NUS National University of Singapore | School of Computing

CS5340
Uncertainty Modeling in AI

Lecture 2:
Fitting Probability Models

Asst. Prof. Harold Soh

AY 23/24

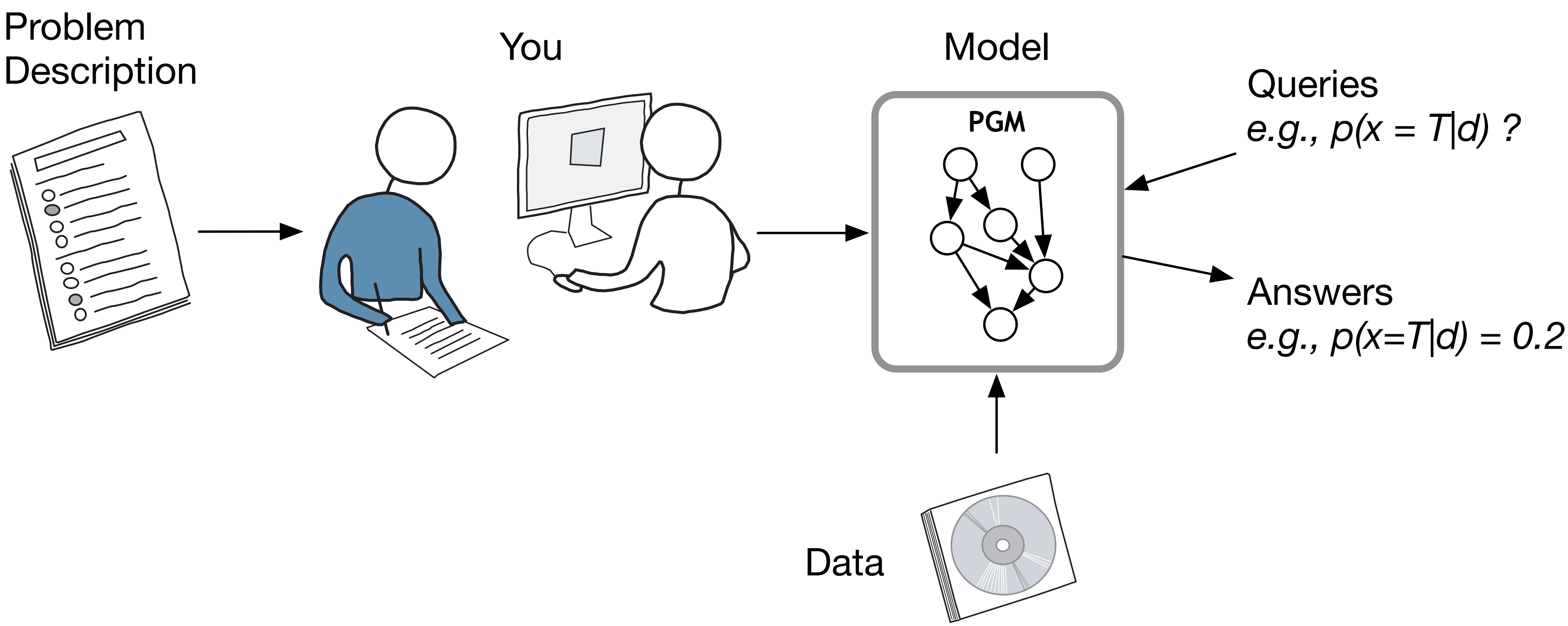
Semester 2

Course Schedule (Tentative)

Week	Date	Lecture Topic	Tutorial
1	16 Jan	Introduction to Uncertainty Modeling + Probability Basics	~~Introduction~~
2	23 Jan	Simple Probabilistic Models	Introduction and Probability Basics
3	30 Jan	Bayesian networks (Directed graphical models)	More Basic Probability
4	6 Feb	Markov random Fields (Undirected graphical models)	DGM modelling and d-separation
5	13 Feb	Variable elimination and belief propagation	MRF + Sum/Max Product
6	20 Feb	Factor graphs	**Quiz 1**
-	-	**RECESS WEEK**	
7	5 Mar	Mixture Models and Expectation Maximization (EM)	Linear Gaussian Models
8	12 Mar	Hidden Markov Models (HMM)	Probabilistic PCA
9	19 Mar	Monte-Carlo Inference (Sampling)	Linear Gaussian Dynamical Systems
10	26 Mar	Variational Inference	MCMC + Sequential VAE
11	2 Apr	Inference and Decision-Making	Diffusion Models
12	9 Apr	Gaussian Processes (Special Topic)	**Quiz 2**
13	16 Apr	Project Presentations	Closing Lecture

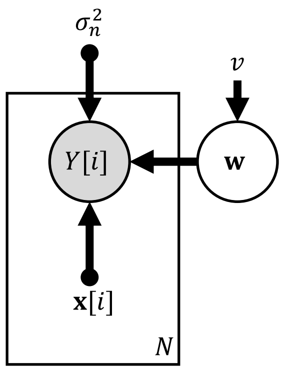
CS5340 in a nutshell

CS5340 is about how to "**represent**" and "**reason**" with uncertainty in a computer.

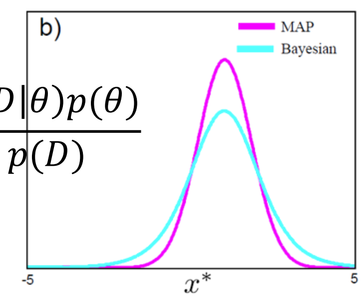


CS5340 in a nutshell

CS5340 is about how to "**represent**" and "**reason**" with uncertainty in a computer.



$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$



Representation: The *language* is probability and probabilistic graphical models (PGM).

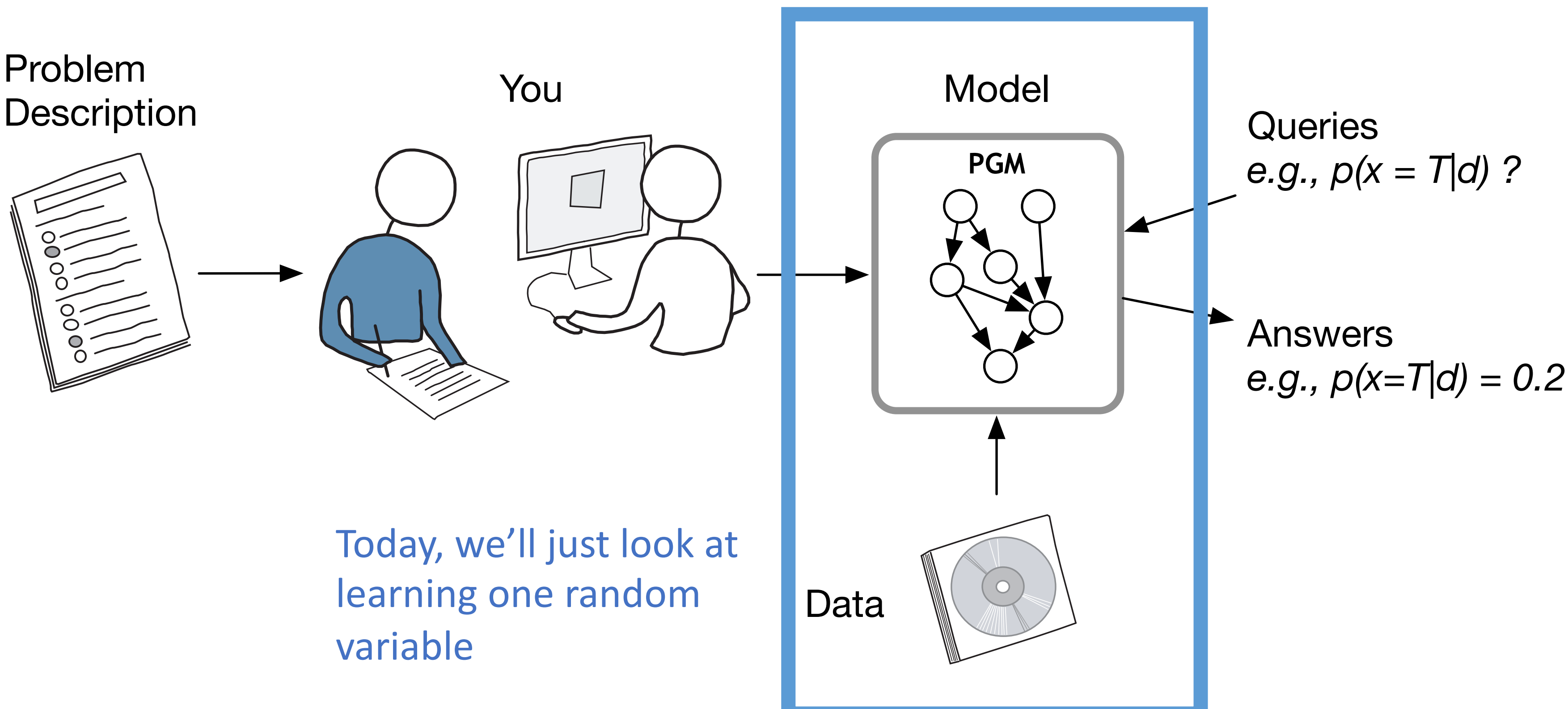
The language is used to model problems.

Reasoning: We use learning and inference algorithms to answer questions.

e.g., Belief-propagation/sum-product, MCMC, and variational Bayes

CS5340 in a nutshell

CS5340 is about how to "**represent**" and "**reason**" with uncertainty in a computer.



Today, we'll just look at learning one random variable

Summary: Sum and Product Rules

- Sum rule:

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

$$p(x) = \sum_{y} p(x,y)$$

- Product/Chain rule:

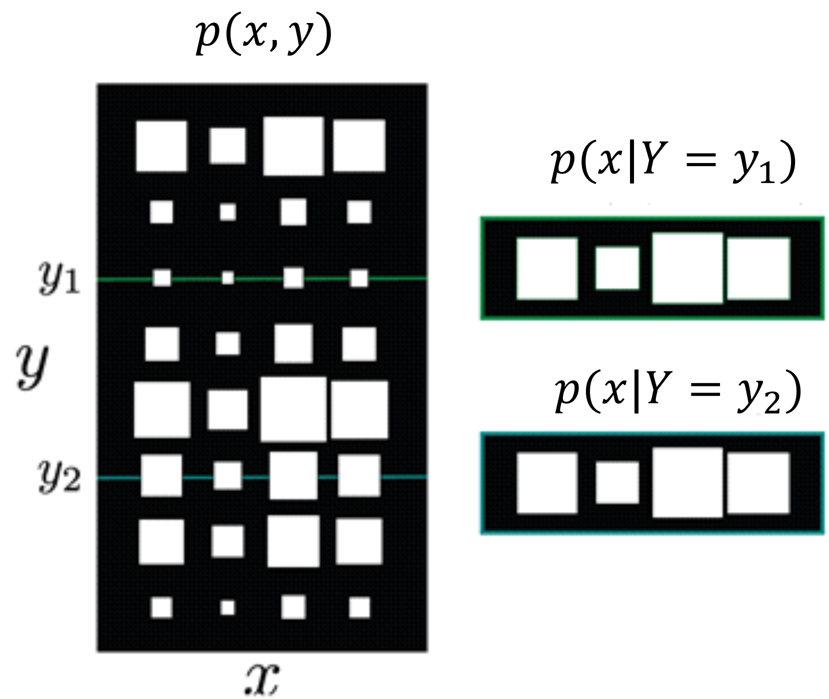
$$p(x,y) = p(x|y)p(y)$$

Probability: Independence

- The independence of X and Y means that every conditional distribution is the same.
- The value of Y tells us nothing about X and vice-versa.

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

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



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Probability: Bayes' Rule

- Recall:

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

Thomas Bayes
1701–1761

- Eliminating $p(x, y)$, we get:

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

- Rearranging:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x, y)dy} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

Image source: "Pattern Recognition and Machine Learning", Christopher Bishop

Probability: Expectation

- The expected or average value of some function $f[x]$ taking into account the distribution of X.

