BAYESIAN LEARNING

- There is a natural interpretation of the regularization procedure in terms of the Bayesian approach:

  Starting point, same as before, the likelihood:

$$p(D|\vec{\theta}) = \prod_{i=1}^{N} \mathcal{N}\left(y_i \mid f(x_i, \vec{\theta}), \sigma^2\right)$$

  But now we need a prior distribution for $\vec{\theta}$:

  Suppose $p(\vec{\theta}) = \mathcal{N}(\vec{\theta} \mid \vec{0}, \alpha^{-1} \mathbb{1})$

$$= \left(\frac{\alpha}{2\pi}\right)^M \exp\left\{-\frac{\alpha}{2}\, \vec{\theta}^T \cdot \vec{\theta}\right\}$$

  Then the posterior $\qquad M = \dim \vec{\theta}$

$$p(\vec{\theta}|D) \propto p(D|\vec{\theta})\, p(\vec{\theta})$$

$$\ln p(\vec{\theta}|D) = \ln p(D|\vec{\theta}) + \ln p(\vec{\theta}) + \text{const.}$$

$$-\ln p(\vec{\theta}|D) = \frac{1}{2\sigma^2}\sum_i \left(y_i - f(x_i,\vec{\theta})\right)^2 + \frac{\alpha}{2}\vec{\theta}^T\cdot\vec{\theta}$$

$$+ \text{const}$$

$$= \frac{1}{\sigma^2}\left\{\frac{1}{2}\sum_i \left(y_i - f(x_i,\vec{\theta})\right)^2 + \frac{\tilde{\alpha}}{2}\vec{\theta}^T\cdot\vec{\theta}\right\}$$

where $\tilde{\alpha} \equiv \alpha\sigma^2$

$= E_{ridge}$!

So we see that obtaining the $\vec{\theta}_{MAP}$ in the Bayesian approach is equivalent to $\vec{\theta}_{ridge}$, for a Gaussian choice of the prior.

Bayesian Linear Regression

We have just seen that the Bayesian approach automatically encodes a regularized fitting. However its most attractive feature is that it provides a natural way to estimate prediction uncertainties (through the "predictive distribution" we saw in lecture #2)

$$p(y_{new} | \vec{x}_{new}, D) = \int d\vec{\theta}\; p(\vec{\theta}|D)\; p(y_{new} | \vec{x}_{new}, \vec{\theta})$$

Remember that in order to compute $p(\vec{\theta}|D)$ in principle you need to compute the normalisation (the "evidence"):

$$\int d\vec{\theta}\; p(D|\vec{\theta})\; p(\vec{\theta})$$

There are only few cases where these integrals are analytical, and in real-life problems this is typically not the case. The most popular academic example is:

Likelihood (1 point): $p(y | \vec{x}, \vec{\theta}) = N(y | f(\vec{x}, \vec{\theta}), \Sigma)$ ← assume known

$$f(\vec{x}, \vec{\theta}) = \vec{\theta}^T \cdot \vec{\phi}(\vec{x}) = \theta_1 \phi_1(\vec{x}) + \dots \theta_M \phi_M(\vec{x})$$

- Prior: $p(\vec{\theta}) = N(\vec{\theta} | \vec{\mu}_0, \Sigma_0)$

So the posterior $p(\vec{\theta}|\vec{y})$ and the 'evidence' $p(y)$ are both Gaussians,

only because since $f(x;\theta)$ is linear in θ, then in the exponent we will have a polynomial of order 2 in θ, which can be recasted as a Gaussian by completing the square. So the result is

$$p(\vec{\theta}|\vec{y}) = N(\vec{\theta}\,|\,\vec{\mu}_N, S_N)$$

where $$\vec{\mu}_N = S_N\left\{\Phi^T C^{-1}\vec{y} + S_0^{-1}\vec{\mu}_0\right\}$$

$$S_N = \left(\Phi^T C^{-1}\Phi + S_0^{-1}\right)^{-1}$$

for the posterior, while for the evidence is:

$$p(\vec{y}) = N\left(\vec{y}\,|\,\Phi\vec{\mu}_0\,,\; C + \Phi S_0 \Phi^T\right)$$

