Bayesian Methods

CS 534: Machine Learning

Material adapted from
Radford Neal's tutorial (http://ftp.cs.utoronto.ca/pub/radford/bayes-tut.pdf),
Zoubin Ghahramni (http://hunch.net/~coms-4771/Zoubin_Ghahramani_Bayesian_Learning.pdf),
Taha Bahadori (http://www-scf.usc.edu/~mohammab/sampling.pdf)

Frequentist vs Bayesian

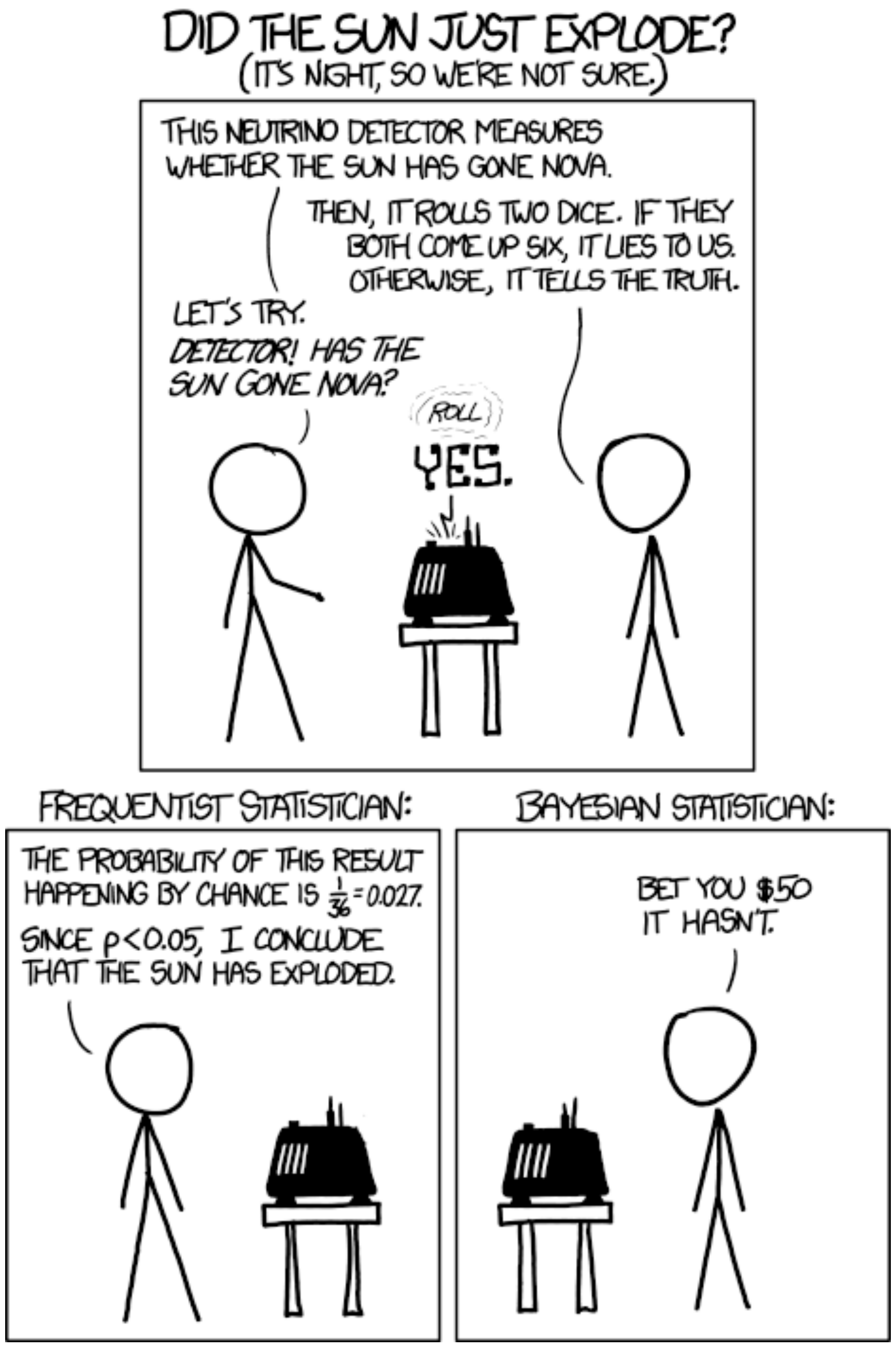
Frequentist

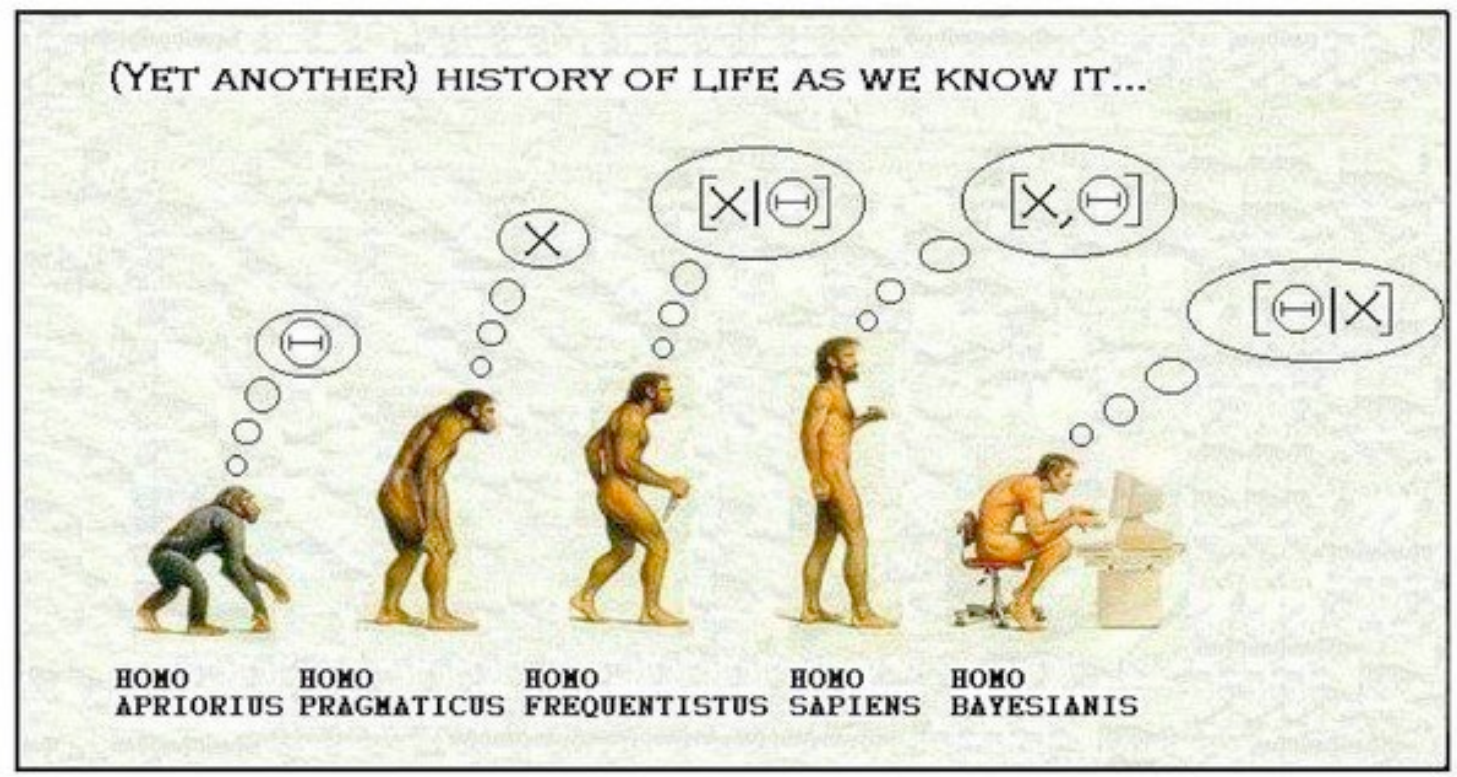
- Data are a repeatable random sample (there is a frequency)
- Underlying parameters remain constant during repeatable process
- Parameters are fixed
- Prediction via the estimated parameter value