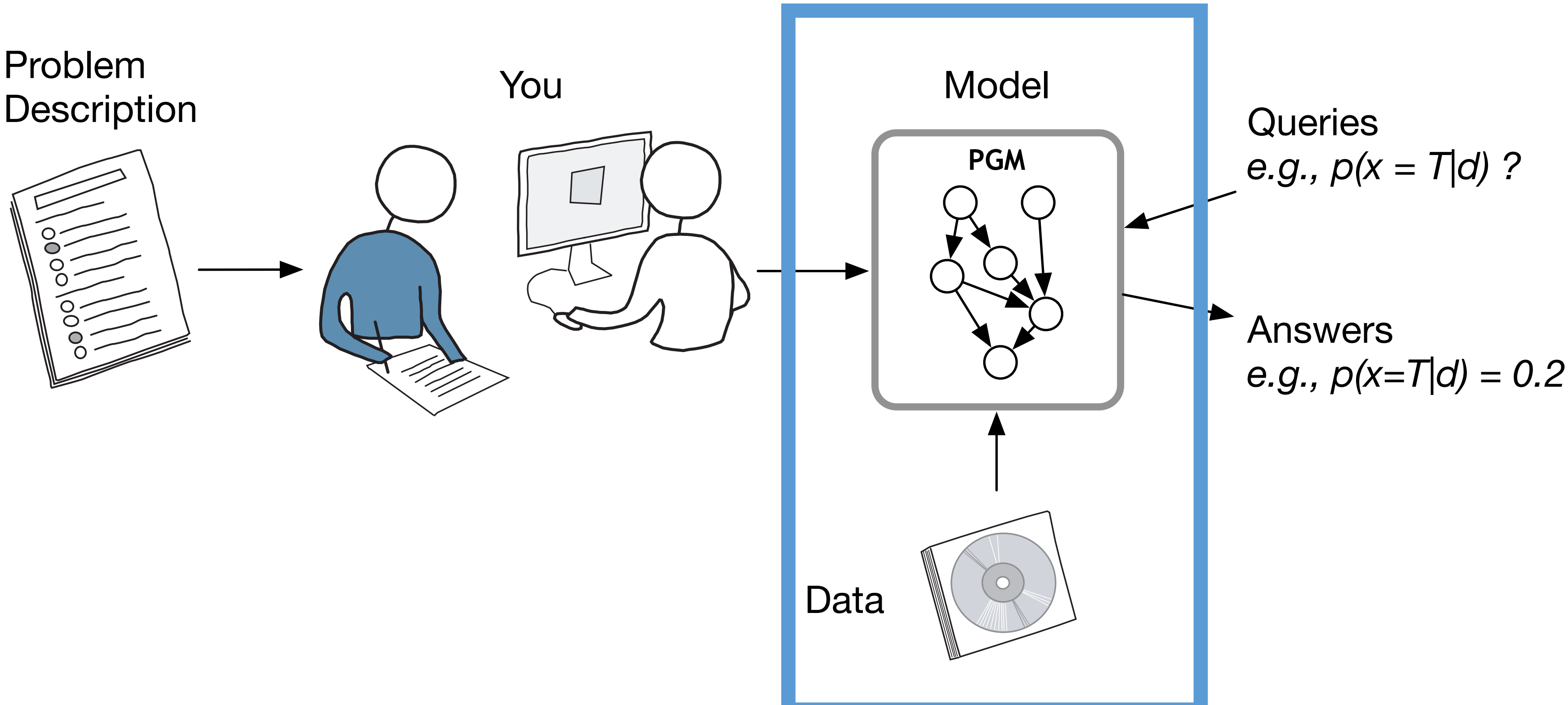
Definition:

$$E\left[f[x]\right] = \sum_x f[x]p(x)$$

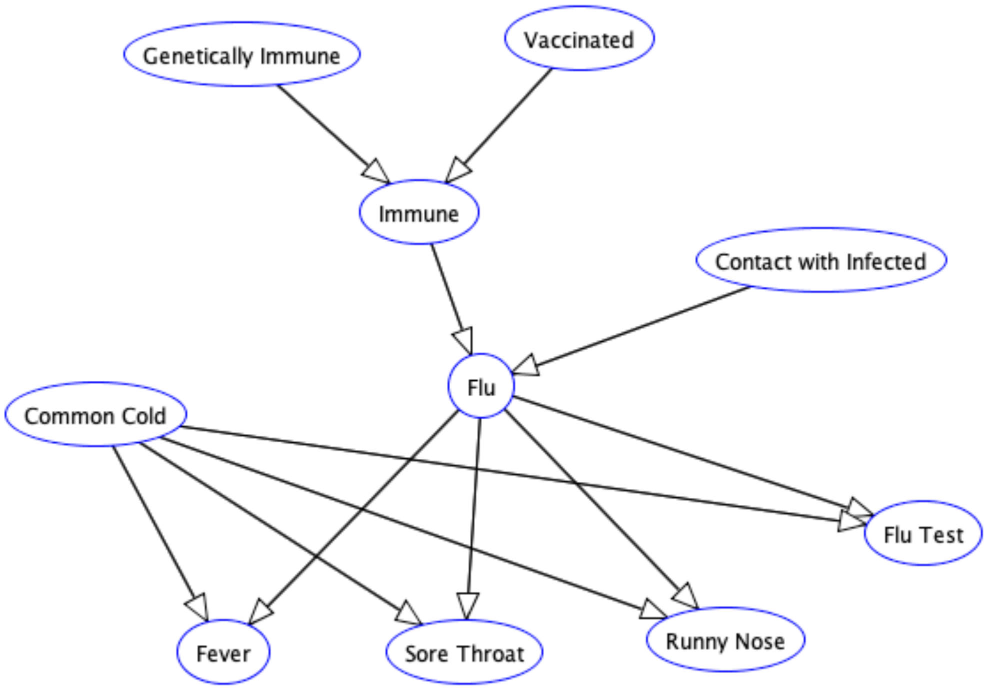
$$E\left[f[x]\right] = \int f[x]p(x)dx$$

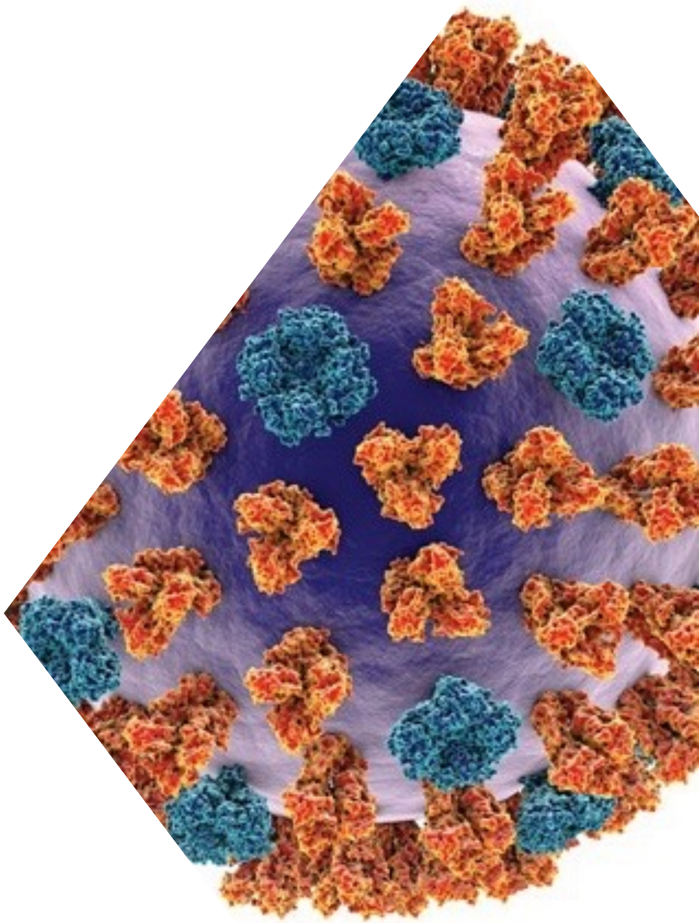
CS5340 in a nutshell

CS5340 is about how to "**represent**" and "**reason**" with uncertainty in a computer.

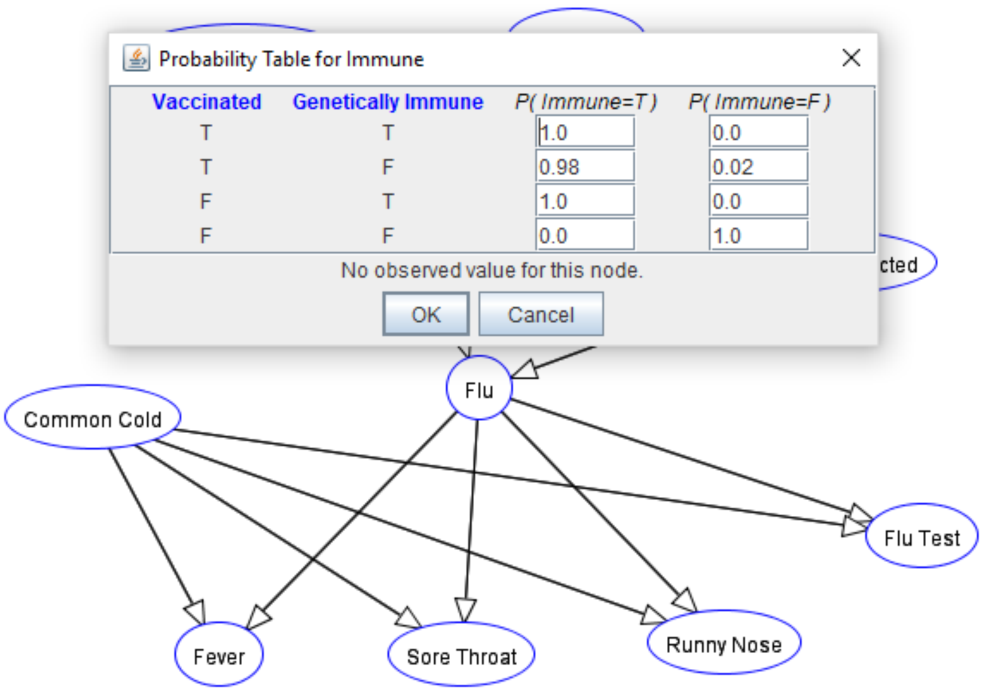


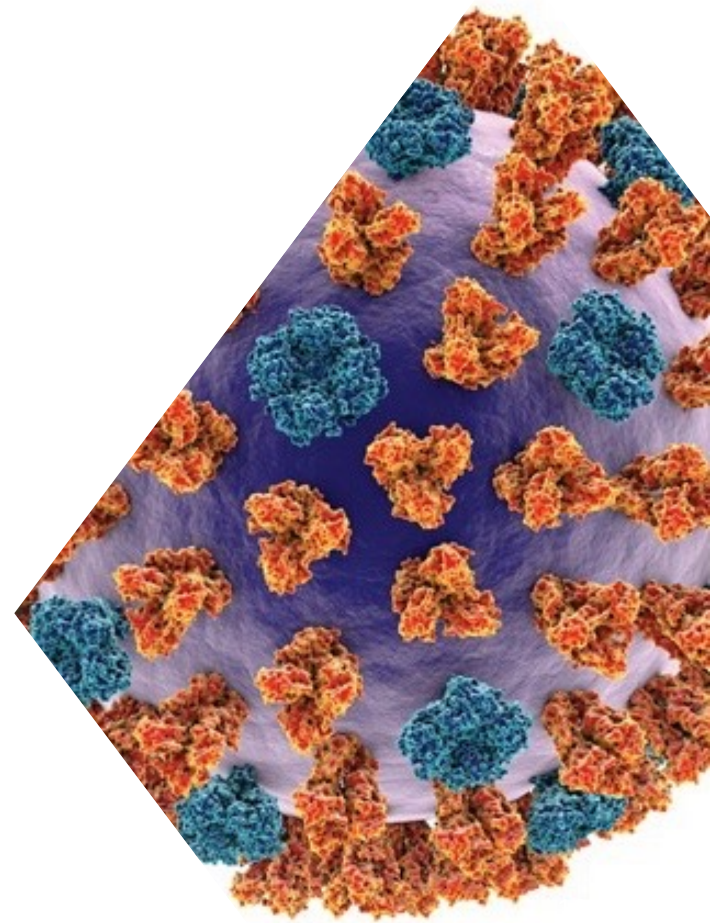
Generative (Causal) Modeling of Relationships between Variables





Generative (Causal) Modeling of Relationships between Variables





Fitting Probability Models

- Focus on parametric probability distributions $p(x|\theta)$.
- How to learn the unknown parameters θ from a set of given data, i.e. instances of the random variable, $\mathcal{D} = \{x[1], \ldots, x[N]\}$.
- And then use those parameters to make predictions.

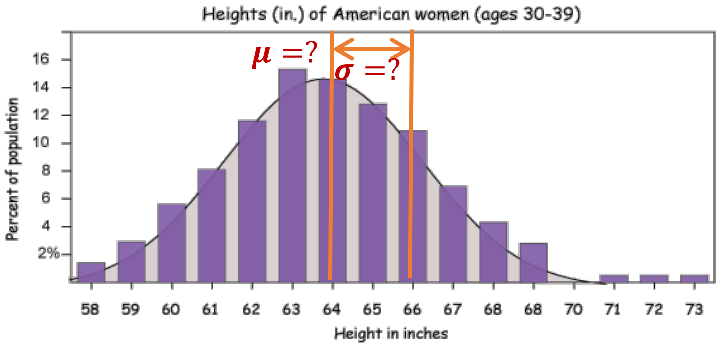


Image source: http://www.drcruzan.com/ProbStat_Distributions.html

Learning Outcomes

- Students should be able to:

1. Use the Maximum Likelihood, Maximum a Posteriori and Bayesian approaches to learn the unknown parameters of probability distributions of a single random variable from data.

