Approximation theoretic advice for supervised learning

PAUL CONSTANTINE
Assistant Professor
Department of Computer Science
University of Colorado, Boulder

activesubspaces.org
@DrPaulynomial

SLIDES AVAILABLE UPON REQUEST

DISCLAIMER: These slides are meant to complement the oral presentation. Use out of context at your own risk.
Even more understanding is lost if we consider each thing we can do to data only in terms of some set of very restrictive assumptions under which that thing is best possible—assumptions we know we CANNOT check in practice.
data
model
noise
parameter
error
uncertainty
overfit
Selected regression / approximation / UQ literature
My personal bibliography


Nobile, Tempone, and Webster, *A sparse grid stochastic collocation method* (SINUM, 2008)


Jones, *A taxonomy of global optimization methods based on response surfaces* (JGO, 2001)

Regression

\[ \{x_i, y_i\} \]
\[ \pi(x, y) \]

**GIVEN**

i.i.d. samples \( \{x_i, y_i\} \)

from **unknown** \( \pi(x, y) \)

**GOAL**

statistically characterize \( y \mid x \)

e.g., \( \mathbb{E}[y \mid x] \), \( \text{Var}[y \mid x] \)
Regression

GIVEN
i.i.d. samples \( \{x_i, y_i\} \)
from unknown \( \pi(x, y) \)

GOAL
statistically characterize \( y \mid x \)
e.g., \( \mathbb{E}[y \mid x], \text{Var}[y \mid x] \)

\( \mathbf{MODE}_{\text{L}} \) (e.g., polynomials)

\[ y = p(x, \theta) + \varepsilon \]
modeled r.v., zero-mean, independent of \( x \)

\[ \mathbb{E}[y \mid x] \]
Regression

\[
\{ x_i, y_i \} \stackrel{\text{\(\sim\)}}{\sim} (x, y) \text{ i.i.d. samples from unknown \(\pi(x, y)\)}
\]

**GIVEN**

i.i.d. samples \(\{ x_i, y_i \}\) from **unknown** \(\pi(x, y)\)

**GOAL**

statistically characterize \(y \mid x\)  

\( \text{e.g., } \mathbb{E}[y \mid x], \text{Var}[y \mid x] \)
Regression

i.i.d. samples \( \{x_i, y_i\} \)

from unknown \( \pi(x, y) \)

GOAL

statistically characterize \( y \mid x \)

e.g., \( \mathbb{E}[y \mid x], \text{Var}[y \mid x] \)

MODEL (e.g., polynomials)

\[
y = p(x, \theta) + \varepsilon \quad \text{modeled r.v., zero-mean, independent of } x
\]

\[\mathbb{E}[y \mid x]\]

FIT (e.g., max likelihood)

\[
\hat{\theta} = \arg\min_{\theta} \sum_i (y_i - p(x_i, \theta))^2
\]

PREDICT

\[
\mathbb{E}[y \mid x^*] \approx p(x^*, \hat{\theta}) = \hat{p}(x^*)
\]
Regression

MODELS (e.g., polynomials)
\[ y = p(x, \theta) + \varepsilon \]
- modeled r.v.,
- zero-mean,
- independent of \( x \)

\[ \mathbb{E}[y|x] \]

FIT (e.g., max likelihood)
\[ \hat{\theta} = \arg\min_{\theta} \sum_i (y_i - p(x_i, \theta))^2 \]

PREDICT
\[ \mathbb{E}[y|x^*] \approx p(x^*, \hat{\theta}) = \hat{p}(x^*) \]

QUANTIFY UNCERTAINTY
\[ \text{Var}[y|x^*] \approx \text{“formula”} \]

GIVEN
i.i.d. samples \( \{x_i, y_i\} \)
from \textbf{unknown} \( \pi(x, y) \)

GOAL
statistically characterize \( y|x \)
e.g., \( \mathbb{E}[y|x], \text{Var}[y|x] \)
Regression

Approximation
Approximation

Does a **unique, best** approximation exist?

\[ p^* = \arg\min_{p \in \mathcal{P}_n} \| p - f \| \]

How does the best error behave?

\[ \| p^* - f \| = e^*(n) \]

Can we construct an approximation?

