Advanced Machine Learning
Lecture 7

Approximate Inference

07.11.2012

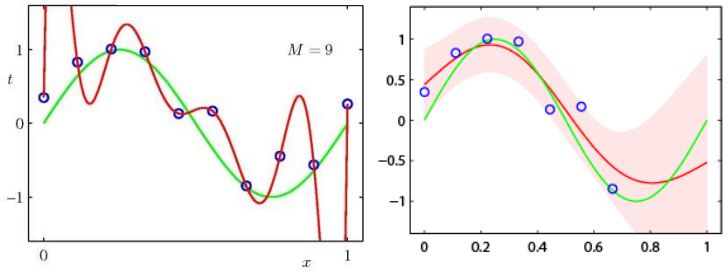
Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de/

leibe@vision.rwth-aachen.de

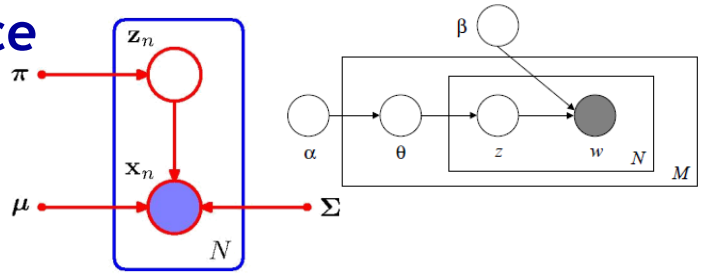
This Lecture: *Advanced Machine Learning*

- Regression Approaches
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Kernels (Kernel Ridge Regression)
  - Gaussian Processes

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



- Bayesian Estimation & Bayesian Non-Parametrics
  - Prob. Distributions, Approx. Inference
  - Mixture Models & EM
  - Dirichlet Processes
  - Latent Factor Models
  - Beta Processes



- SVMs and Structured Output Learning
  - SV Regression, SVDD
  - Large-margin Learning

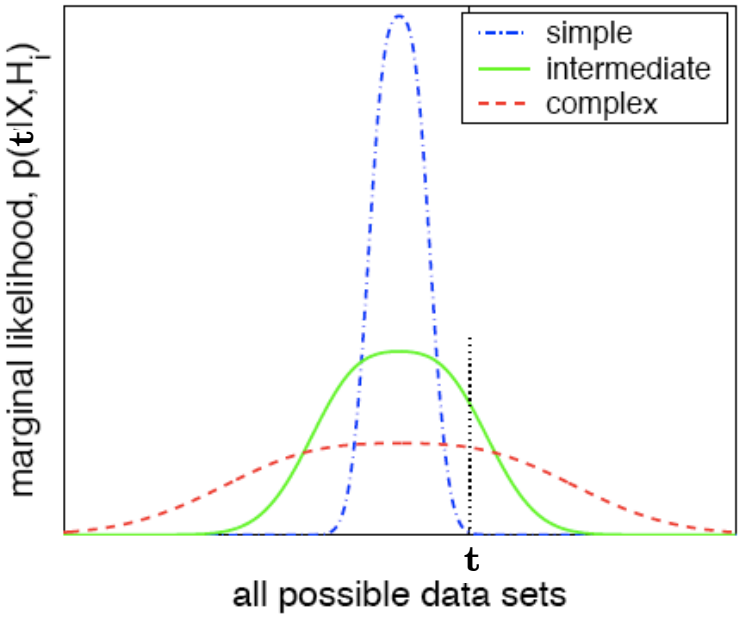
$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

Correction: Bayesian Model Selection

- Discussion
  - Marginal likelihood is main difference to non-Bayesian methods

$$p(\mathbf{t}|X, \mathcal{H}_i) = \int p(\mathbf{t}|X, \theta, \mathcal{H}_i) p(\theta|\mathcal{H}_i) d\theta$$

  - It automatically incorporates a trade-off between the model fit and the model complexity:
    - A simple model can only account for a limited range of possible sets of target values – if a simple model fits well, it obtains a high marginal likelihood.
    - A complex model can account for a large range of possible sets of target values – therefore, it can never attain a very high marginal likelihood.



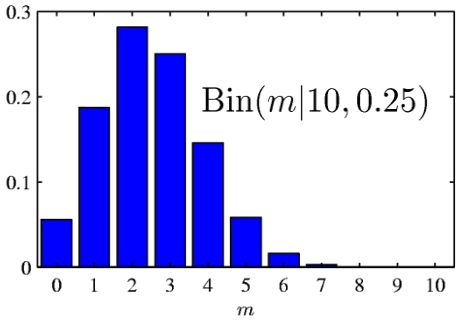
Slide credit: Bernt Schiele
Image source: Rasmussen & Williams, 2006

Recap: Binary Variables

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

$$\begin{aligned} \text{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \text{var}[x] &= \mu(1-\mu) \end{aligned}$$

- **Binomial distribution**
  - **Generalization for m outcomes out of N trials**

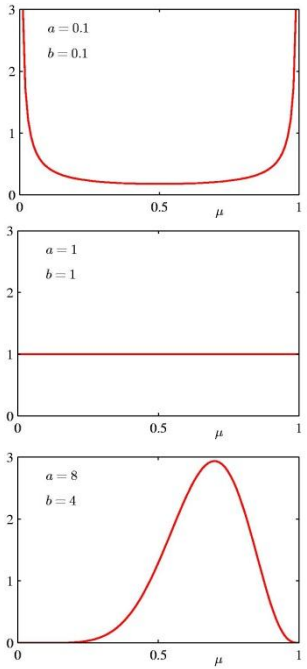


$$\text{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

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

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

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.

Multinomial Variables

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

- **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

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

Recap: Multinomial Variables

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

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

$$\mathbb{E}[m_k] = N\mu_k$$
$$\mathrm{var}[m_k] = N\mu_k(1-\mu_k)$$
$$\mathrm{cov}[m_j m_k] = -N\mu_j\mu_k$$

with the constraints

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

Slide adapted from C. Bishop

Recap: The Dirichlet Distribution

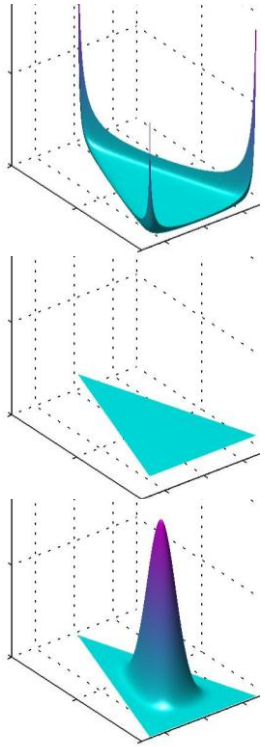
- **Dirichlet Distribution**
  - Multivariate generalization of the Beta distribution

