



ECE 5984: Introduction to Machine Learning

Topics:

- Gaussians
- (Linear) Regression

Readings: Barber 8.4, 17.1, 17.2

Dhruv Batra
Virginia Tech

Administrativa

- HW0
 - Solutions available
- HW1
 - Due on Sun 02/15, 11:55pm
 - <http://inclass.kaggle.com/c/VT-ECE-Machine-Learning-HW1>
- Project Proposal
 - Due: Tue 02/24, 11:55 pm
 - ≤ 2 pages, NIPS format



Recap of last time

Statistical Estimation

- Frequentist Tool
 - Maximum Likelihood
- Bayesian Tools
 - Maximum A Posteriori
 - Bayesian Estimation

MLE

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- $D_1 = \{1, 1, 1, 0, 0, 0\}$
- $D_2 = \{1, 0, 1, 0, 1, 0\}$
- A function of the data $\phi(Y)$ is a sufficient statistic, if the following is true

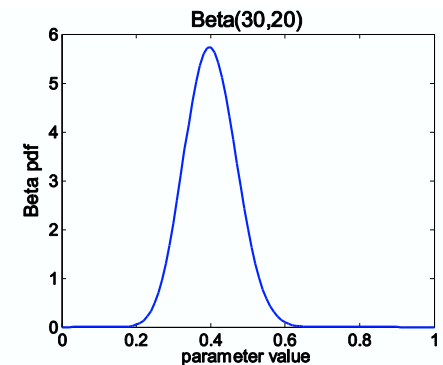
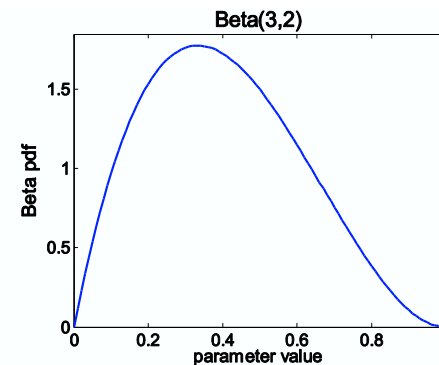
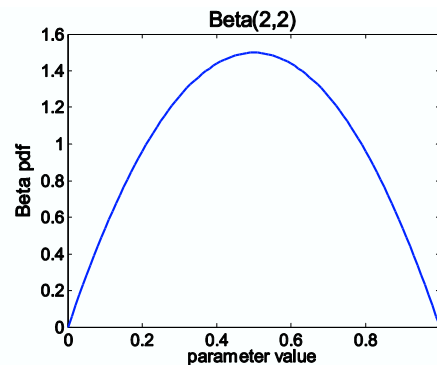
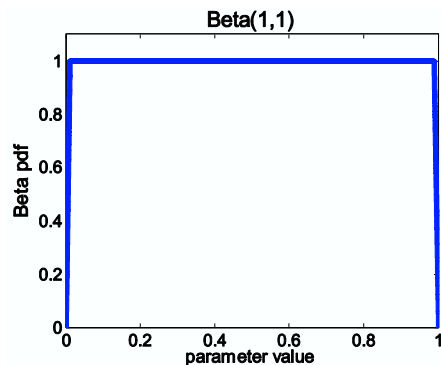
$$\sum_{i \in D_1} \phi(y_i) = \sum_{i \in D_2} \phi(y_i) \quad \Rightarrow \quad L(\theta; D_1) = L(\theta; D_2)$$

Beta prior distribution – $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_H - 1} (1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

- Demo:

- <http://demonstrations.wolfram.com/BetaDistribution/>



MAP for Beta distribution

$$P(\theta | \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

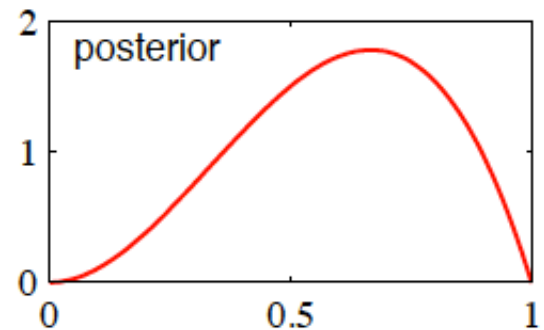
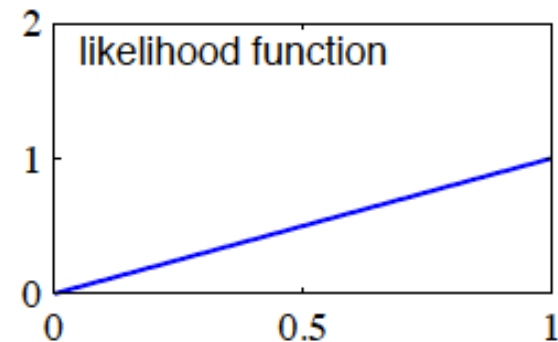
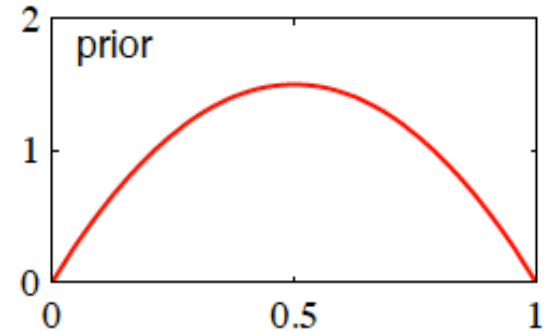
- MAP: use most likely parameter:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} P(\theta | \mathcal{D}) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

- Beta prior equivalent to extra W/L matches
- As $N \rightarrow \infty$, prior is “forgotten”
- **But, for small sample size, prior is important!**

