$$\max_{\theta \in \Theta} p(\mathrm{data}|\theta)\, p(\theta)$$

Bias, Variance, Regularization

› Marco Loog

Past

› Last week, part 1 : focus on linear least squares

Through conditional log-likelihood $$\prod_{i=1}^{N} N(y_i | w^T x_i, \sigma^2)$$

Direct minimization of squared loss $$\sum_{i=1}^{N} (w^T x_i - y_i)^2 = \|Xw - Y\|^2$$

Also MAP estimation, nonlinear extensions,…

Present

- Important additional ingredient : regularization

- Generally, important concept in learning
  - Here exemplified through regression
  - "Simplest case" : L_2
  - Sparsity inducing regularizer

- Bias-variance tradeoff
  - Also within context of least squares regression

Many Dimensions
Few Observations

- What happens?

E.g. assume true covariance of data is I
and consider $\widehat{w} = (X^T X)^{-1} X^T Y = \left(\frac{1}{N} X^T X\right)^{-1} \left(\frac{1}{N} X^T Y\right)$

Eigenvalues of [identity] covariance matrix?
Effect of this on $(X^T X)^{-1}$ and, therefore, $\widehat{w}$?
Do experiments if you do not see or believe…

Some Matlab?

Many Dimensions
Few Observations

› Solution $\widehat{w} = (X^T X)^{-1} X^T Y$ is unstable

Can be all over the place

› Generalization to unseen data can, and will often, be very bad

› How to stabilize the solution? Any ideas?

Stabilization, One Way to Perform

› Here's an idea : keep eigenvalues away from 0

› Add identity to $X^T X$: $\widehat{w} = (X^T X + \lambda I)^{-1} X^T Y$

› Why consider the identity?

Stabilization as Regularization

- Add identity to $X^T X$: $\widehat{w} = (X^T X + \lambda I)^{-1} X^T Y$
- This estimate is, in fact, solution of

$$\min_w \sum_{i=1}^{N} (x_i^T w - y_i)^2 + \lambda ||w||^2$$

Where did we see a very similar solution?

More Matlab?

An Equivalent View

› Instead of solving

$$\min_{w} \sum_{i=1}^{N} (f(x_i, w) - y_i)^2 + \lambda \|w\|^2$$

one can also solve

$$\min_{w} \sum_{i=1}^{N} (f(x_i, w) - y_i)^2$$

$$\text{s.t. } \|w\|^2 \leq \tau$$

Intermezzo?

- Shape of these functions? How do contours look?