2. Apply the assumption independent and identically distributed samples to simplify the parameter learning process.

3. Apply the learned parameters to make predictions.

4. Describe the exponential family and its properties

Learning Parameters

- Common approaches to learn the unknown parameters θ from a set of given data $\mathcal{D} = \{x[1], \dots, x[N]\}$:

Maximum Likelihood Estimation (MLE)

$$\operatorname*{argmax}_{\theta} p(D|\theta)$$

Maximum a Posteriori Estimation (MAP)

$$\operatorname*{argmax}_{\theta} p(\theta|D)$$

Variational Bayes

Monte-Carlo Sampling

"Full" Bayesian Inference

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Computational Cost

(In general and not to scale)

Acknowledgements

- A lot of slides and content of this lecture are adopted from:

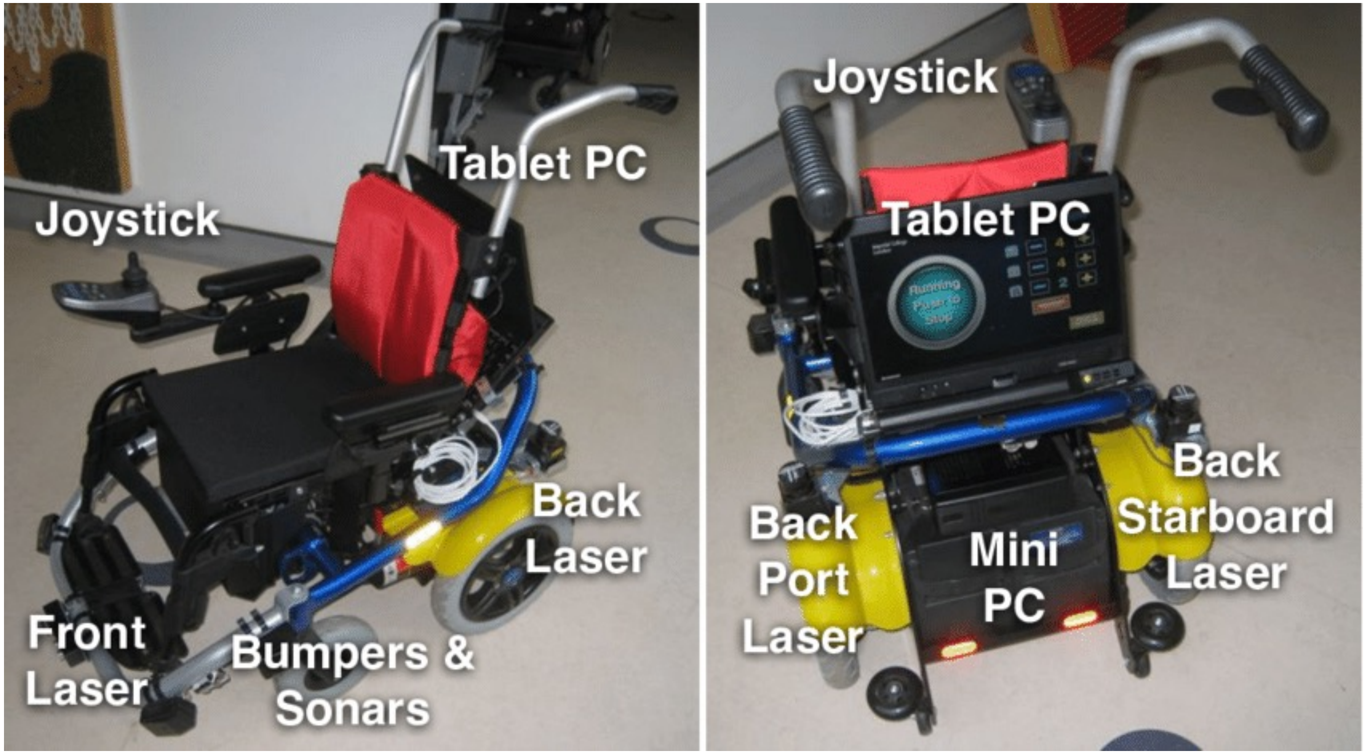
1. "Pattern Recognition and Machine Learning", Christopher Bishop.
2. "Computer Vision: Models, Learning, and Inference", Simon Prince.
3. Lee Gim Hee's CS5340 slides.

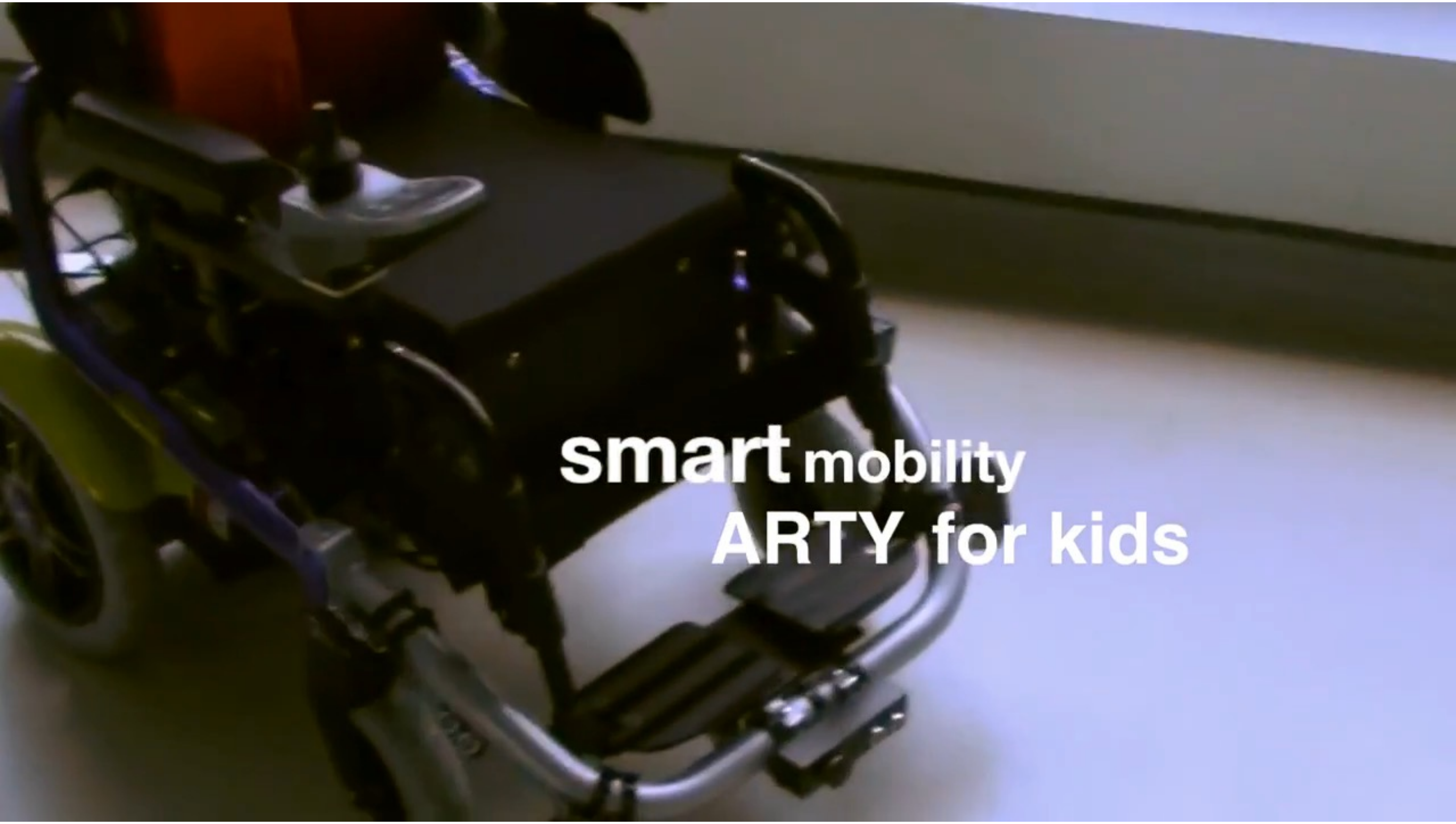
Learning via MLE

Maximum Likelihood Estimation (MLE)

Building a *Smart* Wheelchair for Kids with Disabilities

- https://youtu.be/XbyqU88jmb0




smart mobility
ARTY for kids

Problem: Sensor Uncertainty

- You have a ultrasonic ranger for your robot.
- Like other sensors, there is some error.
- How can you model and estimate the uncertainty of your range readings?
- *Later*: can we predict the range given a noisy reading?

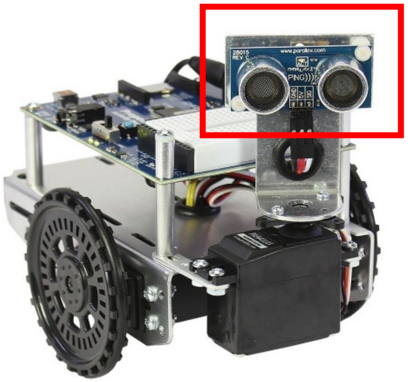
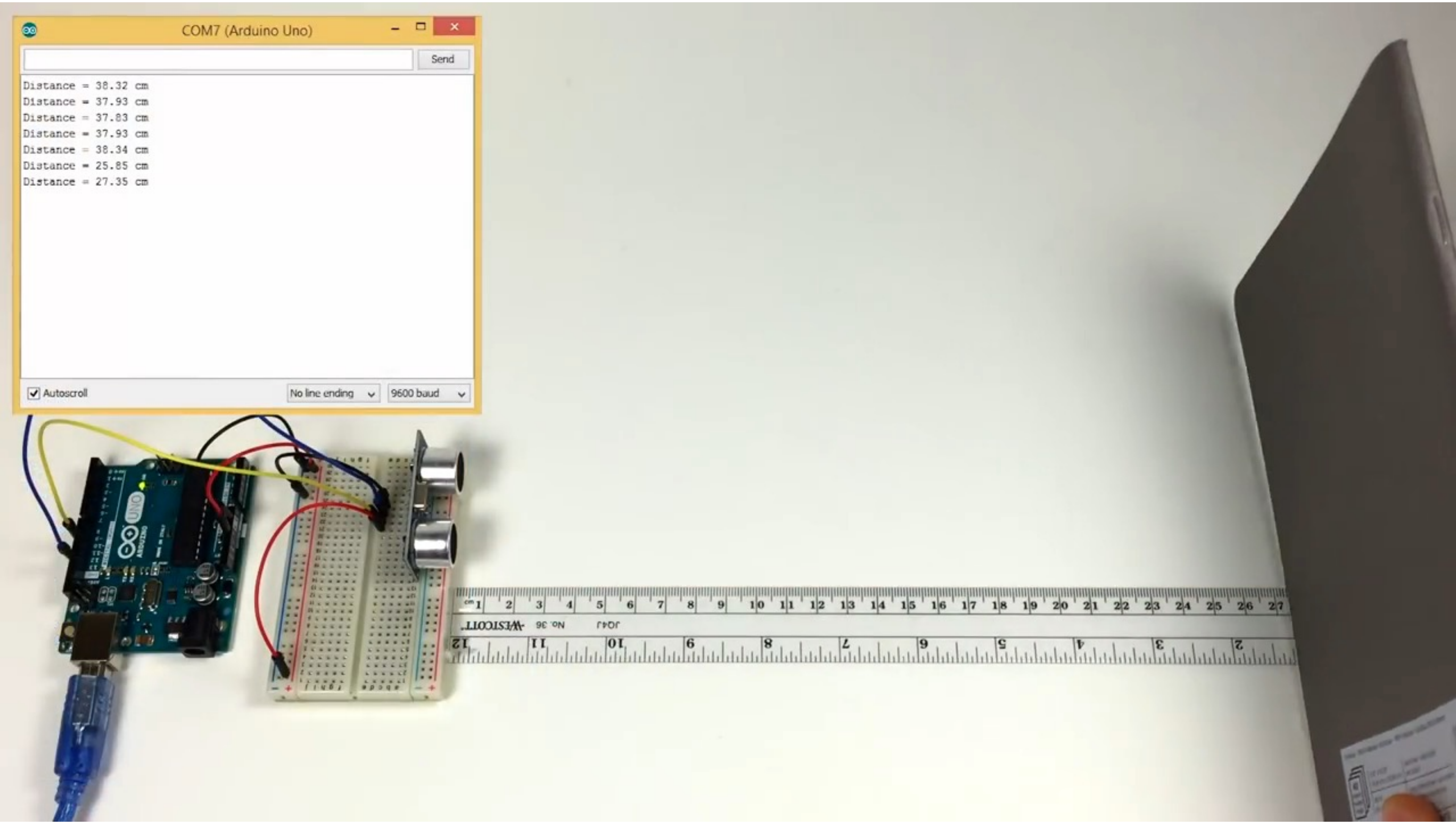


Image credit: https://www.parallax.com/product/910-28015a

Video: https://youtu.be/Ea4CGAw6b_M?t=683


COM7 (Arduino Uno)
Send
Distance = 38.32 cm
Distance = 37.93 cm
Distance = 37.83 cm
Distance = 37.93 cm
Distance = 38.34 cm
Distance = 25.85 cm
Distance = 27.35 cm
Autoscroll
No line ending
9600 baud

Our Model

- (Assumed) Model:
  - Range reading = true range + error
- Formalize:

$$Y = r + X$$
$$X \sim \mathrm{Norm}_x[\mu, \sigma^2]$$

- **Problem:** Don't know parameters $\theta = \{\mu, \sigma^2\}$
- **Solution:** Learn from data!
  - Fix r to some distance (1m)
  - Collect range reading deviations ($x[i] = y[i] - r$)
  - Estimate (learn) parameters $\theta = \{\mu, \sigma^2\}$

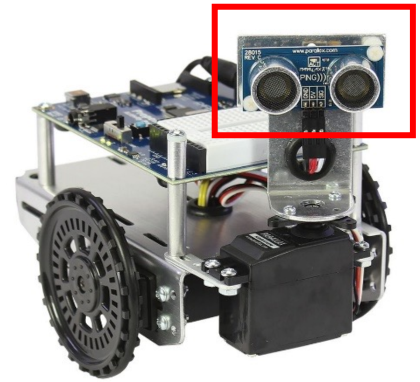


Image credit: https://www.parallax.com/product/910-28015a

Learning Parameters

- Common approaches to learn the unknown parameters θ from a set of given data $\mathcal{D} = \{x[1], \dots, x[N]\}$:

Maximum Likelihood Estimation (MLE)

$$\underset{\theta}{\operatorname{argmax}}\, p(D|\theta)$$

Maximum a Posteriori Estimation (MAP)

$$\underset{\theta}{\operatorname{argmax}}\, p(\theta|D)$$

Variational Bayes

Monte-Carlo Sampling

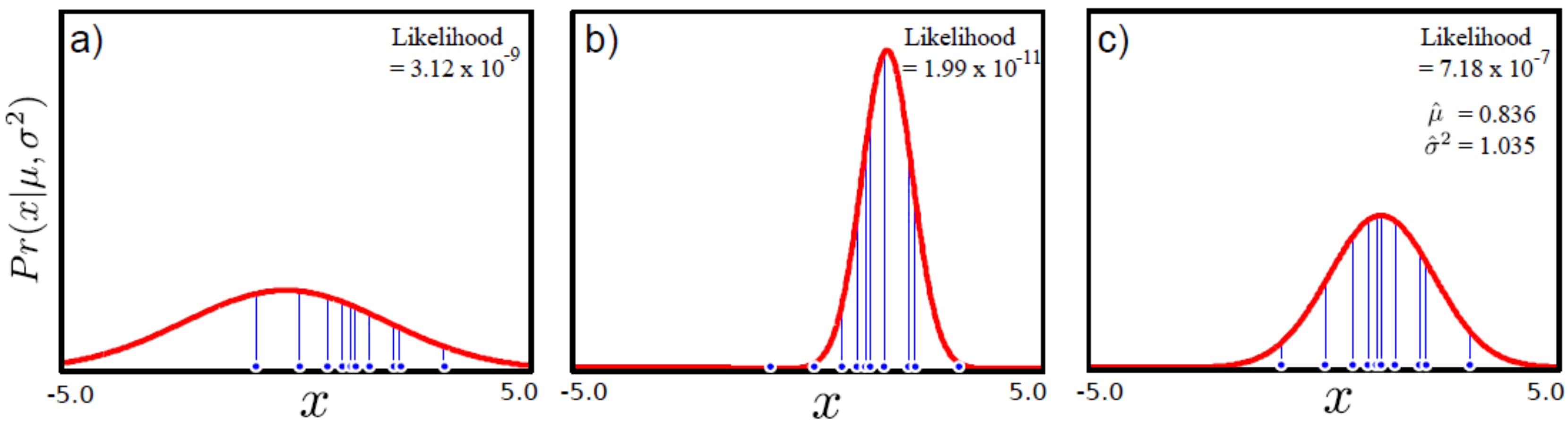
"Full" Bayesian Inference

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Computational Cost

(In general and not to scale)

Maximum Likelihood Estimate: Intuition



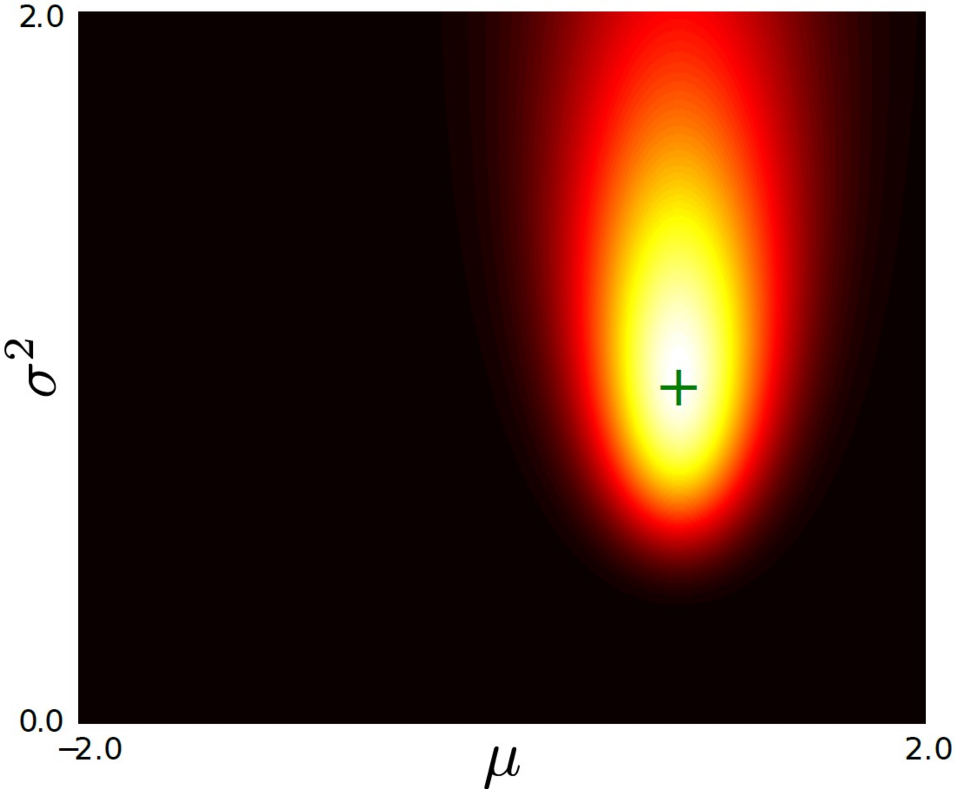
- Blue dots are the observed data $\mathcal{D} = \{x[1], \dots, x[N]\}$.
- Red curves are the Normal distribution for a possible μ and σ^2.
- The likelihood of a set of **independently** sampled data is the **product** of the individual likelihoods $p(x|\mu, \sigma^2)$ (blue vertical lines).
- The maximum likelihood should be correct μ and σ^2

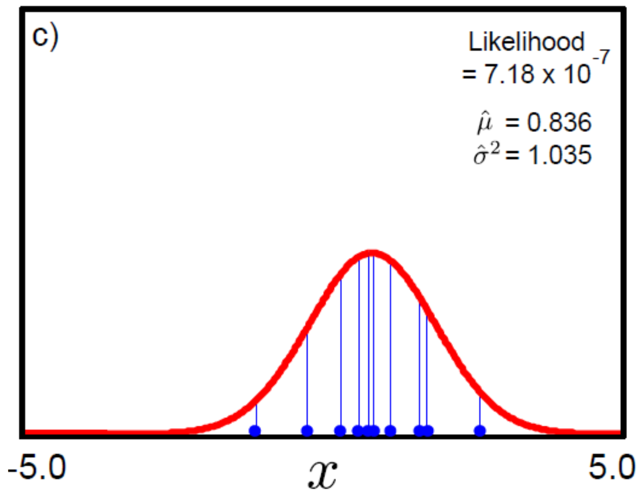
Example 1: Univariate Normal Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

Intuition behind MLE:

$p(x|\mu, \sigma^2)$





Plotted surface of likelihoods as a function of possible parameter values.

ML Solution is at the peak.

Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Maximum Likelihood Estimation

- Given data $\mathcal{D} = \{x[1], \dots, x[N]\}$
- Assume:
  - a set of distributions $\{p_\theta: \theta \in \Theta\}$ where $p_\theta = p(x|\theta)$
  - $\mathcal{D}$ is sample from $X_1, X_2, \dots, X_N \sim p_{\theta^*}$ for some $\theta^* \in \Theta$
  - Random variables $X_1, X_2, \dots, X_N$ are independent and identically distributed (iid) according to p_{θ^*}
- Goal: Estimate θ^*
- The estimate θ_{MLE} is a maximum likelihood estimate (MLE) for θ^* if

$$\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}[p(\mathcal{D}|\theta)]$$

Independent and Identically Distributed (iid)

- Common assumption in many modeling and learning scenarios
- Allows us to decompose the likelihood into products of likelihoods (one for each datum)

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}[p(\mathcal{D}|\theta)]$$

$$= \underset{\theta}{\operatorname{argmax}}\left[\prod_{i=1}^{N} p(X = x[i]\,|\theta)\right] \qquad \text{(i.i.d)}$$

Sensor Uncertainty: MLE

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\left[\prod_{i=1}^{N} p(X = x[i] \,|\theta)\right]$$

In our case, X is Normal / Gaussian distributed.

Fit an univariate normal distribution model to a set of scalar data $\mathcal{D} = \{x[1], \ldots x[N]\}$.

Recall that the univariate normal distribution is given by:

$$p(x) = \mathrm{Norm}_x[\mu, \sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$

Our goal is to find the two unknown parameters μ and σ^2.

Example 1: Univariate Normal Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}[p(x|\theta)]$$

$$= \underset{\theta}{\operatorname{argmax}}\left[\prod_{i=1}^{N} p(x[i] \mid \theta)\right] \qquad \text{(iid)}$$

Likelihood given by pdf

$$p(x|\mu,\sigma^2) = \mathrm{Norm}_x[\mu,\sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$

Example 1: Univariate Normal Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

Algebraically:

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}}[p(x|\mu,\sigma^2)]$$

where

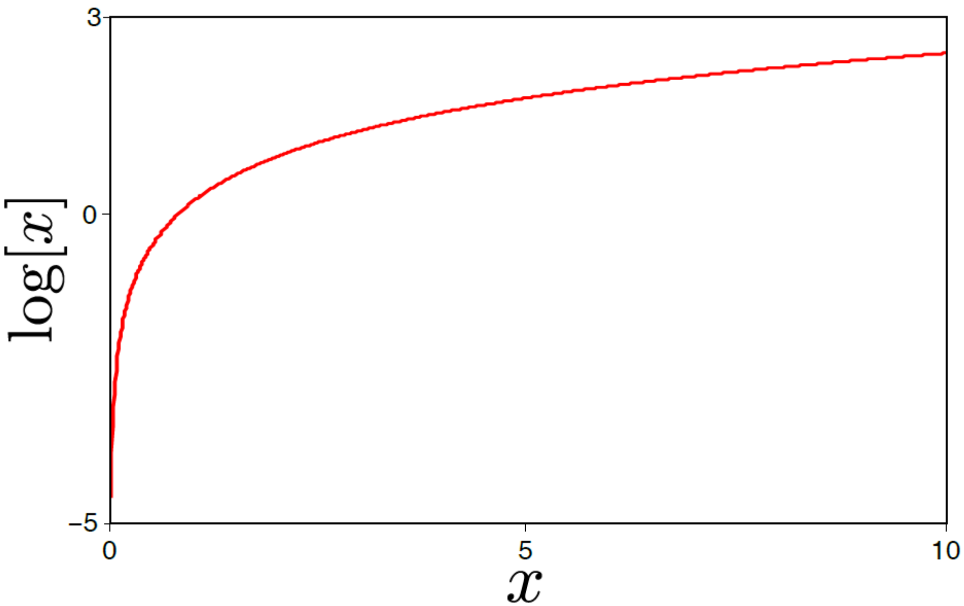
$$p(x|\mu,\sigma^2) = \prod_{i=1}^{N} \operatorname{Norm}_{x[i]}[\mu,\sigma^2],$$

or alternatively, we can maximize the logarithm:

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log\left[\operatorname{Norm}_{x[i]}[\mu,\sigma^2]\right]$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}}\left[-0.5N\log[2\pi] - 0.5N\log\sigma^2 - 0.5\sum_{i=1}^{N}\frac{(x[i]-\mu)^2}{\sigma^2}\right]$$

Why the Logarithm?



- The logarithm is a monotonic transformation.
- Hence, the position of the peak stays in the same place.
- The log likelihood is easier to work with.

Image Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Example 1: Univariate Normal Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log\left[\mathrm{Norm}_{x[i]}[\mu, \sigma^2]\right]$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \underbrace{\left[-0.5N\log[2\pi] - 0.5N\log\sigma^2 - 0.5\sum_{i=1}^{N}\frac{(x[i]-\mu)^2}{\sigma^2}\right]}_{L}$$

Maximization can be done in closed-form by taking derivative w.r.t. the variable and equate to zero:

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{N}\frac{(x[i]-\mu)}{\sigma^2} = \frac{\sum_{i=1}^{N} x[i]}{\sigma^2} - \frac{N\mu}{\sigma^2} = 0, \qquad \frac{\partial L}{\partial \sigma^2} = -\frac{N}{\sigma^2} + \sum_{i=1}^{N}\frac{(x[i]-\mu)^2}{\sigma^4} = 0$$

$$\boxed{\Rightarrow \quad \hat{\mu} = \frac{\sum_{i=1}^{N} x[i]}{N} = \bar{x}, \qquad \Rightarrow \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(x[i]-\mu)^2}{N}}$$

Least Squares Interpretation

Maximum likelihood for the mean of the normal distribution...

$$\hat{\mu} = \underset{\mu}{\operatorname{argmax}} \left[-0.5N\log\left[2\pi\right] - 0.5N\log\sigma^2 - 0.5\sum_{i=1}^{N}\frac{(x[i]-\mu)^2}{\sigma^2} \right]$$

$$= \underset{\mu}{\operatorname{argmax}} \left[-\sum_{i=1}^{N}(x[i]-\mu)^2 \right]$$

$$= \underset{\mu}{\operatorname{argmin}} \left[\sum_{i=1}^{N}(x[i]-\mu)^2 \right]$$

...gives `least squares' fitting criterion.

Let’s try it out.

https://github.com/crslab/CS5340-notebooks

MLE: Properties

- **Easy and fast** to compute
- Nice Asymptotic properties:
  - **Consistent**: if data generated from $f(\theta^*)$, MLE converges to its true value, $\hat{\theta}_{MLE} \to \theta^*$ as $n \to \infty$
  - **Efficient**: there is no consistent estimator that has lower mean squared error than the MLE estimate (achieves Cramer-Rao lower bound)
- **Functional Invariance**: if $\hat{\theta}$ is the MLE of θ^*, and $g(\theta^*)$ is a transformation of θ^* then the MLE for $\alpha = g(\theta^*)$ is $\hat{\alpha} = g(\hat{\theta})$

What is a problem?

- Imagine if you had samples:

  {0.3,-0.1, 1.2, 0.2, -0.2}

  - What is your MLE estimate of the mean and variance?
  - $\mu_{MLE} = 0.28$, $\sigma^2 = 0.245$
- Manual says that on average devices have *zero* bias $\mu = 0$ and variance $\sigma^2 = 0.05$
- How certain are you that your estimate is correct?

Other issues

- MLE is a point estimate i.e., does not represent uncertainty over the estimate
- MLE may overfit.
- MLE does not incorporate prior information.
- Asymptotic results are for the limit and assumes model is correct.
- MLE may not exist or may not be unique

How can we model our *uncertainty* about the parameters estimates μ and σ^2?