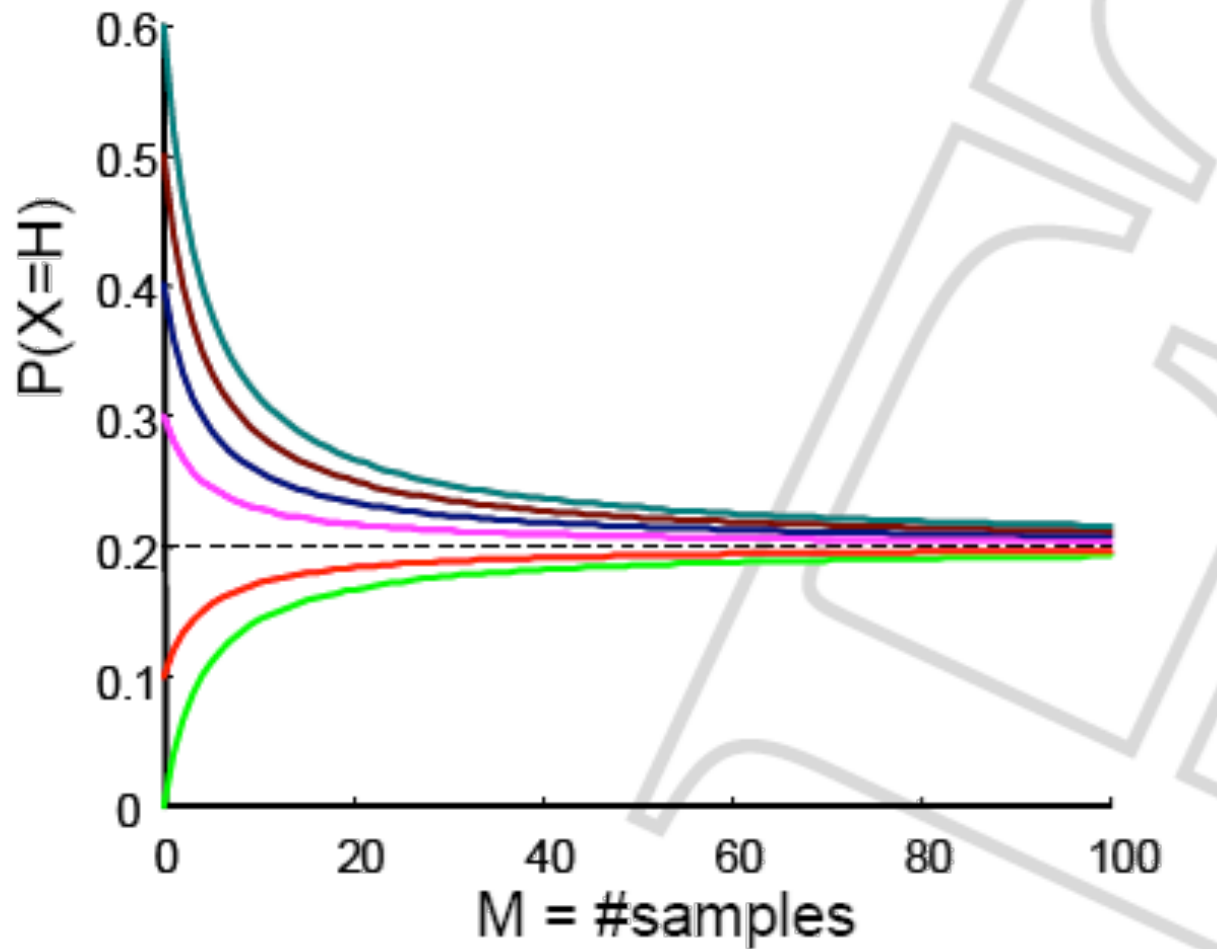
Effect of Prior

- Prior = Beta(2,2)
 - $\theta_{\text{prior}} = 0.5$
- Dataset = {H}
 - $L(\theta) = \theta$
 - $\theta_{\text{MLE}} = 1$
- Posterior = Beta(3,2)
 - $\theta_{\text{MAP}} = (3-1)/(3+2-2) = 2/3$



Effect of Prior

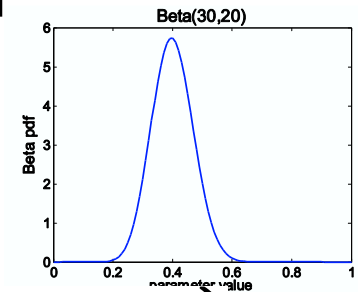
Starting from different priors



Using Bayesian posterior

- Posterior distribution:

$$P(\theta | \mathcal{D}) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



- Bayesian inference:
 - No longer single parameter:

