Recap: GPs with Noise-free Observations

- Assume our observations are noise-free: \( \{ (x_n, f_n) \mid n = 1, \ldots, N \} \)
  - Joint distribution of the training outputs \( f \) and test outputs \( f^* \) according to the prior:
    \[
    \begin{bmatrix}
    f \\
    f^*
    \end{bmatrix}
    \sim \mathcal{N}
    \begin{bmatrix}
    0 \\
    0
    \end{bmatrix},
    \begin{bmatrix}
    K(X, X) & K(X, X^*) \\
    K(X^*, X) & K(X^*, X^*)
    \end{bmatrix}
    \]
  - Calculation of posterior corresponds to conditioning the joint Gaussian prior distribution on the observations:
    \[
    f | X, X^* \sim \mathcal{N}(f^*, \text{cov}[f^*])
    \]
    \[
    \text{cov}[f^*] = K(X^*, X) - K(X^*, X) K(X, X)^{-1} K(X, X^*)
    \]

Recap: GPs with Noisy Observations

- Joint distribution of the observed values and the test locations under the prior:
  \[
  \begin{bmatrix}
  t \\
  f^*
  \end{bmatrix}
  \sim \mathcal{N}
  \begin{bmatrix}
  0 \\
  0
  \end{bmatrix},
  \begin{bmatrix}
  K(X, X) + \sigma_t^2 I & K(X, X^*) \\
  K(X^*, X) & K(X^*, X^*)
  \end{bmatrix}
  \]
  - Calculation of posterior corresponds to conditioning the joint Gaussian prior distribution on the observations:
    \[
    f, X, X^*, t \sim \mathcal{N}(f^*, \text{cov}[f^*])
    \]
    \[
    f^* = \mathbb{E}[f | X, X^*, t]
    \]
    \[
    \text{cov}[f^*] = K(X, X) - K(X, X^*) K(X, X)^{-1} K(X, X^*)
    \]
  - 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 \( k(x, x') \), evaluated on \( X \).

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)
    - \( \theta \) - e.g., controlling prior distribution of \( \mathbf{w} \).
  - Top level
    - Typically discrete set of model structures \( \mathbf{H} \).
- Approach
  - Inference takes place one level at a time.

Recap: Model Selection at Lowest Level

- Posterior of the parameters \( \mathbf{w} \) is given by Bayes’ rule
  \[
  p(\mathbf{w} | t, \theta, \mathbf{H}_i) = \frac{p(t | \mathbf{w}, \theta, \mathbf{H}_i) p(\mathbf{w} | \theta, \mathbf{H}_i)}{p(t | \theta, \mathbf{H}_i)}
  \]
  \[
  p(t | \mathbf{w}, \theta, \mathbf{H}_i) = \frac{p(t | \mathbf{w}, \theta, \mathbf{H}_i)}{p(t | \theta, \mathbf{H}_i)}
  \]
  - with
    - likelihood and
    - prior parameters \( \mathbf{w} \), \( \theta \), \( \mathbf{H}_i \).
    - Denominator (normalizing constant) is independent of the parameters and is called marginal likelihood.
      \[
      p(t | \theta, \mathbf{H}_i) = \int p(t | \mathbf{w}, \theta, \mathbf{H}_i) p(\mathbf{w} | \theta, \mathbf{H}_i) d\mathbf{w}
      \]
Recap: Model Selection at Mid Level

- Posterior of parameters $\theta$ is again given by Bayes’ rule

$$p(\theta | t, X, H_i) = \frac{p(t | X, \theta, H_i) p(\theta | H_i)}{p(t | X, H_i)}$$

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

$$p(t | X, H_i) = \int p(t | X, \theta, H_i) p(\theta | H_i) d\theta$$

Recap: Model Selection at Top Level

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

$$p(H_i | t, X) = \frac{p(t | X, H_i) p(H_i)}{p(t | X)}$$

where
- Again, the denominator of the previous level $p(t | X, H_i)$ plays the role of the likelihood.
- $p(H_i)$ is the prior of the model structure.
- Denominator (normalizing constant) is given by:

$$p(t | X) = \sum_i p(t | X, H_i) p(H_i)$$

Recap: Bayesian Model Selection

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

$$p(t | X, H_i) = \int p(t | X, \theta, H_i) p(\theta | H_i) d\theta$$

- 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.

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.
Topics of This Lecture

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

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.

Sampling from a Gaussian

- Given: 1-dim. Gaussian pdf (probability density function) \( p(x | \mu, \sigma^2) \) and the corresponding cumulative distribution:
  \[
  F_{\mu, \sigma^2}(x) = \int_{-\infty}^{x} p(z | \mu, \sigma^2)dz
  \]
- To draw samples from a Gaussian, we can invert the cumulative distribution function:
  \[
  u \sim \text{Uniform}(0, 1) \Rightarrow F_{\mu, \sigma^2}^{-1}(u) \sim p(x | \mu, \sigma^2)
  \]
  \[
  p(x | \mu, \sigma^2) \\
  F_{\mu, \sigma^2}(x)
  \]

Sampling Idea

- Objective:
  - Evaluate expectation of a function \( f(x) \) w.r.t. a probability distribution \( p(x) \):
    \[
    \mathbb{E}[f] = \int f(x) p(x) dx
    \]
- Sampling idea
  - Draw \( L \) independent samples \( x^{(l)} \) with \( l = 1,..., L \) from \( p(x) \).
  - This allows the expectation to be approximated by a finite sum
    \[
    \hat{f} = \frac{1}{L} \sum_{l=1}^{L} f(x^{(l)})
    \]
  - As long as the samples \( x^{(l)} \) are drawn independently from \( p(x) \), then
    \[
    \mathbb{E}[f] - \mathbb{E}[\hat{f}] = \mathcal{O}(1/L)
    \]
  - \( \Rightarrow \) 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:
  \[
  F(x) = \int_{-\infty}^{x} p(z)dz
  \]
- To draw samples from this pdf, we can invert the cumulative distribution function:
  \[
  u \sim \text{Uniform}(0, 1) \Rightarrow F^{-1}(u) \sim p(x)
  \]
Example 1: Sampling from Exponential Distribution

- Exponential Distribution
  \[ p(y) = \lambda \exp(-\lambda y) \]
  where \( 0 \leq y < \infty \).

