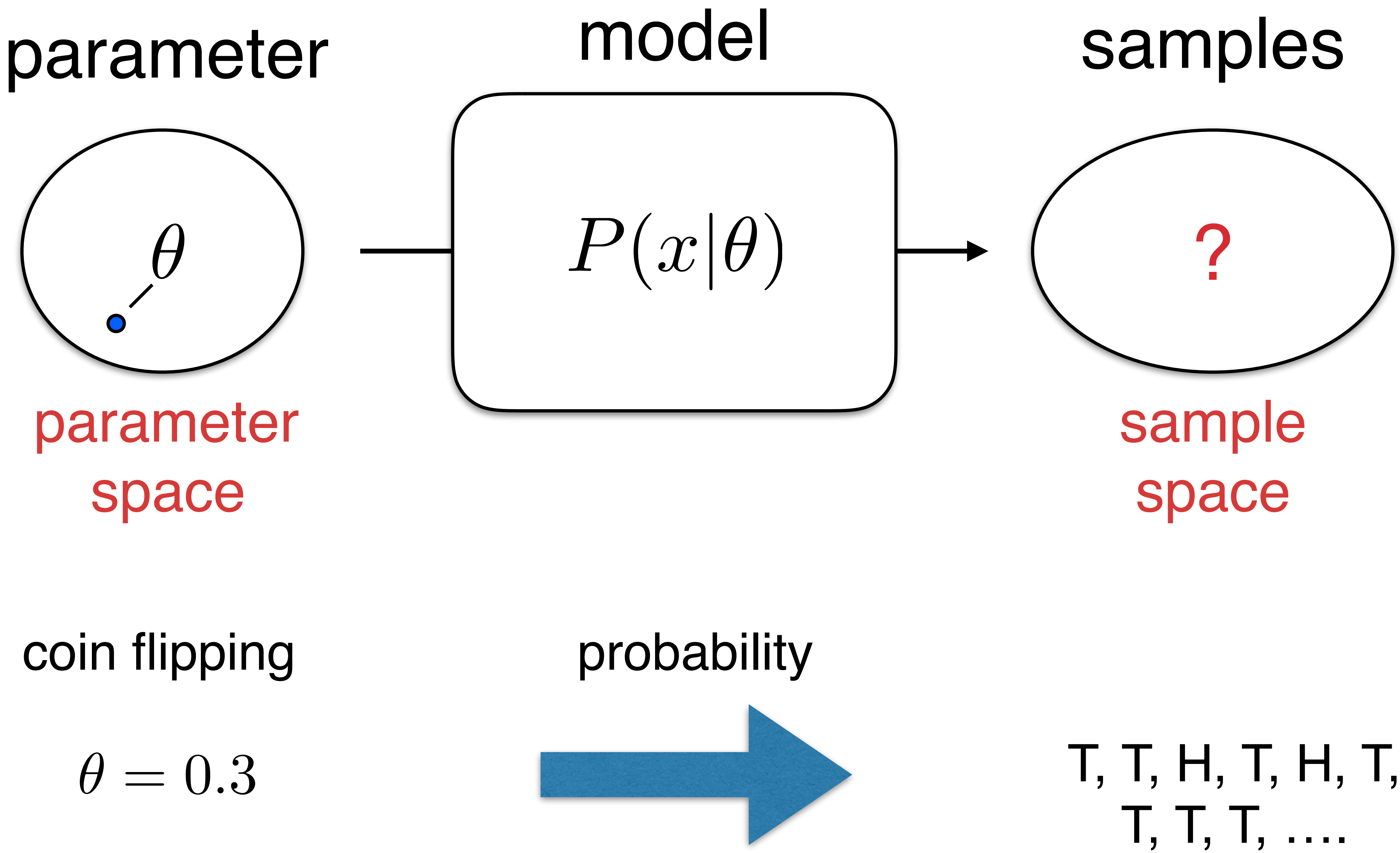
<u>examples</u>

3. reaction times

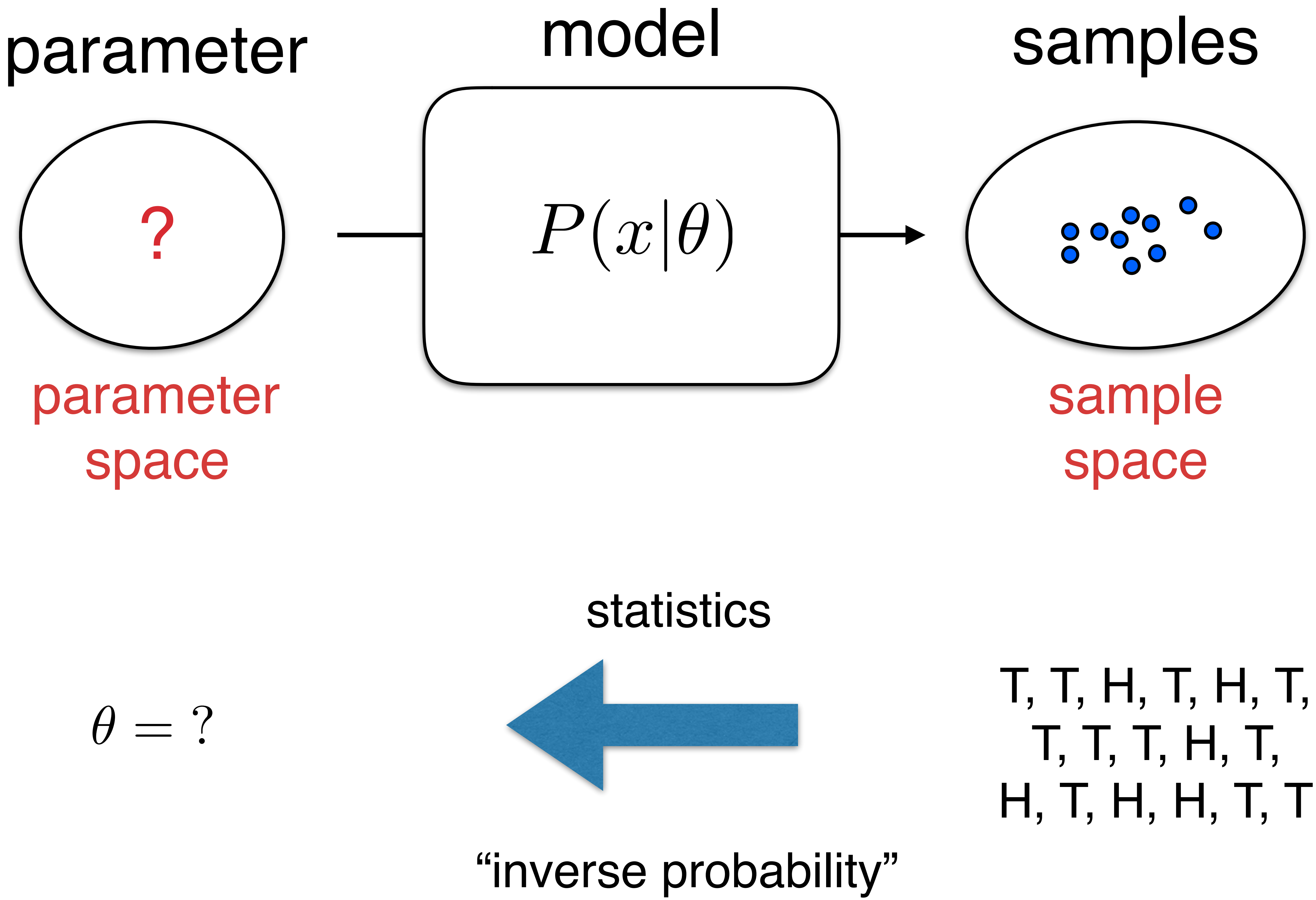
$\theta =$ mean reaction time

X ∈ positive reals

Probability vs. Statistics



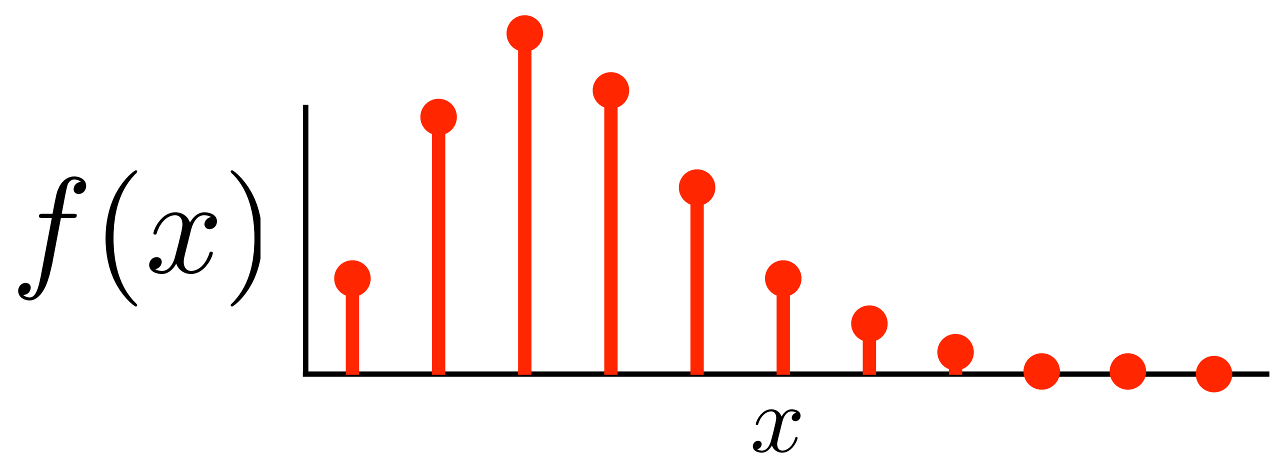
Probability vs. Statistics



discrete probability distribution

takes finite (or countably infinite) number of values, eg $x \in \mathbb{N}$

probability mass function (pmf):



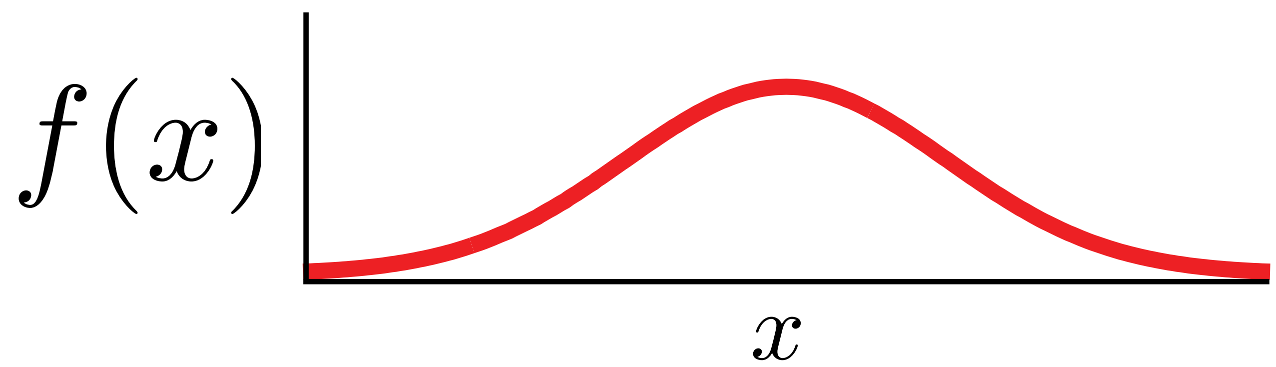
positive and sum to 1:

- $P(x = a) = f(a)$
- $\sum_{i=1}^{N} f(x_i) = 1$

continuous probability distribution

takes values in a continuous space, e.g., $x \in \mathbb{R}$

probability density function (pdf):



positive and integrates to 1:

- $P(x = a) = 0$
- $P(a < x < b) = \int_a^b f(x)\, dx$
- $\int_{-\infty}^{\infty} f(x)\, dx = 1$

some friendly neighborhood probability distributions

Discrete

Bernoulli $$P(x|\,p) = p^x \cdot (1-p)^{(1-x)}$$ coin flipping

binomial $$P(k|\,n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$ sum of n coin flips

Poisson $$P(k|\,\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$ sum of n coin flips with P(heads)=λ/n, in limit n→∞

some friendly neighborhood probability distributions

Continuous

Gaussian

$$P(x|\,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-u)^2}{2\sigma^2}\right]$$

