Machine Learning: Computation versus Statistics

Peter Bartlett
CS, Statistics, Simons Institute for the Theory of Computing
UC Berkeley

November 2, 2019

Berkeley
UNIVERSITY OF CALIFORNIA

SIMONS INSTITUTE
for the Theory of Computing

Machine Learning: Computation versus Statistics

- Tension:
  computation versus statistics
  optimization versus estimation
- A central feature of machine learning (past, present, and **future**).

Machine Learning: Computation versus Statistics

- Tension:
  computation versus statistics
  optimization versus estimation
- A central feature of machine learning (past, present, and **future**).

Outline

1. Deep learning

Machine Learning: Computation versus Statistics

- Tension:
  computation versus statistics
  optimization versus estimation
- A central feature of machine learning (past, present, and **future**).

Outline

1. Deep learning
   - Nonparametric statistical methodology

Machine Learning: Computation versus Statistics

- Tension:
  computation versus statistics
  optimization versus estimation
- A central feature of machine learning (past, present, and **future**).

Outline

1. Deep learning
   - Nonparametric statistical methodology
   - Benign overfitting
   - Implicit regularization

Machine Learning: Computation versus Statistics

- Tension:
  computation versus statistics
  optimization versus estimation
- A central feature of machine learning (past, present, and **future**).

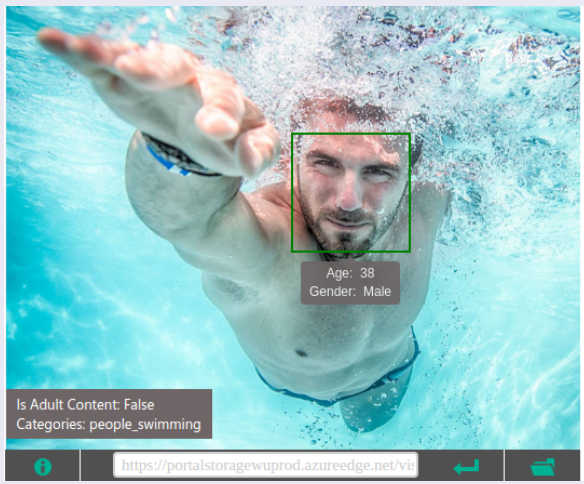
Outline

1. Deep learning
   - Nonparametric statistical methodology
   - Benign overfitting
   - Implicit regularization
2. Computational complexity of estimation problems

Deep Learning Successes

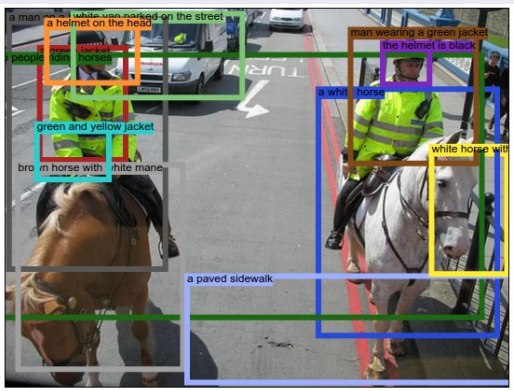
Deep Learning Successes

Computer Vision



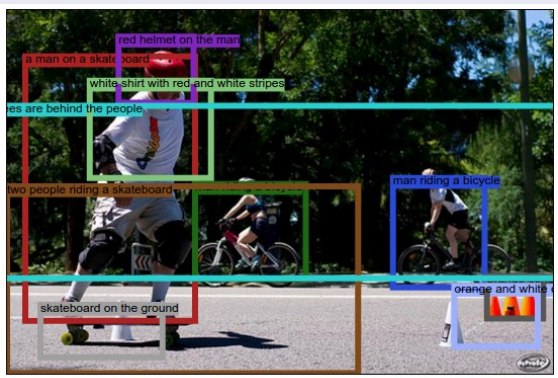
Features:

Feature Name	Value
Description	{ "type": 0, "captions": [{ "text": "a man swimming in a pool of water", "confidence": 0.7850108124440484 }] }
Tags	[{ "name": "water", "confidence": 0.9996442794799805 }, { "name": "sport", "confidence": 0.9504992365837097 }, { "name": "swimming", "confidence": 0.9062818288803101, "hint": "sport" }, { "name": "pool", "confidence": 0.8787588477134705 }, { "name": "water sport", "confidence": 0.631849467754364, "hint": "sport" }]
Image Format	Jpeg
Image Dimensions	1500 x 1155
Clip Art Type	0 Non-clipart
Line Drawing Type	0 Non-LineDrawing
Black & White Image	False



a green jacket. a white horse. a man on a horse. two people riding horses. man wearing a green jacket. the helmet is black. brown horse with white mane. white van parked on the street. a paved sidewalk. green and yellow jacket. a helmet on the head. white horse with white face.

bus parked on the street. a city street scene. front windshield of a bus. man walking on sidewalk. a silver car parked on the street. a city scene. a green traffic light. a building in the background. the bus has a number. a large building. a brick building. red brick building with windows. a blue sign with a white arrow. white lines on the road.



a man on a skateboard. man riding a bicycle. orange cone on the ground. man riding a bicycle. two people riding a skateboard. red helmet on the man. skateboard on the ground. white shirt with red and white stripes. orange and white cone. trees are behind the people.

(microsoft.com) (Johnson et al, 2016)

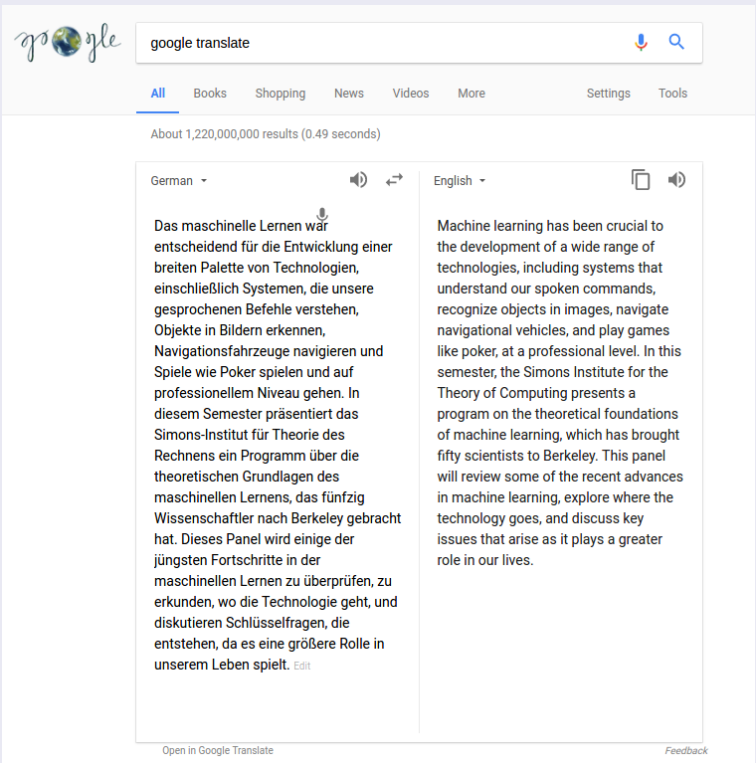
Deep Learning Successes

Speech Recognition

(bgr.com)

Deep Learning Successes

Natural language processing



(google.com, skype.com)

Deep learning to date: an era of craftsmanship

- Engineering solutions to practical problems, striving for good performance on benchmark datasets.

Theoretical Foundations of Deep Learning

Deep learning to date: an era of craftsmanship

- Engineering solutions to practical problems, striving for good performance on benchmark datasets.
- Ingredients: big data + fast computation (GPUs/TPUs) + ... ?

