# Discretization Error Analysis for Statistical Inverse Problems

Nicole Mücke
(Potsdam)

October 11, 2013

# Contents

# Introduction: Linear Inverse Problems with Random Noise

- fields of application:
  - astronomy (blurred images of the Hubble telescope)
  - econometrics (instrumental variables)
  - financial mathematics (model calibration of the volatility)
  - medical image processing (X-ray tomography)

# Introduction: Linear Inverse Problems with Random Noise

- fields of application:
  - astronomy (blurred images of the Hubble telescope)
  - econometrics (instrumental variables)
  - financial mathematics (model calibration of the volatility)
  - medical image processing (X-ray tomography)
- classical inverse problem:

  $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces (of functions)
  $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ linear and bounded
  given the datum $g \in \mathcal{H}_2$, find $f \in \mathcal{H}_1$ such that

$$Af = g \tag{1}$$

# Introduction: Linear Inverse Problems with Random Noise

- fields of application:
  - astronomy (blurred images of the Hubble telescope)
  - econometrics (instrumental variables)
  - financial mathematics (model calibration of the volatility)
  - medical image processing (X-ray tomography)
- classical inverse problem:
  $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces (of functions)
  $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ linear and bounded
  given the datum $g \in \mathcal{H}_2$, find $f \in \mathcal{H}_1$ such that

$$Af = g \tag{1}$$

- mathematically: to solve inverse problem means to invert operator $A$

# Example: Differentiation

- consider problem of recovering the derivative of a function
- setting:
  $\mathcal{H}_1 = H^1([0,1])$ Sobolev space of continuous functions with weak derivative in $L^2([0,1], dx)$
  $\mathcal{H}_2 = L^2([0,1], dx)$

# Example: Differentiation

- consider problem of recovering the derivative of a function
- setting:
  $\mathcal{H}_1 = H^1([0,1])$ Sobolev space of continuous functions with weak derivative in $L^2([0,1], dx)$
  $\mathcal{H}_2 = L^2([0,1], dx)$
- define $A : H^1([0,1]) \longrightarrow L^2([0,1], dx)$ as

$$(Af)(x) = \int_0^x f(t)\, dt \qquad (x \in [0,1])$$

- then $Af = g \iff f = g'$ for any $g \in Ran(A)$

- problem: $A$ may not be invertable (problem is ill-posed)
- way out: use regularization method to obtain approximate solution

- problem: $A$ may not be invertable (problem is ill-posed)
- way out: use regularization method to obtain approximate solution
- in practise, inverse problems further compounded by some additive noise

$$Y = Af + \text{"}noise\text{"}$$

- problem: $A$ may not be invertable (problem is ill-posed)
- way out: use regularization method to obtain approximate solution
- in practise, inverse problems further compounded by some additive noise

$$Y = Af + \text{"}noise\text{"}$$

- three frequently used noise models: deterministic, as a Hilbert space process, as a random variable
- in what follows, noise is modeled as a r.v. with values in $\mathcal{H}_2$

- problem: $A$ may not be invertable (problem is ill-posed)
- way out: use regularization method to obtain approximate solution
- in practise, inverse problems further compounded by some additive noise

$$Y = Af + \text{"}noise\text{"}$$

- three frequently used noise models: deterministic, as a Hilbert space process, as a random variable
- in what follows, noise is modeled as a r.v. with values in $\mathcal{H}_2$
- study statistical regularization methods, leading to a meaningful reconstruction despite the noise and ill-posedness
- References: Engl et al. (2000), Evans/ Stark (2002), Mathé/ Pereverzev (2006), Bissantz/ Holzmann (2008)

# The Model: Nonparametric Inverse Regression

- two sets of variables:
  - input space $\mathcal{X} \subset \mathbb{R}^d$, compact (for simplicity d=1)
  - output space $\mathcal{Y} \subseteq [-M, M] \subset R$
  - build sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

# The Model: Nonparametric Inverse Regression

- two sets of variables:
  input space $\mathcal{X} \subset \mathbb{R}^d$, compact (for simplicity d=1)
  output space $\mathcal{Y} \subseteq [-M, M] \subset R$
  build sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces consisting of functions on $\mathcal{X}$
- $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ linear, bounded and injective

