

ENGSCI 721

INVERSE PROBLEMS

Oliver Maclaren
oliver.maclaren@auckland.ac.nz

MODULE OVERVIEW

3. Preview of the statistical view of inverse problems

[1 lectures]

Bayesians, Frequentists and all that. Basic frequentist analysis.

MODULE OVERVIEW

Inverse Problems (*Oliver Maclaren*) [~8 lectures/2 tutorials]

1. Basic concepts [4 lectures]

Forward vs inverse problems. Well-posed vs ill-posed problems. Algebra and calculus of inverse problems (matrix calculus, generalised inverses etc). Regularisation and trade-offs.

2. More regularisation [3 lectures]

Higher-order Tikhonov regularisation, truncated singular value decompositions, iterative regularisation.

LECTURE 4: TRADE-OFFS AND REGULARISATION

Topics:

- Trade-offs between fit and complexity
- Relation to regularisation
- Tikhonov regularisation, L-curves etc

Eng Sci 721 : Lecture 4

Trade-offs, Constraints & Regularisation

→ We have seen that generalised* inverses give minimum norm least squares solutions

→ We have also seen that this solves existence & uniqueness issues, but does not solve stability issues.

What's going on?

How can we do better?

* really, pseudo inverses

Optimisation formulation revisited

From last time:

We can think of the generalised inverse as the result of a two-step optimisation problem:

1. Find the set of least squares solutions
2. Of these solutions, choose the minimum norm solution.

It turns out that it is better to simultaneously control

- these two objectives & trade-off
- fit to data
 - model 'complexity' (size)
-

Motivation: fitting a polynomial to data

$$\boxed{y = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n}$$

Observations:

$$\boxed{(x_0, y_0), (x_1, y_1), \dots, (x_m, y_m)}$$

leads to

$$y_0 = a_0 + a_1 x_0 + a_2 x_0^2 + \dots + a_n x_0^n$$

$$y_1 = a_0 + a_1 x_1 + a_2 x_1^2 + \dots + a_n x_1^n$$

:

$$y_m = a_0 + a_1 x_m + a_2 x_m^2 + \dots + a_n x_m^n$$

ie

$$\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & - & - & - & x_m^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix}$$

$$\boxed{y_{\text{obs}} = A_{\text{obs}} \theta} \quad \text{linear system}$$

→ linear in parameters!

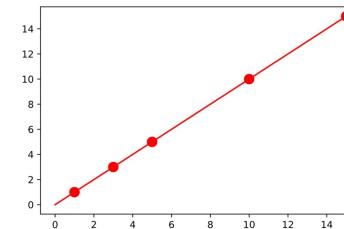
Underdetermined case & Instability

- Eg - high degree polynomial } possibly non-unique.
- plus observation noise } possibly inconsistent

$$\boxed{y = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n} \quad \text{class}$$

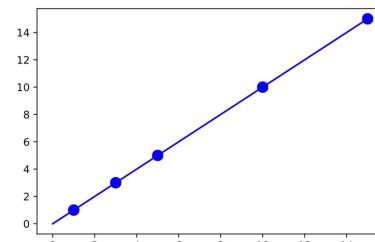
But true model: $a_i = 0 \text{ if } i \neq 1$
 $a_1 = 1$

'True' data: (5 observations)

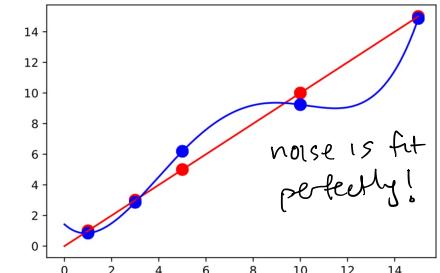


noise
free
recovery

add noise &
recover



noise is fit
perfectly!



unstable! Not reduced
enough!
→ fitting noise
(not reproducible!)

What we seem to want is something
that fits the data less well
→ Huh?

