Bayesian Nonparametrics

Yee Whye Teh
Gatsby Computational Neuroscience Unit, UCL

http://www.gatsby.ucl.ac.uk/~ywteh/teaching/npbayes2012

Bayesian Machine Learning

Probabilistic Machine Learning

- Machine Learning is all about data.
  - Stochastic, chaotic and/or complex process
  - Noisily observed
  - Partially observed
- **Probability theory** is a rich language to express these uncertainties.
  - **Probabilistic models**
- Graphical tool to visualize complex models for complex problems.
- Complex models can be built from simpler parts.
- Computational tools to derive algorithmic solutions.
- Separation of modelling questions from algorithmic questions.

Probabilistic Modelling

- Data: $x_1, x_2, \ldots, x_n$.
- Latent variables: $y_1, y_2, \ldots, y_n$.
- Parameter: θ.
- A probabilistic model is a parametrized joint distribution over variables.

$$P(x_1, \ldots, x_n, y_1, \ldots, y_n | \theta)$$

- Typically interpreted as a **generative model** of data.
- Inference, of latent variables given observed data:

$$P(y_1, \ldots, y_n | x_1, \ldots, x_n, \theta) = \frac{P(x_1, \ldots, x_n, y_1, \ldots, y_n | \theta)}{P(x_1, \ldots, x_n | \theta)}$$

Probabilistic Modelling

- Learning, typically by maximum likelihood:

$$\theta^{\mathrm{ML}} = \operatorname*{argmax}_{\theta} P(x_1, \ldots, x_n | \theta)$$

- Prediction:

$$P(x_{n+1}, y_{n+1} | x_1, \ldots, x_n, \theta)$$

- Classification:

$$\operatorname*{argmax}_{c} P(x_{n+1} | \theta^c)$$

- Visualization, interpretation, summarization.
- Standard algorithms: EM, junction tree, variational inference, MCMC...

Bayesian Modelling

- Prior distribution:

$$P(\theta)$$

- Posterior distribution (both inference and learning):

$$P(y_1, \ldots, y_n, \theta | x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n, y_1, \ldots, y_n | \theta) P(\theta)}{P(x_1, \ldots, x_n)}$$

- Prediction:

$$P(x_{n+1} | x_1, \ldots, x_n) = \int P(x_{n+1} | \theta) P(\theta | x_1, \ldots, x_n) d\theta$$

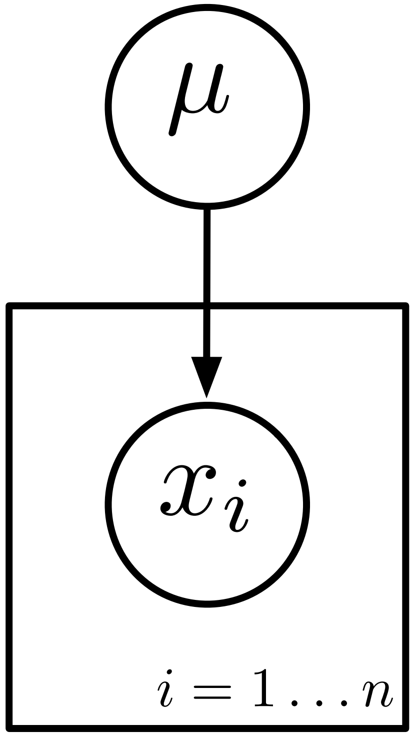
- Classification:

$$P(x_{n+1} | x_1^c, \ldots, x_n^c) = \int P(x_{n+1} | \theta^c) P(\theta^c | x_1^c, \ldots, x_n^c) d\theta^c$$

