{Lecture 3}

Define $h(\vec{x}) = \int dt\, t\, p(t|\vec{x}) = E[t|\vec{x}]$

optimal prediction under squared loss f'n

Then $E[L] = \int d\vec{x}\, p(\vec{x}) [y(\vec{x}) - h(\vec{x})]^2 +$

($y(\vec{x})$: model)

$$+ \int dt\, d\vec{x}\, p(\vec{x}, t) [h(\vec{x}) - t]^2$$

indep. of $y(\vec{x})$, reflects intrinsic noise/scatter in the data

Imagine modeling $h(\vec{x})$ with $y(\vec{x}, \vec{w})$.

Consider multiple (K) datasets of size N.

For each dataset k, we can obtain

$y(\vec{x}, \vec{w}_{ML,k}) \equiv y(\vec{x}; \mathcal{D}_k)$ ($\mathcal{D}_k$: particular dataset)

Now, consider

$(y(\vec{x}; \mathcal{D}_k) - h(\vec{x}))^2 = (y(\vec{x}; \mathcal{D}_k) - E(y(\vec{x}; \mathcal{D}_k)) +$

$+ E(y(\vec{x}; \mathcal{D}_k)) - h(\vec{x}))^2$ ⇒

$E(y(\vec{x}; \mathcal{D}_k)) = \frac{1}{K} \sum_{k=1}^{K} y(\vec{x}; \mathcal{D}_k)$

$\Rightarrow$ $E[(y(\vec{x};\mathcal{D}_k) - h(\vec{x}))^2] =$

$= E[(y(\vec{x};\mathcal{D}_k) - E(y(\vec{x};\mathcal{D}_k)))^2] +$

$+ E[(E(y(\vec{x};\mathcal{D}_k)) - h(\vec{x}))^2] +$

$+ 2E[(y(\vec{x};\mathcal{D}_k) - E(y(\vec{x};\mathcal{D}_k)))(E(y(\vec{x};\mathcal{D}_k)) - h(\vec{x}))]$ (1)

since $E(y) - E(E(y)) = E(y) - E(y) = 0$, so the underbraced factor $(y(\vec{x};\mathcal{D}_k) - E(y(\vec{x};\mathcal{D}_k)))$ is 0; the other factor $(E(y(\vec{x};\mathcal{D}_k)) - h(\vec{x}))$ is indep. of the E-operator

$= \underbrace{E[(E(y) - h(\vec{x}))^2]}_{(E(y) - h(\vec{x}))^2} + E[(y - E(y))^2]$

$(E(y) - h(\vec{x}))^2$: bias^2; $E[(y - E(y))^2]$: variance

So, expected loss $=$ (bias)^2 + variance + noise

(expected loss: to be minimized)

Expect "rigid" models to have high bias / low variance & "flexible" models to have low bias / high variance

above, a single $\vec{x}$ was considered. Over all values of $\vec{x}$,

$(\text{bias})^2 = \int d\vec{x}\, p(\vec{x}) [E[y(\vec{x};\mathcal{D}_k)] - h(\vec{x})]^2$

$\text{variance} = \int d\vec{x}\, p(\vec{x})\, E[(y(\vec{x};\mathcal{D}_k) - E(y(\vec{x};\mathcal{D}_k)))^2]$

