Bayesian model comparison

Model selection from a Bayesian perspective: use probability theory to represent uncertainties in the model choice.

Namely, recall that

$$\overbrace{p(\vec{w} \mid \underset{\substack{\uparrow \\ \text{data}}}{D}, M_i)}^{\text{posterior}} = \frac{p(D \mid \vec{w}, M_i)\, p(\vec{w} \mid M_i)}{p(D \mid M_i)}, \quad \text{where}$$

M_i: model out of a set of L models: $\{M_i\}$ that we wish to compare

$$p(D \mid M_i) = \int d\vec{w}\; p(D \mid \vec{w}, M_i)\, p(\vec{w} \mid M_i)$$

is the model evidence.

Now, we can consider

$$\underbrace{p(M_i \mid D)}_{\text{model posterior}} \sim p(D \mid M_i)\, \underbrace{p(M_i)}_{\substack{\text{often, equal} \\ \text{probs for each} \\ \text{model}}}$$

$\dfrac{p(D \mid M_i)}{p(D \mid M_j)}$ is called a Bayes factor

Finally, the predictive distr'n is given by

$$p(t|\vec{x},\mathcal{D}) = \sum_{i=1}^{L} \overbrace{p(t|\vec{x},M_i,\mathcal{D})}^{\text{predictive distr'n of model } M_i}\, p(M_i|\mathcal{D}) \qquad (**)$$

a mixture distr'n weighted by $p(M_i|\mathcal{D})$

A simple appr'n to (**) is to leave only a single term in the sum corresponding to the most probable model for which $p(M_i|\mathcal{D})$ is at maximum. ← model selection

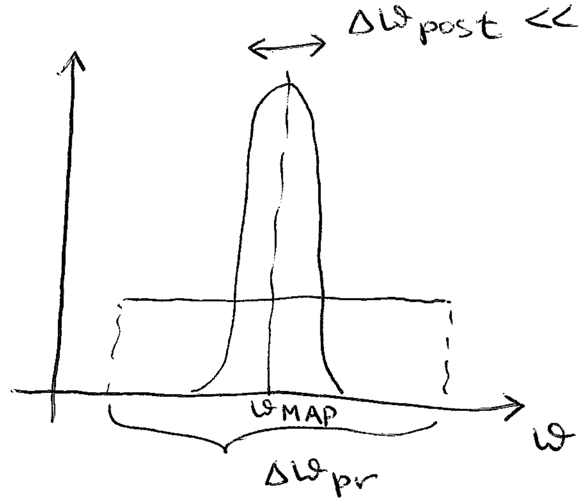
---

Consider a model M_i with a single prm w.

Then $p(w|\mathcal{D}) \sim p(\mathcal{D}|w)\,p(w)$ [dependence on M_i omitted for brevity]

Assume that $p(w|\mathcal{D})$ is peaked @ w_{MAP} with width Δw_{post} and the prior is flat with width Δw_{pr}:

$\Delta w_{post} << \Delta w_{pr}$



$$p(w) = \frac{1}{\Delta w_{pr}}$$

Then $\underbrace{p(\mathcal{D})}_{\text{evidence for } M_i} = \int dw\, p(\mathcal{D}|w)\,p(w) \simeq p(\mathcal{D}|w_{MAP})\,\frac{\Delta w_{post}}{\Delta w_{pr}}$,

or

$$\log P(\mathcal{D}) \simeq \log P(\mathcal{D} \mid \omega_{MAP}) + \underbrace{\log\left(\frac{\Delta\omega_{post}}{\Delta\omega_{pr}}\right)}_{\substack{<0, \text{ acts like a penalty} \\ \text{for model complexity}}}$$

