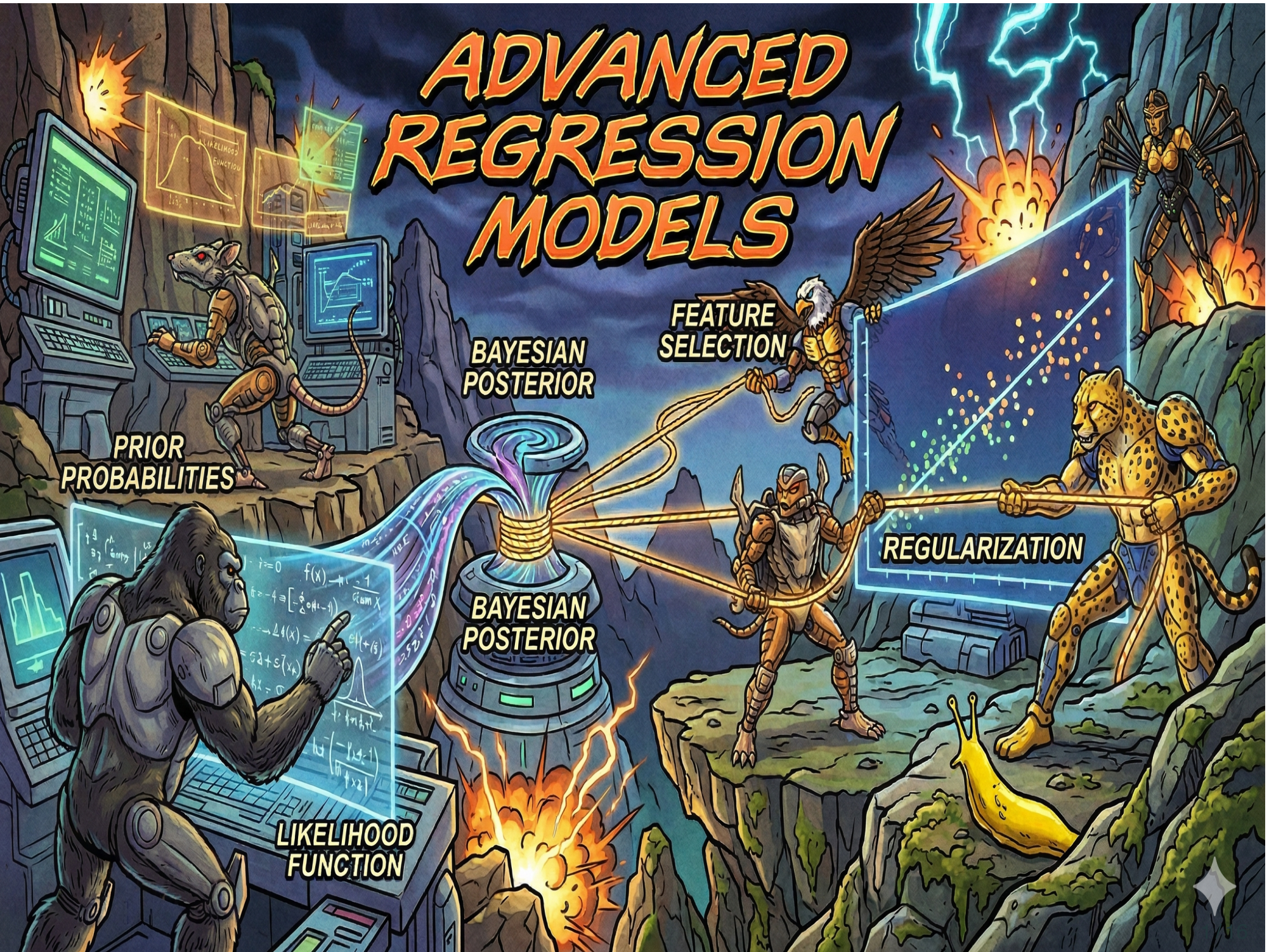

ADVANCED REGRESSION MODELS
PRIOR PROBABILITIES
BAYESIAN POSTERIOR
FEATURE SELECTION
REGULARIZATION
BAYESIAN POSTERIOR
LIKELIHOOD FUNCTION