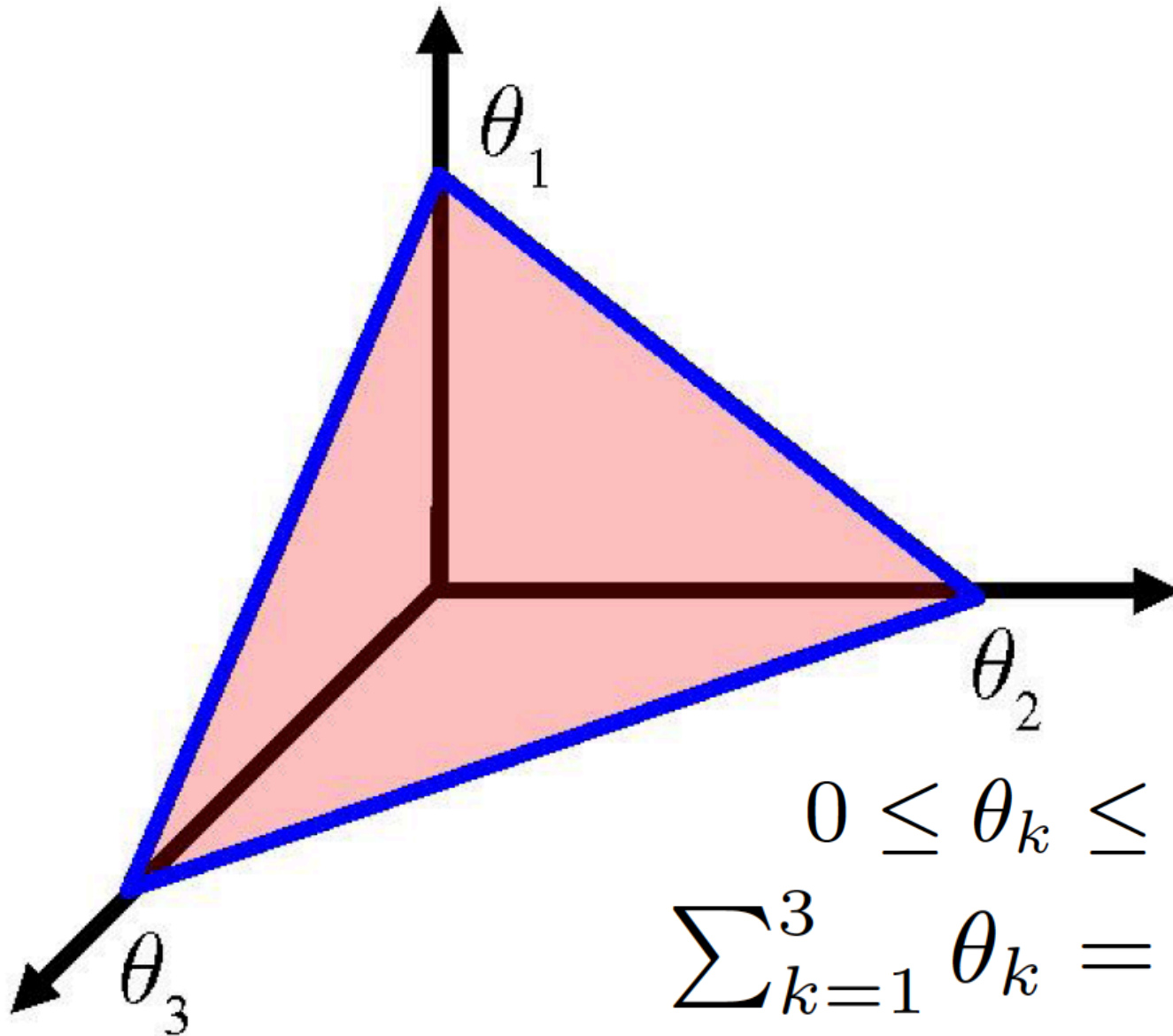
$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta | \mathcal{D}) d\theta$$

- Integral is often hard to compute

Bayesian learning for multinomial

- What if you have a k sided coin???
- Likelihood function if **categorical**:

Simplex



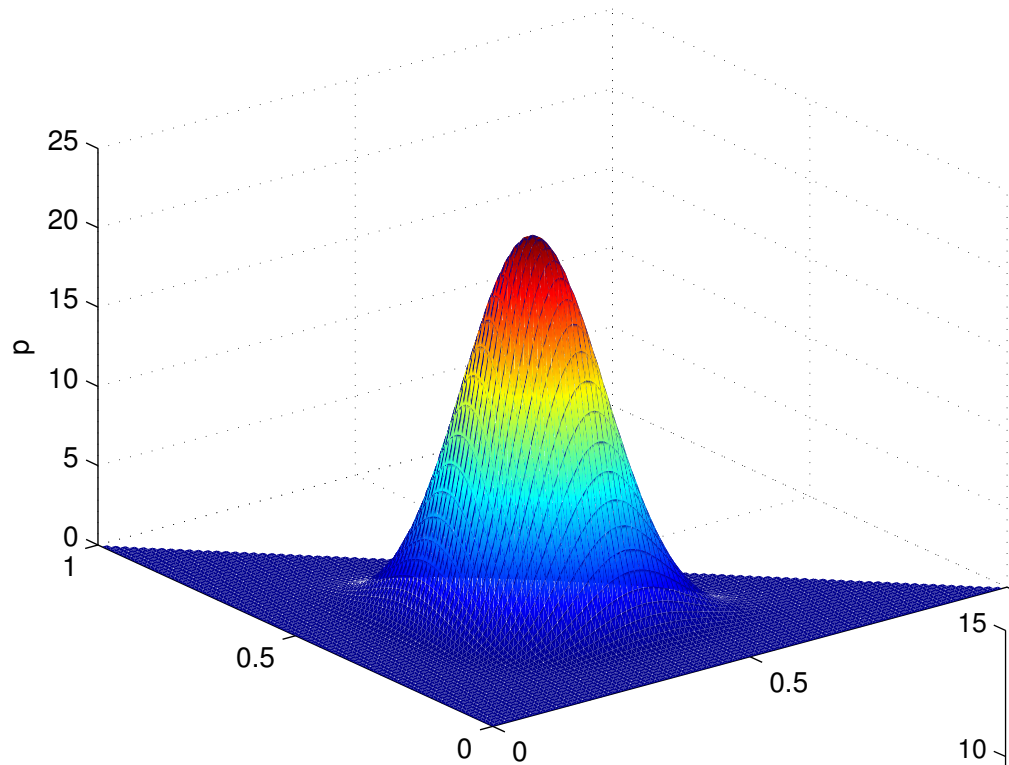
Bayesian learning for multinomial

- What if you have a k sided coin???
- Likelihood function if **categorical**:
- **Conjugate** prior for multinomial is **Dirichlet**:

