

# The Asymptotics of Tikhonov Regularization

Ross A. Lippert  
MIT Department of Mathematics

February 6, 2006

## Learning unknown functions

We have data:  $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ .

$$Y_i \sim r(X_i)$$

$r \in$  Function Space

But we don't know  $r$

Expect future data: we hope  $Y_{\text{new}} \sim r(X_{\text{new}})$ .

1. Review of Tikhonov regularization
2. Gaussian regularized least squares:
  - (a) An experimental exploration of parameter space
  - (b) The mystery of the asymptotic regressors
3. Key point: Polynomials are the asymptotic regressors for general T.R. problems
4. Describe asymptotics of T.R.
  - (a) Formulation in terms of optimization
  - (b) Optimization asymptotics applied to T.R.
5. Conclusions

## Interpolation

An unknown function

$$r(x) = \sum_{i=1}^n c_i b_i(x)$$

where  $b_i(x)$  are the *basis functions*.

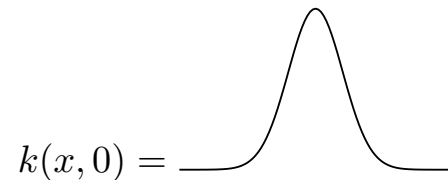
Interpolation: Data  $\{(X_i, Y_i) \in \mathbb{R}^{d+1}\}_{i=1}^n$  and  $r(X_i) = Y_i$

$$Kc = Y \quad \text{where} \quad K = \begin{pmatrix} b_1(X_1) & b_1(X_2) & \cdots \\ b_2(X_1) & b_2(X_2) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

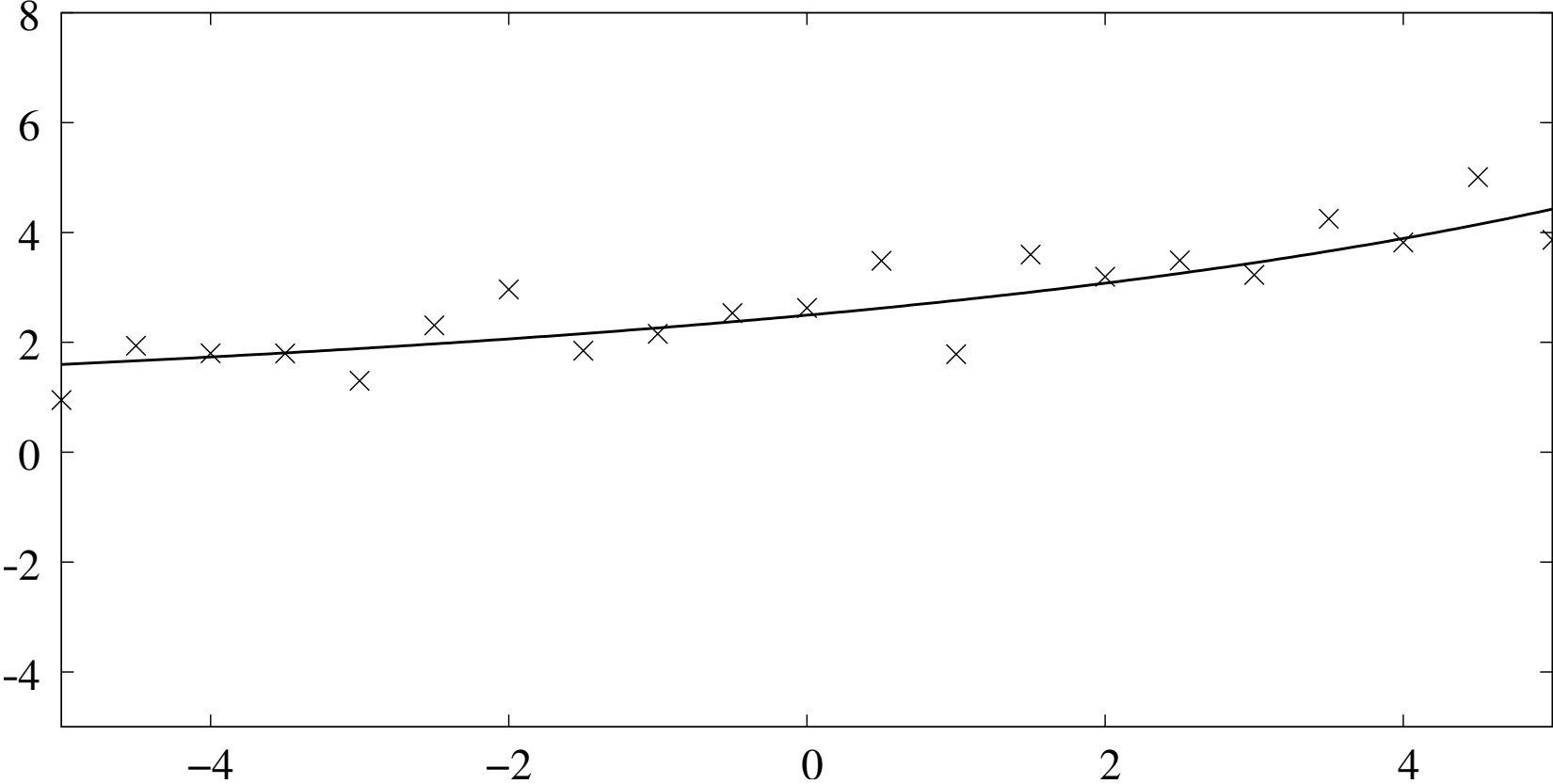
(spd) kernel function:  $k(x, x')$

$$r(x) = \sum_i c_i k(X_i, x) \quad \text{and} \quad K = k(X_*, X_*) = \begin{pmatrix} k(X_1, X_1) & k(X_2, X_1) & \cdots \\ k(X_2, X_1) & k(X_2, X_2) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