**Algorithm:** Given \( f \), compute \( \hat{p} \)

And analyze its error?

\[ \| \hat{p} - f \| \leq C e^*(n) \]
Approximation

GIVEN

a function \( f(x) \)

a known density \( \pi(x) \)

GOAL

find \( \hat{p}(x) \) such that

the error \( ||\hat{p} - f|| \) is small
CONSTRUCTION

choose $x_i$

GIVEN

a function $f(x)$

a **known** density $\pi(x)$

GOAL

find $\hat{p}(x)$ such that

the error $\|\hat{p} - f\|$ is small
CONSTRUCTION

choose $x_i$

compute $y_i = f(x_i)$

GIVEN

a function $f(x)$

a **known** density $\pi(x)$

GOAL

find $\hat{p}(x)$ such that

the error $\|\hat{p} - f\|$ is small
CONSTRUCTION

choose $x_i$

compute $y_i = f(x_i)$

fit $\hat{p} = \arg\min_{p \in \mathcal{P}_n} \sum_i (y_i - p(x_i))^2$

GIVEN

a function $f(x)$

a known density $\pi(x)$

GOAL

find $\hat{p}(x)$ such that

the error $\|\hat{p} - f\|$ is small
\[ \hat{\theta} = \arg\min_{\theta} \sum_{i} (y_i - p(x_i, \theta))^2 \]

\[ \hat{p} = \arg\min_{p \in \mathcal{P}_n} \sum_{i} (y_i - p(x_i))^2 \]
The story of the data and fitted curve is different. But does it matter? YES
REGRESSION VS. APPROXIMATION

What is *error*?
Regression

\[ \hat{p}(x) \pm 2\widehat{\text{se}}[y \mid x] \]

Confidence interval

plug-in estimate of standard error

Approximation

Approximation error

\[ |\hat{p}(x) - f(x)| \]

Error norms

\[ \left( \int |\hat{p}(x) - f(x)|^2 \pi(x) \, dx \right)^{1/2} \]

\[ \sup_x |\hat{p}(x) - f(x)| \]
REGRESSION VS. APPROXIMATION

What is convergence?
As data increases, \textit{root-n consistency}

\[
\hat{\theta} \to \theta \quad \hat{p}(x) \to p(x)
\]

\{ “true” parameters \}
As data increases, *root-n consistency*

\[
\hat{\theta} \rightarrow \theta \\
\hat{p}(x) \rightarrow p(x) \quad \left\{ \text{"true" parameters} \right\}
\]
Regression

Approximation

As data increases, *root-n consistency*

\[
\hat{\theta} \rightarrow \theta \quad \hat{p}(x) \rightarrow p(x)
\]

“true” parameters

As the approximation class grows

\[
\| \hat{p}(x) - f(x) \| \rightarrow 0
\]
Regression

As data increases, \textit{root-n consistency}
\[
\begin{align*}
\hat{\theta} & \to \theta \\
\hat{p}(x) & \to p(x)
\end{align*}
\]
\text{“true” parameters}

Approximation

As the approximation class grows
\[
\|\hat{p}(x) - f(x)\| \to 0
\]
Regression

As data increases, root-n consistency

\[
\hat{\theta} \to \theta \\
\hat{p}(x) \to p(x)
\]

"true" parameters

Approximation

As the approximation class grows

\[
\| \hat{p}(x) - f(x) \| \to 0
\]
Regression

Approximation

As data increases, *root-n consistency*

\[
\hat{\theta} \rightarrow \theta
\]

\[
\hat{p}(x) \rightarrow p(x)
\]

\[
\text{“true” parameters}
\]

As the approximation class grows

\[
\| \hat{p}(x) - f(x) \| \rightarrow 0
\]

Convergence rate depends on \( f(x) \)

- high order derivatives
- size of region of analyticity
- Chebyshev coefficients
- …
Gaussian process regression

Radial basis approximation
Gaussian process regression

GIVEN

pairs \( \{x_i, y_i\} \)
Gaussian process regression

**GIVEN**

pairs \( \{x_i, y_i\} \)

**ASSUME**

\[ y_i = g(x_i, \omega) \]

one realization of a GP
Gaussian process regression

GIVEN

pairs \( \{x_i, y_i\} \)

ASSUME

\[ y_i = g(x_i, \omega) \quad \text{one realization of a GP} \]

\[ y_i = g(x_i, \cdot) \quad \text{among many possible realizations} \]
Gaussian process regression

