Advanced Machine Learning
Lecture 6

Approximate Inference

10.11.2016

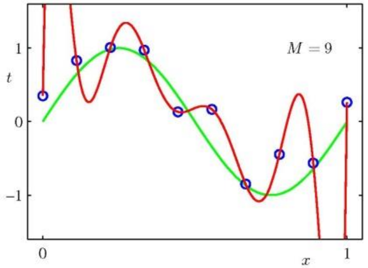
Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de/

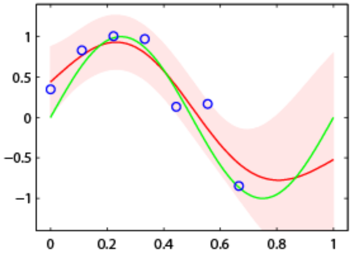
leibe@vision.rwth-aachen.de

This Lecture: *Advanced Machine Learning*

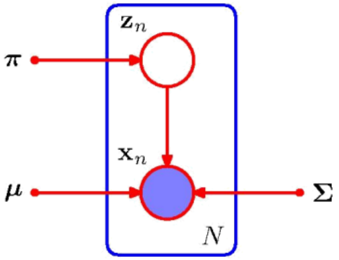
- Regression Approaches
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Gaussian Processes

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

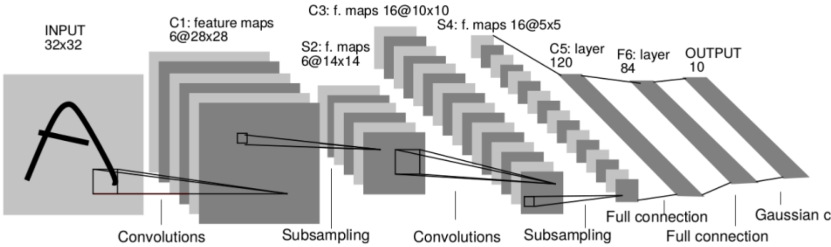




- Learning with Latent Variables
  - Probability Distributions
  - Approximate Inference



- Deep Learning
  - Neural Networks
  - CNNs, RNNs, ResNets, etc.



Recap: GPs with Noise-free Observations

- Assume our observations are noise-free:

$$\{(\mathbf{x}_n, f_n) \mid n = 1, \ldots, N\}$$

  - Joint distribution of the training outputs $\mathbf{f}$ and test outputs $\mathbf{f}_*$ according to the prior:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_\star \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) & K(X,X_\star) \\ K(X_\star,X) & K(X_\star,X_\star) \end{bmatrix}\right)$$

  - Calculation of posterior corresponds to conditioning the joint Gaussian prior distribution on the observations:

$$\mathbf{f}_\star | X_\star, X, \mathbf{f} \sim \mathcal{N}(\bar{\mathbf{f}}_\star, \operatorname{cov}[\mathbf{f}_\star]) \qquad \bar{\mathbf{f}}_\star = \mathbb{E}[\mathbf{f}_\star | X, X_\star, \mathbf{f}]$$

  - with:

$$\begin{aligned} \bar{\mathbf{f}}_\star &= K(X_\star, X)K(X,X)^{-1}\mathbf{f} \\ \operatorname{cov}[\mathbf{f}_\star] &= K(X_\star, X_\star) - K(X_\star, X)K(X,X)^{-1}K(X, X_\star) \end{aligned}$$

Slide adapted from Bernt Schiele

Recap: GPs with Noisy Observations

- Joint distribution of the observed values and the test locations under the prior:

$$\begin{bmatrix} \mathbf{t} \\ \mathbf{f}_\star \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_\star) \\ K(X_\star,X) & K(X_\star,X_\star) \end{bmatrix}\right)$$

  - Calculation of posterior corresponds to conditioning the joint Gaussian prior distribution on the observations:

$$\mathbf{f}_\star | X_\star, X, \mathbf{t} \sim \mathcal{N}(\bar{\mathbf{f}}_\star, \mathrm{cov}[\mathbf{f}_\star]) \qquad \bar{\mathbf{f}}_\star = \mathbb{E}[\mathbf{f}_\star | X, X_\star, \mathbf{t}]$$

  - with:

$$\bar{\mathbf{f}}_\star = K(X_\star, X)\left(K(X,X) + \sigma_n^2 I\right)^{-1} \mathbf{t}$$

$$\mathrm{cov}[\mathbf{f}_\star] = K(X_\star, X_\star) - K(X_\star, X)\left(K(X,X) + \sigma_n^2 I\right)^{-1} K(X, X_\star)$$

⇒ This is the key result that defines Gaussian process regression!