e.g.  $k(x, x') = \exp(-\frac{1}{2}\|x - x'\|^2)$

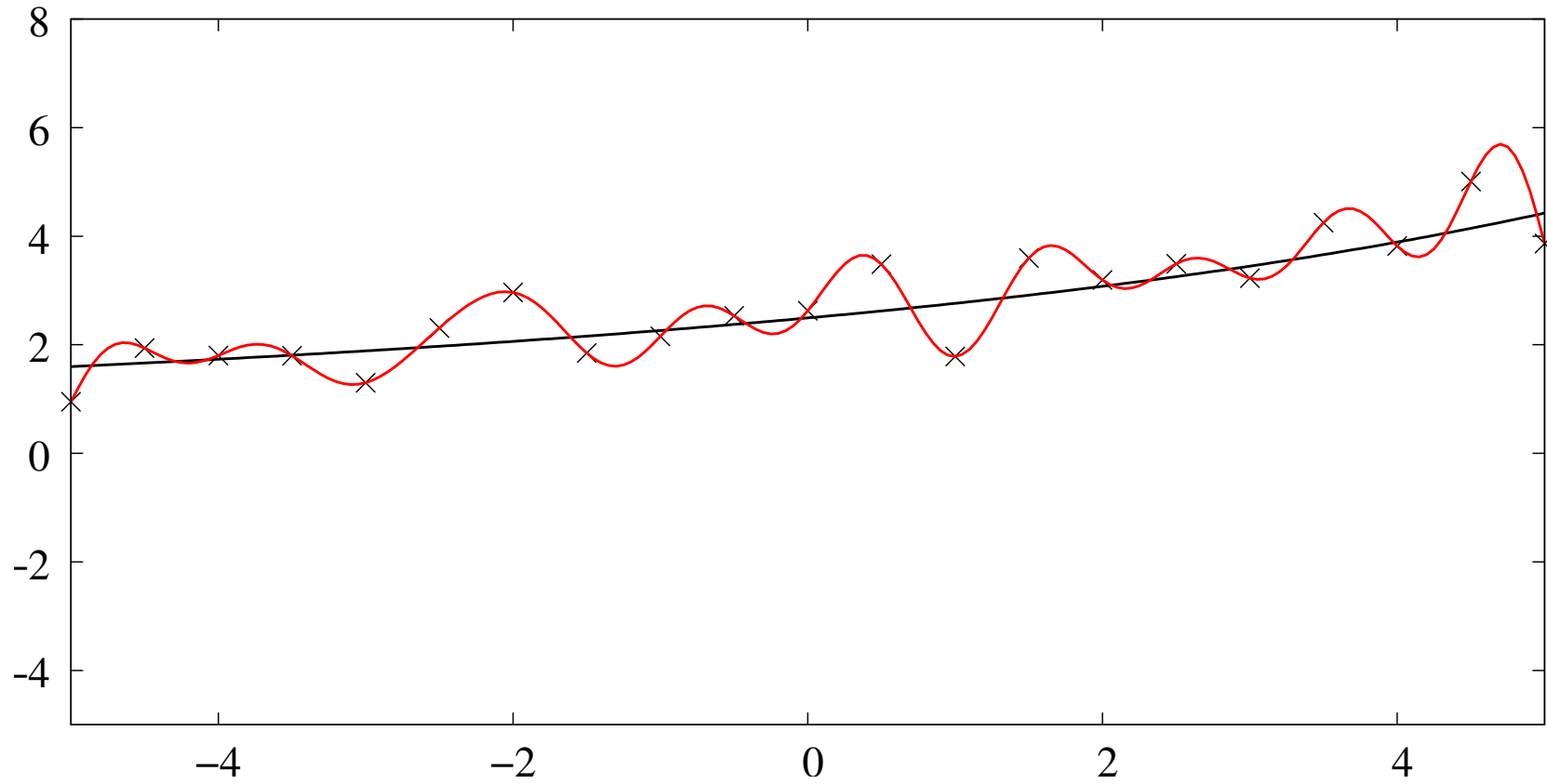


Sample function with measurement errors



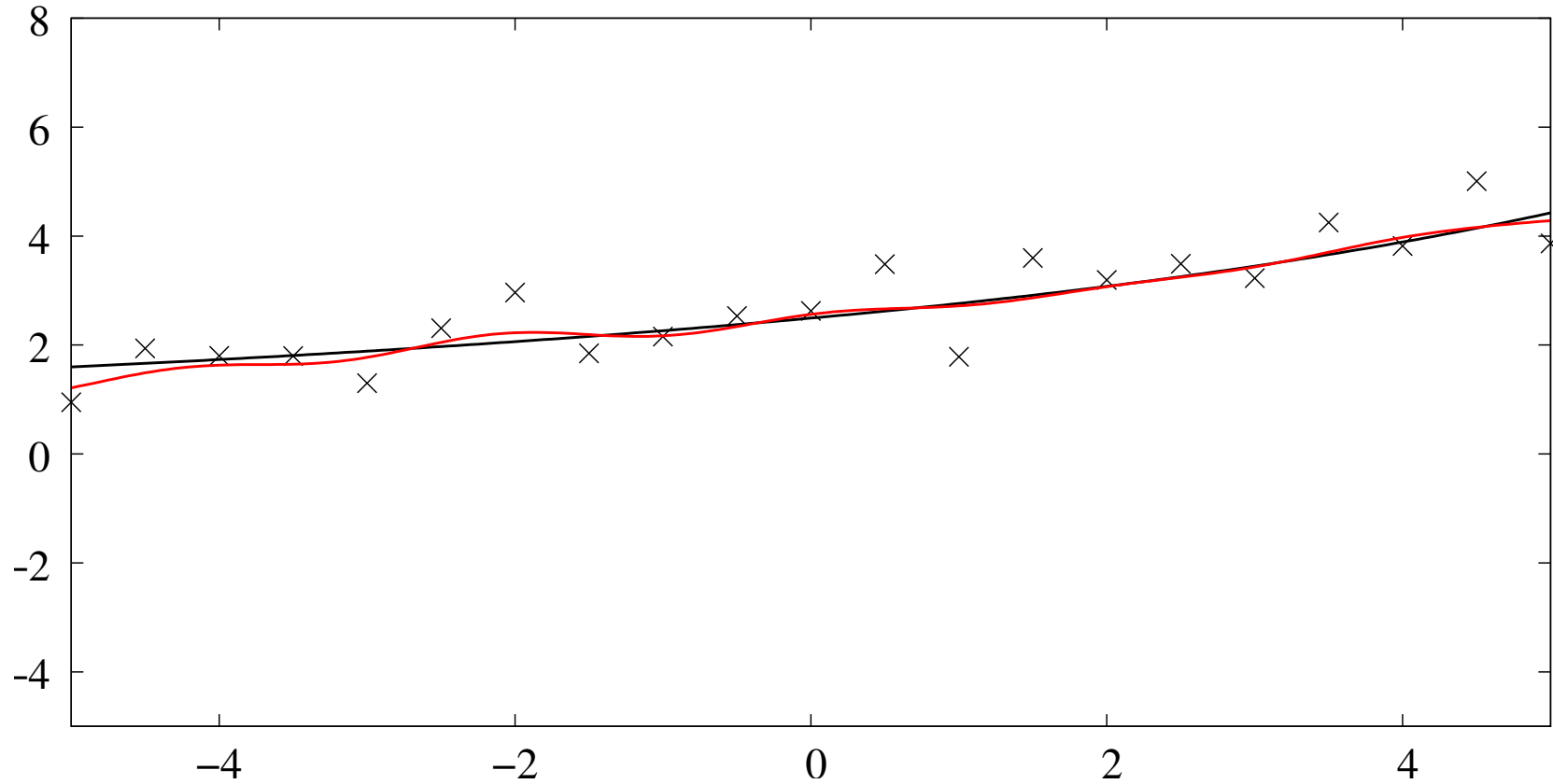
## Shortcomings of interpolation

$c = K^{-1}Y$ : Measurement errors lead to overfitting (and instability)



## Regularized solution

$c = (K + \lambda I_n)^{-1}Y$  gives more stable answer at the expense of fit.



Generalization bounds: this eventually recovers true  $r$ .

## Tikhonov regularization problems

Tikhonov regularization in a *reproducing kernel Hilbert space*

Basis functions:

$$k(x, x') \quad (\text{e.g. } \exp(-\|x - x'\|^2)) \quad \|r\|_k^2 = c^t K c$$

a spd kernel function, (i.e.  $K_{ij} = k(X_i, X_j)$  is spd if all  $X_i$  distinct).

Regression function *represented* by an expansion:

$$r(x) = \sum_{i=1}^n c_i k(X_i, x) = c^t k(X_*, x), \quad \inf_{c \in \mathbb{R}^n} \left\{ \underbrace{\sum_{i=1}^n v_i((Kc)_i)}_{\text{loss}} + \underbrace{\lambda c^t K c}_{\text{regularization}} \right\}$$

$v_i(y_i)$  is the loss coming from point  $i$ .

$$\begin{aligned} v_i(y_i) &= (y_i - Y_i)^2 && \text{least squares regression} \\ &= |y_i - Y_i| && \\ &= \max\{0, 1 - y_i Y_i\} && \text{support vector machine} \\ &= \max\{0, |y_i - Y_i| - \delta\} && \text{support vector regression} \\ &= \log(1 + \exp(-y_i Y_i)) && \text{logistic regression} \end{aligned}$$

Examples: least squares is  $\inf_{c \in \mathbb{R}^n} \{ \|Y - Kc\|^2 + \lambda c^t K c \}$

## Tuning Gaussian regularized least squares

$k_\sigma(x, x') = \exp(-\|x - x'\|^2 / (2\sigma^2))$  Gaussian with bandwidth

$(K_\sigma)_{ij} = k_\sigma(X_i, X_j)$  Kernel matrix

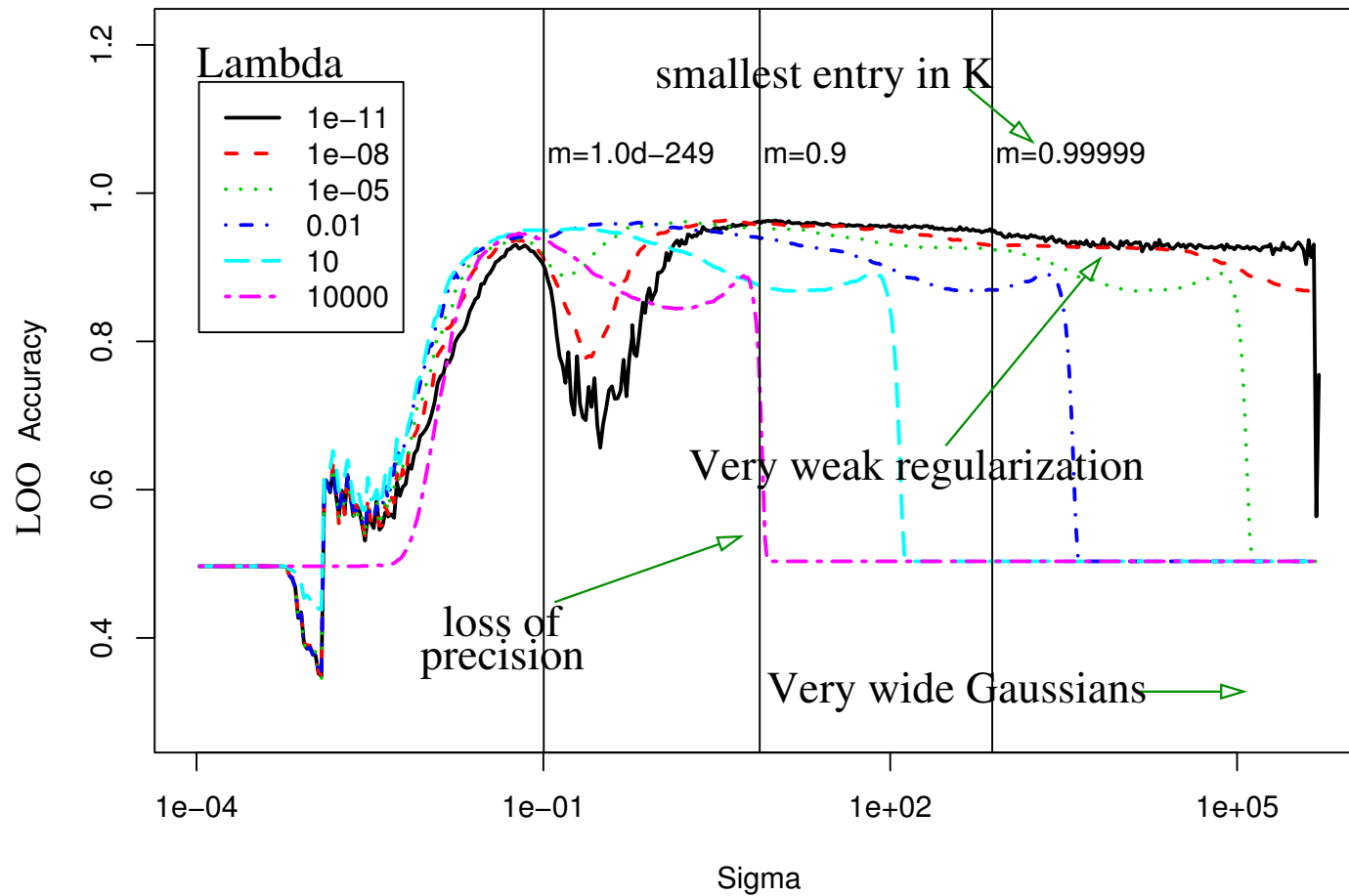
$\inf_{c \in \mathbb{R}^n} \{ \|Y - K_\sigma c\|^2 + \lambda c^t K_\sigma c \}$  optimization

$$c = (K_\sigma + \lambda I)^{-1} Y \quad r(x) = \sum_{i=1}^n c_i k_\sigma(X_i, x)$$

Two free parameters,  $\lambda$  and  $\sigma$ , to tune for optimal performance.

