$$\max_{\theta \epsilon \Theta} p(\text{data}|\theta)p(\theta)$$

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 average x_i is 0
and consider $\widehat{w} = (XX^T)^{-1}XY^T = \left(\frac{1}{N}XX^T\right)^{-1}(\frac{1}{N}XY^T)$

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

Some Matlab…

Many Dimensions
Few Observations

- Solution $\widehat{w} = (XX^T)^{-1}XY^T$ 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 XX^T : $\hat{w} = (XX^T + \lambda I)^{-1} XY^T$
- Why consider the identity?

Stabilization as Regularization

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

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

Where did we see a very similar solution?

More Matlab?

An Equivalent View

› Instead of solving

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

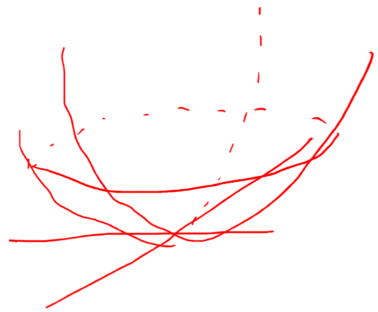
one can also solve

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

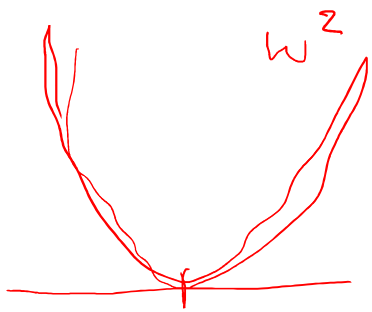
$$\text{s.t.} \ \|w\|^2 \leq \tau$$

Intermezzo?

- Shape of these functions? How do contours look?



