۱.۱) مدل‌های خطی (پرسپترون)

مرحله i ام بروزرسانی :
$$\omega_i = \omega_{i-1} + \underbrace{x^{(l_i)} y^{(l_i)}}_{\text{missclassified یک نمونه‌ای که شده است}}$$

$$\xrightarrow{\text{کل بروزرسانی‌ها}} \begin{cases} \omega_1 = \omega_0 + x^{(l_1)} y^{(l_1)} \\ \vdots \qquad \vdots \qquad \vdots \\ \omega_k = \omega_{k-1} + x^{(l_k)} y^{(l_k)} \end{cases} \qquad (*)$$

دو طرف $(*)$ را نرم گرفته و به توان ۲ می‌رسانیم :

$$\begin{cases} \|\omega_1\|^2 = \|\omega_0\|^2 + \|x^{(l_1)} y^{(l_1)}\|^2 + 2\omega_0^T x^{(l_1)} y^{(l_1)} \\ \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \\ \|\omega_k\|^2 = \|\omega_{k-1}\|^2 + \|x^{(l_k)} y^{(l_k)}\|^2 + 2\omega_{k-1}^T x^{(l_k)} y^{(l_k)} \end{cases}$$

$$\underset{(\text{فرض } \omega_0 = 0)}{\xrightarrow{\text{جمع}}} \|\omega_k\|^2 = \underbrace{\sum_{i=1}^{k} \|x^{(l_i)} y^{(l_i)}\|^2}_{\le kr^2} + \underbrace{2\sum_{i=1}^{k} \omega_{i-1}^T x^{(l_i)} y^{(l_i)}}_{<0}$$

طبق فرض گزاره $\leftarrow \le kr^2$

چون یک داده است که غلط دسته‌بندی شده است $\leftarrow <0$

$$\rightarrow \|\omega_k\|^2 \le kr^2 \qquad (1)$$

حال اینبار روابط (*) را با هم جمع کرده و با فرض $\omega_0 = 0$ خواهیم داشت:

$$\omega_k = x^{(l_1)} y^{(l_1)} + \dots + x^{(l_k)} y^{(l_k)}$$

$$\xrightarrow{\times \omega^{*T}} \omega^{*T} \omega_k = \underbrace{\omega^{*T} x^{(l_1)} y^{(l_1)} + \dots + \omega^{*T} x^{(l_k)} y^{(l_k)}}_{\geq k\gamma \ \text{(طبق فرض)}}$$

$$\left.\begin{array}{l} \xrightarrow{\text{توان ۲}} (\omega^{*T} \omega_k)^2 \geq (k\gamma)^2 \\ \underbrace{(\omega^{*T} \omega_k)^2}_{\text{نامساوی کوشی شوارتز}} \leq \underbrace{\|\omega^*\|^2}_{=1 \ \text{(طبق فرض)}} \|\omega_k\|^2 \end{array}\right\} \rightarrow \|\omega_k\|^2 \geq k^2 \gamma^2 \quad ②$$

$$\xrightarrow{①,②} k^2 \gamma^2 \leq \|\omega_k\|^2 \leq k r^2$$

$$\rightarrow \boxed{k \leq \frac{r^2}{\gamma^2}} \quad \blacksquare$$

۲-۱ (مدل‌های خطی) . پرسش دوم

①

$$X = \begin{bmatrix} x^{(1)T} \\ \vdots \\ x^{(n)T} \end{bmatrix} , \quad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

$$\rightarrow J(\omega) = \| y - X\omega \|^2$$

$$\rightarrow J(\omega) = (y - X\omega)^T (y - X\omega) = y^T y - \underbrace{y^T X\omega}_{= (X\omega)^T y = \omega^T (X^T y)} - \underbrace{\omega^T X^T y}_{= \omega^T (X^T y)} + \omega^T X^T X \omega$$

$$= \omega^T (X^T X) \omega - 2\omega^T (X^T y) + y^T y$$

$$\xrightarrow{\text{بهینه کردن نسبت به } \omega} \nabla_W J(\omega) = 0 \rightarrow \nabla_W \left(\omega^T (X^T X) \omega - 2\omega^T (X^T y) + y^T y \right) = 0$$

$$\xrightarrow{\text{برابر صفر}} 2X^T X\omega - 2X^T y = 0 \rightarrow X^T X \omega = X^T y$$

$$\xrightarrow{\text{با فرض وارون‌پذیری } X^T X} \boxed{\omega = (X^T X)^{-1} X^T y} \quad \blacksquare$$