We are willing to trade-off
some fit to given data
for a 'simpler' or more stable model
↳ 'similar' data
→ similar model

statistical/predictive view:

- want to fit out-of-sample as well as in-sample data
 - ↳ training vs test sets
 - ↳ noise is not exactly reproducible, hence don't try to fit it exactly!
 - ↳ 'bias-variance trade-off'

Trade-offs & regularisation

- o Instead of two-step:

- Stage 1 : $\underset{x}{\text{minimise}} \quad \|y - Ax\| \quad \text{or} \quad \|y - Ax\|^2$

They

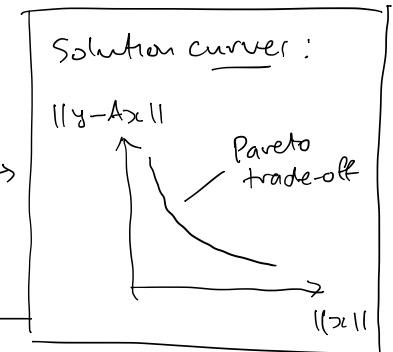
- o Stage 2: minimise $\|x\|$ or $\|x\|^2$ among all solutions to stage 1.

- o Try simultaneous minimisation!

vector/multi objective problem

$$\min_{\mathbf{x}} \quad \left(\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_1, \|\mathbf{x}(1)\| \right) \rightarrow$$

simultaneously



→ allows us to filter noise in underdetermined case, while still 'shrink' or reducing models to get 'simple' solutions

Stepwise revisited

We have seen

- 1. First fit data
- 2. Then reduce model

leads to the (possibly) unstable generalised inverse.

Suppose we instead tried

- 1. First reduce model
- 2. Then fit data.

→ suppose $\min\{\|x\|\} = 0$ & $x=0$ is the soln.

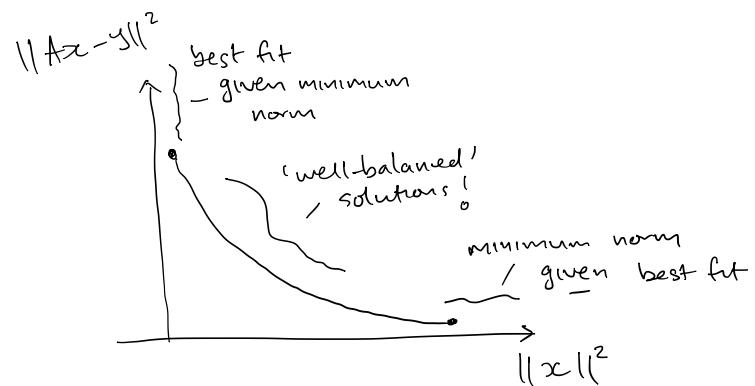
Then for $Ax = y$

→ data error is

$$\|y - Ax\| = \|y\|$$

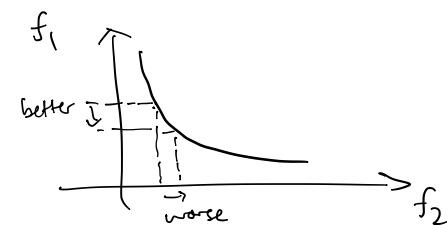
→ not very good either!

Trade-off curve : limits/end points



Pareto solutions : sets of solutions

- equally good, unless we prefer one objective to other
- can't improve one objective without making other worse.



Three (or more) equivalent versions

The bi-objective problem can be written as a standard single objective problem in at least 3 equivalent ways

→ The idea is to introduce an extra parameter that, one way or the other, represents the relative importance, or trade-off, for the two objectives

↳ Think: 'exchange rate' etc.

↳ By varying the exchange rate we 'trace out' all the solutions on the Pareto curve

↳ in particular we control the balance, rather than prioritise one over other as in 2-step.

Version 1: simple model, acceptable fit.

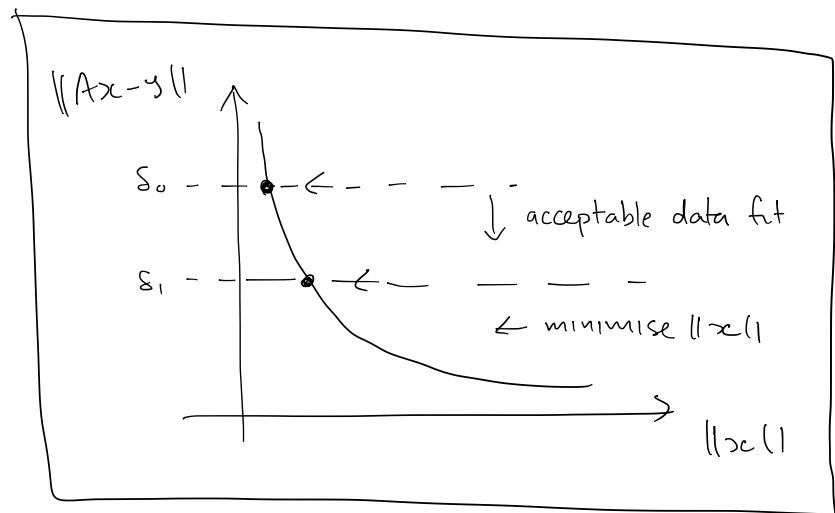
$$\text{la. } \begin{cases} \min \|x\| \\ \text{st. } \|Ax - y\| \leq \delta \end{cases}$$