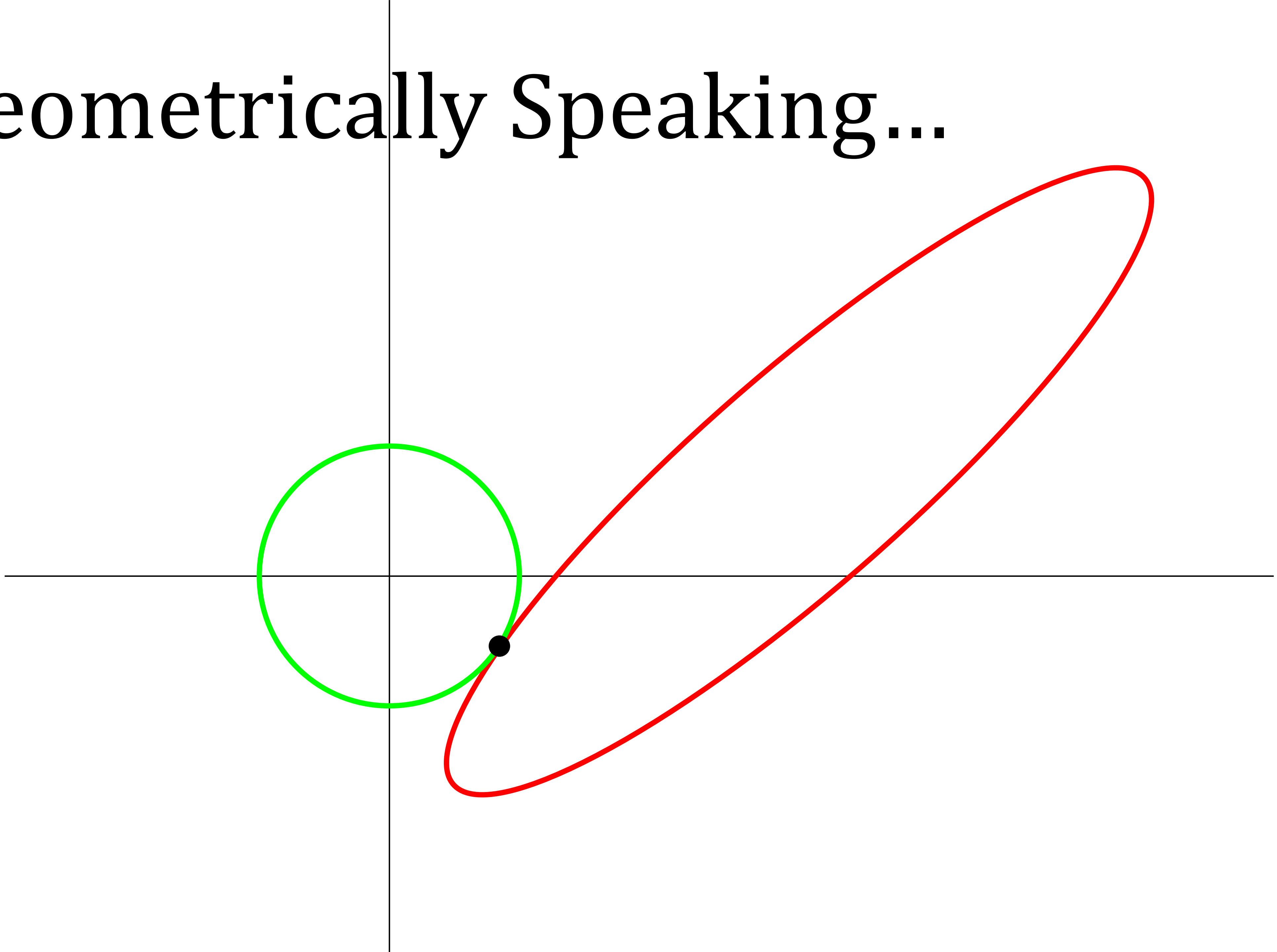
$$\sum_{i=1}^{N}(x_i^T w - y_i)^2 + \lambda\|w\|^2$$

$$\sum_{i=1}^{N}(x_i^T w - y_i)^2$$

$$\|w\|^2$$

Contours?

