PROBABILITY DISTRIBUTIONS
These slides were sourced and/or modified from:
Christopher Bishop, Microsoft UK
Parametric Distributions

- Basic building blocks: \( p(x|\theta) \)
- Need to determine \( \theta \) given \( \{x_1, \ldots, x_N\} \)
- Representation: \( \theta^* \) or \( p(\theta) \) ?

Recall Curve Fitting

\[
p(t|x, x, t) = \int p(t|x, w)p(w|x, t) \, dw
\]
Binary Variables

- Coin flipping: heads=1, tails=0

\[ p(x = 1|\mu) = \mu \]

- Bernoulli Distribution

\[
\begin{align*}
\text{Bern}(x|\mu) &= \mu^x (1 - \mu)^{1-x} \\
\mathbb{E}[x] &= \mu \\
\text{var}[x] &= \mu(1 - \mu)
\end{align*}
\]
Guidelines for Paper Presentations

- Everyone should read the paper prior to the presentation and be prepared to discuss it.
  - What is the objective?
  - What tools from the course are being used?
  - What did you not understand?
Guidelines for Paper Presentations

□ For the presenter:

- Your presentation should be around 10 minutes long – no more than 15! (About 10 slides)
- What is the objective?
- What tools from the course are being used and how?
- What are the key ideas?
- What are the unsolved problems?
- Be prepared to answer questions from other students.
Binary Variables

- **N coin flips:**

  \[ p(m \text{ heads}|N, \mu) \]

- **Binomial Distribution**

  \[
  \text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}
  \]

  \[
  \mathbb{E}[m] \equiv \sum_{m=0}^{N} m \text{Bin}(m|N, \mu) = N\mu
  \]

  \[
  \text{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \text{Bin}(m|N, \mu) = N\mu(1 - \mu)
  \]
Binomial Distribution

Bin(m|10, 0.25)
Parameter Estimation

- **ML for Bernoulli**

- **Given:**

  \[ D = \{ x_1, \ldots, x_N \}, \ m \text{ heads (1), } N - m \text{ tails (0)} \]

  \[
  p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n} (1 - \mu)^{1-x_n}
  \]

  \[
  \ln p(D|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{ x_n \ln \mu + (1 - x_n) \ln(1 - \mu) \}
  \]

  \[
  \mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N}
  \]
Parameter Estimation

- Example: \( D = \{1, 1, 1\} \rightarrow \mu_{\text{ML}} = \frac{3}{3} = 1 \)

- Prediction: all future tosses will land heads up

- Overfitting to \( D \)
Beta Distribution

- Distribution over \( \mu \in [0, 1] \).

\[
\text{Beta}(\mu|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1 - \mu)^{b-1}
\]

\[
\mathbb{E}[\mu] = \frac{a}{a + b}
\]

\[
\text{var}[\mu] = \frac{ab}{(a + b)^2(a + b + 1)}
\]

where \( \Gamma(x) = \int_0^\infty u^{x-1}e^{-u} \, du \)

Note that

\( \Gamma(x + 1) = x\Gamma(x) \)

\( \Gamma(1) = 1 \)

\( \Gamma(x + 1) = x! \) when \( x \) is an integer.
Bayesian Bernoulli

\[ p(\mu|a_0, b_0, \mathcal{D}) \propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \]
\[ = \left( \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \right) \text{Beta}(\mu|a_0, b_0) \]
\[ \propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1} \]
\[ \propto \text{Beta}(\mu|a_N, b_N) \]

\[ a_N = a_0 + m \quad b_N = b_0 + (N - m) \]

The Beta distribution provides the conjugate prior for the Bernoulli distribution.
Beta Distribution

- $a = 0.1$
- $b = 0.1$

- $a = 1$
- $b = 1$

- $a = 2$
- $b = 3$

- $a = 8$
- $b = 4$
Prior \cdot Likelihood = Posterior

\[ \text{prior} \quad \text{likelihood function} \quad \text{posterior} \]
Properties of the Posterior

As the size $N$ of the data set increases

$$a_N \rightarrow m$$
$$b_N \rightarrow N - m$$
$$\mathbb{E}[\mu] = \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\text{ML}}$$
$$\text{var}[\mu] = \frac{a_N b_N}{(a_N + b_N)^2(a_N + b_N + 1)} \rightarrow 0$$
Multinomial Variables

