

Advanced Machine Learning Lecture 7

Approximate Inference

19.11.2015

Bastian Leibe

RWTH Aachen

<http://www.vision.rwth-aachen.de/>

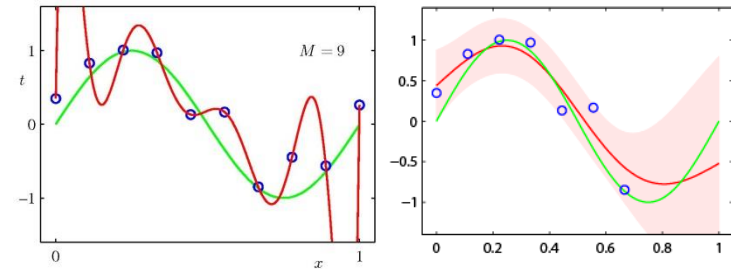
leibe@vision.rwth-aachen.de

This Lecture: *Advanced Machine Learning*

- Regression Approaches

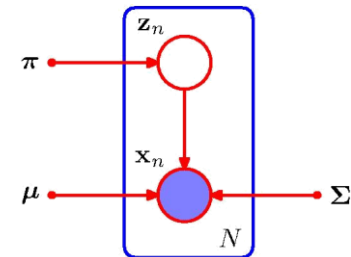
- Linear Regression
- Regularization (Ridge, Lasso)
- Gaussian Processes

$$f : \mathcal{X} \rightarrow \mathbb{R}$$



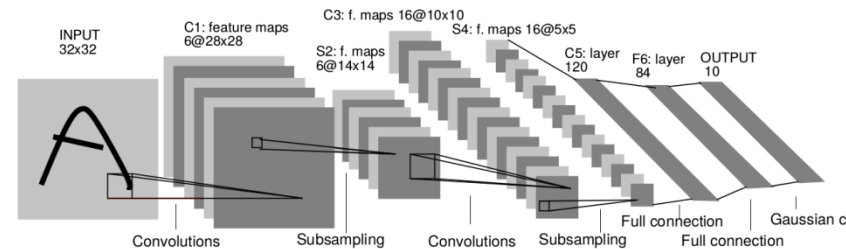
- Learning with Latent Variables

- Probability Distributions
- Approximate Inference
- Mixture Models
- EM and Generalizations



- Deep Learning

- Neural Networks
- CNNs, RNNs, RBMs, etc.



Recap: Binary Variables

- **Bernoulli distribution**

- Probability distribution over $x \in \{0,1\}$:

$$\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$$

$$\mathbb{E}[x] = \mu$$

$$\text{var}[x] = \mu(1-\mu)$$

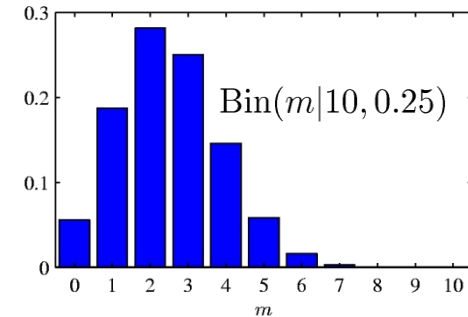
- **Binomial distribution**

- Generalization for m outcomes out of N trials

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

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

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



Recap: The Beta Distribution

- **Beta distribution**

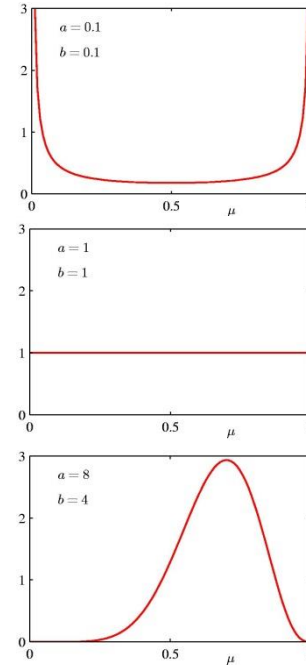
- **Distribution over $\mu \in [0,1]$:**

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

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

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

- where $\Gamma(x)$ is the gamma function, a continuous generalization of the factorial. ($\Gamma(x+1) = x!$ iff x is an integer).



- **Properties**

- The Beta distribution generalizes the Binomial to arbitrary values of a and b , while keeping the same functional form.
- It is therefore a **conjugate prior** for the Bernoulli and Binomial.

Recap: Multinomial Variables

- **Multinomial variables**
 - Variables that can take one of K possible distinct states
 - Convenient: 1-of- K coding scheme: $\mathbf{x} = (0, 0, 1, 0, 0, 0)^T$
- **Generalization of the Bernoulli distribution**
 - Distribution of \mathbf{x} with K outcomes

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

with the constraints

$$\forall k : \mu_k \geq 0 \quad \text{and} \quad \sum_{k=1}^K \mu_k = 1$$

Recap: Multinomial Variables

- **Multinomial Distribution**

- **Variables using 1-of- K coding scheme:** $\mathbf{x} = (0, 0, 1, 0, 0, 0)^T$
- **Joint distribution over m_1, \dots, m_K conditioned on $\boldsymbol{\mu}$ and N**

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

$$\mathbb{E}[m_k] = N \mu_k$$

$$\text{var}[m_k] = N \mu_k (1 - \mu_k)$$

$$\text{cov}[m_j, m_k] = -N \mu_j \mu_k$$

with the constraints

$$\forall k : \mu_k \geq 0 \quad \text{and} \quad \sum_{k=1}^K \mu_k = 1$$

Recap: The Dirichlet Distribution

- **Dirichlet Distribution**

- **Multivariate generalization of the Beta distribution**

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

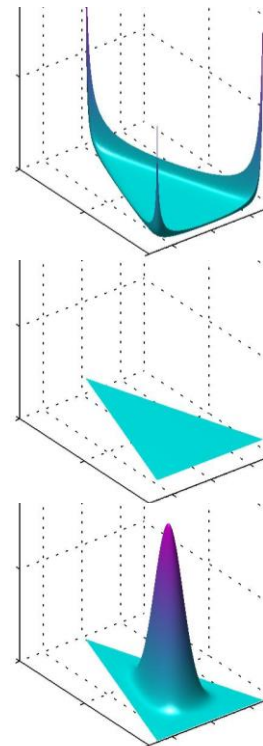
$$\mathbb{E}[\mu_k] = \frac{\alpha_k}{\alpha_0}$$

$$\text{var}[\mu_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$$

$$\text{cov}[\mu_j, \mu_k] = -\frac{\alpha_j \alpha_k}{\alpha_0^2(\alpha_0 + 1)}$$

- **Properties**

- **Conjugate prior** for the Multinomial.
- The Dirichlet distribution over K variables is confined to a $K-1$ dimensional simplex.