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

Example 2: Sampling from Cauchy Distribution

- Cauchy Distribution
  \[ p(y) = \frac{1}{\pi (1 + y^2)} \]

- Transformation sampling
  Inverse of integral can be expressed as a tan function.
  \[ y = h(y)^{-1} = \tan(z) \]
  for a uniformly distributed input variable \( z \).

Note: Efficient Sampling from a Gaussian

- Problem with transformation method
  Integral over Gaussian cannot be expressed in analytical form.
  Standard transformation approach is very inefficient.

- More efficient: Box-Muller Algorithm
  Generate pairs of uniformly distributed random numbers \( z_1, z_2 \in (-1,1) \).
  Discard each pair unless it satisfies \( r^2 = z_1^2 + z_2^2 \leq 1 \).
  This leads to a uniform distribution of points inside the unit circle with \( p(z_1, z_2) = 1/\pi \).

Box-Muller Algorithm (cont’d)

- Box-Muller Algorithm (cont’d)
  For each pair \( z_1, z_2 \), evaluate
  \[ y_1 = z_1 \left(\frac{-2 \ln r^2}{r^2}\right)^{1/2} \]
  \[ y_2 = z_2 \left(\frac{-2 \ln r^2}{r^2}\right)^{1/2} \]
  Then the joint distribution of \( y_1 \) and \( y_2 \) is given by
  \[ p(y_1, y_2) = p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right| \]
  \[ = \left[ \frac{1}{\sqrt{2\pi}} \exp(-y_1^2/2) \right] \left[ \frac{1}{\sqrt{2\pi}} \exp(-y_2^2/2) \right] \]
  \[ \Rightarrow y_1 \text{ and } y_2 \text{ are independent and each has a Gaussian distribution with mean } \mu \text{ and variance } \sigma^2 \].
  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.

Discussion

- **Transformation method**
  - Limited applicability, as we need to invert the indefinite integral of the required distribution \( p(z) \).
  - This will only be feasible for a limited number of simple distributions.

- **More general**
  - Rejection Sampling
  - Importance Sampling

Rejection Sampling

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

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

**Example: Sampling from a Gamma Distrib.**

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

- **Rejection sampling approach**
  - For \( a > 1 \), Gamma distribution has a bell-shaped form.
  - Suitable proposal distribution is Cauchy (for which we can use the transformation method).
  - Generalize Cauchy slightly to ensure it is nowhere smaller than Gamma: \( y = b \tan(c) \) 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
    \[
    b^2 = 2a - 1
    \]
Importance Sampling

**Approach**
- Approximate expectations directly (but does not enable to draw samples from \( p(x) \) directly).
- Goal: \( \mathbb{E}[f] = \int f(x)p(x)dx \)

**Simplistic strategy: Grid sampling**
- Discretize \( x \)-space into a uniform grid.
- Evaluate the integrand as a sum of the form \( \mathbb{E}[f] \approx \sum_{l=1}^{L} f(x^{(l)})p(x^{(l)})dx \)
  - But: number of terms grows exponentially with number of dimensions!

**Typical setting:**
- \( p(x) \) can only be evaluated up to an unknown normalization constant \( p(x) = \frac{\hat{p}(x)}{Z_p} \).
- \( q(x) \) can also be treated in a similar fashion, \( q(x) = \frac{\hat{q}(x)}{Z_q} \).
- Then
  \[
  \mathbb{E}[f] = \int f(x)p(x)dx = \int \frac{\hat{p}(x)}{Z_p} \frac{\hat{q}(x)}{Z_q} q(x)dx \\
  \approx \frac{1}{Z_p Z_q} \sum_{l=1}^{L} \hat{r}_l f(x^{(l)}) \\
  \text{with: } \hat{r}_l = \frac{\hat{p}(x^{(l)})}{\hat{q}(x^{(l)})}
  \]

**Discussion**
- **Observations**
  - Success of importance sampling depends crucially on how well the sampling distribution \( q(x) \) matches the desired distribution \( p(x) \).
  - Often, \( p(x)/f(x) \) 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 \( p(x)/f(x) \) 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 \( q(x) \):
    - Should not be small or zero in regions where \( p(x) \) is significant!

**Ratio of normalization constants can be evaluated**
\[
\frac{Z_p}{Z_q} = \frac{1}{Z_p} \int \hat{p}(x)dx = \int \frac{\hat{p}(x^{(l)})}{\hat{q}(x^{(l)})} q(x)dx \approx \sum_{l=1}^{L} \hat{r}_l
\]
- and therefore
  \[
  \mathbb{E}[f] \approx \sum_{l=1}^{L} w_l f(x^{(l)}) \\
  \text{with}
  \]
  \[
  w_l = \frac{\hat{r}_l}{\sum_{l=1}^{L} \hat{r}_l} = \frac{\hat{p}(x^{(l)})}{\hat{q}(x^{(l)})} q(x^{(l)})
  \]
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/tprnn/book.html)

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