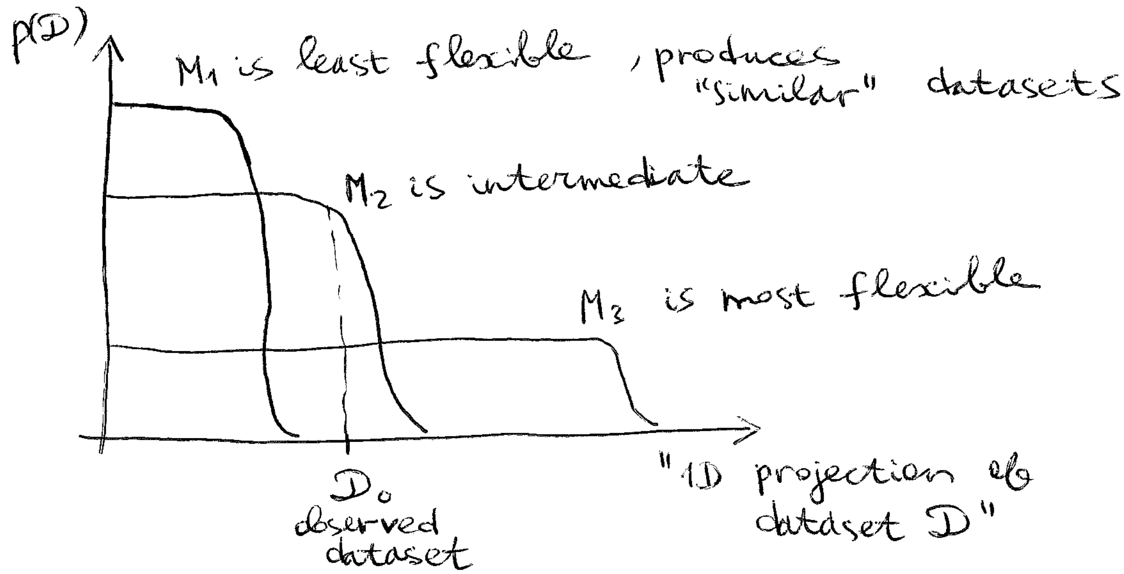
For a model with M prms,

$$\log P(\mathcal{D}) \simeq \log P(\mathcal{D} \mid \omega_{MAP}) + M \log\left(\frac{\Delta\omega_{post}}{\Delta\omega_{pr}}\right)$$

if $\frac{\Delta\omega_{post}}{\Delta\omega_{pr}}$ is the same for each prm.

So, the penalty $\sim M$.

Next, consider 3 models M_1, M_2, M_3 of increasing complexity. Imagine generating synthetic data sets from each of these models: choose prms from $p(\vec{\omega})$ & then generate data from $p(\mathcal{D} \mid \vec{\omega}) \Rightarrow$ compute $p(\mathcal{D})$. (or, more precisely, $p(\mathcal{D} \mid M_i)$, $i = 1,2,3$)



$$\begin{cases} p(\mathcal{D} \mid M_2) > p(\mathcal{D} \mid M_1) & M_1 \text{ did not fit well, too inflexible} \\ p(\mathcal{D} \mid M_2) > p(\mathcal{D} \mid M_3) & M_3 \text{ is too flexible} \end{cases}$$

M₂ wins

The evidence approximation

Consider

$$\underset{\substack{\uparrow \\ \text{training} \\ \text{data}}}{p(t|\vec{t})} = \int d\vec{w}\, d\alpha\, d\beta\, \underbrace{\overbrace{p(t|\vec{w},\beta)}^{\text{likelihood}}\, \underbrace{\overbrace{p(\vec{w}|\vec{t},\alpha,\beta)}^{\text{posterior for } \vec{w}}\, \overbrace{p(\alpha,\beta|\vec{t})}^{\text{posterior for } \alpha,\beta}}_{p(\vec{w},\alpha,\beta|\vec{t})}}_{p(t,\vec{w},\alpha,\beta|\vec{t})}$$

Omitted dependence on $\vec{x}$, $\vec{X}$ for brevity

If $p(\alpha,\beta|\vec{t})$ is sharply peaked around $\hat{\alpha}, \hat{\beta}$, then

$$p(t|\vec{t}) \simeq p(t|\vec{t},\hat{\alpha},\hat{\beta}) = \int d\vec{w}\, p(t|\vec{w},\hat{\beta})\, p(\vec{w}|\vec{t},\hat{\alpha},\hat{\beta})$$

Further, $p(\alpha,\beta|\vec{t}) \sim p(\vec{t}|\alpha,\beta)\, p(\alpha,\beta)$

If $p(\alpha,\beta)$ is flat, we can simply maximize $p(\vec{t}|\alpha,\beta)$ to find $\hat{\alpha}, \hat{\beta}$.

