HW3: Sampling, Markov Chains, and MCMC

Last modified: 2023-03-04 12:05

Due date: Thu Mar 09 at 11:59pm ET

Status: **RELEASED.**

How to turn in: Submit PDF to https://www.gradescope.com/courses/496674/assignments/2716964

Jump to: Problem 1 Problem 2 Problem 3 background 1 background 2

Questions?: Post to the **hw3** topic on the Piazza discussion forums.

Instructions for Preparing your PDF Report

What to turn in: PDF of typeset answers via LaTeX. No handwritten solutions will be accepted, so that grading can be speedy and you get prompt feedback.

Please use provided LaTeX Template: https://github.com/tufts-ml-courses/cs136-23s-assignments/blob/main/unit3_HW/hw3_template.tex

Your PDF should include (in order):

- Cover page with your full name, estimate of hours spent, and Collaboration statement
- Problem 1a, 1b, 1c
- Problem 2a
- Problem 3a, 3b

When you turn in the PDF to gradescope, mark each part via the in-browser Gradescope annotation tool)

Background for Problem 1: Transformations of Vector Random Variables

Recall what we learned about transformations of vector random variables on day 11.

Suppose u is a $D-$length vector *source* random variable with known sample space $\mathcal{U} \subseteq \mathbb{R}^D$ and PDF function $f(u)$, so: $\int_{u \in \mathcal{U}} f(u) du = 1$.

Suppose our target random variable x has a D-dimension vector sample space $\mathcal{X} \subseteq \mathbb{R}^D$.

Suppose further that we have invertible transformation functions S and T, which satisfy:

- $T : \mathcal{U} \rightarrow \mathcal{X}$
- $S : \mathcal{X} \rightarrow \mathcal{U}$
- $S(T(u)) = u$, for all $u \in \mathcal{U}$
- $T(S(x)) = x$, for all $x \in \mathcal{X}$

Then if we take a sample $u \sim f$, and transform it to $x = T(u)$, then we know the PDF of our transformed value x is:

$$p(x) = f(S(x)) | \det (J_S(x)) |$$

where $J_S(x)$ is the $D \times D$ Jacobian matrix of the function $S(x)$ defined below. The symbol $\det$ denotes the *determinant* computation (which always produces a scalar given a square matrix), and the bars denote the absolute value.

$$J_S(x) = \begin{bmatrix} \frac{dS_1}{dx_1} & \frac{dS_1}{dx_2} & \dots & \frac{dS_1}{dx_D} \\ \frac{dS_2}{dx_1} & \frac{dS_2}{dx_2} & \dots & \frac{dS_2}{dx_D} \\ \vdots & & \ddots & \\ \frac{dS_D}{dx_1} & \frac{dS_D}{dx_2} & \dots & \frac{dS_D}{dx_D} \end{bmatrix}$$

Background for Problem 2: Metropolis-Hastings Algorithm

Recall what we learned about MCMC and the Metropolis-Hastings algorithm on day 12.

We are interested in a scalar real random variable Z with sample space $\Omega \subseteq \mathbb{R}$.

We wish we could sample from a target distribution with pdf function $p^*(z)$. However, we cannot evaluate this PDF or sample from the distribution. Instead, we can compute an easier function $\tilde{p}(z)$ for any $z \in \Omega$, where

$$p^*(z) = \underbrace{\frac{1}{\int_{z\in\Omega} \tilde{p}(z)dz}}_{\text{normalization constant}} \tilde{p}(z)$$

Remember that the normalization constant is just a positive scalar value, which does not depend on any specific candidate value $z \in \Omega$ that we might assign to our random variable Z.

Consider the Metropolis-Hastings algorithm for sampling from p^*. We assume we are given a specific proposal distribution Q which, when given a current z value, proposes a new z' value. This distribution Q must satisfy:

- Q is a valid PDF over the sample space Ω
- Q has an easy-to-evaluate PDF function, which we'll denote as $Q(z'|z)$.
- Given any previous value z, Q is easy to draw samples from: $z' \sim Q(\cdot|z)$.

Consider a single propose-accept transition using the Metropolis-Hastings algorithm. The PDF of a transition from z to some new location $z' \neq z$ is:

$$\mathcal{T}(z'|z) = Q(z'|z) \cdot \min(1, \frac{\tilde{p}(z')Q(z|z')}{\tilde{p}(z)Q(z'|z)})$$

This is the compound probability of two events: we first propose the specific different z' value from Q, and then we accept it using the Metropolis-Hastings accept/reject ratio.