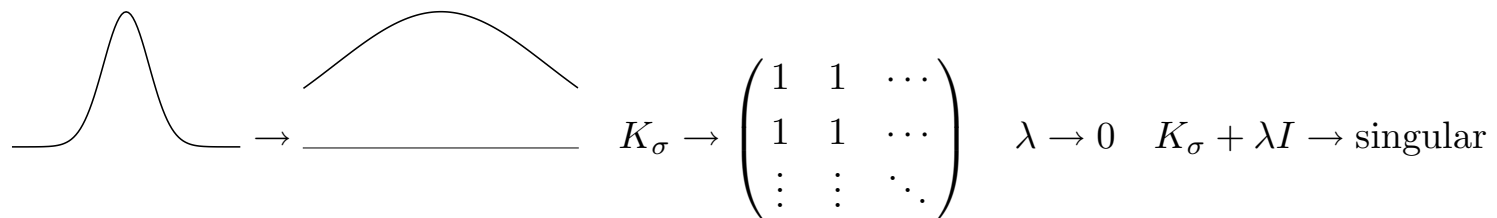
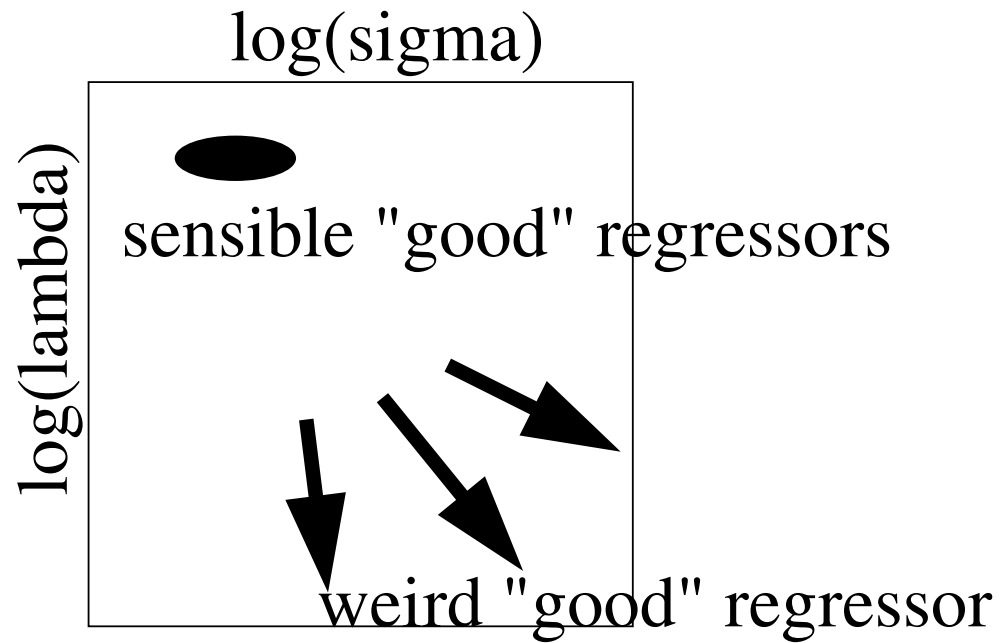
$$\left( k_\sigma(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right), \quad c = (K_\sigma + \lambda I)^{-1}Y, \quad r(x) = \sum_{i=1}^n c_i k_\sigma(X_i, x) \right)$$

### RLSC Results for GALAXY Dataset



## A bit of a mystery

A more schematic view of the observations.



## The result

$$\left( k_\sigma(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right), \quad c = (K_\sigma + \lambda I)^{-1}Y, \quad r(x) = \sum_{i=1}^n c_i k_\sigma(X_i, x) \right)$$

A miracle of cancellation can occur:

$$K_\sigma \rightarrow \begin{pmatrix} 1 & 1 & \cdots \\ 1 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad \text{but also} \quad k_\sigma(X_i, x) \rightarrow 1$$

$$r(x) = c^t k_\sigma(X_*, x) = Y^t \underbrace{(K_\sigma + \lambda I_n)^{-1} k_\sigma(X_*, x)}$$

cancellation  
l'Hopital's rule??

$$(K_\sigma + \lambda I_n)^{-1} k_\sigma(X_*, x) \rightarrow \text{finite}$$

What is going on?

Bottom line: Taylor series + linear algebra ...

## We get polynomial approximation

for  $\sigma \rightarrow \infty$ ,  $\lambda \propto \sigma^{-2p-1}$

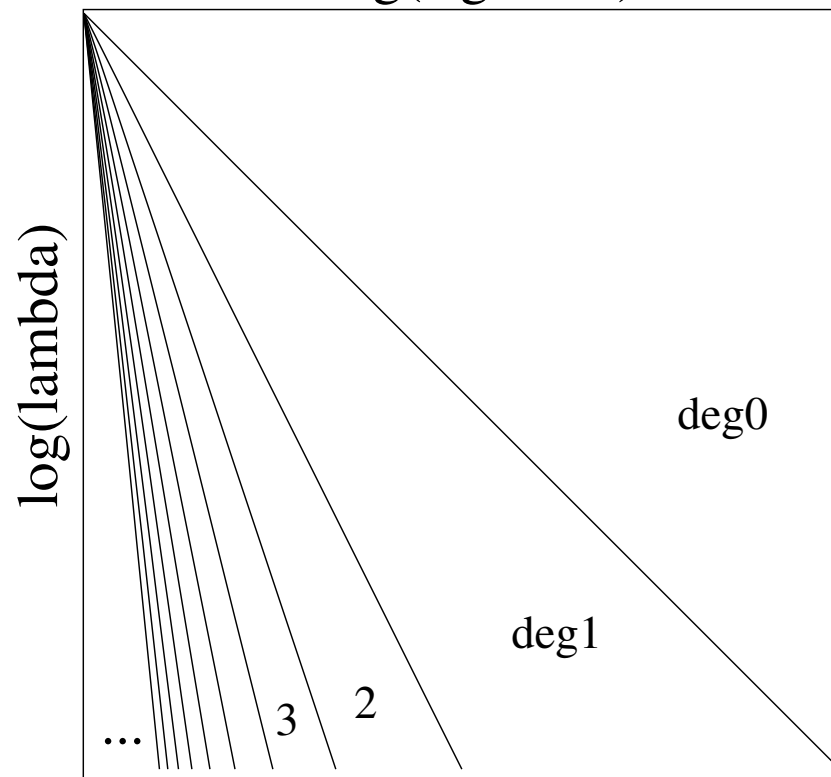
$r(x) \rightarrow$  least squares degree  $p$  polynomial at  $x$

$$\inf_{r \in \text{poly}_p} \{ \|Y - r(X_*)\|^2 \}$$

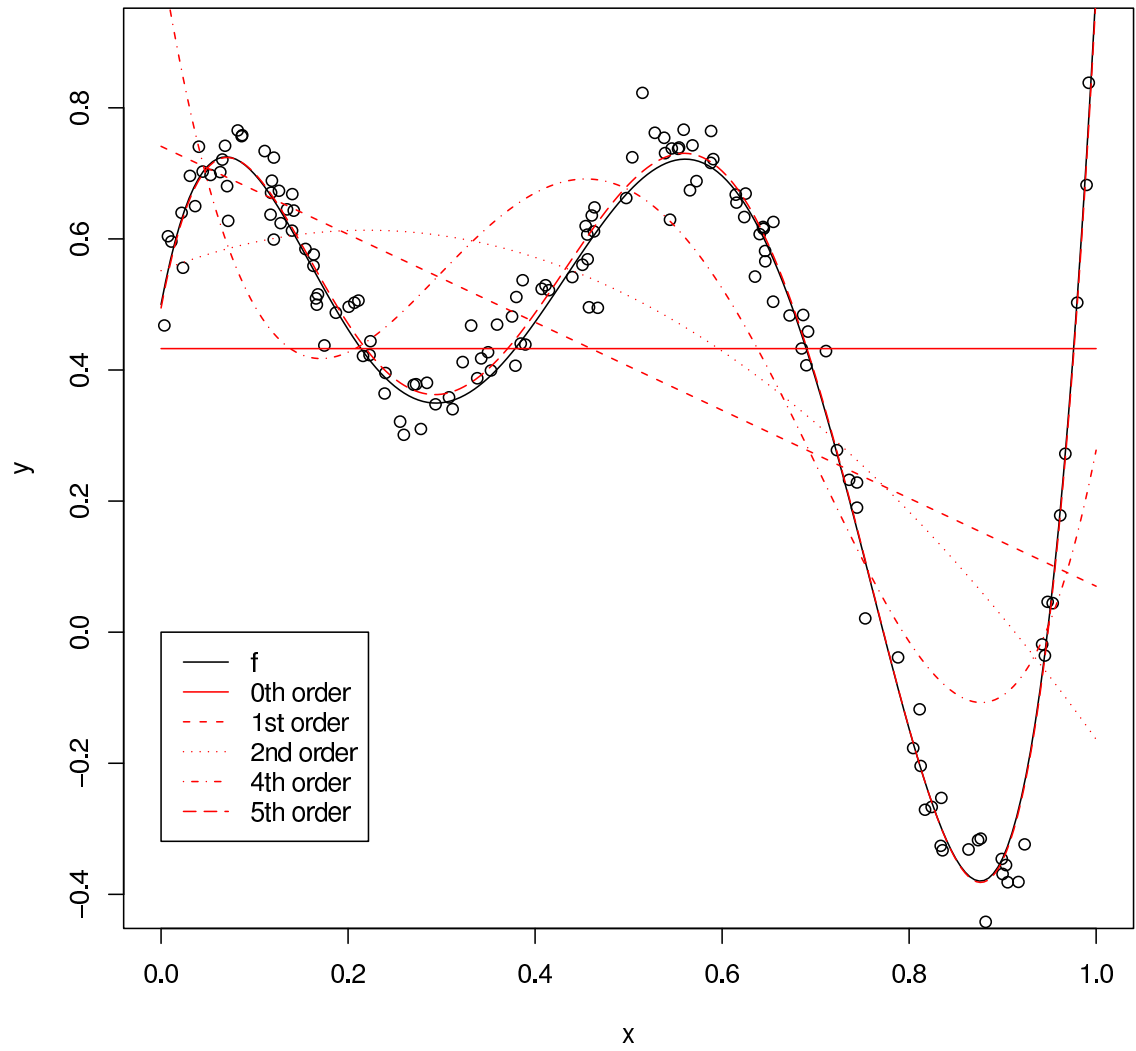
(in 1-dim:)