② مشکل اصلی استفاده از رابطه قبل، زمان‌بر بودن محاسبه وارون ماتریس $X^T X$ است و هم‌چنین حتی ضرب ماتریس‌ها در صورت داشتن ابعاد زیاد، وقت‌گیر است. برای مقابله با این مسئله رویکرد اول این است که برای بدست آوردن ω هم‌چنان از این فرم بسته استفاده کنیم و از روش‌هایی استفاده کنیم که وارون ماتریس را در مرتبه زمانی پایین‌تری بدست می‌آورند.

رویکرد دوم و بهتر این است که از فرم بسته برای بدست آوردن ω استفاده نکنیم، بلکه $J(\omega)$ را با روش‌های تکرار شونده (iterative) بهینه کنیم مانند گرادیان کاهشی (Gradient Descent).

$X, y \longrightarrow$ تعریف مسئله قسمت ① ۳

$F = \operatorname{diag}(F_i) \rightarrow F^{1/2} = \operatorname{diag}(\sqrt{F_i})$

$\rightarrow J(\omega) = \| F^{1/2}(y - X\omega) \|_2^2$

مسئله بخش ① $\;\stackrel{\rightarrow}{=} y^T F y - 2\omega^T X^T F y + \omega^T X^T F X \omega$

$\rightarrow \nabla_\omega J(\omega) = -2X^T F y + 2X^T F X \omega$

$\nabla_\omega J(\omega) = 0 \rightarrow X^T F X \omega = X^T F y \rightarrow \omega = (X^T F X)^{-1} X^T F y$

$(y - \omega^T x)^2 = (y - \omega^T x)(y - x^T\omega) = y^2 - y \underbrace{(\omega^T x + x^T \omega)}_{= x^T\omega} + \omega^T (x x^T)\omega$ ۴

$\rightarrow E_{x,y}[(y-\omega^T x)^2] = \iint [y^2 - 2y(x^T\omega) + \omega^T(xx^T)\omega]\, p(x,y)\, dx\, dy$

$\rightarrow \nabla_\omega E_{x,y}[(y-\omega^T x)^2] = \iint (-2yx + 2xx^T\omega)\, p(x,y)\, dx\, dy$

$\nabla_\omega = 0 \rightarrow \underbrace{\iint y x\, p(x,y)\, dx\, dy}_{C} = \underbrace{\left(\iint x x^T p(x,y)\, dx\, dy\right)}_{R} \omega$

$\rightarrow R\omega = C \rightarrow \omega = R^{-1} C$

بهترین خط کلی : $\omega^* = \underset{\omega}{\operatorname{argmin}}\; E_{x,y}[(y - \omega^T x)^2]$

$\rightarrow$ structural error : $E_{x,y}[(y - \omega^{*T} x)^2]$

تخمین بهترین خط کلی از روی نمونه‌ها : $\hat{\omega} = \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^{n} (y^{(i)} - \omega^T x^{(i)})^2$

→ Approximation Error: $E_x\left[(w^{*T}x - \hat{w}^T x)^2\right]$

حال expected(test) loss را باز می‌کنیم و ثابت می‌کنیم که برابر با حاصل جمع Approximation, structural Error است:

expected(test) loss : $E_{x,y}\left[(y-\hat{w}^T x)^2\right]$

$$= E_{x,y}\left[(y - w^{*T}x + w^{*T}x - \hat{w}^T x)^2\right]$$

$$= E_{x,y}\left[(y-w^{*T}x)^2\right] + E_x\left[(w^{*T}x - \hat{w}^T x)^2\right] + \underbrace{2E_{x,y}\left[(y-w^{*T}x)(w^{*T}x - \hat{w}^T x)\right]}_{\text{ثابت می‌کنیم برابر با ۰ است}}$$

علت: $w^* = \underset{w}{\operatorname{argmin}}\ E_{x,y}\left[(w^T x - y)^2\right]$