Bayesian

- Data are observed from the realized sample
- Parameters are unknown and described probabilistically (random variables)
- Data are fixed
- Prediction is expectation over unknown parameters

The War in Comics





http://conversionxl.com/bayesian-frequentist-ab-testing/

http://www.xkcd.com/1132/

Classic Example: Binomial Experiment

- Given a sequence of coin tosses x_1, x_2, …, x_M, we want to estimate the (unknown) probability of heads

$$P(H) = \theta$$

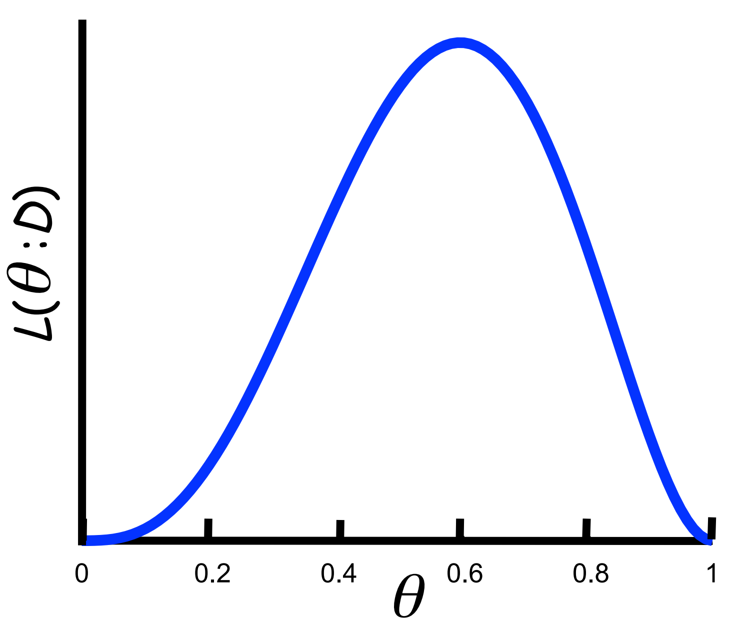
- The instances are independent and identically distributed samples
- Note that x can take on many possible values potentially if we decide to use a multinomial distribution instead

Likelihood Function

- How good is a particular parameter?
  Ans: Depends on how likely it is to generate the data

$$L(\theta; D) = P(D|\theta) = \prod_m P(x_m|\theta)$$

- Example: Likelihood for the sequence H, T, T, H, H



$$\begin{aligned} L(\theta; D) &= \theta(1-\theta)(1-\theta)\theta\theta \\ &= \theta^3(1-\theta)^2 \end{aligned}$$

Maximum Likelihood Estimate (MLE)

- Choose parameters that maximize the likelihood function
  - Commonly used estimator in statistics
  - Intuitively appealing
- In the binomial experiment, MLE for probability of heads

$$\hat{\theta} = \frac{N_H}{N_H + N_T}$$

- Optimization problem approach

Is MLE the only option?

- Suppose that after 10 observations, MLE estimates the probability of a heads is 0.7, would you bet on heads for the next toss?
- How certain are you that the true parameter value is 0.7?
- Were there enough samples for you to be certain?

Bayesian Approach

- Formulate knowledge about situation probabilistically
  - Define a model that expresses qualitative aspects of our knowledge (e.g., forms of distributions, independence assumptions)
  - Specify a **prior** probability distribution for unknown parameters in the model that expresses our beliefs about which values are more or less likely
- Compute the **posterior** probability distribution for the parameters, given observed data
- Posterior distribution can be used for:
  - Reaching conclusions while accounting for uncertainty
  - Make predictions by averaging over posterior distribution

Posterior Distribution

- Posterior distribution for model parameters given the observed data combines the prior distribution with the likelihood function using Bayes' rule

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

- Denominator is just a normalizing constant so you can write it proportionally as