- Predictive distribution is Gaussian whose mean and variance depend on test points X_* and on the kernel $k(\mathbf{x},\mathbf{x}')$, evaluated on X.

Slide adapted from Bernt Schiele

Recap: Bayesian Model Selection for GPs

- Goal
  - Determine/learn different parameters of Gaussian Processes
- Hierarchy of parameters
  - Lowest level
    - $\mathbf{w}$ - e.g. parameters of a linear model.
  - Mid-level (hyperparameters)
    - θ - e.g. controlling prior distribution of $\mathbf{w}$.
  - Top level
    - Typically discrete set of model structures $\mathcal{H}_i$.
- Approach
  - Inference takes place one level at a time.

Slide credit: Bernt Schiele

Recap: Model Selection at Lowest Level

- Posterior of the parameters $\mathbf{w}$ is given by Bayes' rule

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

- with
  - $p(\mathbf{t}\,|\,X,\mathbf{w},\mathcal{H}_i)$ likelihood and
  - $p(\mathbf{w}\,|\,\theta,\mathcal{H}_i)$ prior parameters $\mathbf{w}$,
  - Denominator (normalizing constant) is independent of the parameters and is called marginal likelihood.

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

Slide credit: Bernt Schiele

Recap: Model Selection at Mid Level

- Posterior of parameters θ is again given by Bayes' rule

$$\begin{aligned} p(\theta|\mathbf{t}, X, \mathcal{H}_i) &= \frac{p(\mathbf{t}|X, \theta, \mathcal{H}_i)p(\theta|X, \mathcal{H}_i)}{p(\mathbf{t}|X, \mathcal{H}_i)} \\ &= \frac{p(\mathbf{t}|X, \theta, \mathcal{H}_i)p(\theta|\mathcal{H}_i)}{p(\mathbf{t}|X, \mathcal{H}_i)} \end{aligned}$$

- where
  - The marginal likelihood of the previous level $p(\mathbf{t}\,|\,X,\theta,\mathcal{H}_i)$ plays the role of the likelihood of this level.
  - $p(\theta\,|\,\mathcal{H}_i)$ is the hyperprior (prior of the hyperparameters)
  - Denominator (normalizing constant) is given by:

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

Slide credit: Bernt Schiele

Recap: Model Selection at Top Level

- At the top level, we calculate the posterior of the model

$$p(\mathcal{H}_i|\mathbf{t}, X) \;=\; \frac{p(\mathbf{t}|X, \mathcal{H}_i)p(\mathcal{H}_i)}{p(\mathbf{t}|X)}$$

- where
  - Again, the denominator of the previous level $p(\mathbf{t}\,|\,X,\mathcal{H}_i)$ plays the role of the likelihood.
  - $p(\mathcal{H}_i)$ is the prior of the model structure.
  - Denominator (normalizing constant) is given by:

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

Slide credit: Bernt Schiele

Recap: 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.