Theoretical Foundations of Deep Learning

Deep learning to date: an era of craftsmanship

- Engineering solutions to practical problems, striving for good performance on benchmark datasets.
- Ingredients: big data + fast computation (GPUs/TPUs) + . . . ?

Deep learning theory

- There are huge gaps in our understanding of deep learning.

Theoretical Foundations of Deep Learning

Deep learning to date: an era of craftsmanship

- Engineering solutions to practical problems, striving for good performance on benchmark datasets.
- Ingredients: big data + fast computation (GPUs/TPUs) + . . . ?

Deep learning theory

- There are huge gaps in our understanding of deep learning.
- Practitioners stumbled on methodology that contradicts established statistical wisdom.

Theoretical Foundations of Deep Learning

Deep learning to date: an era of craftsmanship

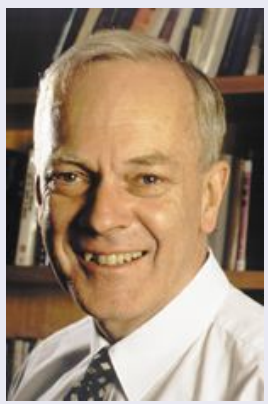
- Engineering solutions to practical problems, striving for good performance on benchmark datasets.
- Ingredients: big data + fast computation (GPUs/TPUs) + . . . ?

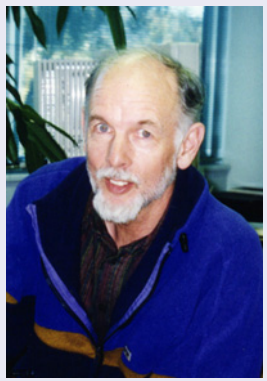
Deep learning theory

- There are huge gaps in our understanding of deep learning.
- Practitioners stumbled on methodology that contradicts established statistical wisdom.
- Key challenge: develop the theoretical foundations that will allow us to understand this new methodology and to improve it.

A Digression

Dept of Systems Engineering, ANU





Brian Anderson

John Moore

A Digression: Model Reference Adaptive Control

United States Patent Office 3,221,229 Patented Nov. 30, 1965

3,221,229
MODEL REFERENCE ADAPTIVE CONTROL SYSTEM

Allen Kezer, North Billerica, and Henry Philip Whitaker, Framingham, Mass., and Paul Vernon Osburn, Oxon Hill, Md., assignors to Massachusetts Institute of Technology, Cambridge, Mass., a corporation of Massachusetts

Filed Jan. 22, 1962, Ser. No. 168,584
7 Claims. (Cl. 318—18)

This invention relates to automatic control systems. It relates to control systems that are self-adaptive, that is, the performance of the control system automatically is improved during its operation. More particularly it relates to an improvement of the Adaptive Control System described in a co-pending application by P. V. Osburn, Serial No. 168,583. The class of control systems to which this invention pertains is particularly appropriate to the control of aircraft and missiles.

The October 1960 issue of "Electro-Technology" magazine contains a survey article on adaptive control systems by R. A. Mathias and R. I. Van Nice. In the article, the term "adaptive control system" is given a wide interpretation. This invention relates in a more restricted sense to adaptive control systems characterized by automatic adjustments to certain parameters in the control loop or loops to compensate for changes in plant or signal characteristics or in both plant *and* signal characteristics. At page 119 of this article, Mathias and Van Nice disclose a model reference adaptive flight control system developed and tested by H. P. Whitaker of the Massachusetts Institute of Technology. The present invention relates to improvements in adaptive control systems of this so-called "MIT" type.

These systems comprise a network of elements each having an input and an output and a number of summing points whereby branches and loops are connected to link a system input to a system output. Each element has a performance function relating its input quantity to its output quantity.