multivariate Gaussian

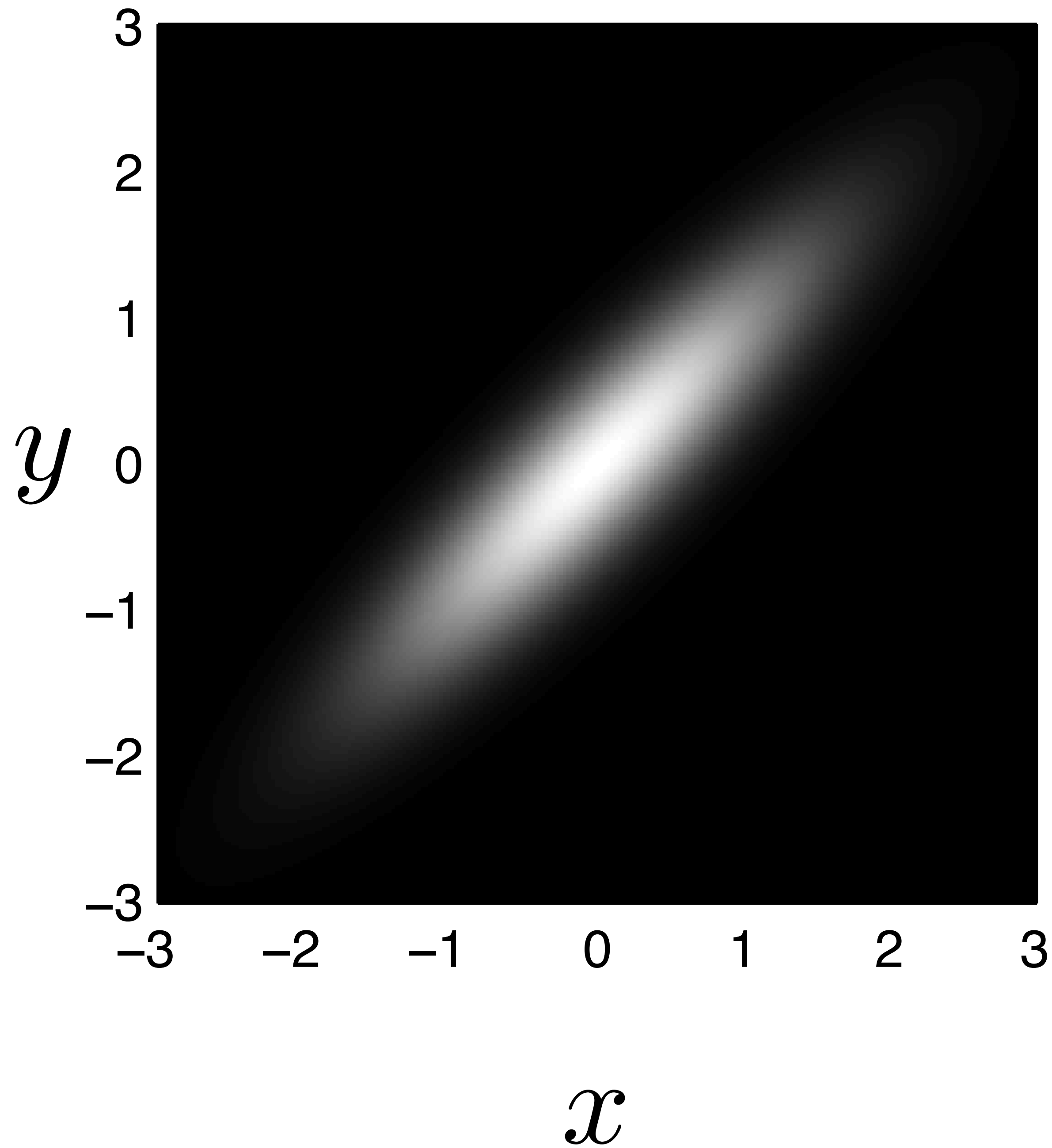
$$P(x_n|\,\mu,\Lambda) = \frac{1}{(2\pi)^{\frac{n}{2}}\,|\Lambda|^{\frac{1}{2}}}\exp\left[-\tfrac{1}{2}(\mathbf{x}-\mu)^T\Lambda^{-1}(\mathbf{x}-\mu)\right]$$

exponential

$$P(x|\,a) = ae^{-ax}$$

joint distribution

$$P(x, y)$$

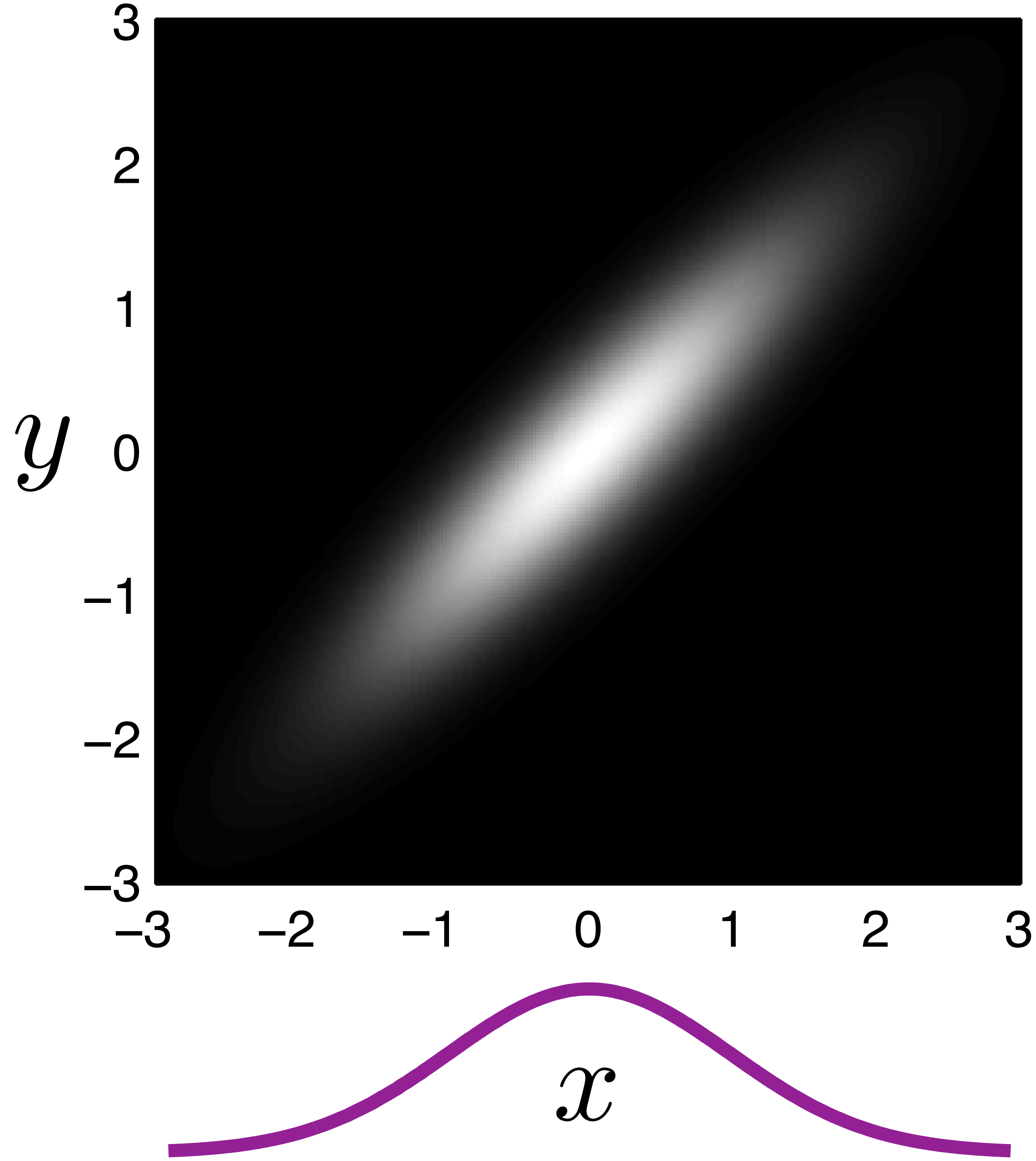


- positive
- sums to 1

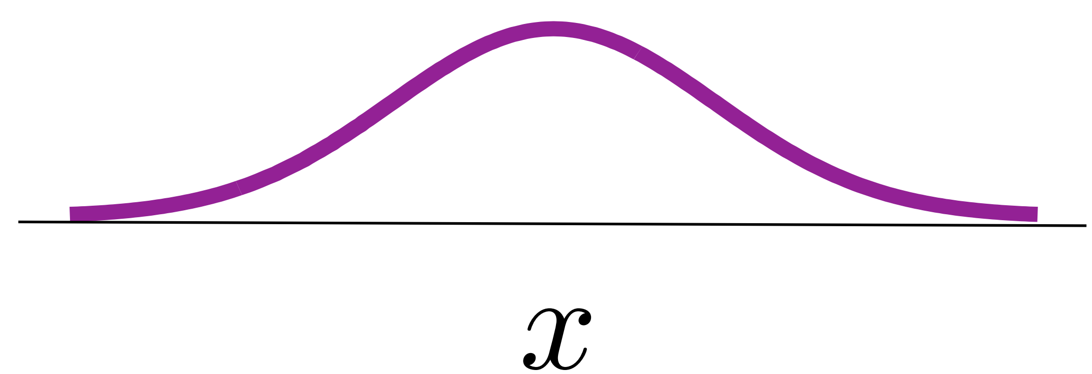
$$\iint P(x, y)\, dx\, dy = 1$$

marginalization ("integration")

$$P(x,y)$$

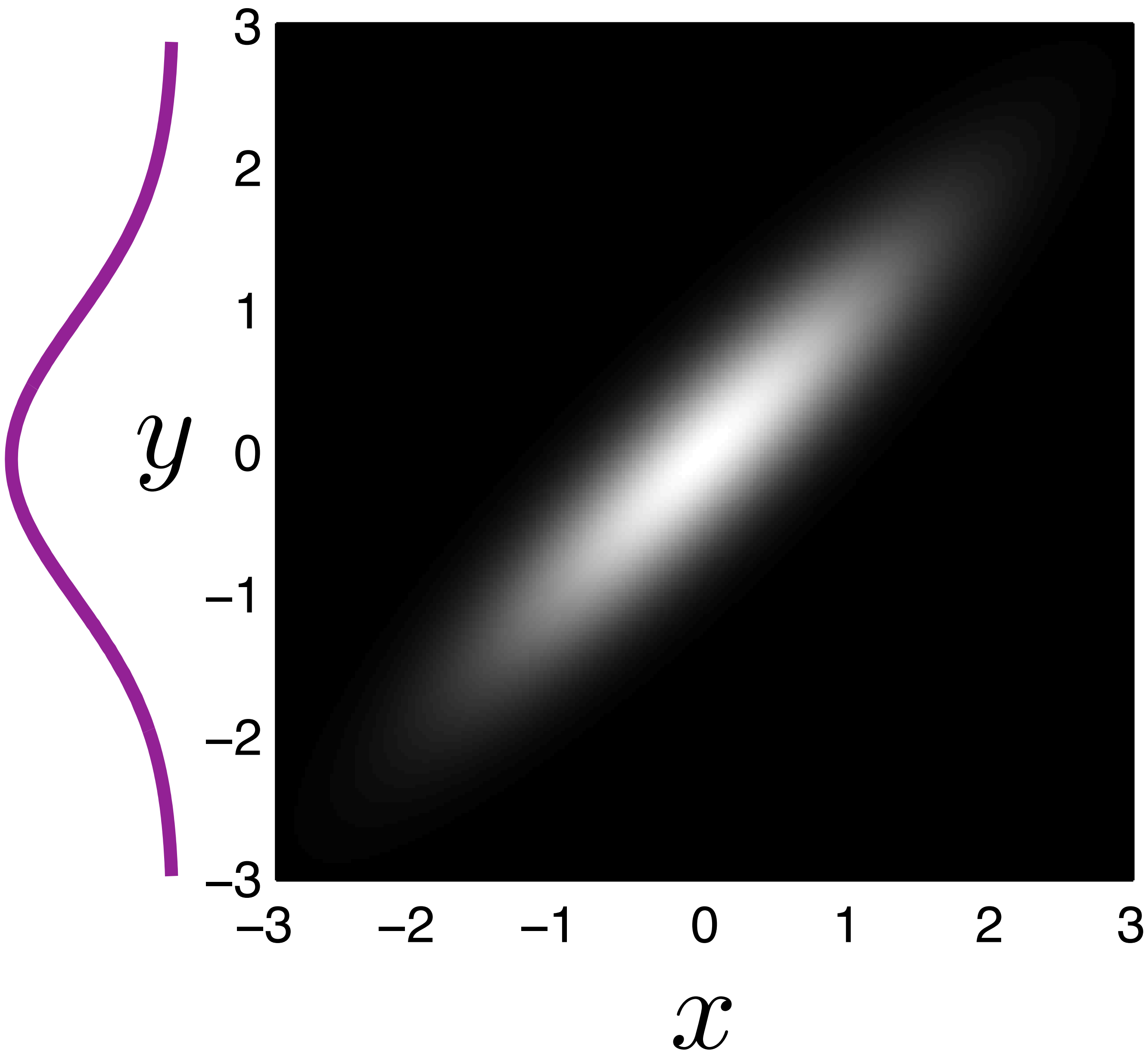


$$P(x) = \int P(x,y)dy$$

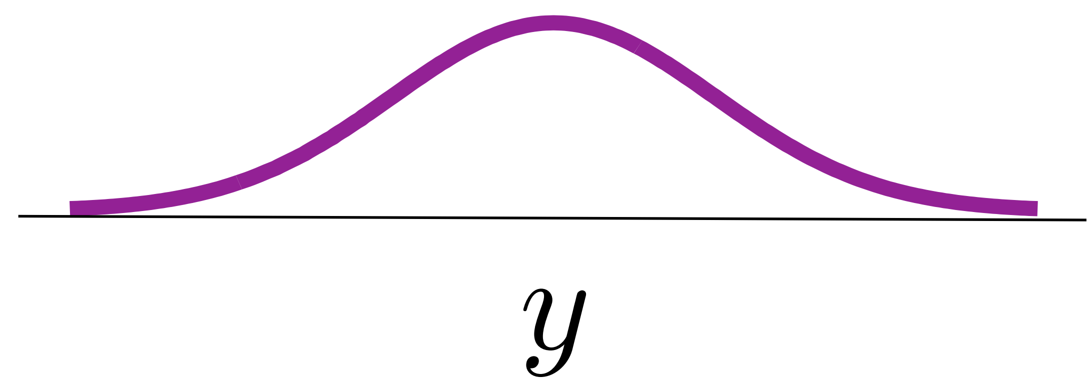


marginalization (“integration”)

$P(x, y)$



$$P(y) = \int P(x, y)dx$$



Joint Probability Distribution

$P(x, y)$

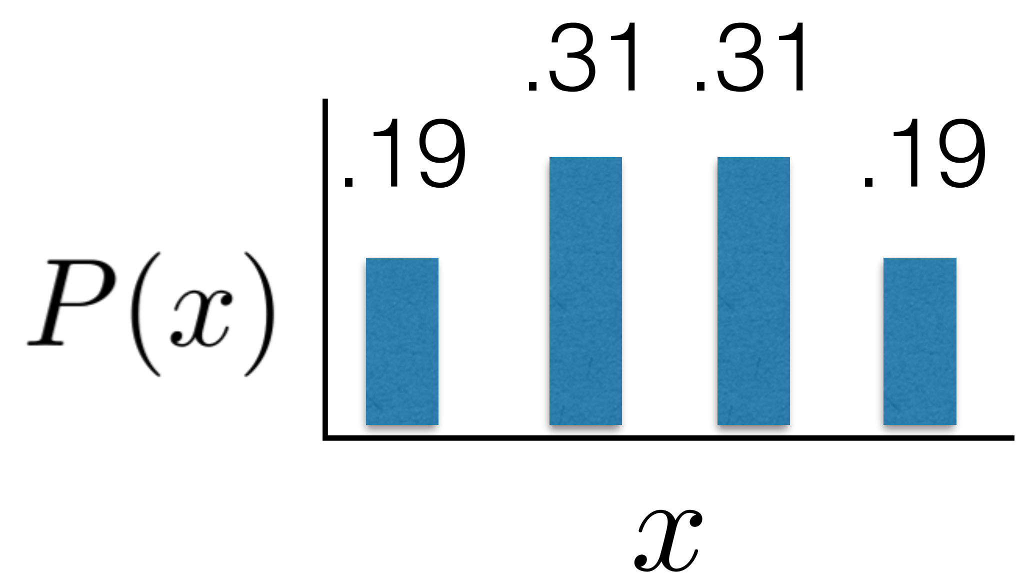
y

.01	.03	.07	.08
.03	.07	.14	.07
.07	.14	.07	.03
.08	.07	.03	.01

x

Sum (Marginalization)

Adds to one!



