ENGSCI 721

INVERSE PROBLEMS

Oliver Maclaren
oliver.maclaren@auckland.ac.nz

MODULE OVERVIEW

Inverse Problems (*Oliver Maclaren*) [~8 lectures/2-3 tutorials]

1. *Basic concepts* [3 lectures]
   Forward vs inverse problems. Well-posed vs ill-posed problems. Algebra of inverse problems (generalised inverses etc). Regularisation and trade-offs.
2. *More regularisation* [3 lectures]
   Higher-order Tikhonov regularisation, truncated singular value decompositions, iterative regularisation.

MODULE OVERVIEW

3. *Statistical view of inverse problems I* [2 lectures]
   Bayesians, Frequentists and all that. Basic frequentist analysis. Linearisation and covariance propagation.

LECTURE 3: TRADE-OFFS AND REGULARISATION

Topics:

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

Eng Sci 721 : Lecture 3

Trade-offs, Constraints & Regularisation

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

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

What's going on?

How can we do better?

---

* really, psendo 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

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

Observations:

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

Leads to

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

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

$$\vdots$$

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

ie

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

ie $y_{obs} = A_{obs}\,\theta$ linear system

→ linear in parameters !

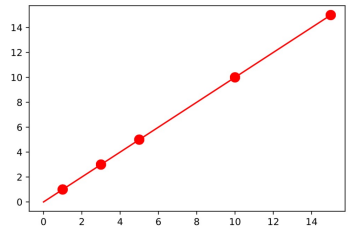
Underdetermined case & instability

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

$$y = a_0 + a_1 x + a_2 x + \dots a_n x^n$$ class

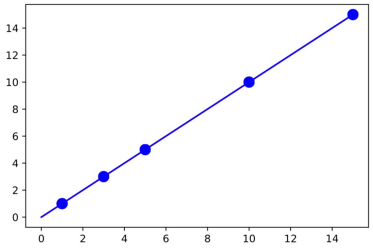
But true model: $a_i = 0$ if $i \neq 1$, $a_1 = 1$

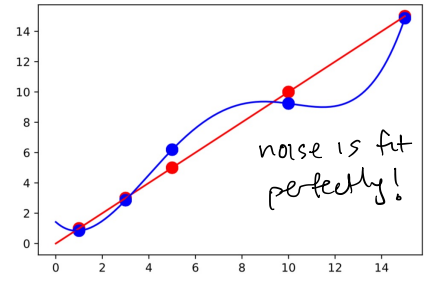
'True' data: (5 observations)



noise free recovery

add noise & recover





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

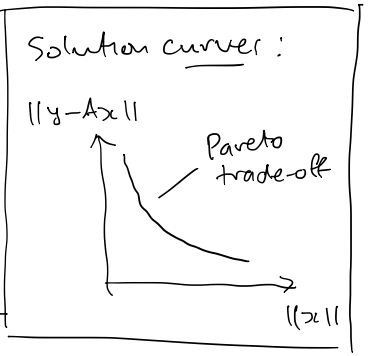
- Instead of two-step:
  - Stage 1: minimise $\|y - Ax\|$ or $\|y - Ax\|^2$ x
  - Then
  - Stage 2: minimise $\|x\|$ or $\|x\|^2$ among all solutions to stage 1.

- Try simultaneous minimisation:

vector/multi objective problem

$$\min_x \left(\|y - Ax\|, \|x(1)\| \right) \rightarrow \text{simultaneously}$$

Solution curves:



→ allows us to filter noise in underdetermined case, while still 'shrinking' 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

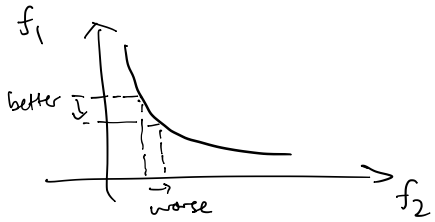
$$\|y - Ax\| = \|y\|$$

→ 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.

