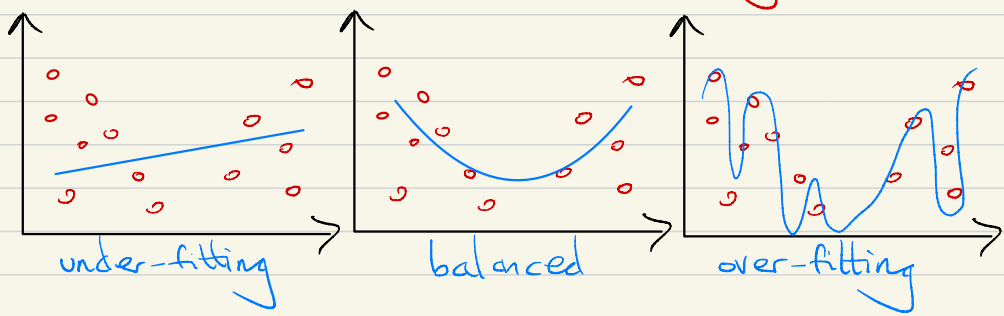
Lecture 18

- Regularization → applying additional constraints to avoid over-fitting.



$$(Y-M\theta)^T C^{-1}(Y-M\theta) \Rightarrow (Y-M\theta)^T C^{-1}(Y-M\theta) + \lambda\theta^T\theta$$

λ: regularization parameter

... called RIDGE REGRESSION ⇒ $\hat{\theta} = (M^T C^{-1} M + \lambda I)^{-1} M^T C^{-1} Y$

NOTE: From a Bayesian perspective, we have just applied a prior constraint on θ such that

$$p(\theta) \propto e^{-\frac{\lambda}{2}\theta^T\theta}$$

GENERALLY ... $\ln L \propto -\frac{1}{2}(Y-M\theta)^T C^{-1}(Y-M\theta) - \frac{1}{2}\lambda \sum_{j=1}^{N_p} |\theta_j|^q$

$q=2$ (RIDGE REGRESSION) ... $q=1$ (LASSO REGRESSION)

Non-linear Regression

- often we can make non-linear problems linear
  e.g $y = e^{\theta x} \Rightarrow \ln y = \theta x$
- but not always ... Levenberg-Marquardt (LM) algorithm uses GRADIENT DESCENT and GAUSS-NEWTON OPTIMIZATION

eg expand $f(x_i|\theta)$ as TAYLOR EXPANSION...

$$f(x_i|\theta) = f(x_i|\theta_0) + J d\theta$$

$$J = \left.\frac{\partial f(x_i|\theta)}{\partial \theta}\right|_{\theta_0}$$

$\Rightarrow$ LM minimizes sum of squares,

$$\sum_i [y_i - f(x_i|\theta_0) - J d\theta]^2$$

... for a perturbation $d\theta$, that is updated by

$$[J^T C^{-1} J + \lambda \, \mathrm{diag}(J^T C^{-1} J)] d\theta = J^T C^{-1} (Y - f(X|\theta))$$

λ: damping parameter

- λ small → Gauss-Newton
- λ big → gradient descent

- Gaussian Process Regression (GPR)

  ↳ GPR treats your data vector as a big random sample taken from a N-D Gaussian distribution.

  ⇒ by learning the covariance between points, GPR can interpolate ... (gives interpolation uncertainty too!)

  NOTE we are not drawing Gaussian shapes on the x-axis ... we model the data itself as a Gaussian draw.

- Total Least Squares ⇒ uncertainties in "x" and "y"