Now, $p(\vec{t}|\alpha,\beta) = \int d\vec{w}\, p(\vec{t}|\vec{w},\beta)\, p(\vec{w}|\alpha) =$

$$= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{D+1}{2}} \int d\vec{w}\, e^{-E(\vec{w})},$$

where

$$E(\vec{w}) = \frac{\beta}{2} \underbrace{\sum_{n=1}^{N} (t_n - \vec{w}^T \vec{\varphi}(\vec{x}_n))^2}_{\|\vec{t} - \Phi\vec{w}\|^2} + \frac{\alpha}{2} \vec{w}^T \vec{w} \quad =$$

$$\overset{(\vec{t}-\Phi\vec{m}_N)^T(\vec{t}-\Phi\vec{m}_N)}{\overbrace{\phantom{\|\vec{t}-\Phi\vec{m}_N\|^2}}}$$

$$\stackrel{\text{ⓔ}}{=} \underbrace{\frac{\beta}{2}\|\vec{t}-\Phi\vec{m}_N\|^2 + \frac{\alpha}{2}\vec{m}_N^T\vec{m}_N}_{= E(\vec{m}_N)} +$$

$$+ \frac{1}{2}(\vec{w}-\vec{m}_N)^T \underbrace{(\alpha I + \beta\Phi^T\Phi)}_{= A = S_N^{-1}} (\vec{w}-\vec{m}_N).$$

Here, $\vec{m}_N = \beta A^{-1}\Phi^T\vec{t} = \beta S_N\Phi^T\vec{t}$

~~Indeed, the last term on the RHS is~~

$$\frac{\alpha}{2}\vec{w}^T\vec{w} - \alpha\vec{w}^T\vec{m}_N + \frac{\alpha}{2}\vec{m}_N^T\vec{m}_N +$$

$$+ \frac{\beta}{2}\vec{w}^T\Phi^T\Phi\vec{w} + \frac{\beta}{2}\vec{m}_N^T\Phi^T\Phi\vec{m}_N -$$

$$- \frac{\beta}{2}\vec{w}^T\Phi^T\Phi\vec{m}_N - \frac{\beta}{2}\vec{m}_N^T\Phi^T\Phi\vec{w}, \text{ so that}$$

$$E(\vec{w}) = \frac{\beta}{2}(\vec{t}-\Phi\vec{m}_N)^T(\vec{t}-\Phi\vec{m}_N) + \frac{\alpha}{2}\vec{w}^T\vec{w} -$$

$$- \alpha\vec{w}^T\vec{m}_N + \alpha\vec{m}_N^T\vec{m}_N + \frac{\beta}{2}\vec{w}^T\Phi^T\Phi\vec{w} +$$

$$+ \frac{\beta}{2}\vec{m}_N^T\Phi^T\Phi\vec{m}_N - \beta(\Phi\vec{w})^T(\Phi\vec{m}_N) =$$

$$= \frac{\beta}{2}\vec{t}^T\vec{t} + \frac{\beta}{2}(\Phi\vec{m}_N)^T(\Phi\vec{m}_N) - \beta(\Phi\vec{m}_N)^T\vec{t} +$$

$$+ \frac{\alpha}{2}\vec{w}^T\vec{w} - \ldots =$$

$$= \frac{\beta}{2}\vec{t}^T\vec{t} + \frac{\beta}{2}(\Phi\vec{w})^T(\Phi\vec{w}) + \frac{\alpha}{2}\vec{w}^T\vec{w}$$

OK, $\Theta(\vec{w}^2)$ terms: "new": $\frac{1}{2}\vec{w}^T A \vec{w} =$

$$= \frac{1}{2}\vec{w}^T(\alpha I + \beta \Phi^T\Phi)\vec{w} =$$

$$= \frac{1}{2}\vec{w}^T \alpha \vec{w} + \frac{\beta}{2}(\Phi\vec{w})^T(\Phi\vec{w}).$$

"old": $\frac{\beta}{2}(\Phi\vec{w})^T(\Phi\vec{w}) + \frac{\alpha}{2}\vec{w}^T\vec{w}$ same!

