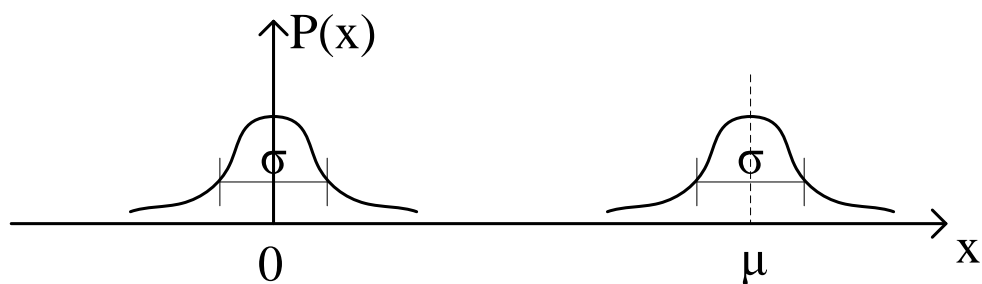


$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 P(x) dx = \int_{-\infty}^{\infty} x^2 P(x) dx$$



無論 μ 的值為何，均不會影響 σ^2 的值

微		積
x	+	$x e^{-kx^2}$
1	-	$\frac{-1}{2k} e^{-kx^2}$

$$\int_{-\infty}^{\infty} x^2 e^{-kx^2} dx = \frac{-1}{k} e^{-kx^2} \Big|_{-\infty}^{\infty} + \frac{1}{2k} \int_{-\infty}^{\infty} e^{-kx^2} dx = (0-0) + \frac{1}{2k} \sqrt{\frac{\pi}{k}}$$

$$\Rightarrow \sqrt{\frac{k}{\pi}} \int_{-\infty}^{\infty} x^2 e^{-kx^2} dx = \sqrt{\frac{k}{\pi}} \frac{1}{2k} \sqrt{\frac{\pi}{k}} = \frac{1}{2k} = \sigma^2$$

$$\Rightarrow k = \frac{1}{2\sigma^2}$$

故我們就可以將 Gaussian distribution 寫為

$$P(x | \mu, \sigma^2) = \sqrt{\frac{1}{2\sigma^2 \pi}} e^{-\frac{x^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

而由於我們希望 Gaussian distribution 是沿著 mean 對稱的，故我們將上式改寫為

$$P(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

故給定 μ, σ^2 ，我們可得到唯一，離 mean 距離相同就有相同機率的分佈

$x \sim N(\mu, \sigma^2)$ //univariate Gaussian

note: Gaussian distribution 是一個很特別的函數，是一種 local 函數，若離 mean 太遠，其機率會小到幾乎可以忽略，很少函數有這種性質。

MLE on Gaussian(期中考大熱門)

若有一組 Data $D = x_1, x_2, \dots, x_n$

$$L(\theta = \mu, \sigma^2 | D) = P(D | \theta) = \prod_{i=1}^n P(x_i | \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

我們要求最大值，需要使用微分，和前面一樣的方法，我們取 \log ， \log 得到最大的 θ 同時也是沒取 \log 最大的 θ

$$\arg \max_{\theta} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = \arg \max_{\theta} \log \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \right) = \arg \max_{\theta} \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} &= \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi}\sigma} + \sum_{i=1}^n \log e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = n \log(2\pi\sigma^2)^{-\frac{1}{2}} - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} \end{aligned}$$

有兩個參數： μ, σ^2 ，因為彼此為獨立的參數(需要證明，上網找有，但太複雜就沒看)，故只需要個別求最大的 \log likelihood 的參數，就是整體最大 \log likelihood 的參數

μ_{MLE}

$$\begin{aligned} \frac{d}{d\mu} \left(-\frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} \right) &= -\frac{d}{d\mu} \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \\ \Rightarrow \sum_{i=1}^n x_i &= n\mu \Rightarrow \mu = \frac{\sum_{i=1}^n x_i}{n} \end{aligned}$$