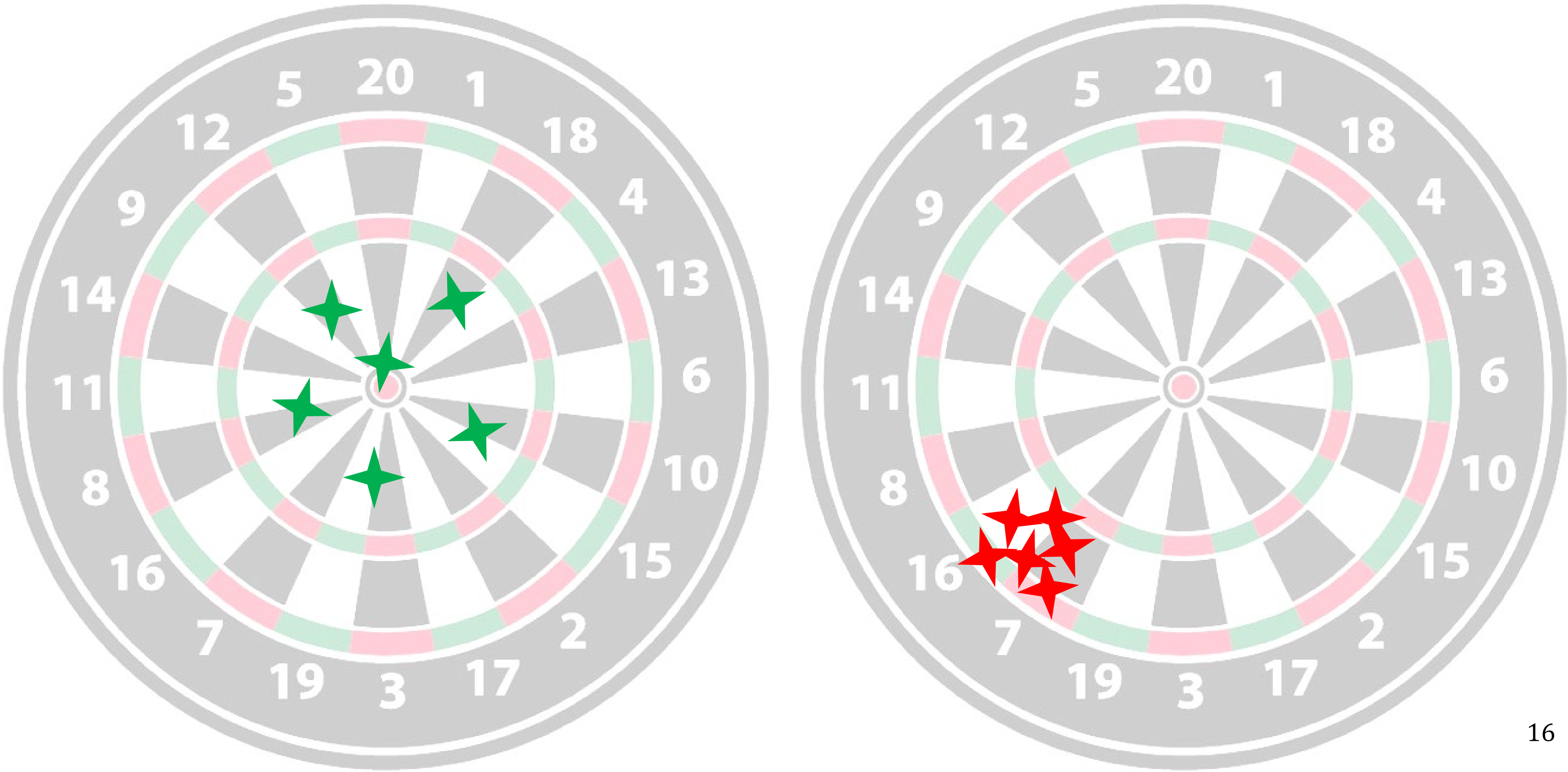
Geometrically Speaking…



Time to Discuss Bias and Variance

Bias and Variance…

› What is what?



Bias-Variance Decomposition

› Assume optimal prediction $f^*(x)$ at x

› Consider error for some estimate $\hat{f}(x)$

Depends on training data

› Consider expected error over different data sets

[Yes, I dropped the x]

$$\mathop{\mathbb{E}}_{\text{data}}\left[\left(f^* - \hat{f}\right)^2\right]$$

Write out…

$$\mathop{\mathbb{E}}_{\text{data}}\left[\left(f^* - \hat{f}\right)^2\right]$$

- Decompose…

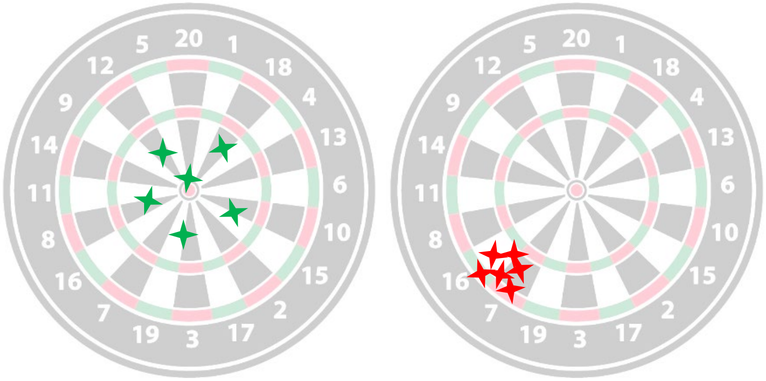
Bias-Variance Decomposition

$$\mathbb{E}\left[\left(f^* - \hat{f}\right)^2\right]$$
$$= \mathbb{E}\left[\left(f^* - \mathbb{E}\hat{f}\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\hat{f} - \hat{f}\right)^2\right]$$
$$= \text{bias}^2 + \text{variance}$$