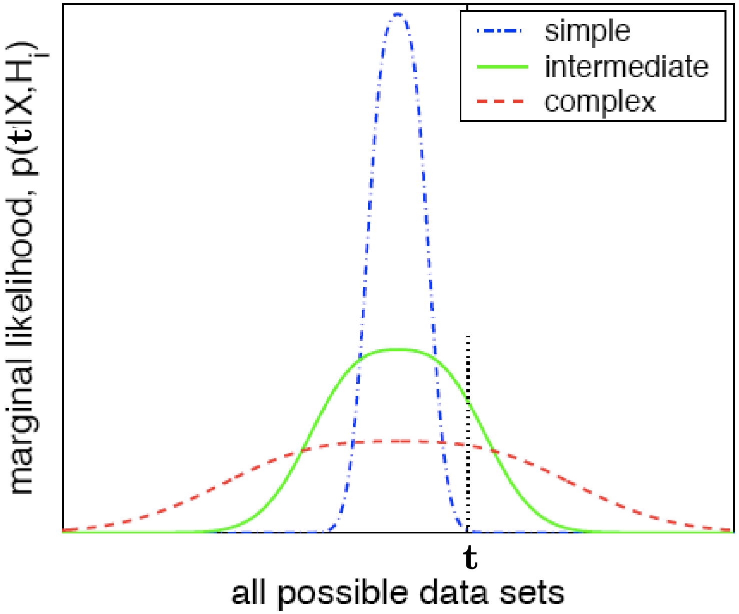


Image source: Rasmussen & Williams, 2006

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

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

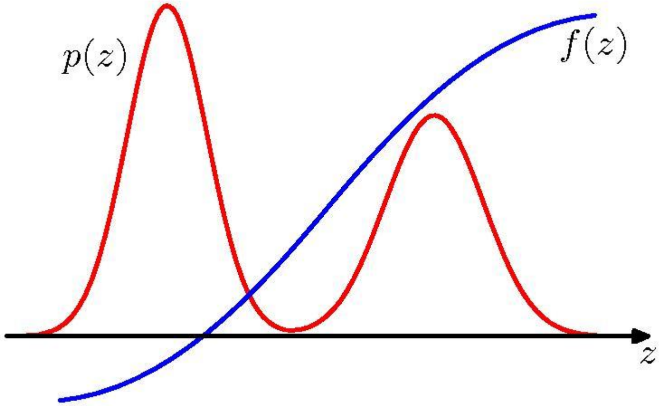
- Sampling approaches
  - Sampling from a distribution
  - Ancestral Sampling
  - Rejection Sampling
  - Importance Sampling

- 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,…,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 $\mathbb{E}[\hat{f}] = \mathbb{E}[f]$

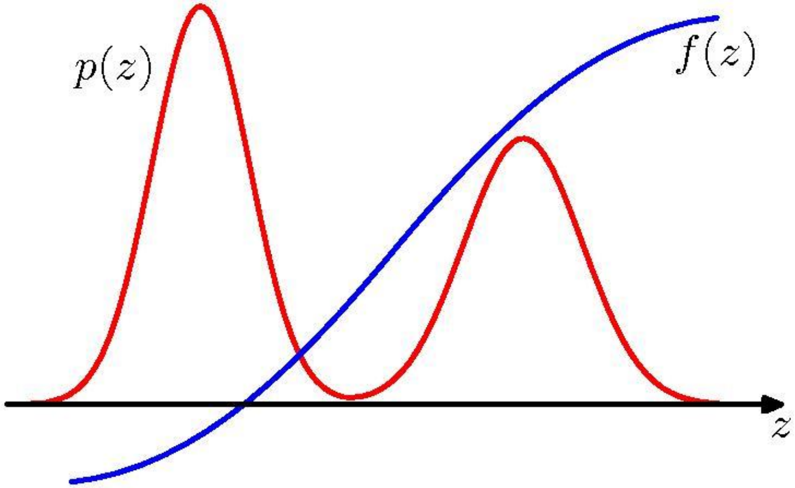
$\Rightarrow$ **Unbiased estimate, independent** of the dimension of $\mathbf{z}$!

Slide adapted from Bernt Schiele

Image source: C.M. Bishop, 2006

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.

