Week 3 Homework Submission

Iuliia Skobleva
Matriculation Number 03723809

November 3, 2019

Optimizing Likelihoods: Monotonic Transforms

Problem 1

The first derivatives of $\theta^t(1-\theta)^h$ and $\log\theta^t(1-\theta)^h$:

1. $\frac{\mathrm{d}}{\mathrm{d}\theta}\theta^t(1-\theta)^h = \theta^{t-1}(1-\theta)^{h-1}(t(1-\theta)-h\theta)$
2. $\frac{\mathrm{d}}{\mathrm{d}\theta}\log\theta^t(1-\theta)^h = \frac{\mathrm{d}}{\mathrm{d}\theta}\Big[\log\theta^t + \log(1-\theta)^h\Big] = \frac{t}{\theta} + \frac{h}{1-\theta}$

The second derivatives are:

1. $\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\theta^t(1-\theta)^h = (1-\theta)^{h-2}\theta^{t-2}[h^2\theta^2 + h\theta(2t(\theta-1)-\theta) + t(t-1)(\theta-1)^2]$
2. $\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log\theta^t(1-\theta)^h = -\frac{t}{\theta^2} + \frac{h}{(1-\theta)^2}$

To calculate the derivatives I used the chain rule and maybe some Wolfram Alpha.

Problem 2

To calculate the local maximum of $f(\theta)$ we need to calculate the first derivative and set it to zero to find $\theta_* = \text{argmax} f(\theta)$. Then, the second derivative $\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f(\theta_*)$ needs to be less than 0. Same logic applies when we are looking for a local maximum of $\log f(\theta)$.

The first derivative is $\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(\theta) = \frac{f'(\theta)}{f(\theta)}$, where $f'(\theta)$ is the inner derivative. Now, setting this derivative to zero leads to $f'(\theta) = 0$, which is aquivalent to $\frac{\mathrm{d}}{\mathrm{d}\theta}f(\theta) = 0$ and this is exactly the condition that needs to be satisfied if we are looking for the local maximum of $f(\theta)$.

Therefore, the local maxima of $f(\theta)$ are identical to the local maxima of $\log f(\theta)$.

Properties of MLE and MAP

Problem 3

We can use the formula from slide 26, namely:

$$\theta_{MAP} = \frac{M + a - 1}{N + M + a + b - 2} \tag{1}$$

We put in $\theta_{MAP} = 0.75$, $a = 6$ and $b = 4$:

$$\frac{M + 6 - 1}{N + M + 6 + 4 - 2} = 0.75 \tag{2}$$

$$\Leftrightarrow M + 5 = 0.75(N + M + 8) \tag{3}$$

$$\Leftrightarrow M = 3N + 1 \tag{4}$$

Under this condition we could get, for example, $M = 7$ and $N = 2$.

Problem 4

- The likelihood function is the Binomial distribution and it is given by

$$p(x = m|N, \theta) = \binom{N}{m} \theta^m (1 - \theta)^{N-m} \tag{5}$$

  To derive the maximum likelihood estimate for θ, we find θ that satisfies $\frac{\mathrm{d} \log p(x=m|N,\theta)}{\mathrm{d}\theta} = 0$. This happens to be equal to $\frac{m}{N}$.

- A prior distribution for θ is given by the Beta distribution with parameters a, b, namely:

$$\mathrm{Beta}(\theta|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1 - \theta)^{b-1} \tag{6}$$

  To find the prior mean of θ we look up the mean of the Beta distribution and it is equal to $\frac{a}{a+b}$.

- The posterior distribution is proportional to the likelihood times prior, like:

$$posterior \propto \theta^m (1 - \theta)^{N-m} \theta^{a-1} (1 - \theta)^{b-1} \tag{7}$$

$$\Leftrightarrow posterior \propto \theta^{m+a-1}(1-\theta)^{N-m+b-1} \tag{8}$$

Comparing this distribution to the beta distribution we find the normalizing factor and get:

$$p(\theta|\mathcal{D}) = \frac{\Gamma(N+a+b)}{\Gamma(m+a)\Gamma(N-m+b)}\theta^{m+a-1}(1-\theta)^{N-m+b-1} \tag{9}$$

We can look up the mean of this posterior distribution as well and it equals $\frac{m+a}{a+N+b}$.

