Assignment 1

Machine learning
708.064 19S

Last name	First name	Matr. Number
Konanur Ramanna	Keerthi Datta	01647641
Gajanovic	Stefan	01431869

1. Probabilistic calculus on Gaussian distributions

① Basic probability calculus in one dimension

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \leftarrow \text{Gaussian distribution } \mathcal{N}(x|\mu,\sigma)$$

a) $H(p) = \frac{1}{2}\left(\log(2\pi e\sigma^2)\right)$

$* \int p(x)dx = 1$

$** \int x p(x)dx = \mu$

$*** \int x^2 p(x)dx = \sigma^2 + \mu^2$

$$H(x) = -\int p(x)\log p(x)\,dx$$

$$= -\int p(x)\log\left(\frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}\right)$$

$$= -\int p(x)\cdot\log\left(\exp\left(-\frac{(x-\mu^2)}{2\sigma^2}\right)\right) - p(x)\cdot\log\sqrt{2\pi\sigma^2}\,dx$$

$$= -\int \log e\cdot p(x)\cdot\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx + \underbrace{\int p(x)}_{=1\,*}\log\sqrt{2\pi\sigma^2}\,dx$$

$$= -\log e\cdot\frac{1}{2\sigma^2}\int p(x)\cdot(-x^2+2x\mu-\mu^2) + \frac{1}{2}\log 2\pi\sigma^2$$

$$= -\log e\cdot\frac{1}{2\sigma^2}\left[\underbrace{\int -p(x)x^2dx}_{-\sigma^2-\mu^{2\,***}} + \underbrace{\int p(x)2x\mu dx}_{2\mu^{2}\,**} + \underbrace{\int p(x)\mu^2dx}_{-\mu^{2}\,*}\right] + \frac{1}{2}\log 2\pi\sigma^2$$

$$= -\log e\,\frac{1}{2\sigma^2}\left[-\sigma^2-\mu^2+2\mu^2-\mu^2\right] + \frac{1}{2}\log 2\pi\sigma^2$$

$$= \log e\,\frac{\sigma^2}{2\sigma^2} + \frac{1}{2}\log 2\pi\sigma^2 = \frac{1}{2}\log e + \frac{1}{2}\log 2\pi\sigma^2 = \underline{\underline{\frac{1}{2}\log(2\pi e\sigma^2)}}$$

b) RVs X and Y and $Z = X + Y$

$$\mathbb{E}(X) = \mu_x \quad \mathbb{E}(Y) = \mu_y$$

$$\mathrm{Var}(X) = \sigma_x^2 \quad \mathrm{Var}(Y) = \sigma_y^2$$

$$\mathbb{E}(Z) = \mathbb{E}(X+Y) = \mu_x + \mu_y$$

$$\mathrm{Var}(Z) = \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = \sigma_x^2 + \sigma_y^2$$

c) $H(p_x, p_y) = \mathbb{E}_{p_x}(-\log_2 p(x)) = -\mathbb{E}(f)$

$$H(p_x, p_y) = \int p(x)\left(\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)\log e\, dx + \frac{1}{2}\log 2\pi\sigma_y^2$$

since $H(p_x, p_y) = \int p(x)\log p(y)$ so we have the expression from task 1.1.a).

$$\int p(x) f(x)\,dx = \mathbb{E}(f(x))$$

$$\Rightarrow \mathbb{E}\left(\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)\log e + \frac{1}{2}\log 2\pi\sigma_y^2$$

$$= \frac{\log e}{2}\underbrace{\mathbb{E}\left((x-\mu_y)^2\right)}_{\sigma_y^2} + \frac{1}{2}\log 2\pi\sigma_y^2$$

$$= \frac{1}{2}\log e + \frac{1}{2}\log 2\pi\sigma_y^2 \Rightarrow H(p_x, p_y) = \frac{1}{2}\log 2\pi e\sigma_y^2 = H(p_y)$$

d) $D(p_x \| p_y) = H(p_x) - H(p_x, p_y)$

$$= \frac{1}{2}\log 2\pi e\sigma_x^2 - \frac{1}{2}\log 2\pi e\sigma_y^2$$

$$= \frac{1}{2}\log\left(\frac{2\pi e\sigma_x^2}{2\pi e\sigma_y^2}\right) = \log\left(\frac{\sigma_x}{\sigma_y}\right)$$