1a. $$\min \|x\| \quad \text{s.t. } \|Ax - y\| \leq \delta$$

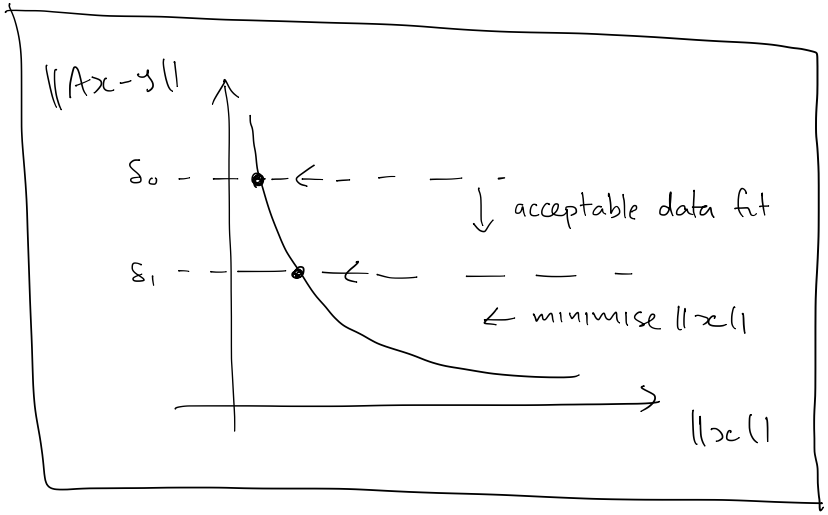
or

1b. $$\min \|x\|^2 \quad \text{s.t. } \|Ax - y\|^2 \leq \delta$$

etc.

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

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



Version 2 : acceptable model, best fit

2a. $\min \|Ax - y\|$ s.t. $\|x\| \leq \epsilon$

or

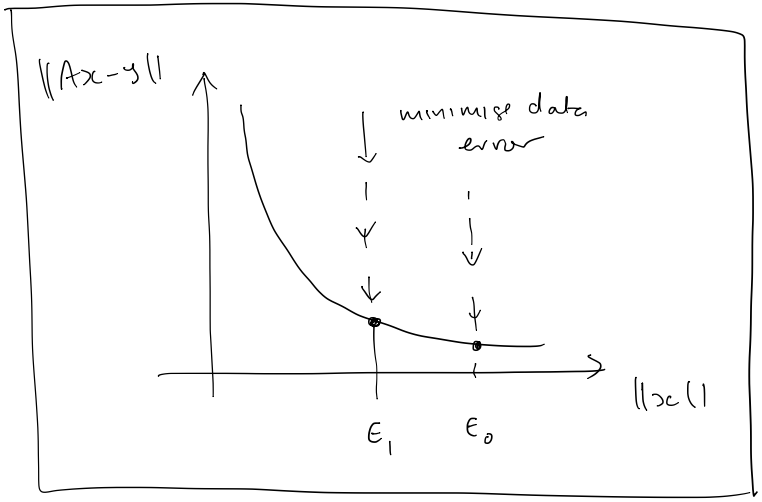
2b. $\min \|Ax - y\|^2$ s.t. $\|x\|^2 \leq \epsilon$

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

etc.

Parameterised by ϵ,

ie 'allowable model complexity' :



Version 3 : weighted sum ('scalarised' version)

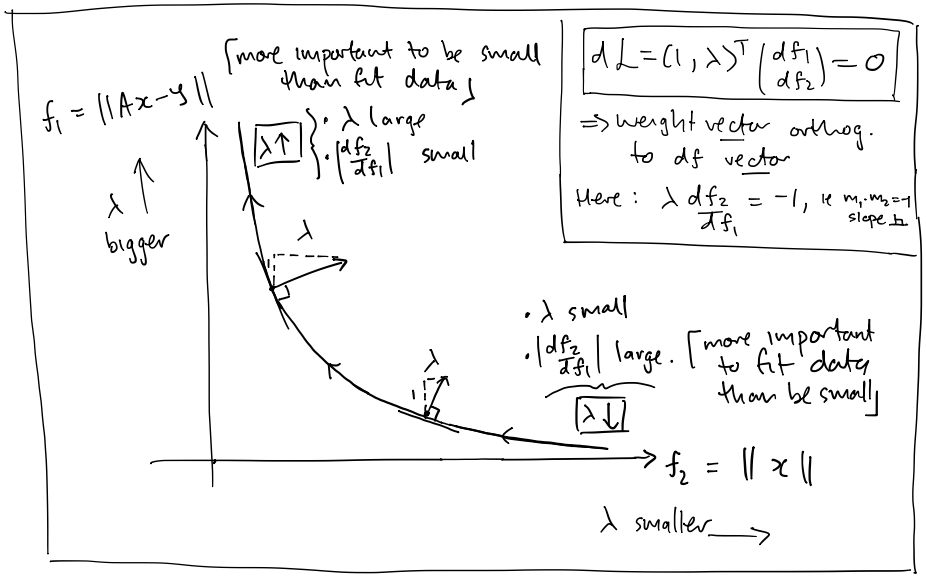
3a. $\min \|Ax - y\| + \lambda \|x\|$, $\lambda > 0$

3b. $\min \|Ax - y\|^2 + \lambda \|x\|^2$ (or $\min \lambda^T f$ for vector of weights & f vector obj.)

