

EEL5840 Fundamental Machine Learning Homework 4

Hudanyun Sheng

Question 1

It is attached in the separate file.

Question 2

It is attached in the separate file.

Question 3

The code is attached at the end of this file.

Suppose the true mean of the Gaussian distribution is 2 ($\mu = 2$). And we are drawing random samples from this Gaussian distribution, with a Gaussian prior distribution on the mean parameters. Assume the variance for the Gaussian likelihood and Gaussian prior is fixed and known. And our goal is to estimate the mean. Since we are to explore how do the initial values of the prior mean, prior variance and likelihood variance interact to effect the final estimate of mean, below shows my experiment design, for each experiment I draw 10 samples.

	μ_0	σ_0	σ
1	4	0.5	0.5
2		0.5	5
3		5	0.5
4		5	5
5	2	0.5	0.5
6		0.5	5
7		5	0.5
8		5	5

Table 1: Experiment design

The results of the experiments described above are attached.

The table of experiment results from first 10 draws from Gaussian:

		Number of draws	1	2	3	4	5	6	7	8	9	10
mu_0 = 4	sigma = 0.5, sigma_0 = 0.5	sample	2.36	1.61	2.42	1.44	1.29	2.36	1.61	2.16	2.70	2.20
		true probability	0.77	0.22	0.80	0.13	0.08	0.76	0.22	0.62	0.92	0.66
		ML probability	0.50	0.23	0.72	0.15	0.14	0.81	0.30	0.69	0.92	0.64
		MAP probability	0.05	0.03	0.52	0.06	0.05	0.73	0.21	0.63	0.92	0.66
		ML estimate	2.36	1.99	2.13	1.96	1.82	1.91	1.87	1.91	1.99	2.02
		MAP estimate	3.18	2.59	2.39	2.23	2.11	2.05	2.01	1.99	1.99	1.99
	sigma = 5, sigma_0 = 0.5	sample	5.06	3.45	3.98	-2.35	-0.49	1.47	-1.44	3.66	13.83	-0.41
		true probability	0.73	0.61	0.65	0.19	0.31	0.46	0.25	0.63	0.99	0.31
		ML probability	0.50	0.44	0.49	0.16	0.31	0.47	0.29	0.65	0.98	0.27
		MAP probability	0.58	0.45	0.50	0.10	0.19	0.32	0.15	0.51	0.98	0.22
		ML estimate	5.06	4.26	4.16	2.53	1.93	1.85	1.38	1.67	3.02	2.67
		MAP estimate	4.01	4.02	4.02	3.97	3.88	3.78	3.65	3.53	3.50	3.44
	sigma = 0.5, sigma_0 = 5	sample	2.03	2.08	1.68	2.81	1.96	1.76	3.09	2.40	2.36	1.50
		true probability	0.52	0.56	0.26	0.95	0.47	0.32	0.99	0.79	0.76	0.16
		ML probability	0.50	0.52	0.31	0.91	0.38	0.28	0.96	0.64	0.59	0.09
		MAP probability	0.48	0.52	0.27	0.93	0.41	0.27	0.98	0.71	0.66	0.09
		ML estimate	2.03	2.05	1.93	2.15	2.11	2.05	2.20	2.23	2.24	2.17
		MAP estimate	2.05	2.05	1.99	2.05	2.07	2.07	2.10	2.13	2.15	2.15
	sigma = 5, sigma_0 = 5	sample	5.94	1.99	2.47	0.11	-5.41	1.78	6.80	10.69	-0.15	-6.14
		true probability	0.78	0.50	0.54	0.35	0.07	0.48	0.83	0.96	0.33	0.05
		ML probability	0.50	0.35	0.42	0.31	0.10	0.55	0.83	0.94	0.28	0.06
		MAP probability	0.58	0.31	0.38	0.25	0.05	0.46	0.82	0.95	0.30	0.05
		ML estimate	5.94	3.96	3.46	2.62	1.02	1.14	1.95	3.05	2.69	1.81
		MAP estimate	4.97	4.46	4.04	3.52	2.74	2.30	2.22	2.40	2.46	2.34