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

- Predictions can be made by integrating with respect to posterior

$$P(\text{new data}|D) = \int_\theta P(\text{new data}|\theta)P(\theta|D)$$

Revisiting Binomial Experiment

- Prior distribution: uniform for θ in [0, 1]
- Posterior distribution:

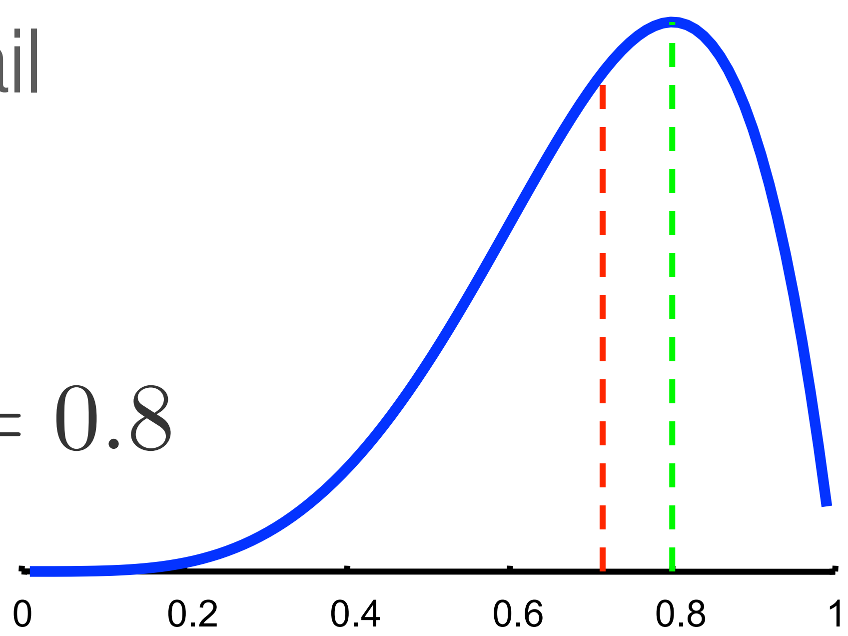
$$P(\theta|x_1, x_2, \cdots, x_M) \propto P(x_1, x_2, \cdots, x_M|\theta) \times 1$$

- Example: 5 coin tosses with 4 heads, 1 tail
  - MLE estimate:

$$P(\theta) = \frac{4}{5} = 0.8, P(x_{M+1} = H|D) = 0.8$$

  - Bayesian prediction:

$$P(x_{M+1} = H|D) = \int \theta P(\theta|D) d\theta = \frac{5}{7}$$



Bayesian Inference and MLE

- MLE and Bayesian prediction differ
- However…
  - IF prior is well-behaved (i.e., does not assign 0 density to any "feasible" parameter value)
  - THEN both MLE and Bayesian prediction converge to the same value as the number of training data increases

Features of the Bayesian Approach

- Probability is used to describe "physical" randomness and uncertainty regarding the true values of the parameters
  - Prior and posterior probabilities represent degrees of belief, before and after seeing the data
- Model and prior are chosen based on the knowledge of the problem and not, in theory, by the amount of data collected or the question we are interested in answering

Priors

- Objective priors: noninformative priors that attempt to capture ignorance and have good frequentist properties
- Subjective priors: priors should capture our beliefs as well as possible. They are subjective but not arbitrary.
- Hierarchical priors: multiple levels of priors
- Empirical priors: learn some of the parameters of the prior from the data (“Empirical Bayes”)
  - Robust, able to overcome limitations of mis-specification of prior
  - Double counting of evidence / overfitting

Computing the Posterior Distribution

- Analytical integration: works when “conjugate” prior distributions can be used, which combine nicely with the likelihood —usually not the case
- Gaussian approximation: works well when there is sufficient data compared to model complexity — posterior distribution is close to Gaussian (Central Limit Theorem) and can be handled by finding its mode
- Markov Chain Monte Carlo: simulate a Markov chain that eventually converges to the posterior distribution —currently the dominant approach
- Variational approximation: cleverer way to approximate the posterior and maybe faster than MCMC but not as general and exact