② Conditional Gaussian

$z \in \mathbb{R}^2 \qquad z = \begin{bmatrix} x \\ y \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$

$\Delta \overset{\text{def.}}{=} \Sigma^{-1} \qquad \Delta = \begin{bmatrix} \lambda_{xx} & \lambda_{xy} \\ \lambda_{yx} & \lambda_{yy} \end{bmatrix}$

a) $p(y)$ using marginal probability distribution.

$$p(y) = \int p(x,y)\,dx$$

$$p(y) = \int \frac{1}{2\pi\sqrt{\Sigma}} \exp\left(-\frac{1}{2}(z-\mu)^T \Delta (z-\mu)\right) dx$$

$$= \int \frac{1}{2\pi\sqrt{\Sigma}} \exp\left(-\frac{1}{2}\begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}^T \Delta \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}\right) dx$$

$$= \frac{1}{2\pi\sqrt{\Sigma}} \int \exp\left(-\frac{1}{2}\left((x-\mu_x)^T \Delta_{xx}(x-\mu_x) + (x-\mu_x)^T \Delta_{xy}(y-\mu_y) + (y-\mu_y)^T \Delta_{yx}(x-\mu_x) + (y-\mu_y)^T \Delta_{yy}(y-\mu_y)\right)\right) dx$$

- completing the squares method:

$$z = x - \mu_x \qquad A = \Delta_{xx} \qquad b = \Delta_{xy}(y-\mu_y)$$

$$c = \frac{1}{2}(y-\mu_y)^T \Delta_{yy}(y-\mu_y)$$

$$\Rightarrow \frac{1}{2} z^T A z + b^T z + c = \frac{1}{2}(z + A^{-1}b)^T A (z + A^{-1}b) + c - \frac{1}{2} b^T A^{-1} b.$$

$$p(y) = \exp\left(-\frac{1}{2}(y-\mu_y)^T \Delta_{yy}(y-\mu_y) + \frac{1}{2}(y-\mu_y)^T \Delta_{yx} \Delta_{xx}^{-1} \Delta_{xy}(y-\mu_y)\right) \cdot \frac{1}{2\pi\sqrt{\Sigma}} \int \exp\left(\frac{1}{2}\left[(x-\mu_x + \Delta_{xx}^{-1}\Delta_{xy}(y-\mu_y))^T \Delta_{xx}(x-\mu_x + \Delta_{xx}^{-1}\Delta_{xy}(y-\mu_y))\right]\right) dx$$

we know: $\frac{1}{2\pi\sqrt{\Sigma}} \int \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) = 1$ so,

$$\int \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) = 2\pi\sqrt{\Sigma}$$

- now can get rid of the integral part

$$p(y) = \frac{1}{2\pi\sqrt{\Sigma}} \cdot 2\pi\sqrt{\Delta_{xx}} \cdot \exp\left(-\frac{1}{2}(y-\mu_y)^T(\Delta_{yy} - \Delta_{yx}\Delta_{xx}^{-1}\Delta_{xy})(y-\mu_y)\right)$$

at this point we can continue to multiply/expand the equation!

b) $p(x|y)$ using the product rule.

$$= \frac{p(x,y)}{p(y)} = \frac{\frac{1}{2\pi|\Sigma|}\cdot \exp\left(-\frac{1}{2}(z-\mu)^T\Delta(z-\mu)\right)}{\int p(x,y)dx} \leftarrow \text{does not depend on } x. = z$$

$$= \frac{1}{z}\cdot\exp\left(-\frac{1}{2}(z-\mu)^T\Delta(z-\mu)\right)$$ following the same expansion procedure from the 1.2.a)

$$= \frac{1}{z}\cdot\exp\Big(-\Big[\frac{1}{2}(x-\mu_x)^T\Delta_{xx}(x-\mu_x)+\frac{1}{2}(x-\mu_x)^T\Delta_{xy}(y-\mu_y)+$$
$$\frac{1}{2}(y-\mu_y)^T\Delta_{yx}(x-\mu_x)+\frac{1}{2}(y-\mu_y)^T\Delta_{yy}(y-\mu_y)\Big]\Big)$$

again following completing the squares.