$\text{noise} = \int dt\, d\vec{x}\, p(\vec{x},t) [h(\vec{x}) - t]^2$

Example

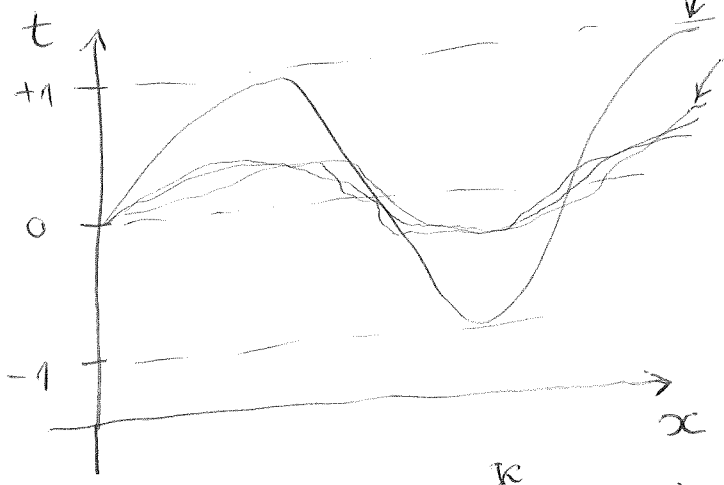
Consider $K = 100$ data sets, each with $N = 25$ points, sampled from $\sin(2\pi x)$ + noise.

To each data set, we fit a model with 24 Gaussian basis functions by minimizing

$$\tilde{E}(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \vec{w}^T\vec{\varphi}(\vec{x}_n))^2 + \frac{\lambda}{2}\vec{w}^T\vec{w}$$

Consider several values of λ:

$\ln\lambda = 2.6$ "λ large"



"true" model $\sin(2\pi x)$

$y(x, \vec{w}_{ML,k})$

average of all these curves is

$$E[y(x;\mathcal{D}_k)] = \frac{1}{K}\sum_{k=1}^{K} y(x;\mathcal{D}_k) \equiv \bar{y}(x)$$

"average prediction"

low variance
high bias

$$h(x) = \frac{1}{K}\sum_{k=1}^{K} t_k(x)$$

"average target"

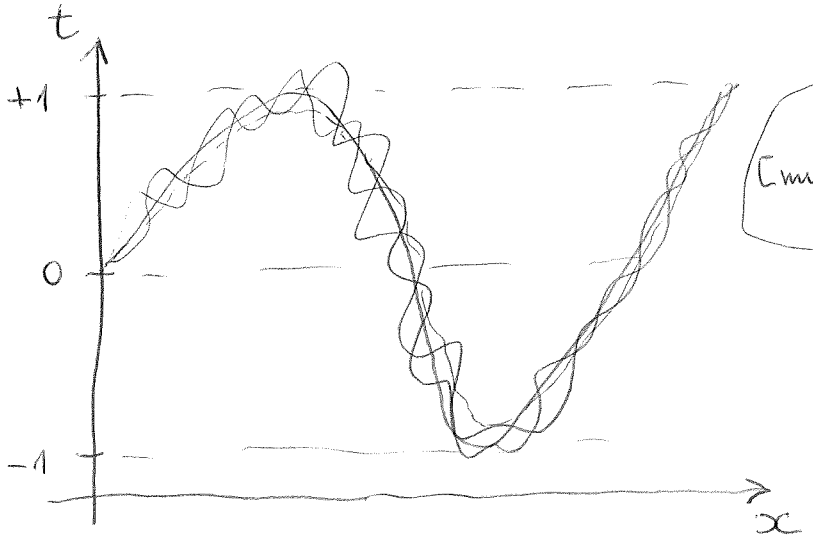
Finally,

$$(\text{bias})^2 = \frac{1}{N}\sum_{n=1}^{N}[\bar{y}(x_n) - h(x_n)]^2$$

$$\text{variance} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}[y(x_n, \vec{w}_{ML,k}) - \bar{y}(x_n)]^2$$

$$\text{noise} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}(h(x_n) - t_k(x_n))^2$$

Now, consider $\ln \lambda = -0.31$ "medium"

