10-701 HW1

Dapeng Zhao
AndrewID: dapengz

18 Sep 2019

1 Probability Review (10 pts) [Derun]

- $P(A) = 0.5 \quad P(B) = 0.3 \quad P(C) = 0.2$
- $A = 2 \quad B = 1 \quad C = 0$
- String ends when C appears

1) **[4 pts]** Find $P(AAB, \overline{BAA})$ (AAB meaning "Substring AAB appears in the string").

Suppose after n trials(letters), the first AAB appears.
In the case of $AAB \cap \overline{BAA}$, the first (n-1) trials must all be A. *Proof by induction: the $(n-3)_{th}$ trial must be A, otherwise String[n-3:n] would be BAA which would be earlier than AAB(String[$n-2$:]). Similar argument can be made for $(k-1 \leftarrow k)$ till $(k=0)$.*

$$\begin{aligned}
P(AAB, \overline{BAA}) &= \sum_{n=3}^{\infty} P(A)^{n-1} P(B) \\
&= P(B) \sum_{n=2}^{\infty} P(A)^n \\
&= P(B)[\sum_{n=0}^{\infty} P(A)^n - \sum_{n=0}^{1} P(A)^n] \\
&= P(B)[\frac{1 - P(A)^{\infty}}{1 - P(A)} - \sum_{n=0}^{1} P(A)^n] \\
&= 0.3(\frac{1-0}{1-0.5} - 1 - 0.5) \\
&= 0.15
\end{aligned}$$

2) **[2 pts]** Find $E(T)$ (T= sum of the string).

Denote the string length as l, value of a non-C trial as t.
$E(t) = P(A|\overline{C}) * A + P(B|\overline{C}) * B = 1.625$
$E(T|l = n+1) = nE(t)$

$$\begin{aligned}
E(T) &= \sum_{n=0}^{\infty} P(l = n+1) E(T|l = n+1) \\
&= \sum_{n=0}^{\infty} P(\overline{C})^n P(C) n E(t) \\
&= P(C)E(t) \sum_{n=0}^{\infty} n P(\overline{C})^n \\
&= P(C)E(t) \frac{P(\overline{C})}{(1 - P(\overline{C}))^2} \\
&= \frac{E(t)}{P(C)} = \frac{1.625 * 0.8}{0.2} = 6.5
\end{aligned}$$

3) **[4 pts]** Find $V(T)$.

Denote $P(A|\overline{C})$ as p, $P(C|\overline{C})$ as q, and $q = 1 - p$.

$$\begin{aligned}
E(T^2|l = n+1) &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} [Ak + B(n-k)]^2 \\
&= \sum_{k=0}^{n} (n+k)^2 \binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=0}^{n} [k(k-1) + (2n+1)k + n^2] \binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=0}^{n} k(k-1) \binom{n}{k} p^k q^{n-k} + (2n+1) \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} + n^2 \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \\
&= n(n-1)p^2 \sum_{k=0}^{n} \binom{n-2}{k-2} p^{k-2} q^{n-k} + (2n+1)np \sum_{k=0}^{n} \binom{n-1}{k-1} p^{k-1} q^{n-k} + n^2 \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \\
&= n(n-1)p^2 + (2n+1)np + n^2 \\
&= (p+1)^2 n^2 + p(1-p)n
\end{aligned}$$

$$\begin{aligned}
E(T^2) &= E(E(T^2|l)) \\
&= \sum_{n=0}^{\infty} P(l = n+1) E(T^2|l = n+1) \\
&= \sum_{n=0}^{\infty} P(\overline{C})^n P(C) [(p+1)^2 n^2 + p(1-p)n] \\
&= (p+1)^2 P(C) \sum_{n=0}^{\infty} n^2 P(\overline{C})^n + p(1-p)P(C) \sum_{n=0}^{\infty} n P(\overline{C})^n \\
&= (p+1)^2 P(C) \frac{P(\overline{C})(1 + P(\overline{C}))}{(1 - P(\overline{C}))^3} + p(1-p)P(C) \frac{P(\overline{C})}{(1 - P(\overline{C}))^2} \\
&= 96
\end{aligned}$$

$$\begin{aligned}
V(T^2) &= E(T^2) - E(T)^2 \\
&= 96 - 6.5^2 \\
&= 53.75
\end{aligned}$$

2 Star Wars a MAP problem (20 pts)[Naji]

$Data : M \sim N(T + \sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$

$Prior\ Belief : T \sim N(\mu_0, \sigma_0^2)$

Given M, find T_{MAP}.