Figure 1: Experimental Results with “wrong” prior

		Number of draws	1	2	3	4	5	6	7	8	9	10
mu_0 = 2	sigma = 0.5, sigma_0 = 0.5	sample	1.42	2.20	2.65	1.70	2.22	1.75	2.05	2.60	2.06	1.48
		true probability	0.12	0.65	0.90	0.28	0.67	0.31	0.54	0.88	0.55	0.15
		ML probability	0.50	0.78	0.87	0.28	0.64	0.32	0.54	0.85	0.49	0.14
		MAP probability	0.28	0.81	0.93	0.32	0.69	0.33	0.56	0.88	0.54	0.14
		ML estimate	1.42	1.81	2.09	1.99	2.04	1.99	2.00	2.07	2.07	2.01
		MAP estimate	1.71	1.76	1.90	1.93	1.97	1.97	1.98	2.00	2.01	2.01
	sigma = 5, sigma_0 = 0.5	sample	1.55	0.74	7.97	5.03	4.70	-5.22	-2.84	3.01	0.26	8.45
		true probability	0.46	0.40	0.88	0.73	0.71	0.07	0.17	0.58	0.36	0.90
		ML probability	0.50	0.47	0.82	0.60	0.56	0.06	0.18	0.59	0.39	0.89
		MAP probability	0.46	0.40	0.88	0.72	0.69	0.07	0.16	0.57	0.36	0.90
		ML estimate	1.55	1.14	3.42	3.82	4.00	2.46	1.70	1.87	1.69	2.36
		MAP estimate	2.00	1.98	2.02	2.09	2.17	2.18	2.16	2.14	2.11	2.13
	sigma = 0.5, sigma_0 = 5	sample	1.79	2.18	2.02	1.82	2.89	2.11	3.37	1.85	2.28	2.79
		true probability	0.34	0.64	0.51	0.36	0.96	0.59	1.00	0.38	0.71	0.94
		ML probability	0.50	0.65	0.52	0.39	0.93	0.48	0.98	0.21	0.52	0.83
		MAP probability	0.50	0.70	0.55	0.39	0.96	0.55	0.99	0.28	0.59	0.88
		ML estimate	1.79	1.99	2.00	1.95	2.14	2.13	2.31	2.25	2.26	2.31
		MAP estimate	1.79	1.92	1.96	1.96	2.02	2.05	2.12	2.15	2.17	2.19
	sigma = 5, sigma_0 = 5	sample	-2.75	5.97	2.36	-1.87	5.87	3.33	0.82	11.39	5.05	1.45
		true probability	0.17	0.79	0.53	0.22	0.78	0.60	0.41	0.97	0.73	0.46
		ML probability	0.50	0.81	0.54	0.29	0.79	0.59	0.41	0.95	0.63	0.37
		MAP probability	0.32	0.86	0.60	0.28	0.82	0.64	0.43	0.97	0.71	0.42
		ML estimate	-2.75	1.61	1.86	0.93	1.92	2.15	1.96	3.14	3.35	3.16
		MAP estimate	-0.38	0.62	1.15	1.07	1.33	1.56	1.65	1.98	2.24	2.41

Figure 2: Experimental Results with “right” prior

We can see from the experiment result that when more samples are drawn, both the ML probability and the MAP probability are approaching the true probability, and at the same time, the both the ML estimate and the MAP estimate are also approaching the true mean, or tend to approach, no matter how initialized. The corresponding figures and analysis details are shown in the coming pages.

- Comparison of results when **prior mean** is initialized to be “wrong” or “right”, with $\sigma = \sigma_0 = 0.5$ fixed.

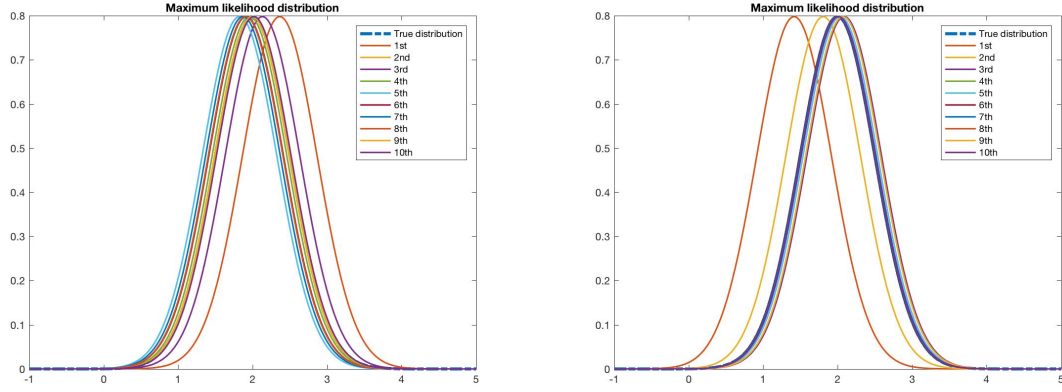


Figure 3: left: ML with “wrong” prior mean(Exp1). right: ML with “right” prior mean(Exp5).