$$= \frac{1}{z}\cdot\exp\Big(-\Big[\frac{1}{2}\Big(x-\mu_x+\Delta_{xx}^{-1}\Delta_{xy}(y-\mu_y)\Big)^T\Delta_{xx}\Big(x-\mu_x+\Delta_{xx}^{-1}\Delta_{xy}(y-\mu_y)\Big)+$$
$$+\frac{1}{2}(y-\mu_y)^T\Delta_{yy}(y-\mu_y)-\frac{1}{2}(y-\mu_y)^T\Delta_{yx}\Delta_{xx}^{-1}\Delta_{xy}(y-\mu_y)\Big]\Big) =$$

$$= \frac{1}{z}\exp\Big(-\frac{1}{2}\Big(x-\underbrace{\mu_x+\Delta_{xx}^{-1}\Delta_{xy}(y-\mu_y)}_{\mu_{x|y}}\Big)^T\underbrace{\Delta_{xx}}_{\Sigma_{x|y}}\Big(x-\mu_x+\Delta_{xx}^{-1}\Delta_{xy}(y-\mu_y)\Big)\Big) \quad *$$

now contains more terms indep. on x.

c) $p(x|y)$ is gaussian $\quad \mu_{x|y} = ? \quad \Sigma_{x|y}$

- partitioned matrices

$$\begin{bmatrix} A & B \\ C & 0 \end{bmatrix}^{-1} = \begin{bmatrix} M^{-1} & -M^{-1}BD^{-1} \\ -D^{-1}CM^{-1} & D^{-1}+D^{-1}CM^{-1}BD^{-1} \end{bmatrix} \text{ where } M = A-BD^{-1}C, \text{ so}$$

$$\begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} = \begin{bmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{bmatrix}^{-1} = \begin{bmatrix} (\Lambda_{xx}-\Lambda_{xy}\Lambda_{yy}^{-1}\Lambda_{yx})^{-1} & -(\Lambda_{xx}-\Lambda_{xy}\Lambda_{yy}^{-1}\Lambda_{yx})^{-1}\Lambda_{xy}\Lambda_{yy}^{-1} \\ -(\Lambda_{yy}^{-1}\Lambda_{yx}(\Lambda_{xx}-\Lambda_{xy}\Lambda_{yy}^{-1}\Lambda_{yx})^{-1} & (\Lambda_{yy}-\Lambda_{yx}\Lambda_{xx}^{-1}\Lambda_{xy})^{-1} \end{bmatrix}$$

by looking at * we can say that $\mu_{x|y}$ is:

$$\mu_x + \underbrace{\Delta_{xx}^{-1}\Delta_{xy}}_{=\alpha}(y-\mu_y) = \mu_x + \frac{-(\Lambda_{xx}-\Lambda_{xy}\Lambda_{yy}^{-1}\Lambda_{yx})^{-1}\Lambda_{xy}\Lambda_{yy}^{-1}}{(\Lambda_{xx}-\Lambda_{xy}\Lambda_{yy}^{-1}\Lambda_{yx})^{-1}}(y-\mu_y) = \mu_x + \Lambda_{xy}\Lambda_{yy}^{-1}(y-\mu_y)$$

$$= \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}\cdot(y-\mu_y) = \mu_x + \alpha\cdot\frac{1}{\Sigma_y}(y-\mu_y)$$

also looking at the * we see that λ_{xx}^{-1} is our $\sigma_{x|y}$ as expected.

$$\sigma_{x|y} = \lambda_{xx}^{-1} = \frac{1}{(\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})^{-1}} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} = \sigma_x - \alpha^2 \cdot \frac{1}{\sigma_y}$$

③a) Having defined mean and covariance, we constructed our transformation matrix Q and used it as "Q * point + mean" to transform the points sampled from normal distribution. We used np.random.rand() to generate 2-D point of mean 0 and variance 1.

Empirical mean from this points is: $[1.957, -0,922]$ while the covariance matrix is: $\begin{bmatrix} 2.696 & 0.751 \\ 0.751 & 1.429 \end{bmatrix}$

plot: see appendix 1.

b) see appendix 2.