Parametric vs. Nonparametric

Parametric vs Nonparametric Models

- Parametric models: finite fixed number of parameters, regardless of the size of the dataset (e.g., mixture of k Gaussians)

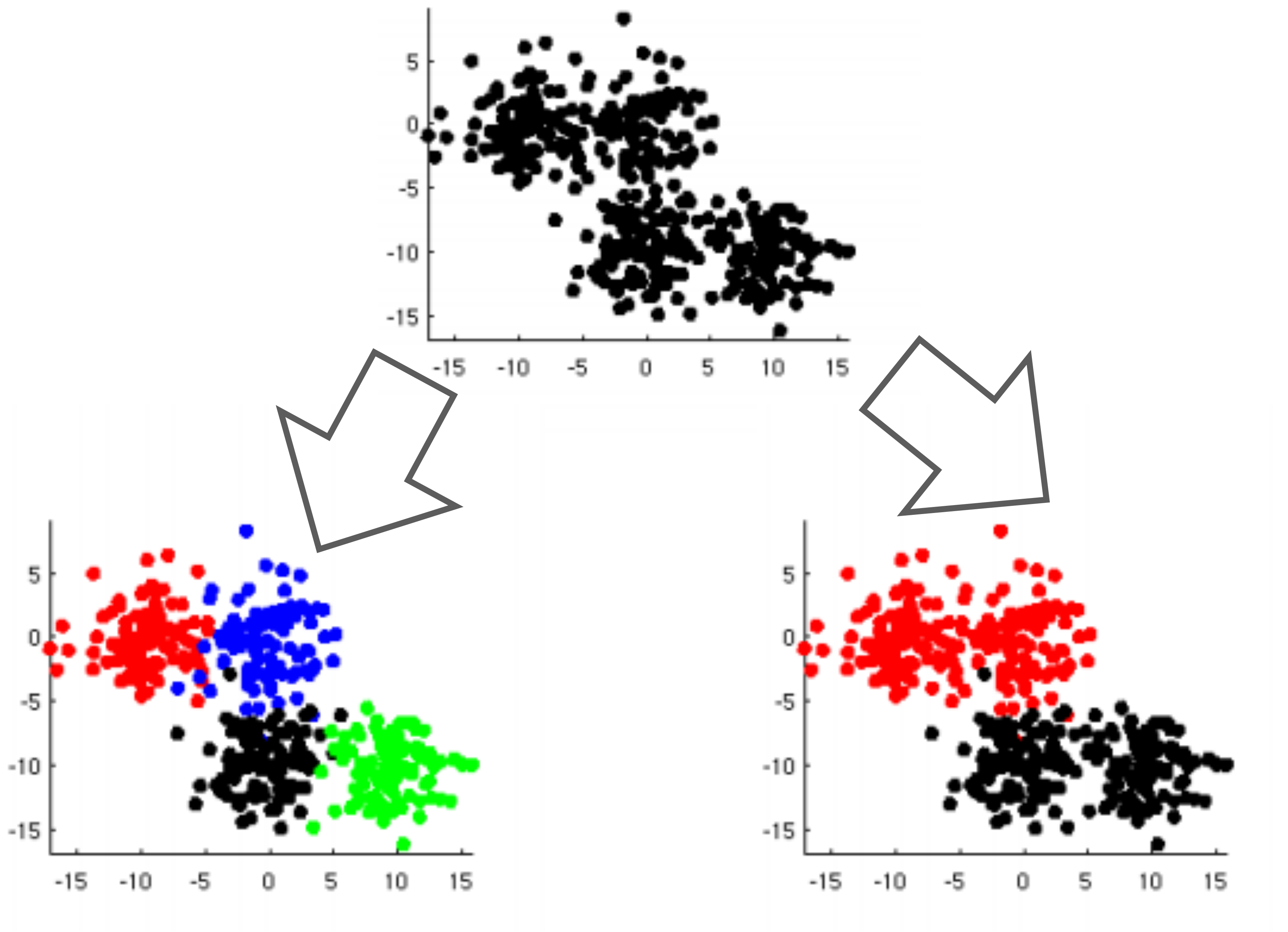
- Non-parametric models: number of parameters are allowed to grow with the data set size, or the predictions depend on the data size