$$\text{lb. } \begin{cases} \min \|x\|^2 \\ \text{st. } \|Ax - y\|^2 \leq \delta \end{cases}$$

smallest model that fits to within δ (not zero!)

etc.

Parameterised by δ , ie 'acceptable data fit' level:



Version 2: acceptable model, best fit

$$2a. \begin{cases} \min \|Ax - y\| \\ \text{s.t. } \|x\| \leq \epsilon \end{cases}$$

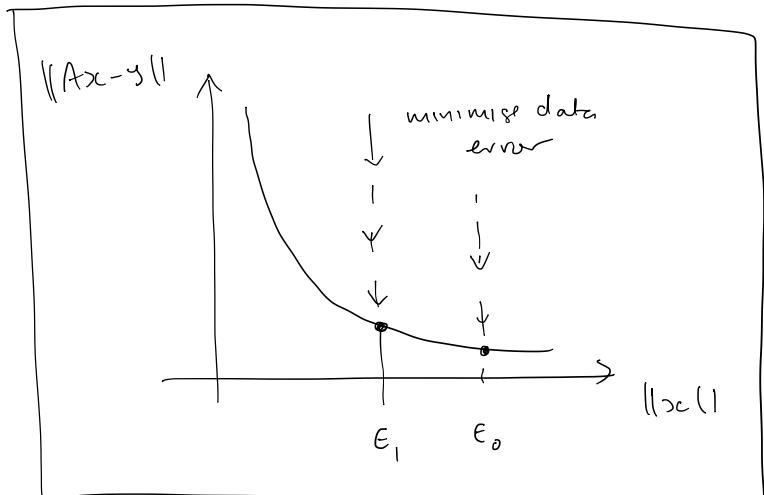
best fit within
acceptable
model limits
(not too
'complex!')

$$2b. \begin{cases} \min \|Ax - y\|^2 \\ \text{s.t. } \|x\|^2 \leq \epsilon \end{cases}$$

etc.

Parameterised by ϵ ,

i.e. 'allowable model complexity' :



Version 3 : weighted sum ('Scalarised' version)

$$3a. \boxed{\min \|Ax - y\| + \lambda \|x\|}$$

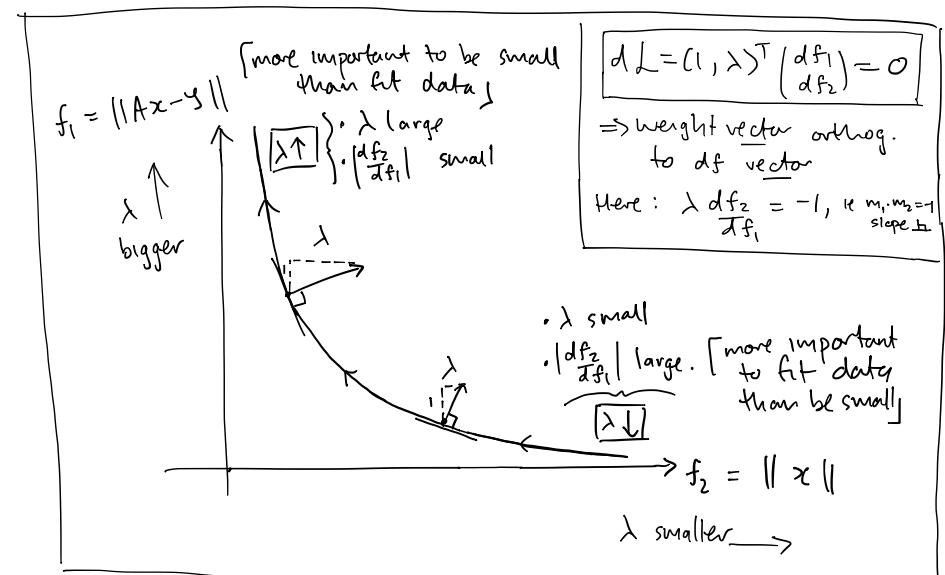
$$3b. \boxed{\min \|Ax - y\|^2 + \lambda \|x\|^2}$$

(or $\min \mathbf{f}^T \mathbf{f}$ for vector of weights \mathbf{f} & vector obj.)

