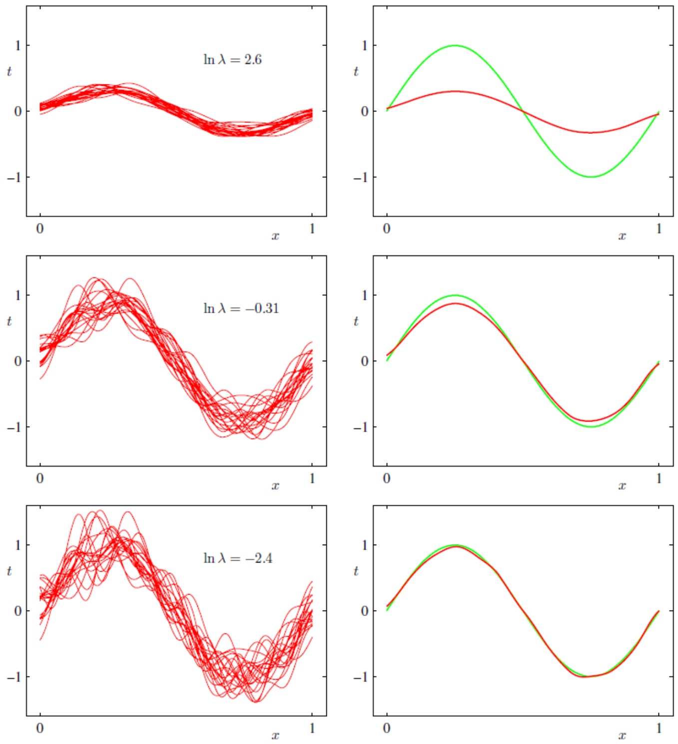
上一堂課的延伸：

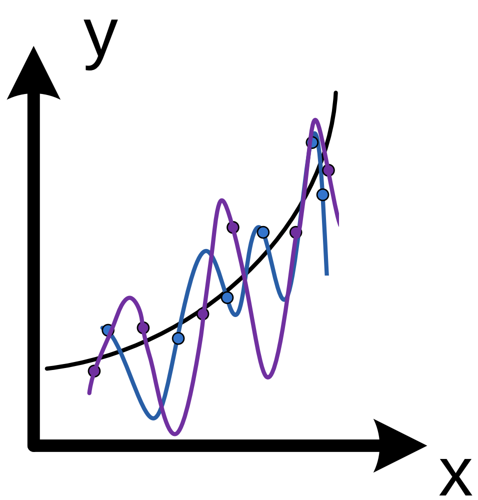
當λ值越大，regression 後會較穩定，但是 bias(偏移量)會較大，用之前學過的 rLSE 來想，λ越大，w(係數)就會較小，就不會因為一兩筆資料而讓 regression 後的結果有大幅改變，但是這樣會讓 regression 後的結果離最佳的(error 最小)的 w 較遠，換句話說，若λ趨近於 0，就幾乎沒有誤差存在。