Data-dependent priors

- When we use (weakly) informative priors, e.g., $\boldsymbol{\beta} \sim N(\mu_0, V_0)$, we are implicitly adding prior information for $\boldsymbol{x}_i^T \boldsymbol{\beta}$. This is usually reasonable in data analysis, as there is usually a reasonable range of $\boldsymbol{\beta}$ that the data supports.
- Sometimes, though, we might want to specify 'default priors' that do not require subjective input.
- An alternative type of weakly informative prior is the so-called 'unit information prior'

$$\boldsymbol{\beta} \mid \sigma^2 \sim N(\hat{\boldsymbol{\beta}}_{MLE}, (\frac{1}{n\sigma^2} \boldsymbol{X}^T \boldsymbol{X})^{-1})$$

and

$$\sigma^2 \sim \text{Inv-Gamma}(\frac{1}{2}, \frac{\hat{\sigma}^2_{MLE}}{2})$$

- Notice that the variance of $\hat{\beta}_{MLE}$ is $(\frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X})^{-1}$. So conceptually, this prior contains the same amount of information as a single observation from the data. However, it does require knowledge of both $\boldsymbol{X}$ and $\boldsymbol{y}$ to construct the prior.

Data-dependent priors

- Another principle for constructing a prior for $\boldsymbol{\beta}$ is based on the idea that the parameter inference should be invariant to changes in the scales of $\boldsymbol{X}$.
- That is, for a $p \times p$ matrix $\boldsymbol{H}$, if we transform $\boldsymbol{X}$ to $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{H}$ and estimate regression coefficients for $\tilde{\boldsymbol{\beta}}$ from $\boldsymbol{y}$ and $\tilde{\boldsymbol{X}}$, the posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{H}\tilde{\boldsymbol{\beta}}$ should be the same because

$$\boldsymbol{X}\boldsymbol{H}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}$$

- A popular prior satisfying the invariant property is the Zellner's g-prior:

$$\boldsymbol{\beta} \mid \sigma^2 \sim N(\mathbf{0}, g\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

  where g is a free parameter. It can be specified or inferred by putting a prior on it. The unit information prior correspond to $g = n$.
- The posterior conditional is

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y} \sim N(\frac{g}{g+1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \frac{g}{g+1}\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

- Another motivation for Zellner's g-prior is that it leads to a closed form of Bayes factor computation for model selection. See Liang et al (2008) "Mixtures of g Priors for Bayesian Variable Selection" for more details.

Conditional v.s independent priors

- In these two examples, we define priors for $\boldsymbol{\beta} \mid \sigma^2$. In general, for almost all normal priors (or scale mixture of normal) of $\boldsymbol{\beta}$, we can model the priors with

$$\boldsymbol{\beta} \mid \sigma^2 \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{V})$$

  for some covariance matrix $\boldsymbol{V}$.
- This is similar to the conjugate normal-inverse-gamma prior v.s. conditionally conjugate normal and inverse gamma priors for simple normal models.
- For low-dimensional problems, usually the prior choices do not matter much, similar to before.
- Moran et al. (2018) "Variance prior forms for high-dimensional Bayesian variable selection" explores the differences of the two approaches in high dimensions and recommend independent priors, i.e., $\boldsymbol{\beta}$'s prior not dependent on σ^2 a priori.

Spike and slab prior

- When there are too many predictors, often we are interested in model selection. Consider the following priors for $\boldsymbol{\beta}$.

$$\beta_j|z_j, \sim (1-z_j)N(0,\tau_0^2) + z_j N(0,\tau_1^2)$$
$$z_j \sim Bern(\pi)$$

  with $\tau_1^2 \gg \tau_0^2 > 0$.
- This is called the spike-and-slab prior, where the first mixture component correspond to a distribution that spikes at 0 and the second component has a wider slab.
- A different parameterization can be used to let $\tau_1^2 = c\tau_0^2$.
- Posterior inference only requires a simple modification to the standard linear regression sampler:

$$z_j|\beta_j, \sigma^2 \sim Bern(\frac{\pi N(\beta_j; 0, \tau_1^2)}{\pi N(\beta_j; 0, \tau_1^2) + (1-\pi)N(\beta_j; 0, \tau_0^2)})$$

- Unfortunately, the posterior of $\boldsymbol{\beta}$ again does not have point mass at 0. But we can obtain the posterior distribution of the inclusion indicator z_j and compute $Pr(z_j = 1|\boldsymbol{y})$ for each variable. For example, we can select any covariates with $Pr(z_j = 1|\boldsymbol{y}) > 0.5$, which is usually called the median probability model.