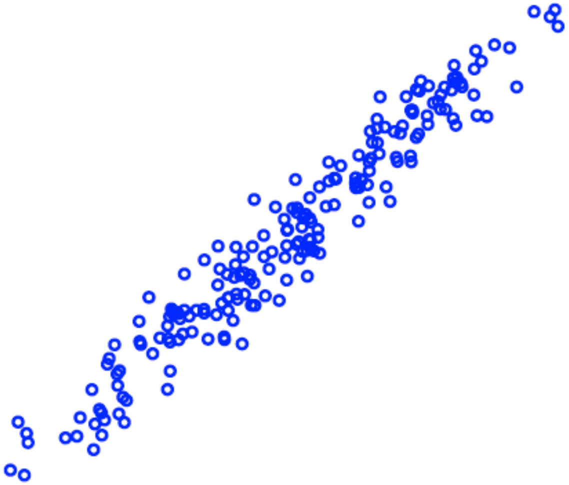
Image source: C.M. Bishop, 2006

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.**



Slide adapted from Bernt Schiele

Sampling from a Gaussian

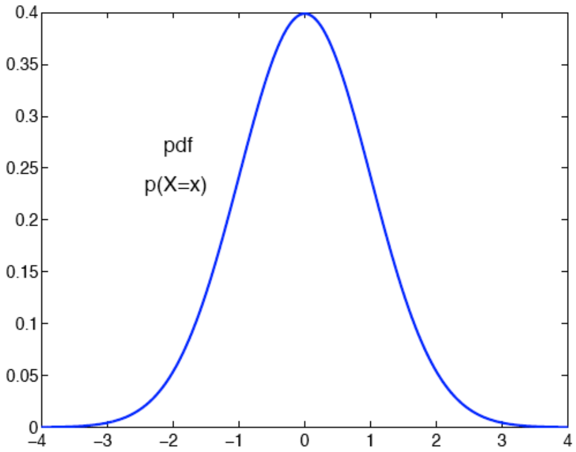
- **Given: 1-dim. Gaussian pdf (probability density function) $p(\mathbf{x}|\mu,\sigma^2)$ and the corresponding cumulative distribution:**

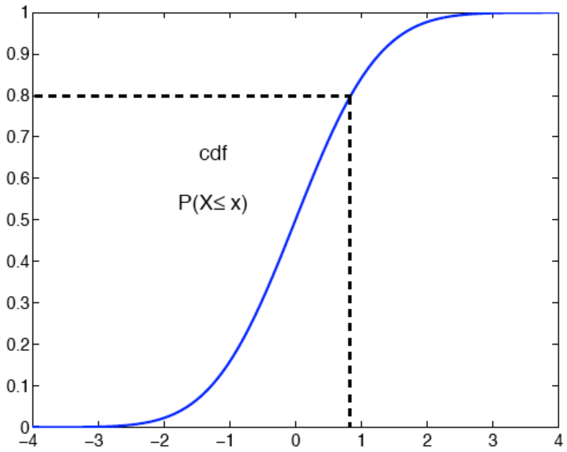
$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} p(x|\mu,\sigma^2)dx$$

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

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

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





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

Slide credit: Bernt Schiele

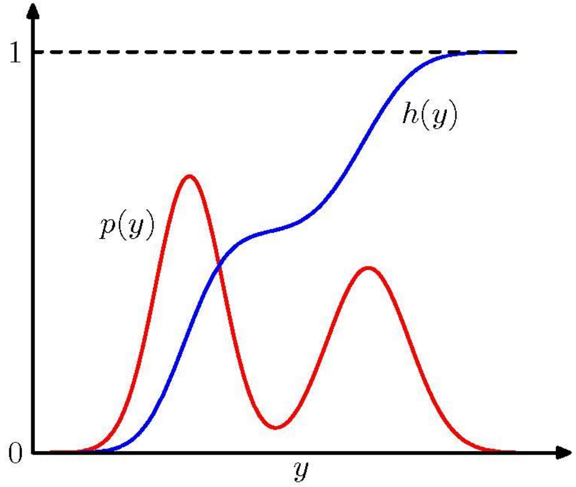
Sampling from a pdf (Transformation method)

- 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

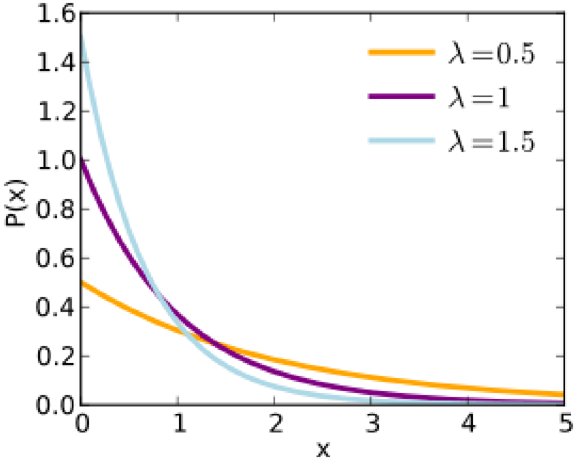
Image source: C.M. Bishop, 2006

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

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

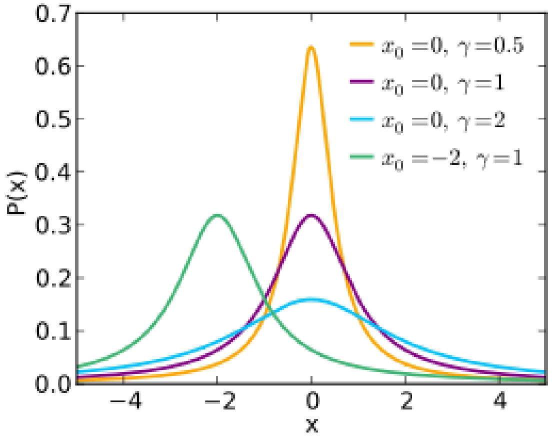
for a uniformly distributed input variable z.