# The Model: Nonparametric Inverse Regression

- two sets of variables:
  input space $\mathcal{X} \subset \mathbb{R}^d$, compact (for simplicity d=1)
  output space $\mathcal{Y} \subseteq [-M, M] \subset R$
  build sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces consisting of functions on $\mathcal{X}$
- $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ linear, bounded and injective
- unknown Borel probability measure

$$\rho(x, y) \;=\; \rho(y|x)\, \rho_X(x)$$

- r.v. $(X_i, Y_i) \sim (X, Y)$ drawn i.i.d. according to $\rho$, $(i = 1, ..., n)$

# The Model: Nonparametric Inverse Regression

- two sets of variables:
  input space $\mathcal{X} \subset \mathbb{R}^d$, compact (for simplicity d=1)
  output space $\mathcal{Y} \subseteq [-M, M] \subset R$
  build sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces consisting of functions on $\mathcal{X}$
- $A : \mathcal{H}_1 \longrightarrow \mathcal{H}_2$ linear, bounded and injective
- unknown Borel probability measure

$$\rho(x, y) \;=\; \rho(y|x)\, \rho_X(x)$$

- r.v. $(X_i, Y_i) \sim (X, Y)$ drawn i.i.d. according to $\rho$, $(i = 1, ..., n)$
- model

$$Y_i \;=\; (Af)(X_i) \;+\; E_i$$

with

$$(Af)(x) \;=\; \mathbb{E}[Y|X = x] \;=\; \int_{\mathcal{Y}} y\rho(dy|x)\,, \qquad x \in \mathcal{X}$$

and hence, the unobservable errors $E_i \sim E$ i.i.d. satisfy

$$\mathbb{E}[E|X = x] = 0$$

- aim: find an estimator $f_{\mathbf{z}}^{\lambda} : \mathcal{X} \longrightarrow \mathbb{R}$ in $\mathcal{H}_1$, based on a given sample $\mathbf{z} \in \mathcal{Z}^n$ and on a parameter $\lambda_n > 0$ such that

$$\rho^{\otimes n} \left( \| f - f_{\mathbf{z}}^{\lambda} \|_{\mathcal{H}_1} \leq \delta(n, \eta) \right) \geq 1 - \eta$$

  for some small $\delta = \delta(n, \eta)$, depending on sample size $n$ and confidence level $\eta \in (0, 1)$
- furthermore, $f_{\mathbf{z}}^{\lambda}$ should approximate true function $f$ more accurate as sample size $n$ grows (consistency)
- but some problem:

- aim: find an estimator $f_{\mathbf{z}}^{\lambda} : \mathcal{X} \longrightarrow \mathbb{R}$ in $\mathcal{H}_1$, based on a given sample $\mathbf{z} \in \mathcal{Z}^n$ and on a parameter $\lambda_n > 0$ such that

$$\rho^{\otimes n} \left( \| f - f_{\mathbf{z}}^{\lambda} \|_{\mathcal{H}_1} \leq \delta(n, \eta) \right) \geq 1 - \eta$$

  for some small $\delta = \delta(n, \eta)$, depending on sample size $n$ and confidence level $\eta \in (0, 1)$
- furthermore, $f_{\mathbf{z}}^{\lambda}$ should approximate true function $f$ more accurate as sample size $n$ grows (consistency)
- but some problem:
- need to assume some prior on the problem: restrict class of admissible Borel prob. distributions and consider for given set $\Omega \subset \mathcal{H}_1$

$$\mathcal{M}^A(\Omega) = \{\rho \ : \ \exists f_\rho \in \Omega : Af_\rho(x) = \mathbb{E}[Y|X = x] \quad \rho_X - a.s.\}$$