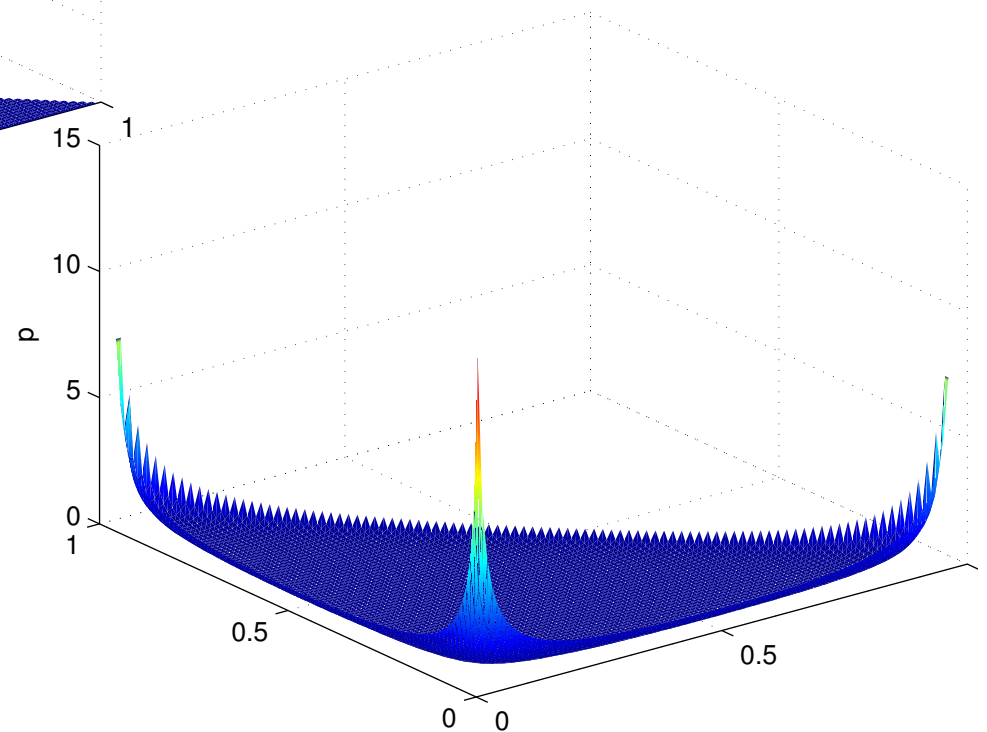
$$\theta \sim \text{Dirichlet}(\beta_1, \beta_2, \dots, \beta_k) \propto \prod_i \theta_i^{\beta_i - 1}$$

Dirichlet Probability Densities

$\alpha=10.00$



$\alpha=0.10$



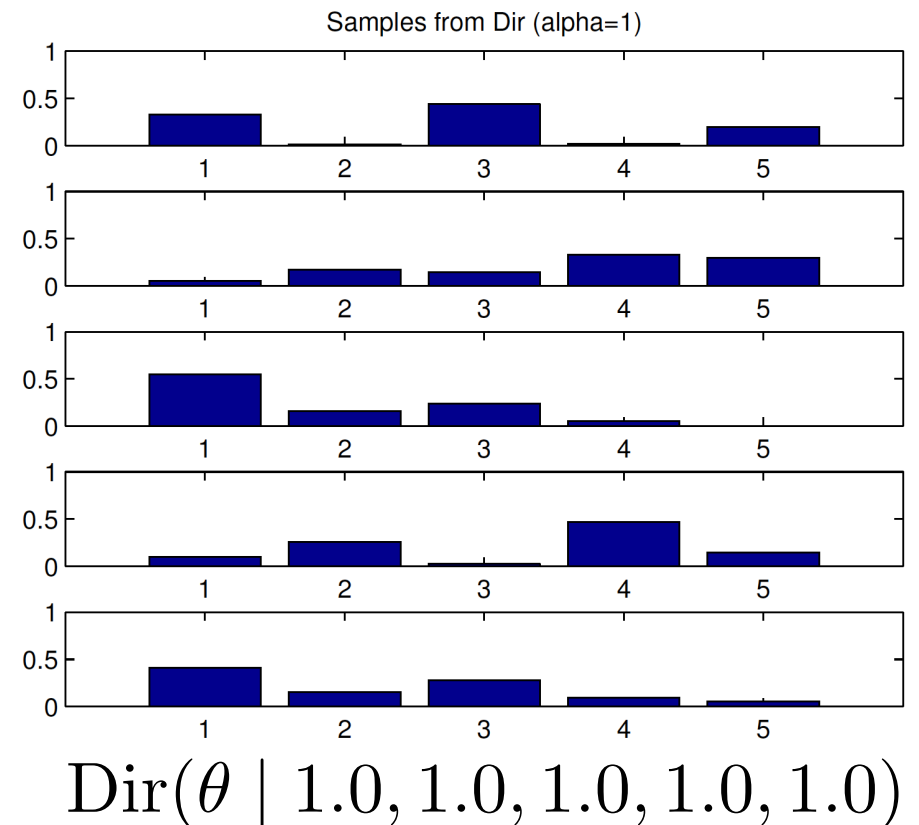
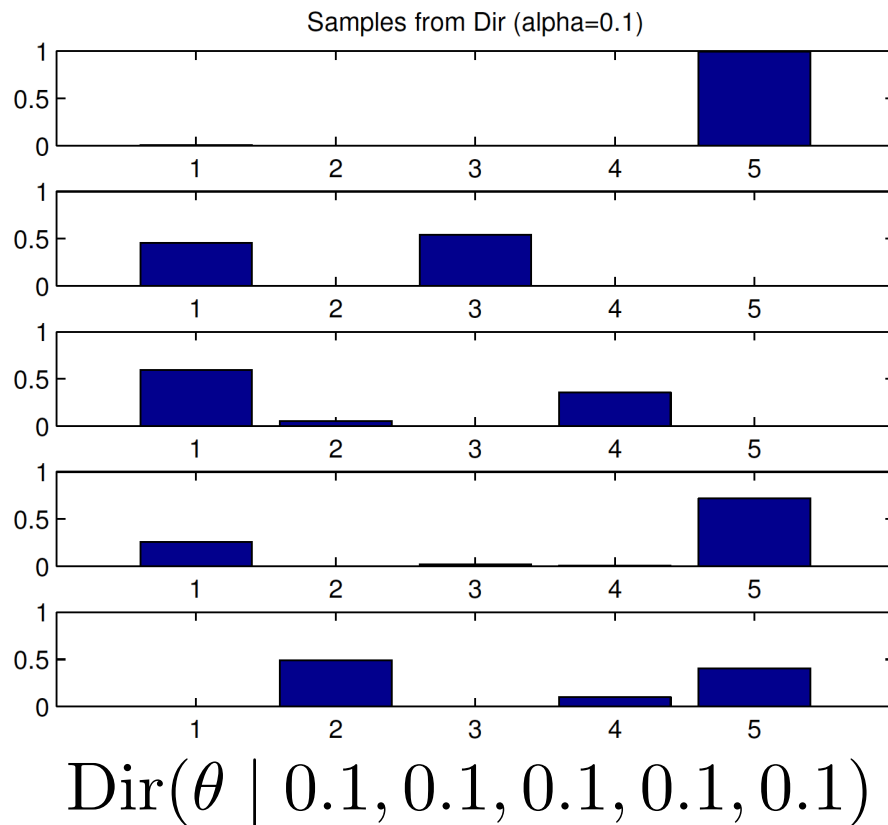
Mean: $\mathbb{E}[\theta_i] = \frac{\beta_i}{\sum_j \beta_j}$