$\Theta(\vec{w}^0)$: "old": $\frac{\beta}{2}\vec{t}^T\vec{t}$

"new": $\frac{\beta}{2}(\vec{t} - \Phi\vec{m}_N)^T(\vec{t} - \Phi\vec{m}_N) + \frac{\alpha}{2}\vec{m}_N^T\vec{m}_N +$

$$+ \frac{1}{2}\vec{m}_N^T A \vec{m}_N = \frac{\beta}{2}\vec{t}^T\vec{t} + \frac{\beta}{2}(\Phi\vec{m}_N)^T(\Phi\vec{m}_N) -$$

$$\underbrace{-\beta(\Phi\vec{m}_N)^T\vec{t}}_{\beta \vec{m}_N^T \Phi^T \vec{t} \;=\; \beta \vec{m}_N^T A \vec{m}_N} + \frac{\alpha}{2}\vec{m}_N^T\vec{m}_N + \frac{1}{2}\vec{m}_N^T A \vec{m}_N =$$

insert AA^{-1}

$$= \frac{\beta}{2}\vec{t}^T\vec{t} + \frac{\beta}{2}(\Phi\vec{m}_N)^T(\Phi\vec{m}_N) + \frac{\alpha}{2}\vec{m}_N^T\vec{m}_N -$$

$$- \frac{1}{2}\underbrace{\vec{m}_N^T A \vec{m}_N}_{\vec{m}_N^T \alpha \vec{m}_N + \beta(\Phi\vec{m}_N)^T(\Phi\vec{m}_N)} =$$

$$= \frac{\beta}{2}\vec{t}^T\vec{t}$$ same!

Finally, $O(\vec{w}^1)$:

"old": $-\beta(\Phi\vec{w})^T\vec{t}$

"new": $$-\vec{m}_N^T A\vec{w} = -\beta\,\vec{t}^T\Phi\underbrace{(A^{-1})^T}_{\shortparallel \, (A^T)^{-1}}\overset{A^T}{\overset{\shortparallel}{A}}\vec{w} =$$

$$= -\beta\,\vec{t}^T\Phi\vec{w} = \underline{\underline{-\beta(\Phi\vec{w})^T\vec{t}}}$$

same!

$$A^T = \alpha I + \beta(\Phi^T\Phi)^T = \alpha I + \beta\Phi^T\Phi = \underline{\underline{A}}.$$

So, all terms match between (3.79) & (3.80).

Note also that

$$p(\vec{t}\,|\,\vec{w},\beta) = \prod_{n=1}^{N}\mathcal{N}(t_n\,|\,\vec{w}^T\vec{\varphi}(\vec{x}_n),\beta^{-1}) =$$

$$= \mathcal{N}(\vec{t}\,|\,\underbrace{\Phi w}_{\substack{N\text{-dim}\\ \text{vector}}},\underbrace{\beta^{-1}I}_{\substack{N\times N \text{ diag.}\\ \text{cov. matrix}}})$$

Then

$$p(\vec{t}\,|\,\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{D+1}{2}} e^{-E(\vec{m}_N)} \times$$

$$\times \underbrace{\int d\vec{w}\; e^{-\frac{1}{2}(\vec{w}-\vec{m}_N)^T A (\vec{w}-\vec{m}_N)}}_{(2\pi)^{\frac{D+1}{2}} \frac{1}{\overbrace{|A|^{1/2}}^{\det(A)}}}$$

complexity penalty

recall that (3.82)

$$E(\vec{m}_N) = \frac{\beta}{2}\|\vec{t} - \Phi \vec{m}_N\|^2 + \frac{\alpha}{2}\vec{m}_N^T \vec{m}_N$$

$$\log p(\vec{t}\,|\,\alpha,\beta) = \underbrace{\frac{D+1}{2}\log\alpha}_{\text{complexity penalty}} + \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) -$$