1-of-K coding scheme: \( \mathbf{x} = (0, 0, 1, 0, 0, 0)^T \)

\[
p(\mathbf{x}|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}
\]

\( \forall k : \mu_k \geq 0 \) and \( \sum_{k=1}^{K} \mu_k = 1 \)

\[
\mathbb{E}[\mathbf{x}|\mu] = \sum_{\mathbf{x}} p(\mathbf{x}|\mu) \mathbf{x} = (\mu_1, \ldots, \mu_K)^T = \mu
\]

\[
\sum_{\mathbf{x}} p(\mathbf{x}|\mu) = \sum_{k=1}^{K} \mu_k = 1
\]
ML Parameter estimation

- Given:

\[ D = \{x_1, \ldots, x_N\} \]

\[ p(D|\mu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{n,k}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{n,k})} = \prod_{k=1}^{K} \mu_k^{m_k} \]

- To ensure \( \sum_k \mu_k = 1 \), use a Lagrange multiplier, \( \lambda \)

\[ \sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left( \sum_{k=1}^{K} \mu_k - 1 \right) \]

\[ \mu_k = -m_k / \lambda \quad \mu_k^{\text{ML}} = \frac{m_k}{N} \]

See Appendix E for a review of Lagrange multipliers.
The Multinomial Distribution

\[
\text{Mult}(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1, m_2, \ldots, m_K} \prod_{k=1}^{K} \mu_k^{m_k}
\]

\[
\mathbb{E}[m_k] = N \mu_k
\]
\[
\text{var}(m_k) = N \mu_k (1 - \mu_k)
\]
\[
\text{cov}(m_j, m_k) = -N \mu_j \mu_k \text{ for } j \neq k
\]

Where
\[
\binom{N}{m_1, m_2, \ldots, m_K} \equiv \frac{N!}{m_1! m_2! \ldots, m_K!}
\]
The Dirichlet Distribution

Conjugate prior for the multinomial distribution.

\[
\text{Dir}(\boldsymbol{\mu} | \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}
\]

\[\alpha_0 = \sum_{k=1}^{K} \alpha_k\]

Since \(\sum_{k=1}^{K} \mu_k = 1\)
Bayesian Multinomial

\[ p(\mu|D, \alpha) \propto p(D|\mu)p(\mu|\alpha) \propto \prod_{k=1}^{K} \mu_{k}^{\alpha_{k} + m_{k} - 1} \]

\[ p(\mu|D, \alpha) = \text{Dir}(\mu|\alpha + m) \]

\[ = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1) \cdots \Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_{k}^{\alpha_{k} + m_{k} - 1} \]
Bayesian Multinomial

\[ \alpha_k = 10^{-1} \]

\[ \alpha_k = 10^0 \]

\[ \alpha_k = 10^1 \]
The Gaussian Distribution

\[ \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi \sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} \]

\[ \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \]
Central Limit Theorem

- The distribution of the sum of $N$ i.i.d. random variables becomes increasingly Gaussian as $N$ grows.
- Example: $N$ uniform $[0,1]$ random variables.
Geometry of the Multivariate Gaussian

\[ \Delta^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \]
where \( \Delta \equiv \) Mahalanobis distance from \( \mu \) to \( x \)

Eigenvector equation: \( \Sigma u_i = \lambda_i u_i \)

where \((u_i, \lambda_i)\) are the \( i \)th eigenvector and eigenvalue of \( \Sigma \).

Note that \( \Sigma \) real and symmetric \( \rightarrow \lambda_i \) real.

Proof?

See Appendix C for a review of matrices and eigenvectors.
Geometry of the Multivariate Gaussian

\[ \Delta^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \]
\[ \Delta = \text{Mahalanobis distance from } \mu \text{ to } x \]

\[ \Sigma^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} u_i u_i^T \]
where \((u_i, \lambda_i)\) are the \(i\)th eigenvector and eigenvalue of \(\Sigma\).