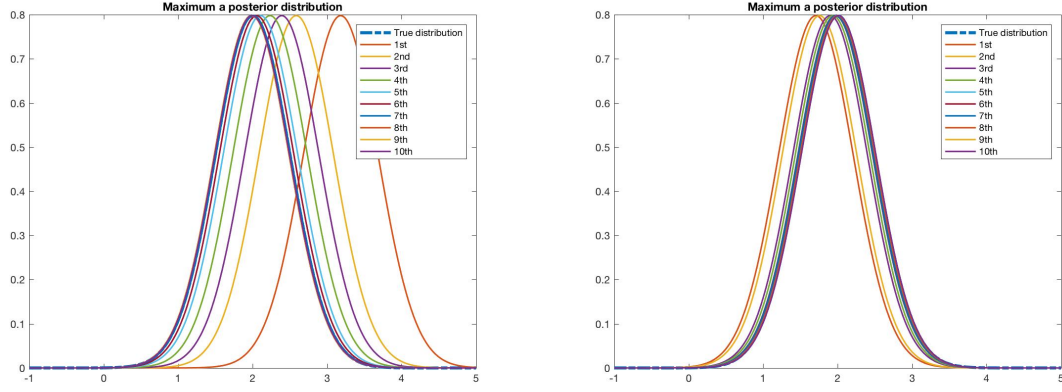


Figure 4: left: MAP with “wrong” prior mean(Exp1). right: MAP with “right” prior mean(Exp5).

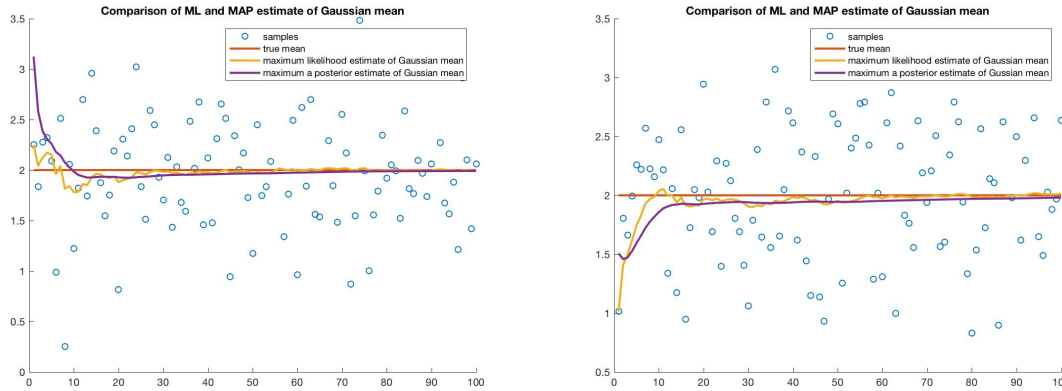


Figure 5: left: comparison with “wrong” prior mean(Exp1). right: comparison with “right” prior mean(Exp5).

Compare the result with experiment 1 and experiment 5 (or equivalently 2 and 6, 3 and 7, 4 and 8), when the

variance for the true distribution and the prior are fixed, and the prior mean is initialized to be “wrong” or “right”, we can see that finally by both approach the distribution from ML and MAP would be very much similar to the true distribution. When the prior of the Gaussian is initialized to be a “wrong” value, at first the MAP distribution is very different from the true distribution, because of the “wrong” initial information, while if the prior of the Gaussian is initialized to be a “right” value, the MAP distribution at first time would seem similar to the true distribution. Since the ML solution based solely on the data, the initialization of the prior mean will not affect the ML distribution. The difference shown in figure are only because of the randomness of sampling.

At the same time we compare the experiment results with larger variance on the prior (experiment 3 and 7), this pair of experiment are ones with larger prior variance compared to the former set.

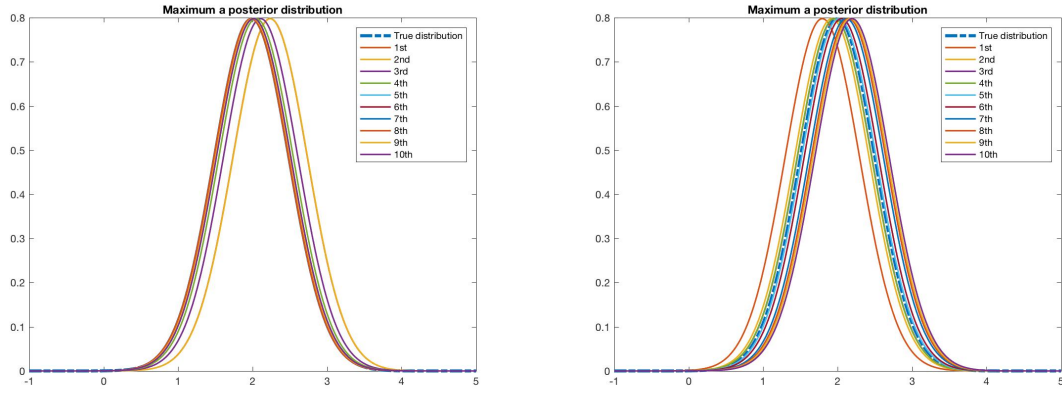


Figure 6: left: comparison with “wrong” prior mean(Exp3). right: comparison with “right” prior mean(Exp7).