$$- E(\vec{m}_N) - \underbrace{\frac{1}{2}\log|A|}_{\text{complexity penalty}}$$

$p(\vec{t}\,|\,\alpha,\beta)$ is basically the model evidence $\Leftrightarrow$

$\Leftrightarrow$ $p(\mathcal{D}\,|\,M_i)$

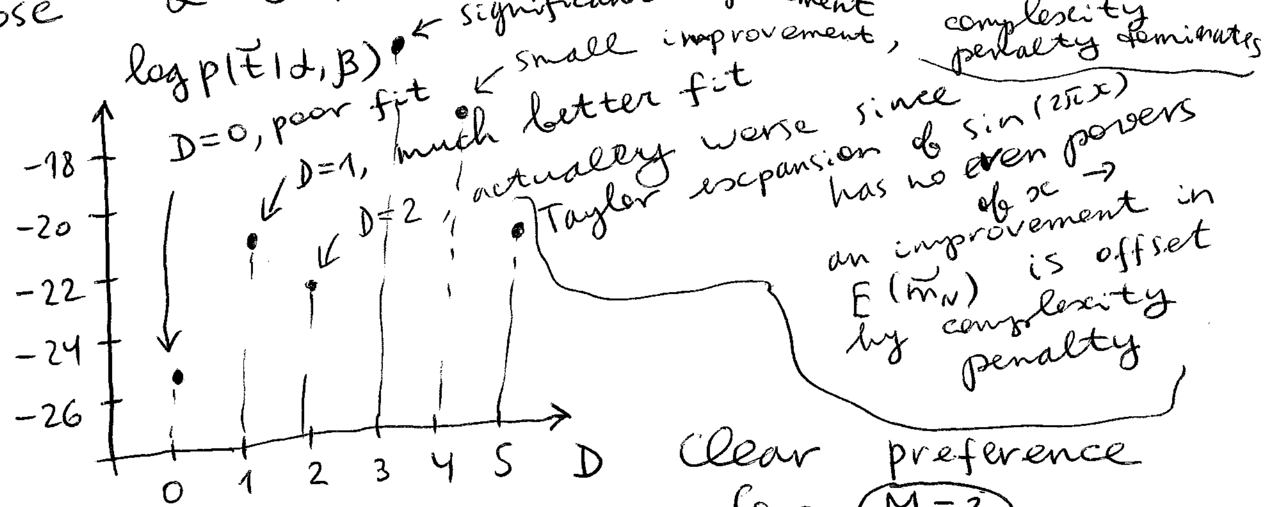
Recall the polynomial regression problem:

fit $y(x,\vec{w}) = \sum_{j=0}^{D} w_j x^j$

to $\sin(2\pi x)$ + noise. $N=10$

and assume β is known as well

Choose $\alpha = 5\times10^{-3}$



Clear preference for $\boxed{M=3}$

Maximizing the evidence function

Consider maximizing $p(\vec{t}\,|\,\alpha, \beta)$ wrt α.

Define $(\beta \, \Phi^T \Phi)\, \vec{u}_i = \lambda_i \vec{u}_i$ eigenvector eq'n

Note that A has eigenvalues of $\alpha + \lambda_i$

Now, consider

$$\frac{\partial}{\partial \alpha} \log |A| = \frac{\partial}{\partial \alpha} \log \left\{ \prod_{i=0}^{D} (\lambda_i + \alpha) \right\} =$$

$$= \sum_i \frac{1}{\lambda_i + \alpha} .$$

Then $\frac{\partial}{\partial \alpha} \log p(\vec{t}\,|\,\alpha, \beta) = \frac{D+1}{2\alpha} - \frac{1}{2} \vec{m}_N^T \vec{m}_N -$

$$- \frac{1}{2} \sum_i \frac{1}{\lambda_i + \alpha} = 0 \quad \text{, or}$$

$$\alpha \, \vec{m}_N^T \vec{m}_N = \underbrace{(D+1)}_{\sum_i \frac{\lambda_i + \alpha}{\lambda_i + \alpha}} - \alpha \sum_i \frac{1}{\lambda_i + \alpha} \equiv \gamma ,$$

where

$$\gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} .$$

So, $\hat{\alpha} = \frac{\gamma}{\vec{m}_N^T \vec{m}_N}$ implicit eq'n since both γ & $\vec{m}_N$ depend on α

Solve iteratively:

choose $\alpha \rightarrow$ compute $\gamma, \vec{m}_N \rightarrow$ update α, etc.

Note that λ_i's can be computed <u>once</u>.