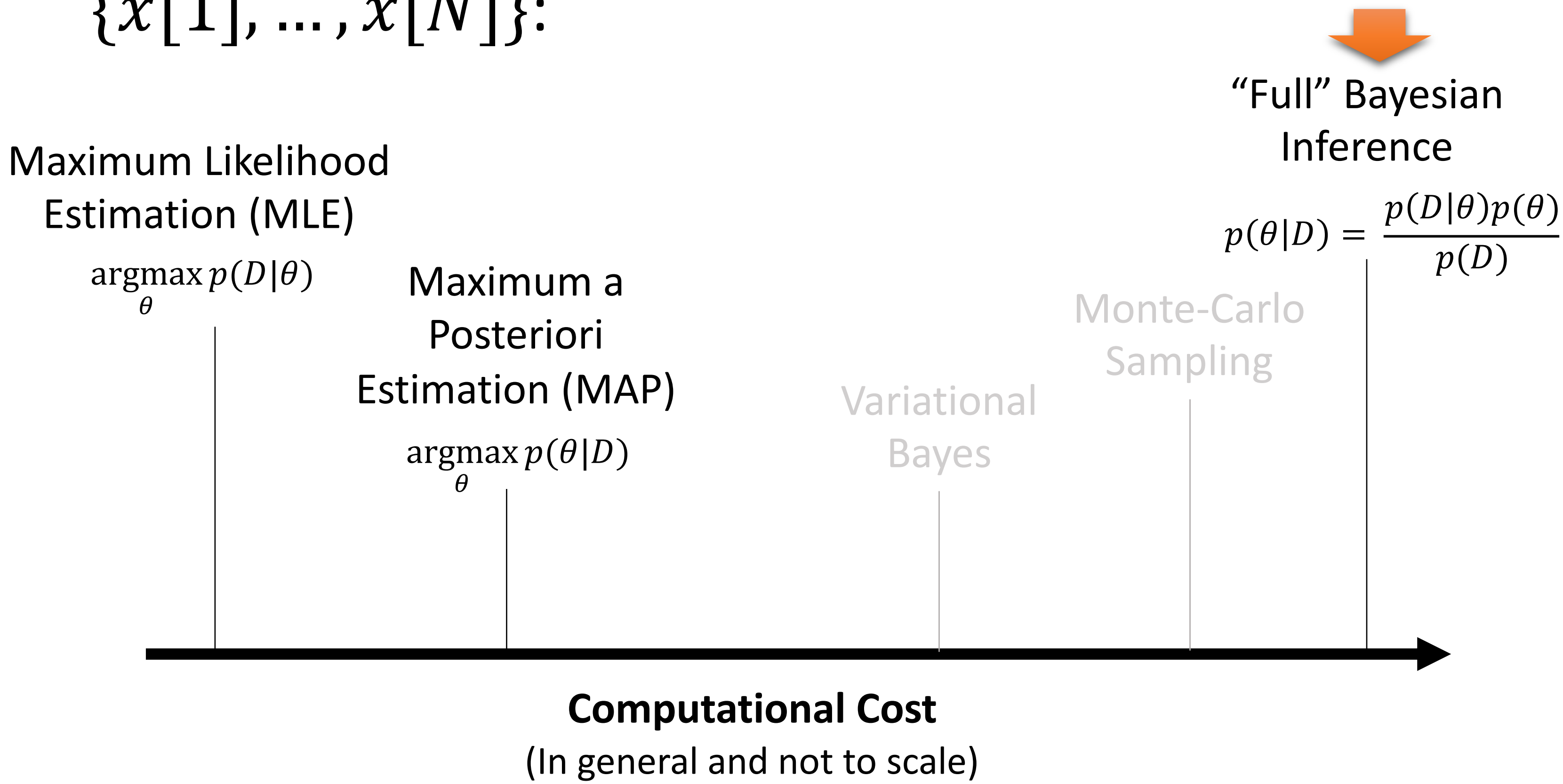
NUS National University of Singapore | School of Computing

Learning via Bayes

Bayesian Inference and Conjugate Models

Learning Parameters

- Common approaches to learn the unknown parameters θ from a set of given data $\mathcal{D} = \{x[1], \ldots, x[N]\}$:



Bayesian Approach

- **Fitting**: Instead of a point estimate $\hat{\theta}$, compute the posterior distribution over all possible parameter values using Bayes' rule:

$$p(\theta|D) = \frac{p(D|\,\theta)p(\theta)}{p(D)}$$

- **Principle**: why pick one set of parameters? There are many values that could have explained the data. Try to capture all of the possibilities.

Our Model

- Possible (Assumed) Model:
  - Range reading = true range + error
- Formalize:

$$Y = r + X$$
$$X \sim \mathrm{Norm}_x[\mu, \sigma^2]$$

- $\theta = \{\mu, \sigma^2\}$ is now a random variable
- Model uncertainty over θ using prior distribution(s).
- Then, find posterior:

$$p(\theta|D) = \frac{\prod_{i=1}^{N} p(x[i] \mid \theta) p(\theta)}{p(D)}$$

- What can be a prior distribution?

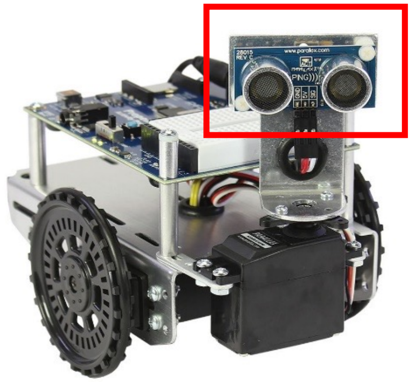


Image credit: https://www.parallax.com/product/910-28015a

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

Compute the posterior distribution using Bayes' rule:

$$p(\theta|x) = \frac{\prod_{i=1}^{N} p(x[i] \mid \theta)p(\theta)}{p(x)} = \frac{\prod_{i=1}^{N} p(x[i] \mid \theta)p(\theta)}{\int \prod_{i=1}^{N} p(x[i] \mid \theta)p(\theta)\, d\theta}$$

where:

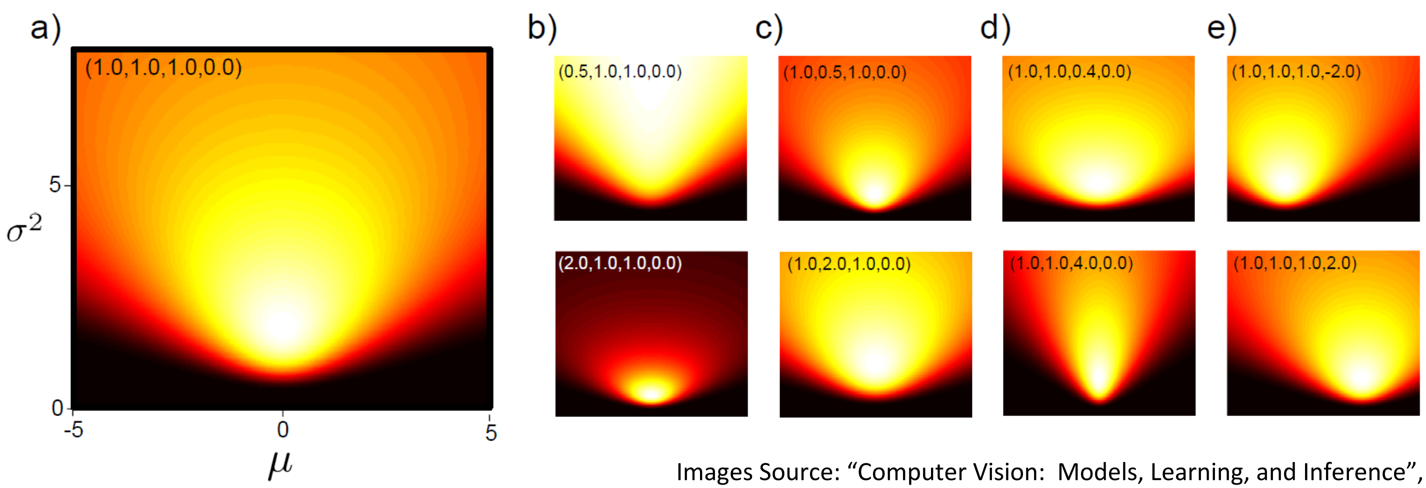
$$\prod_{i=1}^{N} p(x[i] \mid \theta)p(\theta) = \prod_{i=1}^{N} \mathrm{Norm}_{x[i]}[\mu, \sigma^2]\, \underbrace{\mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]}_{\textbf{Conjugate Prior!}}$$

From Lecture 1: Appendix
Normal Inverse Gamma Distribution

$$p(\mu,\sigma^2) = \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}}\frac{\beta^\alpha}{\Gamma[\alpha]}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left[-\frac{2\beta+\gamma(\delta-\mu)^2}{2\sigma^2}\right]$$

$$p(\mu,\sigma^2) = \mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta]$$

- Four hyperparameters α, β, $\gamma > 0$ and $\delta \in \mathbb{R}$.



Images Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Normal Inverse Gamma

- Where does it “come from”?
  - From Normal and Inverse Gamma!
  - Normal is the prior over mean μ
    - $\text{Norm}_{\mu}\,[\delta, s]$
  - Inverse Gamma (IG) is the prior over variance σ^2
    - $\text{InvGam}_{\sigma^2}[\alpha, \beta]$
  - Multiply $\text{Norm}_{\mu}\,[\delta, s]$ and $\text{InvGam}_{\sigma^2}[\alpha, \beta]$ to derive $\text{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]$
    - where $\gamma = \frac{\sigma^2}{s}$

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

Compute the posterior distribution using Bayes' rule:

$$p(\theta|x) = \frac{\prod_{i=1}^{N} p(x[i]|\theta)p(\theta)}{p(x)} = \frac{\prod_{i=1}^{N} p(x[i]|\theta)p(\theta)}{\int \prod_{i=1}^{N} p(x[i]|\theta)p(\theta)\, d\theta}$$

$$\prod_{i=1}^{N} p(x[i]|\theta)p(\theta) = \prod_{i=1}^{N} \mathrm{Norm}_{x[i]}[\mu, \sigma^2]\, \mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]$$

What distribution is $p(\theta|D)$?

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

Compute the posterior distribution using Bayes' rule:

$$p(\theta|x) = \frac{\prod_{i=1}^{N} p(x[i]|\theta)p(\theta)}{p(x)} = \frac{\prod_{i=1}^{N} p(x[i]|\theta)p(\theta)}{\int \prod_{i=1}^{N} p(x[i]|\theta)p(\theta)\, d\theta}$$

$$\prod_{i=1}^{N} p(x[i]|\theta)p(\theta) = \prod_{i=1}^{N} \mathrm{Norm}_{x[i]}[\mu, \sigma^2]\, \mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]$$

$$\boxed{p(\theta|D) = \mathrm{NormInvGam}_{\mu,\sigma^2}\left[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}\right]}$$

NormInvGamma is Conjugate Prior for the Normal.

Posterior Form

$$p(\mu,\sigma^2|D) = \frac{\sqrt{\tilde{\gamma}}}{\sigma\sqrt{2\pi}}\frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma[\tilde{\alpha}]}\left(\frac{1}{\sigma^2}\right)^{\tilde{\alpha}+1}\exp\left[-\frac{2\tilde{\beta}+\tilde{\gamma}\left(\tilde{\delta}-\mu\right)^2}{2\sigma^2}\right]$$

$$p(\theta|D) = \mathrm{NormInvGam}_{\mu,\sigma^2}\left[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}\right]$$

where

$$\tilde{\alpha} = \alpha + \frac{N}{2}, \qquad \tilde{\beta} = \beta + \frac{\sum_i (x[i] - \bar{x})^2}{2} + \frac{N\gamma}{N+\gamma}\frac{(\bar{x}-\delta)^2}{2}.$$

$$\tilde{\delta} = \frac{(\gamma\delta + N\,\bar{x})}{\gamma + N}, \qquad \tilde{\gamma} = \gamma + N, \qquad \bar{x} = \frac{1}{N}\sum_i x[i]$$

Let's try it out.

Bayesian Approach: Properties

- Models uncertainty over parameters.
- Principled way of incorporating prior information.
- Can derive quantities of interest, e.g., $p(x < 10|\mathcal{D})$
- Can perform model selection.

Problem

- “Forced” to select a prior
- What if your initial belief was not conjugate to the normal likelihood?
  - Lognormal, Uniform, Beta …
- Can be computationally intractable

Can we still incorporate prior information into the parameter estimation?