Conditionalization

$$P(x, y)$$

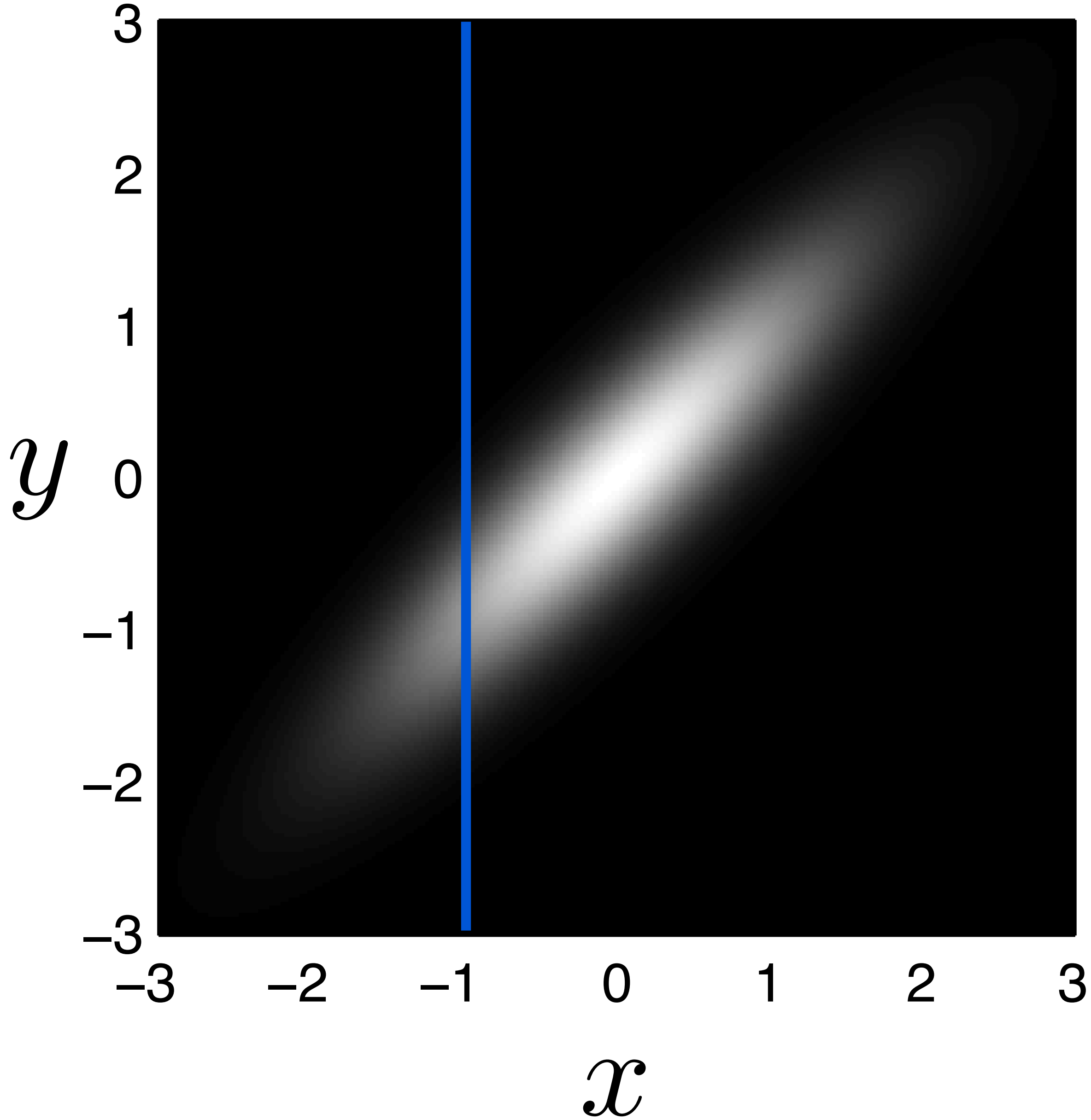
y \ x	-2	-1	1	2
	.01	.03	.07	.08
	.03	.07	.14	.07
	.07	.14	.07	.03
	.08	.07	.03	.01

$$P(y|x = -1) = \frac{P(y, x = -1)}{P(x = -1)}$$

("joint divided by marginal")

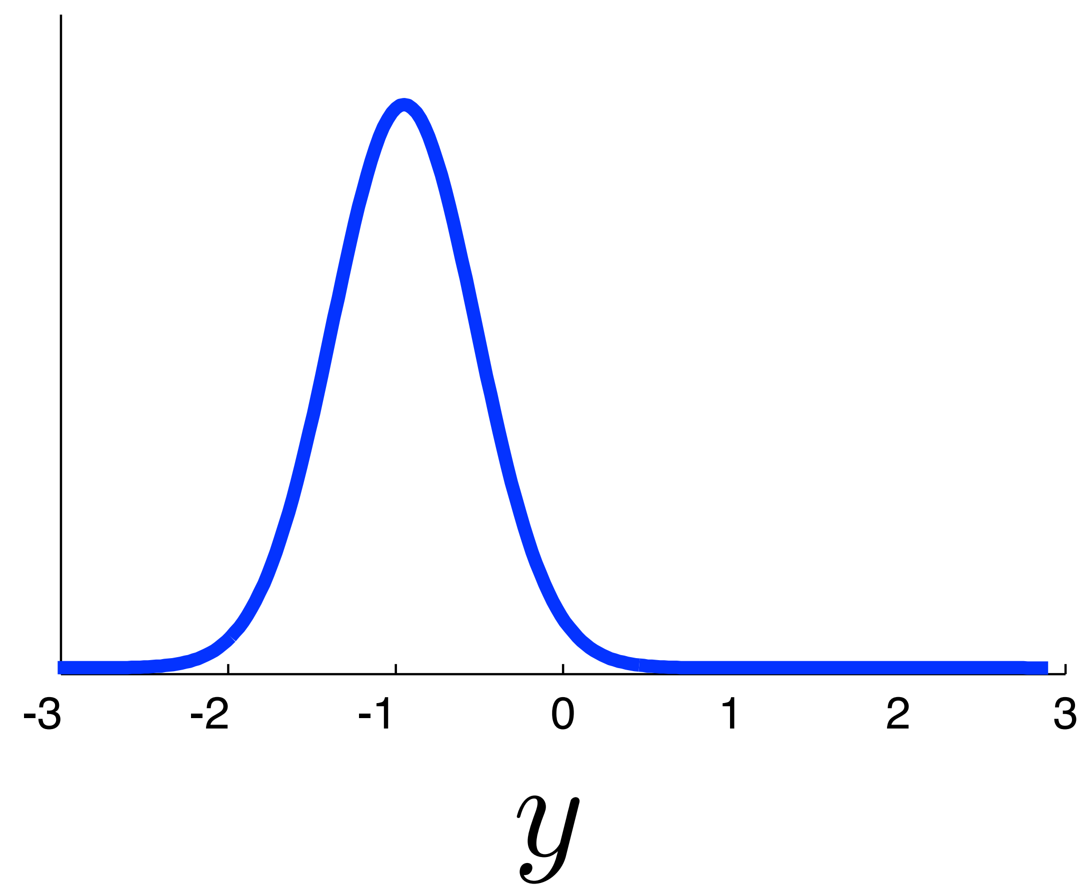
conditionalization ("slicing")

$$P(x, y)$$



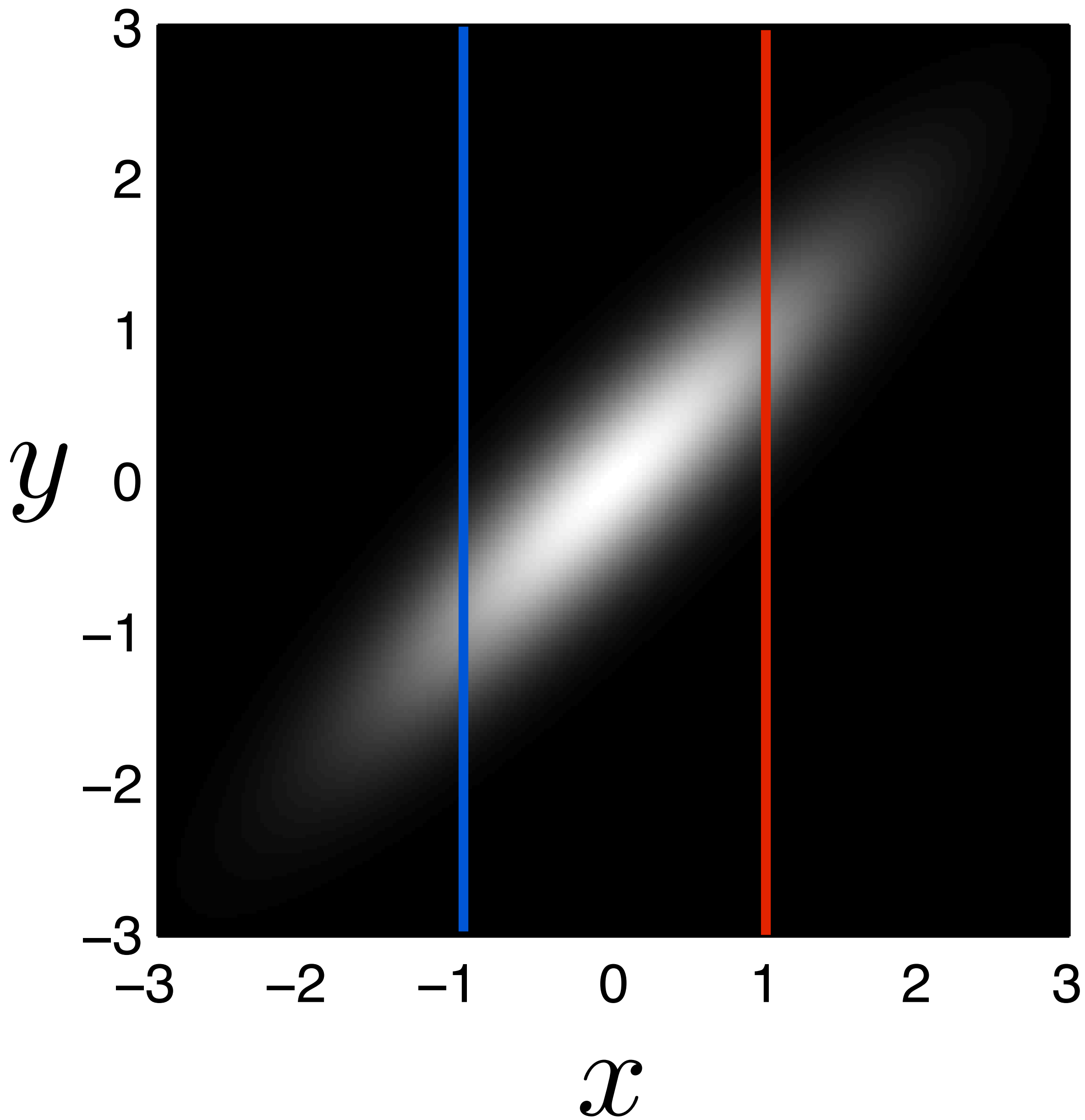
$$P(y|x = -1) = \frac{P(y, x = -1)}{P(x = -1)}$$

("joint divided by marginal")



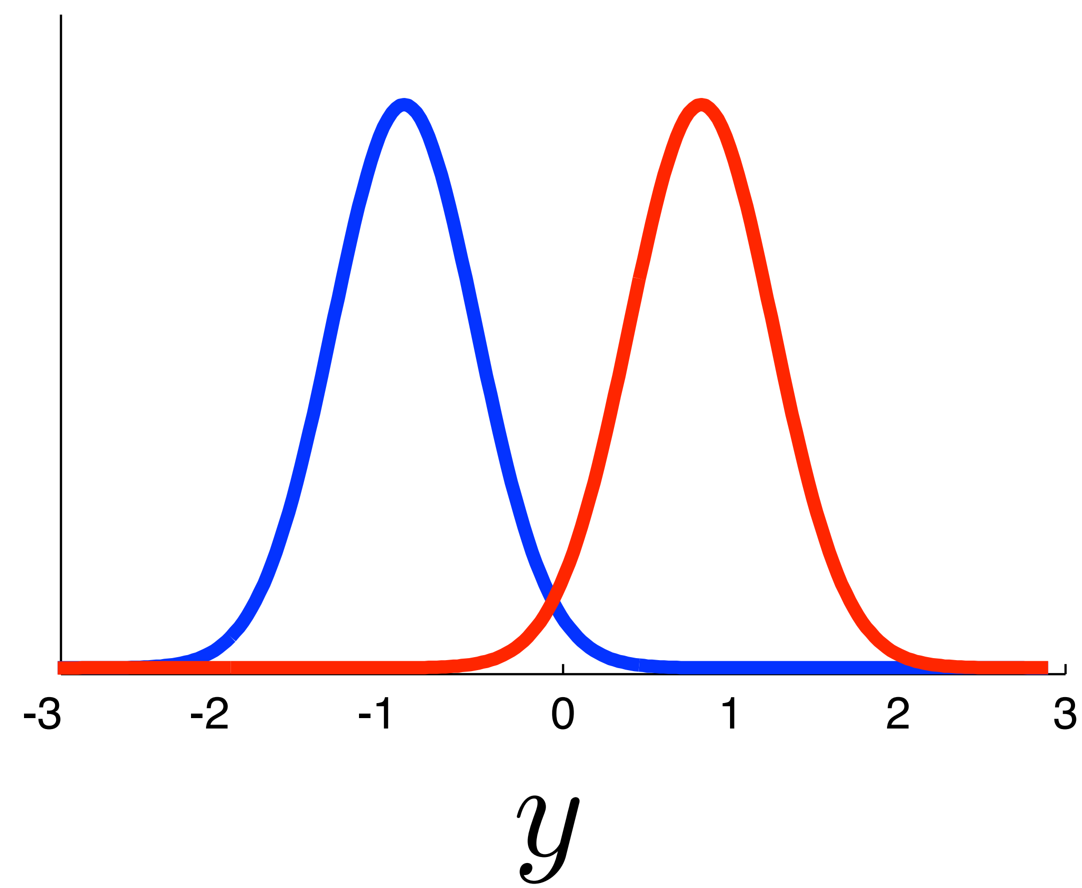
conditionalization (“slicing”)