Programming assignment 3: Probabilistic Inference

In [1]:
```
import numpy as np
import matplotlib.pyplot as plt

from scipy.special import loggamma
%matplotlib inline
```

Your task

This notebook contains code implementing the methods discussed in `Lecture 3: Probabilistic Inference`. Some functions in this notebook are incomplete. Your task is to fill in the missing code and run the entire notebook.

In the beginning of every function there is docstring, which specifies the format of input and output. Write your code in a way that adheres to it. You may only use plain python and `numpy` functions (i.e. no scikit-learn classifiers).

Exporting the results to PDF

Once you complete the assignments, export the entire notebook as PDF and attach it to your homework solutions. The best way of doing that is

1. Run all the cells of the notebook.
2. Export/download the notebook as PDF (File -> Download as -> PDF via LaTeX (.pdf)).
3. Concatenate your solutions for other tasks with the output of Step 2. On a Linux machine you can simply use `pdfunite`, there are similar tools for other platforms too. You can only upload a single PDF file to Moodle.

Make sure you are using `nbconvert` **Version 5.5 or later** by running `jupyter nbconvert --version`. Older versions clip lines that exceed page width, which makes your code harder to grade.

Simulating data

The following function simulates flipping a biased coin.

In [2]:
```
# This function is given, nothing to do here.
def simulate_data(num_samples, tails_proba):
    """Simulate a sequence of i.i.d. coin flips.

    Tails are denoted as 1 and heads are denoted as 0.

    Parameters
    ----------
    num_samples : int
        Number of samples to generate.
    tails_proba : float in range (0, 1)
        Probability of observing tails.

    Returns
    -------
    samples : array, shape (num_samples)
        Outcomes of simulated coin flips. Tails is 1 and heads is 0.
    """
    return np.random.choice([0, 1], size=(num_samples), p=[1 - tails_proba, tails_proba])
```

In [3]:
```
np.random.seed(123)  # for reproducibility
num_samples = 20
tails_proba = 0.7
samples = simulate_data(num_samples, tails_proba)
print(samples)
```

```
[1 0 0 1 1 1 1 1 1 1 1 1 0 1 1 0 0 1 1]
```

Important: Numerical stability

When dealing with probabilities, we often encounter extremely small numbers. Because of limited floating point precision, directly manipulating such small numbers can lead to serious numerical issues, such as overflows and underflows. Therefore, we usually work in the **log-space**.

For example, if we want to multiply two tiny numbers a and b, we should compute $\exp(\log(a) + \log(b))$ instead of naively multiplying $a \cdot b$.

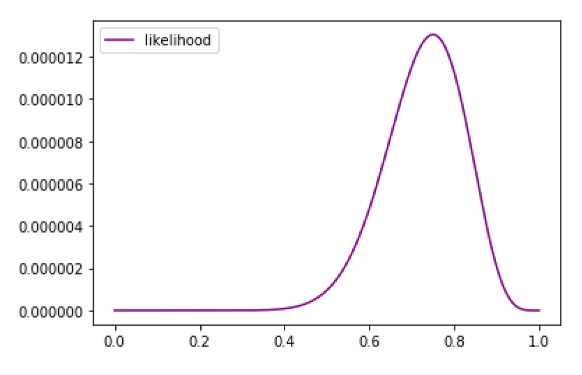
For this reason, we usually compute **log-probabilities** instead of **probabilities**. Virtually all machine learning libraries are dealing with log-probabilities instead of probabilities (e.g. Tensorflow-probability or Pyro).

Task 1: Compute $\log p(\mathcal{D} \mid \theta)$ for different values of θ

In [4]:
```
def compute_log_likelihood(theta, samples):
    """Compute log p(D | theta) for the given values of theta.

    Parameters
    ----------
    theta : array, shape (num_points)
        Values of theta for which it's necessary to evaluate the log-likelihood.
    samples : array, shape (num_samples)
        Outcomes of simulated coin flips. Tails is 1 and heads is 0.

    Returns
    -------
    log_likelihood : array, shape (num_points)
        Values of log-likelihood for each value in theta.
    """
    t = np.count_nonzero(samples == 1)
    h = np.count_nonzero(samples == 0)
    log_likelihood = np.empty(len(theta))

    for i in range(len(theta)):
        log_likelihood[i] = t*np.log(theta[i]) + h*np.log(1 - theta[i])

    return log_likelihood
```

In [5]:
```
x = np.linspace(1e-5, 1-1e-5, 1000)
log_likelihood = compute_log_likelihood(x, samples)
likelihood = np.exp(log_likelihood)
plt.plot(x, likelihood, label='likelihood', c='purple')
plt.legend()
```

Out[5]: <matplotlib.legend.Legend at 0x10f001828>



Note that the likelihood function doesn't define a probability distribution over θ --- the integral $\int_{0}^{1} p(\mathcal{D} \mid \theta) d\theta$ is not equal to one.