- usually $\Omega$ (pre-) compact set

to find such an estimator, we proceed in three steps

1. discretization
2. symmetrization
3. regularization

# Random Discretization



## Assumption:

For any input $x \in \mathcal{X}$ we let the evaluation functional

$$\begin{aligned} S_x^A : \mathcal{H}_1 &\longrightarrow \mathbb{R} \\ f &\longmapsto (Af)(x) \end{aligned}$$

be uniformly (w.r.t. $x \in \mathcal{X}$) bounded, that is there exists a constant $C > 0$ (independent of $x \in \mathcal{X}$) such that

$$|S_x^A(f)| \leq C||f||_{\mathcal{H}_1} \quad .$$

- implies that $Ran(A)$ is a reproducing kernel Hilbert space which can be included in $L^2(\mathcal{X}, \rho_X)$

$$S^A := T \circ A : \mathcal{H}_1 \longrightarrow Ran(A) \hookrightarrow L^2(\mathcal{X}, \rho_X)$$

- implies that $Ran(A)$ is a reproducing kernel Hilbert space which can be included in $L^2(\mathcal{X}, \rho_X)$

$$S^A := T \circ A : \mathcal{H}_1 \longrightarrow Ran(A) \hookrightarrow L^2(\mathcal{X}, \rho_X)$$

- composition $S^A = T \circ A$ is a Carleman operator and we now consider the population case

$$S^A f = g$$

- implies that $Ran(A)$ is a reproducing kernel Hilbert space which can be included in $L^2(\mathcal{X}, \rho_X)$

$$S^A := T \circ A : \mathcal{H}_1 \longrightarrow Ran(A) \hookrightarrow L^2(\mathcal{X}, \rho_X)$$

- composition $S^A = T \circ A$ is a Carleman operator and we now consider the population case

$$S^A f = g$$

- in Caponnetto et al. (2004), a discretization strategy is introduced in the framework of learning theory where $A = Id$ and for inverse problems induced by Carleman operators where the noise is modeled deterministic

- implies that $Ran(A)$ is a reproducing kernel Hilbert space which can be included in $L^2(\mathcal{X}, \rho_X)$

$$S^A := T \circ A : \mathcal{H}_1 \longrightarrow Ran(A) \hookrightarrow L^2(\mathcal{X}, \rho_X)$$

- composition $S^A = T \circ A$ is a Carleman operator and we now consider the population case

$$S^A f = g$$

- in Caponnetto et al. (2004), a discretization strategy is introduced in the framework of learning theory where $A = Id$ and for inverse problems induced by Carleman operators where the noise is modeled deterministic
- adapt this approach and develop a framework to deal simultaneously with the perturbation due to the noise and to the random discretization

- define sampling operator $S_{\mathbf{x}}^{A} : \mathcal{H}_1 \longrightarrow \mathbb{R}^n$ by

$$(S_{\mathbf{x}}^{A} f)_i = (Af)(x_i) \qquad i = 1, \ldots, n \,, \; f \in \mathcal{H}_1$$

- define sampling operator $S_{\mathbf{x}}^A : \mathcal{H}_1 \longrightarrow \mathbb{R}^n$ by

$$(S_{\mathbf{x}}^A f)_i = (Af)(x_i) \qquad i = 1, ..., n \ , \ f \in \mathcal{H}_1$$

- leads to discretized counterpart (sample case)

$$S_{\mathbf{x}}^A f \ = \ \mathbf{y}$$

- true prob. measure is replaced by empirical measure on the sample

- aim: want to compare the population case and the sample case

- aim: want to compare the population case and the sample case
- symmetrized versions are

$$(S_{\mathbf{x}}^A)^\star S_{\mathbf{x}}^A f = (S_{\mathbf{x}}^A)^\star \mathbf{y} \tag{2}$$

and

$$(S^A)^\star S^A f = (S^A)^\star g \tag{3}$$

- advantage: we are now dealing with operators acting on $\mathcal{H}_1$, which can be more easily compared

# Concentration Inequalities:

- sample case can be seen as perturbation (due to random discretization) of population case

## Proposition:

For $n \in \mathbb{N}$ and $0 < \eta < 1$ it holds with probability at least $1 - \eta$

$$|| \, (S_{\mathbf{x}}^{A})^{\star}\mathbf{y} \; - \; (S_{\mathbf{x}}^{A})^{\star} g \, ||_{\mathcal{H}_{\mathbf{1}}} \; \leq \; C_1(\eta) \, \frac{1}{\sqrt{n}}$$

and

# Concentration Inequalities:

- sample case can be seen as perturbation (due to random discretization) of population case