(later in the semester, we will study *approximate* Bayesian inference where we derive an *approximate* posterior distribution)

Learning via MAP

Maximum a Posteriori Estimation(MAP)

Learning Parameters

- Common approaches to learn the unknown parameters θ from a set of given data $\mathcal{D} = \{x[1], \dots, x[N]\}$:

Maximum Likelihood Estimation (MLE)

$$\underset{\theta}{\operatorname{argmax}}\, p(D|\theta)$$

Maximum a Posteriori Estimation (MAP)

$$\underset{\theta}{\operatorname{argmax}}\, p(\theta|D)$$

Variational Bayes

Monte-Carlo Sampling

"Full" Bayesian Inference

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Computational Cost

(In general and not to scale)

Maximum a Posteriori (MAP)

- Given data $\mathcal{D} = \{x[1], \dots, x[N]\}$
- Assume:
  - Joint distribution $p(\mathcal{D}, \theta)$
  - Here θ is a random variable
- Goal: Choose "good" θ
- The estimate θ_{MAP} is a maximum aposteriori estimate (MAP) if

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}[p(\theta|\mathcal{D})]$$

Intuition: The "Peak" of the Posterior

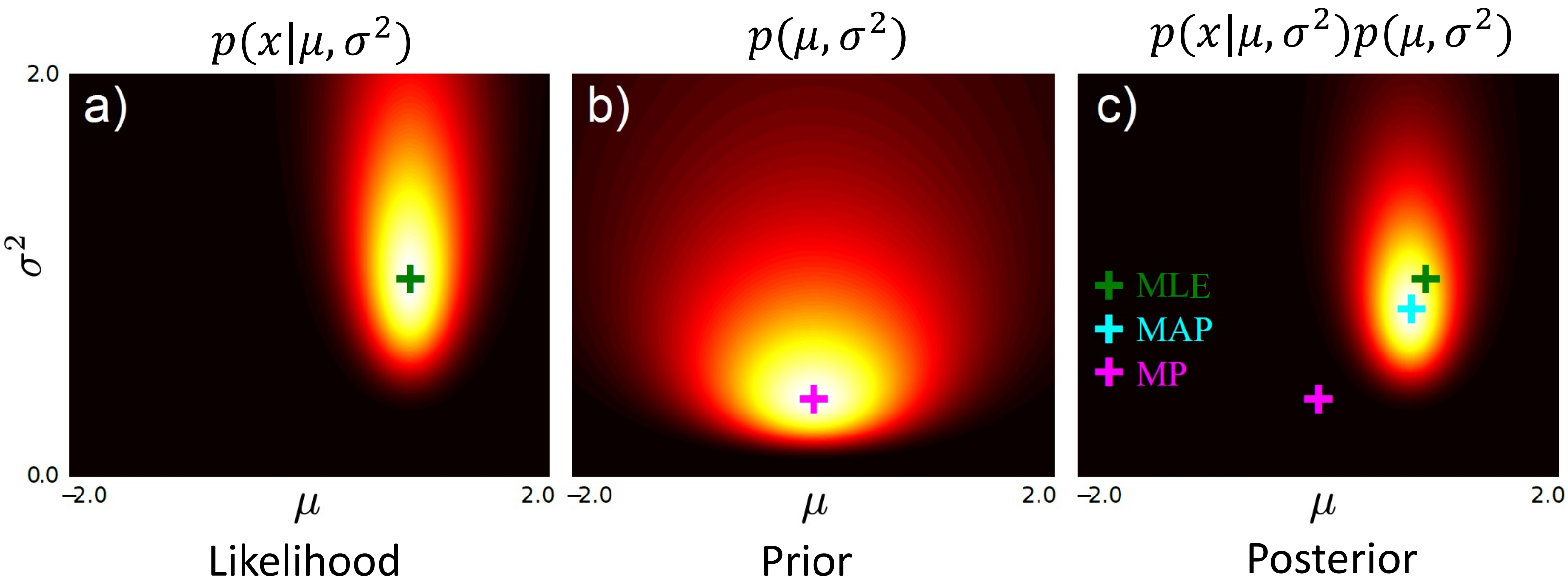


Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Maximum a Posteriori (MAP)

- As the name suggests, we find the unknown parameters θ that maximize the posterior probability $p(\theta|D)$.

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}[p(\theta|D)]$$

$$= \underset{\theta}{\operatorname{argmax}}\left[\frac{p(D|\theta)p(\theta)}{p(D)}\right]$$ (Bayes' rule)

$$= \underset{\theta}{\operatorname{argmax}}\left[\frac{\prod_{i=1}^{N} p(x[i] \mid \theta)\, p(\theta)}{p(D)}\right]$$ (i.i.d)

$$= \underset{\theta}{\operatorname{argmax}}\left[\prod_{i=1}^{N} p(x[i] \mid \theta)\, p(\theta)\right]$$ ($p(D)$ is removed since it is independent of θ)

Example 1: Univariate Normal Distribution

Approach 2: Maximum a Posteriori (MAP)

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \left[\prod_{i=1}^{N} \underbrace{p(x[i] \mid \theta)}_{\text{Likelihood}} \underbrace{p(\theta)}_{\text{Prior}} \right]$$

Likelihood: univariate Normal distribution

$$p(x|\mu, \sigma^2) = \prod_{i=1}^{N} \mathrm{Norm}_{x[i]} [\mu, \sigma^2],$$

Prior: normal inverse gamma distribution

$$p(\mu, \sigma^2) = \mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]$$

(you can try an alternative prior, but we'll use this for now
to compare against the full Bayesian approach)

Example 1: Univariate Normal Distribution

Approach 2: Maximum a Posteriori (MAP)

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \left[\prod_{i=1}^{N} \underbrace{p(x[i] \mid \theta)}_{\text{Likelihood}} \, \underbrace{p(\theta)}_{\text{Prior}} \right]$$

Likelihood: univariate Normal distribution

$$p(x|\mu, \sigma^2) = \prod_{i=1}^{N} \mathrm{Norm}_{x[i]} \left[\mu, \sigma^2\right],$$

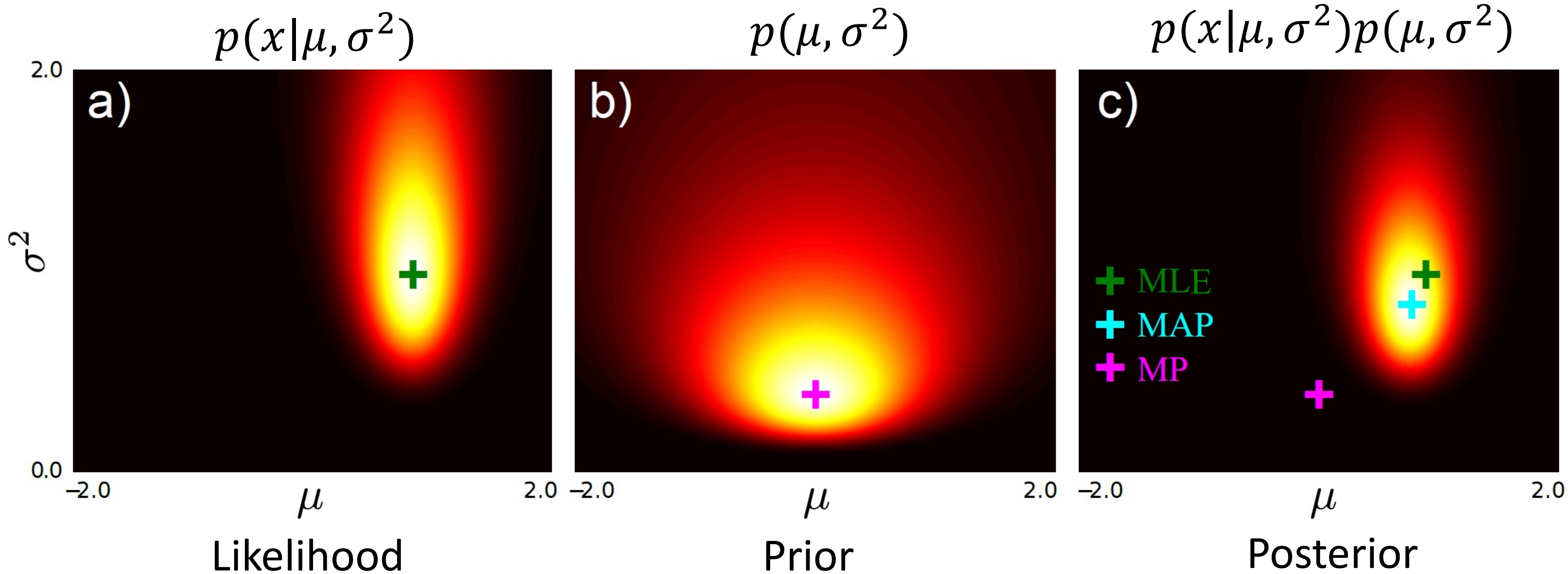
Prior: normal inverse gamma distribution

$$p(\mu, \sigma^2) = \mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]$$

$$= \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma[\alpha]} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{2\beta + \gamma(\delta-\mu)^2}{2\sigma^2}\right]$$

Example 1: Univariate Normal Distribution

Approach 2: Maximum a Posteriori (MAP)



$$\hat{\mu},\hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}}\left[\prod_{i=1}^{N} p(x[i]|\mu,\sigma^2)\,p(\mu,\sigma^2)\right]$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}}\left[\prod_{i=1}^{N} \mathrm{Norm}_{x[i]}[\mu,\sigma^2]\,\mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta]\right]$$

Image Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Example 1: Univariate Normal Distribution

Approach 2: Maximum a Posteriori (MAP)

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\prod_{i=1}^{N} p(x[i] \mid \mu, \sigma^2)\, p(\mu, \sigma^2) \right]$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\prod_{i=1}^{N} \mathrm{Norm}_{x[i]}[\mu, \sigma^2]\, \mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta] \right]$$

Maximize the logarithm:

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\sum_{i=1}^{N} \log\left[\mathrm{Norm}_{x[i]}[\mu, \sigma^2]\right] + \log\left[\mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]\right] \right]$$

Example 1: Univariate Normal Distribution

Approach 2: Maximum a Posteriori (MAP)

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\underbrace{\sum_{i=1}^{N} \log\left[\mathrm{Norm}_{x[i]}[\mu, \sigma^2]\right] + \log\left[\mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]\right]}_{L} \right]$$

Taking derivatives and setting to zero:

$$\frac{\partial L}{\partial \mu} = 0, \qquad \frac{\partial L}{\partial \sigma^2} = 0$$

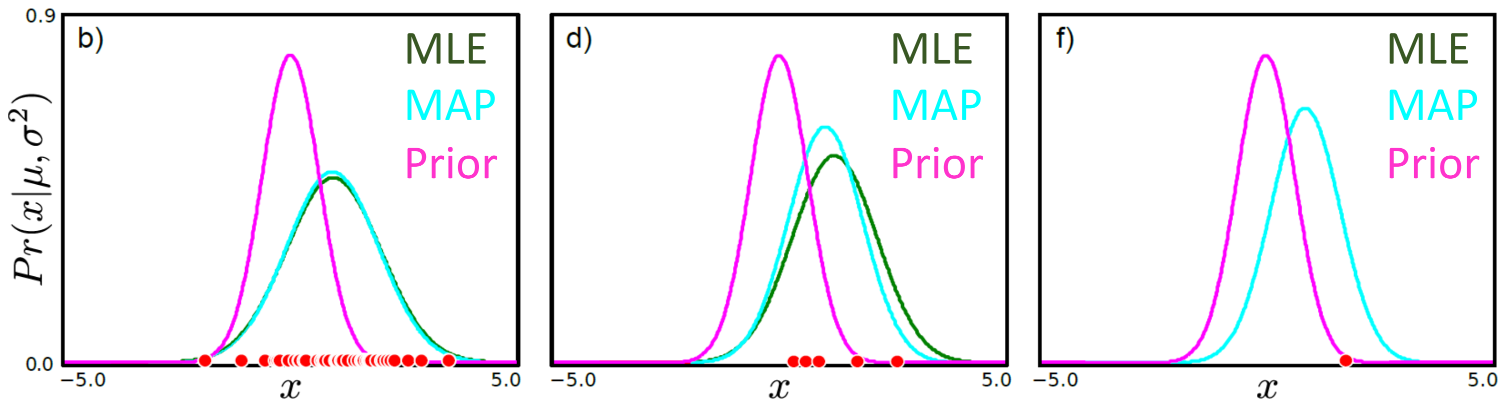
We get:

$$\boxed{\begin{aligned} \hat{\mu} &= \frac{\sum_i x[i] + \gamma\delta}{N + \gamma}, \qquad \hat{\sigma}^2 = \frac{\sum_i (x[i] - \hat{\mu})^2 + 2\beta + \gamma(\delta - \hat{\mu})^2}{N + 3 + 2\alpha} \\ &= \frac{N\bar{x} + \gamma\delta}{N + \gamma} \end{aligned}}$$

Example 1: Univariate Normal Distribution

Approach 2: Maximum a Posteriori (MAP)

More data points → MAP is closer to MLE
Fewer data points → MAP is closer to Prior



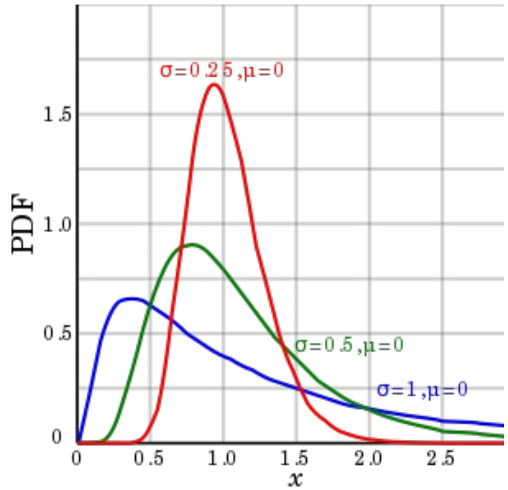
50 data points | 5 data points | 1 data point

Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Let's try it out

Take-Home Exercise: Lognormal Prior

- Say you wanted to use a
  - normal prior for μ
  - lognormal prior for σ^2
- Derive the MAP estimates
  - You can derive a closed-form solution for the mean
  - But would need to optimize for σ^2



- Key steps:
  - Derive $\mathcal{L} = \log p(D|\theta)p(\theta) = \log p(D|\theta) + \log p(\theta)$
  - Then set $\frac{\partial L}{\partial \theta_i} = 0$ for each parameter θ_i

MAP: Properties

- Easy and fast to compute
- Incorporate prior information
- Avoid overfitting ("Regularization")
- As $n \to \infty$, MAP tends to look like MLE
  - but does not have the same nice asymptotic properties.

MAP: Problems

- **Point estimate** (like MLE)
  - Does not capture uncertainty over estimates
  - "Poor man's Bayes"
- Still "forced" to choose prior.
- **NOT Functionally Invariant**: if $\hat{\theta}$ is the MAP of θ^*, and $g(\theta^*)$ is a transformation of θ^* then the MAP for $\alpha = g(\theta^*)$ is **not necessarily** $\hat{\alpha} = g(\hat{\theta})$

Prediction

Maximum Likelihood Estimation (MLE), Maximum a posteriori (MAP), and Bayesian posterior

Predictions for 3 Approaches

Maximum Likelihood Estimate (MLE):

Evaluate new data point x^* under probability distribution with MLE parameters $p(x^*|\theta_{MLE})$.

Maximum a Posteriori (MAP):

Evaluate new data point x^* under probability distribution with MAP parameters $p(x^*|\theta_{MAP})$.