To show this, we approximate $\int_{0}^{1} p(\mathcal{D} \mid \theta) d\theta$ numerically using the rectangle rule.

In [6]:
```
# 1.0 is the length of the interval over which we are integrating p(D | theta)
int_likelihood = 1.0 * np.mean(likelihood)
print(f'Integral = {int_likelihood:.4}')
```

```
Integral = 3.068e-06
```

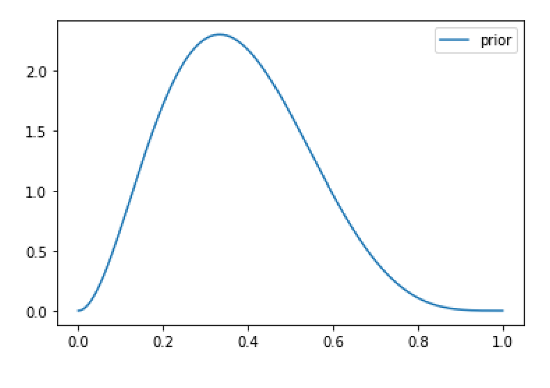
Task 2: Compute $\log p(\theta \mid a, b)$ for different values of θ

The function `loggamma` from the `scipy.special` package might be useful here. (It's already imported - see the first cell)

In [7]:
```
def compute_log_prior(theta, a, b):
    """Compute log p(theta | a, b) for the given values of theta.

    Parameters
    ----------
    theta : array, shape (num_points)
        Values of theta for which it's necessary to evaluate the log-prior.
    a, b: float
        Parameters of the prior Beta distribution.

    Returns
    -------
    log_prior : array, shape (num_points)
        Values of log-prior for each value in theta.

    """

    log_prior = np.empty(len(theta))
    for i in range(len(theta)):
        log_prior[i] = loggamma(a + b) + (a - 1)*np.log(theta[i]) + (b - 1)*np.log(1 - theta[i]) - l
oggamma(a) - loggamma(b)

    return log_prior
```

In [8]:
```
x = np.linspace(1e-5, 1-1e-5, 1000)
a, b = 3, 5

# Plot the prior distribution
log_prior = compute_log_prior(x, a, b)
prior = np.exp(log_prior)
plt.plot(x, prior, label='prior')
plt.legend()
```

Out[8]: <matplotlib.legend.Legend at 0x10f4070f0>



Unlike the likelihood, the prior defines a probability distribution over θ and integrates to 1.

In [9]:
```
int_prior = 1.0 * np.mean(prior)
print(f'Integral = {int_prior:.4}')
```

```
Integral = 0.999
```

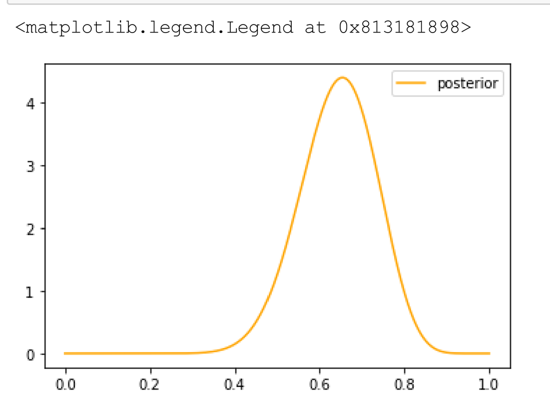
Task 3: Compute $\log p(\theta \mid \mathcal{D}, a, b)$ for different values of θ

The function `loggamma` from the `scipy.special` package might be useful here.

