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

Recap: The Beta Distribution

- Beta distribution
  - Distribution over \( \mu \in [0, 1] \):
    \[
    \text{Beta}(\mu | a, b) = \frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}
    \]
  - \( \mathbb{E}[\mu] = \frac{a}{a+b} \)
  - \( \text{var}[\mu] = \frac{ab}{(a+b)^2 (a+b+1)} \)

- 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: The Dirichlet Distribution

- **Dirichlet Distribution**
  - Multivariate generalization of the Beta distribution
  \[
  \text{Dir}(\mathbf{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \alpha_k^{\alpha_k - 1}
  \]
  - \(\alpha = \sum_{i=1}^{K} \alpha_k\)
  - \(E[\mu_k] = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k}\)
  - \(\text{var}[\mu_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}\)
  - \(\text{cov}[\mu_j, \mu_k] = -\frac{\alpha_j \alpha_k}{\alpha_0^2(\alpha_0 + 1)}\)

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

- **Multi-dimensional case**
  - Mean \(\mu\)
  - Covariance \(\Sigma\)
  \[\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}\Sigma^{1/2}} \exp\left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\}\]

Recap: Bayes’ Theorem for Gaussian Variables

- **Marginal and Conditional Gaussians**
  - Suppose we are given a Gaussian prior \(p(x)\) and a Gaussian conditional distribution \(p(y|x)\) (a linear Gaussian model)
  \[p(x) = \mathcal{N}(x|\mu, \Lambda^{-1})\]
  \[p(y|x) = \mathcal{N}(y|Ax + b, L^{-1})\]
  - From this, we can compute
  \[p(y) = \mathcal{N}(y|A\mu + b, L^{-1} + AA^{-1}A^T)\]
  \[p(x|y) = \mathcal{N}(x|\Sigma(\Lambda + \Lambda^T \Sigma^{-1} A^T)\Sigma^{-1}(y - b) + \Lambda \mu), \Sigma)\]
  where
  \[\Sigma = (\Lambda + \Lambda^T \Sigma^{-1} A^T)\Sigma^{-1}\]
  \(\Rightarrow\) Closed-form solution for (Gaussian) marginal and posterior.

Maximum Likelihood for the Gaussian

- **Maximum Likelihood**
  - Given i.i.d. data \(X = (x_1, \ldots, x_N)^T\), the log likelihood function is given by
  \[\log p(X|\mu, \Sigma) = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma|\]
  - \(-\frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu)\)
  - **Sufficient statistics**
    - The likelihood depends on the data set only through
    \[\sum_{n=1}^{N} x_n, \sum_{n=1}^{N} x_n x_n^T\]
    - Those are the sufficient statistics for the Gaussian distribution.

ML for the Gaussian

- **Setting the derivative to zero**
  \[\frac{\partial}{\partial \mu} \ln p(X|\mu, \Sigma) = \sum_{n=1}^{N} \Sigma^{-1} (x_n - \mu) = 0\]
  - Solve to obtain
    \[\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n\]
  - And similarly, but a bit more involved
    \[\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T\]

- **Comparison with true results**
  - Under the true distribution, we obtain
    \[E[\mu_{ML}] = \mu, E[\Sigma_{ML}] = \frac{N-1}{N} \Sigma\]
  - The ML estimate for the covariance is biased and underestimates the true covariance.
  - Therefore define the following unbiased estimator
    \[\hat{\Sigma} = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T\].
Bayesian Inference for the Gaussian

- Let’s begin with a simple example.
  - Consider a single Gaussian random variable $x$.
  - Assume $\sigma^2$ is known and the task is to infer the mean $\mu$.
  - Given i.i.d. data $X = (x_1, \ldots, x_N)^T$, the likelihood function for $\mu$ is given by
  $$p(X|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2\right\}$$
  - The likelihood function has a Gaussian shape as a function of $\mu$.
  - The conjugate prior for this case is again a Gaussian.
  $$p(\mu) = N(\mu|\mu_0, \sigma_0^2)$$