$$P(x, y)$$

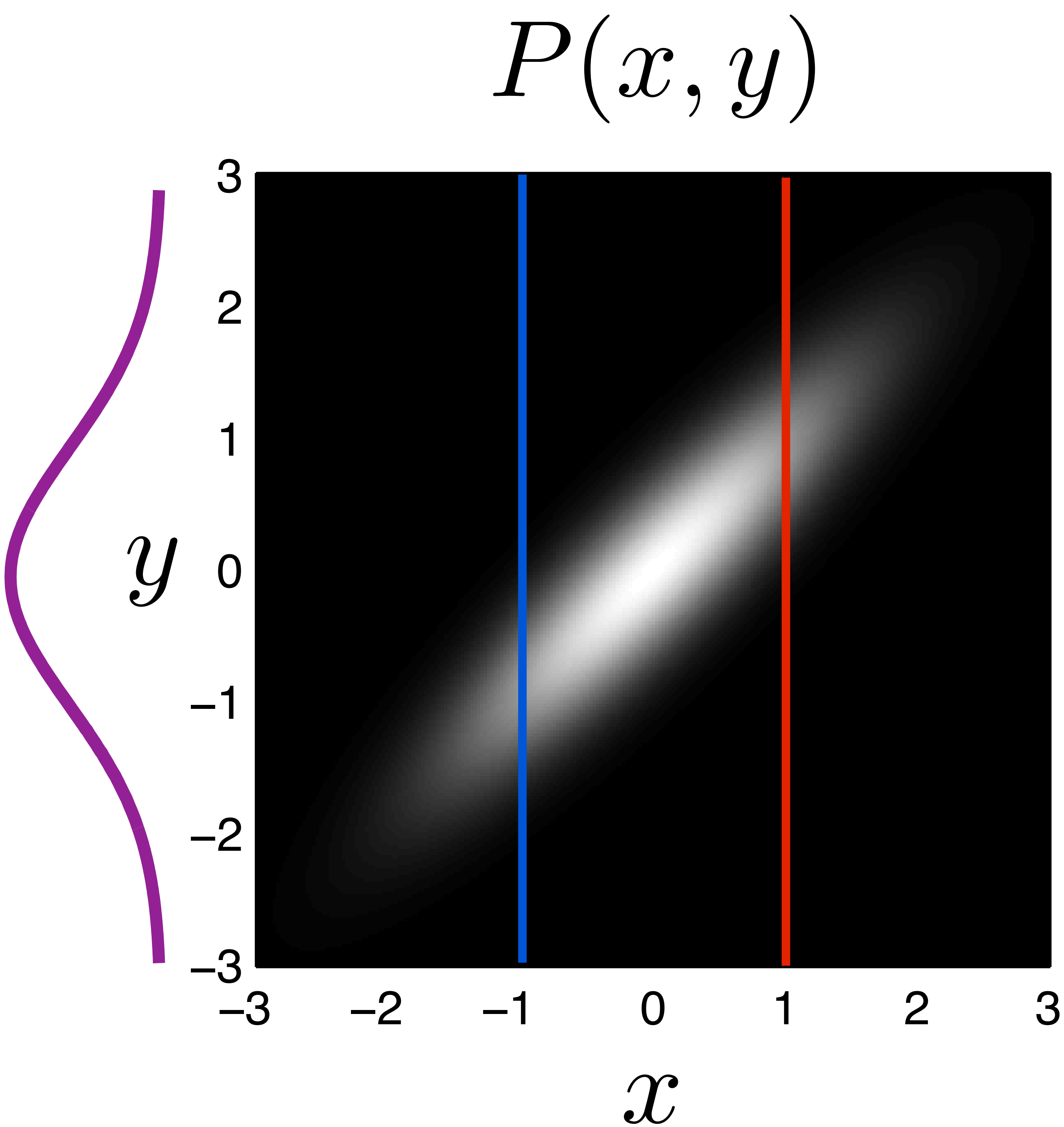


$$P(y|x=1) = \frac{P(y, x=1)}{P(x=1)}$$

(“joint divided by marginal”)

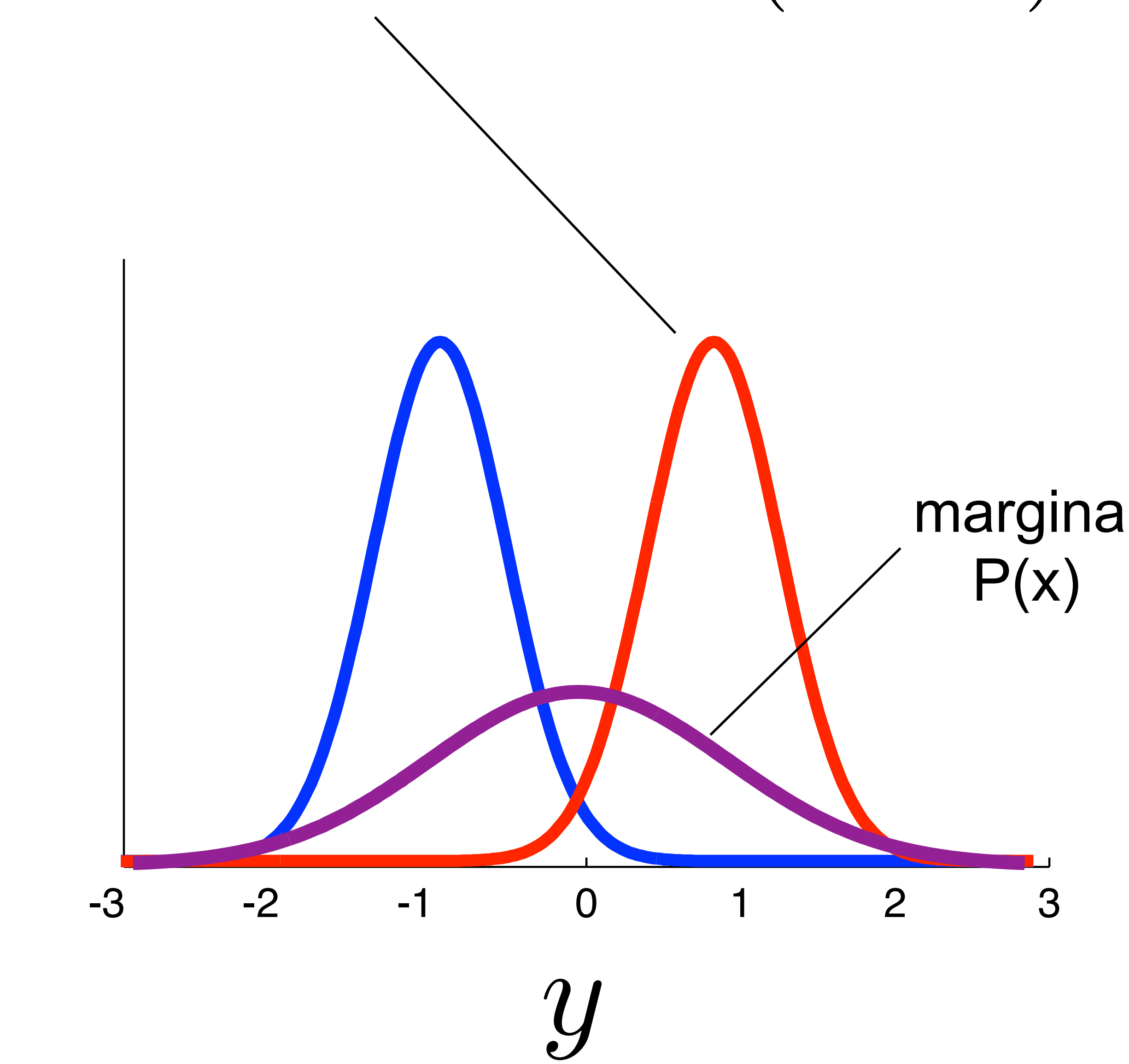


conditionalization ("slicing")