課本 p150
左欄是相同λ，由右欄的綠線加上 error，隨機產生多筆資料，而產生的 regression 結果



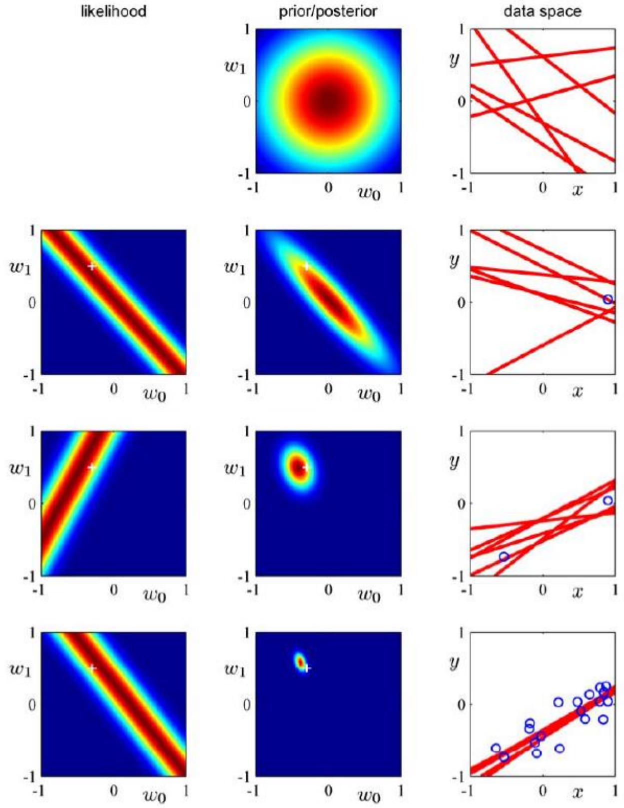

當λ越大，可看到多條紅線幾乎是重疊，不會有不穩定情形發生，但是隨之而來的就是誤差較大，可看到右圖，紅線和綠線差非常多。

當λ越小，不同的資料就會產生不同的結果，不會有不穩定情形發生，但是隨之而來的就是誤差較小，可看到右圖，紅線和綠線差不多。




舉個例子，就像是右圖，黑線是原始的 model，我們隨機產生兩筆有 error(藍點和紫點)的數據，當λ很小時，會產生幾乎沒有 error 的 regression 結果，但是不同筆 data 產生的 regression 結果會相差較大，也就是比較不穩定，當然，當取樣點越來越多時，regression 會越接近原始 model，但是那也要在我們取樣點夠多時才不會發生(和 overfitting 概念一樣)，詳細可見 lession1

課本 p155




上圖是使用上堂課的 MAP 做 online learning 的示意圖，第一欄是 likelihood $P(D\,|\,\mathbf{w})$ ，第二欄是我們給定的 prior $P(\mathbf{w})$ ，第三欄是由 posterior $P(\mathbf{w}\,|\,D)=\dfrac{P(D\,|\,\mathbf{w})P(\mathbf{w})}{P(D)}$ ，產生出多個不同的 **w** 而生成的多項式，在一開始時，我們假定的 prior 是沿著圓點放射出去的同心圓，也就是各個參數 $w_i$ 的 variance 皆相同，此時我們幾乎沒有得到甚麼多餘的資訊。

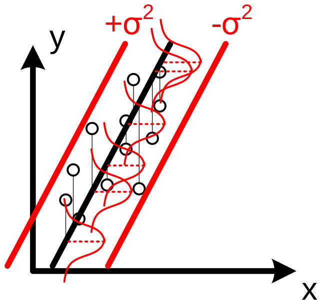
舉個例來說我們在擲一個不知道公不公平的銅板，我們一開始假設正面機率是 0.5，當然，根據我們這個假設如果投擲 10 次，會得到各式各樣不同的結果，和這裡是一樣的概念。

第二列就是我們已經收到第一筆資料(最右邊圖上的藍圓圈)，根據這一筆資料我們可以根據上一次的 prior 可算得 likelihood，而由於 prior 是 multivariate Gaussian，likelihood 是 univariate Gaussian，我們不需要另外算 marginal，因為我們知道結果必為 multivariate Gaussian，就像是 conjugate 的 beta distribution 一樣，我們只需要更新 multivariate Gaussian 的 mean 和 variance 就好，如此我們可以得到第二欄的 posterior，以此當作第三欄的 prior。而因為我們資訊變多了，我們的 mean 和 variance 都會作些許移動，如果資訊沒有偏移太多，variance 理應來說會越來越小，由圖上也可清楚看到，第二列的 posterior 的光暈比第一列還小，故若從這個 posterior 取出多組 **w** 出來，其結果就會呈現如最右邊圖那樣，其直線都會較第一列直線接近。

第三列和第二列做一樣的事，重複得到 data 後可得到第四列的結果，可看到已經得到許多筆資料了(有很多藍圓圈)，因為都和之前分析的 mean 很接近，故可看到其 posterior 的 variance 非常小(幾乎沒有光暈)。

# Predictive distribution

我們固然可以用 MAP 一次一次的作 online learning，找出最有可能的 **w**，再用算出來的 posterior 求出一組 **w**，得到 y，但其實我們根本就不關心 **w** 會怎麼樣，我們有興趣的只有 y，我這前面所做的都只有找出 **w** 的機率，求 y 時 **w** 都已經給定，也就是 $P(y\,|\,\mathbf{w},\theta)$ ，我們其實最有興趣的應該是如同求出 **w** 分布一樣，應該是想要 y 的分布，也就是我們希望除了找出最有可能的直線外，我們也想知道”結果是其他條直線”的結果為何，和 w 無關，也就是 $P(y\,|\,\theta)$ ，也就是想得到這張圖




$$P(Y\,|\,\theta)=\int \mathrm{P}(\mathrm{Y}\,|\,\mathbf{w},\theta)P(\mathbf{w},\theta)d\mathbf{w}$$