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

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

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

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

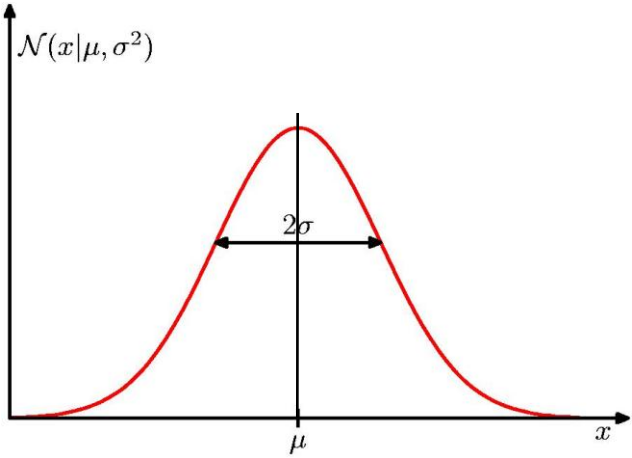


- **Properties**
  - Conjugate prior for the Multinomial.
  - The Dirichlet distribution over K variables is confined to a K-1 dimensional simplex.

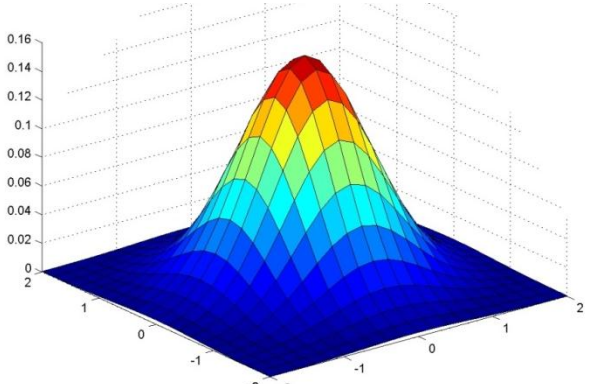
Recap: The Gaussian Distribution

- **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 $\boldsymbol{\mu}$**
  - **Covariance $\boldsymbol{\Sigma}$**



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

Image source: C.M. Bishop, 2006

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)

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}\right) \end{aligned}$$

  - From this, we can compute

$$\begin{aligned} p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \\ p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \end{aligned}$$

    where

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

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

Recap: Bayesian Inference for the Gaussian

- **Univariate conjugate priors**
  - σ^2 **known,** μ **unknown:** $p(\mu)$ **Gaussian**

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

  - μ **is known,** λ **unknown:** $p(\lambda)$ **Gamma**

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

  - **both** μ **and** λ **unknown:** $p(\mu,\lambda)$ **Gaussian-Gamma**

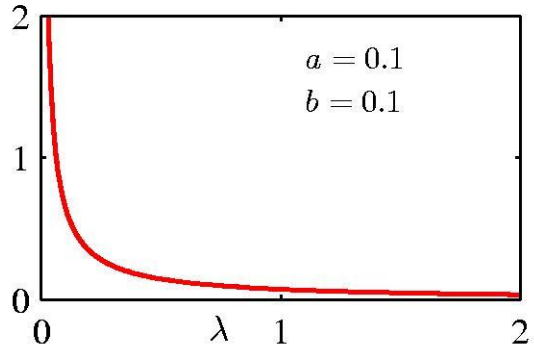
$$p(\mathbf{X}|\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}.$$

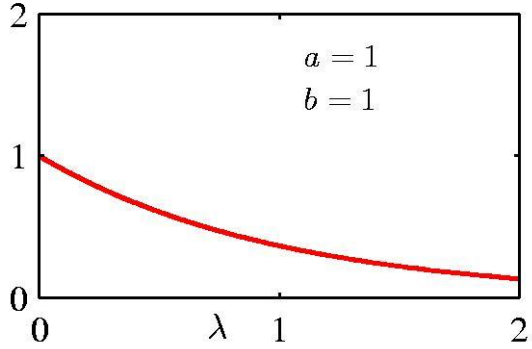
Recap: The Gamma Distribution

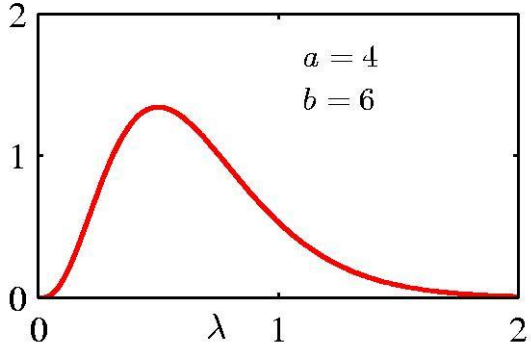
- **Gamma distribution**
  - Product of a power of λ and the exponential of a linear function of λ.

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

- **Properties**
  - Finite integral if $a>0$ and the distribution itself is finite if $a \geq 1$.
  - Moments $\mathbb{E}[\lambda] = \frac{a}{b}$ $\mathrm{var}[\lambda] = \frac{a}{b^2}$
  - Conjugate prior for a Gaussian with known μ and unknown λ.







Slide adapted from C. Bishop

Image source: C.M. Bishop, 2006

Recap: The Gaussian-Gamma Distribution

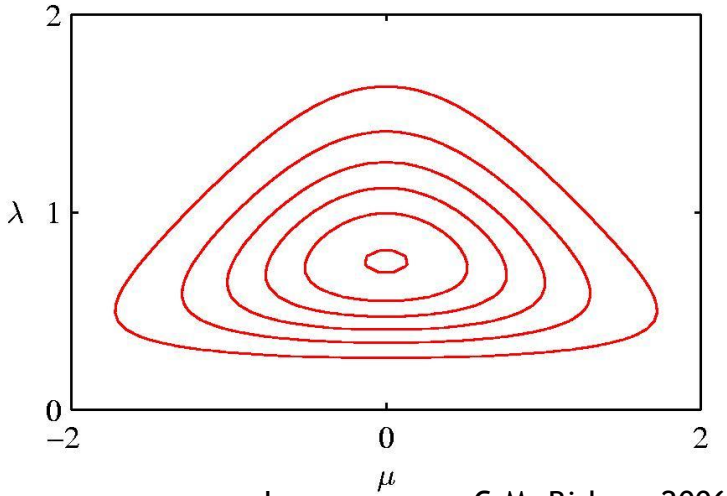
- **Gaussian-Gamma distribution**

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

$$\propto \quad \exp\left\{-\frac{\beta\lambda}{2}(\mu - \mu_0)^2\right\} \lambda^{a-1}\exp\{-b\lambda\}$$

  - Quadratic in μ.
  - Linear in λ.
  - Gamma distribution over λ.
  - Independent of μ.