\[ \Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \]

\[ y_i = u_i^T (x - \mu) \]

or \(y = U(x - \mu)\)
Moments of the Multivariate Gaussian

\[
\mathbb{E}[x] = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} x \, dx
\]

\[
= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int \exp \left\{ -\frac{1}{2} z^T \Sigma^{-1} z \right\} (z + \mu) \, dz
\]

thanks to anti-symmetry of \( z \)

\[
\mathbb{E}[x] = \mu
\]
Moments of the Multivariate Gaussian

\[ \mathbb{E}[xx^T] = \mu \mu^T + \Sigma \]

\[ \text{cov}[x] = \mathbb{E} \left[ (x - \mathbb{E}[x])(x - \mathbb{E}[x])^T \right] = \Sigma \]
Partitioned Gaussian Distributions

\[ p(x) = \mathcal{N}(x | \mu, \Sigma) \]

\[
x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}
\]

\[
\Lambda \equiv \Sigma^{-1} \quad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}
\]
Partitioned Conditionals and Marginals

\[ p(x_a | x_b) = \mathcal{N}(x_a | \mu_{a|b}, \Sigma_{a|b}) \]

\[
\Sigma_{a|b} = \Lambda_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba}
\]

\[
\mu_{a|b} = \Sigma_{a|b} \{ \Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b) \}
\]

\[
= \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b)
\]

\[
= \mu_a + \Sigma_{ab} \Sigma_{bb}^{-1} (x_b - \mu_b)
\]

\[
p(x_a) = \int p(x_a, x_b) \, dx_b
\]

\[ = \mathcal{N}(x_a | \mu_a, \Sigma_{aa}) \]
Partitioned Conditionals and Marginals

\[ p(x_a, x_b) \]

\[ x_b = 0.7 \]

\[ p(x_a | x_b = 0.7) \]

\[ p(x_a) \]
Maximum Likelihood for the Gaussian

- **Given** i.i.d. data $X = (x_1, \ldots, x_N)^T$, the log likelihood function is given by

$$
\ln p(X|\mu, \Sigma) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu)
$$

- **Sufficient statistics**

$$
\sum_{n=1}^{N} x_n \\
\sum_{n=1}^{N} x_n x_n^T
$$
Maximum Likelihood for the Gaussian

- Set the derivative of the log likelihood function to zero,
  \[
  \frac{\partial}{\partial \mu} \ln p(X|\mu, \Sigma) = \sum_{n=1}^{N} \Sigma^{-1}(x_n - \mu) = 0
  \]

- and solve to obtain
  \[
  \mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n.
  \]

- Similarly
  \[
  \Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T.
  \]

  (Recall: If \( x \) and \( a \) are vectors, then \( \frac{\partial}{\partial x}(x^t a) = \frac{\partial}{\partial x}(a^t x) = a \))
Maximum Likelihood for the Gaussian

Under the true distribution

\[ \mathbb{E}[\mu_{ML}] = \mu \]
\[ \mathbb{E}[\Sigma_{ML}] = \frac{N - 1}{N} \Sigma. \]

Hence define

\[ \tilde{\Sigma} = \frac{1}{N - 1} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T. \]
Bayesian Inference for the Gaussian (Univariate Case)

- Assume \( \sigma^2 \) is known. Given i.i.d. data \( x = \{x_1, \ldots, x_N\} \), the likelihood function for \( \mu \) is given by

\[
p(x|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}.
\]

- This has a Gaussian shape as a function of \( \mu \) (but it is not a distribution over \( \mu \)).
Bayesian Inference for the Gaussian (Univariate Case)

- Combined with a Gaussian prior over $\mu$,
  \[ p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2). \]

- this gives the posterior
  \[ p(\mu|x) \propto p(x|\mu)p(\mu). \]

- Completing the square over $\mu$, we see that
  \[ p(\mu|x) = \mathcal{N}(\mu|\mu_N, \sigma_N^2). \]
Bayesian Inference for the Gaussian

\[ \mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}, \quad \mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}x_n \]

\[ \frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}. \]

Shortcut: Get \( \Delta^2 \) in form \( a\mu^2 - 2b\mu + c = a(\mu - b / a)^2 + \text{const} \) and identify \( \mu_N = b / a \)

\[ \frac{1}{\sigma_N^2} = a \]

\[ \mu_N \quad \mu_0 \quad \mu_{ML} \]

\[ \sigma_N^2 \quad \sigma_0^2 \quad 0 \]

Note:

<table>
<thead>
<tr>
<th>( N = 0 )</th>
<th>( N \to \infty )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_N )</td>
<td>( \mu_0 )</td>
</tr>
<tr>
<td>( \sigma_N^2 )</td>
<td>( \sigma_0^2 )</td>
</tr>
</tbody>
</table>
Example: \( p(\mu|x) = \mathcal{N}(\mu|\mu_N, \sigma^2_N) \) for \( N = 0, 1, 2 \) and 10.
Bayesian Inference for the Gaussian