Bayesian Inference for the Gaussian

- Combined with a Gaussian prior over $\mu$
  $$p(\mu) = N(\mu|\mu_0, \sigma_0^2)$$
  - This results in the posterior
  $$p(\mu|X) \propto p(\mu)p(X|\mu)$$
  - Completing the square over $\mu$, we can derive that
  $$p(\mu|X) = N(\mu|\mu_N, \sigma_N^2)$$
  where
  $$\mu_N = \frac{\sigma_0^2 \mu_0 + N\bar{x}}{\sigma_0^2 + N\sigma^2}$$
  $$\sigma_N^2 = \frac{1}{\sigma_0^2 + N\sigma^2}$$

Visualization of the Results

- Bayes estimate:
  $$\mu_N = \frac{\sigma_0^2 \mu_0 + N\bar{x}}{\sigma_0^2 + N\sigma^2}$$
  $$\sigma_N^2 = \frac{1}{\sigma_0^2 + N\sigma^2}$$

  - Behavior for large $N$
    $$\mu_N \rightarrow \mu_0 \quad \sigma_N^2 \rightarrow 0$$
    - $N = 0$ 
    - $N = 10$
    - $N = 20$
    - $N = 100$

The Gamma Distribution

- Gamma distribution
  $$\text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$
- Properties
  - Finite integral if $a>0$ and the distribution itself is finite if $a\geq 1$.
  - Moments
    $$\mathbb{E}[\lambda] = \frac{a}{b} \quad \text{var}[\lambda] = \frac{a}{b^2}$$
  - Visualization

Bayesian Inference for the Gaussian

- More complex case
  - Now assume $\mu$ is known and the precision $\lambda$ shall be inferred.
  - The likelihood function for $\lambda = 1/\sigma^2$ is given by
  $$p(X|\lambda) = \prod_{n=1}^{N} N(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2\right\}$$
  - This has the shape of a Gamma function of $\lambda$.

Bayesian Inference for the Gaussian

- Bayesian estimation
  - Combine a Gamma prior $\text{Gam}(\lambda|a_0, b_0)$ with the likelihood function to obtain
  $$p(\lambda|X) \propto \lambda^{a_0 - 1} \lambda^{N/2} \exp\left\{-b_0 - \frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2\right\}$$
  - We recognize this again as a Gamma function with
    $$a_N = a_0 + N$$
    $$b_N = b_0 + \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^2$$
Bayesian Inference for the Gaussian

- Even more complex case
  - Assume that both $\mu$ and $\lambda$ are unknown
  - The joint likelihood function is given by
    $$p(X|\mu, \lambda) = \prod_{n=1}^{N} \frac{1}{2\pi} \left( \frac{\lambda}{2} \right)^{1/2} \exp \left\{ -\frac{\lambda}{2}(x_n - \mu)^2 \right\}$$
  - $$\propto \left[ \frac{\lambda^{1/2}}{2\pi} \exp \left\{ -\frac{\lambda \mu^2}{2} \right\} \right]^N \exp \left\{ \frac{\lambda}{2} \sum_{n=1}^{N} x_n - \frac{\lambda}{2} \sum_{n=1}^{N} \sigma_n^2 \right\}.$$  

  $\Rightarrow$ Need a prior with the same functional dependence on $\mu$ and $\lambda$.

Bayesian Inference for the Gaussian

- Multivariate conjugate priors
  - $\mu$ unknown, $\Lambda$ known: $p(\mu)$ Gaussian.
  - $\Lambda$ unknown, $\mu$ known: $p(\Lambda)$ Wishart,
    $$W(\Lambda|W, \nu) = B(\Lambda^{(\nu-D)/2} \exp \left\{ -\frac{1}{2} \text{Tr}(W^{-1} \Lambda) \right\}).$$
  - $\Lambda$ and $\mu$ unknown: $p(\mu, \Lambda)$ Gaussian-Wishart,
    $$p(\mu, \Lambda|\mu_0, \beta, W, \nu) = \mathcal{N}(\mu|\mu_0, (\beta\Lambda)^{-1})W(\Lambda|W, \nu).$$

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
    $$p(x|\mu, a, b) = \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a,b) d\tau$$
    $$= \int_0^\infty \mathcal{N} \left( x|\mu, (\eta \lambda)^{-1} \right) \text{Gam}(\eta|\nu/2, \nu/2) d\eta$$
  - This corresponds to an infinite mixture of Gaussians having the same mean, but different precision.