  - Doesn't limit the complexity of our model a priori

  - More flexible and realistic model

  - Better predictive performance

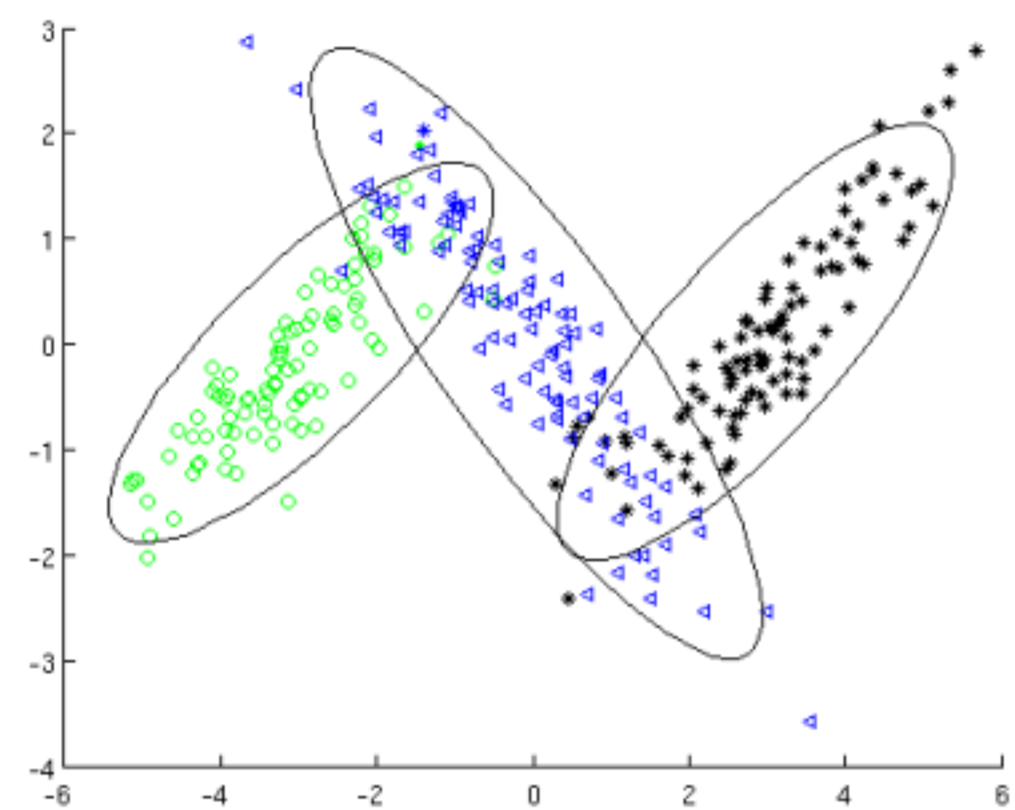
Example: Number of Clusters?



https://www.cs.berkeley.edu/~jordan/courses/294-fall09/lectures/nonparametric/slide