$$r(x) = \sum_{i=0}^p h_i x^i, \quad h = (VV^t)^{-1}VY, \quad V = \begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \\ \vdots & \ddots & \vdots \\ X_1^p & \cdots & X_n^p \end{pmatrix}$$

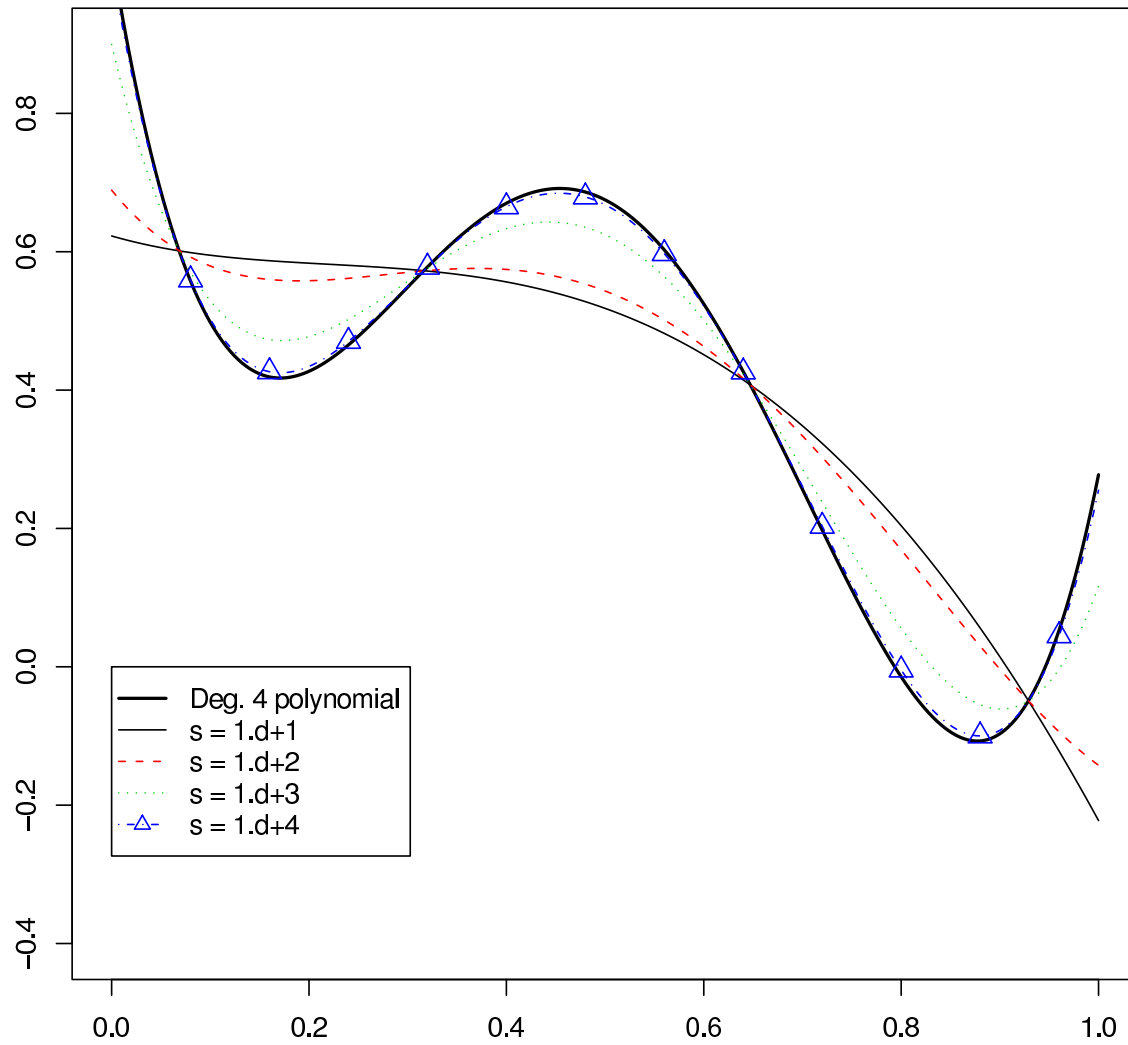
The view at the horizon  
 $\log(\sigma^2)$



**$f(x)$ , Random Sample of  $f(x)$ , and Polynomial Approximations**



4th order solution, and successive approximations.



## A similar horizon result seen for another TR

Keerthi and Lin (2003): Gaussian kernel SVM  $\rightarrow$  linear SVM

Likewise, by Taylor series + quadratic programming.

$\sigma \rightarrow \infty, \lambda \propto \sigma^{-2(1+\epsilon)}, \epsilon \in (0, 1)$ , then  $r(x) \rightarrow$  best fit affine function ( $r(x) = a \cdot x + b$ )

$$\inf_{b \in \mathbb{R}, a \in \mathbb{R}^d} \left\{ \underbrace{\max\{0, 1 - Y_i(a \cdot X_i + b)\}}_{\text{SVM loss}} \right\}$$

$\sigma \rightarrow \infty, \lambda \propto \sigma^{-2}$  then  $r(x) \rightarrow$  regularized least loss affine function

$$\inf_{b \in \mathbb{R}, a \in \mathbb{R}^d} \left\{ \max\{0, 1 - Y_i(a \cdot X_i + b)\} + \underbrace{C\|a\|^2}_{\text{residual regularization}} \right\}$$

**Key point: This is a general property for Tikhonov regularization**

T.R.  $\implies$  fixed degree polynomial approximation

- A general class of spd kernel functions,  $k(x/\sigma, x'/\sigma)$  (not just for Gaussians)
- A general class of loss functions (convex, lower-semicontinuous)
- $\lambda \propto \sigma^{-2p}$  then  $\lfloor p \rfloor$  is the polynomial degree
- If  $p = \lfloor p \rfloor$  then residual regularization

Main ingredients:

- A reformulation which circumvents the basis expansion
- Limits of optimization problems

## Value based formulation

$$r(x) = \sum_{i=1}^n c_i k(X_i, x), \quad \inf_{c \in \mathbb{R}^n} \left\{ \sum_{i=1}^n v_i((Kc)_i) + \lambda c^t Kc \right\}$$

Change of optimization variables,  $y = Kc$  (i.e.  $y_i = r(X_i)$ ) gives a *value-based* form

$$r(x) = \sum_{i=1}^n c_i k(X_i, x), \quad c = K^{-1}y, \quad \inf_{y \in \mathbb{R}^n} \left\{ \sum_{i=1}^n v_i(y_i) + \lambda y^t K^{-1}y \right\}$$

Practical motivation:  $c \rightarrow \infty$  but  $y$  won't diverge.

$$\underbrace{(Y^t(K_\sigma + \lambda I)^{-1}) k_\sigma(X_*, X_0)}_{=c \text{ diverges}} \quad \text{versus} \quad Y^t \underbrace{((K_\sigma + \lambda I)^{-1} k_\sigma(X_*, X_0))}_{\text{finite}}$$

Expansion for  $r(x)$  looks tricky as  $\sigma$  varies,

$$c = K_\sigma^{-1}y, \quad r(x) = \sum_{i=1}^n c_i k_\sigma(X_i, x) = y^t K_\sigma^{-1} k_\sigma(X_*, x)$$

*But we don't actually have to do it.*

## Representing $r(x)$ via optimization

Points:  $X_0 = x, X_1, \dots, X_n$  and Values:  $Y_1, \dots, Y_n$  with Losses:  $v_1(y_1), \dots, v_n(y_n)$ .

What's the difference between T.R. for  $y \in \mathbb{R}^n$  and  $\begin{pmatrix} y_0 \\ y \end{pmatrix} \in \mathbb{R}^{n+1}$ ?

Ans: Nothing.  $y_0 = \sum_{i=1}^n c_i k(X_i, x) = r(x)$  at optimality.