3 a) - Additionally,

Eigen vectors and eigen values of Σ is calculated and multiplied by random number.

It can be written as.

$$Y = \sqrt{\lambda} \cdot \Phi \cdot X$$

where λ = diagnol matrix made up of eigen ~~vectors~~ values of Σ

Φ = Matrix of eigen vectors.

(2.1) - Given data points $x_i \in R^D$, targets $t_i \in R$, the linear transformation $y(x_i, w) = x_i^T w$

- $t_i = y(x_i, w) + \epsilon$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

a] - Each target response 't' becomes a draw from the following gaussian $\rightarrow$ $t \sim \mathcal{N}(y(x, w), \sigma^2)$

$$\Rightarrow p(t \mid x^T w, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(t - x^T w)^2}{2\sigma^2}\right\}$$

- For the data & target $x_1 \ldots x_N$ and $t_1 \ldots t_N$, each data point is assumed to be identical & independent (i.i.d)

$$- p(t_i \mid x_i, w, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(t_i - x_i^T w)^2}{2\sigma^2}\right\}$$

- Taking log likelihood, of data, & solving,

$$\log p(t_i \mid x_i, w, \sigma^2) = -\sum_{i=1}^{N} \frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (t_i - x_i^T w)^2$$

$$\Rightarrow \log p(t_i \mid x_i, w, \sigma^2) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{N}{\sigma^2} \cdot \frac{1}{2N} \cdot \sum_{i=1}^{N} (t_i - x_i^T w)^2$$

where, $\rightarrow \log p(t_1 \ldots t_N \mid x_1 \ldots x_N, w, \sigma^2)$

$$= \alpha E + \beta.$$

$$\hookrightarrow \alpha = -\frac{N}{\sigma^2}, \quad \beta = -\frac{N}{2} \log(2\pi\sigma^2)$$

$$E = \frac{1}{2N} \sum_{i=1}^{N} (t_i - x_i^T w)^2$$

2.1

6] Taking the derivative of log likelihood with respect to α and equating it to 0.

$$\nabla_\alpha \log\left(p(t_i|x_i,\alpha,w)\right) = -\frac{N}{2}\cdot\frac{1}{2\pi\alpha^2}\cdot 4\pi\alpha + \frac{1}{\alpha^3}\left\{\sum_{i=1}^{N}\left(t_i - y(x_i,w)\right)^2\right.$$

$$-\frac{N}{\alpha} + \frac{1}{\alpha^3}\left\{\sum_{i=1}^{N}\left(t_i - y(x_i,w)\right)^2 = 0\right.$$

$$\frac{1}{\alpha^3}\sum_{i=1}^{N}\left(t_i - y(x_i,w)\right)^2 = \frac{N}{\alpha}$$

$$\boxed{\alpha_{ML} = \frac{1}{N}\sum_{i=1}^{N}\left(t_i - y(x_i,w)\right)^2}$$

- Variance is directly proportional to uncertainty
- MLE & L.S are same for Gaussian distribution. But when the distribution is not gaussian, MLE is better than computing least squares.

2.2 Given, posterior distribution = $p(w|x_i, t_i)$

Prior distribution = $p(w) \sim \mathcal{N}(\mu_w, \frac{1}{\alpha} I)$

likelihood = $p(t_i|x_i, w)$

a] $p(w|x_i, t_i) = p(t_i|x_i, w) \cdot p(w)$

WKT, $p(t_i|x_i, w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2\sigma^2} \cdot (t_i - x_i^T w)^T (t_i - x_i^T w)\right\}$

and $p(w) = \frac{1}{\sqrt{2\pi\alpha^{-1}}} \cdot \exp\left\{\frac{-\alpha}{2}(w - \mu_w)^T (w - \mu_w)\right\}$

substituting prior & likelihood, taking log- & solving.

log posterior will be following: (likelihood from)

$$\log p(w|x_i, t_i) = -\frac{1}{2\sigma^2}\|t - Xw\|^2 - \frac{\alpha}{2}\|w - \mu_w\|^2 - \frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log\left(\frac{2\pi}{\alpha}\right)$$