- **Properties**
  - Conjugate prior for a univariate Gaussian where both μ and λ are unknown.



Slide adapted from C. Bishop

Image source: C.M. Bishop, 2006

Recap: Bayesian Inference for the Gaussian

- **Multivariate conjugate priors**
  - μ unknown, Λ known: $p(\mu)$ Gaussian.
  - Λ unknown, μ known: $p(\Lambda)$ Wishart,

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

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

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

Slide adapted from C. Bishop

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 $\mathrm{Gam}(\tau|\,a,b)$.
  - By integrating out the precision, obtain the marginal distribution

$$p(x|\mu,a,b) = \int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|a,b)\mathrm{d}\tau$$
$$= \int_0^\infty \mathcal{N}\left(x|\mu,(\eta\lambda)^{-1}\right)\mathrm{Gam}(\eta|\nu/2,\nu/2)\mathrm{d}\eta$$

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

Slide adapted from C. Bishop

Student's t-Distribution

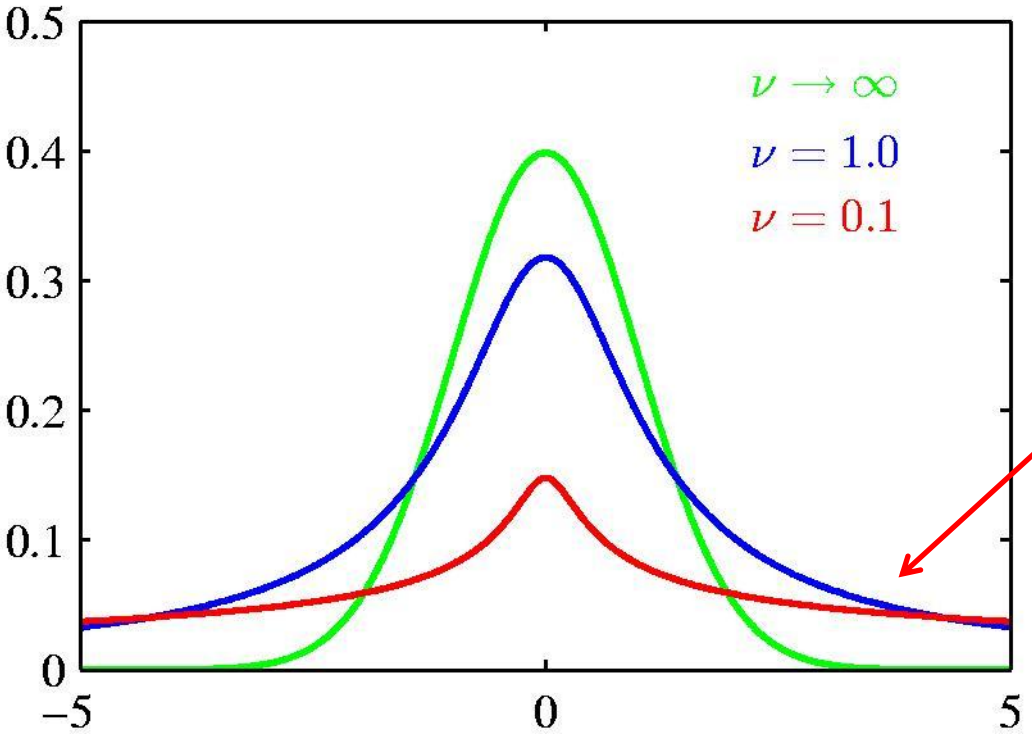
- **Student's t-Distribution**
  - We reparametrize the infinite mixture of Gaussians to get

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

- **Parameters**
  - "Precision" $\lambda = a/b$
  - "Degrees of freedom" $\nu = 2a.$

Slide adapted from C. Bishop

Student's t-Distribution: Visualization



Longer-tailed distribution!

⇒ More robust to outliers…

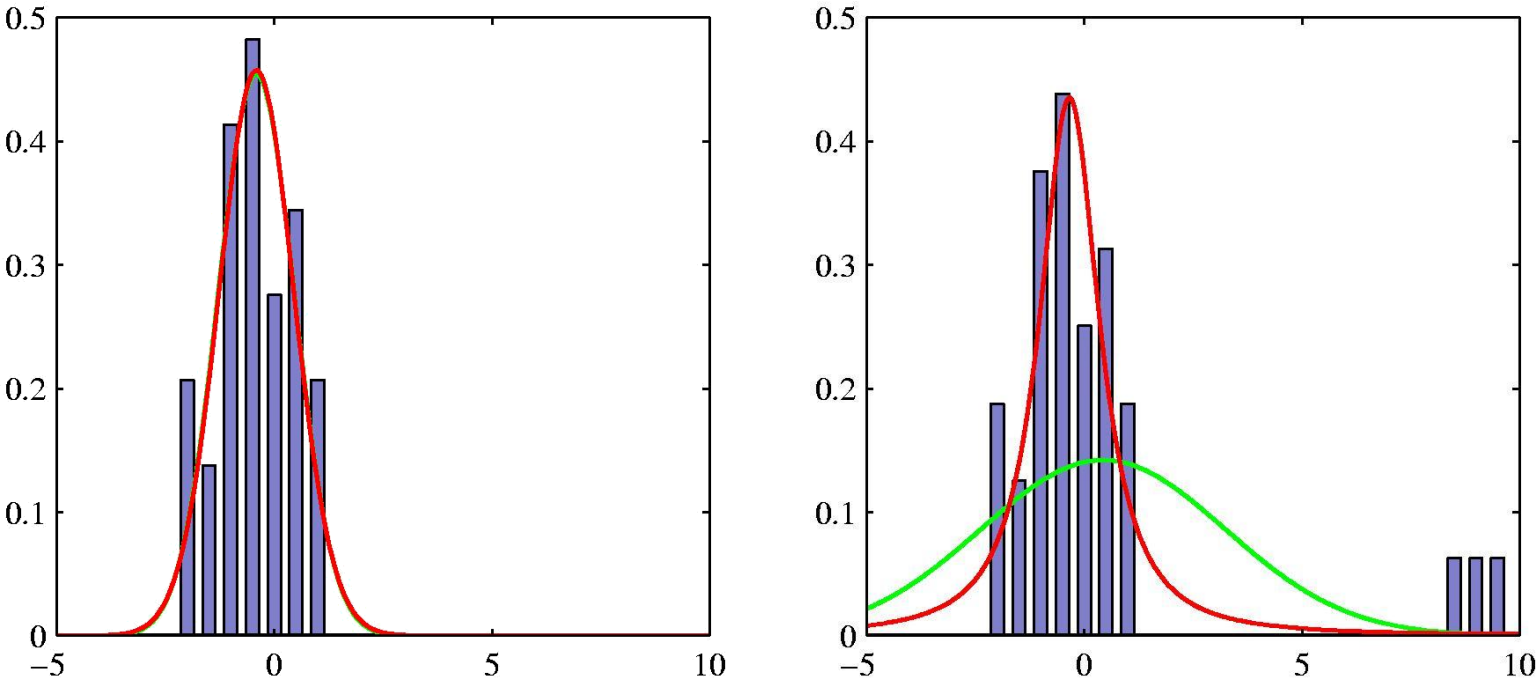
- **Behavior**

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

Image source: C.M. Bishop, 2006

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**

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

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

- **Properties**

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu}, && \text{if } \nu > 1 \\ \mathrm{cov}[\mathbf{x}] &= \frac{\nu}{(\nu-2)}\boldsymbol{\Lambda}^{-1}, && \text{if } \nu > 2 \\ \mathrm{mode}[\mathbf{x}] &= \boldsymbol{\mu}\end{aligned}$$

Slide credit: C. Bishop

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.