مشتق نسبت به w در نقطه w^*، ۰ است

$$2E_{x,y}\left[x(w^{*T}x - y)\right] = 0$$

از طرفی: $E_{x,y}\left[\underbrace{(y-w^{*T}x)}_{\text{عدد}}\underbrace{(w^{*T}x - \hat{w}^T x)}_{(w^*-\hat{w})^T x}\right]$

$$= E_{x,y}\left[(w^*-\hat{w})^T (y - w^{*T}x)\, x\right]$$

$$= (w^*-\hat{w})^T \underbrace{E_{x,y}\left[(y-w^{*T}x)\,x\right]}_{0} = 0$$

$$\Rightarrow E_{x,y}\left[(y-\hat{w}^T x)^2\right] = \underbrace{E_{x,y}\left[(y-w^{*T}x)^2\right]}_{\text{structural error}} + \underbrace{E_x\left[(w^{*T}x - \hat{w}^T x)^2\right]}_{\text{approximation error}} \qquad \blacksquare$$

۱-۳: پرسش سوم (مدل های خطی)

۱) در سوال قبل بدست آوردیم:

$$\omega = (X^TX)^{-1}X^Ty$$

یک ویژگی ← X یک بردار $n\times 1$ است:

$$X^TX = \langle X, X\rangle \quad , \quad X^Ty = \langle X, y\rangle$$

$$X = x_i \rightarrow \boxed{\omega = \frac{\langle x_i, y\rangle}{\langle x_i, x_i\rangle}}$$

ویژگی i ام

۲) X متعامد است ← ستون‌ها برهم عمودند ← X^TX یک ماتریس قطری است که درایه قطر i ام آن $\langle x_i, x_i\rangle$ است

x_i: ستون i ام ماتریس X (ویژگی i ام)

همچنین: $X^Ty = \langle x_i, y\rangle$

$$\rightarrow \omega = \begin{bmatrix} \langle x_1, x_1\rangle & & 0 \\ & \ddots & \\ 0 & & \langle x_n, x_n\rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle x_1, y\rangle \\ \vdots \\ \langle x_n, y\rangle \end{bmatrix}$$

$$\rightarrow \omega = \begin{bmatrix} \vdots \\ \frac{\langle x_i, y\rangle}{\langle x_i, x_i\rangle} \\ \vdots \end{bmatrix}$$ ← با استفاده از قسمت قبل متغیر را حاصل می‌کند ∎

۳)

$$X = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(n)} \end{bmatrix} \rightarrow X^TX = \begin{bmatrix} n & \sum x^{(i)} \\ \sum x^{(i)} & \sum x^{(i)2} \end{bmatrix}$$

CamScanner

$$\rightarrow (X^TX)^{-1} = \frac{1}{n\sum x^{(i)2} - \left(\sum x^{(i)}\right)^2} \begin{bmatrix} \sum x^{(i)2} & -\sum x^{(i)} \\ -\sum x^{(i)} & n \end{bmatrix}$$

$$\rightarrow \omega = \frac{1}{n\sum x^{(i)2} - \left(\sum x^{(i)}\right)^2} \begin{bmatrix} \sum x^{(i)2} & -\sum x^{(i)} \\ -\sum x^{(i)} & n \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x^{(1)} & x^{(2)} & \cdots & x^{(n)} \end{bmatrix} \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

$$= \qquad \begin{bmatrix} \sum y^{(i)} \\ \sum x^{(i)} y^{(i)} \end{bmatrix}$$

$$\frac{\sum x^{(i)}}{n} = \bar{x} \qquad \frac{\sum y^{(i)}}{n} = \bar{y}$$

میانگین های تجربی

$$= \frac{1}{n\left(\sum (x^{(i)} - \bar{x})^2\right)} \begin{bmatrix} n\left(\sum x^{(i)2}\right)\bar{y} - n\bar{x}\sum x^{(i)}y^{(i)} \\ -n^2\bar{x}\bar{y} + n\sum x^{(i)}y^{(i)} \end{bmatrix}$$

$$= \frac{1}{n^2\, var(X)} \begin{bmatrix} n^2\left(\frac{\sum x^{(i)2}}{n}\bar{y} - \bar{x}\frac{\sum x^{(i)}y^{(i)}}{n}\right) \\ n^2\, cov(X,Y) \end{bmatrix}$$