Let’s try it out

Predictions for 3 Approaches

Maximum Likelihood Estimate (MLE):

Evaluate new data point x^* under probability distribution with MLE parameters $p(x^*|\theta_{MLE})$.

Maximum a Posteriori (MAP):

Evaluate new data point x^* under probability distribution with MAP parameters $p(x^*|\theta_{MAP})$.

Bayesian:

Calculate weighted sum of predictions from all possible values of parameters

$$p(x^*|D) = \int p(x^*|\theta)p(\theta|D)d\theta$$

Bayesian Approach

Predictive Density:

$$p(x^*|\mathcal{D}) = \frac{p(x^*, D)}{p(D)}$$ (Conditional probability)

$$= \frac{\int p(x^*, D, \theta)d\theta}{p(D)}$$ (Marginalization)

$$= \frac{\int p(x^*, \theta|D)\cancel{p(D)}d\theta}{\cancel{p(D)}}$$ (Chain Rule)

$$= \int p(x^*|D, \theta)p(\theta|D)d\theta$$ (Chain Rule)

$$= \int p(x^*|\theta)p(\theta|D)d\theta$$ (Conditional Independence)

Bayesian Approach

Predictive Density:

$$p(x^*|D) = \int p(x^*|\theta)p(\theta|D)d\theta$$

Weights

Prediction for each possible θ

Make a prediction that is an (infinite) weighted sum (integral) of the predictions for each parameter value, where weights are the probabilities.

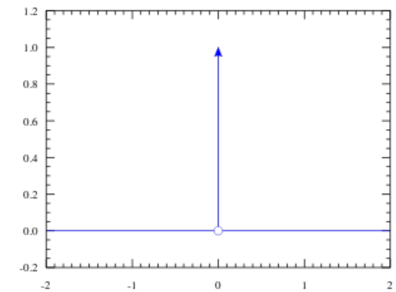
Predictive Densities for 3 Approaches

How to rationalize different forms?

Consider MLE and MAP estimates as probability distributions with zero probability everywhere except at estimate (i.e. delta functions):

$$p(x^*|x) = \int p(x^*|\theta)\delta[\theta - \hat{\theta}]d\theta$$

$$= p(x^*|\hat{\theta})$$



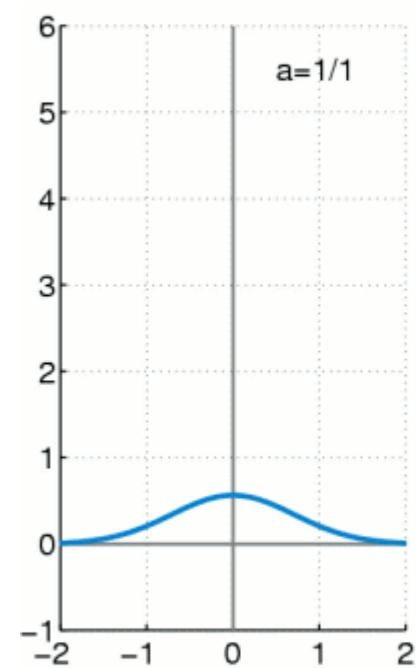


Image Source: Wikipedia

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

Predictive density

Take weighted sum of predictions from different parameter values:

$$p(x^*|D) = \int\int p(x^*|\mu,\sigma^2)p(\mu,\sigma^2|D)d\mu d\sigma^2$$

Posterior: $p(\mu,\sigma^2|D)$

Samples from posterior

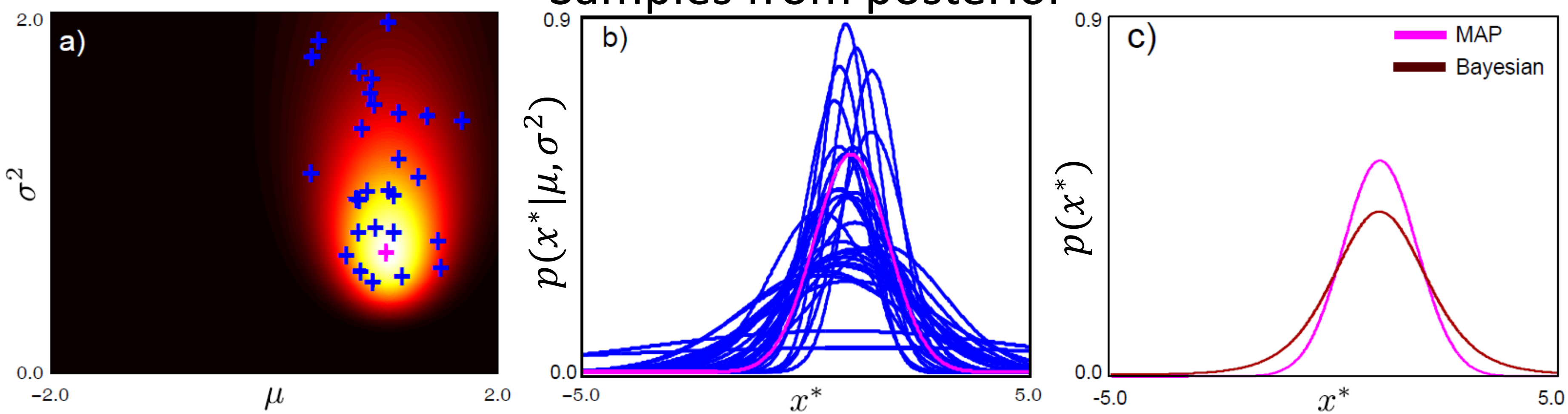


Image Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

Predictive density

Take weighted sum of predictions from different parameter values:

$$p(x^*|x) = t_{2\tilde{\alpha}}\left(x^*|\tilde{\delta}, \frac{\tilde{\beta}(\tilde{\gamma}+1)}{\tilde{\alpha}\tilde{\gamma}}\right)$$

Where $t_{2\tilde{\alpha}}(x^*|\tilde{\delta}, \frac{\tilde{\beta}(\tilde{\gamma}+1)}{\tilde{\alpha}\tilde{\gamma}})$ is the Generalized Student-T distribution with location $\tilde{\delta}$ and scale $\frac{\tilde{\beta}(\tilde{\gamma}+1)}{\tilde{\alpha}\tilde{\gamma}}$.

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

As the training data decreases, the Bayesian prediction becomes less certain but the MAP prediction can be erroneously overconfident.

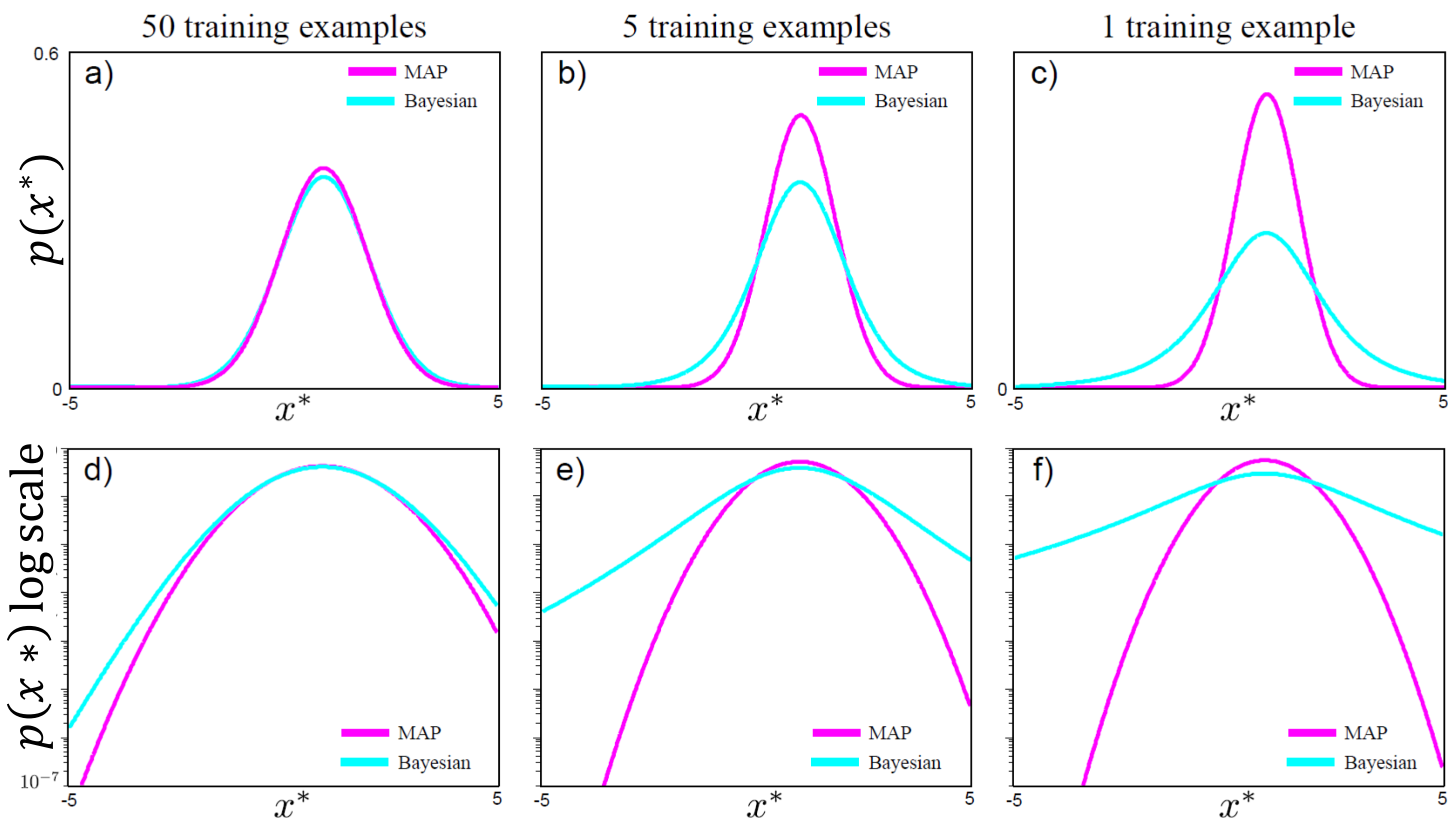
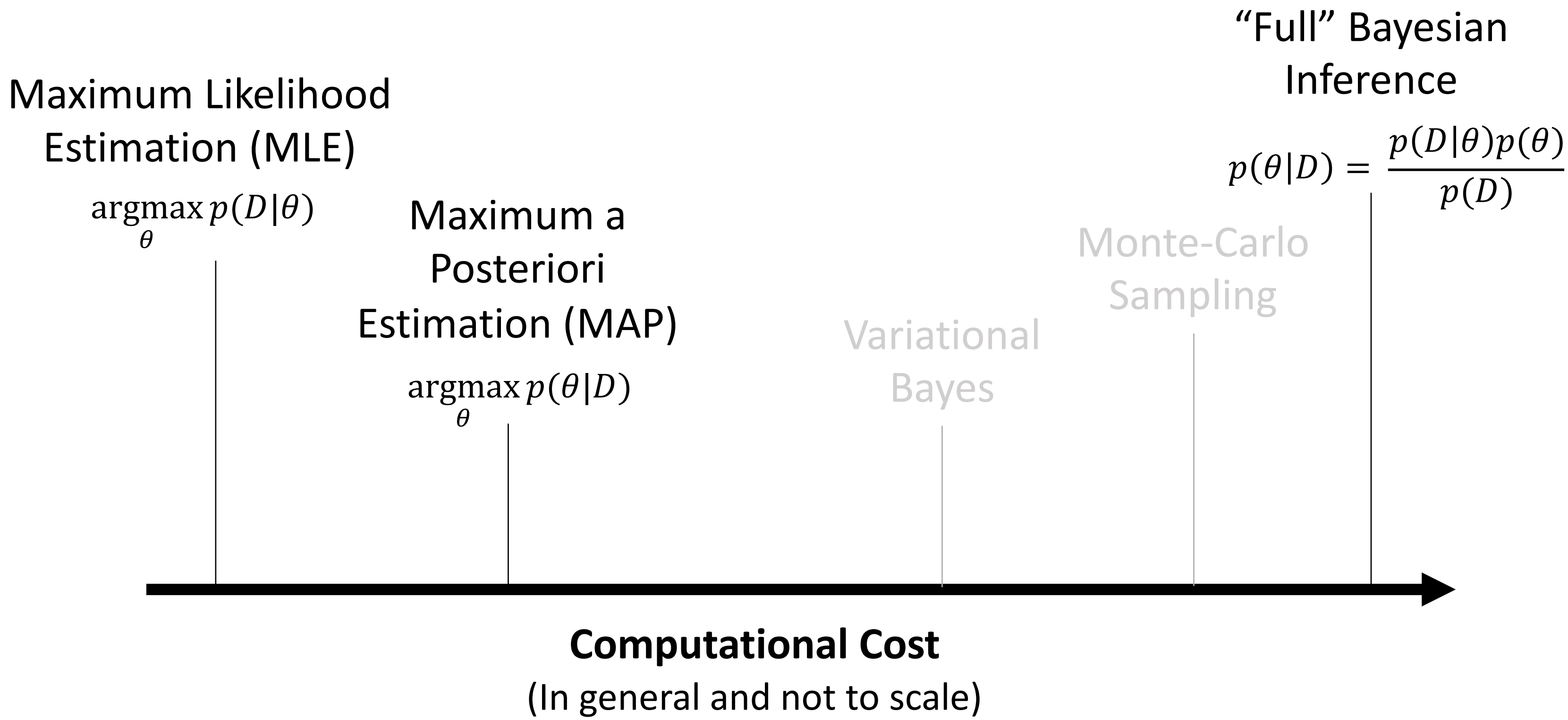


Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Let’s try it out

Learning Parameters

- Common approaches to learn the unknown parameters θ from a set of given data $\mathcal{D} = \{x[1], \ldots, x[N]\}$:



NUS National University of Singapore | School of Computing

Exponential Family

What's an Exponential Family and why should we care?