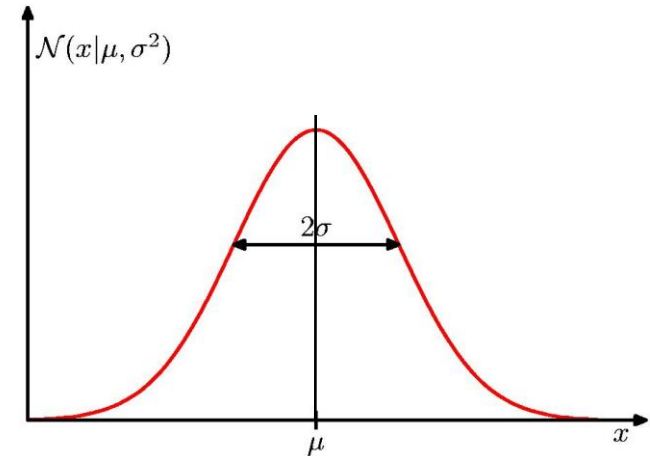


Recap: The Gaussian Distribution

- One-dimensional case

- Mean μ
- Variance σ^2

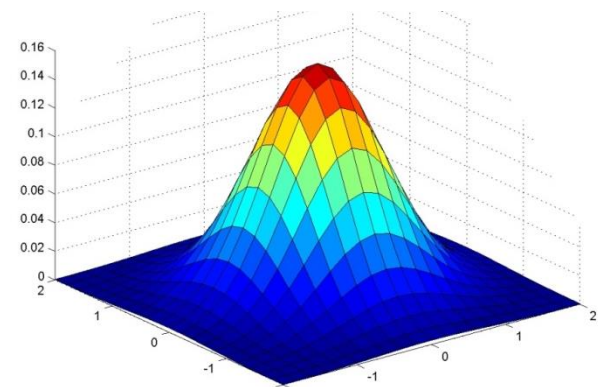
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$



- Multi-dimensional case

- Mean μ
- Covariance Σ

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\}$$



Recap: Bayes' Theorem for Gaussian Variables

- **Marginal and Conditional Gaussians**

- Suppose we are given a Gaussian prior $p(\mathbf{x})$ and a Gaussian conditional distribution $p(\mathbf{y}|\mathbf{x})$ (a **linear Gaussian model**)

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

- From this, we can compute

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T)$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}$$

⇒ Closed-form solution for (Gaussian) marginal and posterior.

Maximum Likelihood for the Gaussian

- **Maximum Likelihood**

- Given i.i.d. data $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T$, the log likelihood function is given by

$$\begin{aligned} \log p(\mathbf{X}|\mu, \Sigma) &= -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| \\ &\quad - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \mu)^T \Sigma^{-1} (\mathbf{x}_n - \mu) \end{aligned}$$

- **Sufficient statistics**

- The likelihood depends on the data set only through

$$\sum_{n=1}^N \mathbf{x}_n \qquad \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T$$

- Those are the **sufficient statistics** for the Gaussian distribution.

ML for the Gaussian

- **Setting the derivative to zero**

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

- **Solve to obtain**

$$\boldsymbol{\mu}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n.$$

- **And similarly, but a bit more involved**

$$\boldsymbol{\Sigma}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^{\text{T}}.$$

ML for the Gaussian

- Comparison with true results
 - Under the true distribution, we obtain

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_{\text{ML}}] &= \boldsymbol{\mu} \\ \mathbb{E}[\boldsymbol{\Sigma}_{\text{ML}}] &= \frac{N-1}{N} \boldsymbol{\Sigma}.\end{aligned}$$

⇒ The ML estimate for the covariance is biased and underestimates the true covariance!

- Therefore define the following unbiased estimator

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^{\text{T}}.$$

Bayesian Inference for the Gaussian

- Let's begin with a simple example

- Consider a single Gaussian random variable x .
- Assume σ^2 is known and the task is to infer the mean μ .
- Given i.i.d. data $\mathbf{X} = (x_1, \dots, x_N)^T$, the likelihood function for μ is given by

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

- The likelihood function has a Gaussian shape as a function of μ .
- ⇒ The **conjugate prior** for this case is again a Gaussian.

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2).$$

Bayesian Inference for the Gaussian

- Combined with a Gaussian prior over μ

$$p(\mu) = \mathcal{N}(\mu | \mu_0, \sigma_0^2).$$

- This results in the posterior

$$p(\mu | \mathbf{x}) \propto p(\mathbf{x} | \mu) p(\mu).$$

- Completing the square over μ , we can derive that

$$p(\mu | \mathbf{x}) = \mathcal{N}(\mu | \mu_N, \sigma_N^2)$$

where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{\text{ML}},$$

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n$$

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

Visualization of the Results

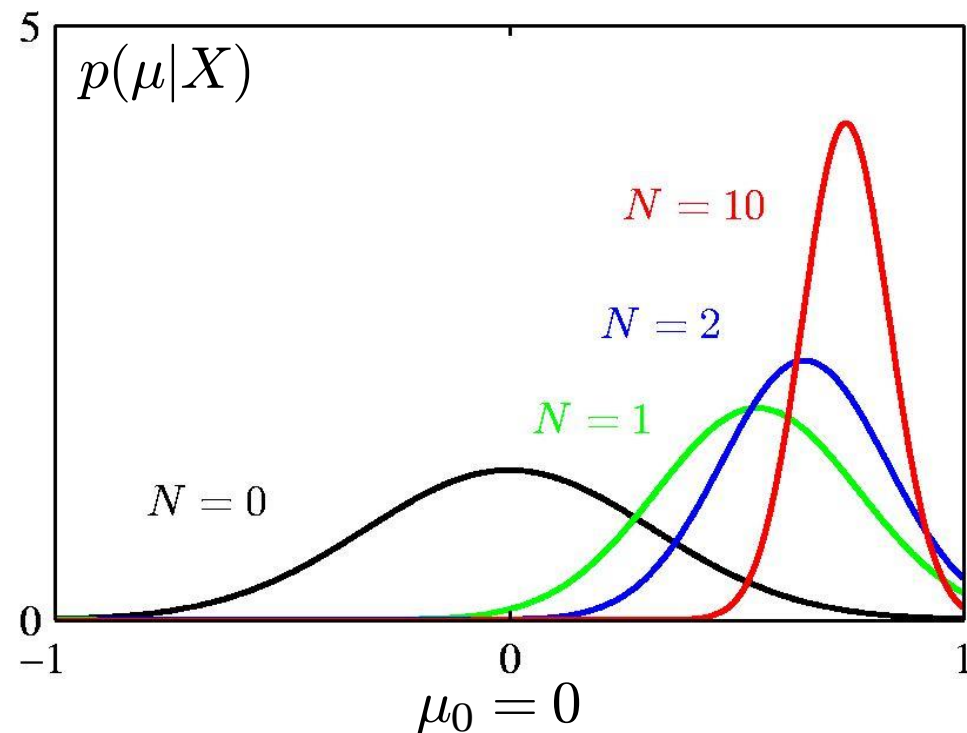
- Bayes estimate:

$$\mu_N = \frac{\sigma^2 \mu_0 + N \sigma_0^2 \mu_{\text{ML}}}{\sigma^2 + N \sigma_0^2}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

- Behavior for large N

	$N = 0$	$N \rightarrow \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0



Bayesian Inference for the Gaussian

- **More complex case**

- Now assume μ is known and the precision λ shall be inferred.
- The likelihood function for $\lambda = 1/\sigma^2$ is given by

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

- This has the shape of a **Gamma function** of λ .

The Gamma Distribution

- **Gamma distribution**

- Product of a power of λ and the exponential of a linear function of λ .