واریانس تجربی

کوواریانس تجربی

$$= \frac{1}{var(X)} \begin{bmatrix} \frac{\sum x^{(i)2}}{n}\bar{y} - \bar{x}\frac{\sum x^{(i)}y^{(i)}}{n} \\ cov(X,Y) \end{bmatrix} \rightarrow \boxed{\omega_1 = \frac{cov(X,Y)}{var(X,Y)}}$$

$$\frac{\sum x^{(i)2}}{n}\bar{y} - \bar{x}\frac{\sum x^{(i)}y^{(i)}}{n} = \frac{\sum x^{(i)2}}{n}\bar{y} - \bar{x}\left(cov(X,Y) + \bar{x}\bar{y}\right)$$

$$= \left(\frac{\sum x^{(i)2}}{n} - \bar{x}^2\right)\bar{y} - \bar{x}\, cov(X,Y) = var(X)\bar{y} - \bar{x}\, cov(X,Y)$$

$$\rightarrow \omega_0 = \bar{y} - \bar{x}\frac{cov(X,Y)}{var(X,Y)} = \bar{y} - \bar{x}\omega_1 = \boxed{E[Y] - E[X]\omega_1} \quad \blacksquare$$

حل. $\hat{w} = \max(|X_۱|, |X_۲|, \dots, |X_n|)$

فرض کنید $\hat{w}$ یک تخمینگر MLE برای w باشد.

$\hat{w} = \arg\max_w p(X_۱, \dots, X_n|w)$.

از آنجایی که $X_۱, \dots, X_n$ همگی iid هستند:

$\hat{w} = \arg\max_w \prod_{i=۱}^{n} p(X_i|w)$.

$X_M = \max(|X_۱|, |X_۲|, \dots, |X_n|)$.

اگر $X_M > w$، آنگاه $\prod_{i=۱}^{n} p(X_i|w) = ۰$ برقرار است.

بنابراین، $w \geqslant X_M$.

$$p(X_i|w) = \frac{۱}{۲w}$$

$$\hat{w} = \arg\max_{w \geqslant X_M} \prod_{i=۱}^{n} p(X_i|w) = \arg\max_{w \geqslant X_M} \frac{۱}{(۲w)^n} = \arg\max_{w \geqslant X_M} \log \frac{۱}{(۲w)^n}$$

$$= \arg\max_{w \geqslant X_M} \log ۱ - n \log(۲w) = \arg\max_{w \geqslant X_M} -n \log(۲w) = \arg\min_{w \geqslant X_M} n \log(۲w)$$

$$= \arg\min_{w \geqslant X_M} \log(w) = \arg\min_{w \geqslant X_M} w = X_M$$

▷

1.1 Answer to Part A

First, we derive the likelihood term:

$$P(x_1,\ldots,x_N|\mu) = \prod_{i=1}^{N} P(x_i|\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Next, we note that log is a monotonically increase function, so we can maximize the log-likelihood:

$$\log\left(P(x_1,\ldots,x_N|\mu)\right) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}$$

We take derivatives of this with respect to μ and find:

$$\frac{d\log\left(P(x_1,\ldots,x_N|\mu)\right)}{d\mu} = \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2}$$

Setting the left hand side equal to zero, we find:

$$\begin{aligned}
0 &= \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} \\
0 &= \sum_{i=1}^{N} (x_i-\mu) \\
\sum_{i=1}^{N} \mu &= \sum_{i=1}^{N} x_i \\
N\mu &= \sum_{i=1}^{N} x_i \\
\hat{\mu} &= \frac{\sum_{i=1}^{N} x_i}{N}
\end{aligned}$$

1.2 Answer to Part B

There are two ways to solve this problem. We first show an easier way, which is sufficient to find the MAP estimator.

We use Bayes' rule to write:

$$P(\mu|x_1,\ldots x_n) = \frac{P(x_1,\ldots,x_N|\mu)P(\mu)}{P(x_1,\ldots,x_N)}$$

From part A, we know that we can write:

$$P(x_1,\ldots,x_N|\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

and we are given that:

$$P(\mu) = \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(\mu-\nu)^2}{2\beta^2}}$$

Thus, we desire to find the value of μ which maximizes:

$$P(\mu|x_1,\ldots x_n) = \frac{\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(\mu-\nu)^2}{2\beta^2}}}{C}$$

where $C = P(x_1,\ldots,x_N)$.