Spike and slab prior

- The previous spike-and-slab prior is sometimes referred to as continuous spike-and-slab prior, to differentiate from the discrete alternative

$$\beta_j|z_j, \sim (1-z_j)\delta_0(\beta_j) + z_j N(0, \tau_1^2)$$

  where δ_0 is a point mass function at 0.
- Previous Gibbs sampler cannot proceed in this case, even though it is the limiting case of $\tau_1 \to 0$, as whenever $z_j = 0$ at some iteration, β_j will be stuck at 0.
- Instead, we can integrate out $\boldsymbol{\beta}$ and σ^2 and sample

$$P(z_j = 1|\mathbf{z}_{-j}, \mathbf{y}) = \frac{p(\mathbf{y}|z_j = 1, \mathbf{z}_{-j})\pi}{p(\mathbf{y}|z_j = 1, \mathbf{z}_{-j})\pi + p(\mathbf{y}|z_j = 0, \mathbf{z}_{-j})(1-\pi)}$$

  where $p(\mathbf{y}|\boldsymbol{z})$ has closed form if conjugate normal inverse gamma prior is used for $(\boldsymbol{\beta}, \sigma^2)$.
- Metropolis-Hasting can also be used where at each iteration, randomly perform one of the two transition moves:
  - (Adding or deleting) Randomly sample j from all $1, 2, ..., p$, and set z_j to $1 - z_j$.
  - (Swapping) Independently draw one 0 and one 1 in $\boldsymbol{z}$ and swap their values.

  Then accept the new $\boldsymbol{z}$ with probability $\min\left(1, \frac{p(\boldsymbol{z}^{new}|\boldsymbol{y})}{p(\boldsymbol{z}^{old}|\boldsymbol{y})}\right)$

Shrinkage methods in regression

- A closely related concept in regression is shrinkage.
- In the least squares fitting procedure, we find $\hat{\boldsymbol{\beta}}$ that minimizes

$$RSS = \sum_i (y_i - \beta_0 - \sum_j \beta_j x_{ij})^2$$

- Ridge regression is a very similar procedure that minimizes

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

  were the additional term is called a shrinkage penalty.
- The penalty is small when β_j are close to zero. So it shrinks the estimates of β_j to 0. $\lambda > 0$ is a tuning parameter that controls the relative impact of the penalty.
- The ridge regression estimator is the same as the posterior mode for the Bayesian regression model with

$$\beta_j | \sigma^2 \sim N(0, \frac{\sigma^2}{\lambda})$$

  and $p(\beta_0) \propto 1$.

Bayesian ridge regression

- The equivalence can be seen from the posterior distribution

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}) \propto (\frac{1}{\sigma})^n (\frac{\sqrt{\lambda}}{\sigma})^p \exp(\frac{1}{2\sigma^2} \sum_i (y_i - \beta_0 - \sum_j \beta_j x_{ij})^2 - \frac{\lambda}{2\sigma^2} \sum_{j=1}^{p} \beta_j^2)$$

- To make notation easier, ignore the intercept in the rest of this slide (we can center all columns of $\boldsymbol{X}$ and also the $\boldsymbol{y}$).
- The posterior mean (or mode) as we have seen before is

$$\mathbb{E}(\boldsymbol{\beta} | \sigma^2, \boldsymbol{y}) = (\frac{\lambda}{\sigma^2} \boldsymbol{I} + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X})^{-1} \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{y} = (\lambda \boldsymbol{I} + \boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

- In the very special case where $\boldsymbol{X}^T \boldsymbol{X} = \boldsymbol{I}$, the posterior mean becomes $\frac{1}{1+\lambda} \hat{\boldsymbol{\beta}}_{MLE}$.
- In the more general case, individual elements of $\mathbb{E}(\boldsymbol{\beta} | \sigma^2, \boldsymbol{y})$ can be larger than $\boldsymbol{\beta}_{MLE}$, but the collection of all coefficients undergoes shrinkage.
- Also notice that $(\lambda \boldsymbol{I} + \boldsymbol{X}^T \boldsymbol{X})$ is always invertible.

Lasso and the Bayesian lasso

- Another very popular shrinkage method is the lasso. It minimizes

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Unlike the ridge regression, lasso estimators are sparse. That is, it can be shown that some of the $\hat{\beta}_j$ are exactly zero. Thus it achieves shrinkage and model selection simultaneously.
- Computation of lasso solution is less obvious, but can be done efficiently with several optimization algorithms.
- From a Bayesian perspective, the lasso estimator is (almost) equivalent to the posterior mode under the Laplace or double exponential distribution,

$$p(\beta_j) = \frac{\lambda}{2} \exp(-\lambda|\beta_j|), \quad j = 1, ..., p$$

  It is exactly equivalent if the noise level σ^2 is assumed known.
- Side note: in frequentist optimization of lasso, the noise level is typically not a major consideration. See Sun and Zhang (2012) “Scaled Sparse Linear Regression” for an example where β and σ^2 are jointly optimized.