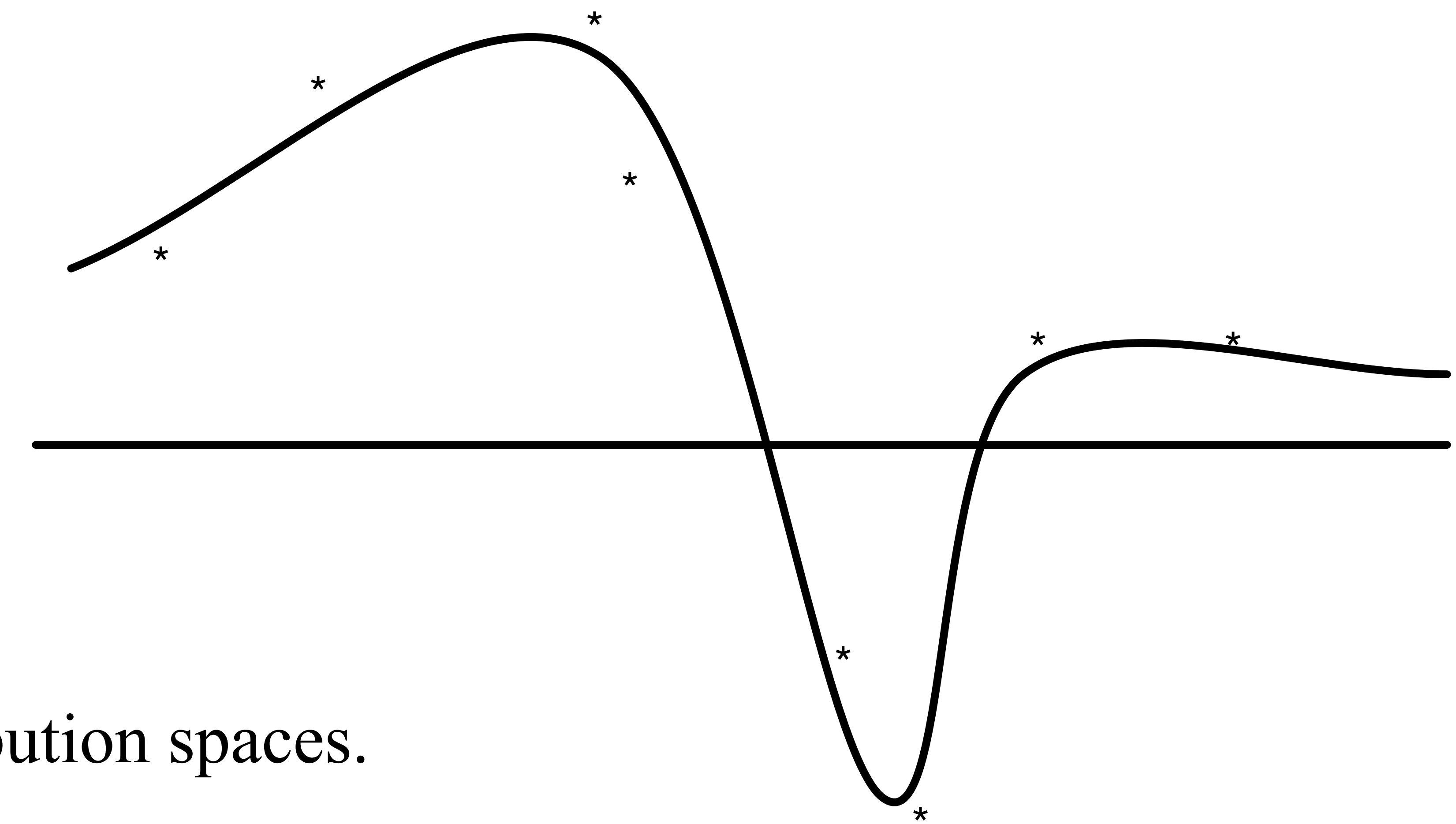
Bayesian Nonparametrics

Nonparametric Statistical Inference

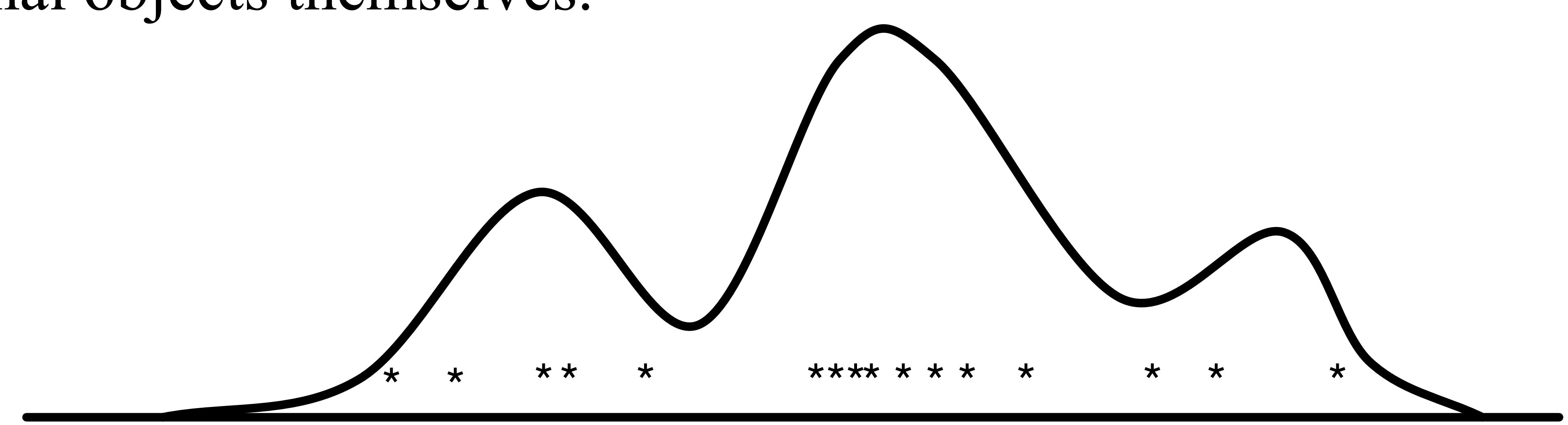
- Draw inferences without making overly restrictive assumptions about underlying distribution.
  - What is $E_\mu[f]$?
  - What is the q'th quantile of μ?
  - Given two distributions μ, v, are they the same?
  - Given two distributions, $X \sim \mu$, $Y \sim v$, is $P(X>Y)>.5$?