Problem 1: From Standard Gaussian to Any Gaussian

Partial Credit Option: If you wish, you can solve all parts of this problem for the specialized $D = 1$ univariate case only, for 85% credit.

Consider a vector r.v. $u \in \mathbb{R}^D$ that follows a *standard* multivariate Gaussian distribution:

$$u \sim \mathcal{N}(0_D, I_D)$$

Here, *standard* means the mean is the all-zero vector and covariance is equal to the $D \times D$ identity matrix I_D). For a standard Gaussian, we can write the PDF as

$$f(u) = (2\pi)^{\frac{-D}{2}} e^{-\frac{1}{2}u^T u}$$

We can then define a source-to-target transformation (using notation of Background above):

- The source-to-target transform is $T(u) = Lu + m$
- The target-to-source transform is $S(x) = L^{-1}(x - m)$

This transform has two parameters:

- $m \in \mathbb{R}^D$, a vector defining a mean
- L, a *positive definite* lower-triangular $D \times D$ square matrix. Note that any lower triangular matrix whose main diagonal has all positive entries is positive definite (and thus invertible).

We can then define $\Sigma = LL^T$, where Σ is by construction a symmetric, positive definite matrix. This means that L is the Cholesky decomposition of Σ. Note that in special case $D = 1$, L is just a scalar standard deviation and Σ is the variance (also a scalar).

This transform has the following **Jacobian**:

$$\begin{aligned} J_S(x) &= \nabla_x S(x) \\ &= \nabla_x[L^{-1}x] - \nabla_x[L^{-1}m] \\ &= L^{-1} \end{aligned}$$

Using the rules for gradients of matrix/vector-valued functions with respect to vectors. See also Bishop PRML App C.

1a: Define $\Sigma = LL^T$. Show the following:

$$|\det\left(L^{-1}\right)| = \frac{1}{(\det\Sigma)^{\frac{1}{2}}}$$

1b: Show that the pdf of x is given by:

$$p(x) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{(\det\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)}$$

where we define $\Sigma = LL^T$.

Hints for 1a - 1b: Feel free to use any of the following identities

- *Hint (i)*: For any *invertible, lower triangular* matrix L, its inverse L^{-1} is also a lower triangular matrix see proof
- *Hint (ii)*: For any *invertible, lower triangular* matrix L, we know: $(L^{-1})^T = (L^T)^{-1}$
- *Hint (iii)*: For any *invertible* matrices A and B, we have $(AB)^{-1} = B^{-1}A^{-1}$
- *Hint (iv)*: Determinant of inverse equals inverse of determinant: $\det(A^{-1}) = \frac{1}{\det A}$
- *Hint (v)*: Determinant of products equals product of determinants: $\det(AB) = (\det A)(\det B)$
- *Hint (vi)*: Determinant of lower triangular matrix is equal to determinant of its transpose: $\det(L) = \det(L^T)$
- *Hint (vii)*: Transpose of products equals product of transposes: $(ABC)^T = C^TB^TA^T$

1c: Complete the Python code snippet below, to show how you can repurpose a standard Gaussian procedure like NumPy's `randn()` to get a sample from any multivariate Gaussian.

```
import numpy as np

def sample_from_mv_gaussian(m_D, Sigma_DD, random_state=np.random):
    ''' Draw sample from multivariate Gaussian

    Args
    ----
    m_D : 1D array, size D
        Mean vector
    Sigma_DD : 2D array, shape (D, D)
        Covariance matrix. Must be symmetric and positive definite.

    Returns
    -------
    x_D : 1D array, size D
        Sampled value of Gaussian with provided mean and covariance
    '''
    D = m_D.size
    L_DD = np.linalg.cholesky(Sigma_DD) # compute L from Sigma
    # GOAL: draw each entry of u_D from standard Gaussian
    u_D = np.zeros(D) # TODO FIXME use random_state.randn(...)
    # GOAL: Want x_D ~ Gaussian(mean = m_D, covar=Sigma_DD)
    x_D = np.zeros(D) # TODO FIXME transform u_D into x_D
    return x_D
```

Problem 2: Detailed Balance

Review the Background on Metropolis-Hastings above.