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

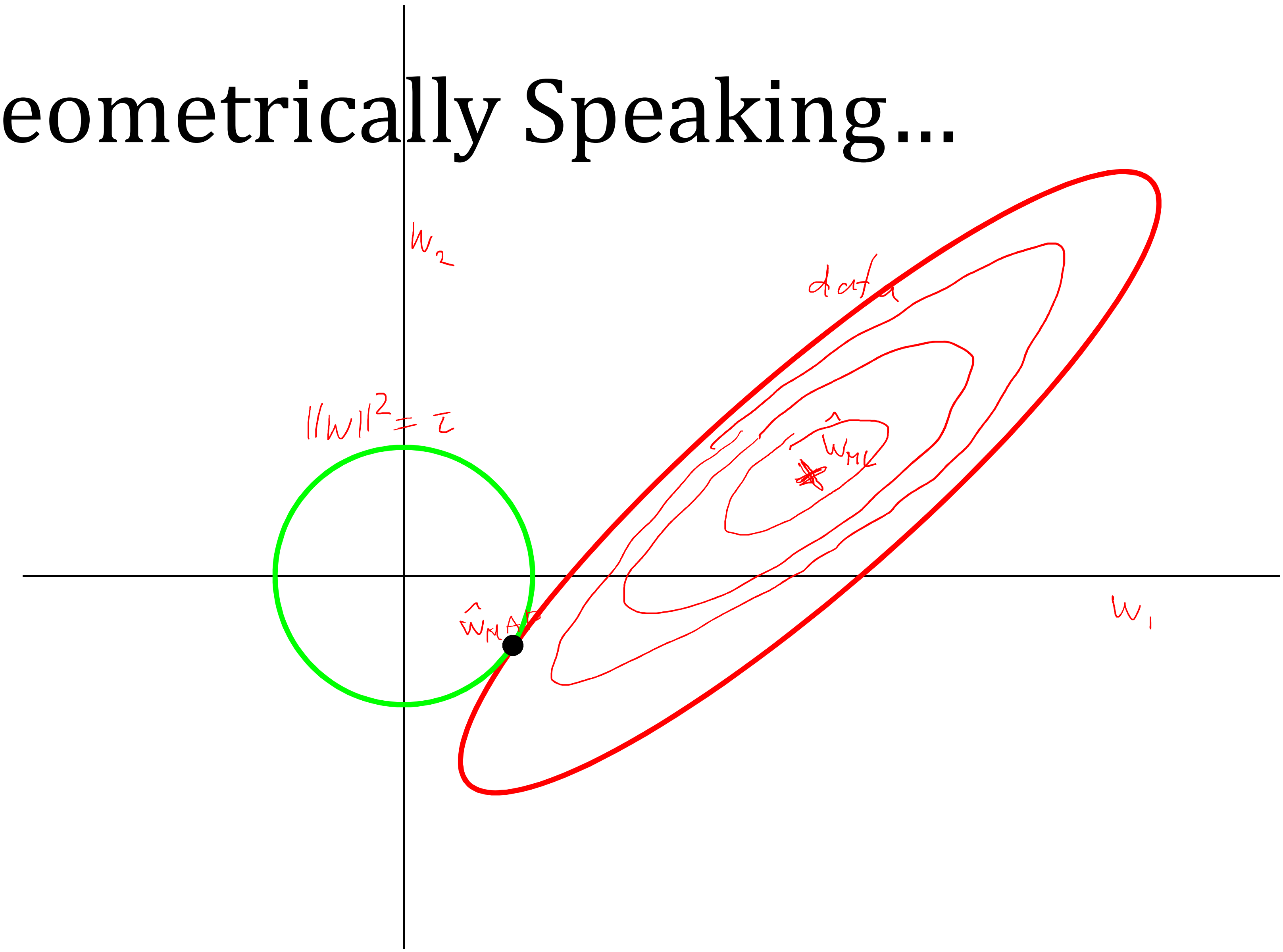


$$\sum_{i=1}^{N}(f(x_i, w) - y_i)^2$$

$$\|w\|^2$$

Contours?

Geometrically Speaking…



Time to Discuss Bias and Variance

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…

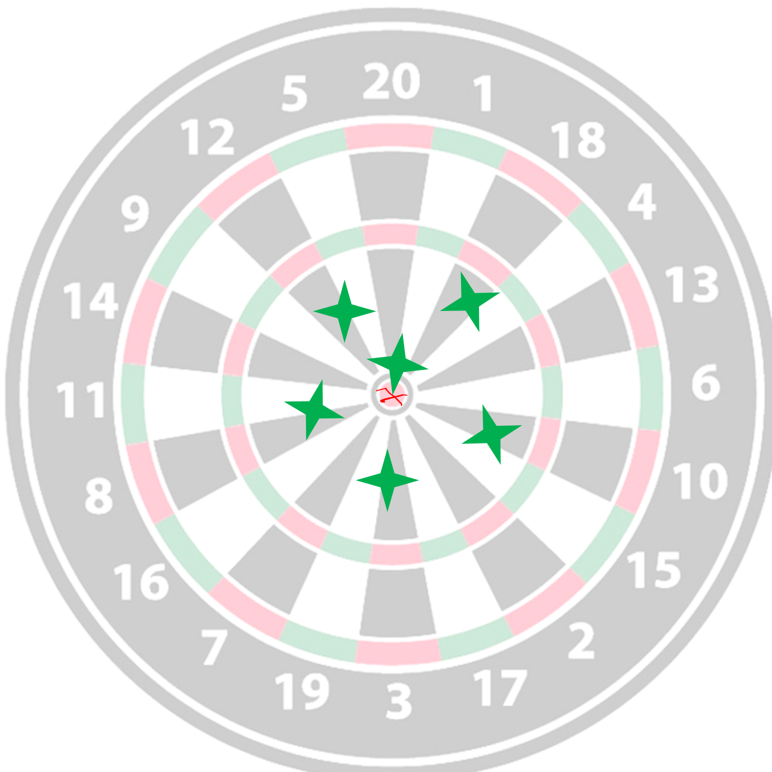
$$= \mathbb{E}\left[\left(f^* - \mathbb{E}\hat{f} + \mathbb{E}\hat{f} - \hat{f}\right)^2\right]$$