The Bayesian lasso

- The naive prior of $p(\beta_j) \propto \frac{\lambda}{2}\exp(-\lambda|\beta_j|)$ sometimes leads to multimodal posterior.
- A more commonly used lasso prior is

$$p(\beta_j|\sigma^2) = \frac{\lambda}{2\sigma}\exp(-\lambda|\beta_j|/\sigma), \quad j = 1, ..., p$$

- Still this is not exactly equivalent to lasso, as we are scaling by σ instead of σ^2. But there is no reason to be exactly equivalent after all. Scaling by σ is more reasonable as it is on the same scale as the linear predictor.
- The double exponential distribution can be written as a mixture of normals as

$$\frac{\lambda}{2}e^{-\lambda|z|} = \int_0^\infty \frac{e^{-z^2/(2s)}}{\sqrt{2\pi s}} \times \frac{\lambda^2}{2}e^{-\lambda^2 s/2}ds$$

- Thus $\beta_j|\sigma^2 \sim DE(\lambda/\sigma)$ is equivalent to

$$\beta_j|\sigma^2, \eta_j^2 \sim N(0, \sigma^2\eta_j^2)$$
$$\eta_j^2|\lambda \sim Exp(\lambda^2/2)$$

The Bayesian lasso

- We may also include the intercept specifically by letting

$$\boldsymbol{Y} \sim N(\mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$$

- The prior can then be completed with $p(\beta_0, \sigma^2) \propto \sigma^{-2}$.
- Let $\boldsymbol{Y}^* = \boldsymbol{Y} - \mathbf{1}_n\beta_0$ and $\mathbf{D} = \text{diag}(\eta_1^2, \ldots, \eta_p^2)$,

$$\boldsymbol{\beta}|\boldsymbol{Y}, \beta_0, \sigma^2, \boldsymbol{\eta}^2 \sim N\left((\boldsymbol{X}^T\boldsymbol{X} + \mathbf{D}^{-1})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*, \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \mathbf{D}^{-1})^{-1}\right)$$

$$\beta_0|\boldsymbol{Y}, \boldsymbol{\beta}, \sigma^2 \sim N\left(\bar{y} - \bar{\boldsymbol{x}}^T\boldsymbol{\beta}, \frac{\sigma^2}{n}\right)$$

$$\sigma^2|\boldsymbol{Y}, \beta_0, \boldsymbol{\beta}, \boldsymbol{\eta}^2 \sim \text{Inv-Gamma}\left(\frac{n+p}{2}, \frac{(\boldsymbol{Y}^* - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}^* - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T\mathbf{D}^{-1}\boldsymbol{\beta}}{2}\right)$$

- For the latent variable, it turns out with a change of variable,

$$\eta_j^{-2}|\beta_j, \sigma^2, \lambda \sim \text{Inverse-Gaussian}\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right)$$

- The posterior distribution of $\boldsymbol{\beta}$, however, is not sparse. In order to achieve model selection, you need to set a cutoff ϵ so that $|\beta_j|$ smaller than ϵ are considered to be 'not selected'.

Disadvantages of lasso

- From a prediction point of view, shrinkage in regression coefficients reduces the variance of the prediction.
- However, the classical lasso estimator has several disadvantages:
  - it cannot select more than n non-zero coefficients (if $p > n$).
  - When covariates are highly correlated, it tends to select only one from the group of correlated covariates, which can lead to higher prediction error.
  - Overshrinkage of large coefficients.
- These limitations are also inherited by the Bayesian lasso...

The case of non-linearity

- To wrap up our regression lecture, we note that non-linear / non-parametric regression models can be easily built up from linear model. Consider the basis function representation

$$y_i = \sum_h \beta_h b_h(\boldsymbol{x}_i) + \epsilon_i$$

  where for each $h = 1, ...,$, $b_h()$ is a fixed basis function that maps the p-dimensional input vector $\boldsymbol{x}_i$ to the real line.
- Usually we specify $b_h()$ on each dimensions of $\boldsymbol{x}$ separately.
- With a flexible collection of basis function, we can capture quite complicated relationship between $\boldsymbol{x}$ and y.
- The priors on β_h can further induce penalization/shrinkage to smoother curves to avoid overfitting.