3b': "Tikhonov" regularisation, Damped least squares, Ridge regression } etc !

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

'Lagrangian' : $L = f_1 + \lambda f_2$ (see 'Lagrange multipliers etc') } $dL = 0$ on Pareto curve ('equally good')



Analytical solution (linear only)

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

---

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

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

$$= \tilde{A} x - \tilde{y}$$, ie augmented residual equation

→ can solve via ordinary least squares!

We want to solve the augmented least squares problem:

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

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

→ Can we?

Note : Ix & $x \in \mathbb{R}^n$ means

I is $n \times n$ identity

A is $m \times n$, where $m < n$ (under-s.

so $\tilde{A} = \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} = \begin{bmatrix} m \times n \\ \\ n \times n \\ \\ \underbrace{\qquad}_{n} \end{bmatrix} \Big\} m+n > n$

⇒ now have tall, over-determined system.

Least squares solution exists if

- columns are LI
- $\tilde{A}^T\tilde{A}$ is invertible

} equiv.

etc

Key:

It 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

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

for

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

→

Regularised normal equations

Note:

$$\tilde{A}^T\tilde{A} =$$

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

&

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

$$= A^Ty$$

So the normal equations become

$$(A^TA + \lambda I)x = A^Ty$$

↑ only extra term.

Explains why works:

A^TA 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

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

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

Cases.

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

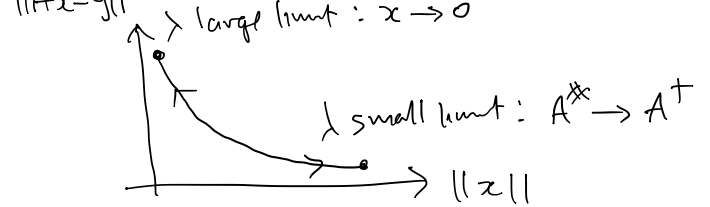
$$A^{\#} \to (A^T A)^{-1} A^T = A^{\dagger}$$

ie unregularised limit is generalised inverse!

2. As $\lambda \to \infty$

$$A^{\#} \to \frac{1}{\lambda} A^T \to 0$$

so $x = A^{\#} y \to 0$



Tikhonov regularisation

The Tikhonov form is

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

→ 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 $A(x)$ ie $\|A(x) - y\|^2 + \lambda \|x\|^2$ etc

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

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

where:

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

ie combining: $Ax = y$ } measured data

& $\sqrt{\lambda} I x = 0$ } 'made up' data.

'additional data model' → direct measurement of x } ideally, influence → 0 as real data increases

---

Statistical interpretation (see later): 'mixed estimation'

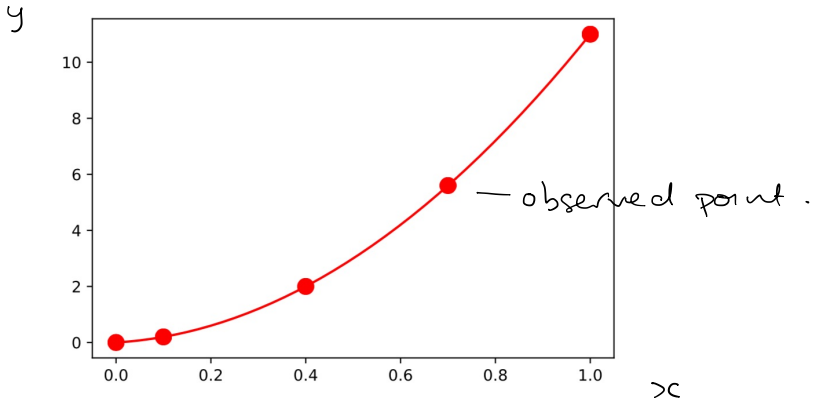
- For frequentist: synthesising data sets (eg product of two likelihoods)
- For Bayesians: combine likelihood & prior.

Important: both 'schools' can do!

Tikhonov Example: polynomial fitting

$y = x + 10x^2$ } True model.