$\Rightarrow$ 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.

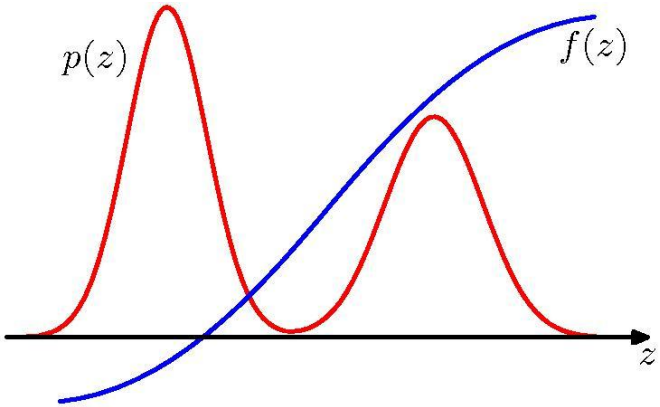
Topics of This Lecture

- 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 $l = 1,\ldots,L$ 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

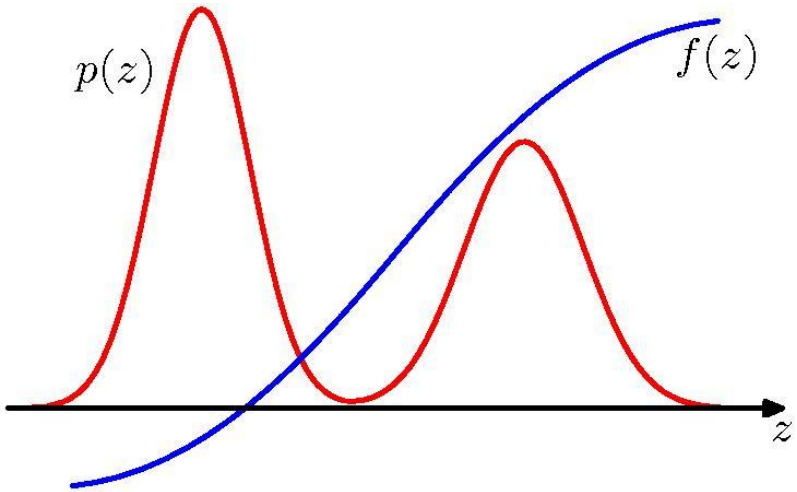
$$\mathbb{E}[\hat{f}] = \mathbb{E}[f]$$

⇒ **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.

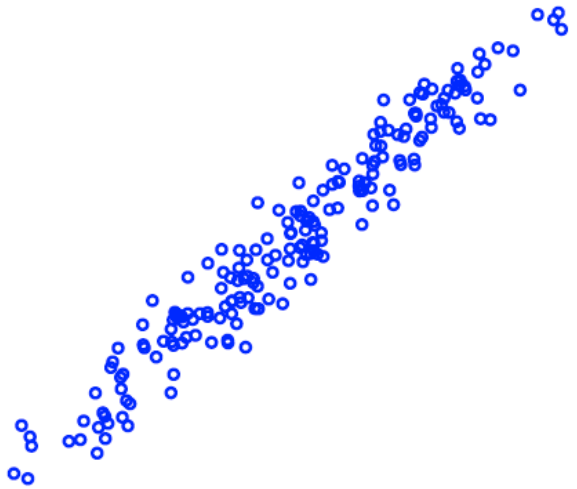
Image source: C.M. Bishop, 2006

Parametric Density Model

- **Example:**
  - **A simple multivariate (d-dimensional) Gaussian model**

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

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



Sampling from a Gaussian

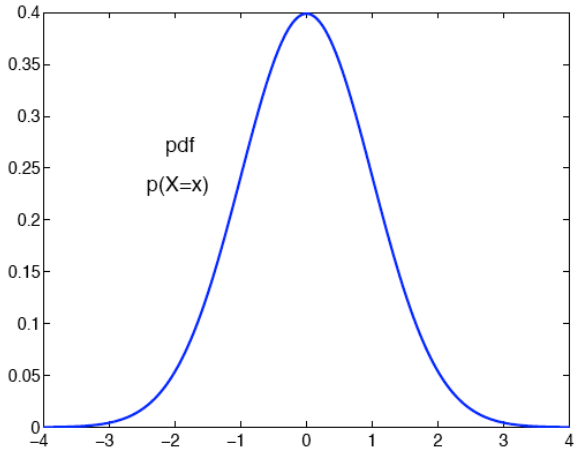
- **Given: 1-dim. Gaussian pdf (probability density function)** $p(\mathbf{x}|\mu,\sigma^2)$ **and the corresponding cumulative distribution:**

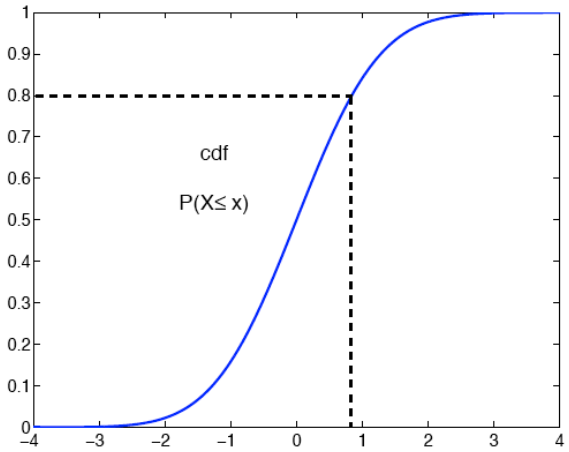
$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} p(x|\mu,\sigma^2)dx$$