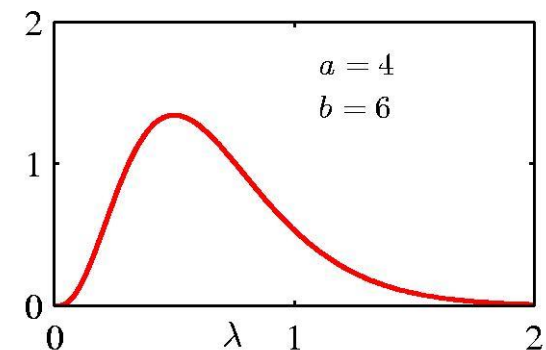
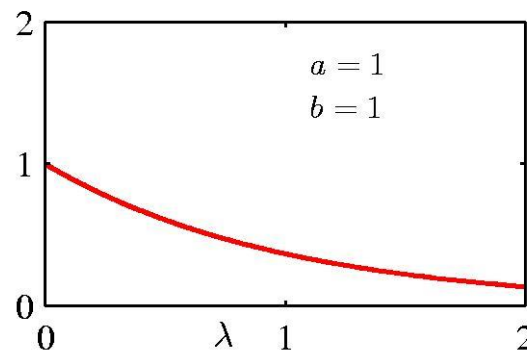
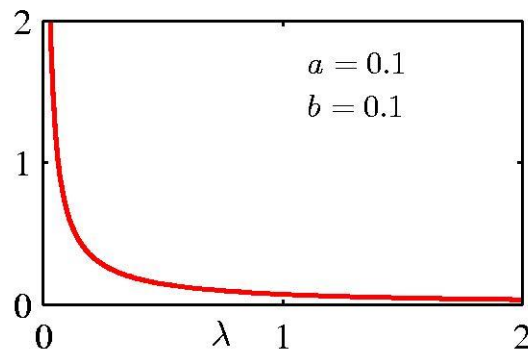
$$\text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

- **Properties**

- Finite integral if $a > 0$ and the distribution itself is finite if $a \geq 1$.

- Moments $\mathbb{E}[\lambda] = \frac{a}{b}$ $\text{var}[\lambda] = \frac{a}{b^2}$

- **Visualization**



Bayesian Inference for the Gaussian

- **Bayesian estimation**

- **Combine a Gamma prior $\text{Gam}(\lambda|a_0, b_0)$ with the likelihood function to obtain**

$$p(\lambda|\mathbf{X}) \propto \lambda^{a_0-1} \lambda^{N/2} \exp \left\{ -b_0 \lambda - \frac{\lambda}{2} \sum_{n=1}^N (x_n - \mu)^2 \right\}$$

- **We recognize this again as a Gamma function $\text{Gam}(\lambda|a_N, b_N)$ 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

- Even more complex case
 - Assume that both μ and λ are unknown
 - The joint likelihood function is given by

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

⇒ Need a prior with the same functional dependence on μ and λ .

The Gaussian-Gamma Distribution

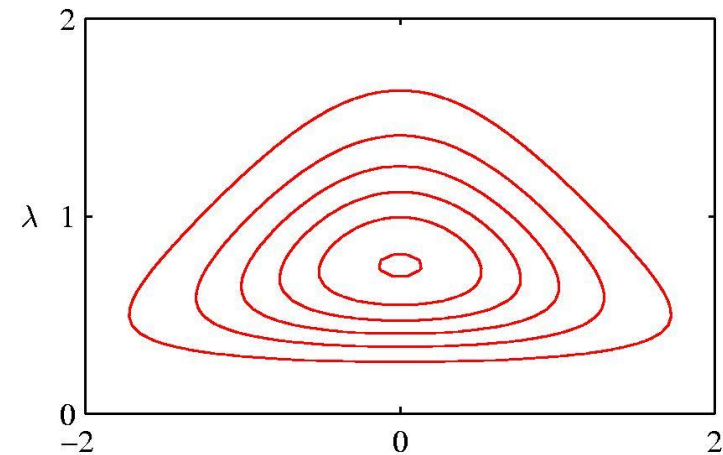
- **Gaussian-Gamma distribution**

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

$$\propto \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu - \mu_0)^2\right\}}_{\text{Quadratic in } \mu} \underbrace{\lambda^{a-1} \exp\{-b\lambda\}}_{\text{Linear in } \lambda}$$

- Quadratic in μ .
- Linear in λ .

- **Visualization**



Bayesian Inference for the Gaussian

- **Multivariate conjugate priors**

- μ unknown, Λ known: $p(\mu)$ **Gaussian**.

- Λ unknown, μ known: $p(\Lambda)$ **Wishart**,

$$\mathcal{W}(\Lambda | \mathbf{W}, \nu) = B |\Lambda|^{(\nu - D - 1)/2} \exp\left(-\frac{1}{2} \text{Tr}(\mathbf{W}^{-1} \Lambda)\right).$$

- Λ and μ unknown: $p(\mu, \Lambda)$ **Gaussian-Wishart**,

$$p(\mu, \Lambda | \mu_0, \beta, \mathbf{W}, \nu) = \mathcal{N}(\mu | \mu_0, (\beta \Lambda)^{-1}) \mathcal{W}(\Lambda | \mathbf{W}, \nu)$$

Recap: Bayesian Inference for the Gaussian

- **Multivariate conjugate priors**

- μ unknown, Λ known: $p(\mu)$ **Gaussian**.

- Λ unknown, μ known: $p(\Lambda)$ **Wishart**,

$$\mathcal{W}(\Lambda | \mathbf{W}, \nu) = B |\Lambda|^{(\nu - D - 1)/2} \exp\left(-\frac{1}{2} \text{Tr}(\mathbf{W}^{-1} \Lambda)\right).$$

- Λ and μ unknown: $p(\mu, \Lambda)$ **Gaussian-Wishart**,

$$p(\mu, \Lambda | \mu_0, \beta, \mathbf{W}, \nu) = \mathcal{N}(\mu | \mu_0, (\beta \Lambda)^{-1}) \mathcal{W}(\Lambda | \mathbf{W}, \nu)$$

Student's t-Distribution

- Gaussian estimation

- The conjugate prior for the precision of a Gaussian is a Gamma distribution.
- Suppose we have a univariate Gaussian $\mathcal{N}(x | \mu, \tau^{-1})$ together with a Gamma prior $\text{Gam}(\tau | a, b)$.
- By integrating out the precision, obtain the **marginal distribution**

$$\begin{aligned} p(x | \mu, a, b) &= \int_0^{\infty} \mathcal{N}(x | \mu, \tau^{-1}) \text{Gam}(\tau | a, b) d\tau \\ &= \int_0^{\infty} \mathcal{N}(x | \mu, (\eta\lambda)^{-1}) \text{Gam}(\eta | \nu/2, \nu/2) d\eta \end{aligned}$$

- This corresponds to an **infinite mixture of Gaussians** having the same mean, but different precision.