- Sequential Estimation

\[
p(\mu|x) \propto p(\mu)p(x|\mu) = \left[ p(\mu) \prod_{n=1}^{N-1} p(x_n|\mu) \right] p(x_N|\mu) \propto \mathcal{N}(\mu|\mu_{N-1}, \sigma^2_{N-1}) p(x_N|\mu)
\]

- The posterior obtained after observing \( N \{ 1 \) data points becomes the prior when we observe the \( N^{th} \) data point.
Bayesian Inference for the Gaussian

Now assume $\mu$ is known. The likelihood function for $\lambda = 1 / \sigma^2$ is given by

$$p(x|\lambda) = \prod_{n=1}^{N} N(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp \left\{ -\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}.$$ 

This has a Gamma shape as a function of $\lambda$. 
Bayesian Inference for the Gaussian

- The Gamma distribution

\[ \text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \]

\[ \mathbb{E}[\lambda] = \frac{a}{b} \quad \text{var}[\lambda] = \frac{a}{b^2} \]
Bayesian Inference for the Gaussian

- Now we combine a Gamma prior, \( \text{Gam}(\lambda|a_0, b_0) \) with the likelihood function for \( \lambda \) to obtain

\[
p(\lambda|x) \propto \lambda^{a_0-1} e^{-\lambda^2/2} \exp \left\{ -b_0 \lambda - \frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}
\]

- which we recognize as \( \text{Gam}(\lambda|a_N, b_N) \) with

\[
a_N = a_0 + \frac{N}{2} \\
b_N = b_0 + \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^2 = b_0 + \frac{N}{2} \sigma_{\text{ML}}^2.
\]
Bayesian Inference for the Gaussian

If both $\mu$ and $\lambda$ are unknown, the joint likelihood function is given by

$$p(x|\mu, \lambda) = \prod_{n=1}^{N} \left( \frac{\lambda}{2\pi} \right)^{1/2} \exp \left\{ -\frac{\lambda}{2} (x_n - \mu)^2 \right\}$$

$$\propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^N \exp \left\{ \lambda \mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2} \sum_{n=1}^{N} x_n^2 \right\}.$$

We need a prior with the same functional dependence on $\mu$ and $\lambda$. 
Bayesian Inference for the Gaussian

The Gaussian-gamma distribution

\[ p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\text{Gam}(\lambda|a, b) \]

\[ \propto \exp \left\{ -\frac{\beta\lambda}{2}(\mu - \mu_0)^2 \right\} \lambda^{a-1} \exp \{-b\lambda\} \]
Bayesian Inference for the Gaussian

- The Gaussian-gamma distribution
Bayesian Inference for the Gaussian

- **Multivariate conjugate priors**
  - $\mu$ unknown, $\Lambda$ known: $p(\mu)$ Gaussian.
  - $\Lambda$ unknown, $\mu$ known: $p(\Lambda)$ Wishart,
    \[
    \mathcal{W}(\Lambda|W, \nu) = B|\Lambda|^{(\nu-D-1)/2} \exp \left( -\frac{1}{2} \text{Tr}(W^{-1}\Lambda) \right).
    \]
  - $\mu$ and $\Lambda$ unknown: $p(\mu,\Lambda)$ Gaussian-Wishart,
    \[
    p(\mu, \Lambda|\mu_0, \beta, W, \nu) = \mathcal{N}(\mu|\mu_0, (\beta\Lambda)^{-1}) \mathcal{W}(\Lambda|W, \nu)
    \]
Student’s t-Distribution

\[
p(x|\mu, a, b) = \int_0^\infty N(x|\mu, \tau^{-1}) \Gamma(\nu/2, \nu/2) \, d\tau \\
= \int_0^\infty N(x|\mu, (\eta \lambda)^{-1}) \Gamma(\eta \nu/2, \nu/2) \, d\eta \\
= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left( \frac{\lambda}{\pi \nu} \right)^{1/2} \left[ 1 + \frac{\lambda(x - \mu)^2}{\nu} \right]^{-\nu/2 - 1/2} \\
= \text{St}(x|\mu, \lambda, \nu)
\]

\[\square \text{where} \]
\[
\lambda = a/b \quad \eta = \tau b/a \quad \nu = 2a.
\]

\[\square \text{Infinite mixture of Gaussians.}\]
Student’s t-Distribution

\[
\begin{array}{c|cc}
\nu \to \infty & \nu = 1 & \nu \rightarrow \infty \\
\text{St}(x|\mu, \lambda, \nu) & \text{Cauchy} & \mathcal{N}(x|\mu, \lambda^{-1})
\end{array}
\]
Student’s t-Distribution