Large Function/Distribution Spaces



- Large function/distribution spaces.
- More straightforward to infer the infinite-dimensional objects themselves.



Novel and Useful Properties

- Many interesting Bayesian nonparametric models with interesting and useful properties:
  - Projectivity, exchangeability.
  - Zipf, Heap and other power laws

    (Pitman-Yor, 3-parameter IBP).
  - Flexible ways of building complex models

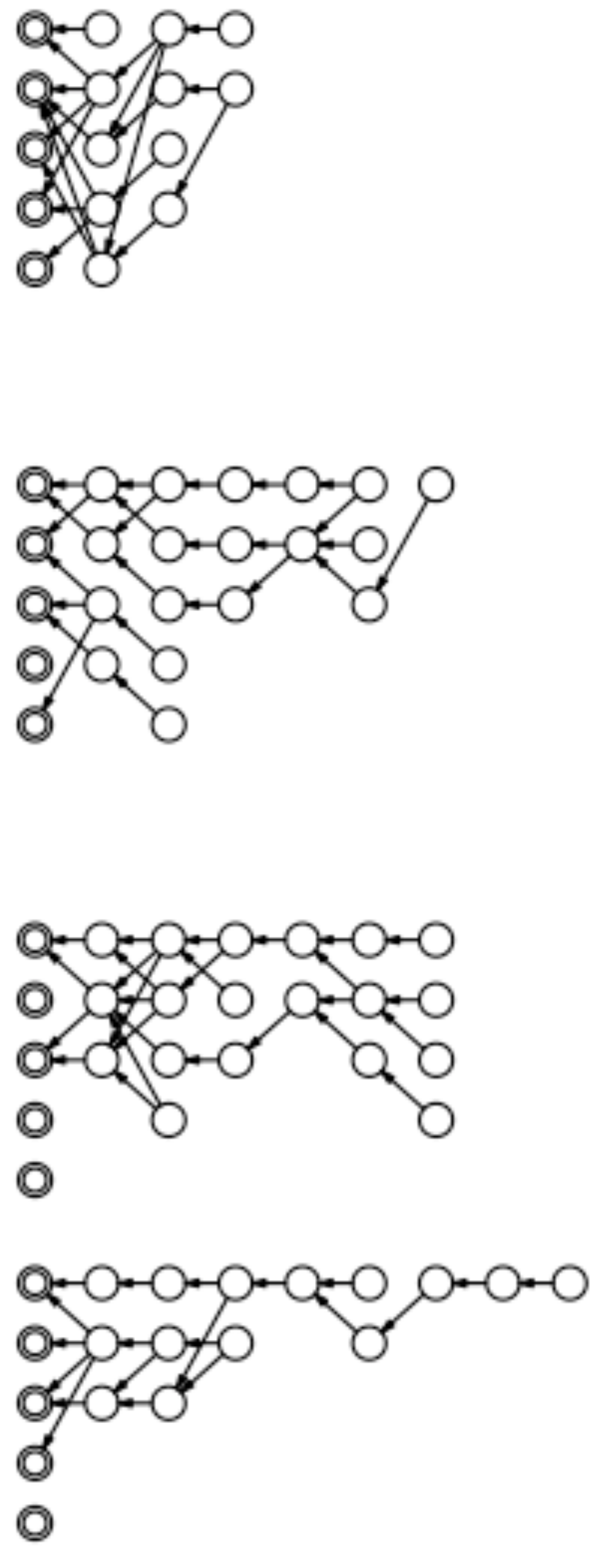
    (Hierarchical nonparametric models, dependent Dirichlet processes).