$N\times 1$ $\quad (N\times D)\times(D\times 1)$ $\quad N\times N$ $\quad (N\times D)\times(D\times D)\times(D\times N)$

In an analogous fashion to the evidence, the predictive distribution is an integral over two Gaussians, and it's found to be:

$$p(y_{new} \mid \vec{x}_{new}, D) = N\left(y_{new} \mid \vec{\mu}_N^T \cdot \vec{\phi}(\vec{x}_{new}), \sigma_N^2(\vec{x}_{new})\right)$$

where $$\sigma_N^2(\vec{x}_{new}) = \sigma^2(x_{new}) + \vec{\phi}(\vec{x}_{new})^T S_N \vec{\phi}(x_{new})$$

statistical noise at this point

So even if σ^2 (the data noise is the same $\forall \vec{x}$, still σ_N^2 would get a dependence on $\vec{x}$ coming from the 2nd term. As we will see, this dependence is such that the variance grows for points $\vec{x}_{new}$ farther away from any of the $\vec{x}_i$ considered in the dataset, and viceversa: the variance of the predictive

distribution shrinks as $\vec{x}_{new}$ gets closer to the $\vec{x}_i$. This is intuitive: as we get further away from the points used in the fit, our predictions become more uncertain.

[Show notebook on Bayesian prediction]

- Before understanding analytically this behaviour with the # of observations (N points), let use visualise in another way the sequential nature of the Bayesian learning:

  The posterior distribution after seeing N datapoints

$$p(\vec{w} \mid \vec{y}_{(N)}) \propto \underbrace{p(\vec{y}_{(N)} \mid \vec{w})}_{\text{joint likelihood of N observations}} p(\vec{w})$$ } prior before any observation

$$\vec{y}_{(N)} = \{y_1, y_2, \dots, y_N\}$$

$$= p(y_N|\vec{w})\, p(\vec{y}_{(N-1)}|\vec{w})\, p(\vec{w})$$

likelihood of the N^{th} observation

joint likelihood of the rest of $N-1$ observations

$$\propto p(y_N|\vec{w})\, p(\vec{w}|\vec{y}_{(N-1)})$$

prior before the N^{th} observation

posterior after seeing $N-1$ observations

So the posterior after one observation becomes the prior to the 2^{nd} observation, and so on...

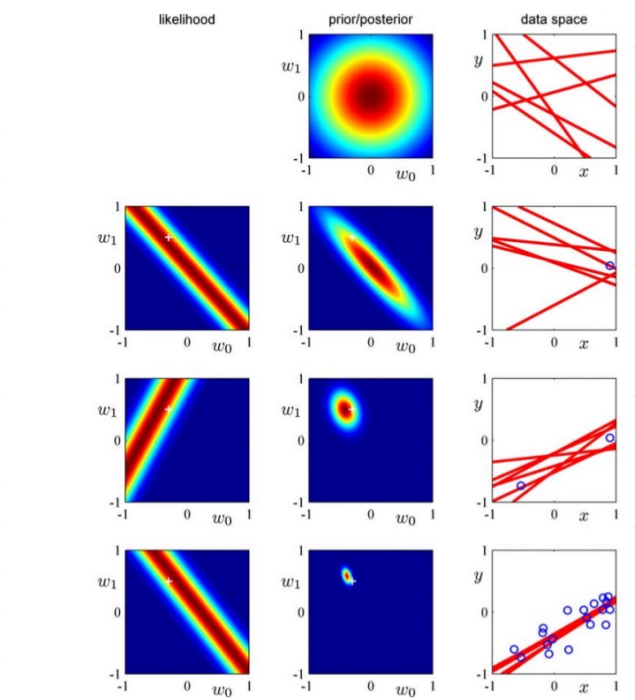


Figure 3.7 Illustration of sequential Bayesian learning for a simple linear model of the form $y(x, \mathbf{w}) = w_0 + w_1 x$. A detailed description of this figure is given in the text.