**Theorem 0.1.** Let  $K' = \begin{pmatrix} k(x, x) & k(x, X_*) \\ k(X_*, x) & K \end{pmatrix}$  and  $K = k(X_*, X_*)$

$$\inf_{y' \in \mathbb{R}^{n+1}} \left\{ \sum_{i=1}^n v_i(y'_i) + \lambda y'^t K'^{-1} y' \right\} = \inf_{y \in \mathbb{R}^n} \left\{ \sum_{i=1}^n v_i(y_i) + \lambda y^t K^{-1} y \right\}$$

$$\text{minimizer: } \begin{pmatrix} r(x) \\ y \end{pmatrix} \quad \text{minimizer: } y$$

where  $r(x) = \sum_{i=1}^n c_i k(X_i, x)$  and  $c = K^{-1} y$ .

An optimization-centric version of the *representer theorem*.

## Representing $r(x)$ via optimization

$$\left( K' = \begin{pmatrix} k(x, x) & k(x, X_*) \\ k(X_*, x) & K \end{pmatrix} \text{ and } K = k(X_*, X_*) \right)$$

$$\inf_{y' \in \mathbb{R}^{n+1}} \left\{ \sum_{i=1}^n v_i(y'_i) + \lambda y'^t K'^{-1} y' \right\} = \inf_{y \in \mathbb{R}^n} \left\{ \sum_{i=1}^n v_i(y_i) + \lambda \cdot \inf_{y_0 \in \mathbb{R}} \{ y'^t K'^{-1} y' \} \right\}$$

**Lemma 0.2.** For  $y \in \mathbb{R}^n$  fixed,

$$\inf_{y_0} \left\{ \begin{pmatrix} y_0 \\ y \end{pmatrix}^t \begin{pmatrix} a & b^t \\ b & C \end{pmatrix}^{-1} \begin{pmatrix} y_0 \\ y \end{pmatrix} \right\} = y^t C^{-1} y$$

where at optimality,  $y_0 = (C^{-1}y)^t b$ .

*Proof.* At optimal  $y_0$ ,

$$\begin{aligned} \frac{d}{dy_0} (y_0 \quad y^t) \begin{pmatrix} a & b^t \\ b & C \end{pmatrix}^{-1} \begin{pmatrix} y_0 \\ y \end{pmatrix} = 0 &\Rightarrow (1 \quad 0) \begin{pmatrix} a & b^t \\ b & C \end{pmatrix}^{-1} \begin{pmatrix} y_0 \\ y \end{pmatrix} = 0 \\ &\Rightarrow \frac{1}{a - b^t C^{-1} b} (1 \quad -b^t C^{-1}) \begin{pmatrix} y_0 \\ y \end{pmatrix} = 0 \\ &\Rightarrow y_0 = b^t C^{-1} y. \end{aligned}$$

Substitution gives the rest. □

## Value-based formulation reduces everything to optimization

Instead of

$$r(x) = \sum_{i=1}^n c_i k(X_i, x), \quad c = K^{-1}y, \quad \inf_{y \in \mathbb{R}^n} \left\{ \sum_{i=1}^n v_i(y_i) + \lambda y^t K^{-1}y \right\}$$

we construct  $r(x)$  via optimization on an augmented system,

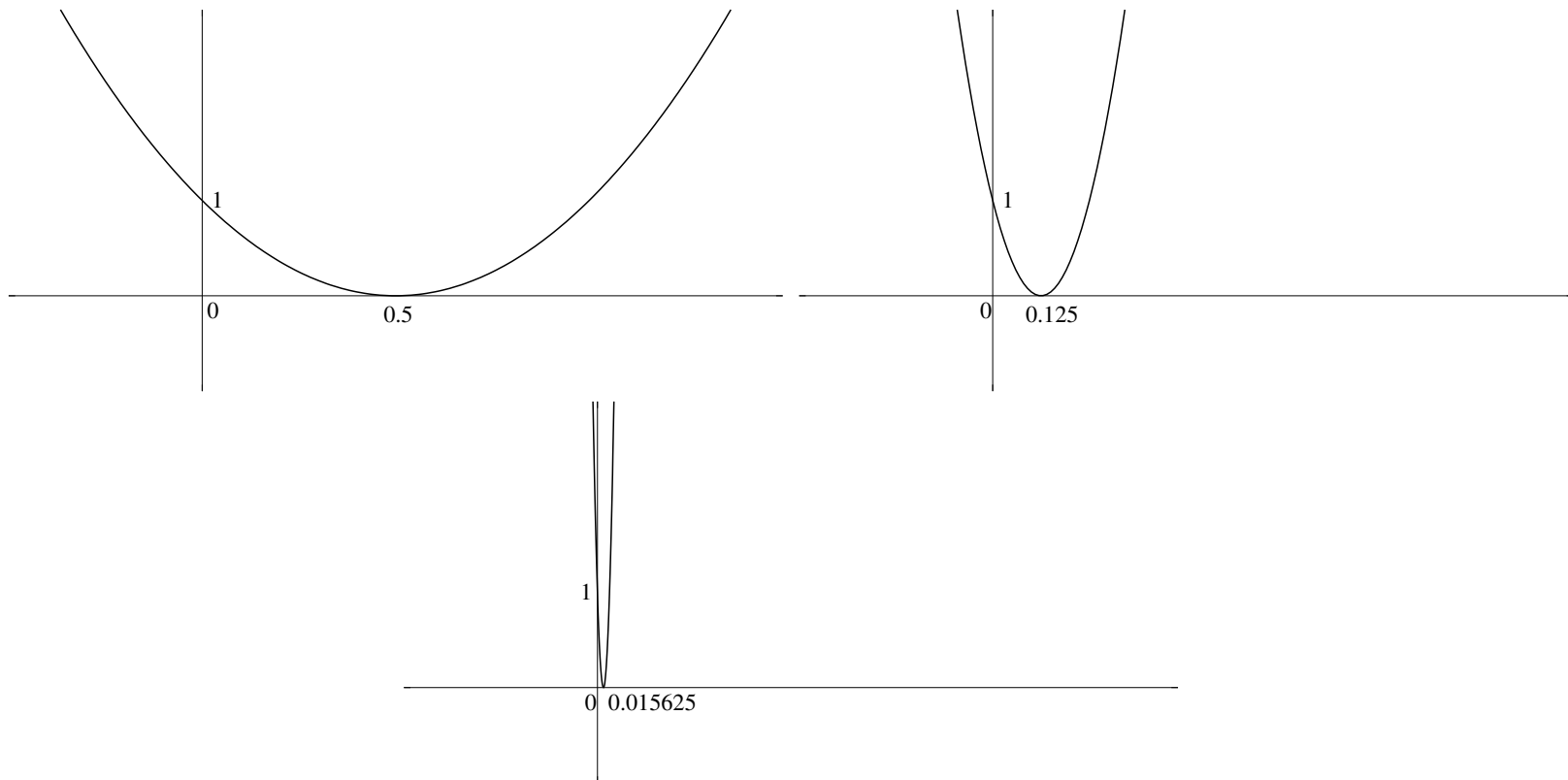
$$r(x) = y_0 \text{ from } \inf_{y \in \mathbb{R}^{n+1}} \left\{ \sum_{i=1}^n v_i(y_i) + \lambda y^t K^{-1}y \right\}, \quad K = \begin{pmatrix} k(x, x) & k(x, X_*) \\ k(X_*, x) & k(X_*, X_*) \end{pmatrix}$$

Asymptotics of T.R. via limits of optimization

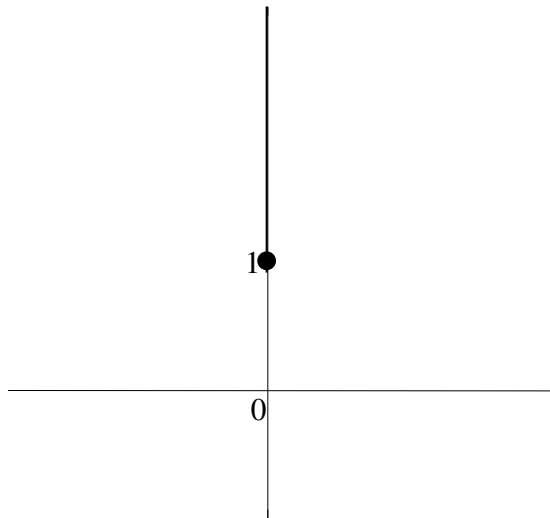
## Asymptotics of optimization: convergence issues

What optimization problem does this become?

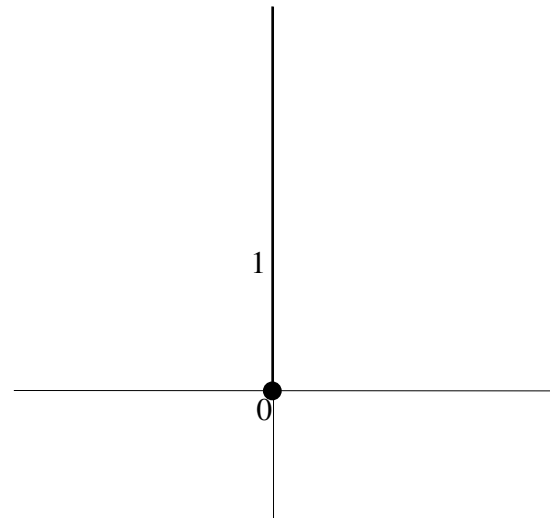
$$f_s(y) = \frac{2}{s}y(y - s) + 1$$



Which limit do we want?