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

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

$p(x|\mu,\sigma^2)$





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

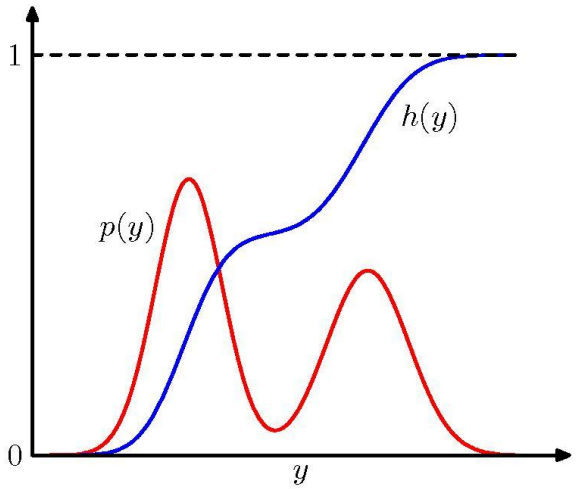
Sampling from a pdf (Transformation method)

- 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:

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



Slide credit: Bernt Schiele

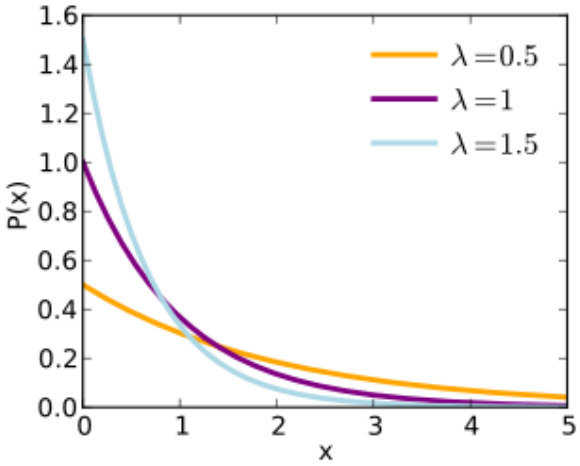
Image source: C.M. Bishop, 2006

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**

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

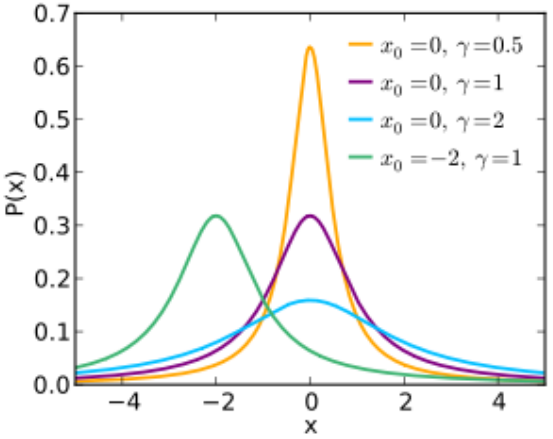
for a uniformly distributed input variable z.

Image source: Wikipedia

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.**

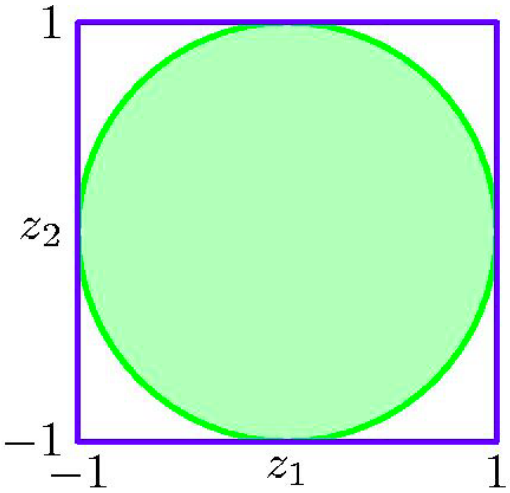
$$y = h(y)^{-1} = \tan(z)$$

for a uniformly distributed input variable z.

Image source: Wikipedia

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 $p(z_1, z_2) = 1/\pi$.

Image source: C.M. Bishop, 2006

Box-Muller Algorithm (cont'd)

- Box-Muller Algorithm (cont'd)
  - For each pair z_1, z_2 evaluate

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

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

$$p(y_1, y_2) = p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right|$$

$$= \left[\frac{1}{\sqrt{2\pi}} \exp(-y_1^2/2) \right] \left[\frac{1}{\sqrt{2\pi}} \exp(-y_2^2/2) \right]$$

  $\Rightarrow$ 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 $\sigma y + \mu \sim \mathcal{N}(\mu, \sigma^2)$.

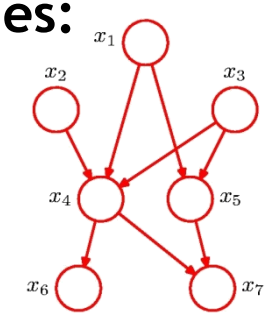
Box-Muller Algorithm (cont'd)

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

Ancestral Sampling

- Generalization of this idea to directed graphical models.
  - Joint probability factorizes into conditional probabilities:

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



- 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.

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

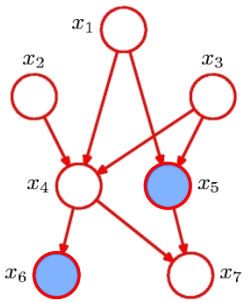
  - Cycle through each of the nodes in order and draw samples from the conditional distribution (where the parent variable is set to its sampled value).

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

Image source: C.M. Bishop, 2006

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.

$\Rightarrow$ 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

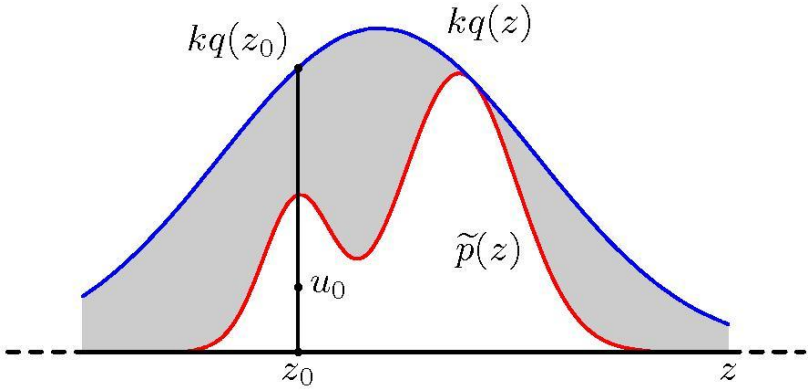
Slide adapted from Bernt Schiele

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: $\forall z : kq(z) \geq \tilde{p}(z)$

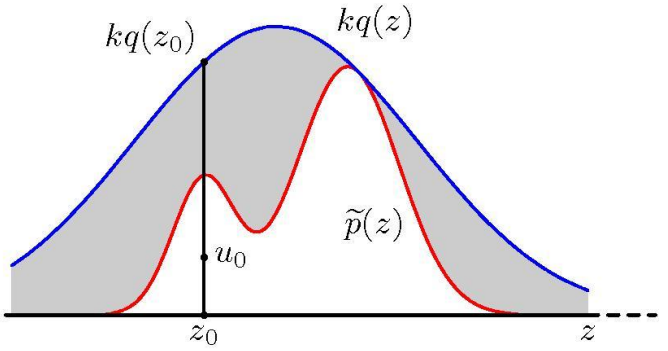


Slide credit: Bernt Schiele

Image source: C.M. Bishop, 2006

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 $\tilde{p}(z)$.