3b': $\left\{ \begin{array}{l} \text{"Tikhonov" regularisation} \\ \text{Damped least squares} \\ \text{Ridge regression} \end{array} \right\}$ etc!

Parameterised by λ , 'trade-off weight' or 'relative importance'

'Lagrangian' : $L = f_1 + \lambda f_2 \quad \left\{ \begin{array}{l} dL = 0 \text{ on Pareto curve} \\ \text{(see "Lagrange multipliers etc")} \end{array} \right\}$ ('equally good')



Analytical solution (linear only)

The form $\boxed{\|Ax - y\|^2 + \lambda \|x\|^2}$ is particularly useful for getting an analytical solⁿ

$$\text{Note : } \|r\|^2 = \|r_1\|^2 + \|r_2\|^2$$

so define the augmented vector

$$\tilde{r} = \begin{bmatrix} Ax - y \\ \sqrt{\lambda} x \end{bmatrix} \Rightarrow \|\tilde{r}\|^2 = \|Ax - y\|^2 + \lambda \|x\|^2$$

This can be written as

$$\tilde{r} = \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x - \begin{bmatrix} y \\ 0 \end{bmatrix}$$

$$= \hat{A}x - \tilde{y} \quad \text{ie augmented residual equation}$$

→ can solve via ordinary least squares!

We want to solve the augmented least squares problem:

$$\boxed{\min \|\tilde{r}\|^2 = \min \|\tilde{A}x - \tilde{y}\|^2}$$

$$\left(\text{ie } \min \left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x - \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|^2 \right)$$

→ Can we?

Note: $I \in \mathbb{R}^n$ means
 I is $n \times n$ identity
 A is $m \times n$, where $m < n$ (under).

$$\text{so } \hat{A} = \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} = \left[\begin{array}{c|c} m \times n & \\ \hline n \times n & \end{array} \right] \quad \left. \right\} m+n > n$$

⇒ now have tall, over-determined system.

Least squares solution exists if

- columns are LI
 - $\tilde{A}^T \tilde{A}$ is invertible
 - etc
- } eqn's.

Key:

If turns out that the columns of \tilde{A} are LI & $\tilde{A}^T \tilde{A}$ is invertible,
regardless of A, as long as $\lambda > 0$!

→ (we'll prove in tutorial / assignment)

So the normal equations

$$\tilde{A}^T \tilde{A} x = \tilde{A}^T \tilde{y} \text{ are always solvable}$$

for

$$\begin{aligned} x &= \tilde{A}^+ \tilde{y} \\ &= (\tilde{A}^T \tilde{A})^{-1} \tilde{A}^T \tilde{y} \end{aligned}$$



Regularised normal equations

Note:
— $\tilde{A}^T \tilde{A} =$

$$[A^T \sqrt{\lambda} I] \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} = (A^T A + \lambda I)$$

$$\begin{aligned} &\& \tilde{A}^T \tilde{y} = [A^T \sqrt{\lambda} I] \begin{bmatrix} y \\ 0 \end{bmatrix} \\ &= A^T y \end{aligned}$$

so the normal equations become

$$(A^T A + \lambda I) x = A^T y$$

↑
only extra term.

Explains why works:

$A^T A$ is positive semi-definite } almost invertible

λI is positive definite, $\lambda > 0$

add together gives positive definite
→ invertible!

Limiting cases

Consider the 'Tikhonov' inverse in

$$x = (A^T A + \lambda I)^{-1} A^T y$$

i.e. $\boxed{x = A^* y}$ where $\boxed{A^* = (A^T A + \lambda I)^{-1} A^T}$

Cases.

1. As $\lambda \rightarrow 0$, keeping $\lambda > 0$

$$A^* \rightarrow (A^T A)^{-1} A^T = A^+$$

i.e. unregularised limit is generalised inverse!

2. As $\lambda \rightarrow \infty$

$$A^* \rightarrow \frac{1}{\lambda} A^T \rightarrow 0$$

$$\text{so } x = \underline{A^* y} \rightarrow 0$$

$\|Ax - y\|$ \nearrow large limit: $x \rightarrow 0$

\searrow small limit: $A^* \rightarrow A^+$

Tikhonov regularisation

The Tikhonov form is

$$\boxed{\min \|Ax - y\|^2 + \lambda \|x\|^2}$$

\rightarrow zeroth order.