Student's t-Distribution

- Student's t-Distribution

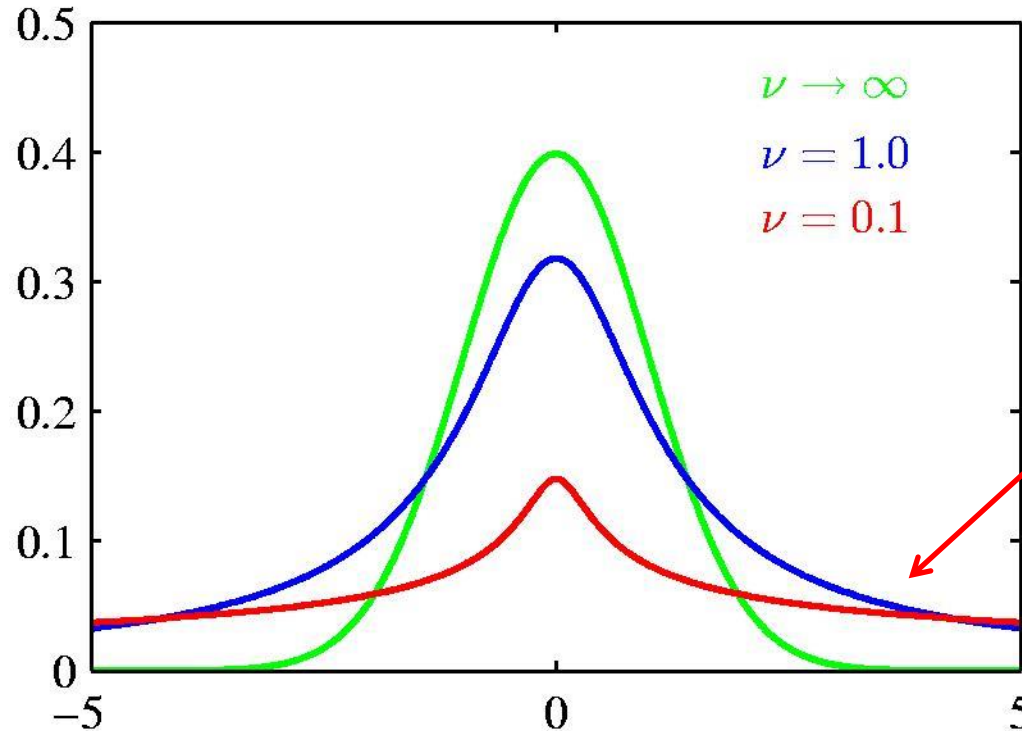
- We reparametrize the infinite mixture of Gaussians to get

$$\text{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda(x - \mu)^2}{\nu}\right]^{-\nu/2 - 1/2}$$

- Parameters

- “Precision” $\lambda = a/b$
- “Degrees of freedom” $\nu = 2a$.

Student's t-Distribution: Visualization

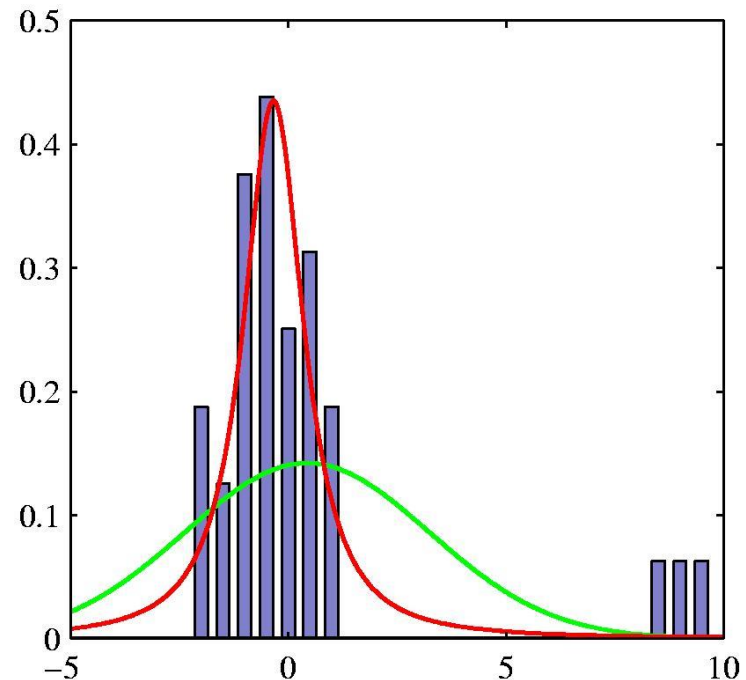
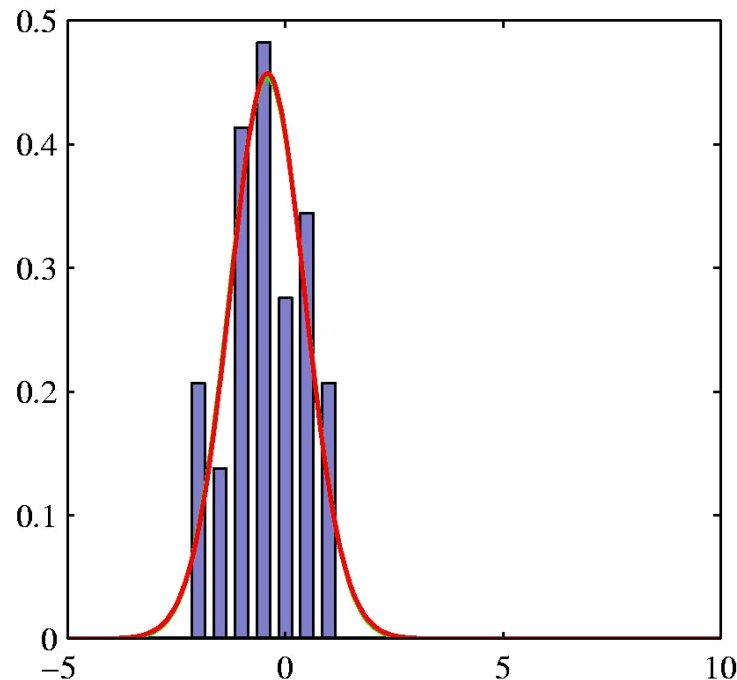


- Behavior

	$\nu = 1$	$\nu \rightarrow \infty$
$\text{St}(x \mu, \lambda, \nu)$	Cauchy	$\mathcal{N}(x \mu, \lambda^{-1})$

Student's t-Distribution

- Robustness to outliers: **Gaussian** vs **t-distribution**.



⇒ The t-distribution is much less sensitive to outliers, can be used for robust regression.

⇒ Downside: ML solution for t-distribution requires EM algorithm.

Student's t-Distribution: Multivariate Case

- Multivariate case in D dimensions

$$\begin{aligned}\text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) &= \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}(\eta|\nu/2, \nu/2) d\eta \\ &= \frac{\Gamma(D/2 + \nu/2)}{\Gamma(\nu/2)} \frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}} \left[1 + \frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2}\end{aligned}$$

where $\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu})$ is the Mahalanobis distance.

- Properties

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

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

$$\text{mode}[\mathbf{x}] = \boldsymbol{\mu}$$

Topics of This Lecture

- **Approximate Inference**
 - Variational methods
 - Sampling approaches
- **Sampling approaches**
 - Sampling from a distribution
 - Ancestral Sampling
 - Rejection Sampling
 - Importance Sampling
- **Markov Chain Monte Carlo**
 - Markov Chains
 - Metropolis Algorithm
 - Metropolis-Hastings Algorithm
 - Gibbs Sampling

Approximate Inference

- **Exact Bayesian inference is often intractable.**
 - **Often infeasible to evaluate the posterior distribution or to compute expectations w.r.t. the distribution.**
 - E.g. because the dimensionality of the latent space is too high.
 - Or because the posterior distribution has a too complex form.
 - **Problems with continuous variables**
 - Required integrations may not have closed-form solutions.
 - **Problems with discrete variables**
 - Marginalization involves summing over all possible configurations of the hidden variables.
 - There may be exponentially many such states.

⇒ **We need to resort to approximation schemes.**

Two Classes of Approximation Schemes

- **Deterministic approximations (Variational methods)**
 - Based on analytical approximations to the posterior distribution
 - E.g. by assuming that it factorizes in a certain form
 - Or that it has a certain parametric form (e.g. a Gaussian).
 - ⇒ Can never generate exact results, but are often scalable to large applications.
- **Stochastic approximations (Sampling methods)**
 - Given infinite computationally resources, they can generate exact results.
 - Approximation arises from the use of a finite amount of processor time.
 - ⇒ Enable the use of Bayesian techniques across many domains.
 - ⇒ But: computationally demanding, often limited to small-scale problems.