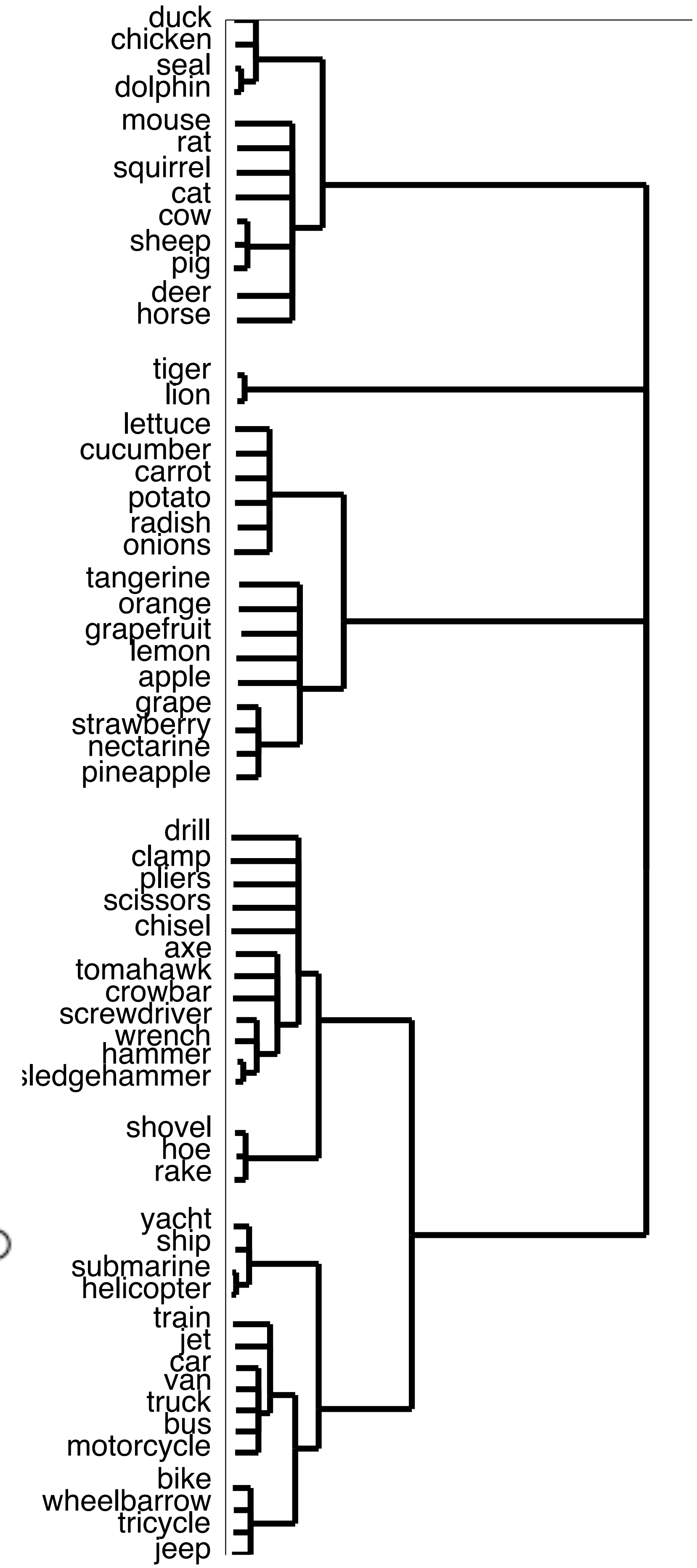
Model Selection and Averaging

- Model selection/averaging typically very expensive computationally.
- Used to prevent overfitting and underfitting.
- But a well-specified Bayesian model should not overfit anyway.
- By using a very large Bayesian model or one that grows with amount of data, we will not underfit either.

Structural Learning

- Learning structures.
- Bayesian prior over combinatorial structures.
- Nonparametric priors sometimes end up simpler than parametric priors.