Note that we assume

$$\frac{\partial}{\partial \alpha}\vec{m}_N \approx 0, \quad \frac{\partial}{\partial \beta}\vec{m}_N \approx 0.$$

Indeed, $\begin{cases} \vec{m}_N = \beta A^{-1}\varphi^T\vec{t}, \\ A = \alpha I + \beta \varphi^T\varphi. \end{cases}$

$$\Downarrow$$

$$\frac{\partial \vec{m}_N}{\partial \alpha} = \beta \underbrace{\frac{\partial A^{-1}}{\partial \alpha}}_{-A^{-1}\frac{\partial A}{\partial \alpha}A^{-1} = -(A^{-1})^2} \varphi^T\vec{t} = -\beta (A^{-1})^2\varphi^T\vec{t}. \leftarrow$$

suppressed by A^{-1} compared to $\vec{m}_N$

derive using $\frac{\partial}{\partial \alpha}[\,A^{-1}A = I\,]$

Further,

$$\frac{\partial \vec{m}_N}{\partial \beta} = A^{-1}\varphi^T\vec{t} + \beta \frac{\partial A^{-1}}{\partial \beta}\varphi^T\vec{t} = A^{-1}\varphi^T\vec{t} - \beta A^{-1}\underbrace{\frac{\partial A}{\partial \beta}}_{\varphi^T\varphi} A^{-1}\varphi^T\vec{t} =$$

$$= A^{-1}\varphi^T\vec{t} - A^{-1}[A - \alpha I]A^{-1}\varphi^T\vec{t} =$$

$$= \alpha (A^{-1})^2\varphi^T\vec{t}. \leftarrow$$

suppressed by A^{-1} compared to $\vec{m}_N$

Now, compute $\frac{\partial}{\partial\beta} \log p(\vec{t}|\alpha,\beta)$.

Note that $\lambda_i \sim \beta \Rightarrow \frac{\partial \lambda_i}{\partial\beta} = \frac{\lambda_i}{\beta}$.

Then $\frac{\partial}{\partial\beta} \log|A| = \frac{\partial}{\partial\beta} \sum_i \log(\lambda_i + \alpha) =$

$= \frac{1}{\beta} \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta}$, and

$$\frac{\partial}{\partial\beta} \log p = \frac{N}{2\beta} - \frac{\gamma}{2\beta} - \frac{1}{2} \sum_{n=1}^{N} (t_n - \vec{m}_N^T \vec{\varphi}(\vec{x}_n))^2 = 0,$$

or

$$\frac{1}{\hat{\beta}} = \frac{1}{N-\gamma} \sum_{n=1}^{N} (t_n - \vec{m}_N^T \vec{\varphi}(\vec{x}_n))^2.$$

Again, solve iteratively:

$\beta \to \vec{m}_N, \gamma \to$ update β, etc.

So, now we have $\hat{\alpha}, \hat{\beta}$:

predictive distribution

$p(t|\vec{t}) = p(t|\vec{t}, \hat{\alpha}, \hat{\beta}) =$

$= \underbrace{\int d\vec{w}\, p(t|\vec{w}, \hat{\beta})\, p(\vec{w}|\vec{t}, \hat{\alpha}, \hat{\beta})}_{\text{same as (3.57)}} = \mathcal{N}(t|\vec{m}_N^T \vec{\varphi}(\vec{x}), \sigma_N^2(\vec{x}))$,

new value of $\vec{x}$

where $\begin{cases} \vec{m}_N = \hat{\beta} S_N \Phi^T \vec{t}, \quad S_N^{-1} = \hat{\alpha} I + \hat{\beta} \Phi^T \Phi, \\ \sigma_N^2(\vec{x}) = \beta^{-1} + \vec{\varphi}^T(\vec{x}) S_N \vec{\varphi}(\vec{x}). \end{cases}$

Effective number of prms

Consider $\alpha = \dfrac{\gamma}{\vec{m}_N^T \vec{m}_N}$ again.