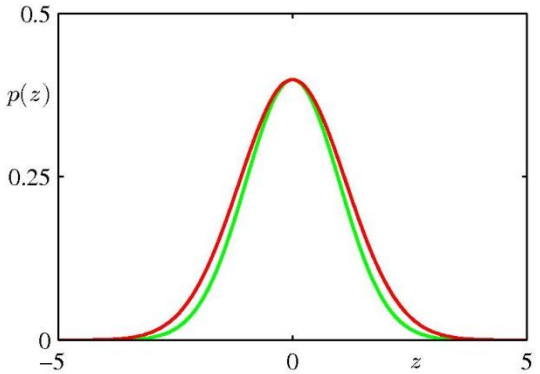
- **Discussion**
  - Original values of $\mathbf{z}$ are generated from the distribution $q(\mathbf{z})$.
  - Samples are accepted with probability $\tilde{p}(z)/kq(z)$

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

  $\Rightarrow k$ 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 $p(accept) \cdot \frac{1}{20000}$



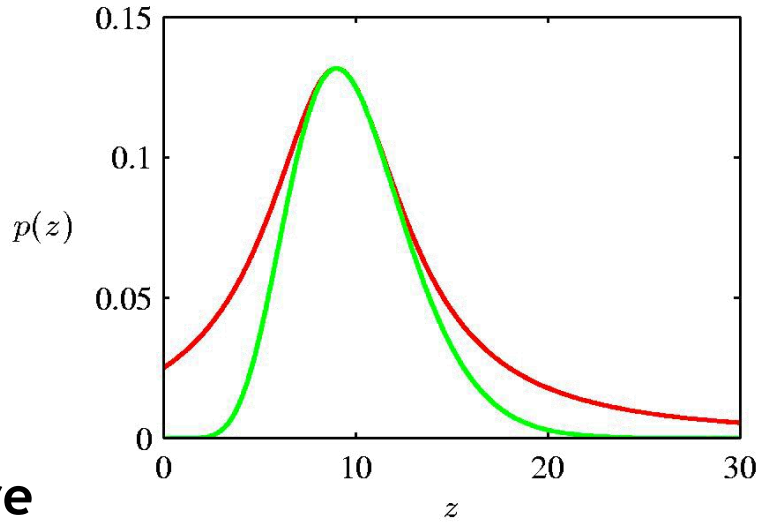
⇒ **Often impractical to find good proposal distributions for high dimensions!**

Example: Sampling from a Gamma Distrib.

- **Gamma distribution**

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

- **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

$$c = a - 1$$
$$b^2 = 2a - 1$$

Image source: C.M. Bishop, 2006

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!

Slide credit: Bernt Schiele

Importance Sampling

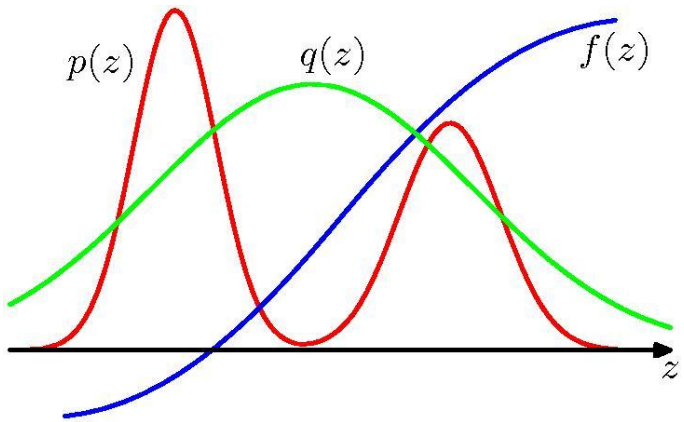
- **Idea**
  - Use a proposal distribution $q(\mathbf{z})$ from which it is easy to draw samples.
  - Express expectations in the form of a finite sum over samples $\{\mathbf{z}^{(l)}\}$ drawn from $q(\mathbf{z})$.

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

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

  - 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**
    $$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z} = \frac{Z_q}{Z_p}\int f(\mathbf{z})\frac{\tilde{p}(\mathbf{z})}{\tilde{q}(\mathbf{z})}q(\mathbf{z})d\mathbf{z}$$
    $$\simeq \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l f(\mathbf{z}^{(l)})$$
  - **with:** $\tilde{r}_l = \frac{\tilde{p}(\mathbf{z}^{(l)})}{\tilde{q}(\mathbf{z}^{(l)})}$

Slide credit: Bernt Schiele

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.

  $\Rightarrow$ 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!

Slide credit: Bernt Schiele

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**

Independent Sampling vs. Markov Chains

- So far
  - We've considered two methods, Rejection Sampling and Importance Sampling, which were both based on independent samples from $q(\mathbf{z})$.
  - However, for many problems of practical interest, it is difficult or impossible to find $q(\mathbf{z})$ with the necessary properties.

- Different approach
  - We abandon the idea of independent sampling.
  - Instead, rely on a Markov Chain to generate dependent samples from the target distribution.
  - Independence would be a nice thing, but it is not necessary for the Monte Carlo estimate to be valid.

Slide credit: Zoubin Ghahramani

MCMC – Markov Chain Monte Carlo

- Overview
  - Allows to sample from a large class of distributions.
  - Scales well with the dimensionality of the sample space.
- Idea
  - We maintain a record of the current state $\mathbf{z}^{(\tau)}$
  - The proposal distribution depends on the current state: $q(\mathbf{z} \mid \mathbf{z}^{(\tau)})$
  - The sequence of samples forms a Markov chain $\mathbf{z}^{(1)}$, $\mathbf{z}^{(2)}$,...
- Setting
  - We can evaluate $p(\mathbf{z})$ (up to some normalizing factor Z_p):

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

  - At each time step, we generate a candidate sample from the proposal distribution and accept the sample according to a criterion.