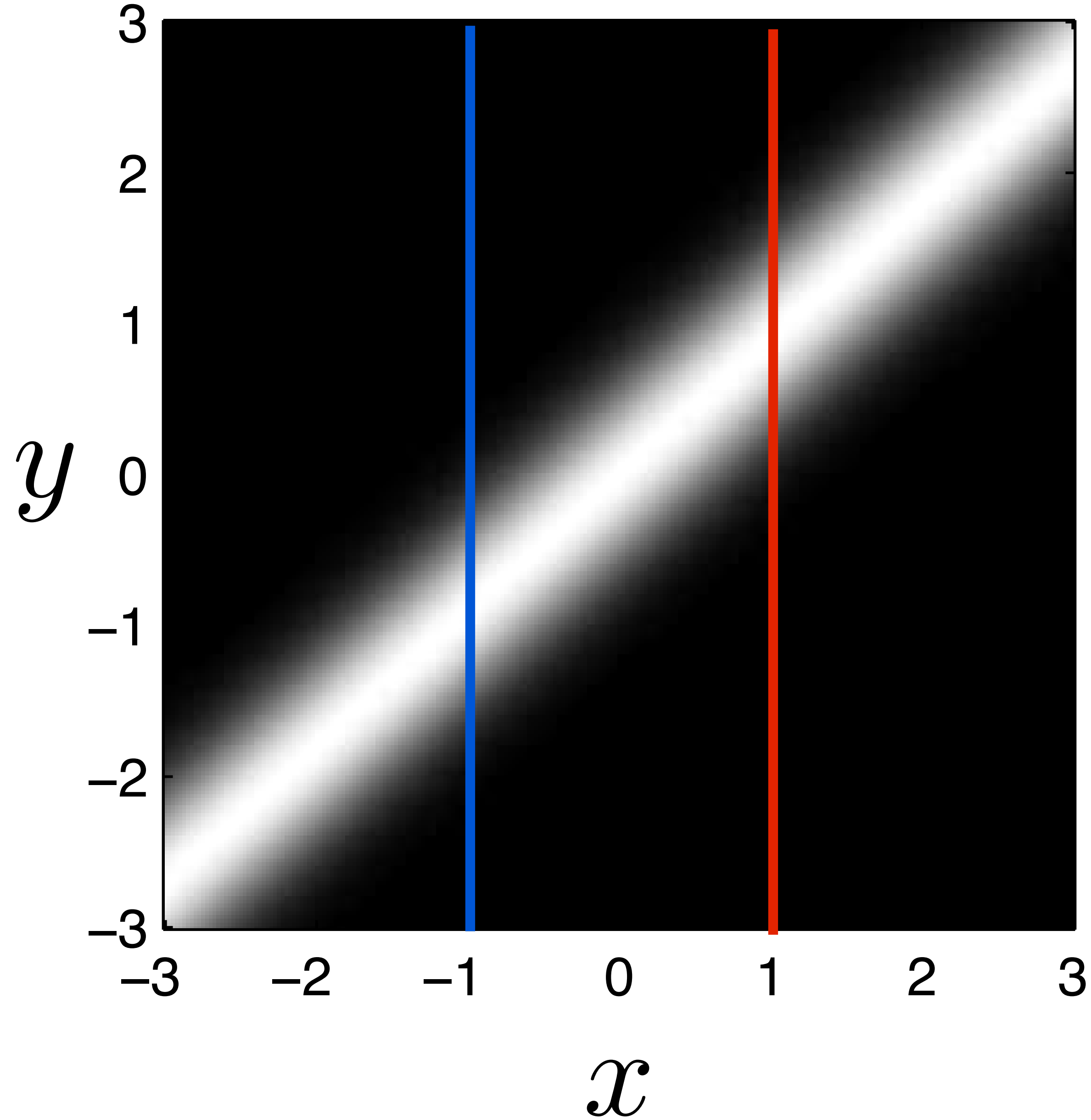
conditional

$$P(y|x=1) = \frac{P(y, x=1)}{P(x=1)}$$

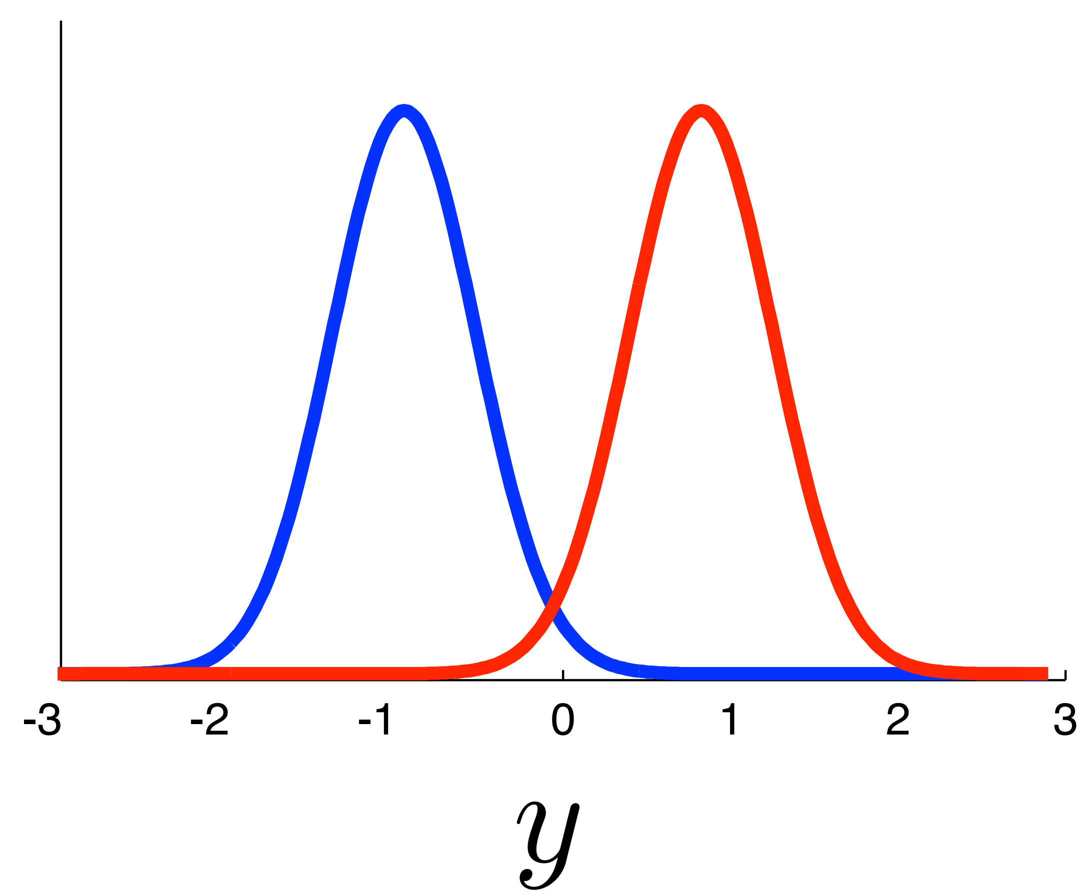


conditional densities

$P(y|x)$

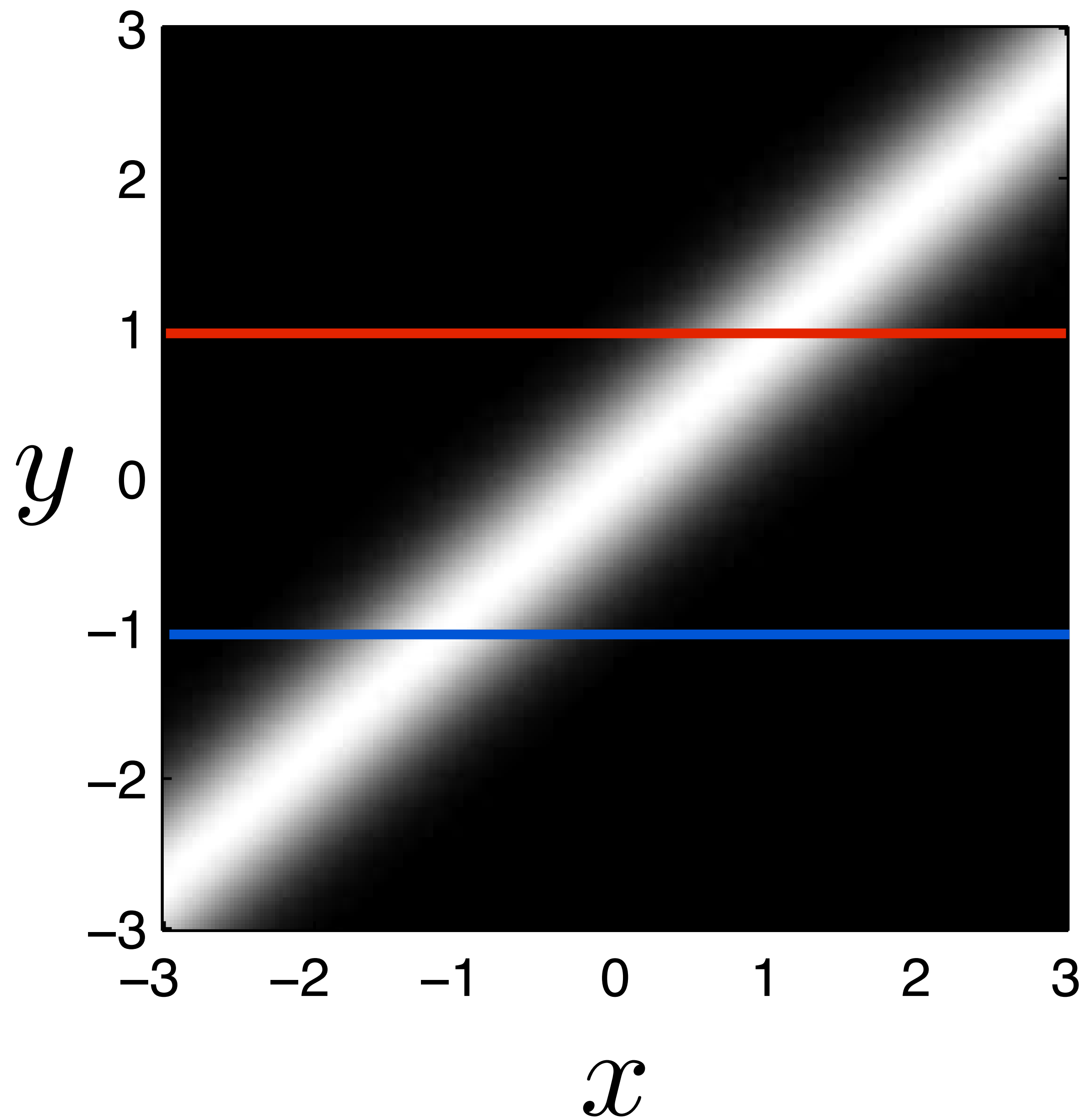


$$P(y|x) = \frac{P(x,y)}{P(x)}$$

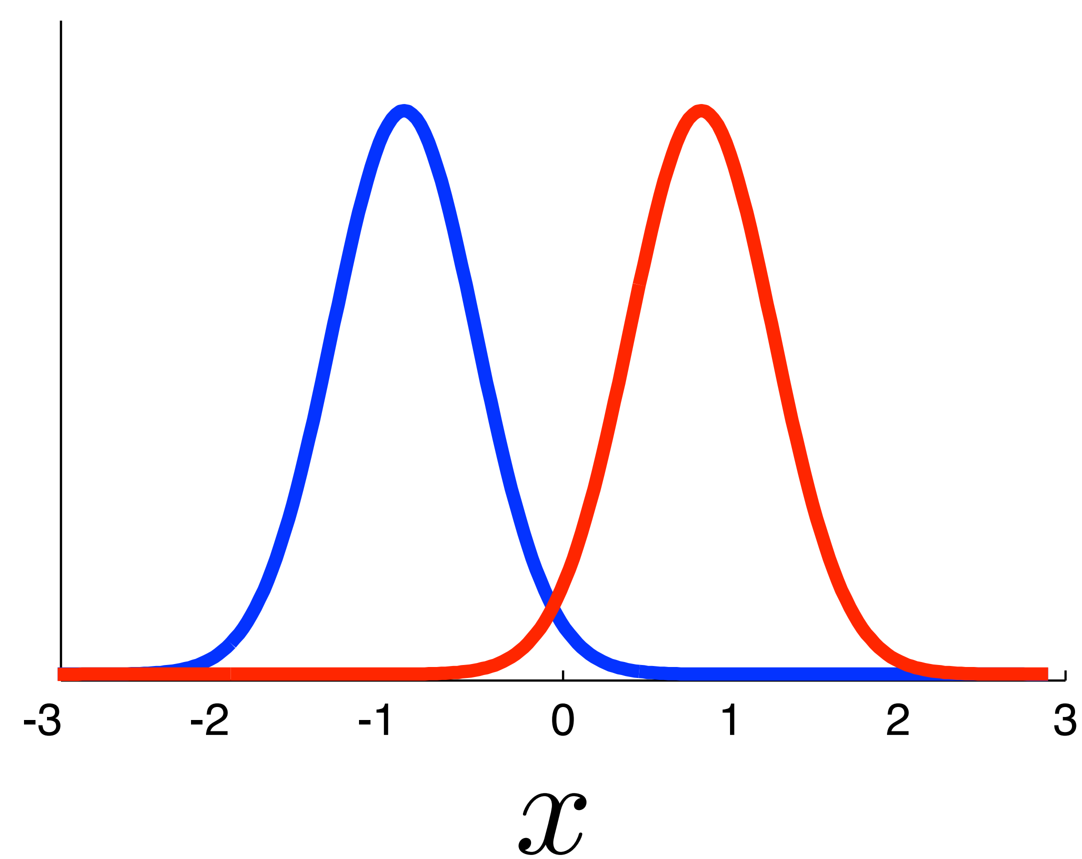


conditional densities

$$P(x|y)$$



$$P(x|y) = \frac{P(x, y)}{P(y)}$$



Estimation

parameter

model

measured dataset

("sample")

$$\underset{\text{parameter}}{\theta} \xrightarrow{\;p(y|\theta)\;} \underset{\text{measured dataset}}{y}$$

An *estimator* is a function $f\ :\ y \longrightarrow \hat{\theta}$

- often we will write $\hat{\theta}(y)$ or just $\hat{\theta}$

Example 1: linear Poisson neuron

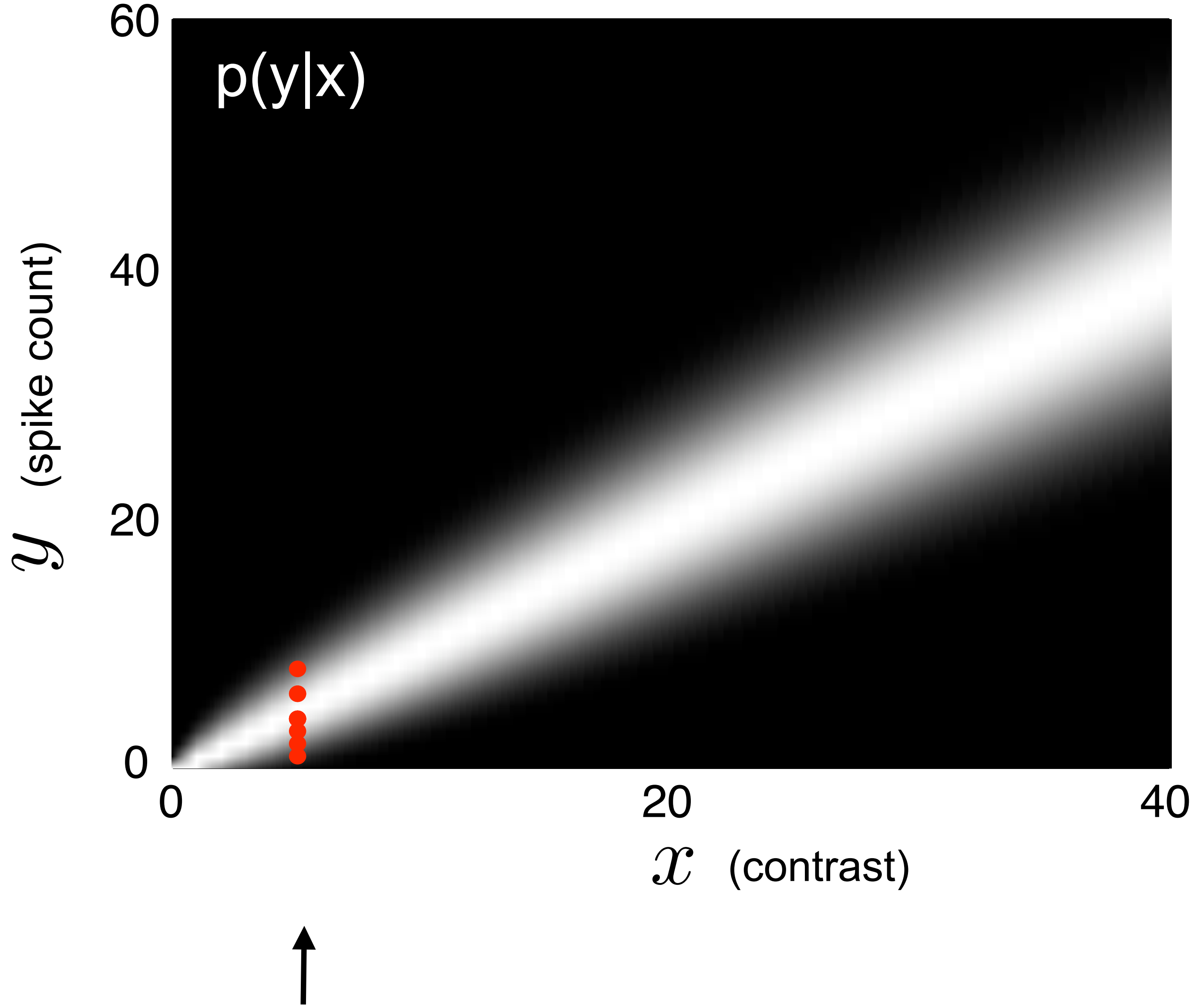
spike count $$y \sim Poiss(\lambda)$$

spike rate $$\lambda = \theta x$$

parameter stimulus

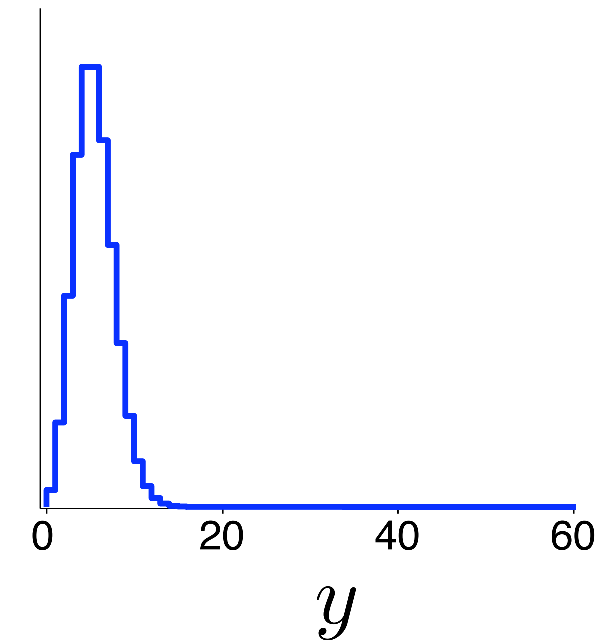
encoding model:
$$\begin{aligned} P(y|x,\theta) &= \frac{1}{y!}\lambda^y e^{-\lambda} \\ &= \frac{1}{y!}(\theta x)^y e^{-(\theta x)} \end{aligned}$$