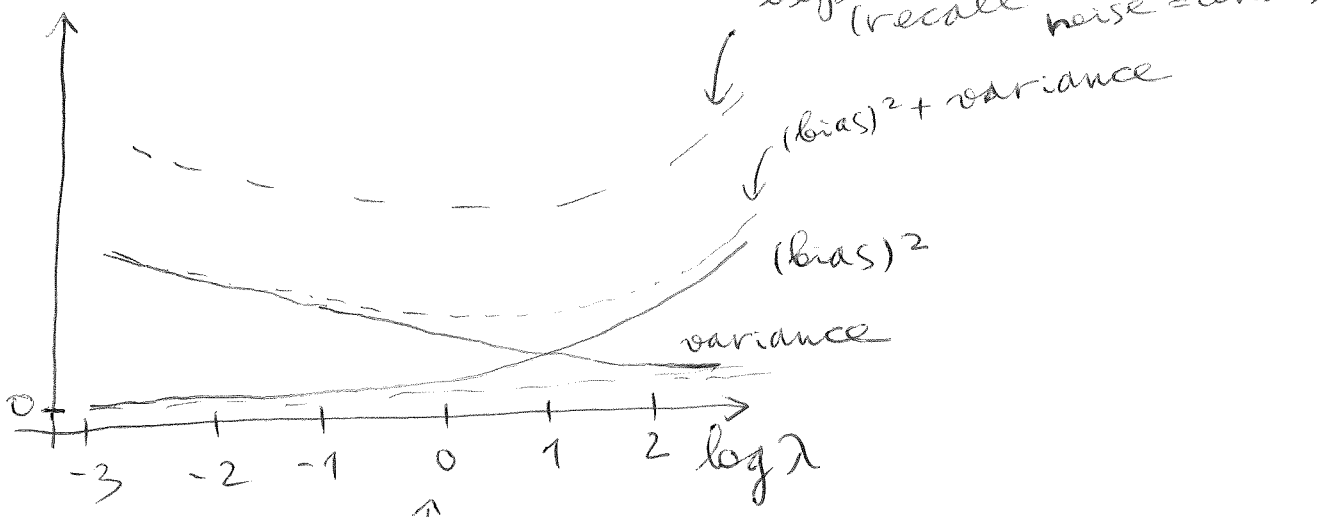


higher variance
[much] lower bias

If we try $\ln \lambda = -2.4$ "low"

very low bias
high variance

Overall,



↑ choose λ to minimize expected loss and either refit the final model on all data or just average K fit with the best value of λ

(separate models)

Bayesian linear regression

Assume $\underbrace{p(\vec{w}|\alpha)}_{\text{prior}} = \mathcal{N}(\vec{w}|\vec{0}, \alpha^{-1}I)$

Then the posterior is gaussian as well, since the likelihood

$$p(\vec{t}|\vec{w}) \propto e^{-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \vec{w}^T\vec{\varphi}(\vec{x}_n))^2}$$

[assume that β is known for now]

is gaussian & we can complete the square:

$$p(\vec{w}|\vec{t}) \sim \underbrace{e^{-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \vec{w}^T\vec{\varphi}(\vec{x}_n))^2}}_{\text{likelihood}} \underbrace{e^{-\frac{\alpha}{2}\vec{w}^T\vec{w}}}_{\text{prior}}$$

Further,

$$p(\vec{w}|\vec{t}) = \mathcal{N}(\vec{w}|\vec{m}_N, S_N)$$

$\vec{m}_N$: $D+1$ vector; S_N: $(D+1)\times(D+1)$ covariance matrix

N = # datapoints

Recall that

$$(*)\quad \mathcal{N}(\vec{x}|\vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\underbrace{|\Sigma|^{1/2}}_{\det(\Sigma)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})}$$

($\vec{\mu}$: mean vector; Σ: covariance matrix)

D-dim gaussian

In our case,

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

$\vec{w}_{MAP} = \vec{m}_N$ (mean = mode for gaussians)

$\alpha \to 0$ (infinitely broad prior):

$\vec{m}_N \to \underbrace{(\varphi^T\varphi)^{-1}\varphi^T\vec{t}}_{\vec{w}_{ML}}$