(from prior; constant)

where $t = [t_1 \ldots t_N]^T$, $t \in R^N$

$X = \text{vert. cat of } x_i^T$ $\quad X \in R^{N \times D}$

using $(a-b)^T(a-b) = \|a-b\|^2$

2.2 b] $\log p(w|x_i,t_i) = -\frac{1}{2\sigma^2}\|t - Xw\|^2$

$$-\frac{\alpha}{2}\|w-\mu_w\|^2 + \text{const.}$$

which can be written as, (ignoring const).

$$\log p(w|x_i,t_i) = -\frac{1}{2\sigma^2}(t-Xw)^T(t-Xw)$$

$$-\frac{\alpha}{2}(w-\mu_w)^T(w-\mu_w)$$

$$= -\frac{1}{2\sigma^2}(t^Tt + w^TX^TXw - w^TX^Tt$$

$$- t^TXw) - \frac{\alpha}{2}(w^Tw - 2\mu_w^Tw + \mu_w^T\mu_w)$$

Taking derivative w.r.t w,

$$\Rightarrow \nabla_w \log p(w|x_i,t_i) = -\frac{1}{2\sigma^2}(0 + 2X^TXw - 2X^Tt) - \frac{\alpha}{2}(2w - 2\mu_w) = 0$$

$$\Rightarrow \frac{-X^TXw + X^Tt}{\sigma^2} - \alpha w + \alpha\mu_w = 0 \quad -(1)$$

$$\left(\alpha + \frac{X^TX}{\sigma^2}\right)w = \frac{X^Tt}{\sigma^2} + \alpha\mu_w$$

$$\boxed{w = \left(\alpha + \frac{X^TX}{\sigma^2}\right)^{-1}\left(\frac{X^Tt}{\sigma^2} + \alpha\mu_w\right)}$$

or solving (1), leads to $w^{MAP} = (X^TX + \lambda I)^{-1}(X^Tt + \lambda\mu_w)$

where $\lambda = \sigma^2\alpha$.

Matrix $A = (X^T X + \lambda I)$ Can be ~~I~~ invertible because, if there exists another matrix B & multiplication of $(A \cdot B)$ is an Identity matrix, which is possible only if both A & B are square matrices. Also, 'I' inside 'A' gives hint that the resulting matrix 'A' is a square matrix.

2.2 C] Posterior distribution

$$p(w|x,t) \propto \prod p(t_i|x_i,w) \cdot p(w)$$

$$p(w|x,t) \propto \exp\left\{-\frac{1}{2\sigma^2}(t - X^T w)^T (t - X^T w)\right\} \cdot \exp\left\{-\frac{\alpha}{2}(w - \mu_w)^T (w - \mu_w)\right\}$$

(ignoring const)

$$\Rightarrow = \exp\left\{-\frac{1}{2\sigma^2}(t^T t + w^T X^T X w - w^T X^T t - t^T X w) - \frac{\alpha}{2}(w^T w - 2\mu_w^T w + \mu_w^T \mu_w)\right\}$$

$$\Rightarrow = \exp\left\{-\left(t^T t + w^T\left(\frac{X^T X}{\sigma^2} + \alpha I\right) w - 2w^T\left(\frac{X^T t}{\sigma^2} + \alpha \mu_w\right)\right)\right\} \quad -(1)$$

$$(w - \mu)^T \Sigma (w - \mu) = w^T \Sigma w - 2\mu^T \Sigma w$$

using the above form,

$$\Sigma = \frac{X^T X}{\sigma^2} + \alpha I \quad -(2)$$

$$\left(\frac{X^T X}{\alpha^2} + \alpha I\right) \mu = \left(\frac{X^T t}{\alpha^2} + \alpha M_0\right)$$

$$\mu = \left(\frac{X^T X}{\alpha^2} + \alpha I\right)^{-1} \left(\frac{X^T t}{\alpha^2} + \alpha M_0\right) \quad \text{— (3)}$$