## Proposition:

For $n \in \mathbb{N}$ and $0 < \eta < 1$ it holds with probability at least $1 - \eta$

$$|| \, (S_{\mathbf{x}}^A)^\star \mathbf{y} \; - \; (S_{\mathbf{x}}^A)^\star g \, ||_{\mathcal{H}_1} \; \leq \; C_1(\eta) \, \frac{1}{\sqrt{n}}$$

and

$$|| \, (S^A)^\star S^A \; - \; (S_{\mathbf{x}}^A)^\star S_{\mathbf{x}}^A \, ||_{\mathcal{L}} \; \leq \; C_2(\eta) \, \frac{1}{\sqrt{n}} \, ,$$

for some $C_1(\eta) > 0$, $C_2(\eta) > 0$.

- similiar result is obtained in the framework of learning theory ($A = Id$) in Caponnetto (2004) and Bauer et al. (2009)

# Regularization

- in context of inverse problems: regularization leads to algorithms for determining approximate solutions to ill-posed problems which are stable with respect to noise

  in deterministic setting see e.g. Bertero et al. (1985, 1988), Engl et al. (2000)

# Regularization

- in context of inverse problems: regularization leads to algorithms for determining approximate solutions to ill-posed problems which are stable with respect to noise

  in deterministic setting see e.g. Bertero et al. (1985, 1988), Engl et al. (2000)

- in context of statistics: regularization refers to techniques allowing to avoid overfitting

  focus mostly on Tikhonov regularization, also called ridge regression/ penalized least squares

  see e.g. Hastie et al. (2001), Wahba (1990)

- key idea: consider family of regularized solutions

$$f_{\mathbf{z}}^{\lambda} = g_{\lambda}(\ (S_{\mathbf{x}}^{A})^{\star} S_{\mathbf{x}}^{A}\ )(S_{\mathbf{x}}^{A})^{\star}\mathbf{y} \tag{4}$$

depending on the regularization parameter $\lambda > 0$ in such a way that

- key idea: consider family of regularized solutions

$$f_{\mathbf{z}}^{\lambda} = g_{\lambda}(\ (S_{\mathbf{x}}^{A})^{\star} S_{\mathbf{x}}^{A}\ )(S_{\mathbf{x}}^{A})^{\star}\mathbf{y} \tag{4}$$

depending on the regularization parameter $\lambda > 0$ in such a way that

1. the map

$$t \mapsto g_{\lambda}(t)$$

approximates the function $t \mapsto 1/t$ as $\lambda \to 0$
implies: $g_{\lambda}(\ (S_{\mathbf{x}}^{A})^{\star} S_{\mathbf{x}}^{A}\ )$ is a family of operators approximating the inverse of $(S_{\mathbf{x}}^{A})^{\star} S_{\mathbf{x}}^{A}$ as $\lambda \to 0$

- key idea: consider family of regularized solutions

$$f_{\mathbf{z}}^{\lambda} = g_{\lambda}(\ (S_{\mathbf{x}}^{A})^{\star} S_{\mathbf{x}}^{A}\ )(S_{\mathbf{x}}^{A})^{\star}\mathbf{y} \tag{4}$$

depending on the regularization parameter $\lambda > 0$ in such a way that

1. the map
$$t \mapsto g_{\lambda}(t)$$
approximates the function $t \mapsto 1/t$ as $\lambda \to 0$
implies: $g_{\lambda}(\ (S_{\mathbf{x}}^{A})^{\star} S_{\mathbf{x}}^{A}\ )$ is a family of operators approximating the inverse of $(S_{\mathbf{x}}^{A})^{\star} S_{\mathbf{x}}^{A}$ as $\lambda \to 0$
2. the function
$$g_{\lambda} : \sigma((S_{\mathbf{x}}^{A})^{\star} S_{\mathbf{x}}^{A}) \longrightarrow \mathbb{R}$$
is bounded
then the spectral theorem ensures that $g_{\lambda}(\ (S_{\mathbf{x}}^{A})^{\star} S_{\mathbf{x}}^{A}\ )$ is a bounded operator