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.

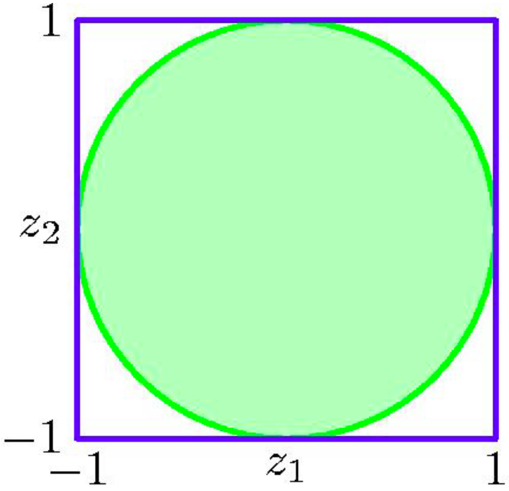
$$y = h(y)^{-1} = \tan(z)$$

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

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

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

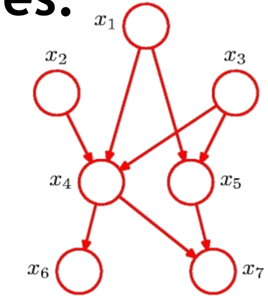
Box-Muller Algorithm (cont'd)

- 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:

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



- 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.

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

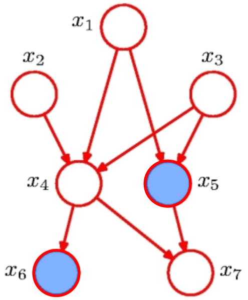
  - Cycle through each of the nodes in order and draw samples from the conditional distribution (where the parent variable is set to its sampled value).

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

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.

  ⇒ 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

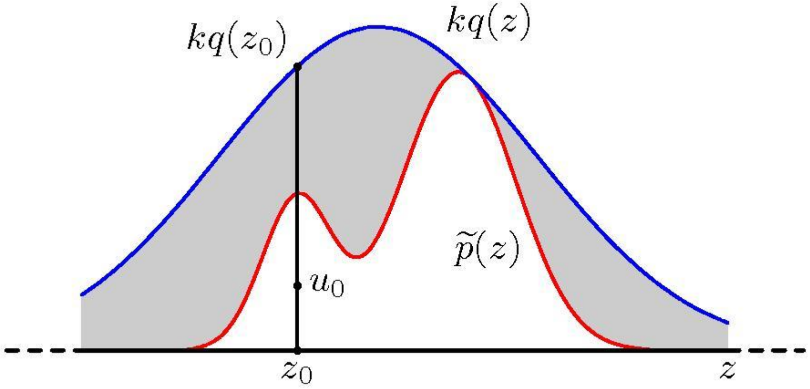
Slide adapted from Bernt Schiele

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

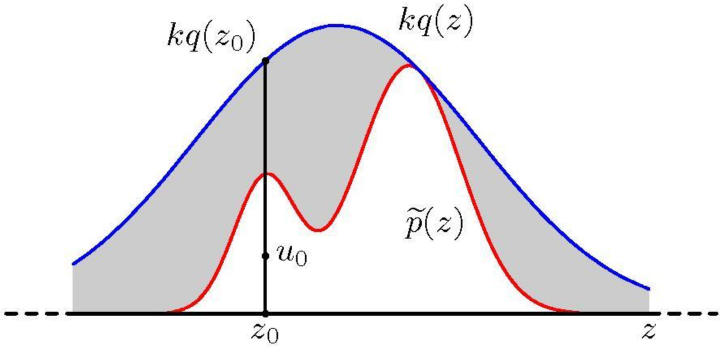


Slide credit: Bernt Schiele

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!