The Gaussian-Gamma Distribution

- Gaussian-Gamma distribution
  $$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\text{Gam}(\lambda|a, b)$$
  $$\propto \exp \left\{ -\frac{\beta\lambda}{2}(\mu - \mu_0)^2 \right\} \lambda^{\nu-1} \exp \left\{ -b\lambda \right\}$$
  - Quadratic in $\mu$.
  - Linear in $\lambda$.

Recap: Bayesian Inference for the Gaussian

- Multivariate conjugate priors
  - $\mu$ unknown, $\Lambda$ known: $p(\mu)$ Gaussian.
  - $\Lambda$ unknown, $\mu$ known: $p(\Lambda)$ Wishart,
    $$W(\Lambda|W, \nu) = B(\Lambda^{(\nu-D)/2} \exp \left\{ -\frac{1}{2} \text{Tr}(W^{-1} \Lambda) \right\}).$$
  - $\Lambda$ and $\mu$ unknown: $p(\mu, \Lambda)$ Gaussian-Wishart,
    $$p(\mu, \Lambda|\mu_0, \beta, W, \nu) = \mathcal{N}(\mu|\mu_0, (\beta\Lambda)^{-1})W(\Lambda|W, \nu).$$

Student’s t-Distribution

- Student’s t-Distribution
  - We reparametrize the infinite mixture of Gaussians to get
    $$St(x|\mu, \lambda, \nu) = \frac{\Gamma((\nu+2)/2)}{\Gamma(\nu/2)} \left( \frac{\lambda}{\nu} \right)^{1/2} \left[ 1 + \frac{\lambda(x - \mu)^2}{\nu} \right]^{-(\nu+2)/2}$$
  - Parameters
    - "Precision" $\lambda = a/b$.
    - "Degrees of freedom" $\nu = 2a$. 

Student’s t-Distribution

- Student’s t-Distribution
  - We reparametrize the infinite mixture of Gaussians to get
    $$St(x|\mu, \lambda, \nu) = \frac{\Gamma((\nu+2)/2)}{\Gamma(\nu/2)} \left( \frac{\lambda}{\nu} \right)^{1/2} \left[ 1 + \frac{\lambda(x - \mu)^2}{\nu} \right]^{-(\nu+2)/2}$$
  - Parameters
    - "Precision" $\lambda = a/b$.
    - "Degrees of freedom" $\nu = 2a$. 

- Student’s t-Distribution
  - We reparametrize the infinite mixture of Gaussians to get
    $$St(x|\mu, \lambda, \nu) = \frac{\Gamma((\nu+2)/2)}{\Gamma(\nu/2)} \left( \frac{\lambda}{\nu} \right)^{1/2} \left[ 1 + \frac{\lambda(x - \mu)^2}{\nu} \right]^{-(\nu+2)/2}$$
  - Parameters
    - "Precision" $\lambda = a/b$.
    - "Degrees of freedom" $\nu = 2a$. 

- Student’s t-Distribution
  - We reparametrize the infinite mixture of Gaussians to get
    $$St(x|\mu, \lambda, \nu) = \frac{\Gamma((\nu+2)/2)}{\Gamma(\nu/2)} \left( \frac{\lambda}{\nu} \right)^{1/2} \left[ 1 + \frac{\lambda(x - \mu)^2}{\nu} \right]^{-(\nu+2)/2}$$
  - Parameters
    - "Precision" $\lambda = a/b$.
    - "Degrees of freedom" $\nu = 2a$. 