The PDF of the Metropolis-Hastings transition distribution from z to some new location $z' \neq z$ is given by:

$$\mathcal{T}(z'|z) = Q(z'|z)\min(1, \frac{\tilde{p}(z')Q(z|z')}{\tilde{p}(z)Q(z'|z)})$$

where $\tilde{p}$ is our desired PDF function (up to a constant), and Q is the proposal PDF.

2a: Show that the Metropolis-Hastings transition distribution $\mathcal{T}$ satisfies **detailed balance** with respect to the target distribution p^*.

That is, show that:

$$p^*(a)\mathcal{T}(b|a) = p^*(b)\mathcal{T}(a|b)$$

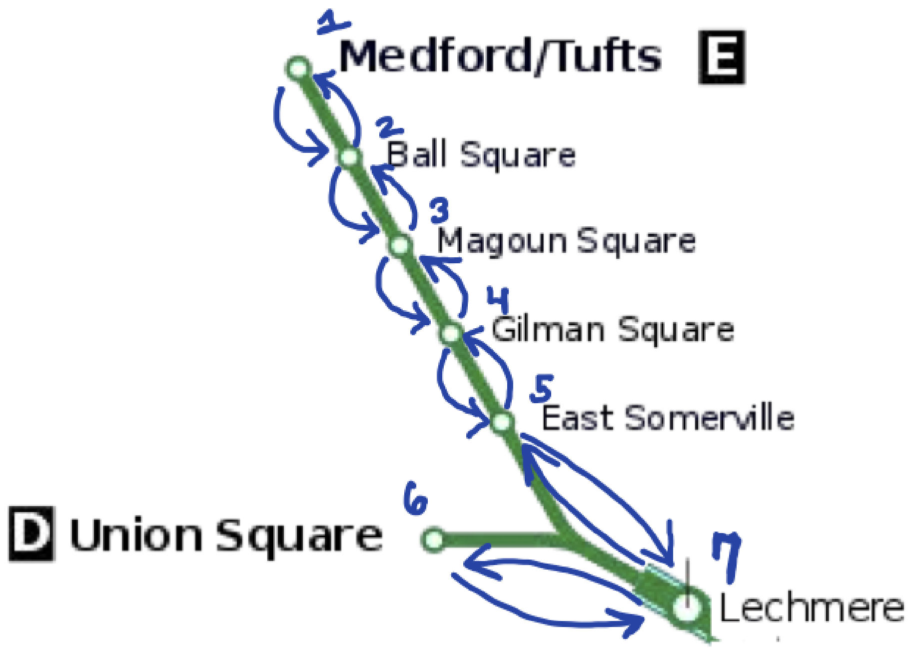
for all possible $a \neq b$, where a, b are any two *distinct* values of random variable z. Note that distinctness is key here, because it means definition of $\mathcal{T}$ above applies.

Hint: You can use following identity, which is true for any $x > 0, y > 0$

$$\frac{\min\left(1, \frac{x}{y}\right)}{\min\left(1, \frac{y}{x}\right)} = \frac{x}{y}$$

Problem 3: Random Walk on the Green Line

To celebrate the arrival of the Green Line Extension at Tufts, consider talking a random "walk" between the 7 stations in the diagram below.



Green Line Extension, imagined as possible states in a Markov chain

Each hand-drawn outgoing edge (blue arrows) from one node indicate all possible transition possibilities. Each has **equal probability**.

This means,

- from the Medford/Tufts station, you have 100% chance of moving to Ball Square
- from Ball Square, you have 50% chance to move to Magoun, and 50% chance of going back to Medford/Tufts.
- from Magoun, you have 50% chance to move on to Gilman, and 50% chance of going back to Ball Square.
- and so on

3a: You start at Medford/Tufts station, and take 1000 steps. What is your probability distribution over ending this journey at each of the 7 stations? Report as a vector (use order of nodes in the diagram, small to large). Round numbers to 3 decimal places.

3b: Is there a unique stationary distribution for this Markov chain? If so, explain why. If not, explain why not.

Hint: Consider repeating 3b, but starting elsewhere. Would you always get the same answer? What if you took 2000 or 3333 steps? Use course vocabulary in your explanation.