Compare the results with these 2 groups of comparison, we can learn that when the variance of the prior is small, it is essential to choose a right value of the prior mean to get a better result of MAP; while when the variance of the prior is large, the choose of prior mean is not that essential.

- Comparison of varying **prior variance** from small to large by comparing the results of experiments 5 and 7, with $\mu_0 = 2$, $\sigma = 0.5$ fixed, $\sigma_0 = 0.5$ (left) or $\sigma_0 = 5$ (right).

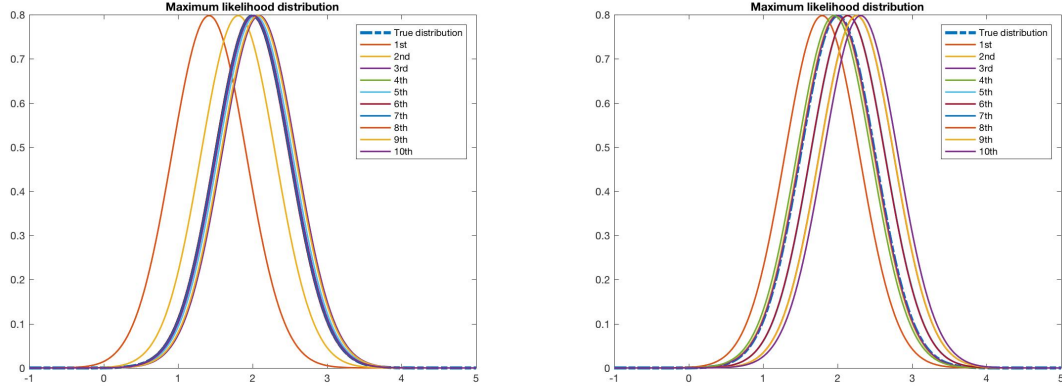


Figure 7: left: ML with small prior variance(Exp5). right: ML with large prior variance(Exp7).

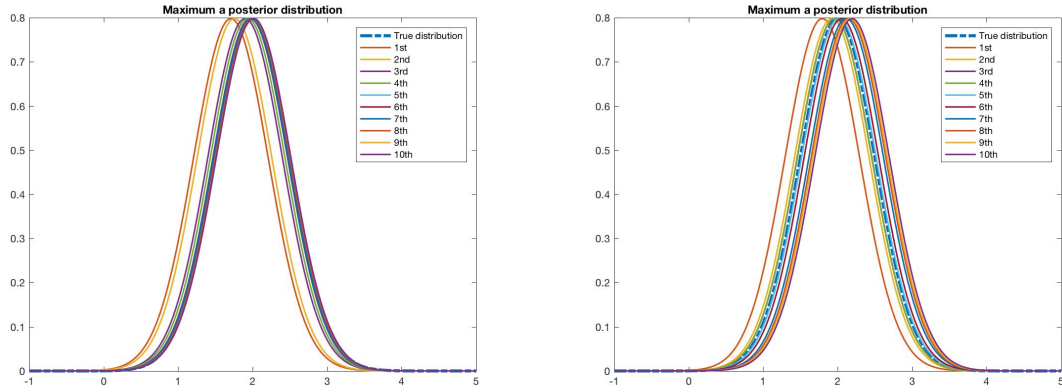


Figure 8: left: MAP with small prior variance(Exp5). right: MAP with large prior variance(Exp7).

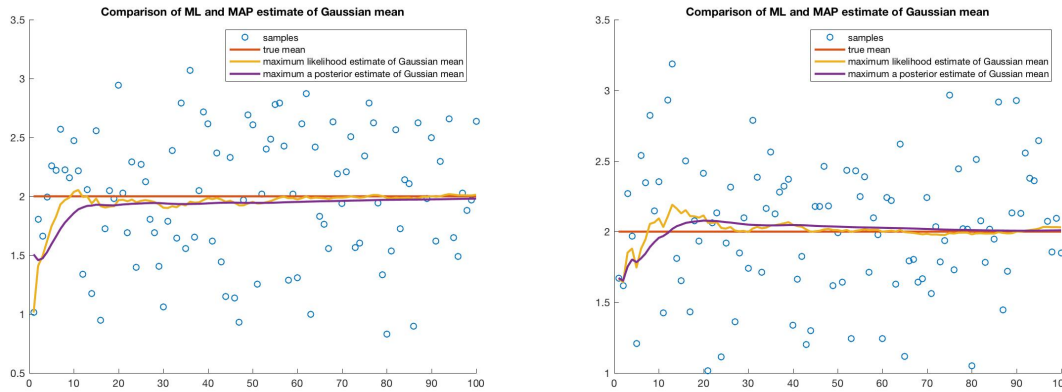


Figure 9: left: comparison with small prior variance(Exp5). right: comparison with large prior variance(Exp7).