**Student's t-Distribution: Visualization**

- **Behavior**
  - $\text{St}(x|\mu, \lambda, \nu) \mid \nu = 1 \rightarrow \text{Cauchy } N(x|\mu, \lambda^{-1})$

  ➔ More robust to outliers...

  ➔ Longer-tailed 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**

  $\text{St}(x|\mu, A, \nu) = \int_0^{\infty} N(x|\mu, (nA)^{-1}) \text{Gam}(\nu/2, \nu/2) \, dn$

  $= \frac{\Gamma(D/2 + \nu/2)}{\Gamma(\nu/2)} \left( \frac{1}{\nu/2} \right)^{D/2} \left( 1 + \frac{\Delta^2}{\nu} \right)^{-D/2 - \nu/2}$

  where $\Delta^2 = (x - \mu)^T A (x - \mu)$ is the Mahalanobis distance.

- **Properties**
  - $E[x] = \mu$, if $\nu > 1$
  - $\text{cov}[x] = \frac{\nu}{\nu - 2} A^{-1}$, if $\nu > 2$
  - $\text{mode}[x] = \mu$

---

**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 computational resources, they can generate exact results.
  - Approximation arises from the use of a finite amount of processor time.

  ⇒ Can use the use of Bayesian techniques across many domains.

  ⇒ But: computationally demanding, often limited to small-scale problems.
Approximate Inference

In general, assume we are given the pdf $p(x)$.

To draw samples from this pdf, we can invert the cdf:

$$F_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

**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.

**Sampling Idea**

- Objective:
  - Evaluate expectation of a function $f(x)$ w.r.t. a probability distribution $p(x)$.
  $$E[f] = \int f(x)p(x)dx$$

- Sampling idea:
  - Draw $L$ independent samples $x^i$ with $i = 1, ..., L$ from $p(x)$.
  - This allows the expectation to be approximated by a finite sum
  $$\hat{E}[f] = \frac{1}{L} \sum_{i=1}^{L} f(x^i)$$
  - As long as the samples $x^i$ are drawn independently from $p(x)$, then
  $$\lim_{L \to \infty} \hat{E}[f] = E[f]$$
  $\Rightarrow$ Unbiased estimate, independent of the dimension of $x$!

**Parametric Density Model**

- Example:
  - A simple multivariate (d-dimensional) Gaussian model
  $$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$
  - This is a "generative" model in the sense that we can generate samples $x$ according to the distribution.

**Sampling from a pdf (Transformation method)**

- In general, assume we are given the pdf $p(x)$ and the corresponding cumulative distribution:
  $$F(x) = \int_{-\infty}^{x} p(t)dt$$

- 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**
  \[ p(y) = \lambda e^{-\lambda y} \]
  where \( 0 \leq y < \infty \).

- **Transformation sampling**
  - Indefinite integral
    \[ h(y) = 1 - \exp(-\lambda y) \]
  - Inverse function
    \[ y = h(y)^{-1} = -\frac{1}{\lambda} \ln(1 - z) \]
  for a uniformly distributed input variable \( z \).

Example 2: Sampling from Cauchy Distrib.

- **Cauchy Distribution**

- **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 \( r^2 = z_1^2 + z_2^2 < 1 \).
  - 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
    \[ y_1 = z_1 \left( \frac{-2 \ln r^2}{r^2} \right)^{1/2} \]
    \[ y_2 = z_2 \left( \frac{-2 \ln r^2}{r^2} \right)^{1/2} \]
  - Then the joint distribution of \( y_1 \) and \( y_2 \) is given by
    \[ p(y_1, y_2) = p(z_1, z_2) \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \]
    \[ = \left[ \frac{1}{\sqrt{2\pi}} \exp\left(-y_1^2/2\right) \right] \left[ \frac{1}{\sqrt{2\pi}} \exp\left(-y_2^2/2\right) \right] \]
    \[ \Rightarrow y_1 \text{ and } y_2 \text{ are independent and each has a Gaussian distribution with mean } \mu \text{ and variance } \sigma^2. \]
  - If \( y \sim N(0,1) \), then \( \sigma y + \mu \sim N(\mu, \sigma^2) \).