Topics of This Lecture

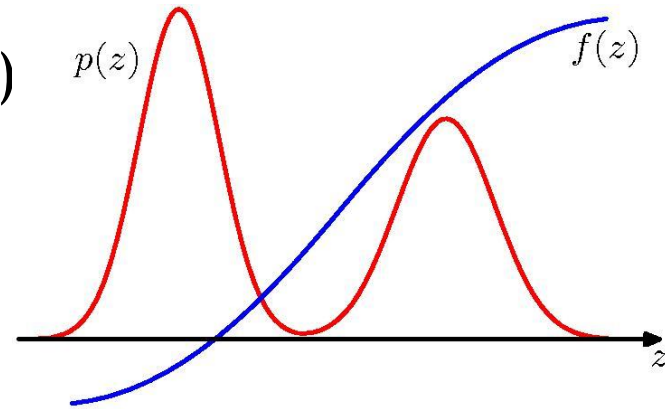
- Approximate Inference
 - Variational methods
 - Sampling approaches
- **Sampling approaches**
 - **Sampling from a distribution**
 - **Ancestral Sampling**
 - **Rejection Sampling**
 - **Importance Sampling**
- Markov Chain Monte Carlo
 - Markov Chains
 - Metropolis Algorithm
 - Metropolis-Hastings Algorithm
 - Gibbs Sampling

Sampling Idea

- **Objective:**

- Evaluate expectation of a function $f(\mathbf{z})$ w.r.t. a probability distribution $p(\mathbf{z})$.

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z}$$



- **Sampling idea**

- Draw L independent samples $\mathbf{z}^{(l)}$ with $l = 1, \dots, L$ from $p(\mathbf{z})$.
- This allows the expectation to be approximated by a finite sum

$$\hat{f} = \frac{1}{L} \sum_{l=1}^L f(\mathbf{z}^{(l)})$$

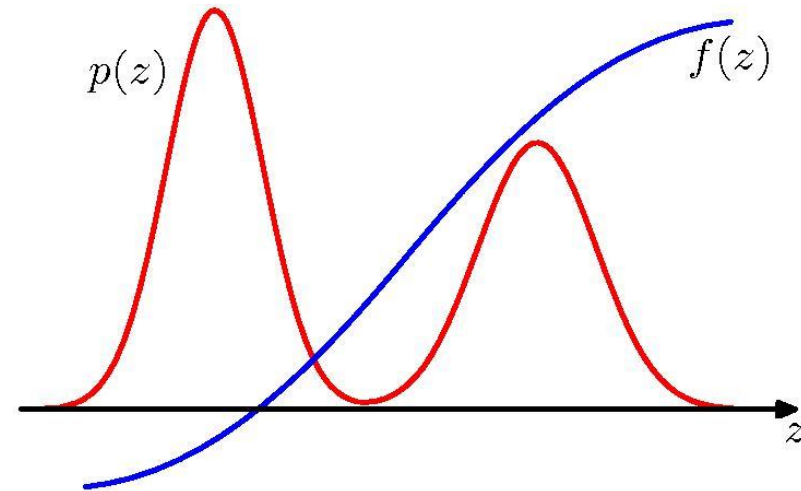
- As long as the samples $\mathbf{z}^{(l)}$ are drawn independently from $p(\mathbf{z})$, then

$$\mathbb{E}[\hat{f}] = \mathbb{E}[f]$$

⇒ **Unbiased estimate, independent** of the dimension of \mathbf{z} !

Sampling - Challenges

- **Problem 1: Samples might not be independent**
⇒ Effective sample size might be much smaller than apparent sample size.



- **Problem 2:**
 - If $f(z)$ is small in regions where $p(z)$ is large and vice versa, the expectation may be dominated by regions of small probability.
 - ⇒ Large sample sizes necessary to achieve sufficient accuracy.

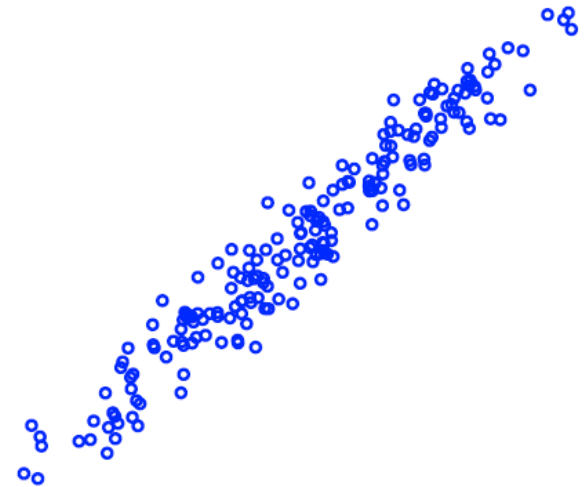
Parametric Density Model

- **Example:**

- A simple multivariate (d-dimensional) Gaussian model

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

- This is a “generative” model in the sense that we can generate samples \mathbf{x} according to the distribution.



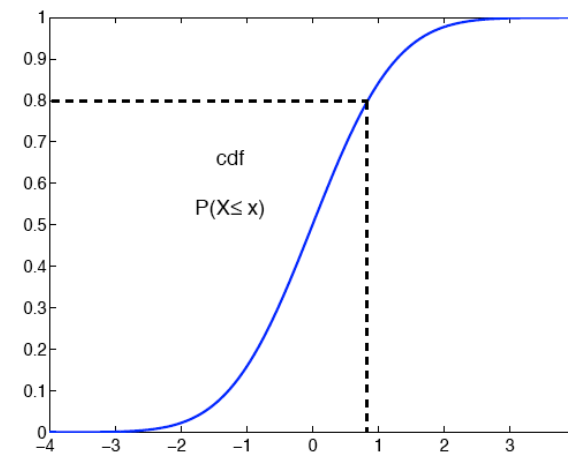
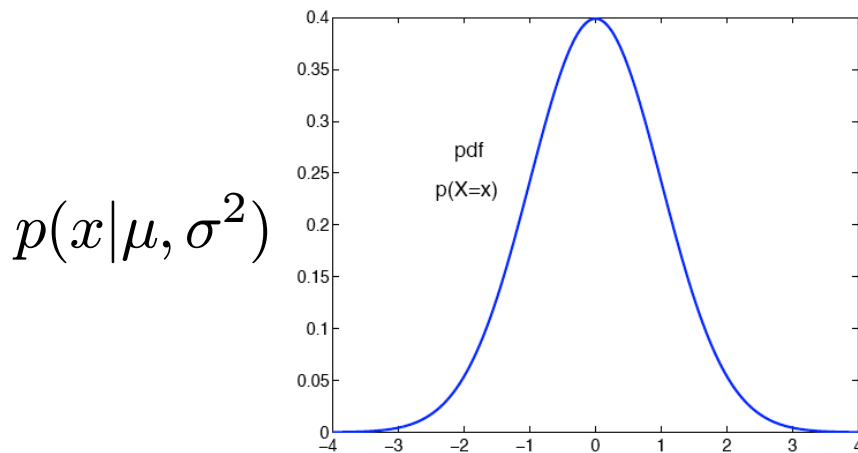
Sampling from a Gaussian