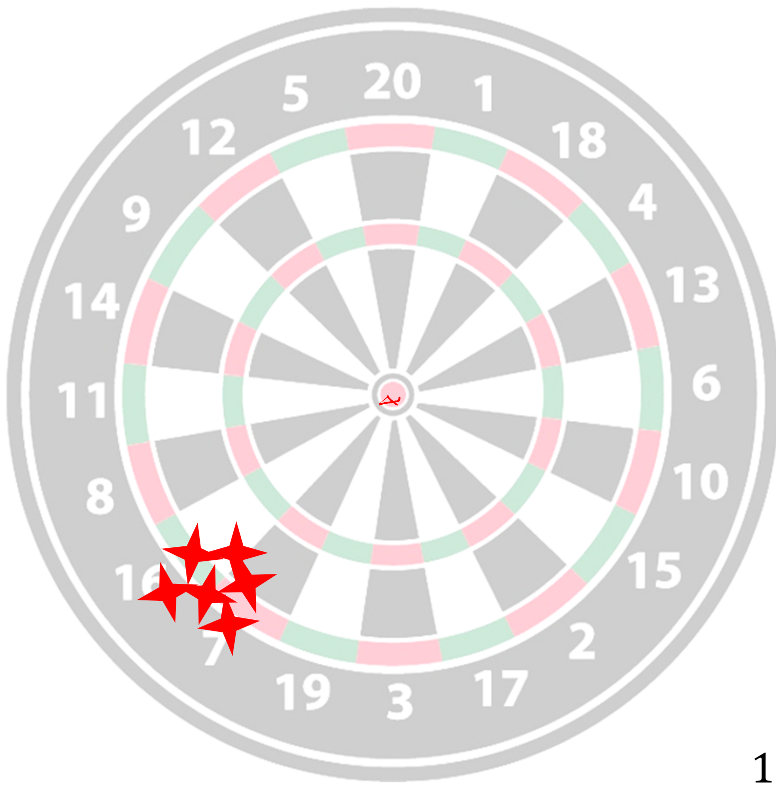
Bias-Variance Decomposition

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

› What does the decomposition tell us?