(2.113) – (2.117):

$$\begin{cases} p(\vec{x}) = \mathcal{N}(\vec{x}\,|\,\vec{\mu}', \Lambda^{-1}) \\ p(\vec{y}|\vec{x}) = \mathcal{N}(\vec{y}\,|\,A\vec{x}+\vec{b},\ L^{-1}) \end{cases} \Rightarrow p(\vec{x}|\vec{y}) = \mathcal{N}(\vec{x}\,|\,\overbrace{S[A^T L(\vec{y}-\vec{b}) + \Lambda\vec{\mu}']}^{\vec{m}},\ S),$$ where

$$S = (\Lambda + A^T L A)^{-1}.$$

Here, $$\begin{cases} p(\vec{w}|\alpha) = \mathcal{N}(\vec{w}\,|\,\vec{0},\ \alpha^{-1} I) & \text{prior} \\ p(\vec{t}|\vec{w}) = \mathcal{N}(\vec{t}\,|\,\varphi\vec{w},\ \beta^{-1} I) & \text{likelihood} \end{cases}$$

↑

Indeed,

$$p(\vec{t}|\vec{w}) \sim e^{-\frac{1}{2}[(t_1 - \underbrace{w_j \varphi_j(\vec{x}_1)}_{\text{sum over } j})\beta\,(t_1 - w_j \varphi_j(\vec{x}_1)) + \ldots + (t_N - w_j \varphi_j(\vec{x}_N))\beta\,(t_N - w_j \varphi_j(\vec{x}_N))]}.$$

Clearly, in eq'n (*) $\Sigma^{-1} = \beta I$ and

$$\underbrace{\vec{\mu}}_{N} = \begin{pmatrix} w_j \varphi_j(\vec{x}_1) \\ \vdots \\ w_j \varphi_j(\vec{x}_N) \end{pmatrix} = \begin{pmatrix} \varphi_{1j} w_j \\ \vdots \\ \varphi_{Nj} w_j \end{pmatrix} = \varphi\vec{w}$$

$$N\left\{\underbrace{\begin{pmatrix} \varphi \end{pmatrix}}_{D+1}\right\}\{\vec{w}\}_{D+1} = \{\vec{\mu}\}_N, \text{ as expected}$$

So, $$\vec{x} \to \vec{w} \Rightarrow \begin{cases} \vec{\mu}' = 0, \\ \Lambda^{-1} = \alpha^{-1} I \end{cases}$$

$$\vec{y} \to \vec{t} \Rightarrow \begin{cases} A = \varphi\,, \quad \vec{b} = 0 \\ L^{-1} = \beta^{-1} I \end{cases}$$

Finally, $S^{-1} = \alpha I + \beta \varphi^T \varphi$ and $\vec{m} = S\varphi^T \beta \vec{t} = \beta S \varphi^T \vec{t}$.

So, $S = S_N$ & $\vec{m} = \vec{m}_N$ from opposite page

We argued before that maximizing $p(\vec{w}|\vec{t})$ is like minimizing $\tilde{E}(\vec{w})$ with $\lambda = \alpha/\beta$.

Example consider $x \to t$ (both 1D)
input target

$y(x, \vec{w}) = w_0 + w_1 x$ linear model

Synthetic data:

$t = f(x, \vec{a}) + \text{noise}$, where

$f(x, \vec{a}) = a_0 + a_1 x$ $\begin{cases} a_0 = -0.3 \\ a_1 = 0.5 \end{cases}$

$\text{noise} = \mathcal{N}(\xi | 0, \sigma^2)$, $\sigma = 0.2$

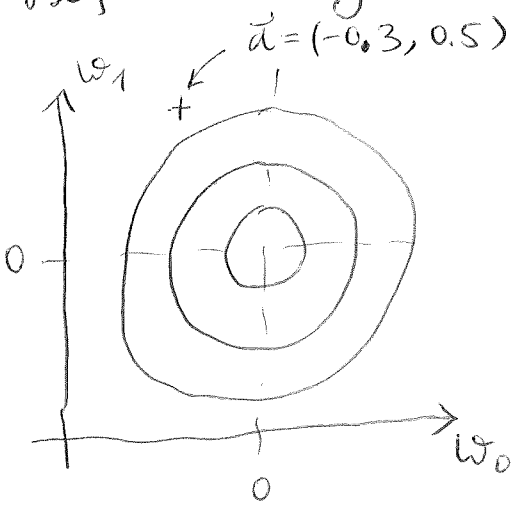
Sampling: choose x_n uniformly in the $[-1, 1]$ range $\Rightarrow$ evaluate $f(x_n, \vec{a})$ $\Rightarrow$ generate noise value ξ_n $\Rightarrow$ get $t_n = f(x_n, \vec{a}) + \xi_n$