CORRELATION MODEL (e.g.)

\[ \kappa(x, x'; \theta) = \exp\left(\frac{|x - x'|^2}{\theta}\right) \]

**GIVEN**

pairs \( \{x_i, y_i\} \)

**ASSUME**

\[ y_i = g(x_i, \omega) \quad \text{one realization of a GP} \]

\[ y_i = g(x_i, \cdot) \quad \text{among many possible realizations} \]
Gaussian process regression

**CORRELATION MODEL** (e.g.)

\[ \kappa(x, x'; \theta) = \exp(|x - x'|^2 / \theta) \]

**GIVEN**

pairs \( \{x_i, y_i\} \)

**ASSUME**

\( y_i = g(x_i, \omega) \) one realization of a GP

\( y_i = g(x_i, \cdot) \) among many possible realizations
Gaussian process regression

**CORRELATION MODEL** (e.g.)

\[
\kappa(x, x'; \theta) = \exp(|x - x'|^2 / \theta)
\]

**GIVEN**

pairs \( \{x_i, y_i\} \)

**ASSUME**

\( y_i = g(x_i, \omega) \)  one realization of a GP

\( y_i = g(x_i, \cdot) \)  among many possible realizations
Gaussian process regression

**GIVEN**

pairs \( \{ x_i, y_i \} \)

**ASSUME**

\[ y_i = g(x_i, \omega) \] one realization of a GP

\[ y_i = g(x_i, \cdot) \] among many possible realizations

**CORRELATION MODEL** (e.g.)

\[ \kappa(x, x'; \theta) = \exp(|x - x'|^2 / \theta) \]

**FIT**

maximize \( \theta \) likelihood\( (\theta; \{ x_i, y_i \}) \)
Gaussian process regression

**CORRELATION MODEL** (e.g.)
\[ \kappa(x, x'; \theta) = \exp(|x - x'|^2 / \theta) \]

**FIT**
\[
\text{maximize } \text{likelihood}(\theta; \{x_i, y_i\})
\]

**PREDICT** (B.L.U.E.)
\[
y(x) = \mathbb{E}[g(x, \cdot) | \{x_i, y_i\}] = \sum_i y_i w_i(x)
\]

**GIVEN**
pairs \( \{x_i, y_i\} \)

**ASSUME**
\( y_i = g(x_i, \omega) \) one realization of a GP

\( y_i = g(x_i, \cdot) \) among many possible realizations
Gaussian process regression

![Diagram showing Gaussian process regression](image)

**CORRELATION MODEL (e.g.)**

\[ \kappa(x, x'; \theta) = \exp(|x - x'|^2 / \theta) \]

**FIT**

maximize \( \theta \) likelihood(\( \theta \); \( \{x_i, y_i\} \))

**PREDICT (B.L.U.E.)**

\[ y(x) = \mathbb{E}[g(x, \cdot) | \{x_i, y_i\}] = \sum_i y_i \omega_i(x) \]

**QUANTIFY UNCERTAINTY**

\[ \text{Var}[g(x, \cdot) | \{x_i, y_i\}] = \text{“formula”} \]

**GIVEN**

pairs \( \{x_i, y_i\} \)

**ASSUME**

\[ y_i = g(x_i, \omega) \] one realization of a GP

\[ y_i = g(x_i, \cdot) \] among many possible realizations
Gaussian process regression

\[ y_i = g(x_i, \omega) \quad \text{one realization of a GP} \]

\[ y_i = g(x_i, \cdot) \quad \text{among many possible realizations} \]

**GIVEN**
pairs \( \{x_i, y_i\} \)

**ASSUME**
\[ y_i = g(x_i, \omega) \quad \text{one realization of a GP} \]

**CORRELATION MODEL (e.g.)**
\[ \kappa(x, x'; \theta) = \exp(|x - x'|^2 / \theta) \]

**FIT**
maximize \( \theta \)
likelihood\( (\theta; \{x_i, y_i\}) \)

**PREDICT (B.L.U.E.)**
\[ y(x) = \mathbb{E}[g(x, \cdot) | \{x_i, y_i\}] \]
\[ = \sum_i y_i w_i(x) \]

**QUANTIFY UNCERTAINTY**
\[ \text{Var}[g(x, \cdot) | \{x_i, y_i\}] = \text{“formula”} \]
Gaussian process regression

\[ y_i = g(x_i, \omega) \] one realization of a GP

\[ y_i = g(x_i, \cdot) \] among many possible realizations

**GIVEN**

pairs \( \{x_i, y_i\} \)