Slide credit: Bernt Schiele

MCMC – Metropolis Algorithm

- **Metropolis algorithm** [Metropolis et al., 1953]
  - Proposal distribution is symmetric: $q(\mathbf{z}_A|\mathbf{z}_B) = q(\mathbf{z}_B|\mathbf{z}_A)$
  - The new candidate sample $\mathbf{z}^*$ is accepted with probability

$$A(\mathbf{z}^\star, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\mathbf{z}^\star)}{\tilde{p}(\mathbf{z}^{(\tau)})}\right)$$

- **Implementation**
  - Choose random number u uniformly from unit interval (0,1).
  - Accept sample if $A(\mathbf{z}^\star, \mathbf{z}^{(\tau)}) > u$.
- **Note**
  - New candidate samples always accepted if $\tilde{p}(\mathbf{z}^\star) \geq \tilde{p}(\mathbf{z}^{(\tau)})$.
    - I.e. when new sample has higher probability than the previous one.
  - The algorithm sometimes accepts a state with lower probability.

Slide credit: Bernt Schiele

MCMC – Metropolis Algorithm

- Two cases
  - If new sample is accepted: $\mathbf{z}^{(\tau+1)} = \mathbf{z}^{\star}$
  - Otherwise: $\mathbf{z}^{(\tau+1)} = \mathbf{z}^{(\tau)}$
  - This is in contrast to rejection sampling, where rejected samples are simply discarded.

$\Rightarrow$ Leads to multiple copies of the same sample!

Slide credit: Bernt Schiele

MCMC – Metropolis Algorithm

- **Property**
  - When $q(\mathbf{z}_A | \mathbf{z}_B) > 0$ for all $\mathbf{z}$, the distribution of $\mathbf{z}^\tau$ tends to $p(\mathbf{z})$ as $\tau \rightarrow \infty$.
- **Note**
  - Sequence $\mathbf{z}^{(1)}$, $\mathbf{z}^{(2)}$,… is not a set of independent samples from $p(\mathbf{z})$, as successive samples are highly correlated.
  - We can obtain (largely) independent samples by just retaining every M^{th} sample.
- **Example: Sampling from a Gaussian**
  - Proposal: Gaussian with $\sigma = 0.2$.
  - Green: accepted samples
  - Red: rejected samples

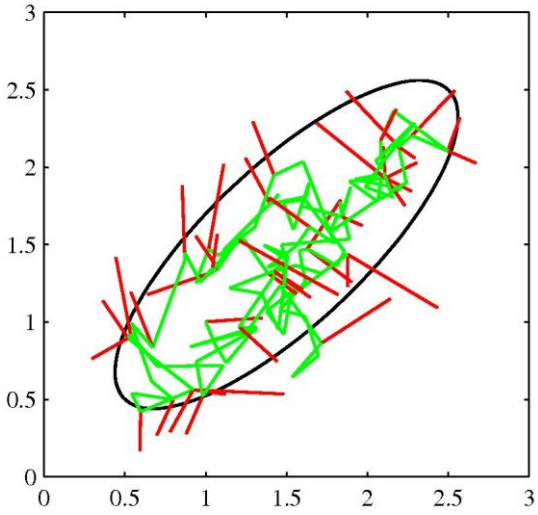


Image source: C.M. Bishop, 2006

Markov Chains

- **Question**
  - How can we show that $\mathbf{z}^{\tau}$ tends to $p(\mathbf{z})$ as $\tau \rightarrow \infty$?

- **Markov chains**
  - First-order Markov chain:

$$p\left(\mathbf{z}^{(m+1)}|\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(m)}\right) = p\left(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)}\right)$$

  - Marginal probability

$$p\left(\mathbf{z}^{(m+1)}\right) = \sum_{\mathbf{z}^{(m)}} p\left(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)}\right) p\left(\mathbf{z}^{(m)}\right)$$

Slide credit: Bernt Schiele

Markov Chains - Properties

- **Invariant distribution**
  - A distribution is said to be invariant (or stationary) w.r.t. a Markov chain if each step in the chain leaves that distribution invariant.
  - Transition probabilities:

$$T\left(\mathbf{z}^{(m)}, \mathbf{z}^{(m+1)}\right) = p\left(\mathbf{z}^{(m+1)} | \mathbf{z}^{(m)}\right)$$

  - Distribution $p^*(\mathbf{z})$ is invariant if:

$$p^\star(\mathbf{z}) = \sum_{\mathbf{z}'} T(\mathbf{z}', \mathbf{z})\, p^\star(\mathbf{z}')$$

- **Detailed balance**
  - Sufficient (but not necessary) condition to ensure that a distribution is invariant:

$$p^\star(\mathbf{z}) T(\mathbf{z}, \mathbf{z}') = p^\star(\mathbf{z}') T(\mathbf{z}', \mathbf{z})$$

  - A Markov chain which respects *detailed balance* is reversible.

Slide credit: Bernt Schiele

Ergodicity in Markov Chains

- **Remark**
  - Our goal is to use Markov chains to sample from a given distribution.
  - We can achieve this if we set up a Markov chain such that the desired distribution is invariant.
  - However, must also require that for $m \to \infty$, the distribution $p(\mathbf{z}^{(m)})$ converges to the required invariant distribution $p^*(\mathbf{z})$ irrespective of the choice of initial distribution $p(\mathbf{z}^{(0)})$. (This property is called **ergodicity**).
  - It can be shown that this is the case for a **homogeneous** Markov chain (i.e., a Markov chain for which the transition probabilities are the same for all m).

MCMC – Metropolis-Hastings Algorithm

- **Metropolis-Hastings Algorithm**
  - Generalization: Proposal distribution not required to be symmetric.
  - The new candidate sample $\mathbf{z}^*$ is accepted with probability

$$A(\mathbf{z}^\star, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\mathbf{z}^\star) q_k(\mathbf{z}^{(\tau)} | \mathbf{z}^\star)}{\tilde{p}(\mathbf{z}^{(\tau)}) q_k(\mathbf{z}^\star | \mathbf{z}^{(\tau)})}\right)$$

  - where k labels the members of the set of possible transitions considered.

- **Note**
  - When the proposal distributions are symmetric, Metropolis-Hastings reduces to the standard Metropolis algorithm.