Mode: $\hat{\theta}_i = \frac{\beta_i - 1}{\sum_j \beta_j - k}$

Dirichlet Probability Densities

- Matlab Demo
 - Written by Iyad Obeid

Dirichlet Samples



Bayesian learning for multinomial

- What if you have a k sided coin???
- Likelihood function if **categorical**:

- **Conjugate** prior for multinomial is **Dirichlet**:

$$\theta \sim \text{Dirichlet}(\beta_1, \beta_2, \dots, \beta_k) \propto \prod_i \theta_i^{\beta_i - 1}$$

- **Observe** n data points, n_i from assignment i , **posterior**:

• **Homework 1!!!!** 😊

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} P(\theta | \mathcal{D}) =$$

Plan for Today

- Gaussians
 - PDF
 - MLE/MAP estimation of mean
- Regression
 - Linear Regression
 - Connections with Gaussians



Gaussians

What about continuous variables?

- Boss says: If I want to bet on continuous variables, like stock prices, what can you do for me?
- **You say: Let me tell you about Gaussians...**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Why Gaussians?

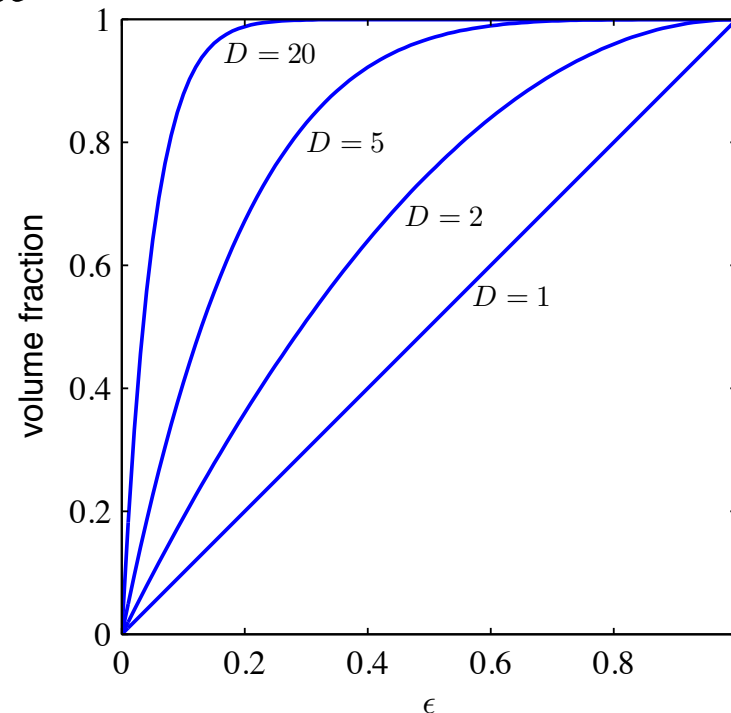
- Why does the entire world seem to always be telling you about Gaussian?
 - Central Limit Theorem!