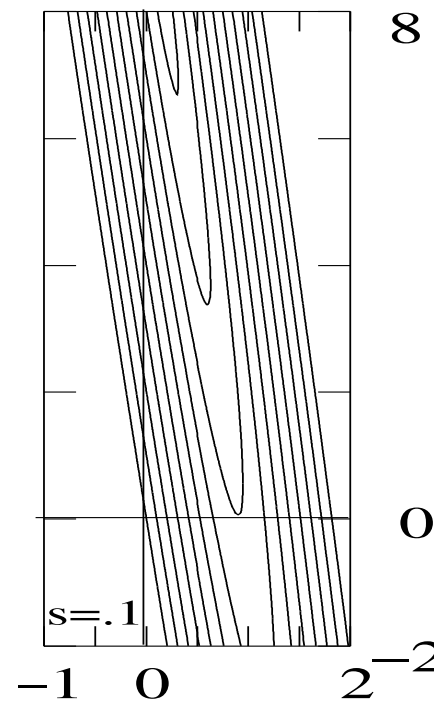
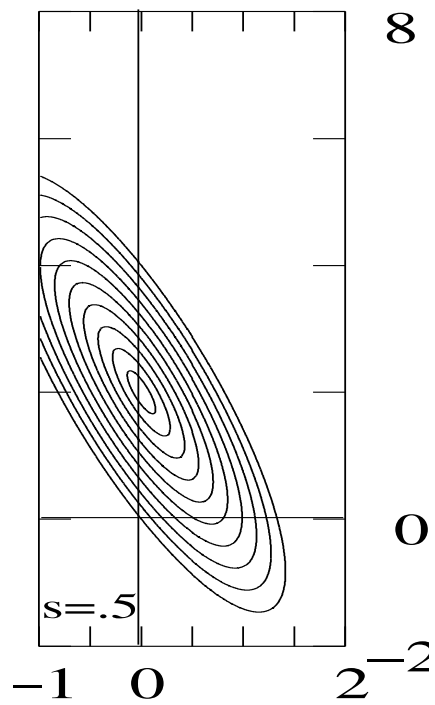
versus



- Clearly the minimizer was  $\rightarrow 0$
- $f_s(0) = 1$  for all  $s > 0$
- $\inf_y f_s(y) = 0$  for all  $s > 0$

## Another puzzle

$$f_s(y_1, y_2) = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}^t \begin{pmatrix} 1+s & s \\ s & s^2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - 2 \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}^t \begin{pmatrix} 1 \\ s \end{pmatrix} + 1$$



## Another puzzle (cont)

$$f_s(y_1, y_2) = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}^t \begin{pmatrix} 1+s & s \\ s & s^2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - 2 \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}^t \begin{pmatrix} 1 \\ s \end{pmatrix} + 1$$
$$\inf f_s = 0, \quad \text{minimizer: } \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{s} \end{pmatrix}$$

But,

$$f_0(y_1, y_2) = y_1^2 - 2y_1 + 1 \quad \text{minimized by } \begin{pmatrix} 1 \\ * \end{pmatrix}$$

and

$$\text{“decoy”}: \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

For any fixed  $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ , the decoy is eventually better:  $f_s \begin{pmatrix} 1 \\ 0 \end{pmatrix} \leq f_s \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$  as  $s \rightarrow 0$ .

## limits of T.R. via limits of optimization

We'll need a notion of *functional limit*,  $f_s \xrightarrow{?} f$  compatible with optimization.

Our wishlist:

- continuous w.r.t. minimization

$$\lim_{s \rightarrow 0} \inf_y \{f_s(y)\} = \inf_y \{f(y)\}$$

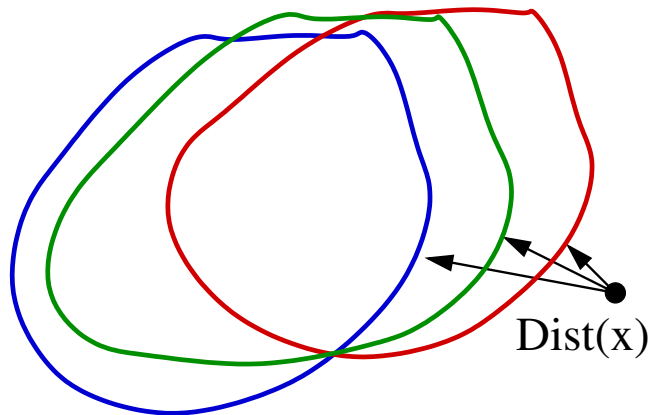
- continuous w.r.t. minimizers

$$\lim_{s \rightarrow 0} \operatorname{argmin}_y \{f_s(y)\} = \operatorname{argmin}_y \{f(y)\}$$

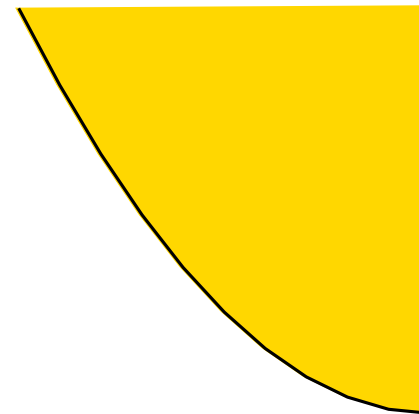
In our problem, as  $\lambda^{-1}, \sigma \rightarrow \infty$ ,

$$\begin{array}{ccc}
 \underbrace{\lambda y^t K_\sigma^{-1} y}_{\text{regularization}} & \xrightarrow{?} & \underbrace{Q(y)}_{\text{limiting regularization}} \\
 \inf_{y \in \mathbb{R}^{n+1}} \left\{ \sum_{i=1}^n v_i(y_i) + \lambda y^t K_\sigma^{-1} y \right\} & \longrightarrow & \inf_{y \in \mathbb{R}^{n+1}} \left\{ \sum_{i=1}^n v_i(y_i) + Q(y) \right\} \\
 \underbrace{\hat{y}^{(\lambda, \sigma)}}_{\text{minimizer for } \lambda, \sigma} & \longrightarrow & \underbrace{\hat{y}}_{\text{fixed degree polynomial in } X_i}
 \end{array}$$

## Epigraphical convergence



epigraph of  $f$



$$f(x) = \begin{cases} x^2 & x \leq 0 \\ \infty & x > 0 \end{cases}$$

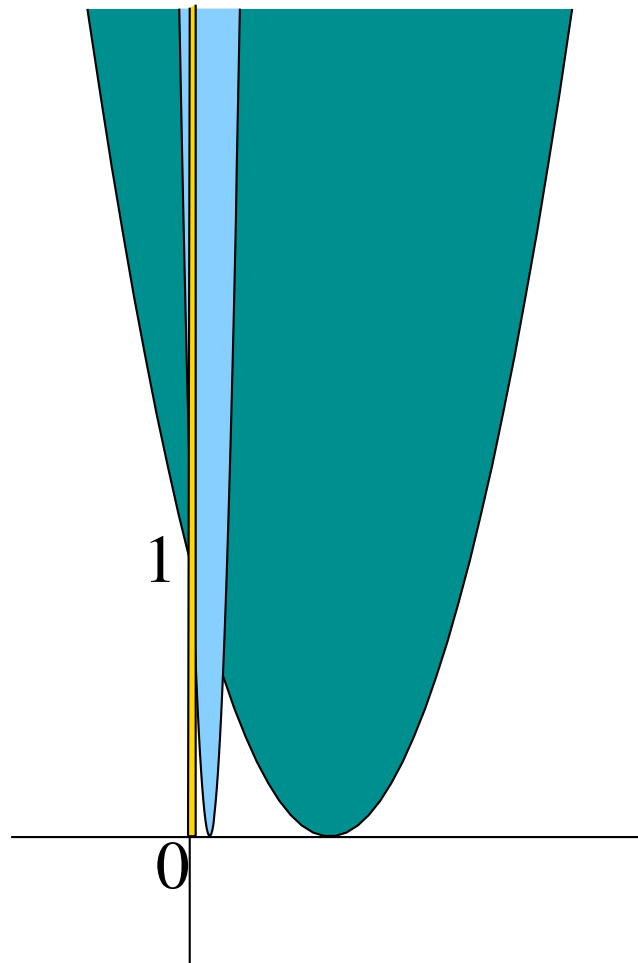
Painlevé-Kuratowski convergence

$$\lim_{s \rightarrow 0} C_s = \left\{ x : \lim_{s \rightarrow 0} \text{dist}(x, C_s) = 0 \right\}$$

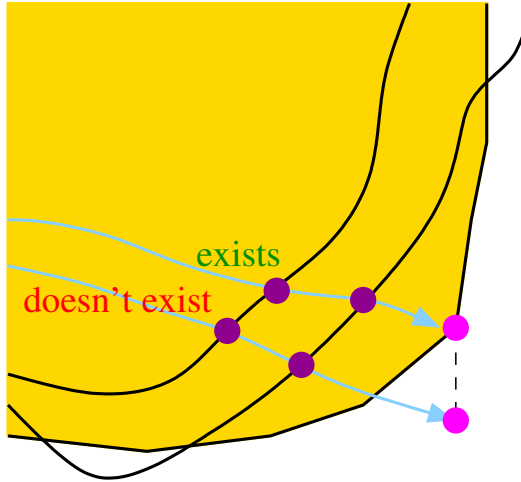
(*lower semi-continuous* is equivalent to *epigraph is closed*, so PK convergence is applicable.)