We note that because we are simply looking for the value of μ that maximizes this expression we can take the log of both side and write:

$$\log\left(P(\mu|x_1,\ldots x_n)\right) = \left(\sum_{i=1}^{N} -\log\left(\sqrt{2\pi\sigma^2}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right) - \log\left(\sqrt{2\pi\beta^2}\right) - \frac{(\mu-\nu)^2}{2\beta^2}$$

Taking the derivative with respect to μ, we have:

$$\frac{\partial\log\left(P(\mu|x_1,\ldots x_n)\right)}{\partial\mu} = \left(\sum_{i=1}^{N} \frac{x_i-\mu}{\sigma^2}\right) - \frac{\mu-\nu}{\beta^2}$$

Setting this equal to zero, we have:

$$\begin{aligned}
0 &= \left(\sum_{i=1}^{N} \frac{x_i-\mu}{\sigma^2}\right) - \frac{\mu-\nu}{\beta^2} \\
\frac{\mu-\nu}{\beta^2} &= \sum_{i=1}^{N} \frac{x_i-\mu}{\sigma^2} \\
\frac{\mu-\nu}{\beta^2} &= \frac{\sum_{i=1}^{N} x_i}{\sigma^2} - \frac{N\mu}{\sigma^2} \\
\frac{\mu}{\beta^2} + \frac{N\mu}{\sigma^2} &= \frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{\nu}{\beta^2} \\
\frac{(\sigma^2+N\beta^2)\mu}{\sigma^2\beta^2} &= \frac{\sigma^2\nu + \beta^2\sum_{i=1}^{N} x_i}{\sigma^2\beta^2} \\
\hat{\mu} &= \frac{\sigma^2\nu + \beta^2\sum_{i=1}^{N} x_i}{\sigma^2 + N\beta^2}
\end{aligned}$$

$$\hat{\mu} = \frac{\sigma^2\nu + \beta^2\sum_{i=1}^N x_i}{\sigma^2 + N\beta^2}$$

The second way involves a little more work. In this way, we first show that the posterior distribution is itself a Gaussian, and we then use the fact that the mean of a Gaussian is where it achieves its maximum value to find the MAP.

We start with the fact that we again want to find:

$$P(\mu|x_1,\dots x_n) = \frac{\left(\prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{\sigma^2}}\right)\frac{1}{\sqrt{2\pi\beta^2}}e^{-\frac{(\mu-\nu)^2}{2\beta^2}}}{C}$$

where $C = P(x_1,\dots,x_N)$. However, instead of simply maximizing this function, we first show that $P(\mu|x_1,\dots x_n)$ is itself a Gaussian. Note, that we can write:

$$\begin{aligned}
P(\mu|x_1,\dots x_n) &\propto \left(\prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)\frac{1}{\sqrt{2\pi\beta^2}}e^{-\frac{(\mu-\nu)^2}{2\beta^2}} \\
&\propto \left(\prod_{i=1}^N e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)e^{-\frac{(\mu-\nu)^2}{2\beta^2}} \\
&= e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^N (x_i-\mu)^2\right)}e^{-\frac{(\mu-\nu)^2}{2\beta^2}} \\
&= e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^N (x_i-\mu)^2\right)-\frac{(\mu-\nu)^2}{2\beta^2}} \\
&= e^{-\frac{\beta^2\left(\sum_{i=1}^N (x_i-\mu)^2\right)+\sigma^2(\mu-\nu)^2}{2\sigma^2\beta^2}}
\end{aligned}$$

At this point, we make use of the hint provided and write:

$$\begin{aligned}
P(\mu|x_1,\dots x_n) &\propto e^{-\frac{1}{2\sigma^2\beta^2}\left(\left[\mu\sqrt{N\beta^2+\sigma^2}-\frac{\sigma^2\nu+\beta^2\sum_{i=1}^N x_i}{\sqrt{N\beta^2+\sigma^2}}\right]^2-\frac{[\sigma^2\nu+\beta^2\sum_{i=1}^N x_i]^2}{N\beta^2+\sigma^2}+\beta^2\left(\sum_{i=1}^N x_i^2\right)+\sigma^2\nu^2\right)} \\
&= e^{-\frac{1}{2\sigma^2\beta^2}\left(\left[\mu\sqrt{N\beta^2+\sigma^2}-\frac{\sigma^2\nu+\beta^2\sum_{i=1}^N x_i}{\sqrt{N\beta^2+\sigma^2}}\right]^2\right)}e^{\frac{1}{2\sigma^2\beta^2}\left(\frac{[\sigma^2\nu+\beta^2\sum_{i=1}^N x_i]^2}{N\beta^2+\sigma^2}+\beta^2\left(\sum_{i=1}^N x_i^2\right)+\sigma^2\nu^2\right)} \\
&\propto e^{-\frac{1}{2\sigma^2\beta^2}\left(\left[\mu\sqrt{N\beta^2+\sigma^2}-\frac{\sigma^2\nu+\beta^2\sum_{i=1}^N x_i}{\sqrt{N\beta^2+\sigma^2}}\right]^2\right)} \\
&= e^{-\frac{N\beta^2+\sigma^2}{2\sigma^2\beta^2}\left(\left[\mu-\frac{\sigma^2\nu+\beta^2\sum_{i=1}^N x_i}{N\beta^2+\sigma^2}\right]^2\right)}
\end{aligned}$$

Note this is just the kernel of a Normal distribution with mean $\frac{\sigma^2\nu+\beta^2\sum_{i=1}^N x_i}{N\beta^2+\sigma^2}$ and variance $\frac{\sigma^2\beta^2}{N\beta^2+\sigma^2}$ (so the posterior is a normal distribution). Therefore, since a normal achieves at its maximum its mean, the MAP estimator must be:

$$\hat{\mu}_{\text{MAP}} = \frac{\sigma^2\nu + \beta^2\sum_{i=1}^N x_i}{N\beta^2 + \sigma^2}$$

1.3 Answer to Part C

The MAP estimator is:

$$\begin{aligned}
\hat{\mu}_{\text{MAP}} &= \frac{\sigma^2\nu + \beta^2\sum_{i=1}^N x_i}{N\beta^2 + \sigma^2} \\
&= \frac{\sigma^2\nu}{N\beta^2+\sigma^2} + \frac{\beta^2\sum_{i=1}^N x_i}{N\beta^2+\sigma^2} \\
&= \frac{\sigma^2\nu}{N\beta^2+\sigma^2} + \frac{\sum_{i=1}^N x_i}{N+\frac{\sigma^2}{\beta^2}} \\
&= \frac{\sigma^2\nu}{N\beta^2+\sigma^2} + \frac{\frac{1}{N}\sum_{i=1}^N x_i}{1+\frac{\sigma^2}{N\beta^2}}
\end{aligned}$$

and the MLE estimator is:

$$\hat{\mu}_{\text{MLE}} = \frac{\sum_{i=1}^N x_i}{N}$$

Notice that as $N \to \infty$, $\frac{\sigma^2}{N\beta^2} \to 0$, $\frac{\sigma^2\nu}{N\beta^2+\sigma^2} \to 0$ and therefore $\hat{\mu}_{\text{MAP}} \to \frac{\sum_{i=1}^N x_i}{N} = \hat{\mu}_{\text{MLE}}$. Thus, as the number of samples increases, the two estimators become identical.

۳.۲: پرسش سوم (MLE & MAP):

$$\beta \sim \Gamma(\alpha_0, \beta_0)$$

$$\forall i \; x_i \sim \Gamma(\alpha, \beta)$$

$$p(\beta \mid x_1, \dots, x_n) \propto p(x_1, \dots, x_n \mid \beta)\, p(\beta) = p(x_1 \mid \beta) \dots p(x_n \mid \beta)\, p(\beta)$$

از طرفی: $p(x_i \mid \beta) = \dfrac{\beta^{\alpha} x_i^{\alpha-1} e^{-\beta x_i}}{\Gamma(\alpha)}$ ، $p(\beta) = \dfrac{\beta_0^{\alpha_0} \beta^{\alpha_0 - 1} e^{-\beta_0 \beta}}{\Gamma(\alpha_0)}$

$$\Rightarrow p(\beta \mid x_1, \dots, x_n) \propto \beta^{n\alpha + \alpha_0 - 1} e^{-\beta(\beta_0 + \sum_{i=1}^{n} x_i)}$$

بنابراین توزیع پسین β نیز از خانواده گاما بوده و پارامترهای جدید از روابط زیر بدست می‌آیند:

$$\alpha_1 = \alpha_0 + n\alpha \quad \text{و} \quad \beta_1 = \beta_0 + \sum_{i=1}^{n} x_i$$