What about prior

- Billionaire says: Wait, I know that the thumbtack is “close” to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way...**

- Rather than estimating a single \( \theta \), we obtain a distribution over possible values of \( \theta \)
Bayesian Learning

- Use Bayes rule:
  \[ P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} \]
- Or equivalently:
  \[ P(\theta \mid D) \propto P(D \mid \theta)P(\theta) \]

If prior is uniform: \[ P(\theta) \propto 1 \]
\[ \Rightarrow P(\theta \mid D) \propto P(D \mid \theta) \]

Bayesian Learning for Thumbtack

- Likelihood function is simply Binomial:
  \[ P(D \mid \theta) = \theta^H (1 - \theta)^T \]
- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Closed-form representation of posterior
  - For Binomial, conjugate prior is Beta distribution
Beta prior distribution – $P(\theta)$

Prior: $\theta \sim \text{Beta}(\beta_H, \beta_T)$

Likelihood function: $P(D \mid \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T}$

Posterior: $P(\theta \mid D) \propto P(D \mid \theta)P(\theta)$

Posterior distribution: $P(\theta \mid D) \sim \text{Beta}((\beta_H + \alpha_H), (\beta_T + \alpha_T))$

Prior: Beta($\beta_H, \beta_T$)

Data: $\alpha_H$ heads and $\alpha_T$ tails

Conjugate: prior: Beta, likelihood: Binomial $\rightarrow$ posterior: Beta

Mean: $\frac{\alpha_H}{\beta_H + \beta_T}$

Mode: $\frac{\alpha_H - 1}{\beta_H + \beta_T - 2}$
Using Bayesian posterior

- Posterior distribution:
  \[ P(\theta \mid D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- Bayesian inference:
  - No longer single parameter:
    \[ E[f(0)] = \int_0^1 f(0) P(0 \mid D) \, d0 \]
  - Integral is often hard to compute

---

MAP: Maximum a posteriori approximation

- Posterior distribution:
  \[ P(\theta \mid D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- As more data is observed, Beta is more certain

- MAP: use most likely parameter:
  \[ \hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(\theta \mid D) \]
  \[ E[f(\theta)] \approx f(\hat{\theta}_{\text{MAP}}) \]

©Carlos Guestrin 2005-2009
MAP for Beta distribution

\[ P(\theta \mid D) = \frac{\theta^{\beta_H + \alpha_H - 1}(1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- MAP: use most likely parameter:
  \[ \hat{\theta}_{MAP} = \arg \max_\theta P(\theta \mid D) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]

- Beta prior equivalent to extra thumbtack flips
- As \( N \rightarrow \infty \), prior is “forgotten”
- But, for small sample size, prior is important!

What you need to know

- Go to the recitation on intro to probabilities
  - And, other recitations too
- Point estimation:
  - MLE
  - Bayesian learning
  - MAP
What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- You say: Let me tell you about Gaussians…

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

Some properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant)
  - \( X \sim N(\mu, \sigma^2) \)
  - \( Y = aX + b \quad \Rightarrow \quad Y \sim N(a\mu + b, a^2 \sigma^2) \)

- Sum of Gaussians
  - \( X \sim N(\mu_X, \sigma_X^2) \)
  - \( Y \sim N(\mu_Y, \sigma_Y^2) \)
  - \( Z = X + Y \quad \Rightarrow \quad Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \)
Learning a Gaussian

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

- Learn parameters
  - Mean \( \mu \)
  - Variance \( \sigma^2 \)

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

MLE for Gaussian

- Prob. of i.i.d. samples \( D=\{x_1, \ldots, x_N\} \):

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

\[
\text{MLE, } \sigma_{\text{MLE}} = \arg \max_{\mu, \sigma} P(D \mid \mu, \sigma)
\]

- Log-likelihood of data:

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

\[
= -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}
\]
Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

\[
\frac{d}{d\mu} \ln P(\mathcal{D} \mid \mu, \sigma) = \frac{d}{d\mu} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right] = 0
\]

\[
\frac{1}{\mu} \sum_{i=1}^{N} (x_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu) = 0
\]

\[
\Rightarrow \hat{\mu}_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

MLE for variance

- Again, set derivative to zero:

\[
\frac{d}{d\sigma} \ln P(\mathcal{D} \mid \mu, \sigma) = \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right] = 0
\]

\[
\frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 = \frac{1}{N^2} \sum_{i=1}^{N} (x_i - \mu)^2
\]

\[
\Rightarrow \hat{\sigma}^2_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2
\]
Learning Gaussian parameters

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

BTW. MLE for the variance of a Gaussian is **biased**
- Expected result of estimation is **not** true parameter!
- Unbiased variance estimator:
\[ \hat{\sigma}^2_{unbiased} = \frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \]

Bayesian learning of Gaussian parameters

- **Conjugate priors**
  - Mean: Gaussian prior
  - Variance: 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**

\[
P(\mu | \eta, \lambda) = \frac{1}{\lambda \sqrt{2\pi}} e^{-\frac{(\mu - \eta)^2}{2\lambda^2}}
\]