5 observations:



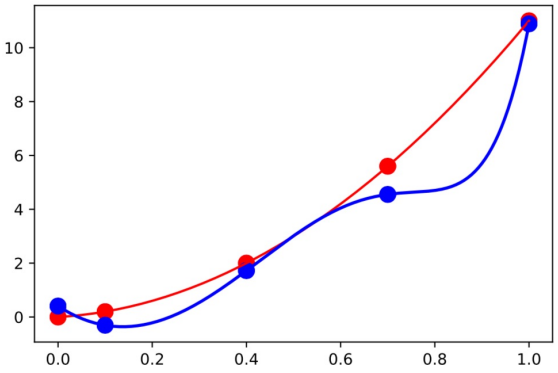
Goal: fit the model

$$y = a_0 + a_1 x + \cdots a_9 x^9$$

parameters > 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!

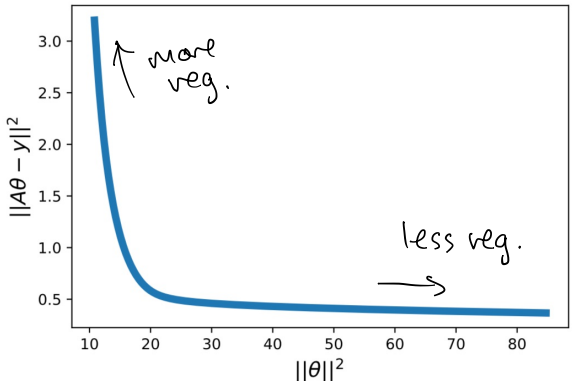
Trade-off curve

- 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:



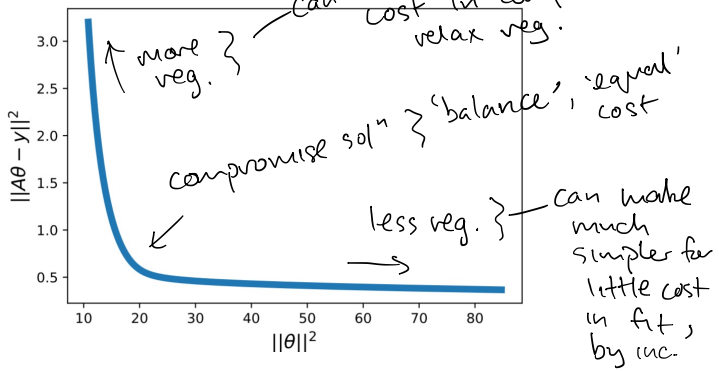
'Choosing' regularisation parameters

- 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.

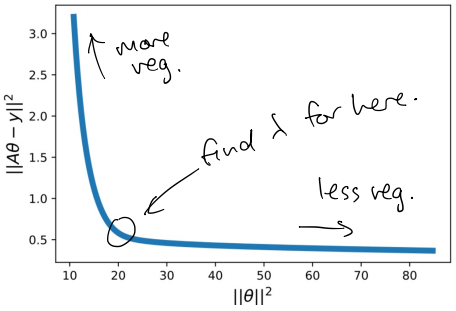
→ Simple idea: choose good 'compromise' soln:

Get:



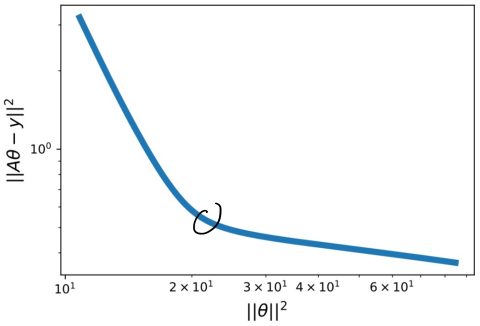
'Choosing' regularisation parameters

Compromise



On log-log plot, called 'L-curve' since often looks (more) L-shaped

↳ find 'elbow':



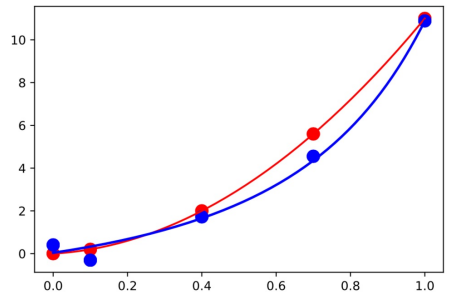
(Here isn't much more L-shaped ....)

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

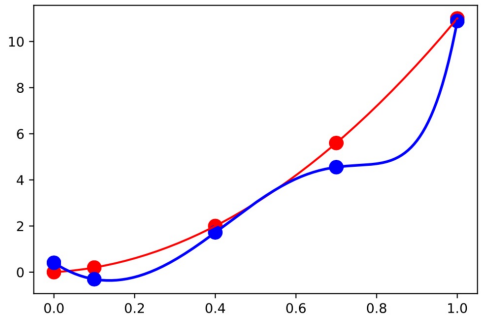
Example regularised soln

(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!