On a previous example

$$f_s(y) = \frac{2}{s}y(y - s) + 1 \xrightarrow{\text{epi}} f(y) = \begin{cases} 0 & y = 0 \\ \infty & y \neq 0 \end{cases}$$



## “Hit or miss” conditions



**Definition 0.3.**  $f_s, f : \mathbb{R}^n \rightarrow (-\infty, \infty]$  for  $s > 0$ .

$$\text{epi-}\lim_{s \rightarrow 0} f_s \equiv f$$

at any  $y_0 \in \mathbb{R}^n$ ,  $f(y_0) < \infty$

$$\exists y(s) : \lim_{s \rightarrow 0} y(s) = y_0 \quad : \quad \lim_{s \rightarrow 0} f_s(y(s)) = f(y_0),$$

$$\forall y(s) : \lim_{s \rightarrow 0} y(s) = y_0 \quad : \quad \liminf_{s \rightarrow 0} f_s(y(s)) \geq f(y_0).$$

at any  $y_0 \in \mathbb{R}^n$ ,  $f(y_0) = \infty$

$$\forall y(s) : \lim_{s \rightarrow 0} y(s) = y_0 \quad : \quad \liminf_{s \rightarrow 0} f_s(y(s)) = \infty$$

The difference from pointwise convergence is the **coupling** between  $y(s)$  and  $f_s$ .

## We get what we want

Limits of minimizers work out.

**Theorem 0.4.** *Let  $f_s, f : \mathbb{R}^n \rightarrow (-\infty, \infty]$  be convex, proper, and lsc, with  $\text{epi-lim}_{s \rightarrow 0} f_s = f$ .*

*If  $f_s, f$  have unique minimizers  $\hat{y}_s, \hat{y}$ ,*

$$\lim_{s \rightarrow 0} \hat{y}_s = \hat{y} \quad \text{and} \quad \lim_{s \rightarrow 0} f_s(\hat{y}_s) = f(\hat{y}).$$

The idea:

$$\begin{aligned} & \lambda y^t K_\sigma^{-1} y \xrightarrow{\text{epi}} Q(y) \\ \Rightarrow \sum_{i=1}^n v_i(y_i) + \lambda y^t K_\sigma^{-1} y & \xrightarrow{\text{epi}} \sum_{i=1}^n v_i(y_i) + Q(y) \end{aligned}$$

in the value-based formulation we only need to study *quadratic forms*.

## The epigraphical convergence of quadratic forms

**Lemma 0.5.**  $Z : [0, \infty) \rightarrow \mathbb{R}^{n \times n}$  continuous,  $Z(s)$  non-singular for  $s > 0$

$$Z(0) = 0.$$

Let  $\begin{pmatrix} A(s) & B(s)^t \\ B(s) & C(s) \end{pmatrix} \in \mathbb{R}^{(m+n) \times (m+n)}$  be continuous pos semi-def for  $s \geq 0$ , and  $C(s)$  posdef.

$$f_s(y_1, y_2) = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}^t \begin{pmatrix} I_m & 0 \\ 0 & Z(s)^{-t} \end{pmatrix} \begin{pmatrix} A(s) & B(s)^t \\ B(s) & C(s) \end{pmatrix} \begin{pmatrix} I_m & 0 \\ 0 & Z(s)^{-1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

epi- $\lim_{s \rightarrow 0} f_s = f$ , where

$$f(y_1, y_2) = \begin{cases} y_1^t \tilde{A} y_1 & y_2 = 0 \\ \infty & y_2 \neq 0 \end{cases}$$

and  $\tilde{A} = A(0) - B(0)^t C(0)^{-1} B(0)$ .

(Pointwise reaches a similar conclusion with  $A(0)$  instead of  $\tilde{A}$ .)

## Factorized form form

(this form is pedagogically better for what comes next)

**Corollary 0.6.** For  $Z(s), N(s), M(s)$  continuous for  $s \geq 0$ ,

$$Z(0) = 0, \quad N(0) = N_0 \quad \text{and} \quad M(s) = \begin{pmatrix} D & 0 \\ E & F \end{pmatrix} \begin{pmatrix} D^t & E^t \\ 0 & F^t \end{pmatrix} + O(s)$$

with  $D, F$  invertible.

$$\begin{aligned} f_s(y_1, y_2) &= \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}^t \begin{pmatrix} N(s)^t & 0 \\ 0 & Z(s)^{-t} \end{pmatrix} M(s)^{-1} \begin{pmatrix} N(s) & 0 \\ 0 & Z(s)^{-1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\ &= \begin{pmatrix} N(s)y_1 \\ Z(s)^{-1}y_2 \end{pmatrix}^t \left( \begin{pmatrix} D^{-t} & -D^{-t}E^tF^{-t} \\ 0 & F^{-t} \end{pmatrix} \begin{pmatrix} D^{-1} & 0 \\ -F^{-1}ED^{-1} & F^{-1} \end{pmatrix} + O(s) \right) \begin{pmatrix} N(s)y_1 \\ Z(s)^{-1}y_2 \end{pmatrix} \end{aligned}$$

then  $\text{epi-lim}_{s \rightarrow 0} f_s = f$ , where

$$f(y_1, y_2) = \begin{cases} \|D^{-1}N_0y_1\|^2 & y_2 = 0 \\ \infty & y_2 \neq 0 \end{cases}.$$

*Proof.*

$$\begin{aligned}
f_s(y_1, y_2) &= \begin{pmatrix} N(s)y_1 \\ Z(s)^{-1}y_2 \end{pmatrix}^t \left( \begin{pmatrix} D^{-t} & -D^{-t}E^tF^{-t} \\ 0 & F^{-t} \end{pmatrix} \begin{pmatrix} D^{-1} & 0 \\ -F^{-1}ED^{-1} & F^{-1} \end{pmatrix} + O(s) \right) \begin{pmatrix} N(s)y_1 \\ Z(s)^{-1}y_2 \end{pmatrix} \\
&= \left\| \begin{pmatrix} D^{-1} & 0 \\ -F^{-1}ED^{-1} & F^{-1} \end{pmatrix} \begin{pmatrix} N(s)y_1 \\ Z(s)^{-1}y_2 \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} N(s)y_1 \\ Z(s)^{-1}y_2 \end{pmatrix} \right\|_{O(s)}^2 \\
&= \|D^{-1}N(s)y_1\|^2 + \|F^{-1}(ED^{-1}N(s)y_1 - Z(s)^{-1}y_2)\|^2 + (1 + \|Z(s)^{-1}y_2\|^2)O(s)
\end{aligned}$$

(hit) If  $y_1(s) = y_1$  and  $y_2(s) = Z(s)ED^{-1}N(s)y_1$  then  $(y_1(s), y_2(s)) \rightarrow (y_1, 0)$  and

$$f_s(y_1(s), y_2(s)) = \|D^{-1}N(s)y_1\|^2 + O(s)$$

and  $\lim_{s \rightarrow 0} f_s(y_1(s), y_2(s)) = \|D^{-1}N(0)y_1\|^2$

(miss) If  $(y_1(s), y_2(s)) \rightarrow (y_1, 0)$ ,

$$\begin{aligned}
f_s(y_1(s), y_2(s)) &\geq \|D^{-1}N(s)y_1(s)\|^2 + (c_1 + O(s))\|Z(s)^{-1}y_2(s)\|^2 \\
&\quad - (c_2 + O(s))\|Z(s)^{-1}y_2(s)\| + O(s)
\end{aligned}$$

and  $\liminf_{s \rightarrow 0} f_s(y_1(s), y_2(s)) \geq \|D^{-1}N_0y_1(0)\|^2$ .

If  $y_2 \neq 0$  and  $(y_1(s), y_2(s)) \rightarrow (y_1, y_2)$

$$f_s(y_1(s), y_2(s)) \geq (c_3 + O(s))\|Z(s)^{-1}y_2(s)\|^2 + O(1)$$

and  $\liminf_{s \rightarrow 0} f_s(y_1(s), y_2(s)) = \infty$ .

□

## Epi-convergence of the regularization term (1-dim)

Analytic expansion:

$$k(x, x') = \sum_{i,j \geq 0} a_{ij} x^i x'^j = \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \end{pmatrix}^t \underbrace{\begin{pmatrix} a_{00} & a_{01} & a_{02} & \cdots \\ a_{10} & a_{11} & a_{12} & \cdots \\ a_{20} & a_{21} & a_{22} & \cdots \\ \vdots & \vdots & \ddots & \ddots \end{pmatrix}}_{\text{coefficient matrix } A} \begin{pmatrix} 1 \\ x' \\ x'^2 \\ \vdots \end{pmatrix}$$

require  $A_m = \begin{pmatrix} a_{00} & \cdots & a_{0m} \\ \vdots & \ddots & \vdots \\ a_{m0} & \cdots & a_{mm} \end{pmatrix}$  positive definite for all  $m$ .

A Cholesky form makes the residuals easier to express

$$k(x, x') = \sum_{i \geq 0} (xx')^i g_i(x) g_i(x') = \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \end{pmatrix}^t \underbrace{\begin{pmatrix} g_{00} & 0 & 0 & \cdots \\ g_{01} & g_{10} & 0 & \cdots \\ g_{02} & g_{11} & g_{20} & \cdots \\ \vdots & \vdots & \ddots & \ddots \end{pmatrix}}_{\text{coefficient matrix } G} \begin{pmatrix} g_{00} & g_{01} & g_{02} & \cdots \\ 0 & g_{10} & g_{11} & \cdots \\ 0 & 0 & g_{20} & \cdots \\ \vdots & \vdots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} 1 \\ x' \\ x'^2 \\ \vdots \end{pmatrix}$$

require  $g_{i0} = g_i(0) > 0$ .