其實就是 marginalize 接下來我們就要開始之前沒有推導完的，若 prior 是 multivariate Gaussian，posterior 是 univariate Gaussian，marginalize 後也還會是 multivariate Gaussian。

其中

$$\mathrm{P}(Y\,|\,\mathbf{w},\theta)\sim N(Y\,|\,X\mathbf{w},a^{-1})$$

$$P(\mathbf{w},\theta)\sim N(\mathbf{w}\,|\,\boldsymbol{\mu},\Lambda^{-1})$$

$$\begin{aligned}
&\int N(Y\,|\,X\mathbf{w},a^{-1})N(\mathbf{w}\,|\,\mu,\Lambda^{-1})d\mathbf{w}\\
&=\int e^{\frac{-a}{2}(X\mathbf{w}-Y)^2}e^{\frac{-1}{2}(\mathbf{w}-\mu)^T\Lambda(\mathbf{w}-\mu)}d\mathbf{w}\\
&=\int e^{\frac{-1}{2}a(X\mathbf{w}-Y)^T(X\mathbf{w}-Y)+\frac{-1}{2}(\mathbf{w}-\mu)^T\Lambda(\mathbf{w}-\mu)}d\mathbf{w}\\
&=\int e^{\frac{-1}{2}(a(\mathbf{w}^TX^TX\mathbf{w}-2\mathbf{w}^TX^TY+Y^TY)+(\mathbf{w}^T\Lambda\mathbf{w}-2\mathbf{w}^T\Lambda\mu+\mu^T\mu))}d\mathbf{w}\\
&=\int e^{\frac{-1}{2}(\mathbf{w}^T(aX^TX+\Lambda)\mathbf{w}-2\mathbf{w}^T(aX^TY+\Lambda\mu)+aY^TY+\mu^T\mu)}d\mathbf{w}
\end{aligned}$$

一樣的，我們拿指數項出來看，試圖將其轉為 quadratic form

$$\begin{aligned}
&(\mathbf{w}-\mathbf{m})^TC(\mathbf{w}-\mathbf{m})\\
&=\mathbf{w}^TC\mathbf{w}-2\mathbf{w}^TC\mathbf{m}+\mathbf{m}^TC\mathbf{m}
\end{aligned}$$

我們使用類似配方法的方式，試圖將對應項對齊

可得

$$C = aX^T X + \Lambda$$
$$C\mathbf{m} = aX^T Y + \Lambda\boldsymbol{\mu} \Rightarrow m = C^{-1}(aX^T Y + \Lambda\boldsymbol{\mu})$$

故指數項可整理成

$$\frac{-1}{2}(\mathbf{w}^T C\mathbf{w} - 2\mathbf{w}^T C\mathbf{m} + \mathbf{m}^T C\mathbf{m} - \mathbf{m}^T C\mathbf{m} + aY^T Y + \boldsymbol{\mu}^T\boldsymbol{\mu})$$

marginalize 的積分式子會變為

$$\int e^{\frac{-1}{2}(\mathbf{w}^T C\mathbf{w} - 2\mathbf{w}^T C\mathbf{m} + \mathbf{m}^T C\mathbf{m} - \mathbf{m}^T C\mathbf{m} + aY^T Y + \boldsymbol{\mu}^T\boldsymbol{\mu})} d\mathbf{w}$$
$$= \int e^{\frac{-1}{2}(\mathbf{w}-\mathbf{m})^T C(\mathbf{w}-\mathbf{m}) + \frac{-1}{2}(-\mathbf{m}^T C\mathbf{m} + aY^T Y + \boldsymbol{\mu}^T\boldsymbol{\mu})} d\mathbf{w} = \int e^{\frac{-1}{2}(\mathbf{w}-\mathbf{m})^T C(\mathbf{w}-\mathbf{m})} e^{\frac{1}{2}(\mathbf{m}^T C\mathbf{m} - aY^T Y - \boldsymbol{\mu}^T\boldsymbol{\mu})} d\mathbf{w} = e^{\frac{-1}{2}(-\mathbf{m}^T C\mathbf{m} + aY^T Y + \boldsymbol{\mu}^T\boldsymbol{\mu})} \int e^{\frac{-1}{2}(\mathbf{w}-\mathbf{m})^T C(\mathbf{w}-\mathbf{m})} d\mathbf{w}$$