Example: A Frequentist Approach

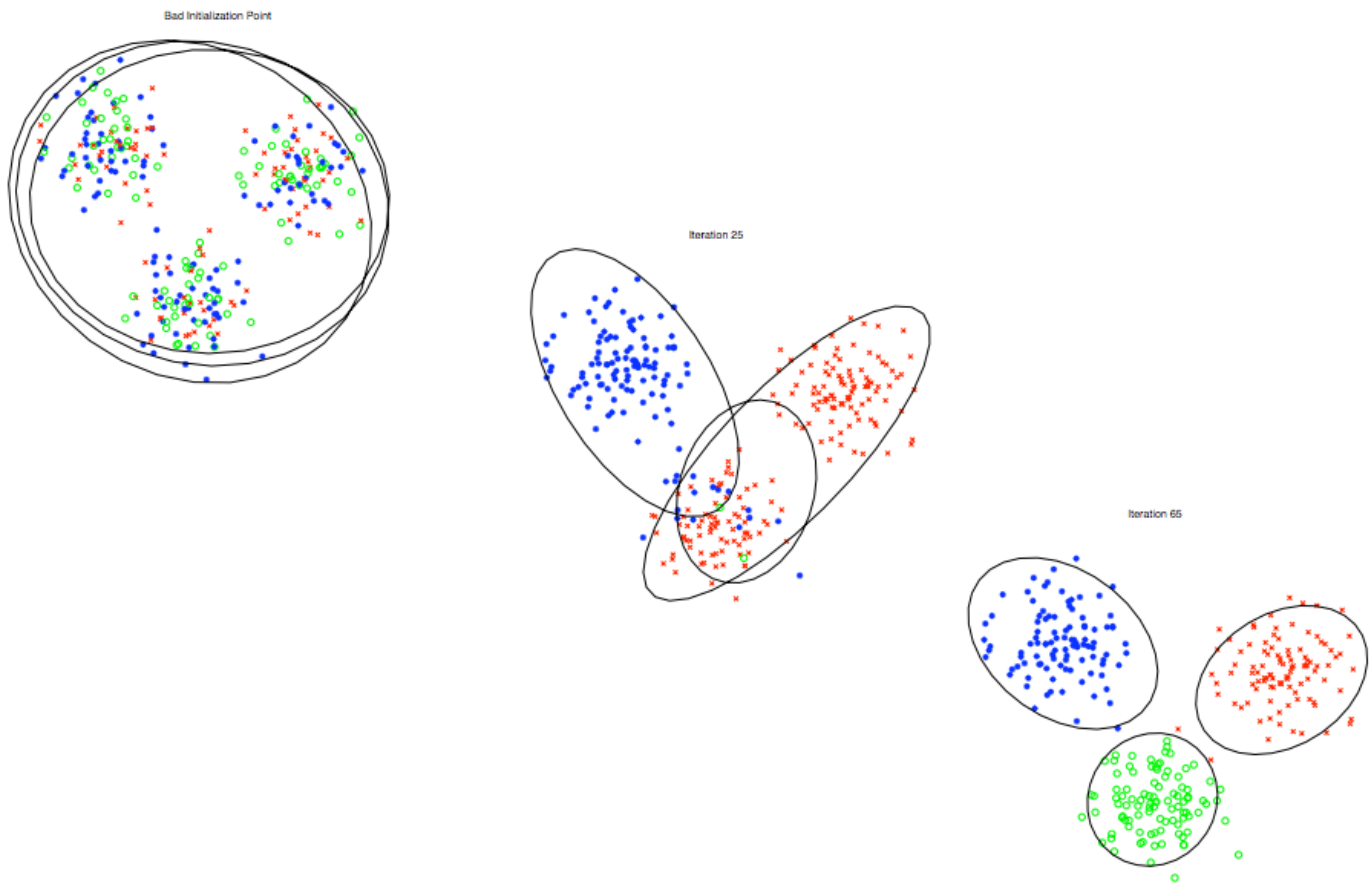
- Gaussian mixture model with K mixtures
  - Distribution over the K classes
  - Each cluster has a mean and covariance
- Use Expectation Maximization (EM) to maximize the likelihood with respect to distribution and cluster points



Example: Bayesian Parametric Approach

- Bayesian Gaussian mixture models with K mixtures
  - Distribution over classes that is drawn from a Dirichlet
  - Each cluster has a mean and covariance that is a Normal-Inverse-Wishart distribution
- Use sampling or variational inference to learn posterior