和我們對 **mean** 的認知一樣，故若我們有一堆資料，假設其呈現高斯分佈，其最大的 **mean** 即為所有 **data** 相加後除以資料個數

$$\sigma_{MLE}$$

為方便起見，令 $\sigma^2 = s$

$$\frac{d}{ds} \left(\frac{-n}{2} \log(2\pi s) - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2s} \right) = \frac{d}{ds} \left(\frac{-n}{2} \log 2\pi + \frac{-n}{2} \log s - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2s} \right) = \frac{-n}{2s} + \sum_{i=1}^n \frac{(x_i - \mu)^2}{2s^2} = 0$$

$$\Rightarrow \frac{1}{2s^2} \sum_{i=1}^n (x_i - \mu)^2 = \frac{n}{2s} \Rightarrow s = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} = \sigma^2$$

也和我們之前認知的 variance 相同

Conjugate prior of Gaussian

Gaussian distribution 也具有之前講過的 Conjugate 性質，即若給 Gaussian distribution 形式的 prior，其 posterior 也是 Gaussian distribution

若有一組 Data $D = x_1, x_2, \dots, x_n$

給定 prior $N(\mu_0, \sigma_0^2)$

圖片來源：維基百科

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters
Normal with known variance σ^2	μ (mean)	Normal	μ_0, σ_0^2	$\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2} \right) / \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right),$ $\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1}$
Normal with known precision τ	μ (mean)	Normal	μ_0, τ_0	$\left(\tau_0 \mu_0 + \tau \sum_{i=1}^n x_i \right) / (\tau_0 + n\tau), \tau_0 + n\tau$
Normal with known mean μ	σ^2 (variance)	Inverse gamma	α, β ^[note 5]	$\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}$
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	ν, σ_0^2	$\nu + n, \frac{\nu \sigma_0^2 + \sum_{i=1}^n (x_i - \mu)^2}{\nu + n}$
Normal with known mean μ	τ (precision)	Gamma	α, β ^[note 3]	$\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}$
Normal ^[note 6]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0, \nu, \alpha, \beta$	$\frac{\nu \mu_0 + n \bar{x}}{\nu + n}, \nu + n, \alpha + \frac{n}{2},$ $\beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n \nu}{\nu + n} \frac{(\bar{x} - \mu_0)^2}{2}$ • \bar{x} is the sample mean

這張圖可知道若我們已知 data 中的 variance，我們希望找出最符合 sample space 的 mean，給定 prior 為 Gaussian distribution，可以得到 posterior 形式為 Gaussian distribution

note:但通常，我們沒辦法知道實際背景的 mean 及 variance，所以通常我們得到的 posterior 只能式 Normal-Inverse gamma function，但是這太難推導，故我們只推圖片中第一列的 posterior

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

上課題外話：

通常我們使用 bayesian 的原因是為了避免 overfitting，在 data 不足時，很容易會產生 overfitting 現象，像是若我們擲兩次銅板，若我們使用 MLE，我們就會得到 100%正面的機率，但若是使用 bayesian，若 prior 選的好，我們可以得到正面機率較高但不是 100%的機率

$$\begin{aligned} P(D | \mu)P(\mu) &= \prod_{i=1}^n P(x_i | \mu, \sigma) \cdot N(\mu | \mu_0, \sigma_0^2) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right) e^{\frac{-1}{2\sigma^2}(x_1-\mu)^2} \left(\frac{1}{\sqrt{2\pi}\sigma} \right) e^{\frac{-1}{2\sigma^2}(x_2-\mu)^2} \dots \left(\frac{1}{\sqrt{2\pi}\sigma} \right) e^{\frac{-1}{2\sigma^2}(x_n-\mu)^2} \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_0} \right) e^{\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{\sum_{i=1}^n \frac{-1}{2\sigma^2}(x_i-\mu)^2} \left(\frac{1}{\sqrt{2\pi}\sigma_0} \right) e^{\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}} = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \left(\frac{1}{\sqrt{2\pi}\sigma_0} \right) e^{\sum_{i=1}^n \frac{-1}{2\sigma^2}(x_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}} \end{aligned}$$

note: 小提醒一下， μ_0, σ_0^2 是 prior， σ 是已知， μ 是我們想要得到最符合 sample space 的 mean

我們先不看常數項，現在我們要做的是這區塊

$$\left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \left(\frac{1}{\sqrt{2\pi}\sigma_0} \right) e^{\sum_{i=1}^n \frac{-1}{2\sigma^2}(x_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$
$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