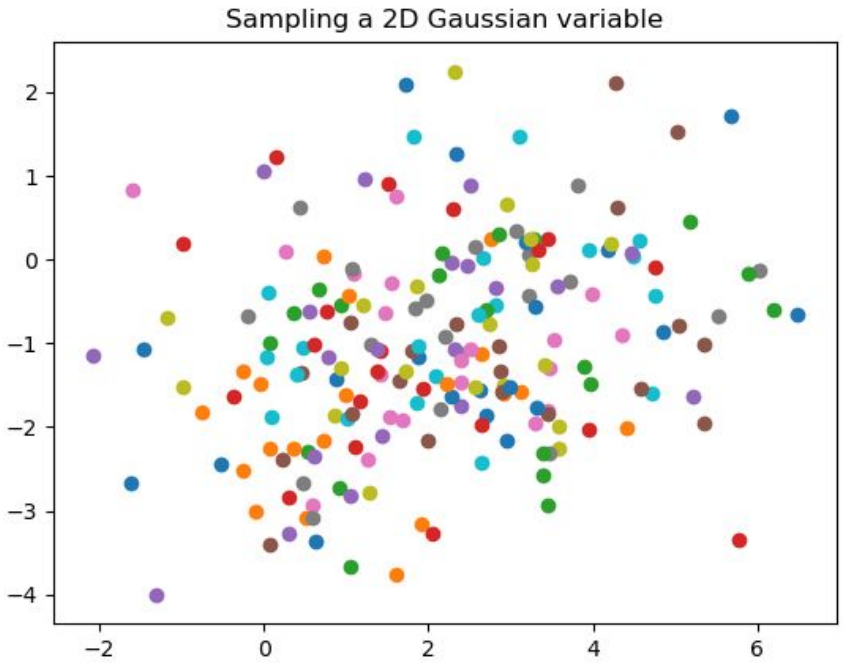
Posterior distribution is also normal distribution as seen in (1), with mean 'μ' given in (3) & Covariance given in (2)

- Mean of the posterior distribution is exactly same as the maximum a posterior as seen in equation (3) & previous computation in 2.2.b.

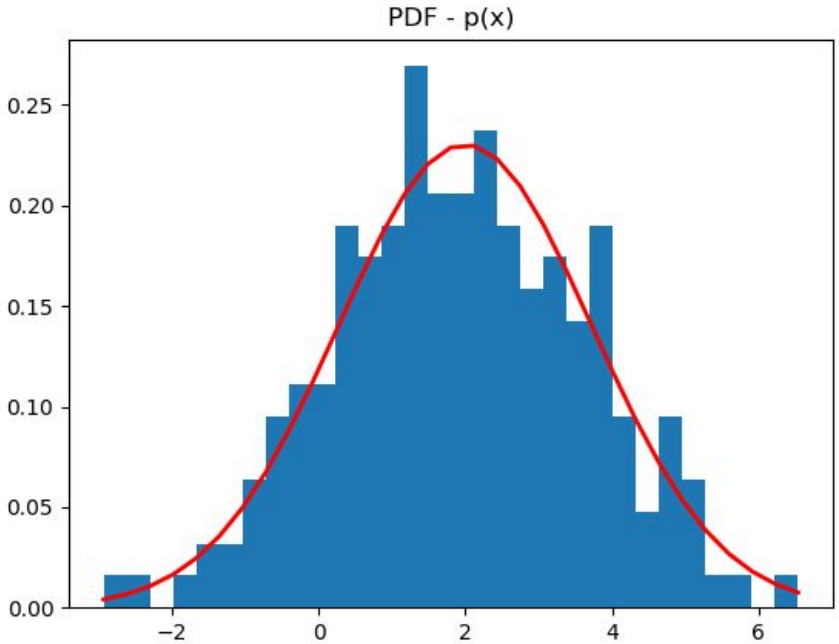
3] plots - Implementation results are shown in Appendix.