Recall that λ_i's are eigenvalues of a positive definite matrix: $\lambda_i > 0$.

$$0 < \frac{\lambda_i}{\lambda_i + \alpha} \leq 1 \implies 0 < \gamma \leq D+1$$

If $\lambda_i \gg \alpha \implies \dfrac{\lambda_i}{\lambda_i+\alpha} \simeq 1$, and

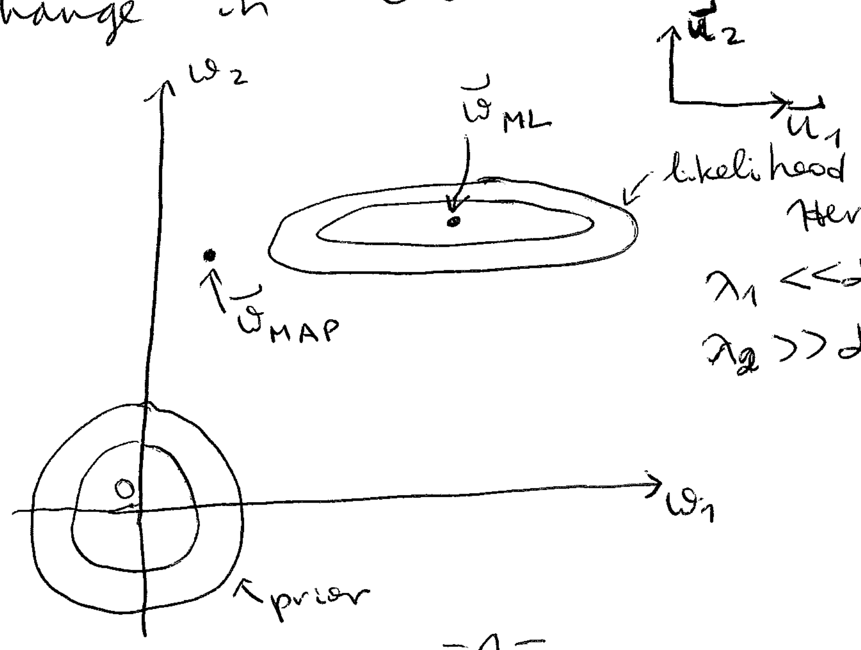
$w_i \simeq w_i^{ML}$.

Conversely, if $\lambda_i \ll \alpha \implies \dfrac{\lambda_i}{\lambda_i+\alpha} \simeq 0$,

and $w_i \simeq 0$.

Indeed, recall that $A = \alpha I + \beta \Phi^T \Phi$ is the curvature in the $\vec{w}$ space since $E(\vec{w}) = E(\vec{m}_N) + \frac{1}{2}(\vec{w}-\vec{m}_N)^T A (\vec{w}-\vec{m}_N)$.

Thus eigenvalues of A, $\lambda_i + \alpha$, will determine how quickly $\vec{w}$'s change in various directions.



Here,

$\lambda_1 \ll \alpha \implies w_{MAP,1} \simeq 0$

$\lambda_2 \gg \alpha \implies w_{MAP,2} \approx w_{ML,2}$

Thus γ determines the eff. # prms.

$\hookrightarrow \gamma$ non-zero prms+ $+[(D+1)-\gamma]$ nearly zero prms

Note that

$$\underbrace{\frac{1}{\hat{\beta}}}_{\text{"MAP" value}} = \frac{1}{N-\gamma}\sum_{n=1}^{N}(t_n - \vec{m}_N^T \vec{\varphi}(\vec{x}_n))^2$$

Recall that

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \vec{w}_{ML}^T \vec{\varphi}(\vec{x}_n))^2$$

$N \to N-\gamma$ in the Bayesian result, correcting for the "ML bias".

If we consider the $N \gg D+1$ limit, $\gamma = M$ since $\beta \Phi^T \Phi \uparrow$ as $N\uparrow$ due to a sum $\sum_{n=1}^{N}$. Thus $\lambda_i \uparrow$ as $N\uparrow$, $\forall i$

and $\frac{\lambda_i}{\lambda_i+\alpha} \to 1 \Rightarrow \gamma \to M$.

In this case,

$$\left[\quad \hat{\alpha} \simeq \frac{M}{\vec{m}_N^T \vec{m}_N}, \quad \hat{\beta}^{-1} \simeq \frac{1}{N}\sum_{n=1}^{N}(t_n - \vec{m}_N^T \vec{\varphi}(\vec{x}_n))^2. \right]$$

still iterative but easier since γ does not need to be re-computed.

Ex.:

(β known for simplicity)