Bayesian additive regression tree

- The basis function may be further made random. A very popular class of models is the Bayesian additive regression tree (BART) models.

$$y_i = \sum_{j=1}^{m} g(X; T_j, M_j) + \epsilon_i$$

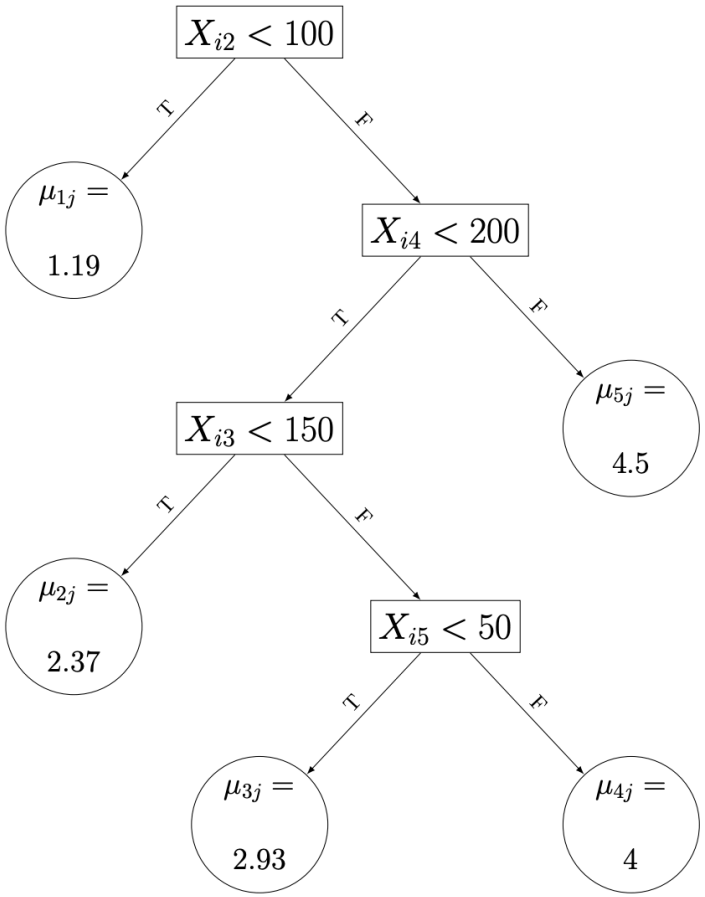
where $g(X; T_j, M_j)$ represent the j-th regression tree model. T_j is a binary tree structure and $M_j = \{\mu_{1j}, \mu_{2j}, ..., \}$ are the terminal node parameters associated with T_j. For example,

$$g(X; T_j, M_j) = \mu_{1j} \mathbf{1}_{X_{i2} < 100} + \mu_{2j} \mathbf{1}_{X_{i4} < 200}$$

is a regression tree with two nodes.

A single regression tree

Figure 1: Example of a regression tree $g(X; T_j, M_j)$ where μ_{kj} is the mean parameter of the k^{th} node for the j^{th} regression tree.



BART priors and posterior computation

- The BART model proposed in Chipman et al (2010) proceeds with a set of default priors that are quite robust for continuous $\boldsymbol{x}$.
- For $p(T_j)$, a shrinkage prior is used so that the probability of a node at depth d being the terminal node is $1 - \alpha(1+d)^{-\beta}$ where α and β are tuning parameters.
- The splitting variable for each node has a uniform prior on $\{1, ..., p\}$ and the splitting point has a uniform prior on all possible values for each x_j.
- Then a conjugate normal prior is used for μ_{ij} and inv-Gamma prior is used for σ^2.
- The number of trees m is typically fixed to a large enough number.
- Posterior inference is carried out by iteratively updating each tree using the residuals keeping the other trees fixed, i.e., let $\boldsymbol{R}_j = \{y_i - \sum_{j' \neq j} g(X; T_{j'}, M_{j'})\}_{i=1,...,n}$
  1. Given all but the j-th tree, sample $T_j | \boldsymbol{R}_j, \sigma^2$ using a M-H step.
     - the proposal can be simple: either remove a node or add a node.
     - this step integrates out M_j using conjugacy
  2. Given T_j and the rest of the trees, update $\mu_{\cdot,j} | T_j, \boldsymbol{R}_j, \sigma^2$, from the conjugate normal posterior.
  3. Given the trees and residuals, update σ^2 from the conjugate inverse Gamma posterior.