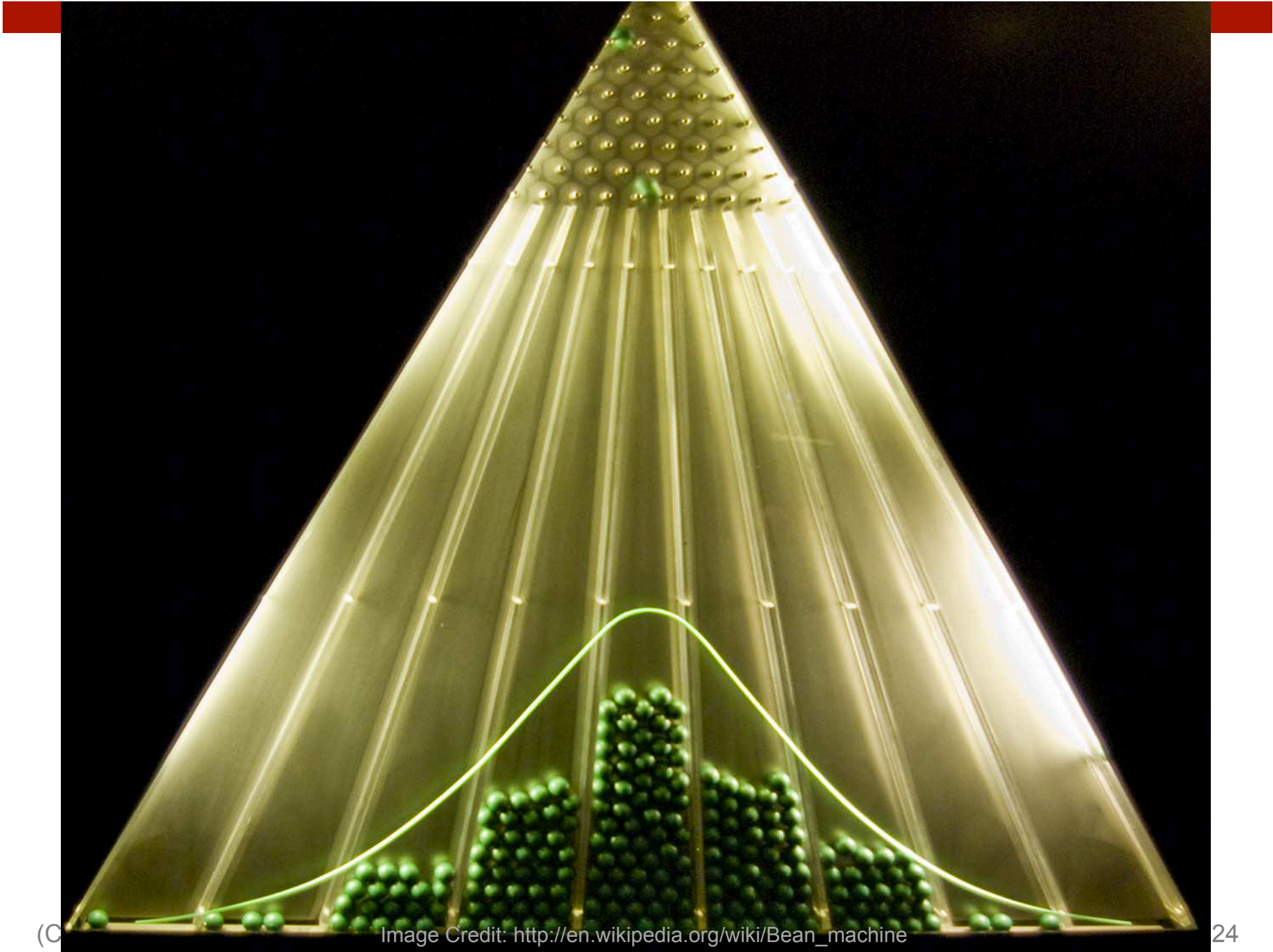
Central Limit Theorem

- Simplest Form
 - X_1, X_2, \dots, X_N are IID random variables
 - Mean μ , variance σ^2
 - Sample mean S_N approaches Gaussian for large N
- Demo
 - <http://www.stat.sc.edu/~west/javahtml/CLT.html>

Curse of Dimensionality

- Consider: Sphere of radius 1 in d-dims
- Consider: an outer ϵ -shell in this sphere
- What is $\frac{\textit{shell volume}}{\textit{sphere volume}}$?





(C

Image Credit: http://en.wikipedia.org/wiki/Bean_machine

Why Gaussians?

- Why does the entire world seem to always be harping on about Gaussians?
 - Central Limit Theorem!
 - They're easy (and we like easy)
 - Closely related to squared loss (will see in regression)
 - Mixture of Gaussians are sufficient to approximate many distributions (will see it clustering)

Some properties of Gaussians

- Affine transformation
 - multiplying by scalar and adding a constant
 - $X \sim N(\mu, \sigma^2)$
 - $Y = aX + b \quad \rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$
- Sum of Independent Gaussians
 - $X \sim N(\mu_X, \sigma_X^2)$
 - $Y \sim N(\mu_Y, \sigma_Y^2)$
 - $Z = X + Y \quad \rightarrow Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

Learning a Gaussian

- Collect a bunch of data
 - Hopefully, i.i.d. samples
 - e.g., exam scores
- Learn parameters
 - Mean
 - Variance

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1, \dots, x_N\}$:

$$P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^N \prod_{i=1}^N e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\begin{aligned} \ln P(\mathcal{D} \mid \mu, \sigma) &= \ln \left[\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^N \prod_{i=1}^N e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \right] \\ &= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2} \end{aligned}$$

Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

$$\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) = \frac{d}{d\mu} \left[-N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

MLE for variance

- Again, set derivative to zero:

$$\begin{aligned}\frac{d}{d\sigma} \ln P(\mathcal{D} \mid \mu, \sigma) &= \frac{d}{d\sigma} \left[-N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2} \right] \\ &= \frac{d}{d\sigma} \left[-N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^N \frac{d}{d\sigma} \left[\frac{(x_i - \mu)^2}{2\sigma^2} \right]\end{aligned}$$

Learning Gaussian parameters

- MLE:

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\hat{\sigma}_{MLE}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

Bayesian learning of Gaussian parameters

- Conjugate priors
 - Mean: Gaussian prior
 - Variance: Inverse Gamma or Wishart Distribution
- Prior for mean:

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{(\mu-\eta)^2}{2\lambda^2}}$$

MAP for mean of Gaussian

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{(\mu-\eta)^2}{2\lambda^2}} \quad P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^N e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