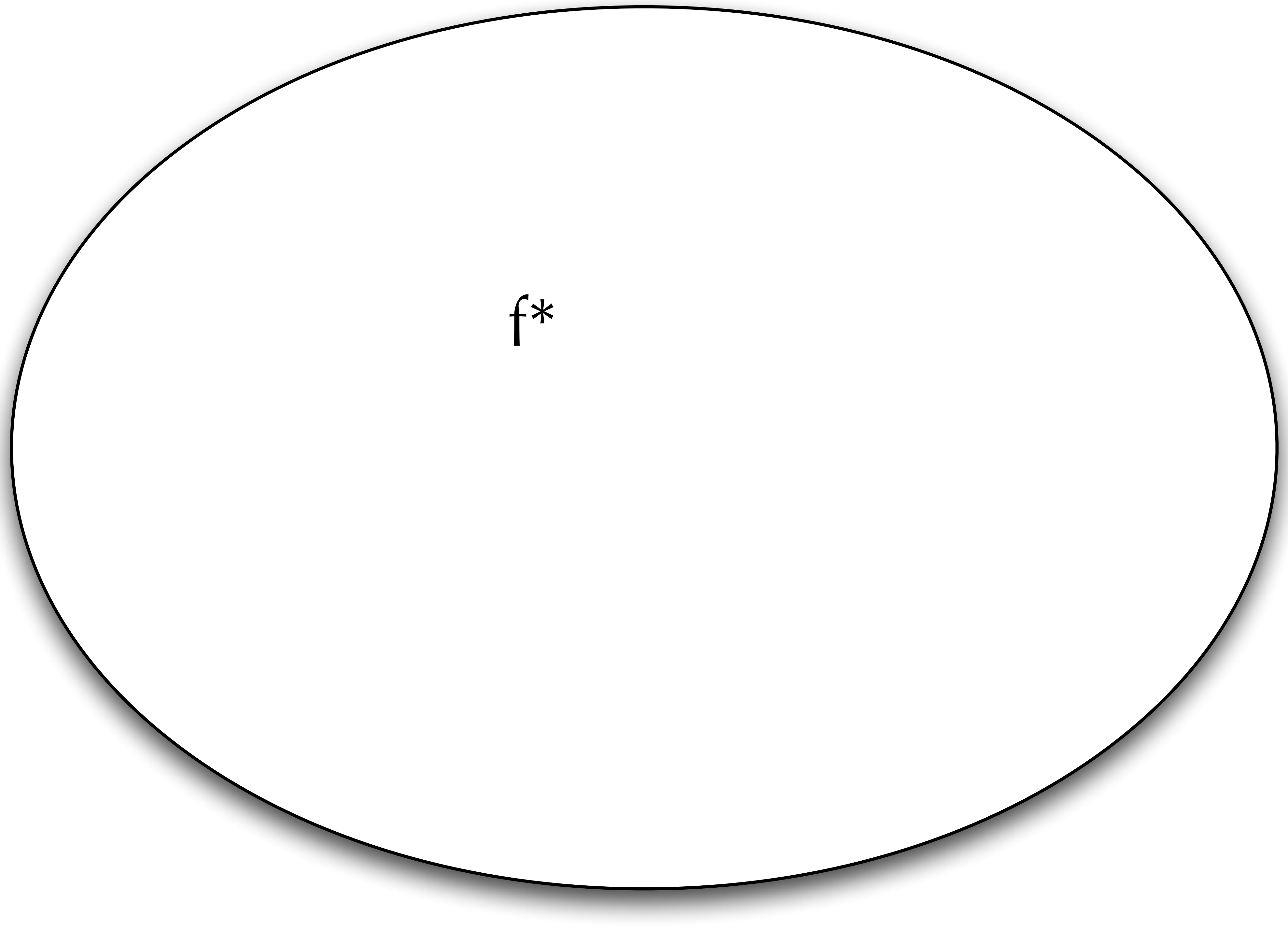


[Adams et al 2010, Blundell et al 2010]

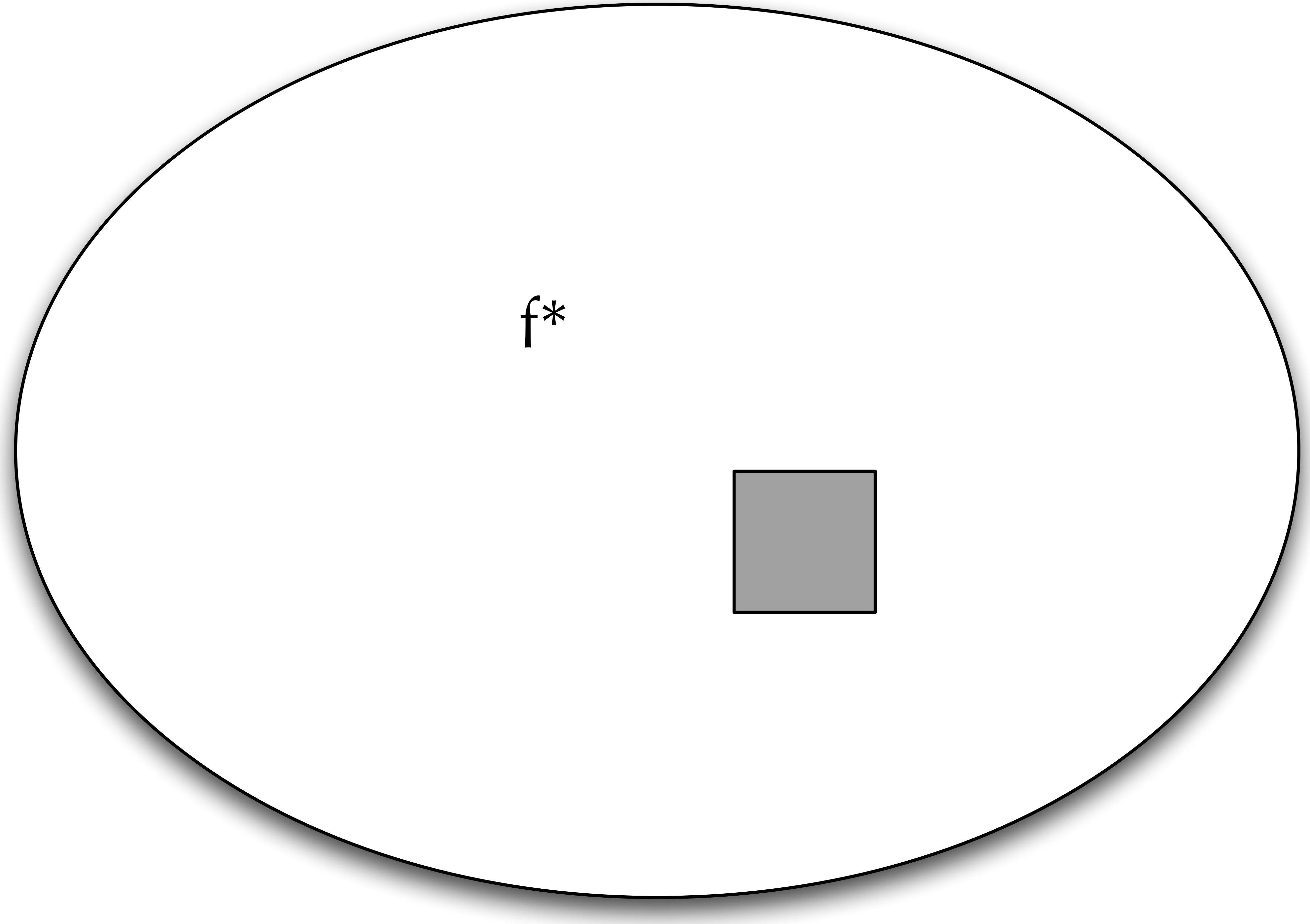
Bayesian Nonparametrics

- What is Bayesian nonparametrics?
  - Coverage: Bayesian modelling over large families of distributions.
  - Rich prior: Prior assumptions made explicit. Flexible framework allowing for rich structures in prior.