Box-Muller Algorithm (cont’d)

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

Ancestral Sampling

- **Generalization of this idea to directed graphical models.**
  - Joint probability factorizes into conditional probabilities:
    \[ p(x) = \prod_{k=1}^{K} p(x_k | p_{a_k}) \]

- **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).
    \[ \hat{x}_n \sim p(x_n | p_{a_n}) \]
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.

Rejection Sampling

- Assumptions
  - Sampling directly from \( p(z) \) is difficult.
  - But we can easily evaluate \( p(z) \) (up to some normalization factor \( Z_p \)):
    \[
    p(z) = \frac{1}{Z_p} \tilde{p}(z)
    \]

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

Rejection Sampling - Discussion

- Limitation: high-dimensional spaces
  - For rejection sampling to be of practical value, we require that \( k\tilde{p}(z) \) be close to the required distribution, so that the rate of rejection is minimal.

- Artificial example
  - Assume that \( p(z) \) is Gaussian with covariance matrix \( \sigma^2 I \)
  - Assume that \( \tilde{p}(z) \) is Gaussian with covariance matrix \( \sigma^2 I \)
  - Obviously: \( \sigma^2 \geq \sigma^2 \)
  - In \( D \) dimensions: \( k = (\sigma/\sigma)_D \)
    - Assume \( \sigma \) is just 1% larger than \( \sigma \).
    - \( D = 1000 \Rightarrow k = 1.01^{1000} \geq 20,000 \)
    - And \( p(\text{accept}) = \frac{1}{k} \)

  \( \Rightarrow \) Often impractical to find good proposal distributions for high dimensions!

Example: Sampling from a Gamma Distib.

- 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
    \[
    e = \alpha - 1 \quad b^2 = 2a - 1
    \]
Importance Sampling

- **Approach**
  - Approximate expectations directly (but does not enable to draw samples from \( p(\mathbf{x}) \) directly).
  - Goal:
    \[
    \mathbb{E}[f] = \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}
    \]
- **Simplistic strategy: Grid sampling**
  - Discretize \( x \)-space into a uniform grid.
  - Evaluate the integrand as a sum of the form
    \[
    \mathbb{E}[f] \approx \frac{1}{L} \sum_{l=1}^{L} f(\mathbf{x}^{(l)}) p(\mathbf{x}^{(l)}) d\mathbf{x}
    \]
  - But: number of terms grows exponentially with number of dimensions!

Importance Sampling

- **Typical setting:**
  - \( p(\mathbf{x}) \) can only be evaluated up to an unknown normalization constant
    \[ p(\mathbf{x}) = \frac{\pi(\mathbf{x})}{Z_{p}} \]
  - \( q(\mathbf{x}) \) can also be treated in a similar fashion.
    \[ q(\mathbf{x}) = \frac{\pi(\mathbf{x})}{Z_{q}} \]
  - Then
    \[
    \mathbb{E}[f] = \frac{Z_{q}}{Z_{p}} \frac{1}{L} \sum_{l=1}^{L} \tilde{r}_{l} f(\mathbf{x}^{(l)})
    \]
  - with: \( \tilde{r}_{l} = \frac{\pi(\mathbf{x}^{(l)})}{q(\mathbf{x}^{(l)})} \)

Importance Sampling - Discussion

- **Observations**
  - Success of importance sampling depends crucially on how well the sampling distribution \( q(\mathbf{x}) \) matches the desired distribution \( p(\mathbf{x}) \).
  - Often, \( p(\mathbf{x})/q(\mathbf{x}) \) is strongly varying and has a significant proportion of its mass concentrated over small regions of \( x \)-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 \( p(\mathbf{x})/q(\mathbf{x}) \) 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 \( q(\mathbf{x}) \):
    - Should not be small or zero in regions where \( p(\mathbf{x}) \) is significant!
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

- 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)

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