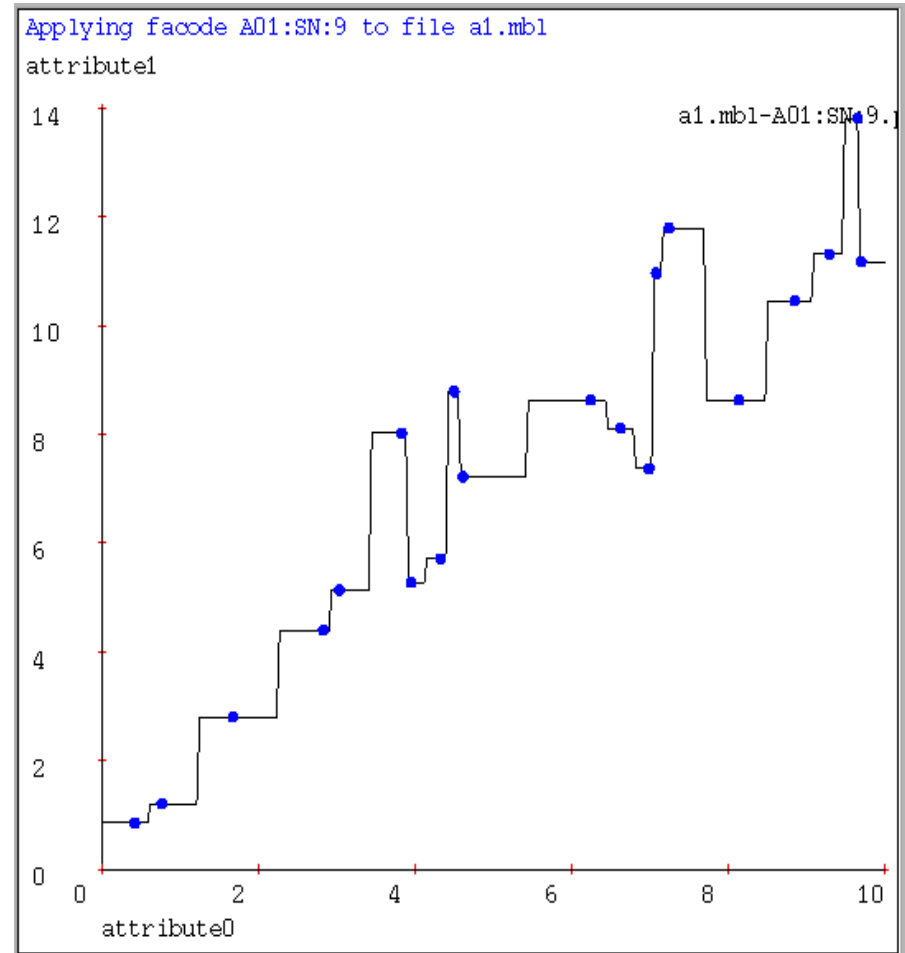
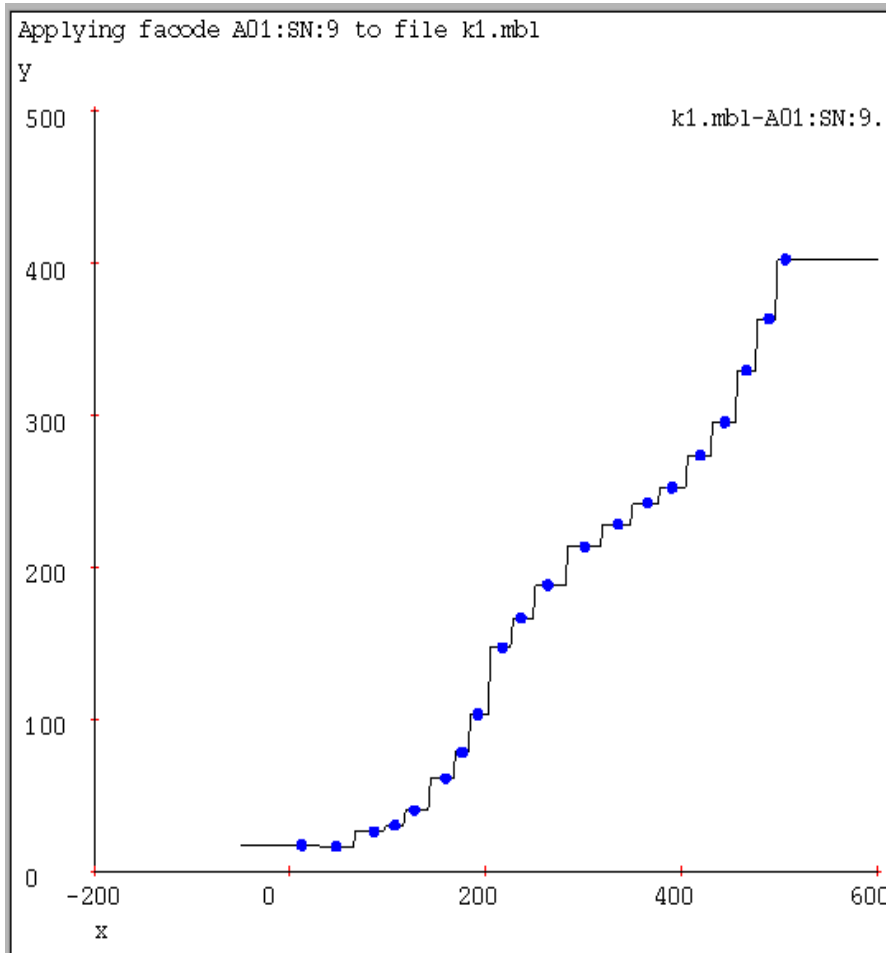
$$\frac{d}{d\mu} [\ln P(\mathcal{D} \mid \mu)P(\mu)] = \frac{d}{d\mu} [\ln P(\mathcal{D} \mid \mu) + \ln P(\mu)]$$



New Topic: Regression

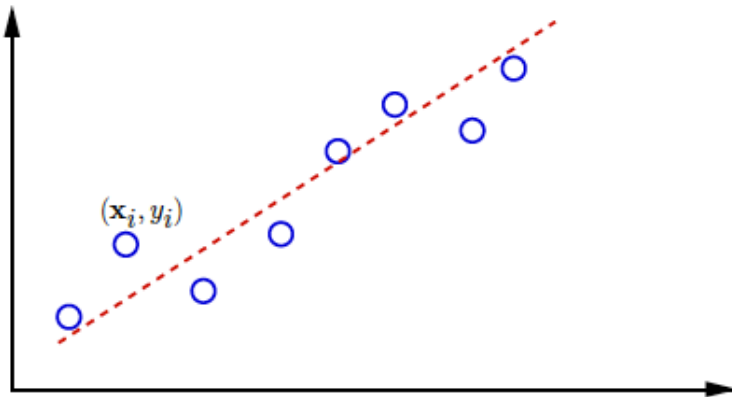
1-NN for Regression

- Often bumpy (overfits)