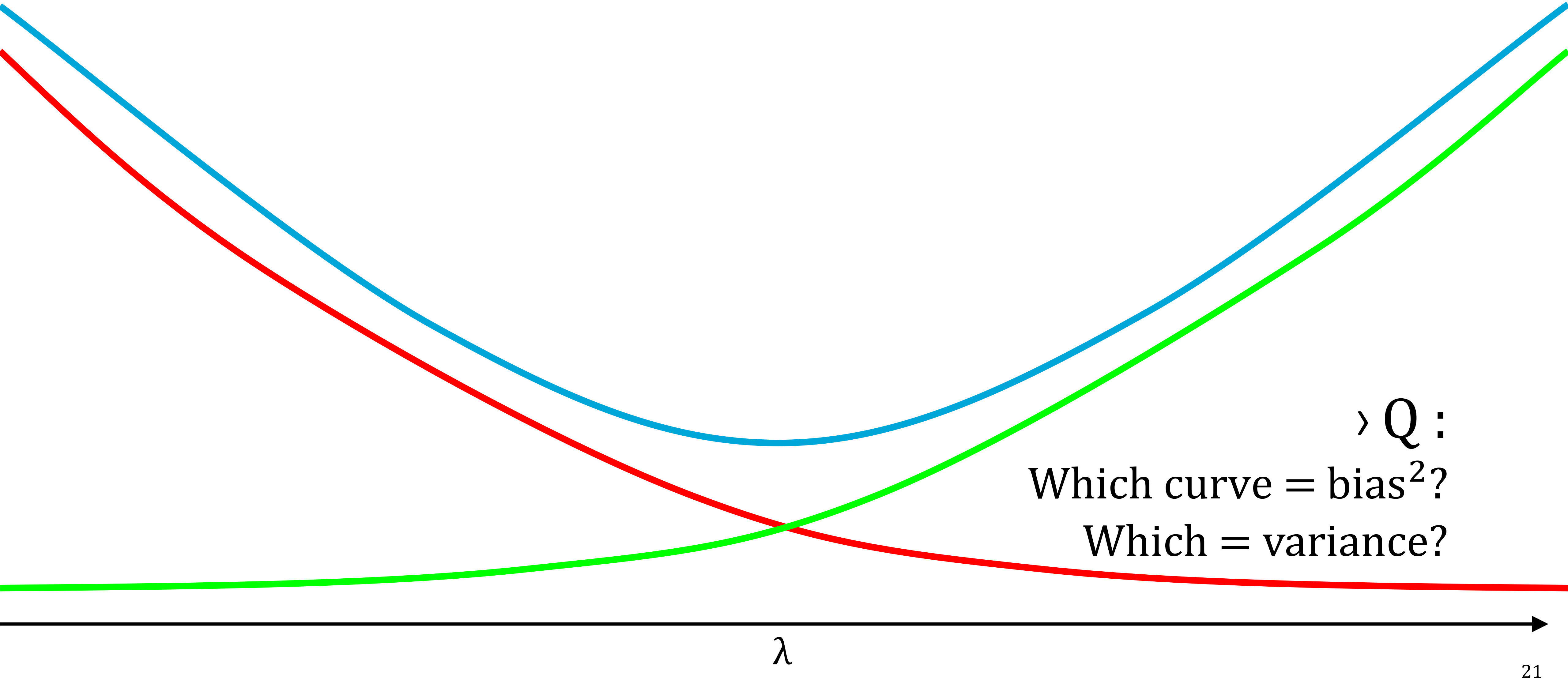
› What does the decomposition tell us?

The Tradeoff

› So, how can we control bias and variance?



Expected Loss and Bias-Variance



› Q :

Which curve = bias^2?

Which = variance?

Regularized Risk

› General approach to regularization

$$\min_{w} \sum_{i=1}^{N} \ell(f(x_i, w), y_i) + R(f)$$

Extension of our "general framework"

Different considerations give different R

Various links : MAP, MDL, SRM, etc.