- Robustness to outliers: Gaussian vs t-distribution.
Student’s t-Distribution

□ The $D$-variate case:

\[
\mathcal{St}(x|\mu, \Lambda, \nu) = \int_0^\infty \mathcal{N}(x|\mu, (\eta\Lambda)^{-1})\Gamma\text{am}(\eta|\nu/2, \nu/2) \, d\eta
\]

\[
= \frac{\Gamma(D/2 + \nu/2)}{\Gamma(\nu/2)} \frac{|\Lambda|^{1/2}}{(\pi\nu)^{D/2}} \left[ 1 + \frac{\Delta^2}{\nu} \right]^{-D/2-\nu/2}
\]

□ where

\[
\Delta^2 = (x - \mu)^T \Lambda (x - \mu)
\]

□ Properties:

\[
\mathbb{E}[x] = \mu, \quad \text{if } \nu > 1
\]

\[
\text{cov}[x] = \frac{\nu}{(\nu - 2)} \Lambda^{-1}, \quad \text{if } \nu > 2
\]

\[
\text{mode}[x] = \mu
\]
Periodic variables

- **Examples:** time of day, direction, ...
- **We require**

\[
\begin{align*}
  p(\theta) & \geq 0 \\
  \int_{0}^{2\pi} p(\theta) \, d\theta &= 1 \\
  p(\theta + 2\pi) &= p(\theta).
\end{align*}
\]
von Mises Distribution

- This requirement is satisfied by

\[ p(\theta | \theta_0, m) = \frac{1}{2\pi I_0(m)} \exp \{m \cos(\theta - \theta_0)\} \]

- where

\[ I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp \{m \cos \theta\} \, d\theta \]

- is the 0\textsuperscript{th} order modified Bessel function of the 1\textsuperscript{st} kind.

(The von Mises distribution is the intersection of an isotropic bivariate Gaussian with the unit circle)
von Mises Distribution

- $m = 5, \theta_0 = \pi/4$
- $m = 1, \theta_0 = 3\pi/4$

- $m = 5, \theta_0 = \pi/4$
- $m = 1, \theta_0 = 3\pi/4$
Maximum Likelihood for von Mises

Given a data set, \( \mathcal{D} = \{\theta_1, \ldots, \theta_N\} \), the log likelihood function is given by

\[
\ln p(\mathcal{D}|\theta_0, m) = -N \ln(2\pi) - N \ln I_0(m) + m \sum_{n=1}^{N} \cos(\theta_n - \theta_0).
\]

Maximizing with respect to \( \mu_0 \) we directly obtain

\[
\theta_0^{\text{ML}} = \tan^{-1} \left\{ \frac{\sum_n \sin \theta_n}{\sum_n \cos \theta_n} \right\}.
\]

Similarly, maximizing with respect to \( m \) we get

\[
\frac{I_1(m_{ML})}{I_0(m_{ML})} = \frac{1}{N} \sum_{n=1}^{N} \cos(\theta_n - \theta_0^{\text{ML}})
\]

which can be solved numerically for \( m_{ML} \).
Mixtures of Gaussians

- Old Faithful data set

![Graph showing Old Faithful data set with single Gaussian and mixture of two Gaussians](image-url)
Mixtures of Gaussians

- Combine simple models into a complex model:

\[ p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x | \mu_k, \Sigma_k) \]

Component Mixing coefficient

\[ \forall k : \pi_k \geq 0 \quad \sum_{k=1}^{K} \pi_k = 1 \]
Mixtures of Gaussians

(a) Graph showing overlapping ellipses with labels 0.5, 0.3, and 0.2.

(b) Graph showing a complex contour plot.

(c) 3D plot graph illustrating a distribution.

Probability Distributions
Mixtures of Gaussians

- Determining parameters $\mu$, $\sigma$ and $\pi$ using maximum log likelihood

$$
\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_{k} \mathcal{N}(x_{n} | \mu_{k}, \Sigma_{k}) \right\}
$$

Log of a sum; no closed form maximum.

- Solution: use standard, iterative, numeric optimization methods or the expectation maximization algorithm (Chapter 9).