- Given: 1-dim. Gaussian pdf (probability density function) $p(\mathbf{x} | \mu, \sigma^2)$ and the corresponding cumulative distribution:

$$F_{\mu, \sigma^2}(x) = \int_{-\infty}^x p(x | \mu, \sigma^2) dx$$

- To draw samples from a Gaussian, we can invert the cumulative distribution function:

$$u \sim \text{Uniform}(0, 1) \Rightarrow F_{\mu, \sigma^2}^{-1}(u) \sim p(x | \mu, \sigma^2)$$



$$F_{\mu, \sigma^2}(x)$$

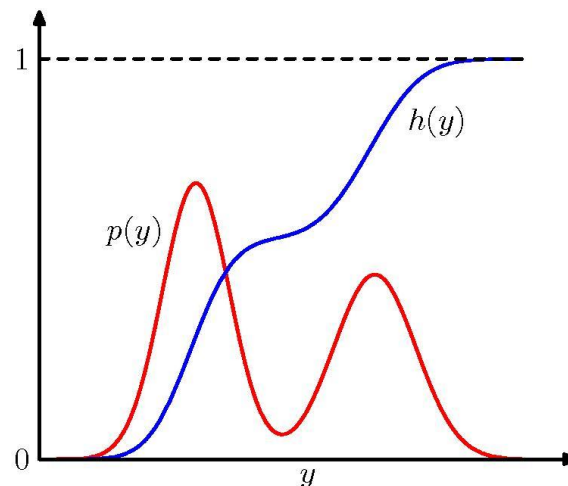
Sampling from a pdf (Transformation method)

- In general, assume we are given the pdf $p(\mathbf{x})$ and the corresponding cumulative distribution:

$$F(x) = \int_{-\infty}^x p(z) dz$$

- To draw samples from this pdf, we can invert the cumulative distribution function:

$$u \sim \text{Uniform}(0, 1) \Rightarrow F^{-1}(u) \sim p(x)$$

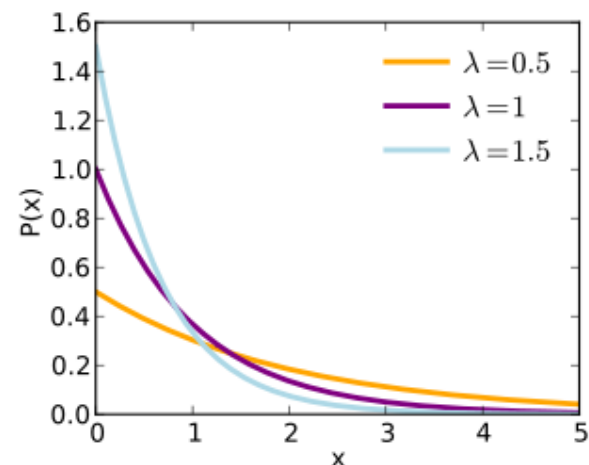


Example 1: Sampling from Exponential Distrib.

- **Exponential Distribution**

$$p(y) = \lambda \exp(-\lambda y)$$

where $0 \leq y < \infty$.



- **Transformation sampling**

- Indefinite Integral

$$h(y) = 1 - \exp(-\lambda y)$$

- Inverse function

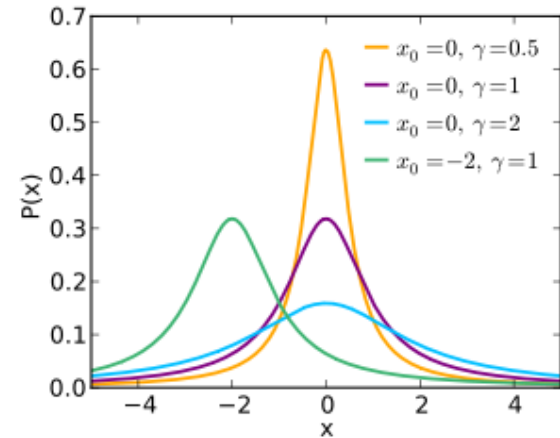
$$y = h(y)^{-1} = -\lambda^{-1} \ln(1 - z)$$

for a uniformly distributed input variable z .

Example 2: Sampling from Cauchy Distrib.

- Cauchy Distribution

$$p(y) = \frac{1}{\pi} \frac{1}{1 + y^2}$$



- Transformation sampling

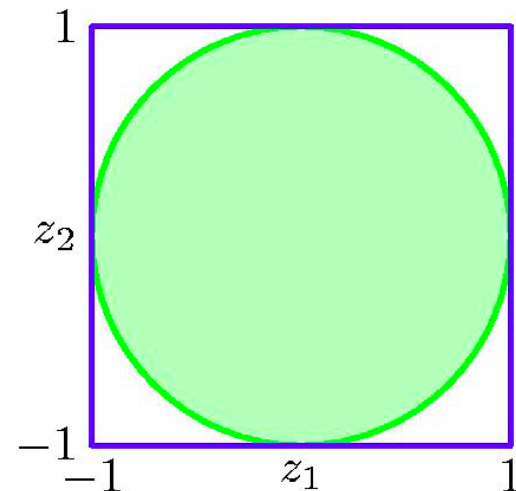
- Inverse of integral can be expressed as a tan function.

$$y = h(y)^{-1} = \tan(z)$$

for a uniformly distributed input variable z .

Note: Efficient Sampling from a Gaussian

- Problem with transformation method
 - Integral over Gaussian cannot be expressed in analytical form.
 - Standard transformation approach is very inefficient.
- More efficient: **Box-Muller Algorithm**
 - Generate pairs of uniformly distributed random numbers $z_1, z_2 \in (-1, 1)$.
 - Discard each pair unless it satisfies $r^2 = z_1^2 + z_2^2 \leq 1$.
 - This leads to a uniform distribution of points inside the unit circle with $p(z_1, z_2) = 1/\pi$.



Box-Muller Algorithm (cont'd)

- **Box-Muller Algorithm (cont'd)**

- For each pair z_1, z_2 evaluate

$$y_1 = z_1 \left(\frac{-2 \ln r^2}{r^2} \right)^{1/2} \quad y_2 = z_2 \left(\frac{-2 \ln r^2}{r^2} \right)^{1/2}$$

- Then the joint distribution of y_1 and y_2 is given by

$$\begin{aligned} p(y_1, y_2) &= p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right| \\ &= \left[\frac{1}{\sqrt{2\pi}} \exp(-y_1^2/2) \right] \left[\frac{1}{\sqrt{2\pi}} \exp(-y_2^2/2) \right] \end{aligned}$$

$\Rightarrow y_1$ and y_2 are independent and each has a Gaussian distribution with mean μ and variance σ^2 .

- If $y \sim \mathcal{N}(0,1)$, then $\sigma y + \mu \sim \mathcal{N}(\mu, \sigma^2)$.

Box-Muller Algorithm (cont'd)

- **Multivariate extension**

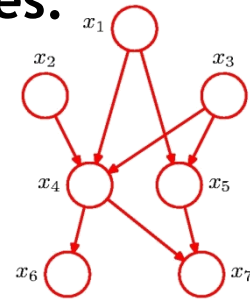
- If \mathbf{z} is a vector valued random variable whose components are independent and Gaussian distributed with $\mathcal{N}(0,1)$,
- Then $\mathbf{y} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z}$ will have mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
- Where $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ is the **Cholesky decomposition** of $\boldsymbol{\Sigma}$.

Ancestral Sampling

- Generalization of this idea to directed graphical models.