$$\text{mean}(y) = \theta x$$
$$\text{var}(y) = \theta x$$

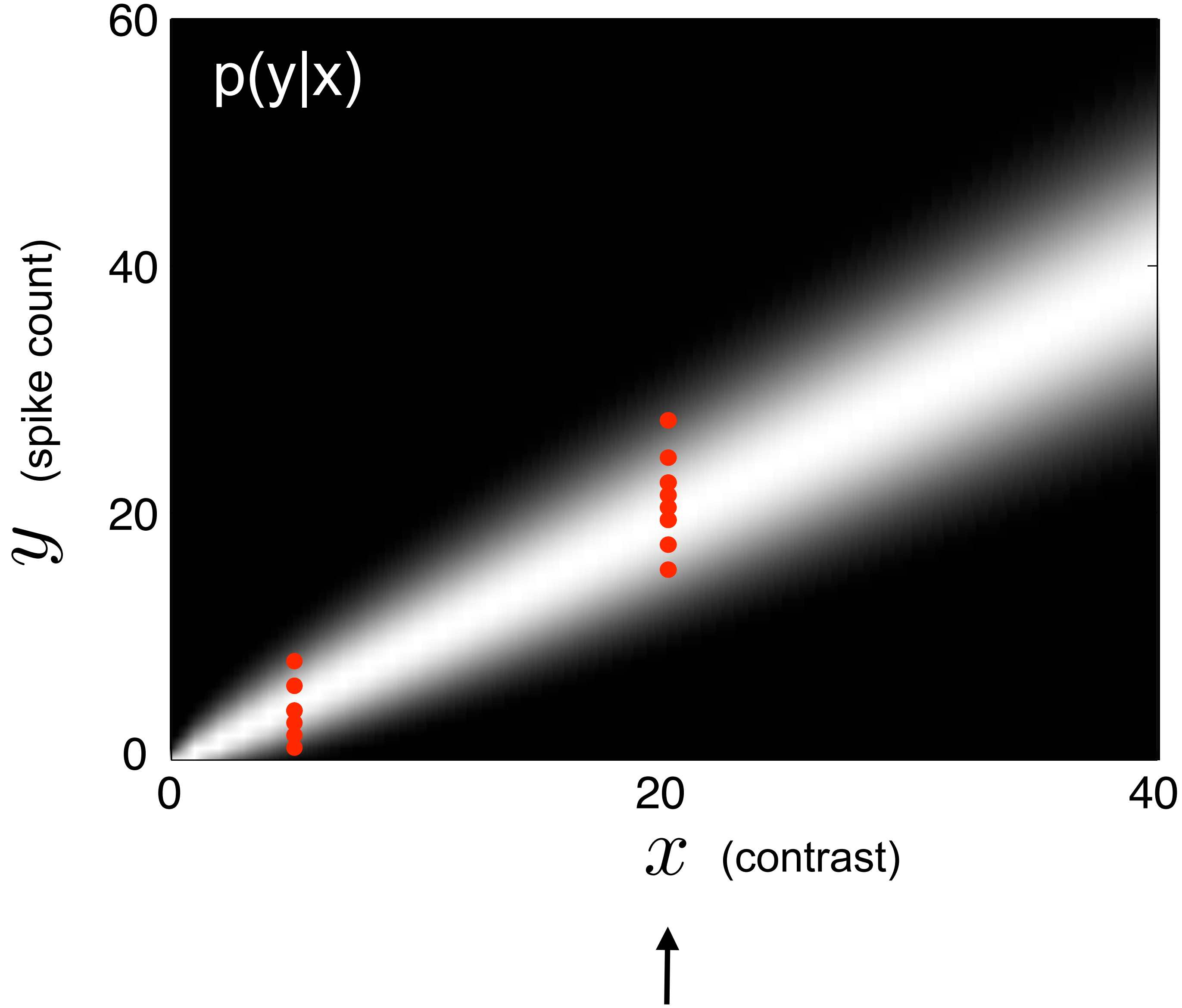


conditional distribution

$p(y|x=5)$

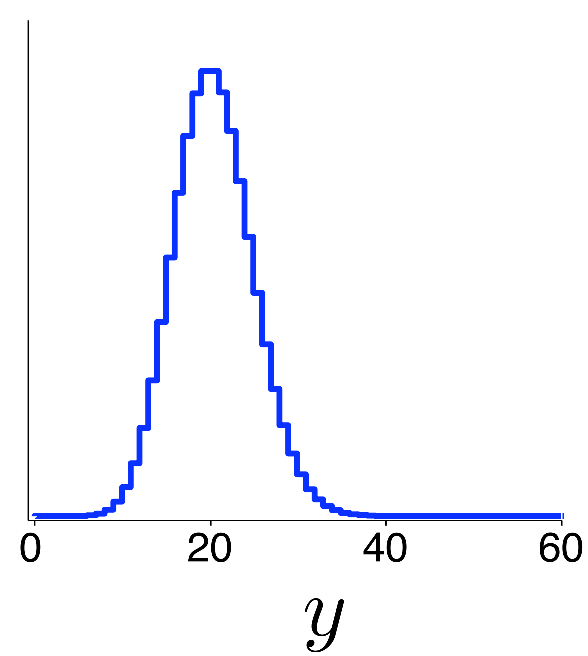


$$\mathrm{mean}(y) = \theta x$$
$$\mathrm{var}(y) = \theta x$$

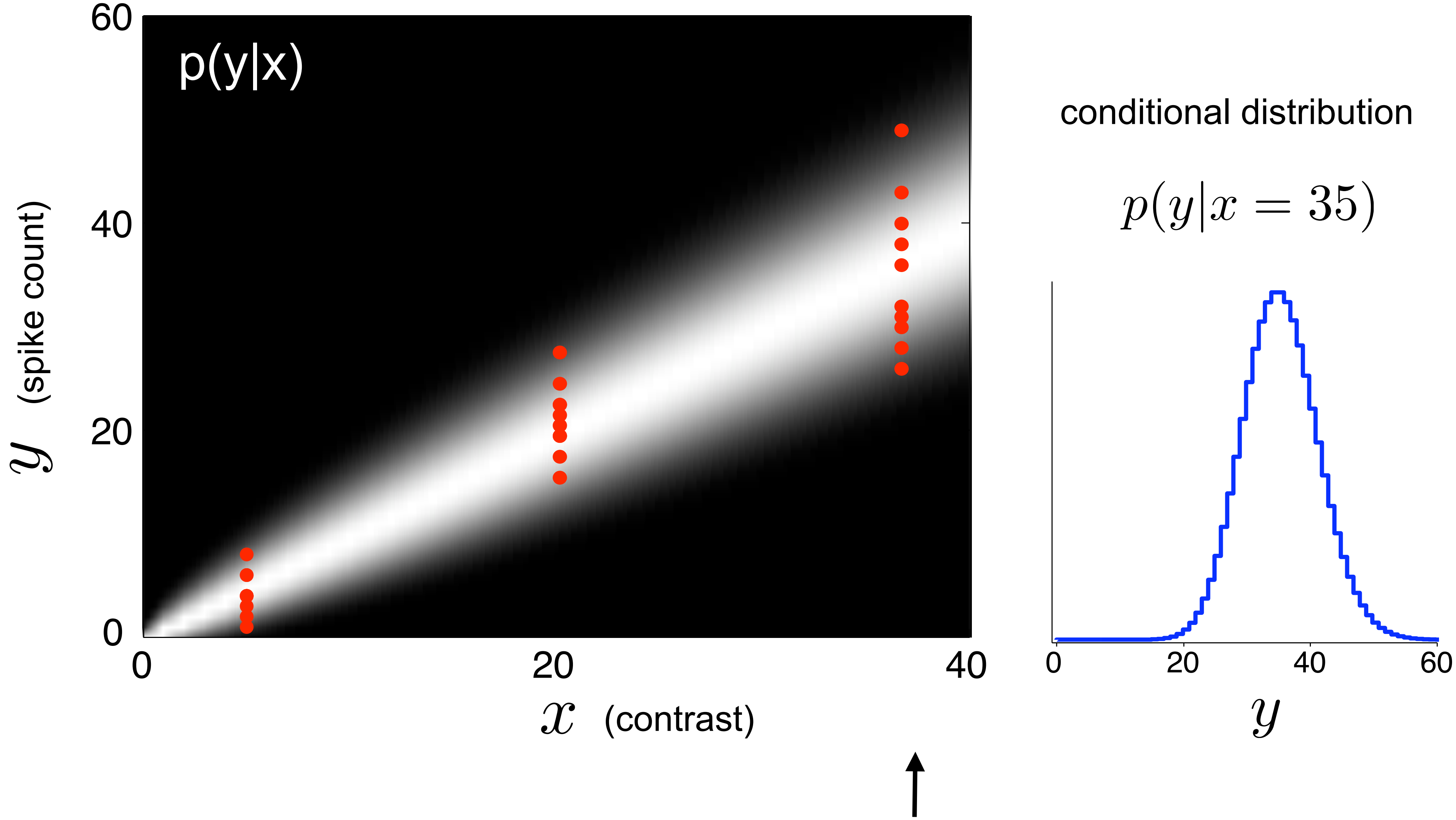


conditional distribution

$$p(y|x = 20)$$

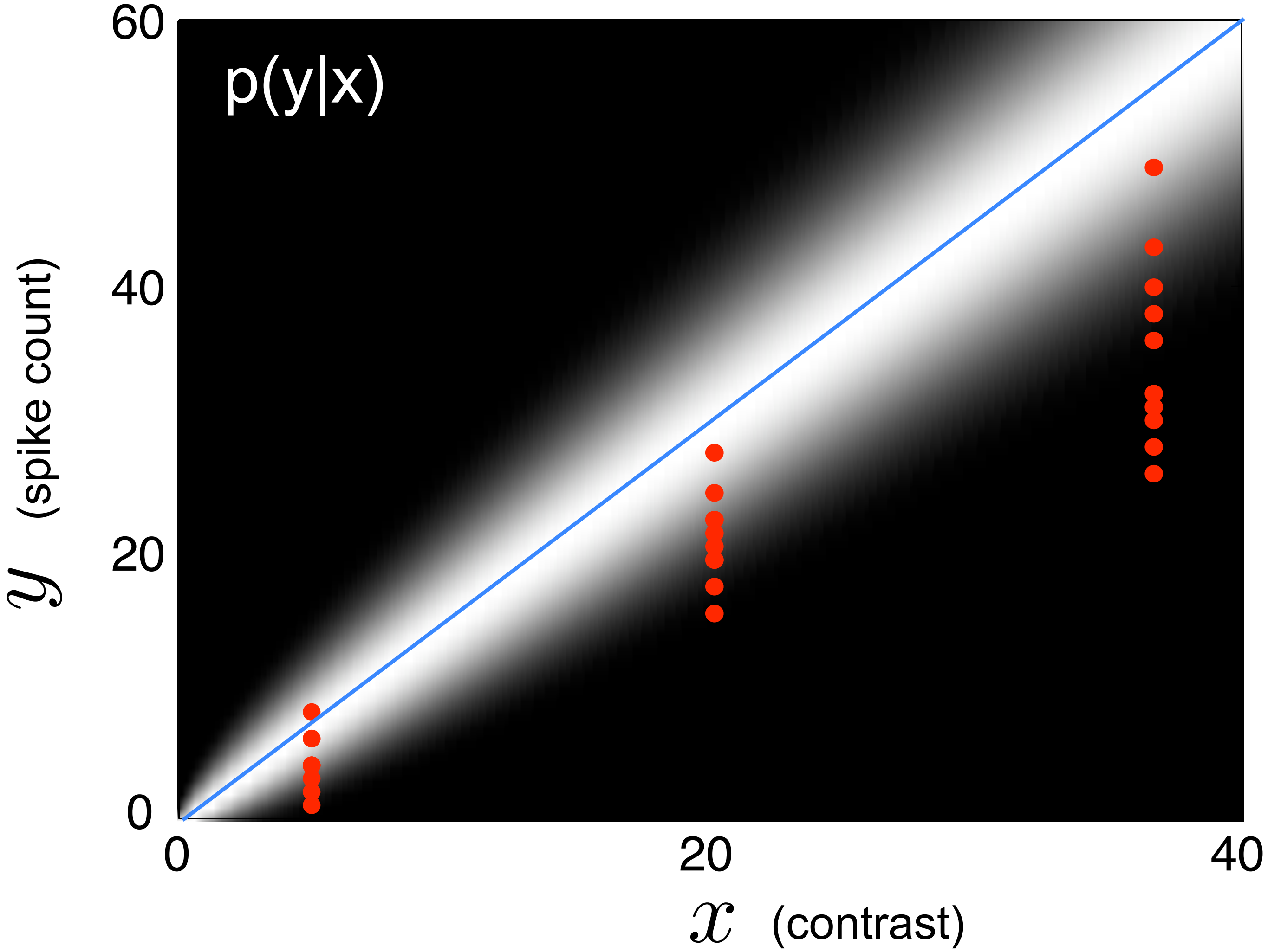


$$\mathrm{mean}(y) = \theta x$$
$$\mathrm{var}(y) = \theta x$$



Maximum Likelihood Estimation:

- given observed data (Y, X), find θ that maximizes $P(Y|X, \theta)$

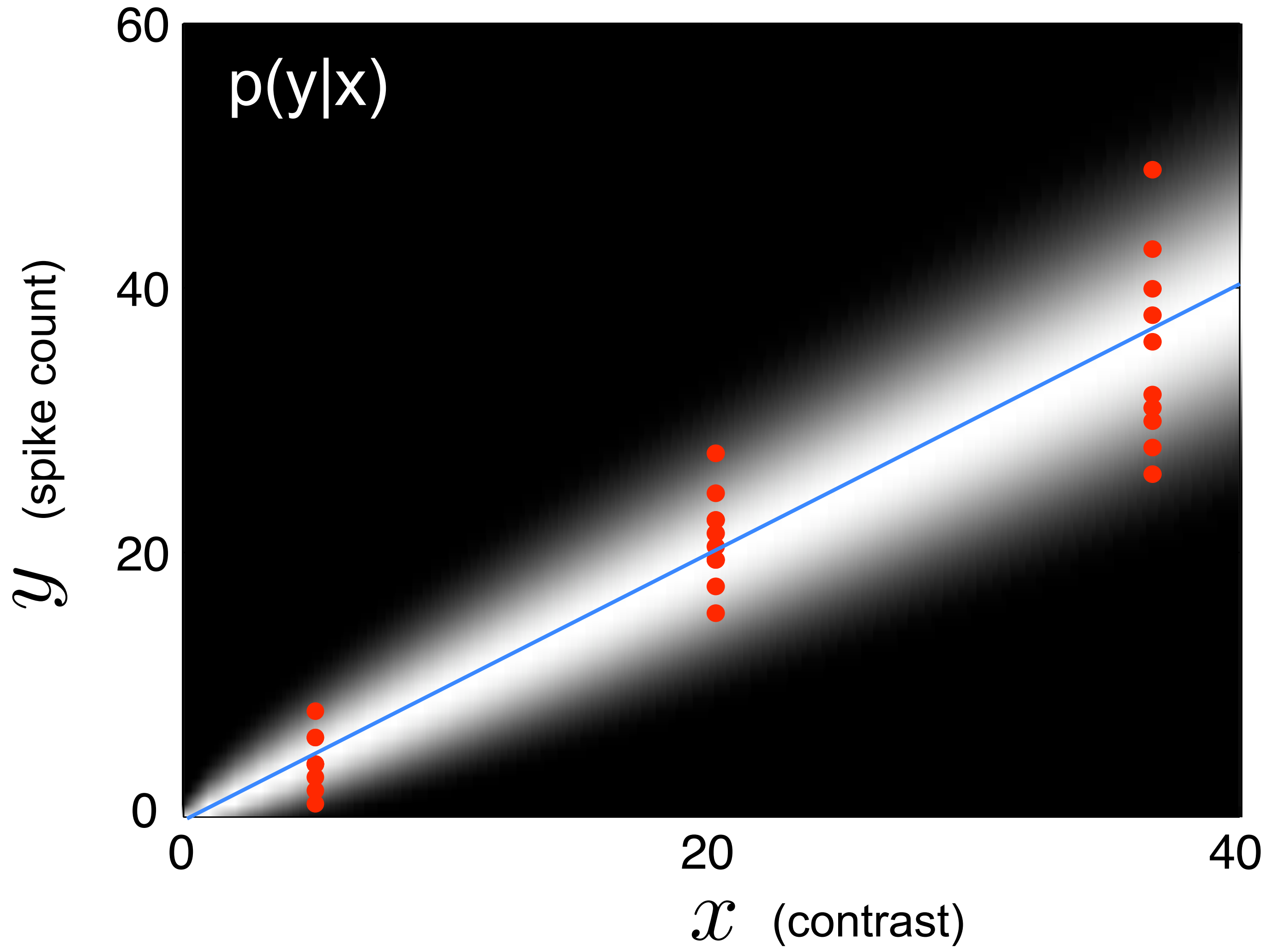


$y \sim \mathrm{Poiss}(\theta x)$

$\theta = 1.5$

Maximum Likelihood Estimation:

- given observed data (Y, X), find θ that maximizes $P(Y|X, \theta)$

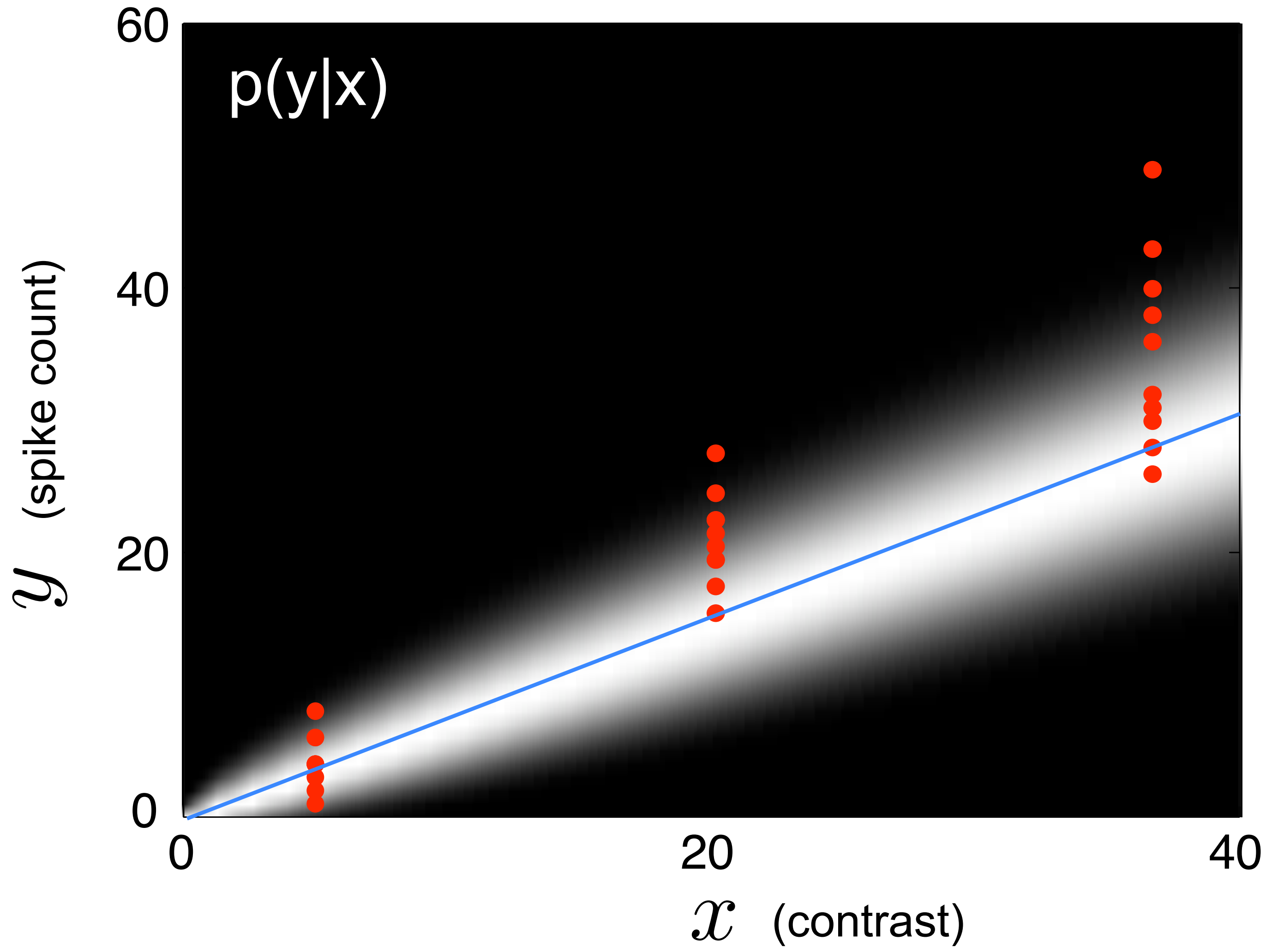


$y \sim \mathrm{Poiss}(\theta x)$

$\theta = 1$

Maximum Likelihood Estimation:

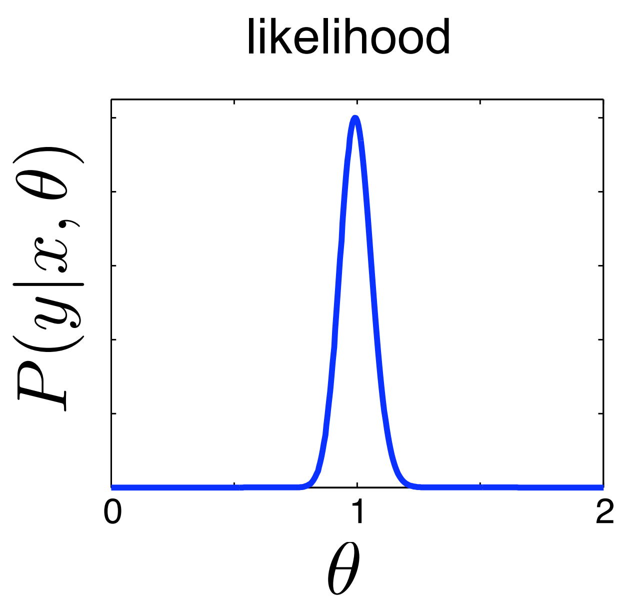
- given observed data (Y, X), find θ that maximizes $P(Y|X, \theta)$



$y \sim \text{Poiss}(\theta x)$

$\theta = 0.5$

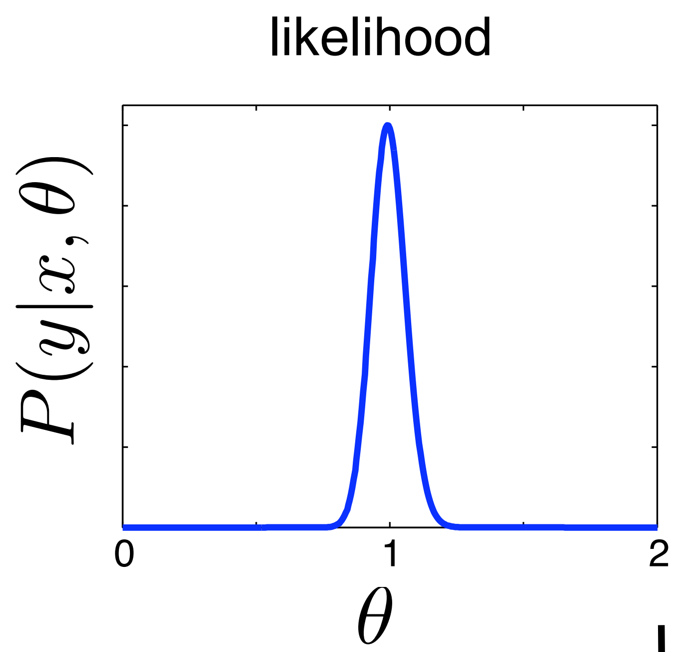
Likelihood function: $P(Y|X,\theta)$ as a function of θ



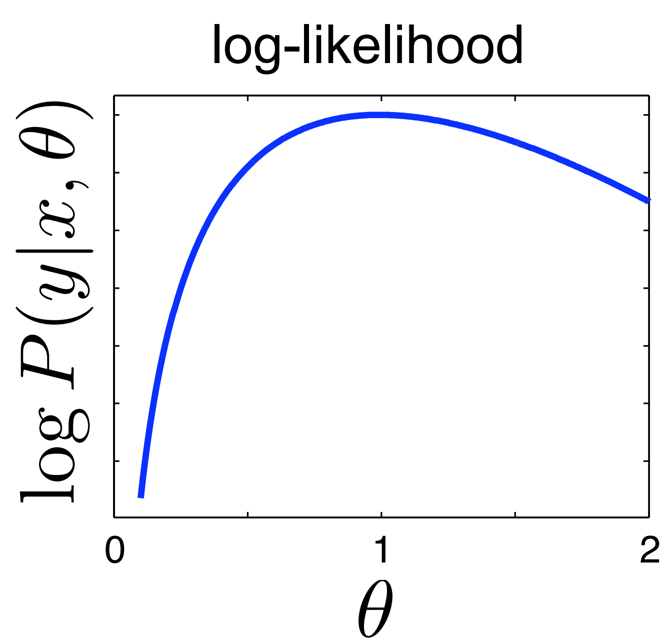
Because data are independent:

$$P(Y|X,\theta) = \prod_i P(y_i|x_i,\theta)$$
$$= \prod \frac{1}{y_i!}(\theta x_i)^{y_i} e^{-(\theta x_i)}$$

Likelihood function: $P(Y|X,\theta)$ as a function of θ



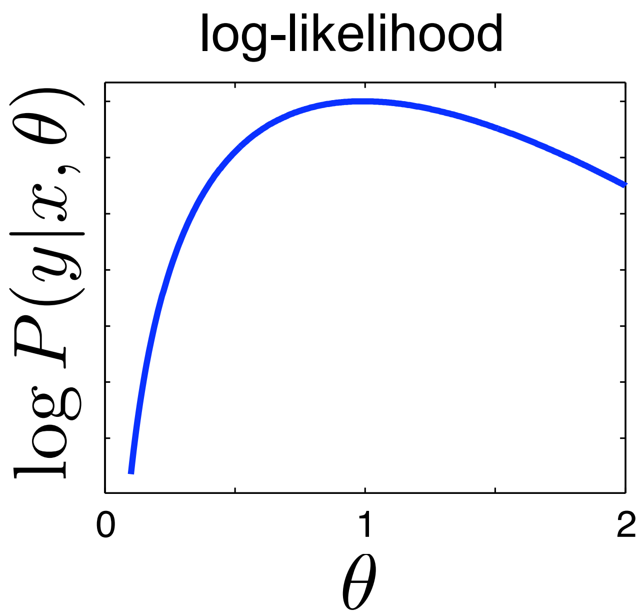
log



Because data are independent:

$$P(Y|X,\theta) = \prod_i P(y_i|x_i,\theta)$$
$$= \prod \frac{1}{y_i!}(\theta x_i)^{y_i} e^{-(\theta x_i)}$$

$$\log P(Y|X,\theta) = \sum_i \log P(y_i|x_i,\theta)$$
$$= \sum y_i \log\theta - \theta x_i + c$$



$$\log P(Y|X,\theta) = \sum_i \log P(y_i|x_i,\theta)$$
$$= \sum y_i \log\theta - \theta x_i + c$$
$$= \log\theta(\sum y_i) - \theta(\sum x_i)$$

- Closed-form solution (exists for “exponential family” models)

$$\frac{d}{d\theta}\log P(Y|X,\theta) = \frac{1}{\theta}\sum y_i - \sum x_i = 0$$
$$\implies \quad \hat{\theta}_{ML} = \frac{\sum y_i}{\sum x_i}$$

Example 2: linear Gaussian neuron

spike count $y \sim \mathcal{N}(\mu, \sigma^2)$

spike rate $\mu = \theta x$

parameter stimulus

encoding model: $$P(y|x,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\theta x)^2}{2\sigma^2}}$$

Log-Likelihood $$\log P(Y|X,\theta) = -\sum \frac{(y_i - \theta x_i)^2}{2\sigma^2} + c$$

Differentiate and set to zero:

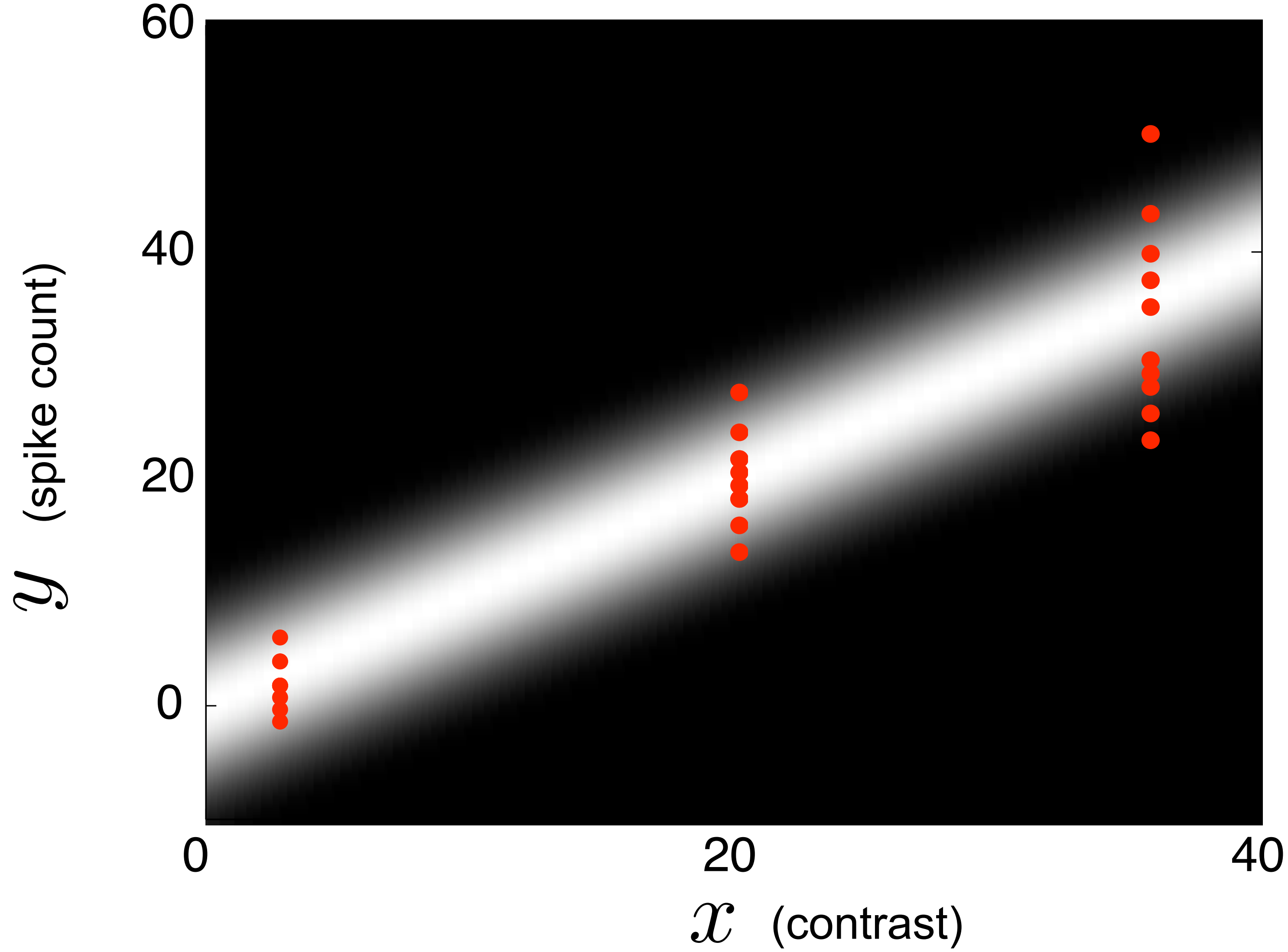
$$\frac{d}{d\theta} \log P(Y|X,\theta) = -\sum \frac{(y_i - \theta x_i)x_i}{\sigma^2} = 0$$

Maximum-Likelihood Estimator: $$\hat{\theta}_{ML} = \frac{\sum y_i x_i}{\sum x_i^2}$$

("Least squares regression" solution)

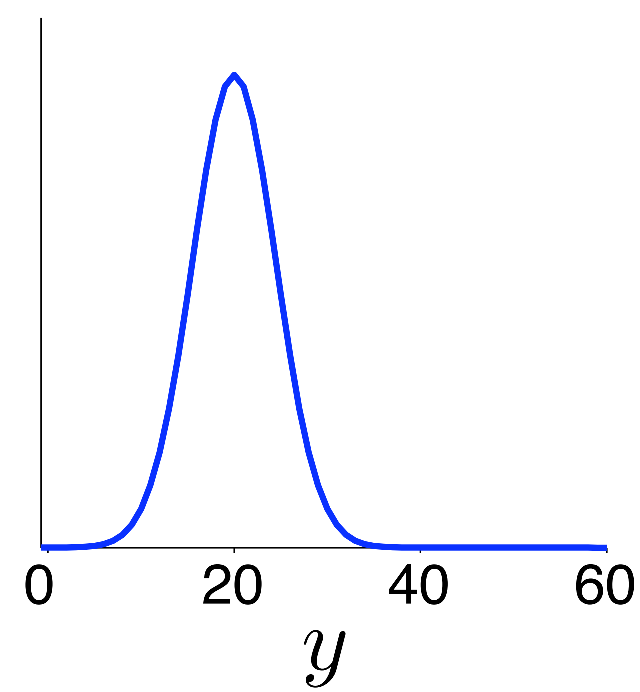
(Recall that for Poisson, $\hat{\theta}_{ML} = \frac{\sum y_i}{\sum x_i}$)

$$\mathrm{mean}(y) = \theta x$$
$$\mathrm{var}(y) = \sigma^2$$



encoding distribution

$$p(y|x = 20)$$



All slices have same width

Log-Likelihood $$\log P(Y|X,\theta) = -\sum \frac{(y_i - \theta x_i)^2}{2\sigma^2} + c$$

Differentiate and set to zero:

$$\frac{d}{d\theta} \log P(Y|X,\theta) = -\sum \frac{(y_i - \theta x_i)x_i}{\sigma^2} = 0$$

Maximum-Likelihood Estimator: $$\hat{\theta}_{ML} = \frac{\sum y_i x_i}{\sum x_i^2}$$

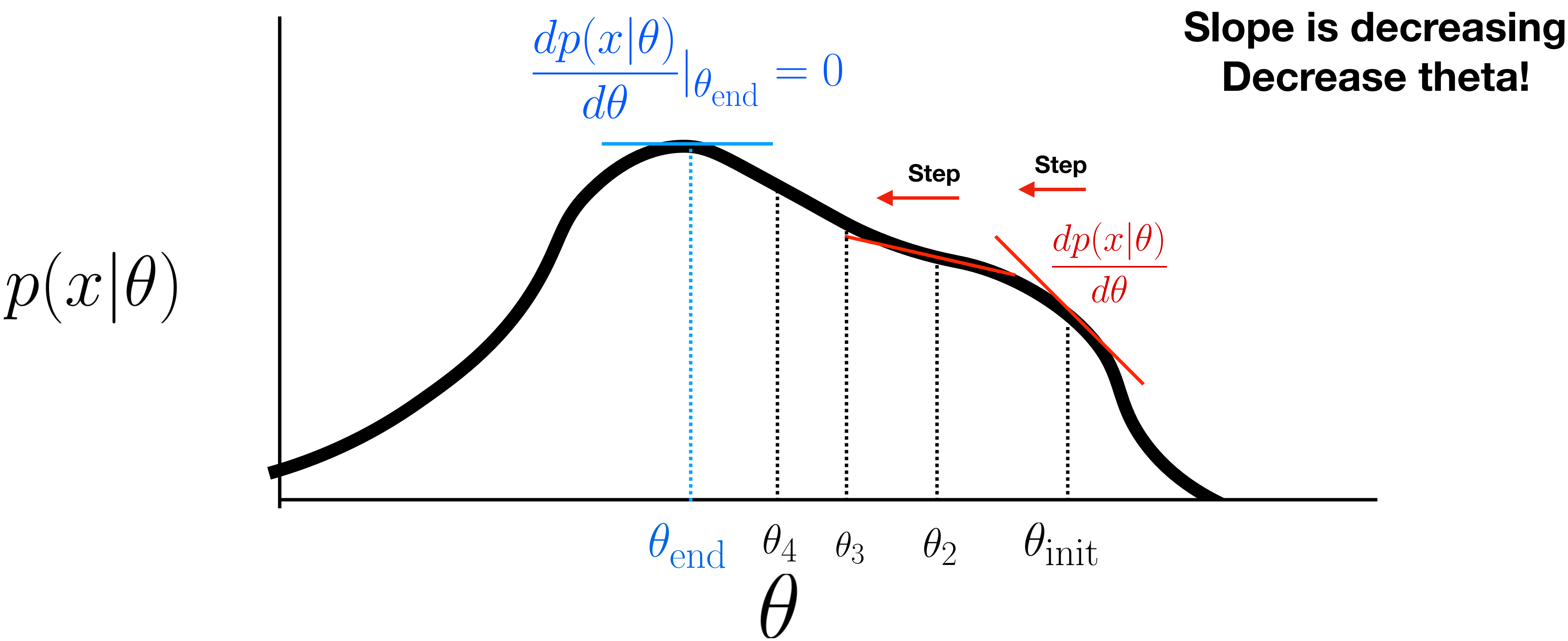
("Least squares regression" solution)

(Recall that for Poisson, $\hat{\theta}_{ML} = \dfrac{\sum y_i}{\sum x_i}$)

What if the model, $p(m|\theta)$ is very complicated??

- What if we can't solve $\frac{d}{d\theta}p(x|\theta) = 0$
- This is analogous to not knowing a closed form solution for the differential equations for dynamics
- We can use derivative based estimation methods!

Gradient Ascent (optimization)

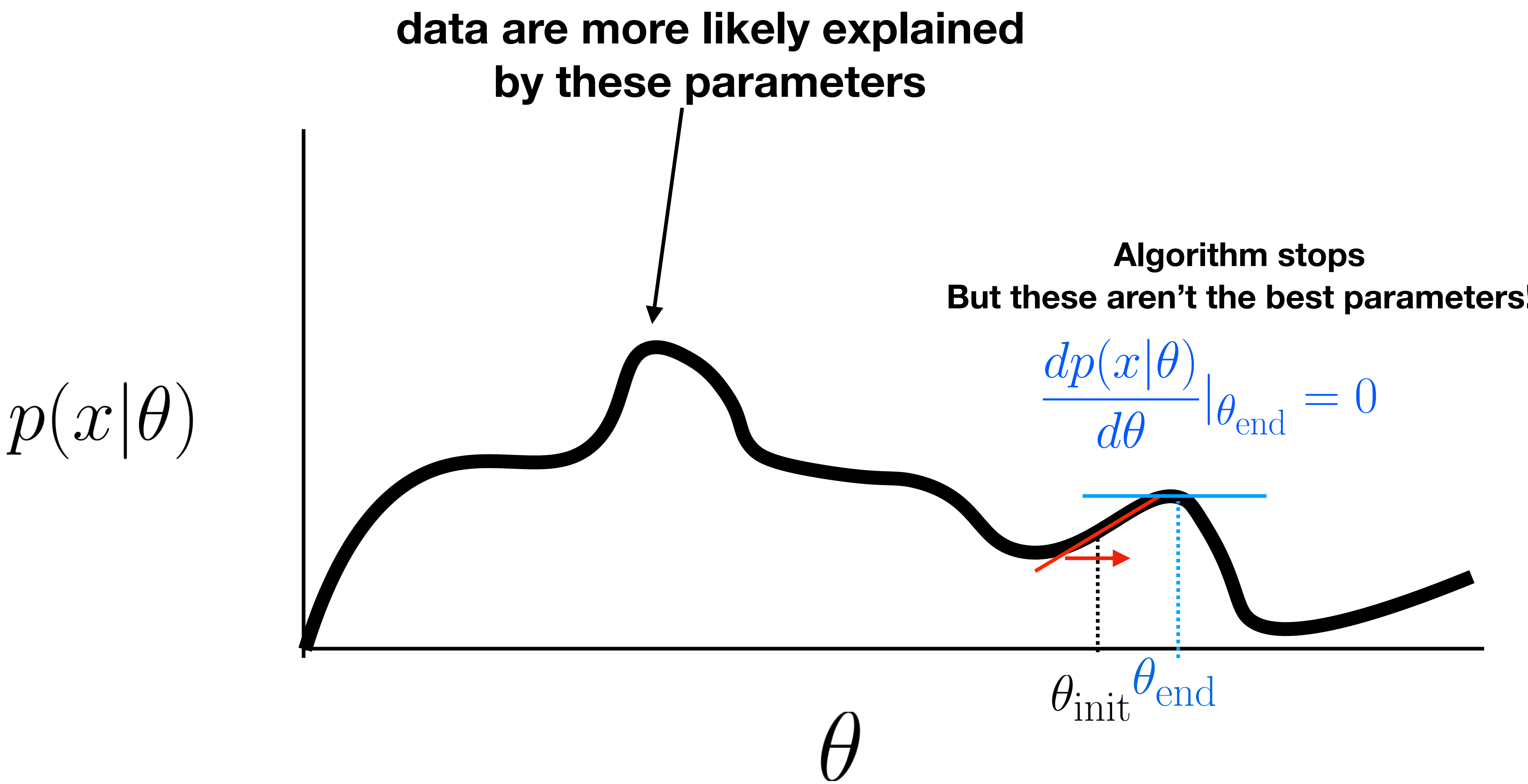


$$\theta_{\text{new}} = \theta_{\text{old}} + \alpha * \frac{dp(x|\theta)}{d\theta}$$

Learning Rate
α

Continue process until derivative is approximately 0, or until theta no longer changes

The Problem of local optima



- Optimization procedures often get 'stuck' in local optima
- This is a challenge for the statistics community, as we always want the value with the *highest* probability that best explains our data (global optimum)

- Sometimes, p(x|theta) is very complicated, and this optimization can be time-consuming or numerically challenging

- *Many* algorithms have been designed to solve this optimization problem for a variety of conditions
    - Newtons Method, Nelder-Mead, Stochastic Gradient Descent, AdaGrad, L-BFGS, ADAM, CG

- This process of optimization for probabilistic models is ubiquitous in statistics, and is the primary means by which we use models to fit data