因為 Gaussian distribution 的形式是 $Ae^{\frac{-(x-\mu)^2}{B}} = e^{\frac{-(x-\mu)^2}{B} + C}$ ，我們要先將指數項做成平方項

$$\begin{aligned}
& \sum_{i=1}^n \frac{-1}{2\sigma^2} (x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} = \frac{-1}{2\sigma^2} \sum_{i=1}^n (x_i^2 + \mu^2 - 2x_i\mu) - \frac{1}{2\sigma_0^2} (\mu^2 + \mu_0^2 - 2\mu\mu_0) \\
&= \frac{-1}{2} \mu^2 \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) + \mu \left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) - \frac{1}{2} \left(\frac{\sum_{i=1}^n x_i^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2} \right) \\
&= \frac{-1}{2} \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \left(\mu^2 - \frac{2 \left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)} \mu + \frac{\left(\frac{\sum_{i=1}^n x_i^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2} \right)}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)} \right) \\
&= \frac{-1}{2} \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \left(\mu^2 - \frac{2 \left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)} \mu + \frac{\left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)^2}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)} + \frac{\left(\frac{\sum_{i=1}^n x_i^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2} \right)}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)} - \frac{\left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)^2}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)} \right)
\end{aligned}$$

$$\text{令 } \mu_n = \frac{\left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)}$$

$$\Rightarrow \sum_{i=1}^n \frac{-1}{2\sigma^2} (x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} = \frac{-1}{2} \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) (\mu - \mu_n)^2 + \frac{\left(\frac{\sum_{i=1}^n x_i^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2} \right)}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)} - \mu_n^2$$

$$= \frac{-1}{2} \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) (\mu - \mu_n)^2 + D$$

$$\Rightarrow e^{\frac{-1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)(\mu - \mu_n)^2 + D} = Ae^{\frac{-1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)(\mu - \mu_n)^2}, A = e^D$$

得到 $P(D|\theta)P(\theta)$ 後，marginal 就簡單了，也就是將所有可能的參數 θ 值所得的機率全部加總

$$\text{marginal: } P(D) = \int_{-\infty}^{\infty} P(D|\theta')P(\theta') d\theta' = \int_{-\infty}^{\infty} P(D|\mu')P(\mu') d\mu'$$

$$= \int_{-\infty}^{\infty} Ae^{\frac{-1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)(\mu' - \mu_n)^2} d\mu' = A \int_{-\infty}^{\infty} e^{\frac{-1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu'^2} d\mu' = A \int_{-\infty}^{\infty} e^{-k\mu'^2} d\mu', \quad \text{let } k = \frac{-1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)$$

$$= A\sqrt{\frac{\pi}{k}} = A\sqrt{\frac{\pi}{\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}}$$

note:

$\int_{-\infty}^{\infty} Ae^{\frac{-1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)(\mu' - \mu_n)^2} d\mu' = A \int_{-\infty}^{\infty} e^{\frac{-1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu'^2} d\mu'$ ，這一步可將 $(\mu' - \mu_n)^2$ 轉為 μ'^2 的理由和這份筆記一開始求 σ^2 時的地方很像，無論 mean 在哪裡，marginal 的值都不會改變

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{Ae^{\frac{-1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)(\mu - \mu_n)^2}}{A\sqrt{\frac{\pi}{\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}}}, \text{let } \sigma_n^2 = \frac{1}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}$$

$$= \frac{1}{\sqrt{2\pi\sigma_n}} e^{\frac{-(\mu - \mu_n)^2}{2\sigma_n^2}} = N(\mu_n, \sigma_n)$$

Prior

Posterior

$$N(\mu, \sigma^2) \longrightarrow N(\mu_n = \sigma_n^2 \left(\frac{n}{\sigma^2} \mu_{MLE} + \frac{1}{\sigma_0^2} \mu_0 \right), \sigma_n^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}})$$