Appendix

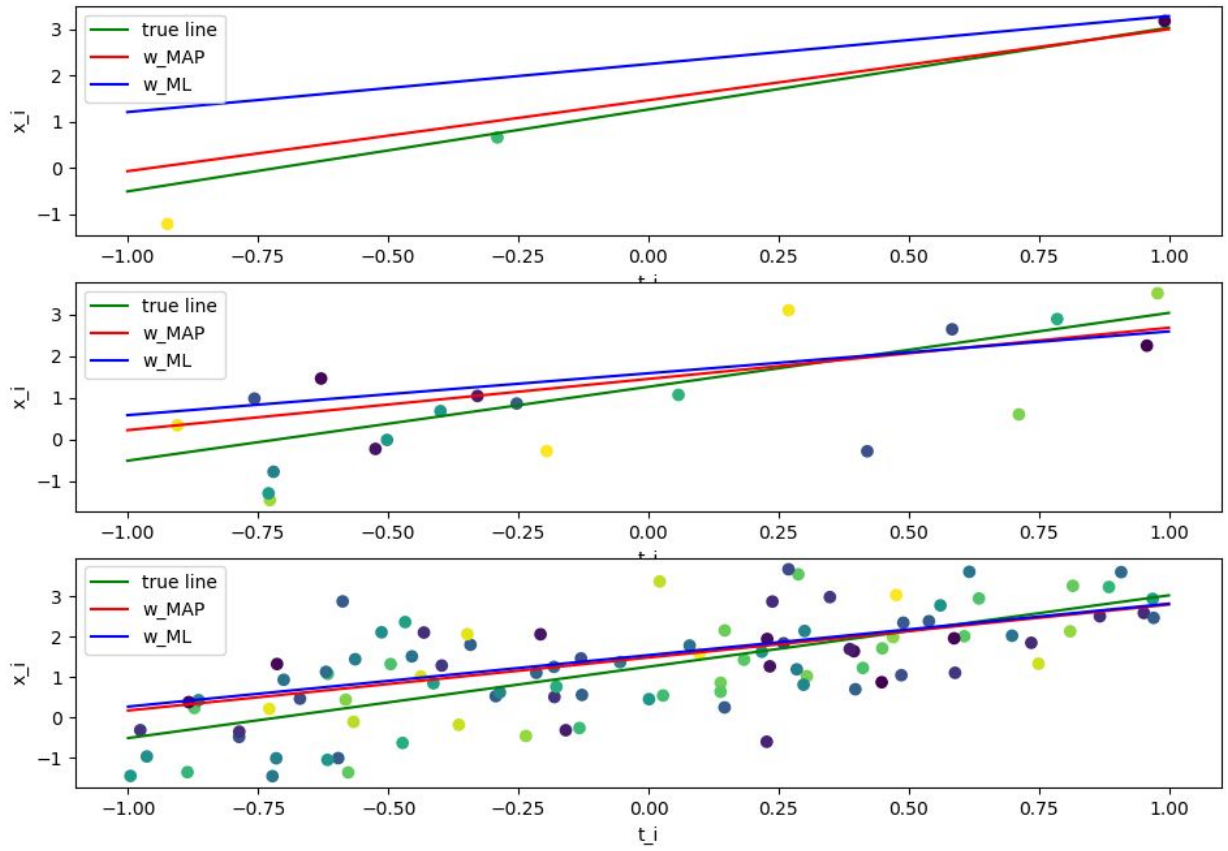
1.3.a Sampling a 2D Gaussian variable



1.3.b



3.



Results:

```
Empirical mean:  [[ 1.80159223]
 [-1.23820129]]
empirical cov:  [[ 3.3531353   0.93371641]
 [ 0.93371641  1.58085217]]
(2,)
For num_points =  3
MAP Estimate  [ 1.46092064  1.5337766 ]
ML Estimate  [ 2.24220258  1.03709472]
Posterior mean:  [ 1.46092064  1.5337766 ] Posterior cov:  [[ 0.0839218   0.00142275]
 [ 0.00142275  0.0769472 ]]
(2,)
For num_points =  20
MAP Estimate  [ 1.4485576   1.22674695]
ML Estimate  [ 1.58391414  1.00293012]
Posterior mean:  [ 1.4485576   1.22674695] Posterior cov:  [[ 0.05641378  0.00355921]
 [ 0.00355921  0.03355789]]
(2,)
For num_points =  100
MAP Estimate  [ 1.48731372  1.31387621]
ML Estimate  [ 1.5471219   1.27688859]
Posterior mean:  [ 1.48731372  1.31387621] Posterior cov:  [[ 0.02527043  0.00062409]
 [ 0.00062409  0.00910632]]
```