Denote:
$\mu_M = T + \sum_{i=1}^{n} \mu_i \ \ , \quad \sigma_M^2 = \sum_{i=1}^{n} \sigma_i^2$

$$\begin{aligned} T_{\text{MAP}} &= \underset{T}{\text{argmax}}\, P(M|T)P(T) \\ &= \underset{T}{\text{argmax}}(\log P(M|T) + \log P(T)) \\ &= \underset{T}{\text{argmin}}[\frac{(\mu_M - m)^2}{2\sigma_M^2} + \frac{(T-\mu_0)^2}{2\sigma_0^2}] \end{aligned}$$

Denote $g(T)$ as $\frac{(M-\mu_M)^2}{2\sigma_M^2} + \frac{(T-\mu_0)^2}{2\sigma_0^2}$, and make $\frac{\partial g}{\partial T} = 0$.

$$\frac{\partial g}{\partial T} = \frac{1}{\sigma_M^2}(\mu_M - m) + \frac{1}{\sigma_0^2}(T - \mu_0) = 0$$

Solve the equation above:

$$T_{\text{MAP}} = \frac{(M - \sum_{i=1}^{n} \mu_i)\sigma_0^2 + \mu_0 \sum_{i=1}^{n} \sigma_i^2}{\sum_{i=1}^{n} \sigma_i^2 + \sigma_0^2}$$

3 MLE and MAP (20 pts)[Derun & Justin]

3.1 MLE with Exponential Family [5 pts]

$P(x|\theta^*) = h(x)\exp(\theta^*\,\phi(x) - A(\theta^*))$
Given X_n, Find $\widehat{\theta}_{\text{MLE}}$.

$$\begin{aligned}\widehat{\theta}_{\text{MLE}} &= \underset{\theta}{\text{argmax}} \prod_{i=1}^{n} P(x_i|\theta) \\ &= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \log(P(x_i|\theta)) \\ &= \underset{\theta}{\text{argmax}} \frac{\sum_{i=1}^{n} \theta\phi(x_i)}{n} - A(\theta)\end{aligned}$$

Denote $g(\theta)$ as $\frac{\sum_{i=1}^{n} \theta\phi(x_i)}{n} - A(\theta)$, and make $\frac{\partial g}{\partial \theta} = 0$.

$$\frac{\partial g}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) - \frac{\partial A}{\partial \theta} = 0$$

$$\frac{\partial A}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$$

$$\theta = (A')^{-1}(\frac{1}{n}\sum_{i=1}^{n}\phi(x_i))$$

3.2 MLE and MAP with Weibull Distribution [15 pts]

1. **[5 pts]**

$$f(x) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}\exp\{-(x/\lambda)^k\}, \quad x \geq 0, k > 0, \lambda > 0.$$

Given X_n, find $\hat{\lambda}_{\text{MLE}}$.

$$\begin{aligned}\hat{\lambda}_{\text{MLE}} &= \underset{\lambda}{\text{argmax}} \prod_{i=1}^{n} f(x_i) \\ &= \underset{\lambda}{\text{argmax}} \sum_{i=1}^{n} \log f(x_i) \\ &= \underset{\lambda}{\text{argmin}}\ nk\log(\lambda) + \sum_{i=1}^{n} x_i^k \lambda^{-k}\end{aligned}$$

Denote $g(\lambda)$ as $nk\log(\lambda) + \sum_{i=1}^{n} x_i^k\lambda^{-k}$, and make $\frac{\partial g}{\partial \lambda} = 0$.