**ASSUME**

\[ y_i = g(x_i, \omega) \] one realization of a GP

\[ y_i = g(x_i, \cdot) \] among many possible realizations

**CORRELATION MODEL (e.g.)**

\[ \kappa(x, x'; \theta) = \exp(|x - x'|^2 / \theta) \]

**FIT**

maximize likelihood(\( \theta; \{x_i, y_i\} \))

**PREDICT (B.L.U.E.)**

\[ y(x) = \mathbb{E}[g(x, \cdot) | \{x_i, y_i\}] = \sum_i y_i w_i(x) \]

**QUANTIFY UNCERTAINTY**

\[ \text{Var}[g(x, \cdot) | \{x_i, y_i\}] = \text{“formula”} \]
Gaussian process regression

Radial basis approximation
Radial basis approximation

GIVEN

a queryable function $f(x)$
Radial basis approximation

**GIVEN**

a queryable function \( f(x) \)

centers \( x_1, \ldots, x_n \)
Radial basis approximation

**GIVEN**

a queryable function \( f(x) \)

centers \( x_1, \ldots, x_n \)

**GOAL**

find \( s(x) \) such that

the error \( \| s - f \| \) is small
QUERY THE FUNCTION

\[ y_i = f(x_i) \]

GIVEN

a queryable function \( f(x) \)

centers \( x_1, \ldots, x_n \)

GOAL

find \( s(x) \) such that

the error \( \| s - f \| \) is small
QUERY THE FUNCTION
\[ y_i = f(x_i) \]

CHOOSE KERNEL
\[ \kappa(x, x'; \varepsilon) = \exp(\varepsilon |x - x'|^2) \]

GIVEN a queryable function \( f(x) \)
centers \( x_1, \ldots, x_n \)

GOAL find \( s(x) \) such that
the error \( ||s - f|| \) is small
Radial basis approximation

QUERY THE FUNCTION
\[ y_i = f(x_i) \]

CHOOSE KERNEL
\[ \kappa(x, x'; \varepsilon) = \exp(\varepsilon |x - x'|^2) \]

DEFINES BASIS
\[ \phi_i(x) = \kappa(x, x_i; \varepsilon) \]

GIVEN
a queryable function \( f(x) \)
centers \( x_1, \ldots, x_n \)

GOAL
find \( s(x) \) such that
the error \( \|s - f\| \) is small
QUERY THE FUNCTION

\[ y_i = f(x_i) \]

CHOOSE KERNEL

\[ \kappa(x, x'; \varepsilon) = \exp(\varepsilon |x - x'|^2) \]

DEFINES BASIS

\[ \phi_i(x) = \kappa(x, x_i; \varepsilon) \]

COMPUTE COEFFICIENTS

\[ K a = f \]

\[ K_{ij} = \kappa(x_i, x_j), \ f_i = f(x_i) \]

GIVEN

a queryable function \( f(x) \)

centers \( x_1, \ldots, x_n \)

GOAL

find \( s(x) \) such that

the error \( \|s - f\| \) is small
QUERY THE FUNCTION
\[ y_i = f(x_i) \]

CHOOSE KERNEL
\[ \kappa(x, x'; \varepsilon) = \exp(\varepsilon |x - x'|^2) \]

DEFINES BASIS
\[ \phi_i(x) = \kappa(x, x_i; \varepsilon) \]

COMPUTE COEFFICIENTS
\[ K a = f \]
\[ K_{ij} = \kappa(x_i, x_j), \; f_i = f(x_i) \]

PREDICT
\[ s(x) = \sum_i a_i \phi_i(x) \]

Radial basis approximation

GIVEN
a queryable function \( f(x) \)
centers \( x_1, \ldots, x_n \)

GOAL
find \( s(x) \) such that
the error \( \|s - f\| \) is small
Gaussian process regression

Radial basis approximation
The **story** of the data and fitted curve is **different**. 
*But does it matter? YES*
Comments on error and convergence

GP conditional variance is **NOT** error (except possibly under some very specific conditions).

RBF error estimates are *asymptotic* in the *fill distance*.

Both approaches make practically unverifiable assumptions about the origin of the data generating function/process.