Exponential Family

- An exponential family (ExpFam) is a set of probability distributions $\{p_\theta : \theta \in \Theta\}$ with the form

$$p_\theta(x) = \frac{h(x)\exp[\eta(\theta)^\top s(x)]}{Z(\theta)}$$

- where:
  - $\theta \in \Theta \subseteq \mathbb{R}^k, x \in \mathbb{R}^d$
  - Natural parameters: $\eta(\theta): \Theta \to \mathbb{R}^m$
  - Sufficient statistics: $s(x): \mathbb{R}^d \to \mathbb{R}^m$
  - Base Measure (Support and scaling): $h(x): \mathbb{R}^d \to [0, \infty)$
  - Partition function: $\mathrm{Z}(\theta): \Theta \to [0, \infty)$

Natural/Canonical form

- An exponential family is in its natural (canonical) form if it is parameterized by its natural parameters:

$$p_\eta(x) = p(x|\eta) = \frac{h(x)\exp[\eta^\top s(x)]}{Z(\eta)}$$

(Compare against $p_\theta(x) = \frac{h(x)\exp[\eta(\theta)^\top s(x)]}{Z(\theta)}$)

ExpFam: So What?!

- Always has conjugate prior!
- Has fixed number of sufficient statistics that summarize iid data (of arbitrary amount!)
- Posterior predictive distribution always has closed form solution (provided $Z(\theta)$ is closed-form).

The Partition function $Z(\eta)$

$$p_\eta(x) = p(x|\eta) = \frac{h(x)\exp[\eta^\top s(x)]}{Z(\eta)}$$

- Also called the normalizer:

$$Z(\eta) = \int h(x)\exp[\eta^\top s(x)]dx$$

- Why? To get normalized distribution:

$$\int p(x|\eta)dx = 1$$
$$\int h(x)\frac{\exp[\eta^\top s(x)]}{Z(\eta)}dx = 1$$
$$Z(\eta) = \int h(x)\exp[\eta^\top s(x)]dx$$

- Aside: sometimes, people write $g(\eta) = 1/Z(\eta)$ and the canonical form becomes:

$$p(x|\eta) = h(x)\, g(\eta)\exp[\eta^\top s(x)]$$

The log Partition function $A(\eta)$

Alternatively, we can specify the log partition function:

$$p_\eta(x) = p(x|\eta) = h(x)\exp[\eta^\top s(x) - A(\eta)]$$

- Is the log of the partition function:

$$A(\eta) = \log Z(\eta) = \log[\int h(x)\exp[\eta^\top s(x)]dx]$$

- Why? To get normalized distribution for any η:

$$\int p(x|\eta)dx = \int h(x)\exp[\eta^\top s(x) - A(\eta)] = 1$$

$$\exp[-A(\eta)]\int h(x)\exp[\eta^\top s(x)]dx = 1$$

$$\exp[A(\eta)] = \int h(x)\exp[\eta^\top s(x)]dx$$

$$A(\eta) = \log[\int h(x)\exp[\eta^\top s(x)]dx]$$

Moments of Sufficient Statistics

- For any exponential family distribution:

$$\mathbb{E}[s(x)] = \nabla \log \mathrm{Z}(\eta) = \nabla A(\eta)$$

- Higher order moments of $s(x)$ given by higher order derivatives.
- If $s(x) = x$ (natural exponential family), we can find moments of x simply by differentiation!

MLE of Parameters of ExpFam

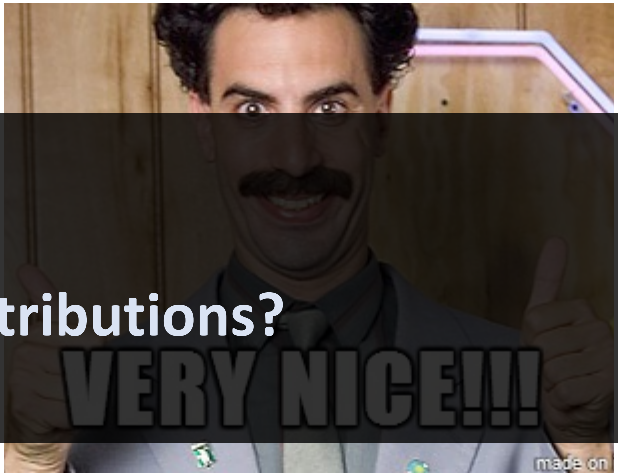
- In addition, the maximum likelihood estimator η_{MLE} satisfies:

$$\nabla A(\eta_{MLE}) = \frac{1}{N}\sum_{n=1}^{N} s(x_n)$$

- We can use this to solve for η_{MLE}
    - Note that A is convex. Proof in Extra Readings.
- The MLE only depends only on sufficient statistics $s(x)$

ExpFam: So What?!

- Always has conjugate prior!
- Has fixed number of sufficient statistics that summarize iid data (of arbitrary amount!)
- Posterior predictive distribution always has closed form solution (provided $Z(\theta)$ is closed-form).



Great!

But how can I find ExpFam distributions?

Is Gaussian an ExpFam?

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

$$p(x) = \mathrm{Norm}_x[\mu, \sigma^2]$$

Rearrange to fit the ExpFam form:

$$p_\eta(x) = p(x|\eta) = \frac{h(x)\exp[\eta^\top s(x)]}{Z(\eta)}$$

the idea is to match the terms into the:

Natural parameters: $\eta(\theta)$

Sufficient statistics: $s(x)$

Base measure: $h(x)$

Partition function: $Z(\eta)$

or Log Partition function: $A(\eta)$

$$p(x|\eta) = h(x)\exp[\eta^\top s(x) - A(\eta)]$$

Many Distributions are ExpFam

PMFs

- Bernoulli
- Binomial
- Categorical/Multinoulli
- Poisson
- Multinomial
- Negative Binomial
- …

PDFs

- Normal
- Gamma & Inverse Gamma
- Wishart & Inverse Wishart
- Beta
- Dirichlet
- lognormal
- Exponential
- …

Exercise: Find a family of distributions that is not ExpFam.

Exponential Family

- An exponential family (ExpFam) is a set of probability distributions $\{p_\theta : \theta \in \Theta\}$ with the form

$$p_\theta(x) = \frac{h(x)\exp[\eta(\theta)^\top s(x)]}{Z(\theta)}$$

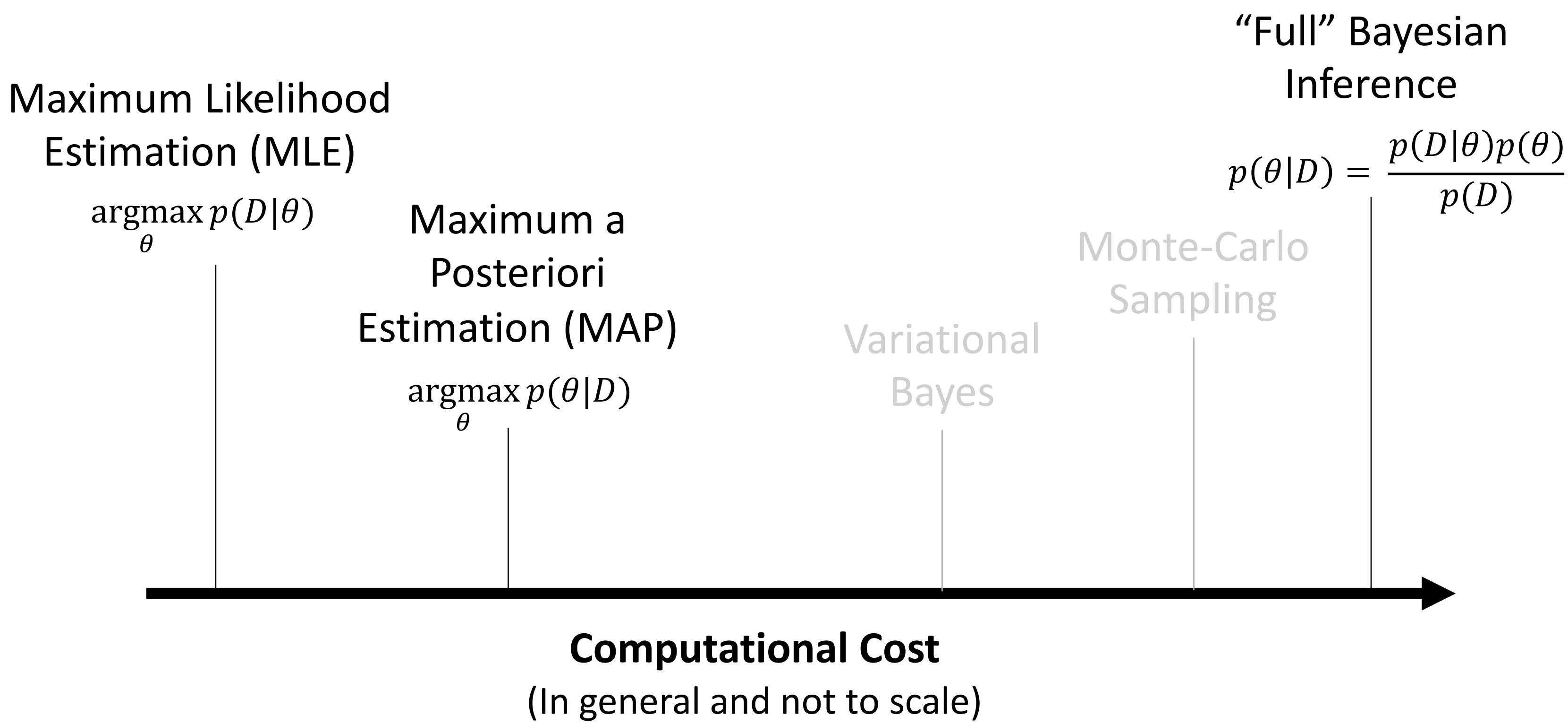
- where:
  - $\theta \in \Theta \subseteq \mathbb{R}^k, x \in \mathbb{R}^d$
  - Natural parameters: $\eta(\theta): \Theta \to \mathbb{R}^m$
  - Sufficient statistics: $s(x): \mathbb{R}^d \to \mathbb{R}^m$
  - Base Measure (Support and scaling): $h(x): \mathbb{R}^d \to [0, \infty)$
  - Partition function: $\mathrm{Z}(\theta): \Theta \to [0, \infty)$

Recap

MLE, MAP, Bayesian Inference, and Exponential Families

Learning Parameters

- Common approaches to learn the unknown parameters θ from a set of given data $\mathcal{D} = \{x[1], \ldots, x[N]\}$:



Exponential Family

- An exponential family (ExpFam) is a set of probability distributions $\{p_\theta : \theta \in \Theta\}$ with the form

$$p_\theta(x) = \frac{h(x)\exp[\eta(\theta)^\top s(x)]}{Z(\theta)}$$

- where:
  - $\theta \in \Theta \subseteq \mathbb{R}^k, x \in \mathbb{R}^d$
  - Natural parameters: $\eta(\theta) : \Theta \to \mathbb{R}^m$
  - Sufficient statistics: $s(x) : \mathbb{R}^d \to \mathbb{R}^m$
  - Base Measure (Support and scaling): $h(x) : \mathbb{R}^d \to [0, \infty)$
  - Partition function: $Z(\theta) : \Theta \to [0, \infty)$

Learning Outcomes

- Students should be able to:

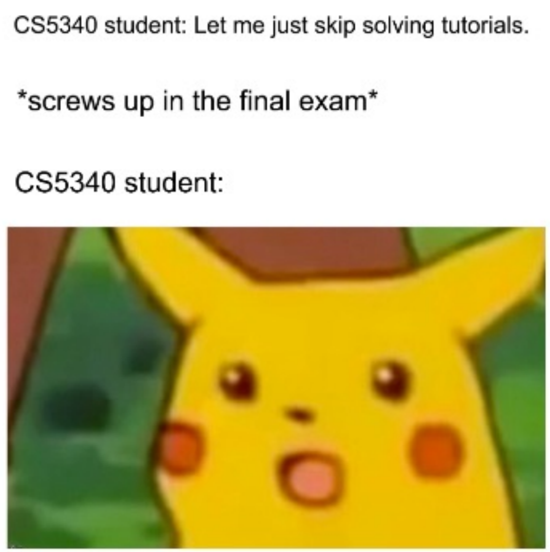
    1. Use the Maximum Likelihood, Maximum a Posteriori and Bayesian approaches to learn the unknown parameters of probability distributions of a single random variable from data.

    2. Apply the assumption independent and identically distributed samples to simplify the parameter learning process.

    3. Apply the learned parameters to make predictions.

    4. Describe the exponential family and its properties

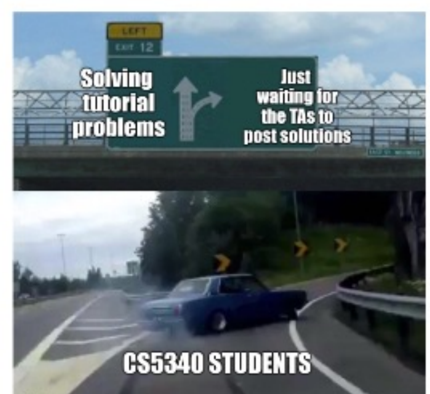
A Discrete Example:
CS5340 Meme of the Year



(a) Surprised Pikachu

(b) Two Buttons Dilemma

(c) Distracted Boyfriend



(d) Left Exit 12

NUS National University of Singapore | School of Computing

CS5340 Meme of the Year

- Model and learn parameters

ID	Template Name	# Votes
1	Surprised Pikachu	25
2	Two Buttons Dilemma	12
3	Distracted Boyfriend	30
4	Left Exit 12	10

Table 1: Votes received by each template by CS5340 students