$$\frac{\partial g}{\partial \lambda} = \frac{nk}{\lambda} - \frac{k}{\lambda^{k+1}}\sum_{i=1}^{n} x_i^k = 0$$

Therefore,

$$\hat{\lambda}_{\text{MLE}} = (\frac{1}{n}\sum_{i=1}^{n} x_i^k)^{\frac{1}{k}}$$

2. **[8 pts]**

$$t = \lambda^k$$

$$f(t) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\frac{1}{t})^{\alpha+1}\exp\{-\beta/t\} \quad (\alpha > 0, \beta > 0).$$

Given X_i, find $P(t|D)$ and $\tilde{\lambda}_{\text{MAP}}$.

$f(x_i|t) = \dfrac{kx^{k-1}}{t}\exp\{-x^k/t\}$

$$\begin{aligned}
\tilde{t}_{\text{MAP}} &= \underset{t}{\text{argmax}}(\prod_{i=1}^{n} f(x_i|t))f(t) \\
&= \underset{t}{\text{argmax}} \sum_{i=1}^{n} \log f(x_i|t) + \log f(t) \\
&= \underset{t}{\text{argmin}}\ n\log t + \sum_{i=1}^{n} x_i^k/t + (\alpha+1)\log t + \beta/t
\end{aligned}$$

Denote $h(t)$ as $n\log t + \sum_{i=1}^{n} x_i^k/t + (\alpha+1)\log t + \beta/t$, and make $\frac{\partial h}{\partial t} = 0$.

$$\frac{\partial h}{\partial t} = \frac{n+\alpha+1}{t} - \frac{\beta + \sum_{i=1}^{n} x_i^k}{t^2} = 0$$

Therefore,

$$\tilde{t}_{\text{MAP}} = \frac{\beta + \sum_{i=1}^{n} x_i^k}{n+\alpha+1}$$

$$\tilde{\lambda}_{\text{MAP}} = (\frac{\beta + \sum_{i=1}^{n} x_i^k}{n+\alpha+1})^{\frac{1}{k}}$$

3. **[2 pts]** Assume $\sum_{i=1}^{n} x_i^k \to \infty$ as $n \to \infty$ for Weibull distribution. Compare $\hat{\lambda}$ and $\tilde{\lambda}$ as $n \to \infty$ and describe your findings.

$$\lim_{n\to\infty} \hat{\lambda}_{\text{MLE}} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^k\right)^{\frac{1}{k}}$$

$$\lim_{n\to\infty} \tilde{\lambda}_{\text{MAP}} = \left(\frac{\beta + \sum_{i=1}^{n} x_i^k}{n+\alpha+1}\right)^{\frac{1}{k}}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^k\right)^{\frac{1}{k}}$$

When $\lim_{n\to\infty}$, $\hat{\lambda}_{\text{MLE}}$ and $\tilde{\lambda}_{\text{MAP}}$ become the same. Because in the MAP estimate, when data set is large enough, the prior belief is weighted very little and the influence of prior belief gets diluted and eventually washed off, as if there is no prior belief, which results in MLE and MAP estimates becoming the same.

4 Fun with Linear Regression (20 pts)[Naji & Derun]

$$y = w^T x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\text{Gaussian: } p(\epsilon; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\epsilon-\mu)^2}$$

$$\text{Laplacian: } p(\epsilon; \mu, b) = \frac{1}{2b} e^{-\frac{|\epsilon-\mu|}{b}}$$

Given $D = \{(y^{(i)}, x^{(i)})\}_{i=1}^n$, find w.

4.1 MLE

(a) **[5 pts]** Compute the likelihood of the data, $L(w) := \prod_{i=1}^n P(y^{(i)}|x^{(i)}, \sigma, w)$.