- Joint probability factorizes into conditional probabilities:

$$p(\mathbf{x}) = \prod_{k=1}^K p(x_k | \text{pa}_k)$$



- Ancestral sampling

- Assume the variables are ordered such that there are no links from any node to a lower-numbered node.
 - Start with lowest-numbered node and draw a sample from its distribution.

$$\hat{x}_1 \sim p(x_1)$$

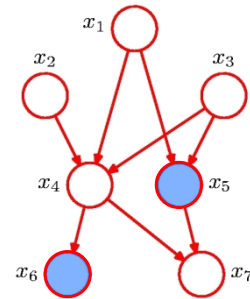
- Cycle through each of the nodes in order and draw samples from the conditional distribution (where the parent variable is set to its sampled value).

$$\hat{x}_n \sim p(x_n | \text{pa}_n)$$

Logic Sampling

- **Extension of Ancestral sampling**
 - Directed graph where some nodes are instantiated with observed values.
- **Use ancestral sampling, except**
 - When sample is obtained for an observed variable, if they agree then sample value is retained and proceed to next variable.
 - If they don't agree, whole sample is discarded.
- **Result**
 - Approach samples correctly from the posterior distribution.
 - However, probability of accepting a sample decreases rapidly as the number of observed variables increases.

⇒ Approach is rarely used in practice.



Discussion

- Transformation method
 - Limited applicability, as we need to invert the indefinite integral of the required distribution $p(\mathbf{z})$.
 - This will only be feasible for a limited number of simple distributions.
- More general
 - Rejection Sampling
 - Importance Sampling

Rejection Sampling

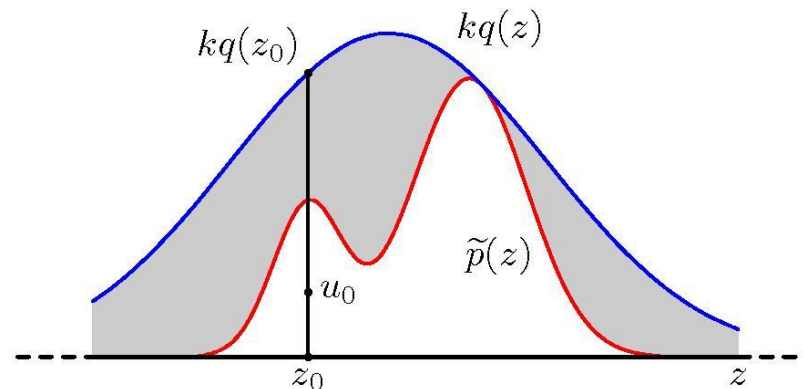
- Assumptions

- Sampling directly from $p(\mathbf{z})$ is difficult.
- But we can easily evaluate $p(\mathbf{z})$ (up to some normalization factor Z_p):

$$p(\mathbf{z}) = \frac{1}{Z_p} \tilde{p}(\mathbf{z})$$

- Idea

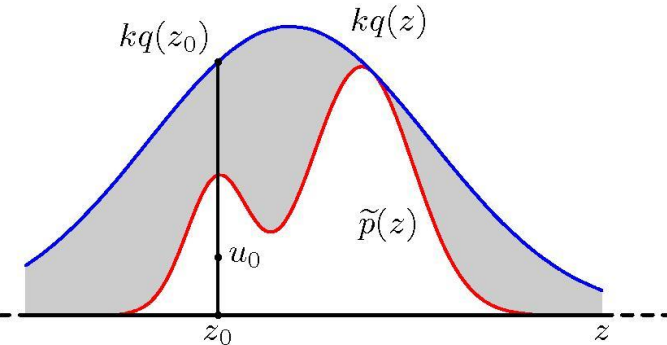
- We need some simpler distribution $q(\mathbf{z})$ (called **proposal distribution**) from which we can draw samples.
- Choose a constant k such that: $\forall z : kq(z) \geq \tilde{p}(z)$



Rejection Sampling

• Sampling procedure

- Generate a number z_0 from $q(z)$.
- Generate a number u_0 from the uniform distribution over $[0, kq(z_0)]$.
- If $u_0 > \tilde{p}(z_0)$ reject sample, otherwise accept.
 - Sample is rejected if it lies in the grey shaded area.
 - The remaining pairs (u_0, z_0) have uniform distribution under the curve $\tilde{p}(z)$.



• Discussion

- Original values of z are generated from the distribution $q(z)$.
- Samples are accepted with probability $\tilde{p}(z)/kq(z)$

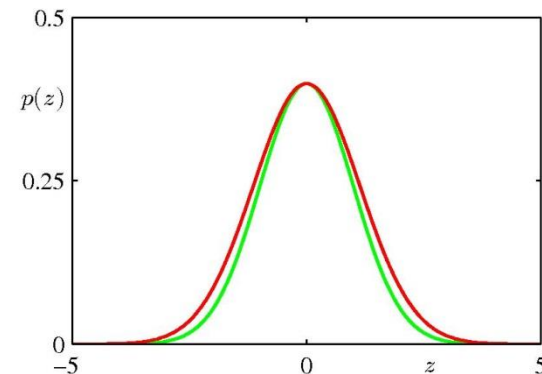
$$p(\text{accept}) = \int \frac{\tilde{p}(z)}{kq(z)} q(z) dz = \frac{1}{k} \int \tilde{p}(z) dz$$

$\Rightarrow k$ should be as small as possible!

Rejection Sampling - Discussion

- **Limitation: high-dimensional spaces**
 - For rejection sampling to be of practical value, we require that $kq(z)$ be close to the required distribution, so that the rate of rejection is minimal.
- **Artificial example**
 - Assume that $p(\mathbf{z})$ is Gaussian with covariance matrix $\sigma_p^2 I$
 - Assume that $q(\mathbf{z})$ is Gaussian with covariance matrix $\sigma_q^2 I$
 - Obviously: $\sigma_q^2 \geq \sigma_p^2$
 - In D dimensions: $k = (\sigma_q / \sigma_p)^D$.
 - Assume σ_q is just 1% larger than σ_p .
 - $D = 1000 \Rightarrow k = 1.01^{1000} \geq 20,000$
 - And $p(\text{accept}) \cdot \frac{1}{20000}$

⇒ Often impractical to find good proposal distributions for high dimensions!



Example: Sampling from a Gamma Distrib.

- Gamma distribution

$$\text{Gam}(z|a, b) = \frac{1}{\Gamma(a)} b^a z^{a-1} \exp(-bz) \quad a > 1$$

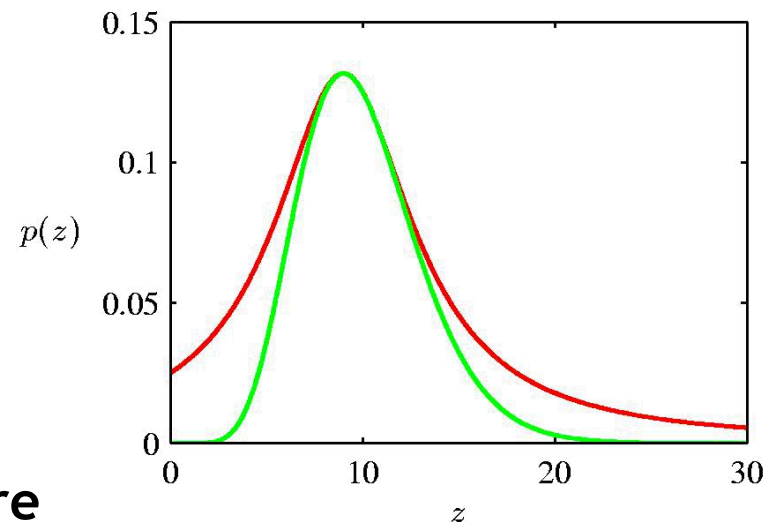
- Rejection sampling approach

- For $a > 1$, Gamma distribution has a bell-shaped form.
- Suitable proposal distribution is Cauchy (for which we can use the transformation method).
- Generalize Cauchy slightly to ensure it is nowhere smaller than Gamma: $y = b \tan y + c$ for uniform y .
- This gives random numbers distributed according to

$$q(z) = \frac{k}{1 + (z - c)^2/b^2}$$

with optimal
rejection rate for

$$\begin{aligned} c &= a - 1 \\ b^2 &= 2a - 1 \end{aligned}$$



Importance Sampling

- Approach

- Approximate expectations directly
(but does not enable to draw samples from $p(\mathbf{z})$ directly).