由於積分式內是一個 multivariate Gaussian，故積分後值為 1，上式改寫為

$$e^{\frac{-1}{2}(-\mathbf{m}^T \mathbf{m} + aY^T Y + \boldsymbol{\mu}^T\boldsymbol{\mu})}$$

故

$$\int N(y \mid X\mathbf{w}, a^{-1}) N(\mathbf{w} \mid \mu, \Lambda^{-1}) d\mathbf{w} = e^{\frac{-1}{2}(-\mathbf{m}^T C\mathbf{m} + aY^T Y + \boldsymbol{\mu}^T\boldsymbol{\mu})}$$

我們再針對指數項做改寫，希望能寫成 quadratic form，但再修改前，我們先要處理 $\mathbf{m}^T\mathbf{m}$ 這一項

$$\mathbf{m}^T C\mathbf{m}$$
$$= (C^{-1}(aX^T Y + \Lambda\boldsymbol{\mu}))^T C C^{-1}(aX^T Y + \Lambda\boldsymbol{\mu}) = (aX^T Y + \Lambda\boldsymbol{\mu})^T (C^{-1})^T (aX^T Y + \Lambda\boldsymbol{\mu}) = (aY^T X + \Lambda^T\boldsymbol{\mu}^T) C^{-1}(aX^T Y + \Lambda\boldsymbol{\mu})$$
$$= a^2 Y^T X C^{-1} X^T Y + 2aY^T X C^{-1}\Lambda\boldsymbol{\mu} + \boldsymbol{\mu}^T \Lambda^T C^{-1}\Lambda\boldsymbol{\mu}$$

$$-\mathbf{m}^T C\mathbf{m} + aY^T \mathrm{Y} + \boldsymbol{\mu}^T\boldsymbol{\mu} = -a^2 Y^T X C^{-1} X^T Y - 2aY^T X C^{-1}\Lambda\boldsymbol{\mu} - \boldsymbol{\mu}^T \Lambda^T C^{-1}\Lambda\boldsymbol{\mu} + aY^T Y + \boldsymbol{\mu}^T\boldsymbol{\mu}$$
$$= Y^T (a - a^2 X C^{-1} X^T) Y - 2aY^T X C^{-1}\Lambda\boldsymbol{\mu} + \ldots$$

我們不繼續做常數項式由於剛剛的經驗讓我們知道，後面常數的修正項到最後只會成為 exponential 前面的係數，最後做 normalize 後會再修正，前面的係數是甚麼並不重要。

為了之後推導方便起見，令 $\lambda = a - a^2 X C^{-1} X^T$

和前面推導的方式一樣 $Y^T \lambda Y - 2a\boldsymbol{\mu}^T \Lambda^T C^{-1} X^T Y + \ldots = (\mathrm{Y} - m')^T C'(Y - m) = Y^T C' Y - 2m' C' Y + \ldots$

故

$$C' = \lambda$$
$$C' m' = aXC^{-1}\Lambda\boldsymbol{\mu} \Rightarrow m' = aC'^{-1} XC^{-1}\Lambda\boldsymbol{\mu} = a\lambda^{-1} XC^{-1}\Lambda\boldsymbol{\mu} = a\lambda^{-1} X(aX^T X + \Lambda)^{-1}\Lambda\boldsymbol{\mu}$$

而根據 Sherman –Morrison formula

$C = aX^T X + \Lambda$

則$C^{-1} = \Lambda^{-1} - \dfrac{\Lambda^{-1} a X^T X \Lambda^{-1}}{1 + aX\Lambda^{-1}X^T}$

$$CC^{-1} = (aX^T X + \Lambda)(\Lambda^{-1} - \frac{\Lambda^{-1} a X^T X \Lambda^{-1}}{1 + aX\Lambda^{-1}X^T})$$
$$= aX^T X\Lambda^{-1} - \frac{aX^T X\Lambda^{-1} aX^T X\Lambda^{-1}}{I + aX\Lambda^{-1}X^T} + I - \frac{\Lambda\Lambda^{-1} aX^T X\Lambda^{-1}}{I + aX\Lambda^{-1}X^T} = aX^T X\Lambda^{-1} - \frac{aX^T(I + X\Lambda^{-1}aX^T)X\Lambda^{-1}}{I + aX\Lambda^{-1}X^T} + I$$
$$= aX^T X\Lambda^{-1} - aX^T X\Lambda^{-1} + I = I$$