$y \sim N(w^T x, \sigma^2)$

$$L(w) := \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - w^T x)^2}{2\sigma^2}\right)$$

(b) **[5 pts]** Compute the log-likelihood of the data, $\ell(w)$ and argue why the solution of the problem

$$\min_w \|Xw - Y\|_2^2$$

yields the maximizer of the likelihood, $L(w)$. **Explicitly** define X and Y.

$$l(w) = n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{n} \frac{(x^T w - y)^2}{2\sigma^2}$$

$X = [x_1\ x_2\ \dots\ x_n]^T$, shape: n*p

$Y = [y_1\ y_2\ \dots\ y_n]^T$, shape: n*1

$$\begin{aligned} w_{\text{MLE}} &= \operatorname*{argmax}_w L(w) \\ &= \operatorname*{argmax}_w l(w) \\ &= \operatorname*{argmin}_w \sum_{i=1}^{n} (x^T w - y)^2 \\ &= \operatorname*{argmin}_w \sum_{i=1}^{n} (x^T w - y)^2 \\ &= \operatorname*{argmin}_w \|Xw - Y\|_2^2 \end{aligned}$$

As shown above, the solution of the problem $\min_w \|Xw - Y\|_2^2$ is also the solution for $\max_w L(w)$.

4.2 MAP estimator with Laplacian Prior

$w_i \sim \mathcal{L}(0, \rho)$.

(a) **[5 pts]** Describe how to derive w_{MAP}.

$$\begin{aligned} w_{\text{MAP}} &= \underset{w}{\text{argmax}}\ P(w|D) \\ &= \underset{w}{\text{argmax}}\ P(D|w)P(w) \\ &= \underset{w}{\text{argmax}}\ [\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - w^T x)^2}{2\sigma^2}\right)][\prod_{i=1}^{p} \frac{1}{2\rho} \exp\left(-\frac{|w_i|}{\rho}\right)] \end{aligned}$$

Find the w that can maximize $P(D|w)P(w)$, this w will be the MAP estimate.

(b) **[5 pts]** Compute the log-posterior of the data and argue why the solution of the problem

$$\min_w \|Xw - Y\|_2^2 + \lambda \|w\|_1 \tag{1}$$

yields the minimizer of the posterior. **Explicitly** define X, Y and λ.

The log-posterior of the data is:

$$\begin{aligned} lp(w) &= \log[P(w|D)] \\ &= \log[P(D|w)P(w)/P(D)] \\ &= n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{n} \frac{(x^T w - y)^2}{2\sigma^2} + p \log \frac{1}{2\rho} - \sum_{i=1}^{p} \frac{|w_i|}{\rho} - \log P(D) \\ w_{\text{MAP}} &= \underset{w}{\text{argmax}}\ P(w|D) \\ &= \underset{w}{\text{argmax}}\ P(D|w)P(w) \\ &= \underset{w}{\text{argmax}}\ lp(w) \\ &= \underset{w}{\text{argmin}} \|Xw - Y\|_2^2 + 2\sigma^2 \sum_{i=0}^{p} \frac{|w_i|}{\rho} \\ &= \underset{w}{\text{argmin}} \|Xw - Y\|_2^2 + 2\sigma^2 \frac{\|w\|_1}{\rho} \end{aligned}$$

The X and Y are the same with 4.1(b):

$X = [x_1\ x_2\ \ldots\ x_n]^T$, shape: n*p

$Y = [y_1\ y_2\ \ldots\ y_n]^T$, shape: n*1

,while

$\lambda = \frac{2\sigma^2}{\rho}$.

As shown above, the solution of the problem $\min_w \|Xw - Y\|_2^2 + \lambda\|w\|_1$ is also the solution for $\max_w\ P(w|D)$.

5 Programming Exercise (24 pts) [Justin & Naji]

5.1 Maximum Likelihood Estimation

(a) **[8 pts]**

$$t = w_f f + w_\alpha \alpha + w_c c + w_U U + w_s s + w_0 + \epsilon, \qquad \epsilon \sim \mathcal{N}(0,1)$$
$$t = w'^T x' + \epsilon$$

$\psi(x)$ is the probability density function of the Normal distribution $\mathcal{N}(0,1)$

(i) **[4pts]** Given $\{X_i\}_{i=1}^N$, derive the likelihood function.