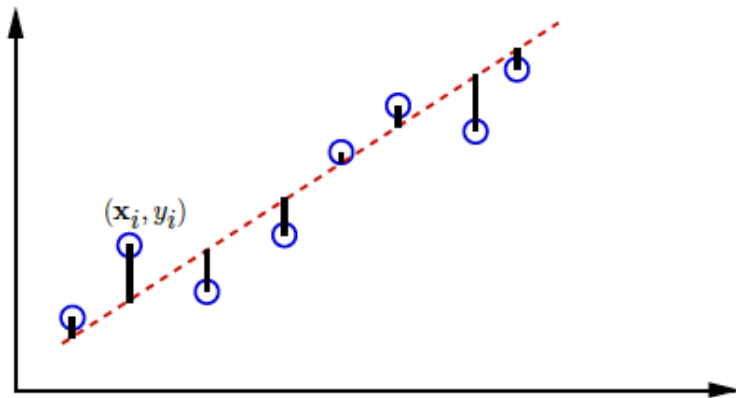
Linear fitting to data

- We want to fit a linear function to an observed set of points $X = [\mathbf{x}_1, \dots, \mathbf{x}_N]$ with associated labels $Y = [y_1, \dots, y_N]$.
 - Once we fit the function, we want to use it to *predict* the y for new \mathbf{x} .



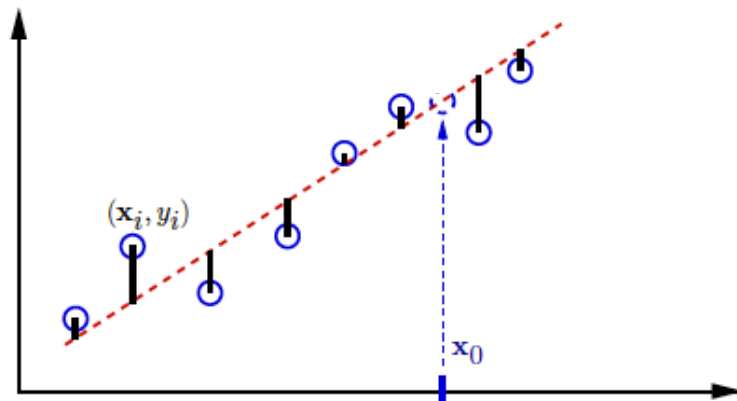
Linear fitting to data

- We want to fit a linear function to an observed set of points $X = [\mathbf{x}_1, \dots, \mathbf{x}_N]$ with associated labels $Y = [y_1, \dots, y_N]$.
 - Once we fit the function, we want to use it to *predict* the y for new \mathbf{x} .
- Least squares (LSQ) fitting criterion: find the function that minimizes sum (or average) of square distances between actual y s in the training set and predicted ones.



Linear fitting to data

- We want to fit a linear function to an observed set of points $X = [\mathbf{x}_1, \dots, \mathbf{x}_N]$ with associated labels $Y = [y_1, \dots, y_N]$.
 - Once we fit the function, we want to use it to *predict* the y for new \mathbf{x} .
- Least squares (LSQ) fitting criterion: find the function that minimizes sum (or average) of square distances between actual y s in the training set and predicted ones.



The fitted line is used as a predictor

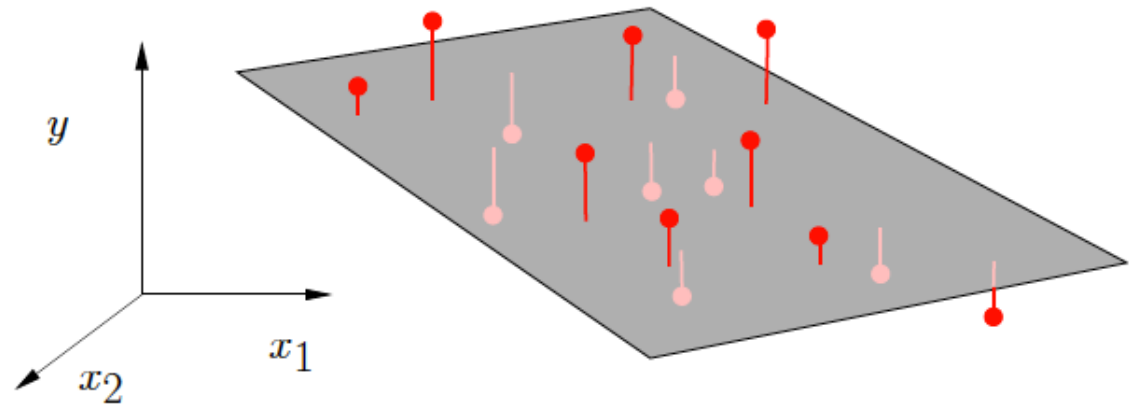
Linear Regression

- Demo
 - <http://hspm.sph.sc.edu/courses/J716/demos/LeastSquares/LeastSquaresDemo.html>

Linear functions

- General form: $f(\mathbf{x}; \mathbf{w}) = w_0 + w_1x_1 + \dots + w_dx_d$
- 1D case ($\mathcal{X} = \mathbb{R}$): a line

- $\mathcal{X} = \mathbb{R}^2$: a plane



- *Hyperplane* in general, d -D case.

Least squares: estimation

- We need to minimize w.r.t. \mathbf{w}

$$\begin{aligned} L(\mathbf{w}, \mathbf{X}) &= L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \\ &= \frac{1}{N} \sum_{i=1}^N (y_i - w_0 - w_1 x_{i1} - \dots - w_d x_{id})^2 \end{aligned}$$

- Necessary condition to minimize L : derivatives w.r.t. w_0, w_1, \dots, w_d must be zero.

Least squares in matrix form

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & & \vdots & \\ 1 & x_{N1} & \cdots & x_{Nd} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_d \end{bmatrix}.$$

- Predictions: $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$, errors: $\mathbf{y} - \mathbf{X}\mathbf{w}$, empirical loss:

$$L(\mathbf{w}, \mathbf{X}) = \frac{1}{N} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$