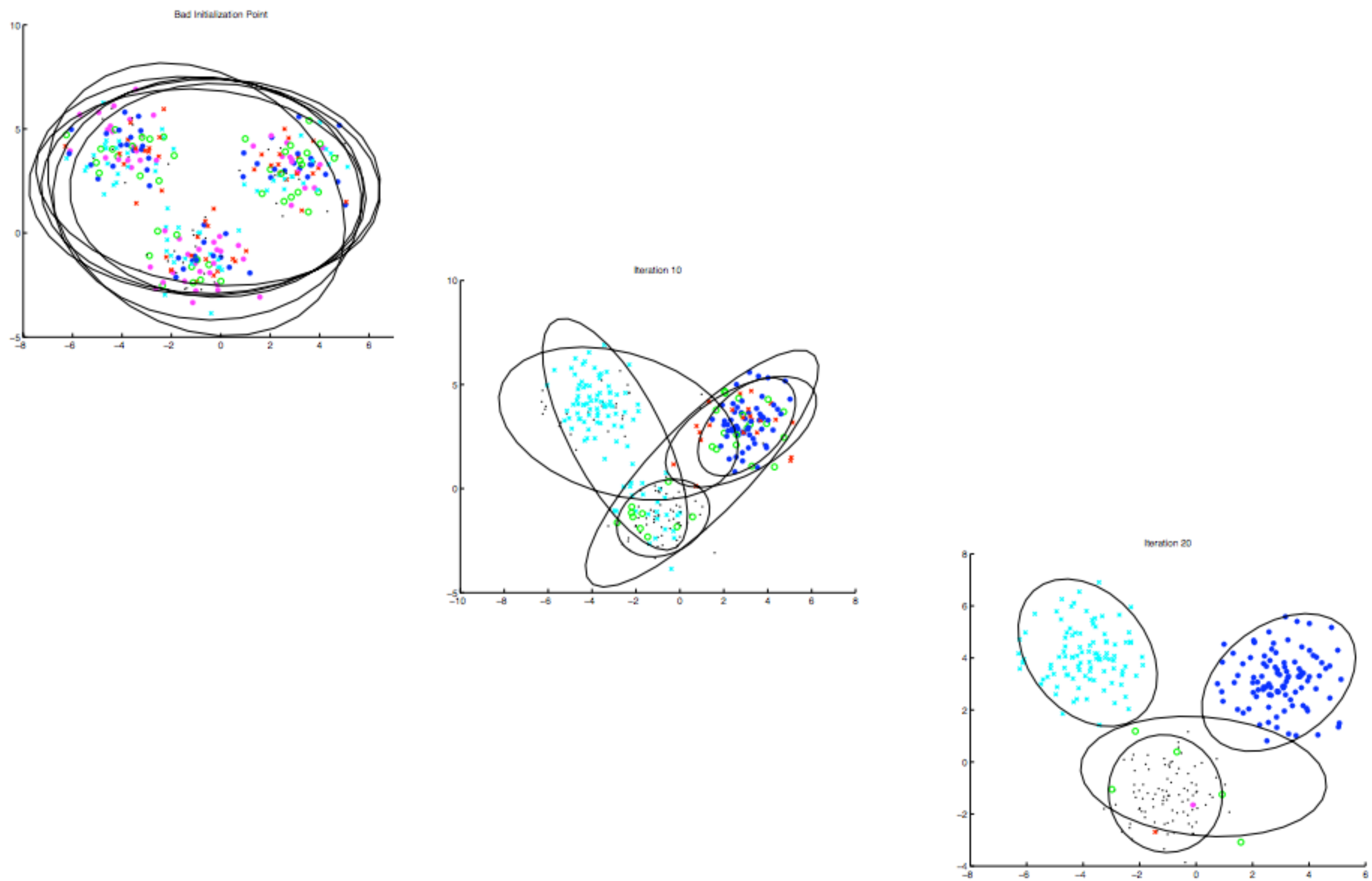
Example: Bayesian Parametric Approach



Example: Nonparametric Bayesian Approach

- Likelihood term looks identical to the parametric case
- Prior distribution uses the Dirichlet Process
    - Flexible, non-parametric prior over infinite number of clusters and their parameters
    - Distribution over distributions
- Use Gibbs sampling to find the right distributions

Example: Nonparametric Bayesian Approach



Limitations and Criticisms of Bayesian Methods

- It is hard to come up with a prior (subjective) and the assumptions may be wrong
- Closed world assumption: need to consider all possible hypotheses for the data before observing the data
- Computationally demanding (compared to frequentist approach)
- Use of approximations weakens coherence argument