More general Tikhonov

$$\boxed{\min \|Ax - y\|^2 + \lambda \|Cx - b\|^2}$$

where C : weighting matrix
 or
 differential operator
 or
 etc. } "prior information or constraints on models"

Also: ✓ other norms for data &
 or model space (can mix
 & match)

✓ nonlinear models $A(x)$ i.e.
 $\|A(x) - y\|^2 + \lambda \|x\|^2$ etc

Note: regularisation is essentially
the same as making up
(or incorporating) additional
data!

Amounts to $\tilde{A}x = \tilde{y}$,

where:

$$\tilde{A} = \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}, \quad \tilde{y} = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

i.e. combining: $Ax = y$ } measured data

& $\sqrt{\lambda} I x = 0$ } 'made up' data.
 'additional data model'
 → direct measurement of x

{ ideally,
 influences
 → 0 as
 real data
 increases

Statistical interpretation (see later): 'mixed'
 ↴ estimation

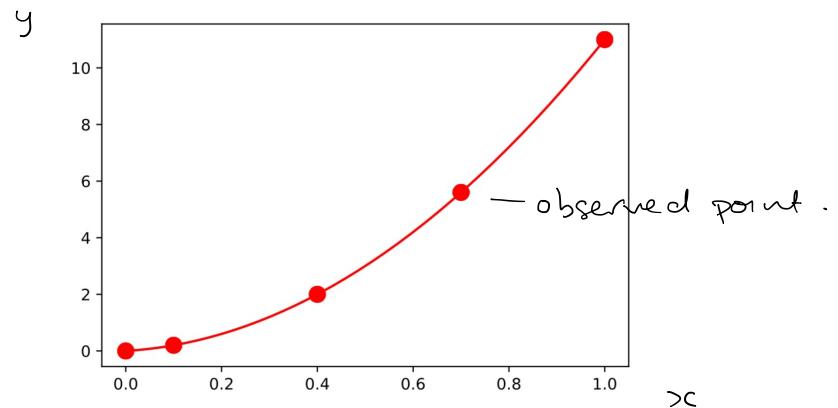
- For frequentist: synthesising data sets }
 (eg product of two likelihoods)
- For Bayesians: combine likelihood & prior.

Important: both 'schools' can do!

Tikhonov Example : polynomial fitting

$$y = xc + 10x^2 \quad \} \text{True model.}$$

5 observations:



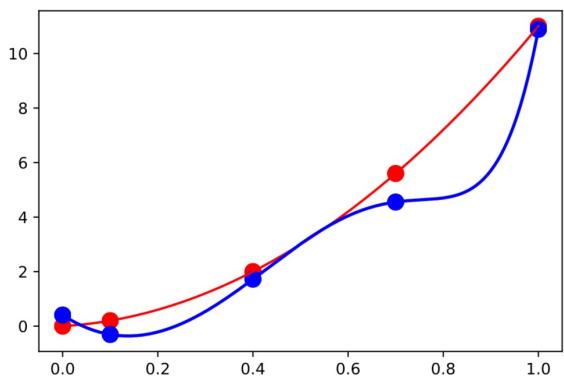
Goal: fit the model

$$\boxed{y = a_0 + a_1 x + \dots + a_q x^q}$$

parameters \rightarrow observations !

→ will fit any noise too

Tikhonov Example : polynomial fitting



Red: true

Blue: noisy

Note: exactly fitting noise

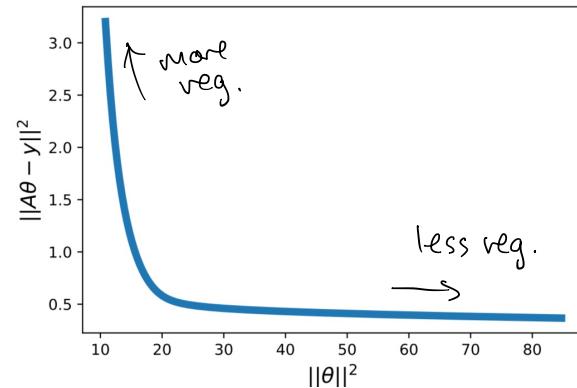
- will get very different solution if observed again
- ↳ fit new noise etc
- 'unstable'

Solution: consider regularisation!