$t_i \sim N(w'^T X_i', 1)$

Likelihood function:

$$L(D|w) = \prod_{i=1}^{n} \psi(t_i)$$
$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i'^T w' - t_i)^2}{2}\right)$$

Log-likelihood function:

$$l(D|w) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i'^T w' - t_i)^2}{2}\right)$$
$$= n \log \frac{1}{\sqrt{2\pi}} - \sum_{i=1}^{n} \frac{(X_i'^T w' - t_i)^2}{2}$$
$$= n \log \frac{1}{\sqrt{2\pi}} - \frac{\|X'^T w' - t\|_2^2}{2}$$

(ii) **[4pts]** Write plotMLE(X, w_α) to plot log-likelihood function confined to w_α.

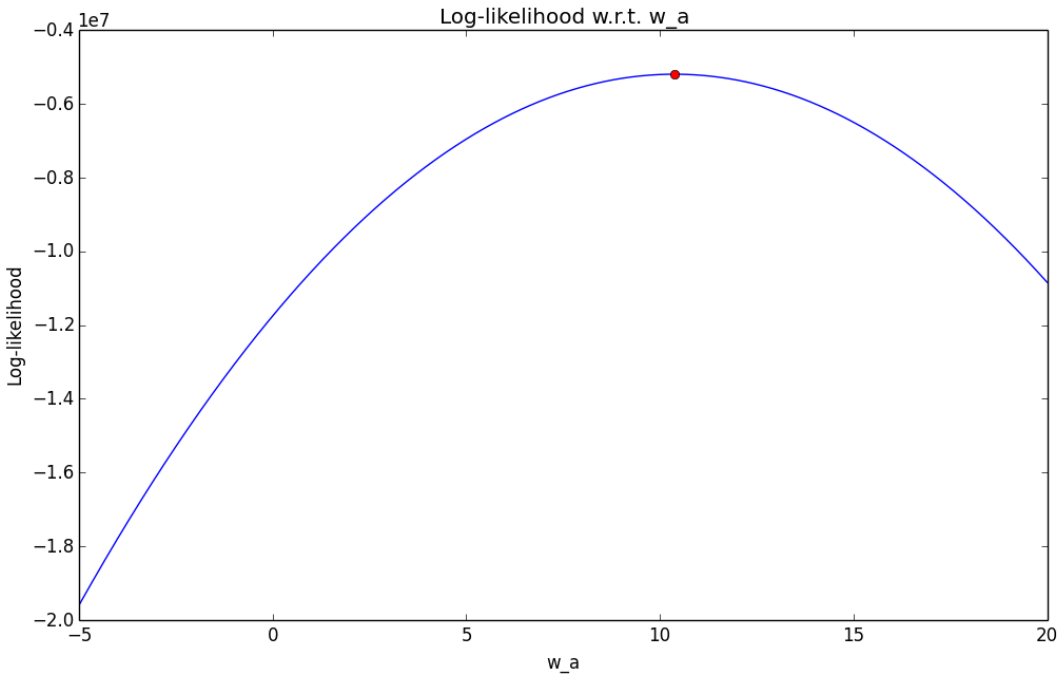
```
def plotMLE(data,Wa):
n,m = len(data),len(Wa)
l = np.zeros(m,dtype='float32')
for j in range(m):
    for i in range(n):
        l[j] += 0.0 - np.log(2*np.pi*eps_var) - (data[i,5]-Wa[j]*data[i,1]-eps_mean)**2/(2*e

l_argmax = np.argmax(l)

plt.plot(Wa,l,'-')
plt.plot(Wa[l_argmax],l[l_argmax],'ro')
plt.xlabel('w_a')
plt.ylabel('Log-likelihood')
plt.show()
```

(b) **[4 pts]** Use `airfoil_sel_noise.data` and $-5.0, -4.99, -4.98, ..., 19.97, 19.98, 19.99, 20.0$ for w_α to plot. Write down observation and estimated w_α.