In [10]:
```
def compute_log_posterior(theta, samples, a, b):
    """Compute log p(theta | D, a, b) for the given values of theta.

    Parameters
    ----------
    theta : array, shape (num_points)
        Values of theta for which it's necessary to evaluate the log-prior.
    samples : array, shape (num_samples)
        Outcomes of simulated coin flips. Tails is 1 and heads is 0.
    a, b: float
        Parameters of the prior Beta distribution.

    Returns
    -------
    log_posterior : array, shape (num_points)
        Values of log-posterior for each value in theta.
    """
    t = np.count_nonzero(samples == 1)
    h = np.count_nonzero(samples == 0)
    log_posterior = np.empty(len(theta))
    for i in range(len(theta)):
        log_posterior[i] = loggamma(a + b + t + h) + (t + a - 1)*np.log(theta[i]) + (h + b - 1)*np.l
og(1 - theta[i]) - loggamma(t + a) - loggamma(h + b)

    return log_posterior
```

In [11]:
```
x = np.linspace(1e-5, 1-1e-5, 1000)

log_posterior = compute_log_posterior(x, samples, a, b)
posterior = np.exp(log_posterior)
plt.plot(x, posterior, label='posterior', c='orange')
plt.legend()
```

Out[11]: <matplotlib.legend.Legend at 0x813181898>



Like the prior, the posterior defines a probability distribution over θ and integrates to 1.

In [12]:
```
int_posterior = 1.0 * np.mean(posterior)
print(f'Integral = {int_posterior:.4}')
```

```
Integral = 0.999
```

Task 4: Compute θ_{MLE}

In [13]:
```
def compute_theta_mle(samples):
    """Compute theta_MLE for the given data.

    Parameters
    ----------
    samples : array, shape (num_samples)
        Outcomes of simulated coin flips. Tails is 1 and heads is 0.

    Returns
    -------
    theta_mle : float
        Maximum likelihood estimate of theta.
    """
    t = np.count_nonzero(samples == 1)
    h = np.count_nonzero(samples == 0)

    theta_mle = t / (t + h)
    return theta_mle
```

In [14]:
```
theta_mle = compute_theta_mle(samples)
print(f'theta_mle = {theta_mle:.3f}')
```

```
theta_mle = 0.750
```

Task 5: Compute θ_{MAP}

In [15]:
```
def compute_theta_map(samples, a, b):
    """Compute theta_MAP for the given data.

    Parameters
    ----------
    samples : array, shape (num_samples)
        Outcomes of simulated coin flips. Tails is 1 and heads is 0.
    a, b: float
        Parameters of the prior Beta distribution.

    Returns
    -------
    theta_map : float
        Maximum a posteriori estimate of theta.
    """
    t = np.count_nonzero(samples == 1)
    h = np.count_nonzero(samples == 0)

    theta_map = (t + a - 1) / (t + h + a + b - 2)
    return theta_map
```

In [16]:
```
theta_map = compute_theta_map(samples, a, b)
print(f'theta_map = {theta_map:.3f}')
```

```
theta_map = 0.654
```

Putting everything together

Now you can play around with the values of `a`, `b`, `num_samples` and `tails_proba` to see how the results are changing.

In [17]:
```
num_samples = 20
tails_proba = 0.7
samples = simulate_data(num_samples, tails_proba)
a, b = 3, 5
print(samples)
```

```
[1 1 1 1 1 1 1 0 0 1 0 1 1 1 1 1 1 1 1 1]
```

In [18]:
```
plt.figure(figsize=[12, 8])
x = np.linspace(1e-5, 1-1e-5, 1000)

# Plot the prior distribution
log_prior = compute_log_prior(x, a, b)
prior = np.exp(log_prior)
plt.plot(x, prior, label='prior')

# Plot the likelihood
log_likelihood = compute_log_likelihood(x, samples)
likelihood = np.exp(log_likelihood)
int_likelihood = np.mean(likelihood)
# We rescale the likelihood - otherwise it would be impossible to see in the plot
rescaled_likelihood = likelihood / int_likelihood
plt.plot(x, rescaled_likelihood, label='scaled likelihood', color='purple')

# Plot the posterior distribution
log_posterior = compute_log_posterior(x, samples, a, b)
posterior = np.exp(log_posterior)
plt.plot(x, posterior, label='posterior')

# Visualize theta_mle
theta_mle = compute_theta_mle(samples)
ymax = np.exp(compute_log_likelihood(np.array([theta_mle]), samples)) / int_likelihood
plt.vlines(x=theta_mle, ymin=0.00, ymax=ymax, linestyle='dashed', color='purple', label=r'$\theta_{M
LE}$')


# Visualize theta_map
theta_map = compute_theta_map(samples, a, b)
ymax = np.exp(compute_log_posterior(np.array([theta_map]), samples, a, b))
plt.vlines(x=theta_map, ymin=0.00, ymax=ymax, linestyle='dashed', color='orange', label=r'$\theta_{M
AP}$')

plt.xlabel(r'$\theta$', fontsize='xx-large')
plt.legend(fontsize='xx-large')
plt.show()
```