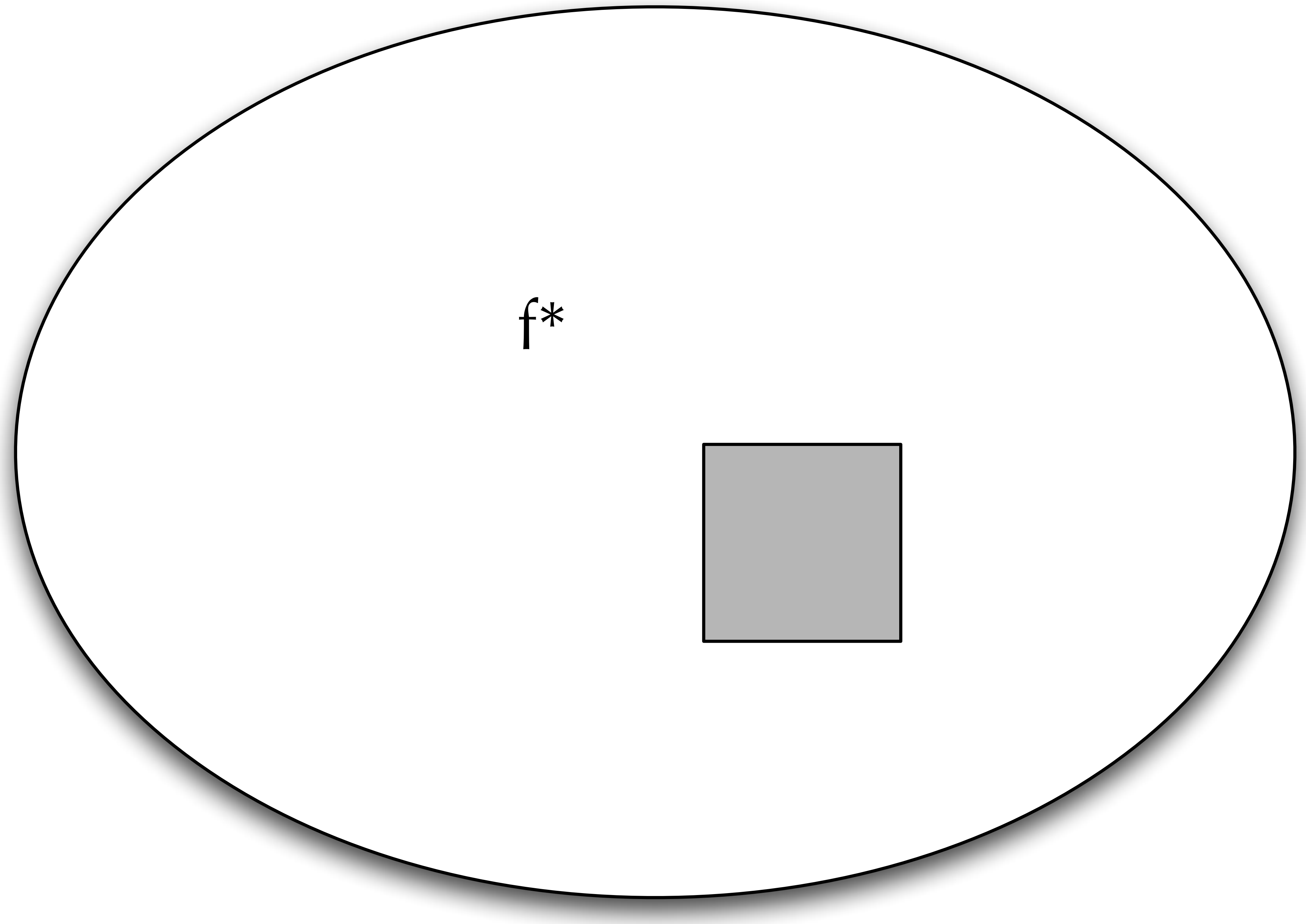
Bayesian Nonparametrics



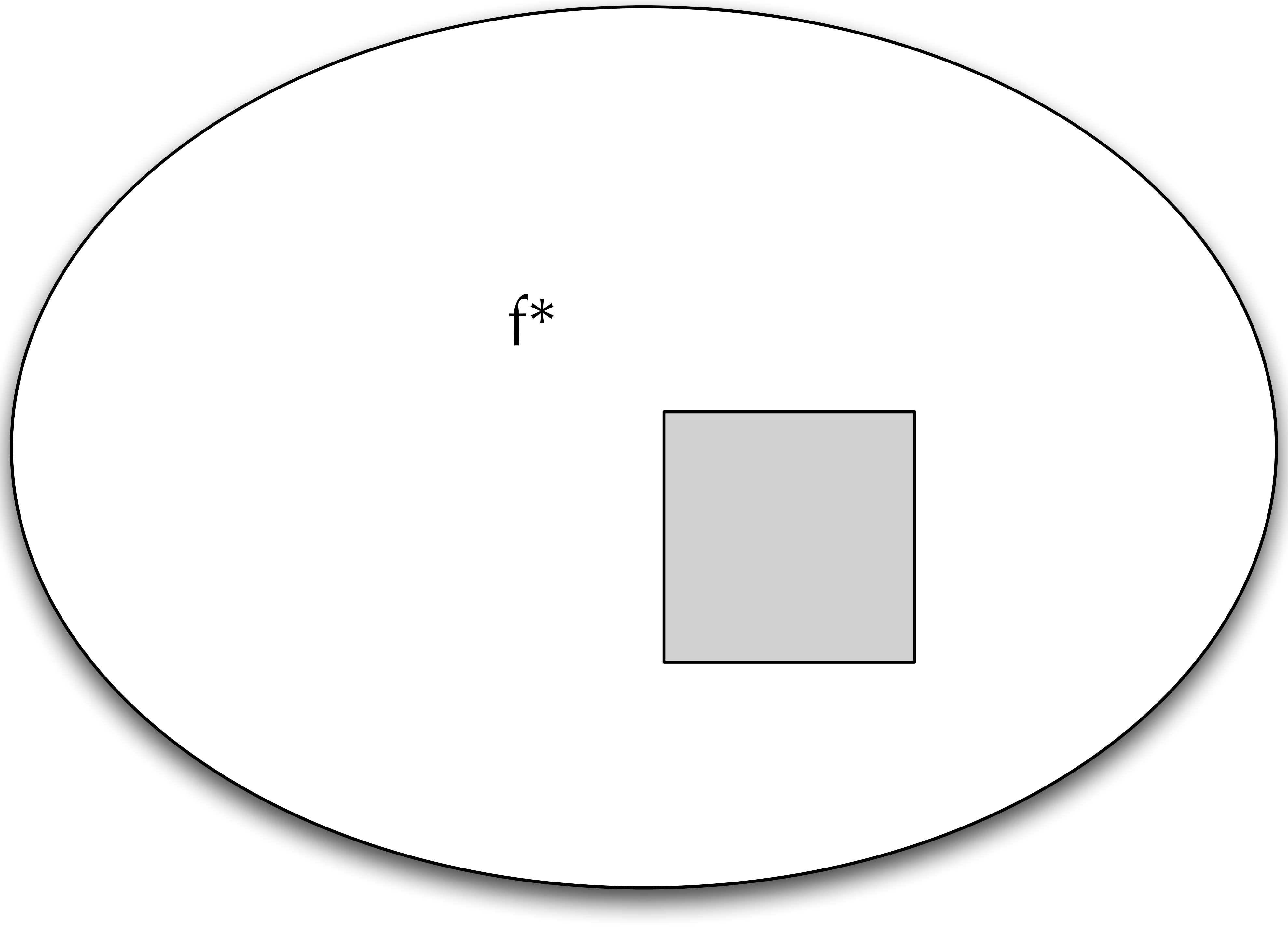
Bayesian Nonparametrics



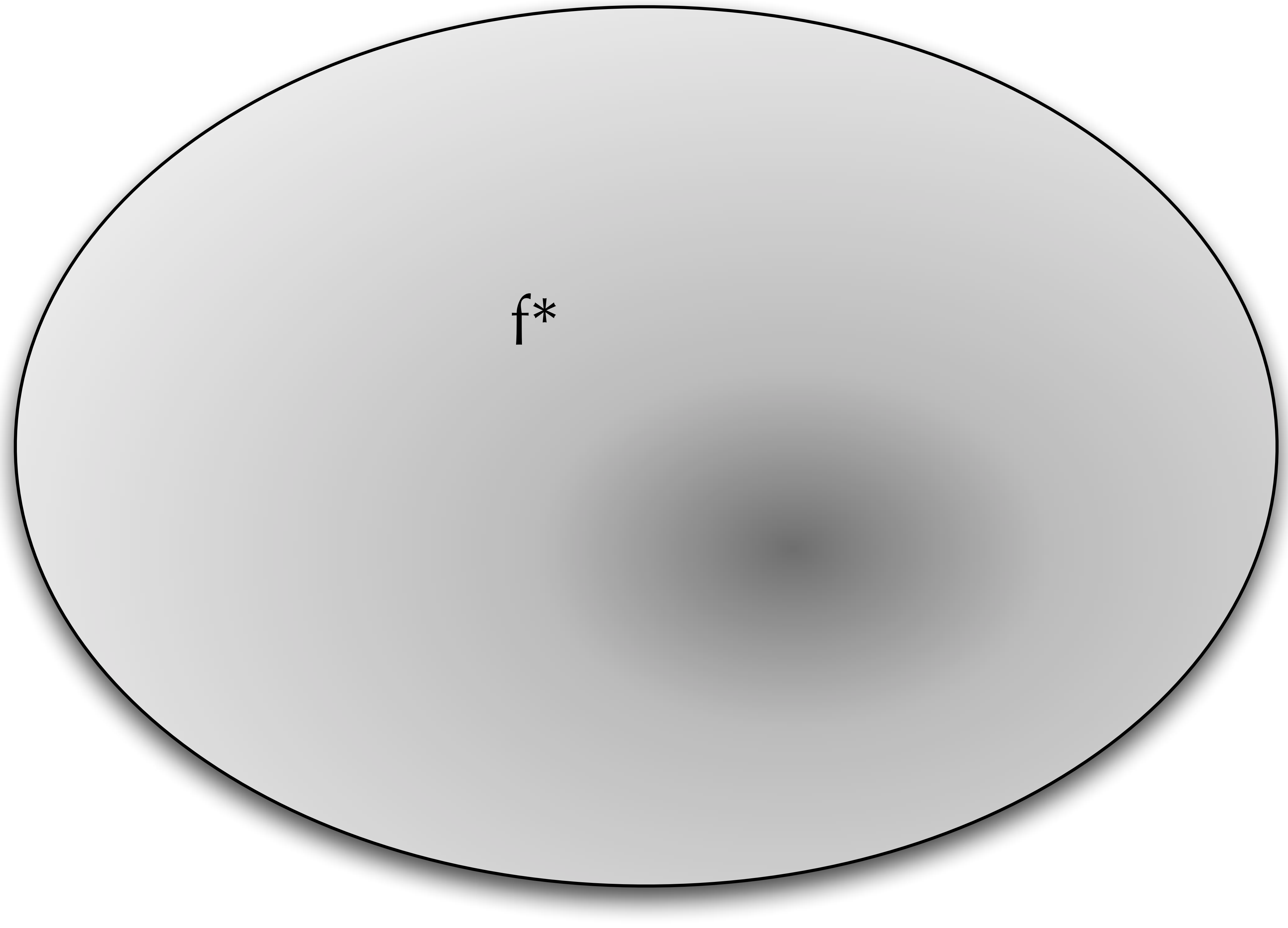
Bayesian Nonparametrics



Bayesian Nonparametrics



Bayesian Nonparametrics



Are Nonparametric Models Nonparametric?

- Nonparametric just means *not parametric*: *cannot be described by a fixed set of parameters*.
  - Nonparametric models still have parameters, they just have an infinite number of them.
- No free lunch: *cannot learn from data unless you make assumptions*.
  - Nonparametric models still make modelling assumptions, they are just less constrained than the typical parametric models.
- Models can be nonparametric in one sense and parametric in another: **semiparametric** models.

Issues with Bayesian Nonparametrics

- Modelling:
  - Classes of nonparametric priors suitable for modelling data.
- Algorithms:
  - Efficiently compute the posterior.
- Theory:
  - Asymptotic and finite sample guarantees for Bayesian nonparametrics.

Previous Tutorials and Reviews

- Tutorials: Mike Jordan NIPS 2005, Zoubin Ghahramani UAI 2005, Peter Orbanz MLSS 2009, Teh MLSS 2007, 2009, 2011, Orbanz & Teh NIPS 2011.
- Introduction to Dirichlet process [Teh 2010], nonparametric Bayes [Orbanz & Teh 2010, Gershman & Blei 2011], hierarchical Bayesian nonparametric models [Teh & Jordan 2010].
- Bayesian nonparametrics book [Hjort et al 2010].

Tiny Bit of Probability Theory

- A σ-**algebra** Σ is a family of subsets of a set Θ such that
  - Σ is not empty;
  - if $A \in \Sigma$ then $\Theta \backslash A \in \Sigma$;
  - if $A_1, A_2, \ldots \in \Sigma$ then $\cup_i A_i \in \Sigma$.
- (Θ, Σ) is a **measure space** and $A \in \Sigma$ are the **measurable sets**.
- A **measure** μ over (Θ, Σ) is a function $\mu : \Sigma \rightarrow [0, \infty]$ such that
  - $\mu(\emptyset) = 0$;
  - if $A_1, A_2, \ldots \in \Sigma$ are disjoint then $\mu(\cup_i A_i) = \Sigma_i \, \mu(A_i)$;
  - a **probability measure** is one where $\mu(\Theta) = 1$.
- Everything we consider here will be measurable.

Tiny Bit of Probability Theory

- Given two measure spaces (Θ, Σ) and (Δ, Φ) a function $f : \Theta \rightarrow \Delta$ is **measurable** if $f^{-1}(A) \in \Sigma$ for every $A \in \Phi$.
- If P is a probability measure on (Θ, Σ), a **random variable** X taking values in Δ is simply a measurable function $X : \Theta \rightarrow \Delta$.
  - This of the probability space (Θ, Σ, P) as a black-box random number generator, and X as a fixed function taking random samples in Θ and producing random samples in Δ.
  - The probability of an event $A \in \Phi$ is $P(X \in A) = P(X^{-1}(A))$.
- A **stochastic process** is simply a collection of random variables $\{X_i\}_{i \in I}$ over the same measure space (Θ, Σ), where I is an index set.
  - I can be an infinite (even uncountably infinite) set.