As a caricature:

- statisticians try to reduce the error by finding a better model (e.g., solve the fitting problem better)
- mathematicians try to reduce the error with more queries (i.e., sample into asymptopia)
Which one is a computer simulation?
What is error in a computer simulation?

Sacks et al. (1989)

Design and Analysis of Computer Experiments
Jerome Sacks, William J. Welch, Toby J. Mitchell and Henry P. Wynn

Abstract. Many scientific phenomena are now investigated by complex computer models or codes. A computer experiment is a number of runs of the code with various inputs. A feature of many computer experiments is that the output is deterministic—rerunning the code with the same inputs gives identical observations. Often, the codes are computationally expensive to run, and a common objective of an experiment is to fit a cheaper predictor of the output to the data. Our approach is to model the deterministic output as the realization of a stochastic process, thereby providing a statistical basis for designing experiments (choosing the inputs) for efficient prediction. With this model, estimates of uncertainty of predictions are also available. Recent work in this area is reviewed, a number of applications discussed, and the extension to multiprocessor computer models considered.

Kennedy and O’Hagan (2001)

Bayesian calibration of computer models
Marc C. Kennedy and Anthony O’Hagan
University of Sheffield, UK

[Read before The Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 13th, 2000, Professor P. J. Diggle in the Chair]

Summary. We consider prediction and uncertainty analysis for systems which are approximated using complex mathematical models. Such models, implemented as computer codes, are often generic in the sense that by a suitable choice of some of the model’s input parameters the code can be used to predict the behaviour of the system in a variety of specific applications. However, in any specific application the values of necessary parameters may be unknown. In this case, physical
What is error in a computer simulation?

Oberkampf et al. (2002)

von Neumann and Goldstine Bulletin of the AMS (1947)

[h/t Joe Grcar]
The von Neumann and Goldstine Catechism

“This analysis of the sources of errors should be objective and strict inasmuch as completeness is concerned, but when it comes to the defining, classifying, and separating of the sources, a certain subjectiveness and arbitrariness is unavoidable. With these reservations, the following enumeration and classification of sources of errors seems to be adequate and reasonable.”

Mathematical model

Observations and parameters

Finitistic approximations

Round-off
This analysis of the sources of errors should be objective and strict inasmuch as completeness is concerned, but when it comes to the defining, classifying, and separating of the sources, a certain subjectiveness and arbitrariness is unavoidable. With these reservations, the following enumeration and classification of sources of errors seems to be adequate and reasonable.

Mathematical model

Observations and parameters

Finitistic approximations

Round-off

NOTES

How well math model approximates reality

Model-form error
"This analysis of the sources of errors should be objective and strict inasmuch as completeness is concerned, but when it comes to the defining, classifying, and separating of the sources, a certain subjectiveness and arbitrariness is unavoidable. With these reservations, the following enumeration and classification of sources of errors seems to be adequate and reasonable."

Mathematical model

Observations and parameters

Finitistic approximations

Round-off

NOTES

Forward and inverse UQ

Most of the UQ methods literature
“This analysis of the sources of errors should be objective and strict inasmuch as completeness is concerned, but when it comes to the defining, classifying, and separating of the sources, a certain subjectiveness and arbitrariness is unavoidable. With these reservations, the following enumeration and classification of sources of errors seems to be adequate and reasonable.”

Mathematical model

Observations and parameters

Finitistic approximations

Round-off

NOTES

Asymptotics from classical numerical analysis

Deterministic numerical noise

“Computational noise in deterministic simulations is as ill-defined a concept as can be found in scientific computing.”

Moré and Wild (2011)
Summary thoughts

Computer models are deterministic. In my opinion, approximation tools are better suited.

But computational noise is really annoying, if you take it seriously.

LOTS of fundamental research opportunities for applying statistical methods to noise-less data---i.e., the approximation setting.

What does Bayes have to do with it?
Practical advice

Everyone, civilized conversation and argumentation!!!

Statisticians, include numerical experiments that demonstrate asymptotic convergence of testing error.

Numerical analysts, convergence analysis of statistical standard error and bootstrap standard error in the context of constructive approximation.

Write three review papers:
- Regression for numerical analysts
- Approximation for statisticians
- Reconciling perspective with authors from both communities
QUESTIONS?

Why should we care?

What do you do in practice?