- How does the posterior of $\vec{w}$ behaves with the # of points, N?

$$p(\vec{\theta}|\vec{y}) = \mathcal{N}(\vec{\theta}\,|\,\vec{\mu}_N, S_N)$$

Just for simplicity. Results in general still valid

Assume covariance matrix of likelihood $C = \sigma^2 \cdot \mathbb{1}$

$$\vec{\mu}_N = S_N \left\{ \sigma^{-2} \Phi^T \cdot \vec{y} + S_0^{-1} \vec{\mu}_0 \right\}$$

$$S_N = \left(\sigma^{-2} \Phi^T \Phi + S_0^{-1} \right)^{-1}$$

1) $N = 0$ (i.e. no observations yet)

$$S_N = \left(0 + S_0^{-1} \right)^{-1} = S_0$$

$$\vec{\mu}_N = S_N \left\{ 0 + S_0^{-1} \vec{\mu}_0 \right\} = \vec{\mu}_0$$

So you get back the prior

2) $N \to \infty$

$\Phi^T \Phi$ grows with N { This matrix is less and less singular

also $\Phi^T \cdot \vec{y}$ grows { more terms in the scalar product, which grows { modulo fine tuning

so $S_N \approx \left(\sigma^{-2}\Phi^T\Phi\right)^{-1}$ ⇒ variance decreases with N

$$\vec{\mu}_N \approx S_N\left\{\sigma^{-2}\Phi^T\cdot\vec{y}\right\}$$

$$\approx \left(\Phi^T\cdot\Phi\right)^{-1}\Phi^T\cdot\vec{y} \quad \text{as } N\to\infty$$

↳ The mean of the posterior tends to the MLE solution! → No prior dependence

- **Bayesian Model comparison**

Even though the introduction of a prior plays the role of a regularizer when doing the fit, and a highly parametrised model can be fitted to some simple data without doing overfit, this may still not be a good model:

In Bayesian approach we use the Evidence (denominator of the Bayes theorem) to compare among different models

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

$p(D|\theta)$ → $p(D|\text{model})$

So imagine you have 2 models m_1 & m_2, with Evidences $p(D/m_1)$ and $p(D/m_2)$

$$p(m_i/D) \propto p(D/m_i)\, p(m_i)$$

$p(m_i)$: prior of the model itself (not of its parameters!)

If all models are assumed to have the same prior, Then the most probable model among m_1 and m_2 is obtained by computing the "Bayes factor"

$$B \equiv p(D/m_1) \Big/ p(D|m_2)$$

for a sufficiently large B (e.g $> \mathcal{O}(10)$) one says There is a strong "evidence" in favour of m_1

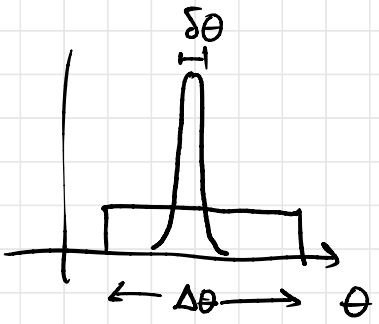
Finally, lets get an intuition about which are the features of a model which may penalize it from the evidence point of view.

$$p(D|m) = \int d\vec{\theta}\; p(D|\vec{\theta}, m)\, p(\vec{\theta}|m)$$

Suppose the following simplifications (which however do not change the general picture)

- only 1 parameter
- likelihood sharply peaked around $\theta = \theta^*$, with width $= \delta\theta$
- flat prior $p(\theta) = \frac{1}{\Delta\theta}$

$$p(D|m) \simeq \frac{\delta\theta}{\Delta\theta}\, p(D|\theta^*)$$



* Penalization of models having large prior uncertainties

• Now assume we have P parameters

$$\vec{\theta} = \{\theta_k\}_{k=1}^{P}$$

Suppose a similar ratio

$$\frac{\delta\theta_k}{\Delta\theta_k} \quad \forall\, k$$

Then the evidence

$$p(D|m) \approx p(D|\vec{\theta}^*) \left[\frac{\delta\theta}{\Delta\theta}\right]^M$$

* Complex models with many parameters are penalized because of the 2nd factor. However typically these larger models fit better the data, so $p(D|\vec{\theta}^*)$ is larger.

⇒ Optimal model should have an intermediate complexity