In this set of comparison I set the prior of mean to be the “right” value, i.e. $\mu_0 = 2$. We have already known that the likelihood distribution based solely on the data, thus the changing of σ_0 would affect only the MAP result. When enough sample been drawn, both the μ_{ML} and μ_{MAP} would converge to the true mean of the distribution.

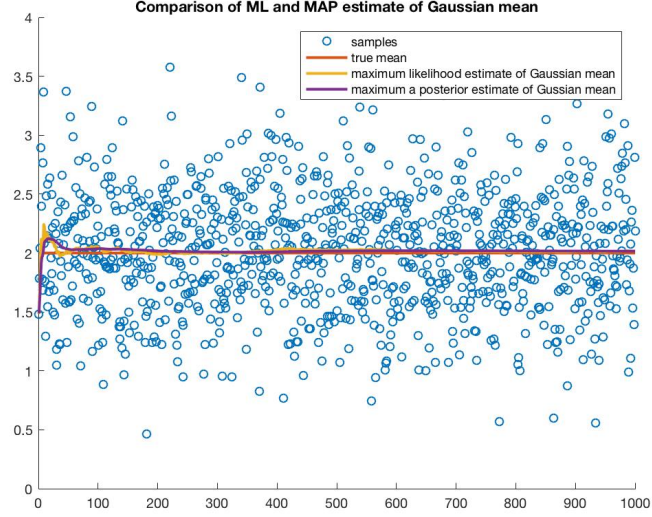


Figure 10: When enough sample being drawn, even start with a big variance of the prior, the estimators would still converge to the true mean

- In this case of comparison of varying **prior variance** from small to large, we can also look on the results when the prior of mean being “wrong” value, i.e. $\mu_0 = 4$. Comparing the results of experiments 1 and 3.

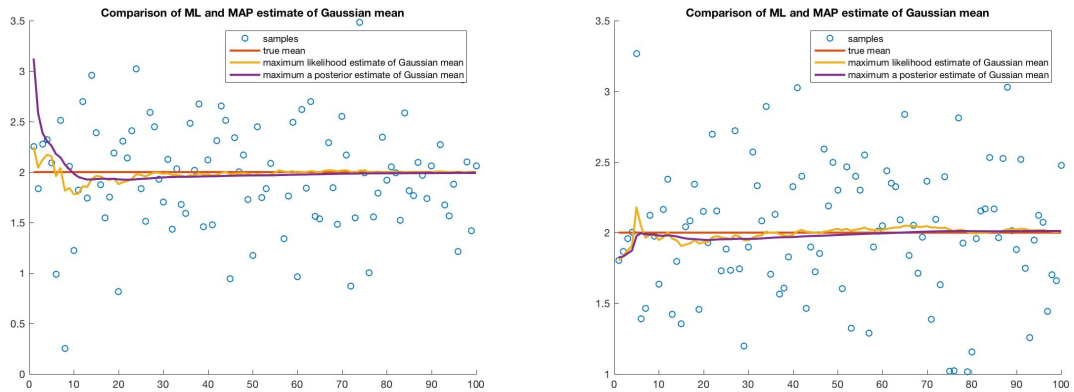


Figure 11: left: comparison with small prior varianc(Exp1). right: comparison with large prior variance(Exp3).

- Comparison of varying the **likelihood variance** from small to large by comparing the results of experiments 5 and 6, with $\mu_0 = 2$, $\sigma_0 = 0.5$ fixed.

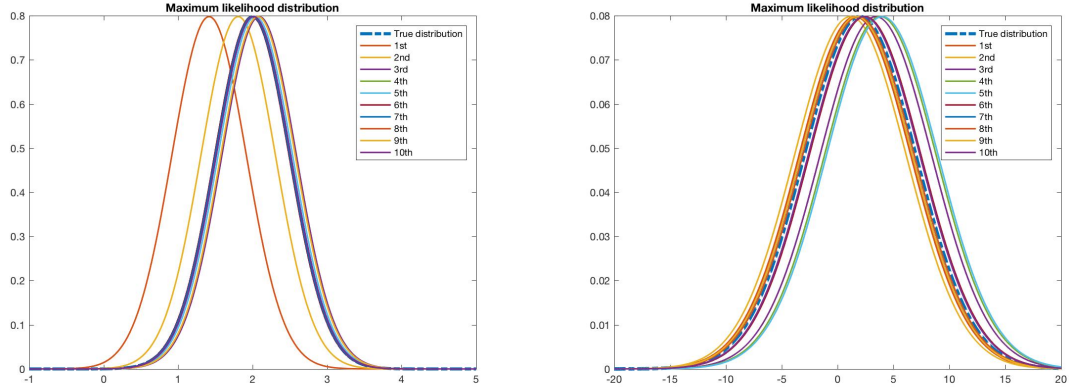


Figure 12: left: ML with small likelihood variance(exp 5). right: ML with large likelihood variance(exp 6).