- Goal:
$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

- Simplistic strategy: Grid sampling

- Discretize \mathbf{z} -space into a uniform grid.
- Evaluate the integrand as a sum of the form

$$\mathbb{E}[f] \simeq \sum_{l=1}^L f(\mathbf{z}^{(l)})p(\mathbf{z}^{(l)})d\mathbf{z}$$

- But: number of terms grows exponentially with number of dimensions!

Importance Sampling

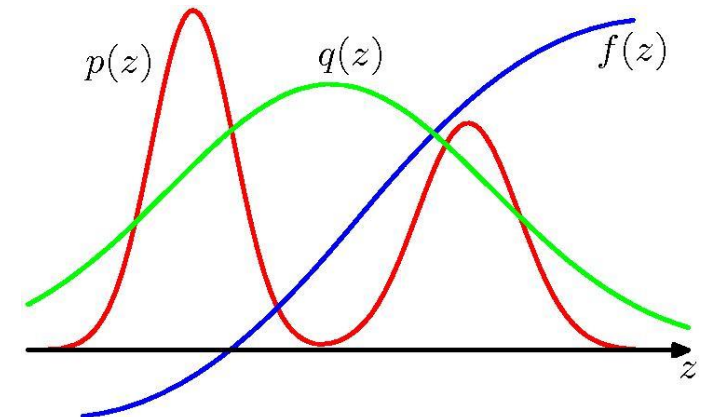
- **Idea**

- Use a proposal distribution $q(\mathbf{z})$ from which it is easy to draw samples.
- Express expectations in the form of a finite sum over samples $\{\mathbf{z}^{(l)}\}$ drawn from $q(\mathbf{z})$.

$$\begin{aligned}\mathbb{E}[f] &= \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})d\mathbf{z} \\ &\approx \frac{1}{L} \sum_{l=1}^L \frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})} f(\mathbf{z}^{(l)})\end{aligned}$$

- **with importance weights**

$$r_l = \frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})}$$



Importance Sampling

- **Typical setting:**

- $p(\mathbf{z})$ can only be evaluated up to an unknown normalization constant

$$p(\mathbf{z}) = \tilde{p}(\mathbf{z})/Z_p$$

- $q(\mathbf{z})$ can also be treated in a similar fashion.

$$q(\mathbf{z}) = \tilde{q}(\mathbf{z})/Z_q$$

- **Then**

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z} = \frac{Z_q}{Z_p} \int f(\mathbf{z}) \frac{\tilde{p}(\mathbf{z})}{\tilde{q}(\mathbf{z})} q(\mathbf{z})d\mathbf{z}$$

$$\approx \frac{Z_q}{Z_p} \frac{1}{L} \sum_{l=1}^L \tilde{r}_l f(\mathbf{z}^{(l)})$$

- **with:** $\tilde{r}_l = \frac{\tilde{p}(\mathbf{z}^{(l)})}{\tilde{q}(\mathbf{z}^{(l)})}$

Importance Sampling

- Ratio of normalization constants can be evaluated

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q} \int \tilde{p}(\mathbf{z}) d\mathbf{z} = \int \frac{\tilde{p}(\mathbf{z}^{(l)})}{\tilde{q}(\mathbf{z}^{(l)})} q(\mathbf{z}) d\mathbf{z} \simeq \frac{1}{L} \sum_{l=1}^L \tilde{r}_l$$

- and therefore

$$\mathbb{E}[f] \simeq \sum_{l=1}^L w_l f(\mathbf{z}^{(l)})$$

- with

$$w_l = \frac{\tilde{r}_l}{\sum_m \tilde{r}_m} = \frac{\frac{\tilde{p}(\mathbf{z}^{(l)})}{\tilde{q}(\mathbf{z}^{(l)})}}{\sum_m \frac{\tilde{p}(\mathbf{z}^{(m)})}{\tilde{q}(\mathbf{z}^{(m)})}}$$

Importance Sampling - Discussion

- **Observations**

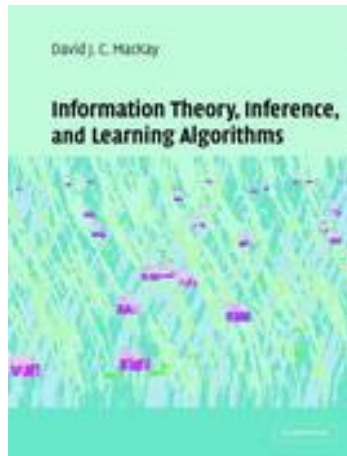
- Success of importance sampling depends crucially on how well the sampling distribution $q(\mathbf{z})$ matches the desired distribution $p(\mathbf{z})$.
- Often, $p(\mathbf{z})f(\mathbf{z})$ is strongly varying and has a significant proportion of its mass concentrated over small regions of \mathbf{z} -space.
⇒ Weights r_l may be dominated by a few weights having large values.
- Practical issue: if none of the samples falls in the regions where $p(\mathbf{z})f(\mathbf{z})$ is large...
 - The results may be arbitrary in error.
 - And there will be no diagnostic indication (no large variance in r_l)!
- Key requirement for sampling distribution $q(\mathbf{z})$:
 - Should not be small or zero in regions where $p(\mathbf{z})$ is significant!

Topics of This Lecture

- Approximate Inference
 - Variational methods
 - Sampling approaches
- Sampling approaches
 - Sampling from a distribution
 - Ancestral Sampling
 - Rejection Sampling
 - Importance Sampling
- **Markov Chain Monte Carlo**
 - **Markov Chains**
 - **Metropolis Algorithm**
 - **Metropolis-Hastings Algorithm**
 - **Gibbs Sampling**

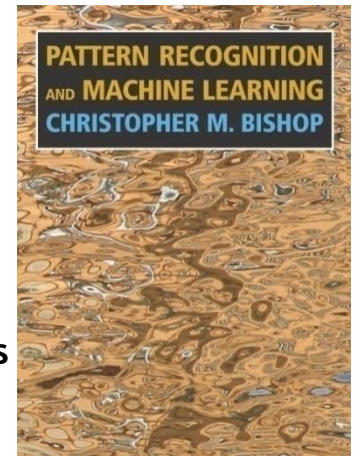
References and Further Reading

- Sampling methods for approximate inference are described in detail in Chapter 11 of Bishop's book.



David MacKay
Information Theory, Inference, and Learning Algorithms
Cambridge University Press, 2003

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006



- Another good introduction to Monte Carlo methods can be found in Chapter 29 of MacKay's book (also available online: <http://www.inference.phy.cam.ac.uk/mackay/itprnn/book.html>)