Introducing Sparsity

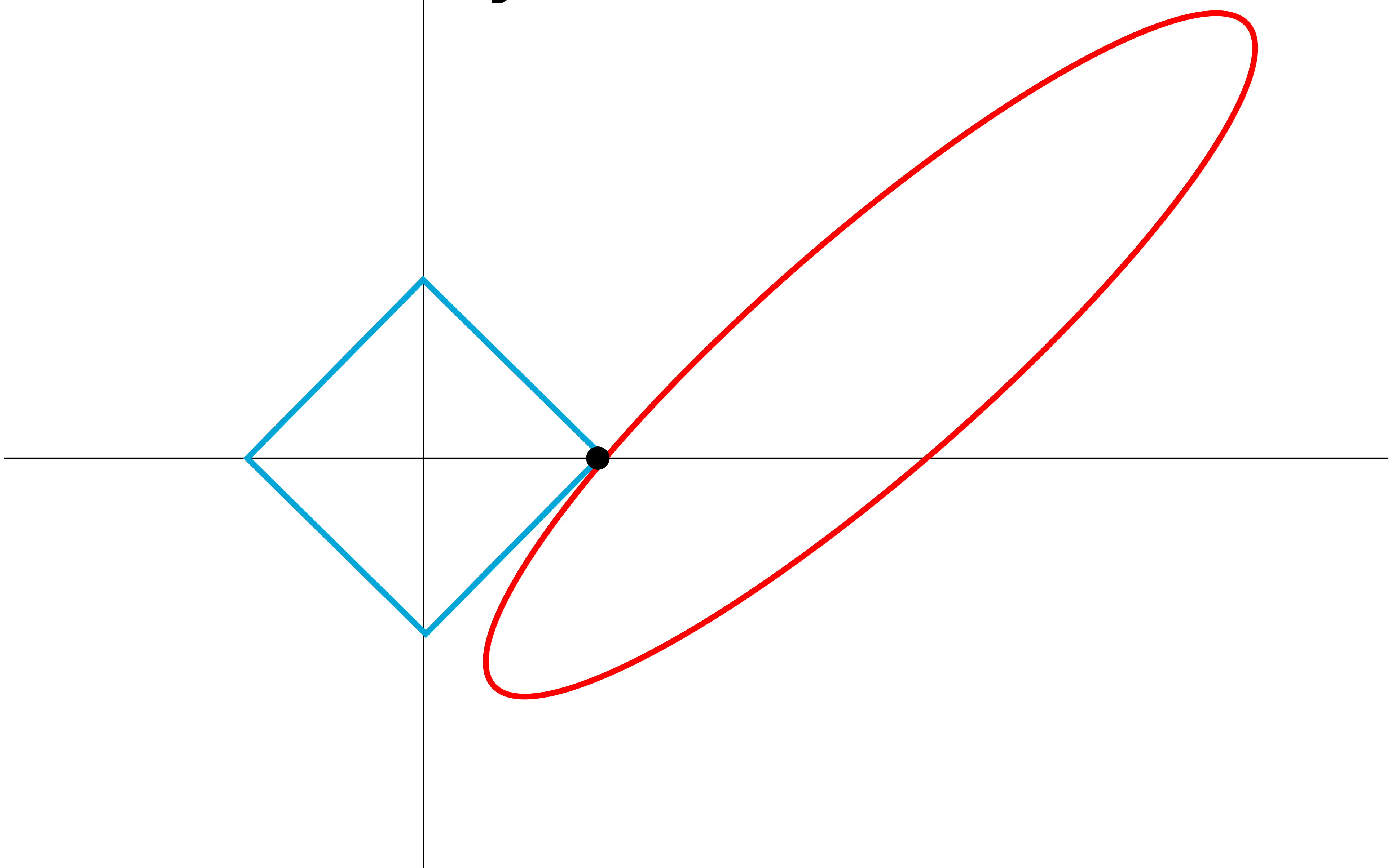
- For a change, let us consider

$$\min_{w} \sum_{i=1}^{N} (f(x_i, w) - y_i)^2$$

$$\text{s.t.} \ \ \|w\|_1 \ \leq \tau$$

What is the shape of $\|w\|_1$? Contours?
What is the effect of this change of norm?

The Geometry…



Again the Equivalent View…

› Include sparsifying norm as an additive term

$$\min_{w} \sum_{i=1}^{N} (f(x_i, w) - y_i)^2 + \lambda \|w\|_1$$

› Matlab “demo” [time permits…]

Final Remarks

- Sparsity by regularization due to Tibshirani
  - Least absolute shrinkage and selection operator or lasso
  - Also performs feature selection [week 6]
- Regularization framework also for classification…
- How to set λ / τ?
- Bias-variance returns next week
  - And at many other points in your life…
- Note the pseudo-inverse [e.g. Exercise 3.5]

› Q?