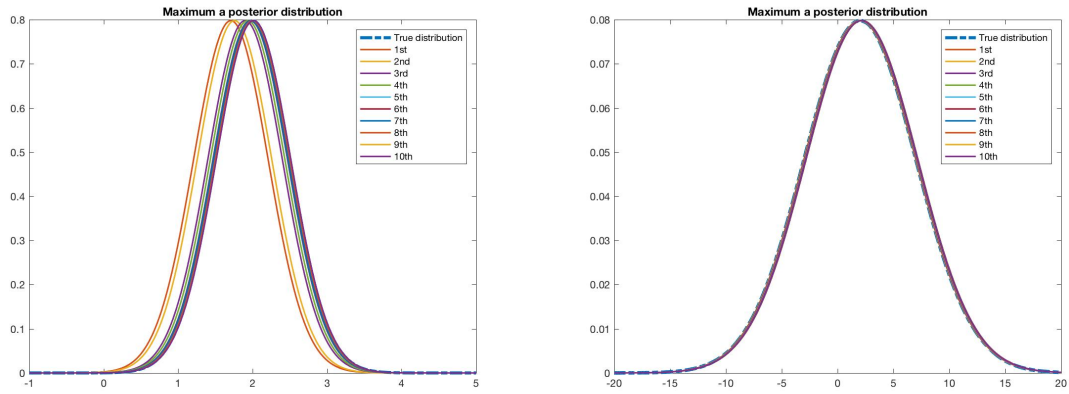


Figure 13: left: MAP with small likelihood variance(Exp 5). right: MAP with large likelihood variance(exp 6).

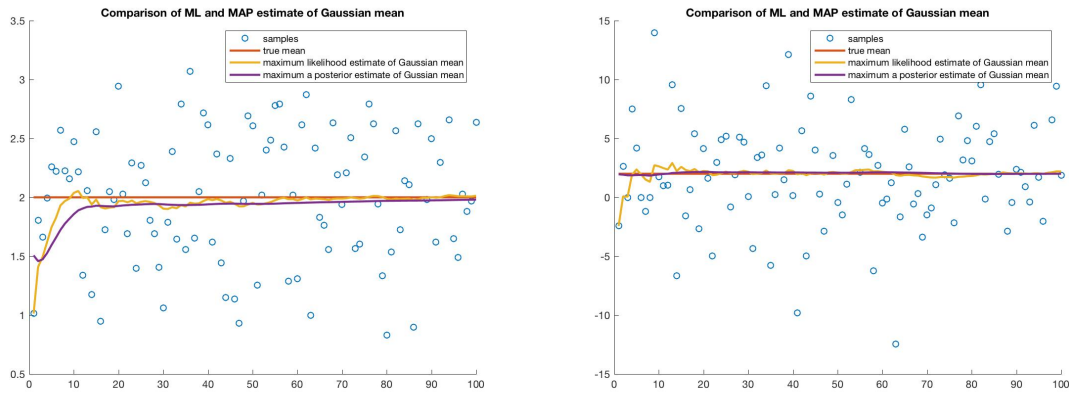


Figure 14: left: comparison with small likelihood variance(Exp 5). right: with large likelihood variance(exp 6).

In this set of comparison I set the prior of mean to be the “right” value, i.e. $\mu_0 = 2$. In this case, when the variance of the likelihood is comparatively large, the MAP estimate of the mean converge much quickly compared to when the variance of the likelihood is small.

- In this case of comparison of varying **likelihood variance** from small to large, we can also look on the results when the prior of mean being “wrong” value, i.e. $\mu_0 = 4$. Comparing the results of experiments 1 and 2.

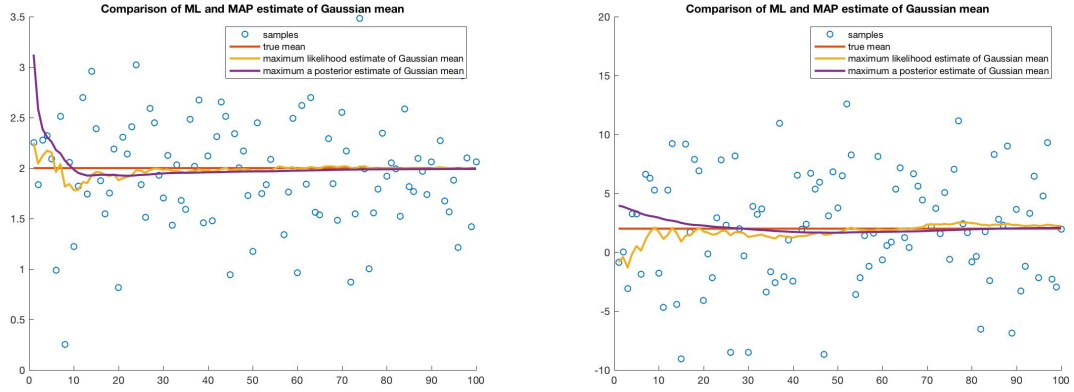


Figure 15: left: comparison with small prior varianc(Exp1). right: comparison with large prior variance(Exp2).