- estimator $f_{\mathbf{z}}^{\lambda}$ depends on given sample $\mathbf{z} \in \mathcal{Z}^n$ (in particular on sample size $n$) and on regularization parameter $\lambda > 0$
- difficulty in applying this method: choice of parameter $\lambda > 0$
- final estimator defined with parameter choice rule either
  a-priori: $\lambda = \lambda_n$ or
  a-posteriori: $\lambda = \lambda_{\mathbf{z},n}$

# Rate of Convergence

- aim: study reconstruction error

$$|| f - f_{\mathbf{z}}^{\lambda} ||_{\mathcal{H}_1}$$

and investigate rate of convergence of estimated solution to true function as the sample size $n \to \infty$

# Rate of Convergence

- aim: study reconstruction error

$$|| f - f_{\mathbf{z}}^{\lambda} ||_{\mathcal{H}_1}$$

  and investigate rate of convergence of estimated solution to true function as the sample size $n \to \infty$
- recall: need to impose some source condition

$$f \in \Omega_{r,R} := \{ f \in \mathcal{H}_1 : f = ((S^A)^\star S^A)^r h , \ ||h||_{\mathcal{H}_1} \leq R \}$$

  with $0 < r$, $0 < R$

# Non-asymptotic upper Bound:

## Proposition:

Let $\lambda > 0$ and $\rho \in \mathcal{M}(\Omega_{r,R})$ with $0 < r \leq q$. Then with probability at least $1 - \eta$

$$|| f - f_{\mathbf{z}}^{\lambda} ||_{\mathcal{H}_1} \leq C_{r,R} C_{\eta} \left( \lambda^r + \frac{1}{\lambda\sqrt{n}} \right) , \tag{5}$$

for some $C_{r,R} > 0$ and $C_{\eta} > 0$.

# Non-asymptotic upper Bound:

## Proposition:

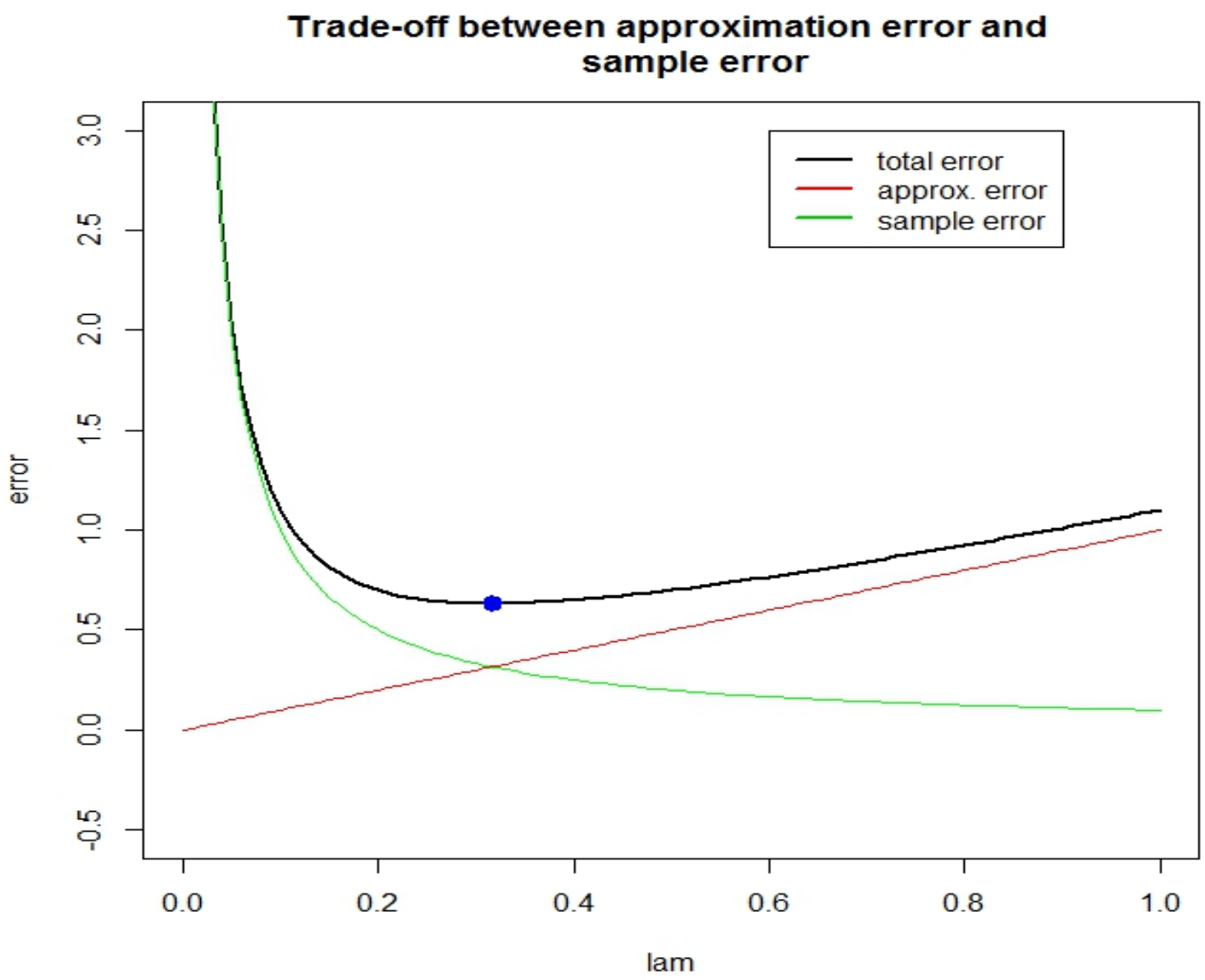
Let $\lambda > 0$ and $\rho \in \mathcal{M}(\Omega_{r,R})$ with $0 < r \leq q$. Then with probability at least $1 - \eta$