而我們對 μ_n 做些觀察

$$\mu_n = \sigma_n^2 \left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) = \sigma_n^2 \left(\frac{1}{\sigma^2} n \mu_{MSE} + \frac{1}{\sigma_0^2} \mu_0 \right) = \sigma_n^2 \frac{n}{\sigma^2} \mu_{MSE} + \frac{\sigma_n^2}{\sigma_0^2} \mu_0$$

$$\text{由於 } \sigma_n^2 \frac{n}{\sigma^2} + \frac{\sigma_n^2}{\sigma_0^2} = \sigma_n^2 \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) = 1$$

則 μ_n 必介於 μ_{MSE} 和 μ_0 之間

解釋：

若 $x_3 = ax_1 + bx_2, a, b > 0$ 且 $a + b = 1$

則 $x_1 \leq x_3 \leq x_2$ or $x_2 < x_3 < x_1$

證明：

其實實在是懶的證，還是證一下

if $x_1 \leq x_2$

$$\Rightarrow x_3 = ax_1 + bx_2 \leq ax_2 + bx_2 = (a+b)x_2 = x_2$$

$$x_3 = ax_1 + bx_2 \geq ax_1 + bx_1 = (a+b)x_1 = x_1$$

$$\Rightarrow x_1 \leq x_3 \leq x_2$$

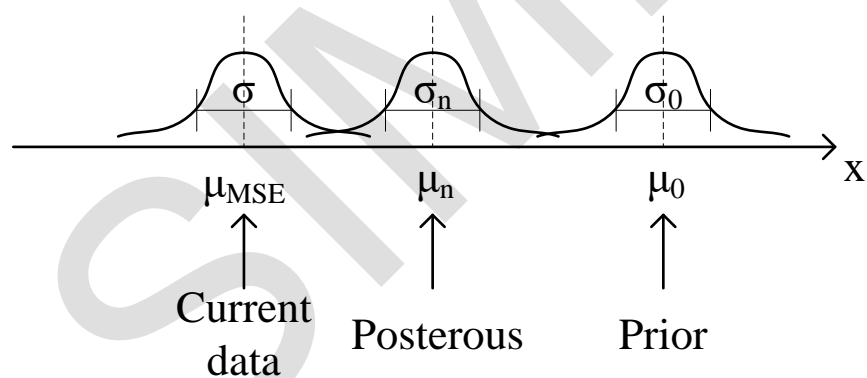
else // ($x_1 > x_2$)

$$x_3 = ax_1 + bx_2 > ax_2 + bx_2 = (a+b)x_2 = x_2$$

$$x_3 = ax_1 + bx_2 < ax_1 + bx_1 = (a+b)x_1 = x_1$$

$$\Rightarrow x_2 < x_3 < x_1$$

圖示：



當 $n \rightarrow 0$

$$\Rightarrow \sigma_n^2 = \frac{1}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)} = \sigma_0^2$$

$$\mu_n = \sigma_n^2 \left(\frac{n}{\sigma^2} \mu_{MLE} + \frac{1}{\sigma_0^2} \mu_0 \right) = \sigma_0^2 \left(\frac{n}{\sigma^2} \mu_{MLE} + \frac{1}{\sigma_0^2} \mu_0 \right) = \mu_0$$

$$P(\theta | D) = \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(\mu - \mu_n)^2}{2\sigma_n^2}} = \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{-(\mu - \mu_0)^2}{2\sigma_0^2}} = \text{prior}$$

當沒有新的資料量時，其分佈和原本 **prior** 的分佈是相同的(廢話)

當 $n \rightarrow \infty$

$$\Rightarrow \sigma_n^2 = \frac{1}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)} = 0$$

$$\mu_n = \sigma_n^2 \left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) = \frac{1}{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)} \left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) = \frac{1}{\frac{n}{\sigma^2}} \left(\frac{\sum_{i=1}^n x_i}{\sigma^2} \right) = \frac{\sum_{i=1}^n x_i}{n}$$

當資料量極大，甚至和 **sample space** 一樣大時，我們就不會有模稜兩可的灰色地帶，資料的 **mean** 就是 **sample space** 的 **mean**，也不會有 **variance**，因為已經沒有“機率”可言了，我們已經能夠百分之百準確預估現象的發生(就像統計全世界男女比例，如果我們只取樣一個小區塊 e.g. 美國，我們只能大概的推測出全世界的男女比，但是若我們手中有全世界人口的資料，我們就能百分之百肯定男女比為多少)