I also included some discussion with regard to the posterior distribution of μ .

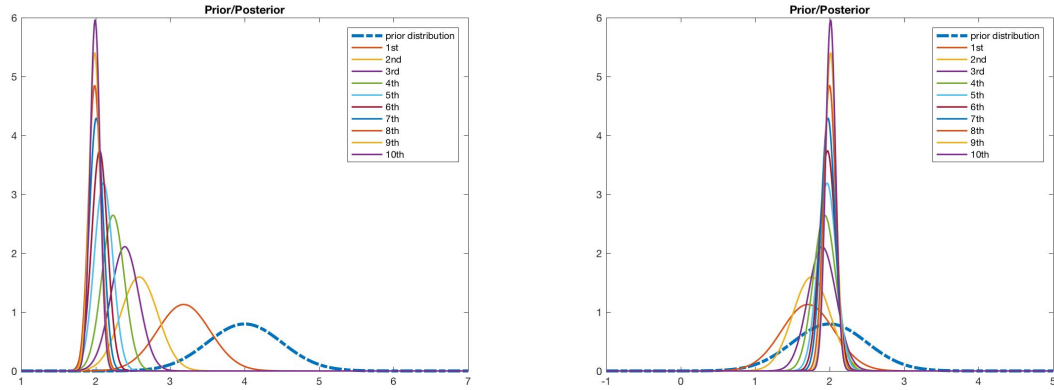


Figure 16: left: posterior distribution of μ (Exp 1). right: posterior distribution of μ (exp 5).

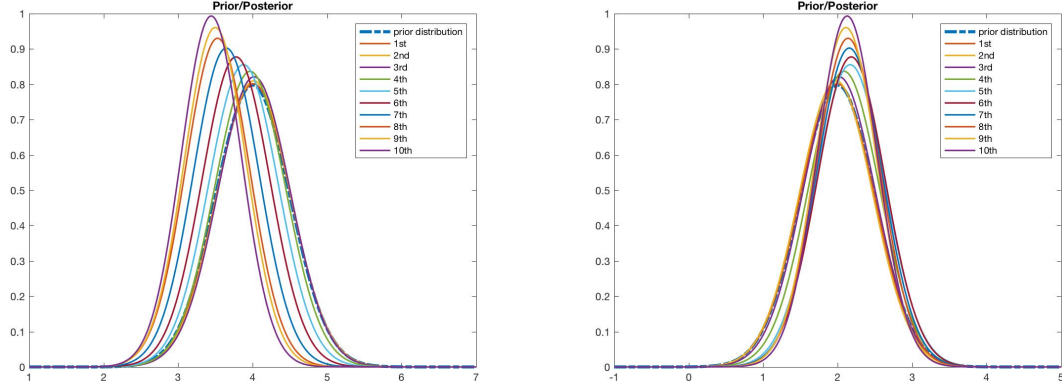


Figure 17: left: posterior distribution of $\mu(\text{Exp } 2)$. right: posterior distribution of $\mu(\text{exp } 6)$.

Based on this comparison, no matter the initialization of prior mean is initialized to a right value or wrong value, after enough iterations, the posterior distribution is center at the true mean, with the variance much smaller adding new sample. Which means that when enough sample is drawn, we become much confident with our posterior distribution of μ . As is also seen from the formula: $\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$. As N becomes larger, the σ_N becomes smaller with every new sample added, which means we are much confident with our estimate.

To conclude, the experiment results justified what we learned from the formulas of posterior mean and posterior variance. Look at the formula of the posterior mean:

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

we learned that the posterior mean is a weighted average of the prior mean and maximum likelihood estimate of the mean. And the reciprocal of the posterior variance is also a weighted combination of the reciprocal of the prior variance and the reciprocal of the likelihood variance. Thus, several point are stated here:

1. When more samples are generated from the distribution we are to estimate the mean from, the weight for the maximum likelihood estimator becomes larger, thus when getting much sample, the maximum a posterior solution would converge to the maximum likelihood solution. And the posterior variance becomes smaller with samples added, which means we become much confident with the knowledge regarding the mean we are to estimate.
2. When the likelihood variance is small, much weight is put on the ML solution, thus the MAP solution get much influence from the data rather than the prior. So in this case, even a wrong prior on the mean is assumed, when the variance of likelihood is small, the MAP is able to converge to ML solution quickly. While conversely, when the variance of likelihood is large, and the assumption on the prior is wrong, it would take longer time for MAP to converge.
3. When the prior variance is small, much weight is put on the prior, which means it is essential to get a prior mean close to the true mean, otherwise it may takes long for MAP solution to converge.