$$|| f - f_{\mathbf{z}}^{\lambda} ||_{\mathcal{H}_1} \leq C_{r,R} C_{\eta} \left( \lambda^r + \frac{1}{\lambda \sqrt{n}} \right) , \tag{5}$$

for some $C_{r,R} > 0$ and $C_{\eta} > 0$.

- first term in (5) corresponds to approximation error while the second corresponds to sample error
- aim: find value of $\lambda$ (a-priori) balancing out the trade-off

# Trade- Off


Trade-off between approximation error and sample error
total error
approx. error
sample error
error
lam
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0

# Rate of Convergence and Consistency:

## Proposition:

Under the assumptions made above, choose

$$\lambda_n = \left(\frac{1}{\sqrt{n}}\right)^{\frac{1}{r+1}} .$$

Then

# Rate of Convergence and Consistency:

## Proposition:

Under the assumptions made above, choose

$$\lambda_n = \left(\frac{1}{\sqrt{n}}\right)^{\frac{1}{r+1}} .$$

Then

$$\lim_{\tau\to\infty} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}^{A}(\Omega_{r,R})} \rho^{\otimes n}( \| f - f_{z}^{\lambda_n} \|_{\mathcal{H}_1} > \tau\, a_n ) = 0 ,$$

with

$$a_n = \left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+1}} .$$

# Rate of Convergence and Consistency:

## Proposition:

Under the assumptions made above, choose

$$\lambda_n = \left(\frac{1}{\sqrt{n}}\right)^{\frac{1}{r+1}} .$$

Then

$$\lim_{\tau\to\infty} \limsup_{n\to\infty} \sup_{\rho\in\mathcal{M}^{A}(\Omega_{r,R})} \rho^{\otimes n}( \| f - f_z^{\lambda_n} \|_{\mathcal{H}_1} > \tau\, a_n ) = 0 ,$$

with

$$a_n = \left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+1}} .$$

- the bigger the smoothness parameter $r \in (0, q]$ the faster the rate of convergence
- parameter $q$ called qualification of method $g_\lambda$

# Enjoy your meal !