Trade-off curve

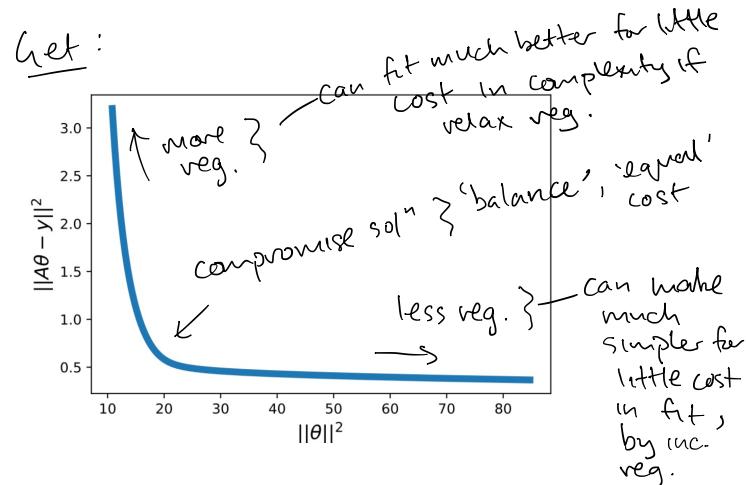
- generate a grid of λ values
(typically log-spaced, since 'asymptotic')
- Solve using ordinary least squares
for augmented system $\tilde{A}\theta = \tilde{y}$
→ I just used $\theta = \tilde{A}^+ \tilde{y}$ via
numpy's pseudo-inverse
- For nonlinear, need to solve
with iterative etc (see later)

Get:



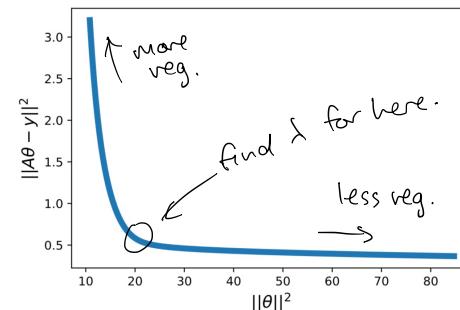
'Choosing' regularisation parameters

- The entire Pareto curve represents more info than any one solution
→ family of solutions as $f(\lambda)$
- But often want one 'good' solution.
→ Need to choose a single λ for these cases.
→ Simple idea: choose good 'compromise' soln:



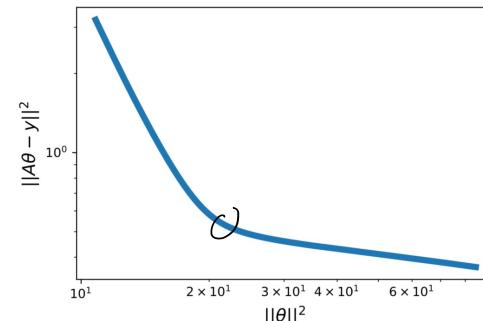
'Choosing' regularisation parameters

Compromise



On log-log plot, called 'L-curve'

since often looks (more) L-shaped
→ find 'elbow':



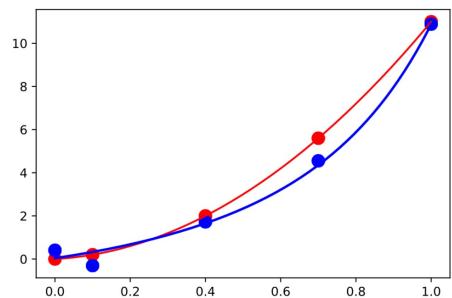
(There isn't much more L-shaped ...)

→ see Aster et al or Hansen for better ex.)

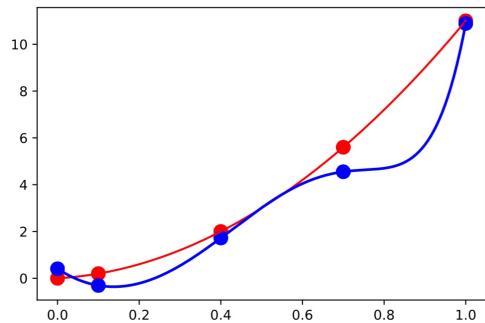
Example regularised solⁿ

(λ chosen from curve 'by eye' here!)

Regularised



Compare to unregularised:



Note: in both cases we are fitting
— an order 9 polynomial to
5 data points!