為方便起見，

令$\alpha = X\Lambda^{-1}X^T$

$$C^{-1} = \Lambda^{-1} - \frac{\Lambda^{-1} aX^T (X\Lambda^{-1}X^T (X^T)^{-1})}{1 + a\alpha} = \Lambda^{-1} - \frac{\Lambda^{-1} aX^T \alpha (X^T)^{-1}}{1 + a\alpha}$$

代入$\lambda$得

$$\lambda = a - a^2 XC^{-1}X^T = a - a^2 X(\Lambda^{-1} - \frac{\Lambda^{-1} aX^T \alpha (X^T)^{-1}}{1 + a\alpha})X^T$$
$$= a - a^2 (X\Lambda^{-1}X^T - X\frac{\Lambda^{-1} aX^T \alpha (X^T)^{-1}}{1 + a\alpha}X^T)$$
$$= a - a^2(\alpha - X\frac{\Lambda^{-1} aX^T \alpha}{1 + a\alpha}) = a - a^2(\alpha - \frac{a\alpha^2}{1 + a\alpha}) = a - a^2 \frac{\alpha}{1 + a\alpha}$$
$$= \frac{a + a^2\alpha - a^2\alpha}{1 + a\alpha} = \frac{a}{1 + a\alpha}$$

則$\lambda^{-1} = \dfrac{1 + a\alpha}{a} = \dfrac{1}{a} + \alpha = \dfrac{1}{a} + X\Lambda^{-1}X^T$

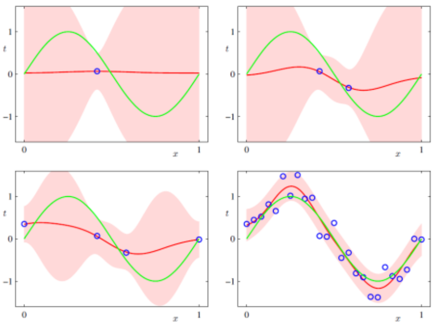
則

$$m' = a\lambda^{-1}XC^{-1}\Lambda\boldsymbol{\mu} = (a\lambda^{-1}XC^{-1}\Lambda\boldsymbol{\mu})^{\mathrm{T}} = a\boldsymbol{\mu}^{\mathrm{T}}\lambda^{-1}\Lambda C^{-1}X^T = a\boldsymbol{\mu}^{\mathrm{T}}\lambda^{-1}\Lambda(\Lambda^{-1} - \frac{\Lambda^{-1}aX^T\alpha(X^T)^{-1}}{1 + a\alpha})X^T = a\boldsymbol{\mu}^{\mathrm{T}}\lambda^{-1}(X^T - \frac{aX^T\alpha}{1 + a\alpha}) = a\boldsymbol{\mu}^{\mathrm{T}}\lambda^{-1}X^T\frac{1}{1 + a\alpha}$$
$$= a\boldsymbol{\mu}^{\mathrm{T}}(\frac{1}{a} + \alpha)X^T\frac{1}{1 + a\alpha} = a\boldsymbol{\mu}^{\mathrm{T}}\frac{1 + a\alpha}{a}X^T\frac{1}{1 + a\alpha} = \boldsymbol{\mu}^{\mathrm{T}}X^T$$

最終，做 marginalize 後的分布為

$$N(\boldsymbol{\mu}^{\mathrm{T}}X^T, \frac{1}{a} + X\Lambda^{-1}X^T) = N(X\boldsymbol{\mu}, \frac{1}{a} + X\Lambda^{-1}X^T)$$

課本 p157



原始是綠線，若只做 MAP，只會得到紅線，在點很少時，我們的信心較不足(co-variance 較大)，可看到紅色區間的範圍較大，當點變多的時候，信心較夠，紅色區間就會縮小