Code that used in this report:

```
clear; close all; clc

exp = 3;

%% true mean and standard deviation of the distribution

num_x = -1:0.01:5;% -20:0.01:20;%

mu = 2;

sigma = 0.5; %5;%standard deviation

true_dist = normpdf(num_x, mu, sigma);

figure(1), plot(num_x, true_dist, '-.', 'Linewidth', 3), ...

hold on, saveas(gcf, ['trueDist' num2str(exp) '.jpg'])

figure(2), plot(num_x, true_dist, '-.', 'Linewidth', 3), ...

hold on, saveas(gcf, ['trueDist' num2str(exp) '.jpg'])

%% initialize the prior of the Gaussian mean

num_mu = -1:0.01:5;

mu_0 = 2;%4;%

sigma_0 = 0.5; %standard deviation5; %

prior = normpdf(num_mu, mu_0, sigma_0);

figure(3), plot(num_mu, prior, '-.', 'Linewidth', 3), ...

hold on, saveas(gcf, ['prior' num2str(exp) '.jpg'])

%% initialize the posterior of the Gaussian mean

mu_N = mu_0;

sigma_N = sigma_0;

posterior = normpdf(num_mu, mu_N, sigma_N);

%%

num_Sample = 100;

sample = [];

true_prob = [];

ML_prob = [];

MAP_prob = [];
```

```

mu_ML = [];
mu_MAP = [];
for iter = 1:num_Sample
    mu_0 = mu_N; %update the prior mean with the posterior mean from the previous draw
    sigma_0 = sigma_N; %update the prior variance(standard deviation) with the posterior
    sample = [sample normrnd(mu,sigma)]; %get a sample from the true distribution
    N = length(sample);
    mu_ML = [mu_ML mean(sample)]; %ML solution for the Gaussian mean
    ML = normpdf(num_x, mu_ML(N), sigma);
    figure(1), plot(num_x, ML, 'Linewidth', 1.5), hold on
    mu_N = mu_0*sigma^2/(N*sigma_0^2 + sigma^2) ...
+ mu_ML(N)*N*sigma_0^2/(N*sigma_0^2 + sigma^2); %MAP solution for the Gaussian mean
    sigma_N = sqrt(1/(1/sigma_0^2+N/sigma^2)); %standard deviation of the posterior
    MAP = normpdf(num_x, mu_N, sigma);
    figure(2), plot(num_x, MAP, 'Linewidth', 1.5), hold on
    mu_MAP = [mu_MAP mu_N];
    posterior_ = normpdf(num_mu, mu_N, sigma_N);
    figure(3), plot(num_mu, posterior_, 'Linewidth', 1.5), hold on
    display(['random data point from Gaussian is ' num2str(sample(N))])
    display(['Maximum likelihood solution of mean is ' num2str(mu_ML)])
    display(['Maximum a posterior solution of mean is ' num2str(mu_N)])
    true_prob = [true_prob cdf('Normal',sample(N),mu,sigma)];
    ML_prob = [ML_prob cdf('Normal',sample(N),mu_ML(N),sigma)];
    MAP_prob = [MAP_prob cdf('Normal',sample(N),mu_N,sigma)];
    display(['True probablity of drawing ' num2str(sample(N)) ' ...
from Gaussian is ' num2str(cdf('Normal',sample(N),mu,sigma))])
    display(['ML probablity of drawing ' num2str(sample(N)) ' ...
from Gaussian is ' num2str(cdf('Normal',sample(N),mu_ML(N),sigma))])
    display(['MAP probability of drawing ' num2str(sample(N)) ' ...

```

```

        from Gaussian is ' num2str(cdf('Normal',sample(N),mu_N,sigma)))]
end

figure(1), title('Maximum likelihood distribution '), ...
saveas(gcf, ['ML' num2str(exp) '.jpg'])

figure(2), title('Maximum a posterior distribution '), ...
saveas(gcf, ['MAP' num2str(exp) '.jpg'])

figure(3), title('Prior/Posterior '), ...
saveas(gcf, ['posterior ' num2str(exp) '.jpg'])

tosave = [sample;true_prob; ML_prob; MAP_prob; mu_ML; mu_MAP];

figure(4), scatter(1:N, sample), hold on, plot(mu*ones(1,N), 'Linewidth', 2), hold on, pl

    hold on, plot(mu_MAP, 'Linewidth', 2)

title('Comparison of ML and MAP estimate of Gaussian mean')

saveas(gcf, [num2str(num_Sample) 'comparison ' num2str(exp) '.jpg'])

```