The estimated w_α is 10.37. Observation: The function is quadratic, and the estimated w_α is on the peak of the function.


Log-likelihood w.r.t. w_a
1e7
−0.4
−0.6
−0.8
−1.0
−1.2
−1.4
−1.6
−1.8
−2.0
Log-likelihood
−5
0
5
10
15
20
w_a

.

5.2 Maximum a Posteriori Estimation

(a) **[4 pts]**
Prior belief: $w_i \sim \mathcal{N}(\mu, \sigma)$
Now with these assumptions, write the function `logPosterior`(X,w_α,w_U,w_c,w_s,w_f, μ,σ) that outputs the log-posterior function $\ell(\mathbf{w})$.

Posterior function:

$$P(w|D) = P(D|w)P(w)$$
$$= [\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i'^T w' - t_i)^2}{2}\right)][\prod_{i=1}^{6} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(w_i' - \mu)^2}{2\sigma^2}\right)]$$

Log-posterior function:

$$l(D|w) = n \log \frac{1}{\sqrt{2\pi}} - [\sum_{i=1}^{n} \frac{(X_i'^T w' - t_i)^2}{2}] + 6 \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{6} \frac{(w_i' - \mu)^2}{2\sigma^2}$$
$$= n \log \frac{1}{\sqrt{2\pi}} - \frac{\|X'^T w' - t\|_2^2}{2} + 6 \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{\|w' - \vec{\mu}\|_2^2}{2\sigma^2}$$

```
def logPosterior(X,wa,wu,wc,ws,wf,w0,mean,sigma):
    n = len(X)
    var = 1.0*sigma**2
    w = np.array([wa,wu,wc,ws,wf,w0])
    t = X[:,5]
    X[:,5] = np.ones(n)
    l = 0
    for i in range(n):
        l += 0.0 - 0.5*np.log(2*np.pi) - (t[i]-np.matmul(w,X[i]))**2/2
    for i in range(len(w)):
        l += 0.0 - 0.5*np.log(2*np.pi*var) - (w[i]-mean)**2/(2*var)
    return l
```

(b) **[4 pts]** Use the given dataset, function `logPosterior` and maximization method to report the MAP estimate on three scenarios ($\mu = 0$, $\mu = 10$ and $\mu = 500$; $\sigma = 1$).

when $\mu = 0$: $w = \begin{bmatrix} -1.23861276e-03 \\ -6.06403948e-01 \\ -3.35693049e+01 \\ 1.21490238e-01 \\ -1.51119253e+01 \\ 1.31010080e+02 \end{bmatrix}$

when $\mu = 10$: $w = \begin{bmatrix} -1.23690898e-03 \\ -6.17601310e-01 \\ -3.30365152e+01 \\ 1.20329608e-01 \\ -6.28866304e+00 \\ 1.30975483e+02 \end{bmatrix}$

when $\mu = 500$: $w = \begin{bmatrix} -1.15823720e-03 \\ -1.16690834e+00 \\ -6.92565807e+00 \\ 6.40112866e-02 \\ 4.26035629e+02 \\ 1.29284086e+02 \end{bmatrix}$

(c) **[4 pts]** Do you see any significant differences between the MAP estimates as you change the μ values? Explain why they're different or why they're not and report the MAP estimate for each value of μ.

No. While some dimensions of w vary a bit with different mean, they are in general similar.
The dataset is big enough, so even though the final MAP estimates are still affected by our prior belief on the mean of w's distribution, they are more influenced by the experiment data which is the same set for all 3 cases. Therefore, similar estimates should be yielded.

6 Collaboration

1. Did you receive any help whatsoever in solving this assignment? No.
2. If you answered Yes to the previous question, give full details below (e.g., "Jane Doe explained to me what is asked in Question 3.4")
3. Did you give any help whatsoever in solving this assignment? No
4. If you answered Yes to the previous question, give full details below (e.g., "I pointed Joe Smith to section 2.3 since he didn't know how to proceed with Question 2").