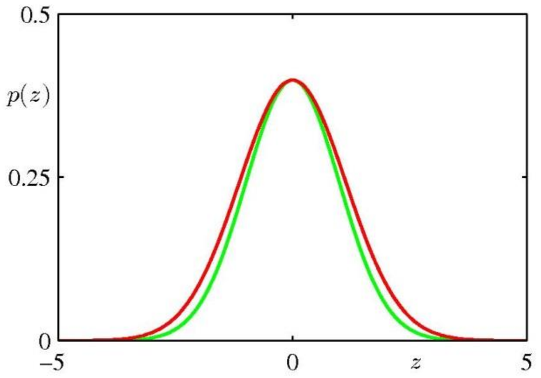
Slide credit: Bernt Schiele

Image source: C.M. Bishop, 2006

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 $p(accept) \cdot \dfrac{1}{20000}$



$\Rightarrow$ **Often impractical to find good proposal distributions for high dimensions!**

Slide credit: Bernt Schiele

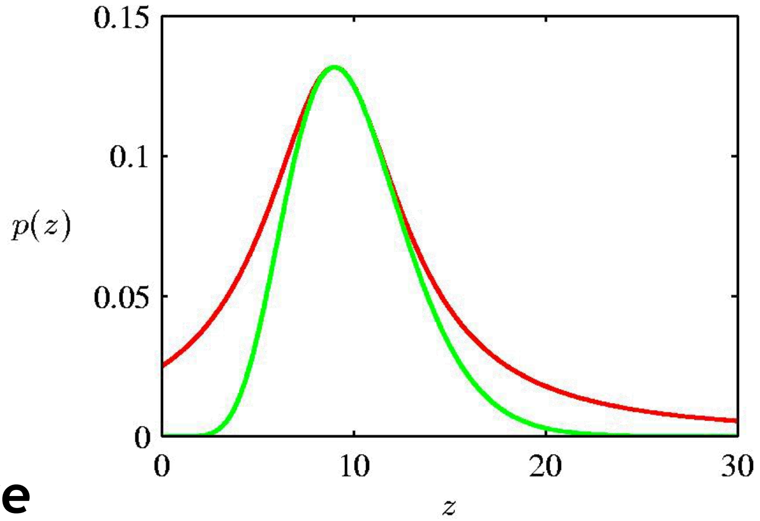
Image source: C.M. Bishop, 2006

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

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

Image source: C.M. Bishop, 2006

Importance Sampling

- Approach
  - Approximate expectations directly (but does <u>not</u> 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

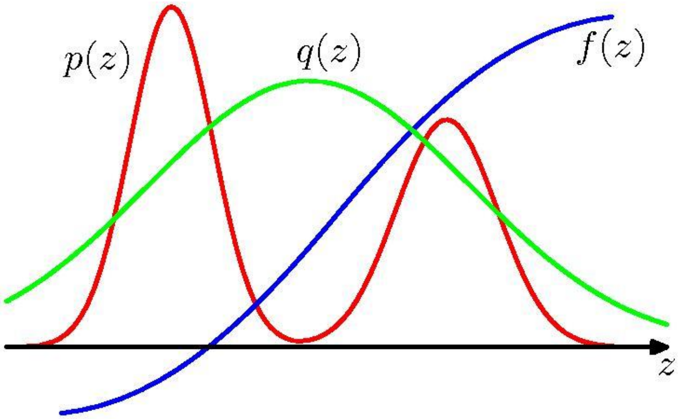
Importance Sampling

- **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 $q(\mathbf{z})$.

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

  - with importance weights

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



Slide credit: Bernt Schiele

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

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

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

Slide credit: Bernt Schiele

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!

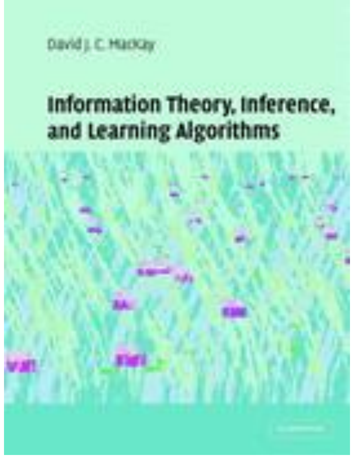
Slide credit: Bernt Schiele

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

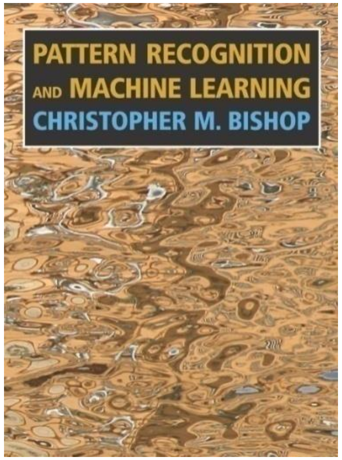
References and Further Reading

- 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)