The Whitaker system makes it possible to design a control system which adjusts its own controllable parameters, so that its dynamic performance satisfies the system specifications in the presence of changing operating characteristics. The novel feature is a reference model which stores the system's specifications and permits closed-loop control of the parameters through the use of response-error functions measured during the normal operating responses of the system.

Optimum or fully adapted performances are achieved when the measured error functions have values corresponding to a specified performance index. Use of the model permits design flexibility, since the model can be made to change with the operating modes of the vehicle and can exhibit non-linear characteristics if the systems specifications required these features. If the specifications are relatively loose the system model can be crude and simple, but, on the other hand, greater control of performance, and hence greater flexibility, can be achieved by designing a model of increased sophistication. The self-improving process in these prior-art systems has employed test pulses, sampling of the error quantities, cross-correlation techniques, and in some cases a process of searching for a minimum point of an error function. With these prior-art techniques it has been possible to achieve adaptation within two or three time constants of the dominant response modes of the system.

It is an object of the instant invention to make a significant reduction in the convergence time and to eliminate much of the complexity of the previous mechanisms. Features of the invention are that adaptation is continuous, using the normal inputs of the system without any special test inputs; and the adaptive features are mechanized with simple equipment.

Another feature of the invention is the extension of the model reference principle. In addition to the model system, which is ordinarily in the form of an electrical network, this invention features in the adaptation circuit additional networks each of which has the characteristics of a filter which is the reciprocal of certain of the forward path elements, one such filter being required for each of the adaptable parameters of the control system. Other features and advantages of the invention will be apprehended from the following specifications and annexed drawings of which:

FIG. 1 is a block diagram of the model reference adaptive control system;

FIG. 2 is a mathematical block diagram of the system of FIG. 1;

FIG. 3 is a graph showing typical variation of integral-squared criterion within the adjustable parameter;

FIGS. *4a* and *4b* are generalized control system block diagrams represented by the single block **10** in FIG. 2;

FIG. 5 is a functional diagram illustrating a feature of the invention;

FIG. 6 is a mathematical block diagram of an aircraft roll control system to which the invention is applied by way of specific example;

FIGS. *7a*, *7b*, *7c*, and *7d* are graphs of error functions for the adaptive roll system of FIG. 6;

FIGS. *8a*, *8b*, *8c*, and *8d* are graphs of the system response of the adaptive system of FIG. 6 for supersonic flight conditions;

FIGS. *9a*, *9b*, *9c*, and *9d* are graphs of the system response for subsonic flight conditions;

FIG. 10 is a block diagram illustrating an alternative embodiment of the invention;

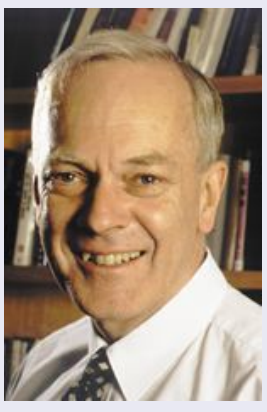
FIG. 11 is a block diagram illustrating another alternative embodiment of the invention.

FIG. 1 is a simplified functional diagram of an adaptive system. The dotted box **10** incloses the stabilization and directional control system **11**, the aircraft **12** upon which the system **11** operates, an output quantity indicating unit **13**, and a feedback path **14**. The stabilization and directional control system **11** comprises various gyroscopes, amplifiers, and actuators which are necessary to convert the pilot's directional commands **15** to motions of the aircraft control surfaces. Equipment, perhaps involving gyroscopes, is included in the output quantity indicating unit **13**, whereby the attitude **16** of the aircraft **12** is converted into suitable electrical signals for feedback **14** to the control system **11**. A dotted box **20** encloses the elements of the adaptation unit **20** which comprise an error signal comparator **21**, a performance analyzer **22**, and an adjusting servo **23**, the output **24** of which adjusts the variable parameters of the stabilization and directional control system **11**. The performance analyzer **22** receives as inputs the Response Error Quantity from the comparator **21**, and certain intermediate signals **26** from the control system **11**, as hereinafter explained. Associated with the adaptation unit **20** is the performance reference model **30** designed to respond to the directional control commands **15** as the aircraft **12** is desired to respond. Typically the model **30** will itself be subject to reference settings **31** wherein take-off weight, changes in balance, and other such variables are accommodated. A second feedback path **34** carries the electrical indications of the aircraft attitude outputs back to the error signal comparator **21** where they may be compared to the electrical outputs **35** of the performance reference model.