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

\[
\frac{d}{d\mu} \left[ \ln P(D | \mu) P(\mu) \right] = \frac{d}{d\mu} \left[ \ln P(D | \mu) + \ln P(\mu) \right]
\]

\[
\frac{\partial}{\partial \mu} \ln P(D | \mu) = \sum \frac{(x_i - \mu)}{\sigma^2}
\]

\[
\frac{\partial}{\partial \mu} \left( \frac{1}{\lambda \sqrt{2\pi}} e^{-\frac{(\mu - \eta)^2}{2\lambda^2}} \right) = \frac{N\mu + \mu}{\lambda^2} + \frac{\eta - \mu}{\lambda^2} = 0
\]

\[
\begin{align*}
\mu &= \frac{1}{N\lambda^2 + \lambda^2} \left( \sum x_i + \frac{N \eta}{\lambda^2} \right) \\
\end{align*}
\]

---

**Prediction of continuous variables**

- Billionaire says: Wait, that’s not what I meant!
- You says: Chill out, dude.
- He says: I want to predict a continuous variable for continuous inputs: I want to predict salaries from GPA.
- You say: I can regress that…

![Graph showing regression of GPA on salary]
The regression problem

- **Instances:** \( \langle x_i, t_i \rangle \)
- **Learn:** Mapping from \( x \) to \( t(x) \)
- **Hypothesis space:**
  - Given, basis functions
  - Find coeffs \( w = \{w_1, \ldots, w_k\} \)
  - Why is this called linear regression???
    - model is linear in the parameters
  - Precisely, minimize the residual squared error:
    \[
    w^* = \arg \min_w \sum_{j=1}^{N} \left( t(x_j) - \sum_i w_i h_i(x_j) \right)^2
    \]

The regression problem in matrix notation

- \( w^* = \arg \min_w \left( Hw - t \right)^T (Hw - t) \)
- \( \text{basis functions} \quad \text{weights} \quad \text{measurements} \)

\( H = \begin{bmatrix} h_1(x_1) & \ldots & h_K(x_1) \\ \vdots & \ddots & \vdots \\ h_1(x_N) & \ldots & h_K(x_N) \end{bmatrix} \)
\( w = \begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix} \)
\( t = \begin{bmatrix} t_1 \\ \vdots \\ t_N \end{bmatrix} \)
Regression solution = simple matrix operations

\[ w^* = \arg \min_w (Hw - t)^T(Hw - t) \]

residual error

solution: \[ w^* = (H^T H)^{-1} H^T t = A^{-1} b \]

where \( A = H^T H \)

\[ b = H^T t \]

- \( k \times k \) matrix for \( k \) basis functions
- \( k \times 1 \) vector

Announcements 1

- Readings associated with each class
  - See course website for specific sections, extra links, and further details
  - Visit the website frequently

- Recitations
  - Thursdays, 5:00-6:20pm in Gates Hillman 6115

- Special recitation on Matlab
  - Today! 5:00-6:20pm GHC 6115
Announcement 2

- First homework out later today!
  - Download from course website!
  - Start early!!!
  - Due Sept. 30th Wednesday
- Also, HW0!
  - Due this Thursday! Just to make sure you can access the submission directory
- To expedite grading:
  - there are 4 questions
  - please hand in 4 stapled separate parts, one for each question
- Privacy policy for returning homeworks and exams:
  - We write grades in second page of homework or exam
  - We want to handout graded homeworks in class, but to do that CMU requires you to sign a waiver acknowledging that someone may turn the page and find your grade
    - If you are not comfortable with this possibility, let us know and your homework will be available for pick up from Michelle Martin at GHC 8001

Billionaire (again) says: Why sum squared error???
You say: Gaussians, Dr. Gateson, Gaussians...

Model: prediction is linear function plus Gaussian noise

- \( t = \sum_i w_i h_i(x) + \varepsilon \)
- \( \varepsilon \sim N(0, \sigma^2) \)

Learn \( w \) using MLE

\[
P(t \mid x, w, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{[t-\sum_i w_i h_i(x)]^2}{2\sigma^2}}
\]
Maximizing log-likelihood

Maximize: \[ \arg \max_w \ln P(D | \mathbf{w}, \sigma) = \ln \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{j=1}^N \frac{1}{\sigma^2} e^{-\frac{1}{2\sigma^2} \left(t_j - \sum_i w_i h_i(x_j)\right)^2} \]

\[ = \arg \min_w \frac{1}{\sigma^2} \sum_{j=1}^N \left( t_j - \sum_i w_i h_i(x_j) \right)^2 \]

Least-squares Linear Regression is MLE for Gaussians!!!
Applications Corner 2

- Measure temperatures at some locations
- Predict temperatures throughout the environment

Applications Corner 3

- Predict when a sensor will fail
  - based several variables
    - age, chemical exposure, number of hours used,…