## Expansion examples

$$\begin{aligned}k(x, x') &= \exp(-\|x - x'\|^2/2) = \exp(x \cdot x') e^{-\frac{1}{2}\|x\|^2} e^{-\frac{1}{2}\|x'\|^2} \\&= \sum_{i \geq 0} \frac{(x \cdot x')^i}{i!} e^{-\frac{1}{2}\|x\|^2} e^{-\frac{1}{2}\|x'\|^2} \\&= \sum_{i \geq 0} (x \cdot x')^i \underbrace{\left( \frac{1}{\sqrt{i!}} e^{-\frac{1}{2}\|x\|^2} \right)}_{g_i(x)} \left( \frac{1}{\sqrt{i!}} e^{-\frac{1}{2}\|x'\|^2} \right) \\ \Rightarrow g_i(0) &= 1/\sqrt{i!}\end{aligned}$$

$$\begin{aligned}k(x, x') &= \frac{1}{1 - x \cdot x'} \\&= 1 + x \cdot x' + (x \cdot x')^2 + (x \cdot x')^3 + \dots \\ \Rightarrow g_i(0) &= 1\end{aligned}$$

Non-example:  $k(x, x') = 1 + x \cdot x' + (x \cdot x')^2$

## Epi-convergence of the regularization term (1-dim)

Expansion:  $k(x, x') = \sum_{i \geq 0} (xx')^i g_i(x) g_i(x')$

**Theorem 0.7.** *With  $\{X_i\}_{i=0}^n$  distinct.  $\sigma(s) = s^{-1}$ ,  $\lambda(s) = s^{2p}$  for  $0 < p < n$*

$$(K_\sigma)_{ij} = k(X_i/\sigma, X_j/\sigma) \quad \text{and} \quad f_s(y) = \lambda(s) y^t K_{\sigma(s)}^{-1} y$$

*Then  $\text{epi-lim}_{s \rightarrow 0} f_s = f$  where*

$$f(y) = \underbrace{\left\{ \begin{array}{ll} 0 & y_i = \sum_{j=0}^{\lfloor p \rfloor} c_j X_i^j \\ \infty & y_i \neq \sum_{j=0}^{\lfloor p \rfloor} c_j X_i^j \end{array} \right\}}_{y \text{ is a poly of degree } \lfloor p \rfloor} + \underbrace{\left\{ \begin{array}{ll} 0 & p \notin \mathbb{Z} \\ (c_p/g_p(0))^2 & p \in \mathbb{Z} \end{array} \right\}}_{\text{residual regularization}}.$$

$$\begin{aligned}
\text{Proof. } K_{\sigma(s)} &= \begin{pmatrix} 1 & sX_0 & s^2X_0^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ 1 & sX_n & s^2X_n^2 & \cdots \end{pmatrix} GG^t \begin{pmatrix} 1 & \cdots & 1 \\ sX_0 & \cdots & sX_n \\ s^2X_0^2 & \cdots & s^2X_n^2 \\ \vdots & \ddots & \vdots \end{pmatrix} \\
&= \begin{pmatrix} 1 & \cdots & 1 \\ X_0 & \cdots & X_n \\ X_0^2 & \cdots & X_n^2 \\ \vdots & \ddots & \vdots \end{pmatrix}^t \begin{pmatrix} 1 & 0 & \cdots \\ 0 & s & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} GG^t \begin{pmatrix} 1 & 0 & \cdots \\ 0 & s & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \\ X_0 & \cdots & X_n \\ X_0^2 & \cdots & X_n^2 \\ \vdots & \ddots & \vdots \end{pmatrix} \\
&= \begin{pmatrix} 1 & \cdots & 1 \\ X_0 & \cdots & X_n \\ \vdots & \ddots & \vdots \\ X_0^n & \cdots & X_n^n \end{pmatrix}^t \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s^n \end{pmatrix} (G_n G_n^t + O(s)) \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s^n \end{pmatrix} \underbrace{\begin{pmatrix} 1 & \cdots & 1 \\ X_0 & \cdots & X_n \\ \vdots & \ddots & \vdots \\ X_0^n & \cdots & X_n^n \end{pmatrix}}_V \\
\underbrace{s^{2p} y^t K_{\sigma(s)}^{-1} y}_{f_s(y)} &= y^t V^{-1} \begin{pmatrix} s^p & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s^{p-n} \end{pmatrix} (G_n^t G_n + O(s))^{-1} \begin{pmatrix} s^p & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s^{p-n} \end{pmatrix} V^{-t} y
\end{aligned}$$

Setup for the corollary

$$f_s(y) = y^t V^{-1} \begin{pmatrix} s^p & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & s^{p-n} \end{pmatrix} (G_n^t G_n + O(s))^{-1} \begin{pmatrix} s^p & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & s^{p-n} \end{pmatrix} V^{-t} y$$

$$G_n = \begin{pmatrix} D & 0 \\ E & F \end{pmatrix} = \begin{pmatrix} G_{[p]} & 0 \\ * & * \end{pmatrix}, \quad N(s) = \begin{pmatrix} s^p & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & s^{p-[p]} \end{pmatrix}, \quad Z(s) = \begin{pmatrix} s^{1+[p]-p} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & s^{n-p} \end{pmatrix},$$

$$\text{Hence } f_s \xrightarrow{\text{epi}} f \text{ where } f(Vc) = f\left(\sum_{i=0}^n c_i X_*^i\right) = \begin{cases} \|G_{[p]}^{-1} Nc\|^2 & c_{i>p} = 0 \\ \infty & c_{i>p} \neq 0 \end{cases}.$$

If  $p \notin \mathbb{Z}$ ,  $N_0 = 0_{[p]+1}$ , hence

$$f\left(y = \sum_{i=0}^n c_i X_*^i\right) = \begin{cases} 0 & c_{i>p} = 0 \\ \infty & c_{i>p} \neq 0 \end{cases}$$

$$\text{otherwise } N_0 = \begin{pmatrix} 0_{[p]} & 0 \\ 0 & 1 \end{pmatrix} \text{ and } G_p^{-1} = \begin{pmatrix} * & 0 \\ * & g_{p0} \end{pmatrix}^{-1} = \begin{pmatrix} * & 0 \\ * & 1/g_{p0} \end{pmatrix}$$

$$f\left(y = \sum_{i=0}^n c_i X_*^i\right) = \begin{cases} (c_p/g_{p0})^2 & c_{i>p} = 0 \\ \infty & c_{i>p} \neq 0 \end{cases}$$

□

## Epi-convergence of the regularization term (d-dim)

Monomials:  $x^I = x_1^{I_1} x_2^{I_2} \cdots x_d^{I_d}$

Degree:  $|I| = I_1 + I_2 + \cdots + I_d$

Multinomial coefficients:  $\binom{P}{I} = \frac{P!}{I_1! \cdots I_d!}$

Expansion:

$$k(x, x') = \sum_{i \geq 0} (x \cdot x')^i g_i(x) g_i(x')$$

(this form chosen for ease of exposition; posdef Taylor coeff's are actually enough)

**Theorem 0.8.** *If  $\sigma(s) = s^{-1}$ ,  $\lambda(s) = s^{2p}$  for  $0 < p < (\max \text{ degree})$  then  $\text{epi-lim}_{s \rightarrow 0} f_s = f$  where*

$$f(y) = \underbrace{\left\{ \begin{array}{ll} 0 & y_i = \sum_{|J| \leq [p]} c_J X_i^J \\ \infty & y_i \neq \sum_{|J| \leq [p]} c_J X_i^J \end{array} \right\}}_{y \text{ is a poly of degree } [p]} + \underbrace{\left\{ \begin{array}{ll} 0 & p \notin \mathbb{Z} \\ g_p(0)^{-2} \sum_{|J|=p} \binom{p}{J}^{-1} c_J^2 & p \in \mathbb{Z} \end{array} \right\}}_{\text{residual regularization}}.$$

## Putting it all together

The regularization epi-converges, so does the objective function

$$\sum_{i=1}^n v_i(y_i) + f_s(y) \xrightarrow{\text{epi}} \sum_{i=1}^n v_i(y_i) + f(y)$$

$$\begin{aligned} \inf_{y \in \mathbb{R}^{n+1}} \left\{ \sum_{i=1}^n v_i(y_i) + \lambda y^t K_\sigma^{-1} y \right\} &\rightarrow \inf_{c_J: |J| \leq \lfloor p \rfloor} \left\{ \sum_{i=1}^n v_i \left( \sum_J c_J X_i^J \right) \right\} \\ &\text{or} \\ &\rightarrow \inf_{c_J: |J| \leq p} \left\{ \sum_{i=1}^n v_i \left( \sum_J c_J X_i^J \right) + g_p(0)^{-2} \sum_{|J|=p} \binom{p}{J}^{-1} c_J^2 \right\} \end{aligned}$$

$$\text{with } r(x) = y_0 \rightarrow \sum_J c_J x^J.$$

## Conclusions

- fixed degree polynomial approximation lies on the frontier of Tikhonov regularization problems
- a general phenomenon for common (convex, lsc) loss terms
- a wide variety of spd kernel functions have the pd Taylor expansions
- suggests why popular approximation ideas for  $K_\sigma$  work well for large  $\sigma$

## Future work

- value-based formulations of Tikhonov regularization problems simplify other results, e.g. problems with parametrized kernel functions
- epi-convergence could be used to study linear systems  $A(s)x = b(s)$  where  $A(s)$  is becoming singular.

## Acknowledgements

- Ryan Rifkin, my collaborator
- Adrian Lewis and Michel Goemans for answering convex analysis questions
- Roger Wets for explaining how epi-convergence works