Bias and Variance…

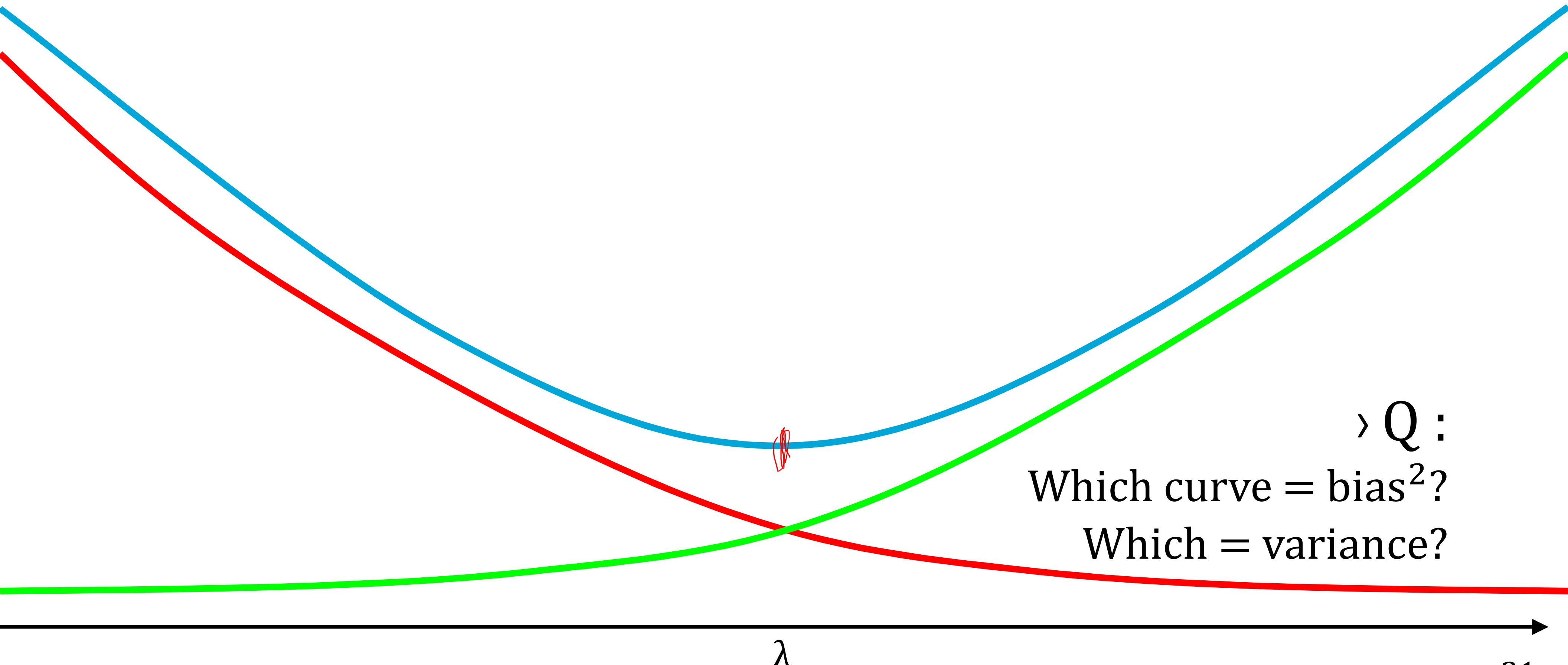




The Tradeoff

› 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

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

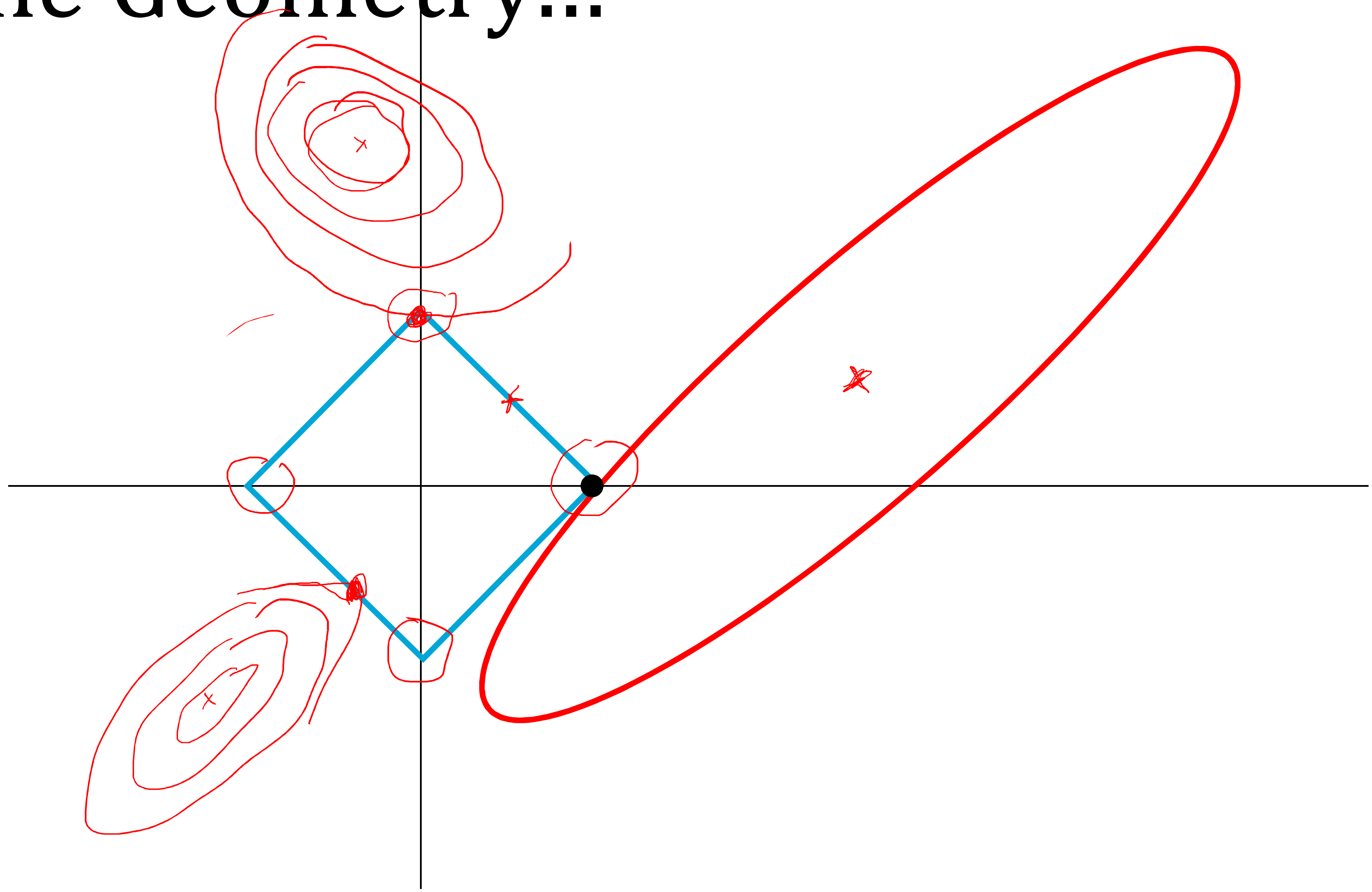
$\|w\|_1 =$

$= |w_1| + |w_2| + \ldots + |w_d|$

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

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 5]
- 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?