N. Bertschinger
M. Kaschube
V. Ramesh

Machine Learning II
Exercise Sheet

Exercise 1. Monte-Carlo integration: *Consider the expectation of $f(x)$ under some distribution with density $p(x)$, i.e.*

$$\mathbb{E}_p[f] = \mu = \int f(x)p(x)dx$$

Then, if samples $x_1, \ldots, x_N$ from $p(x)$ are available the above expectation can be approximated as

$$\mathbb{E}_p[f] \approx \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

Show that this Monte-Carlo estimate *is* **unbiased**, *i.e.* $\mathbb{E}_p[\hat{\mu}] = \mu$

2 points

Exercise 2. Inversion sampling: *Consider sampling from a standard normal distribution:*

1. *Does your favorite library function, i.e.* `numpy.random.normal`, *use inversion sampling? If not, which algorithm is used instead?*

   1 point

2. *How does the Box-Muller method work?*

   3 points

Exercise 3. Rejection sampling: *Consider the problem of sampling from a multi-variate Gaussian with mean zero and diagonal covariance matrix $\sigma_p^2 \mathbf{I}_D$ where $\mathbf{I}_D$ denotes the $D \times D$ identity matrix. As a sampling density, we choose another Gaussian with zero mean and covariance $\sigma_q^2 \mathbf{I}_D$.*

1. *The condition, $cq \geq p$ now requires that $\sigma_q^2 \geq \sigma_p^2$. What is the optimal, i.e. minimal, choice for c?*

   1 point

2. *How does the acceptance rate behave, when you increase the dimension D? Illustrate your findings with an example where σ_q exceeds σ_p by 1%.*

   3 points

N. Bertschinger
M. Kaschube
V. Ramesh

Machine Learning II
Exercise Sheet

Exercise 1. Metropolis-Hastings: *Show that the Markov chain defined by the Metropolis-Hastings algorithm satisfies detailed balance.*

3 points

Exercise 2. MCMC: *Sample from a multi-modal distribution, e.g. a mixture of two one-dimensional Gaussian:*

$$p(x) = \frac{1}{3}\mathcal{N}(-\frac{\mu}{2}, 1) + \frac{2}{3}\mathcal{N}(\frac{\mu}{2}, 1)$$

1. *Use the Metropolis-Hastings algorithm with a Gaussian of width σ as proposal distribution.*
2. *Illustrate the effect of the proposal width σ and the separation μ between the two modes on the sampling efficiency.*

5 points

Exercise 3. Gibbs sampling: *Show that Gibbs sampling is a special case of Metropolis-Hastings sampling, i.e. show that its proposal is always accepted.*

2 points

N. Bertschinger
M. Kaschube
V. Ramesh

Machine Learning II
Exercise Sheet

Exercise 1. Conjugate priors: *Show that the Dirichlet distribution $p(\theta)$ with parameters $(\alpha_1, \ldots, \alpha_K)$ is the conjugate prior for the categorical variable Z, i.e.*

$$p(Z = k|\theta) = \theta_k \quad \forall k = 1, \ldots, K$$

1. *Compute the posterior parameters, i.e.*

$$p(\theta_1, \ldots, \theta_K|\mathcal{D})$$

where $\mathcal{D} = (z_1, \ldots, z_N)$.

2. *Why are the Dirichlet parameters called* pseudo-counts*?*

3 points

Exercise 2. Gaussian mixtures:

1. *Explore the collapsed Gibbs sampler on different data sets, e.g. with two components of varying size and separation. When does it work well and when does it fail to mix?*

3 points

2. *Explain, e.g. in pseudo-code, how you would implement the uncollapsed Gibbs sampler, i.e. sampling from the full posterior*

$$p((z_n)_{n=1}^K, \{\mu_k\}_{k=1}^K, \{\Lambda_k\}_{k=1}^K|\mathcal{X})$$

Hint: Which conditional independencies can you exploit?

3 points

3. *Compare the uncollapsed and collapsed Gibbs sampler. Which algorithm would you expect to mix better and why?*

1 points