Slide credit: Bernt Schiele

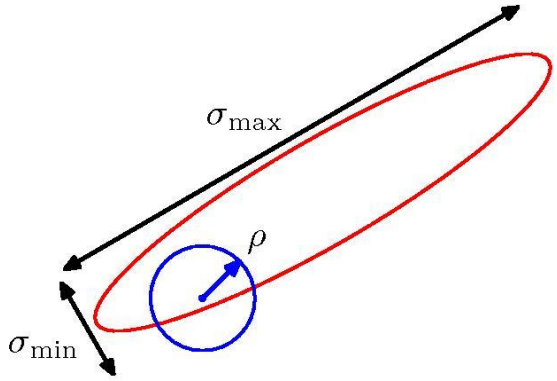
MCMC – Metropolis-Hastings Algorithm

- **Properties**
  - **We can show that $p(\mathbf{z})$ is an invariant distribution of the Markov chain defined by the Metropolis-Hastings algorithm.**
  - **We show detailed balance:**

$$\begin{aligned} p(\mathbf{z})q_k(\mathbf{z}|\mathbf{z}')A_k(\mathbf{z}',\mathbf{z}) &= \min\{p(\mathbf{z})q_k(\mathbf{z}|\mathbf{z}'), p(\mathbf{z}')q_k(\mathbf{z}'|\mathbf{z})\} \\ &= \min\{p(\mathbf{z}')q_k(\mathbf{z}'|\mathbf{z}), p(\mathbf{z})q_k(\mathbf{z}|\mathbf{z}')\} \\ &= p(\mathbf{z}')q_k(\mathbf{z}'|\mathbf{z})A_k(\mathbf{z},\mathbf{z}') \end{aligned}$$

MCMC – Metropolis-Hastings Algorithm

- Schematic illustration
  - For continuous state spaces, a common choice of proposal distribution is a Gaussian centered on the current state.



  ⇒ What should be the variance of the proposal distribution?
    - Large variance: rejection rate will be high for complex problems.
    - The scale ρ of the proposal distribution should be as large as possible without incurring high rejection rates.

    ⇒ ρ should be of the same order as the smallest length scale $\sigma_{\min}$.

  - This causes the system to explore the distribution by means of a random walk.
    - Undesired behavior: number of steps to arrive at state that is independent of original state is of order $(\sigma_{\max}/\sigma_{\min})^2$.
    - Strong correlations can slow down the Metropolis algorithm!

Image source: C.M. Bishop, 2006

Gibbs Sampling

- Approach
  - MCMC-algorithm that is simple and widely applicable.
  - May be seen as a special case of Metropolis-Hastings.
- Idea
  - Sample variable-wise: replace $\mathbf{z}_i$ by a value drawn from the distribution $p(z_i | \mathbf{z}_{\setminus i})$.
    - This means we update one coordinate at a time.
  - Repeat procedure either by cycling through all variables or by choosing the next variable.

Slide credit: Bernt Schiele

Gibbs Sampling

- Example
  - Assume distribution $p(z_1, z_2, z_3)$.
  - Replace $z_1^{(\tau)}$ with new value drawn from $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)})$
  - Replace $z_2^{(\tau)}$ with new value drawn from $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)})$
  - Replace $z_3^{(\tau)}$ with new value drawn from $z_3^{(\tau+1)} \sim p(z_3 | z_1^{(\tau+1)}, z_2^{(\tau+1)})$
  - And so on…

Slide credit: Bernt Schiele

Gibbs Sampling

- Properties
  - The factor that determines the acceptance probability in the Metropolis-Hastings is determined by

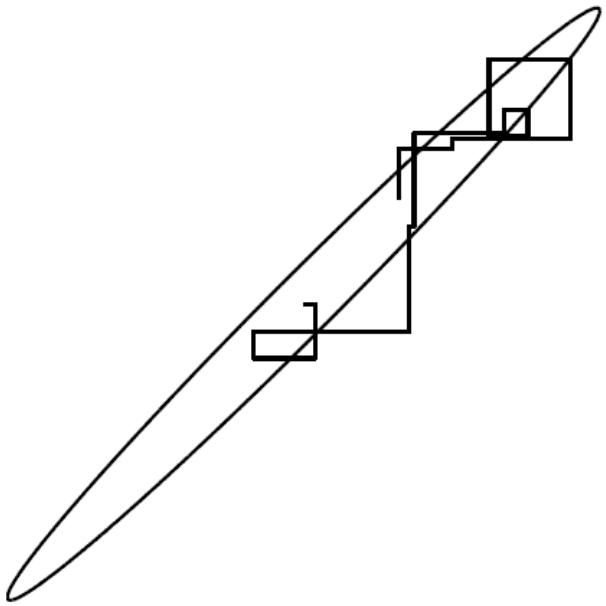
$$A(\mathbf{z}^\star, \mathbf{z}) = \frac{p(\mathbf{z}^\star) q_k(\mathbf{z}|\mathbf{z}^\star)}{p(\mathbf{z}) q_k(\mathbf{z}^\star|\mathbf{z})} = \frac{p(z_k^\star|\mathbf{z}_{\setminus k}^\star) p(\mathbf{z}_{\setminus k}^\star) p(z_k^\star|\mathbf{z}_{\setminus k}^\star)}{p(z_k|\mathbf{z}_{\setminus k}) p(\mathbf{z}_{\setminus k}) p(z_k|\mathbf{z}_{\setminus k})} = 1$$

  - I.e. we get an algorithm which always accepts!
  - If you can compute (and sample from) the conditionals, you can apply Gibbs sampling.
  - The algorithm is completely parameter free.
  - Can also be applied to subsets of variables.

Slide credit: Zoubin Ghahramani

Gibbs Sampling

- Example
  - 20 iterations of Gibbs sampling on a bivariate Gaussian.



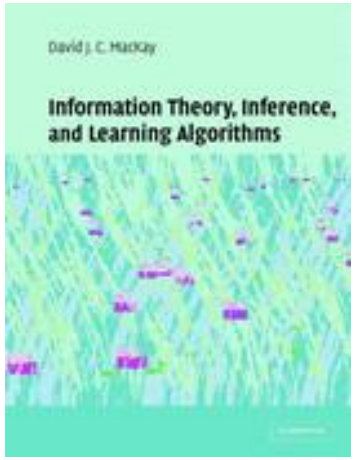
  - Note: strong correlations can slow down Gibbs sampling.

Summary: Approximate Inference

- Exact Bayesian Inference often intractable.
- Rejection and Importance Sampling
  - Generate independent samples.
  - Impractical in high-dimensional state spaces.
- Markov Chain Monte Carlo (MCMC)
  - Simple & effective (even though typically computationally expensive).
  - Scales well with the dimensionality of the state space.
  - Issues of convergence have to be considered carefully.
- Gibbs Sampling
  - Used extensively in practice.
  - Parameter free
  - Requires sampling conditional distributions.

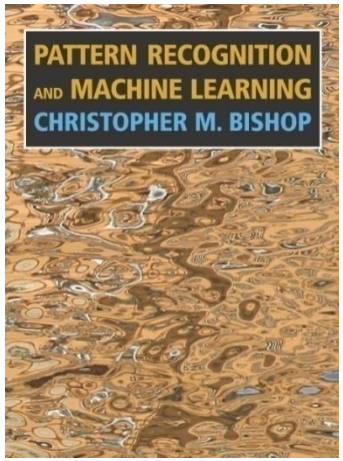
References and Further Reading

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



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

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



- 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)