$n = 1, \ldots, N$

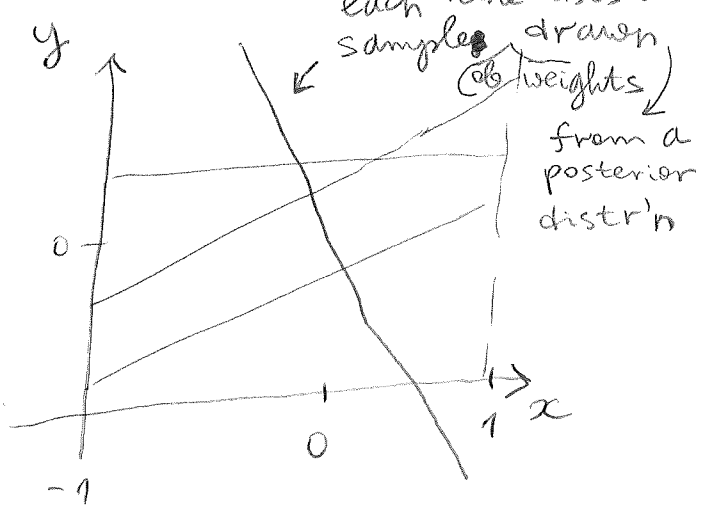
$\nearrow$ these will be relaxed later

- assume $\beta = \frac{1}{\sigma^2} = 25$ is known exactly.
- Choose $\alpha = 2.0$

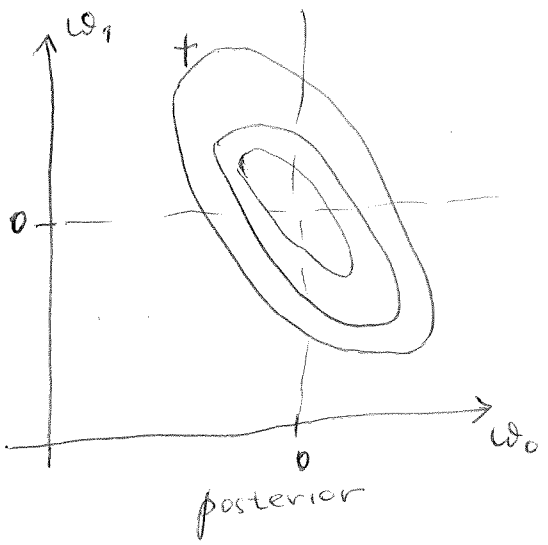
Before any data is observed:

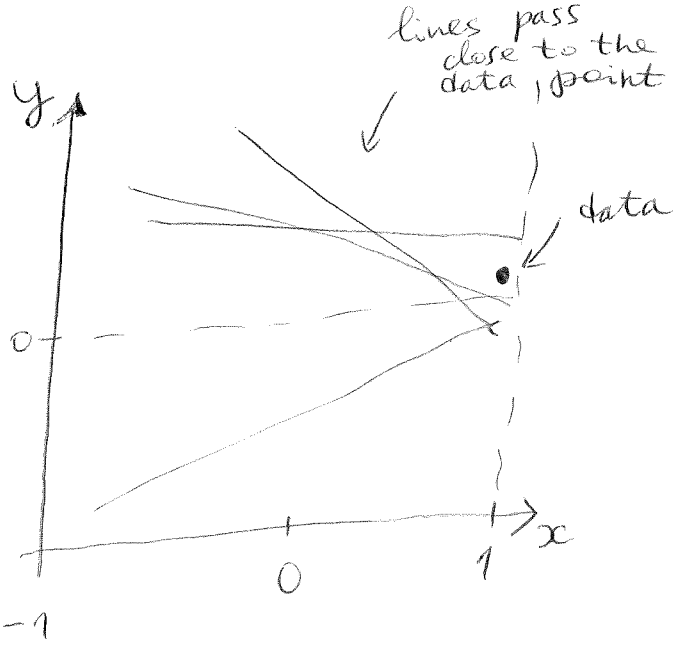


posterior = prior

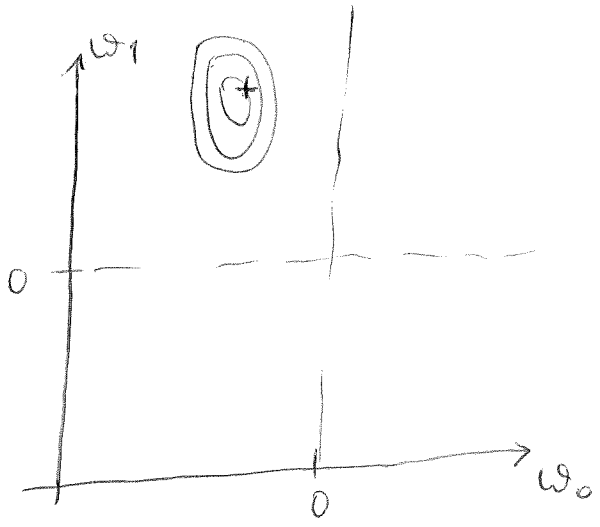


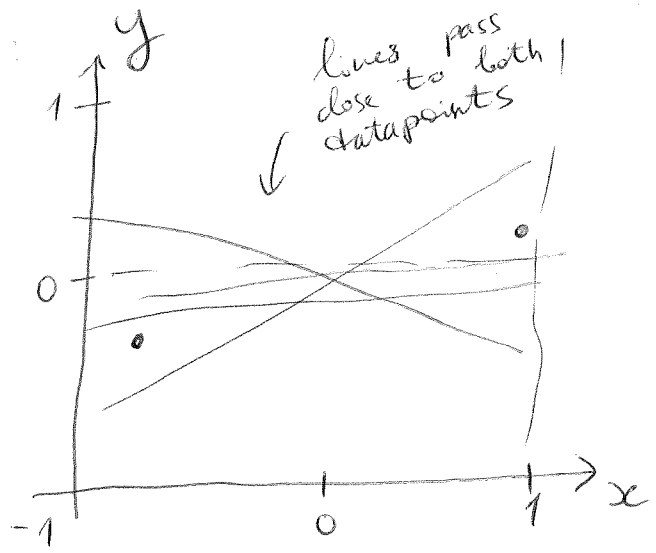
One data point:



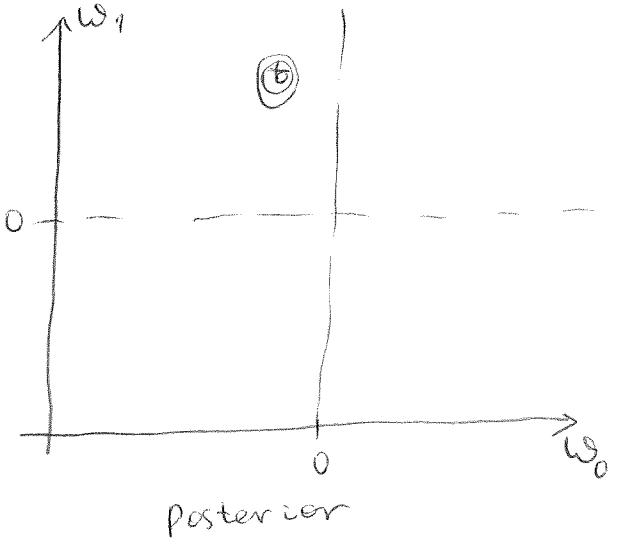


two data points:





Many data points:



Posterior