In these systems there are two different kinds of error quantities. One is the true system error quantity, [*e*],

Whitaker's "MIT Rule" circa 1961

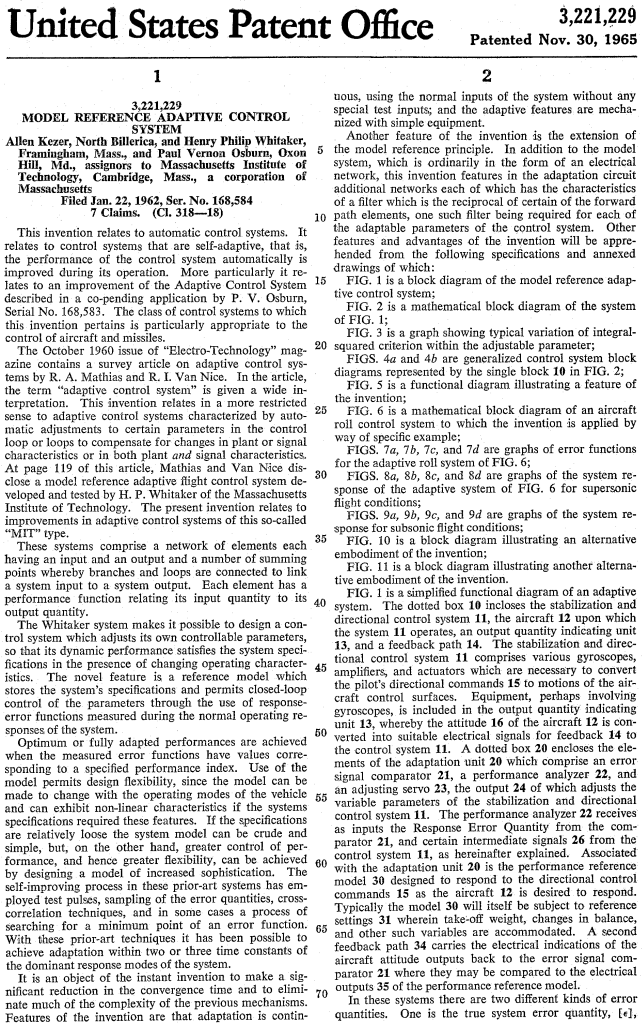
Dept of Systems Engineering, ANU



Brian Anderson

John Moore

A Digression: Model Reference Adaptive Control



United States Patent Office 3,221,229 Patented Nov. 30, 1965

1 2

3,221,229
MODEL REFERENCE ADAPTIVE CONTROL SYSTEM
Allen Kezer, North Billerica, and Henry Philip Whitaker, Framingham, Mass., and Paul Vernon Osburn, Oxon Hill, Md., assignors to Massachusetts Institute of Technology, Cambridge, Mass., a corporation of Massachusetts
Filed Jan. 22, 1962, Ser. No. 168,584
7 Claims. (Cl. 318—18)

This invention relates to automatic control systems. It relates to control systems that are self-adaptive, that is, the performance of the control system automatically is improved during its operation. More particularly it relates to an improvement of the Adaptive Control System described in a co-pending application by P. V. Osburn, Serial No. 168,583. The class of control systems to which this invention pertains is particularly appropriate to the control of aircraft and missiles.

The October 1960 issue of "Electro-Technology" magazine contains a survey article on adaptive control systems by R. A. Mathias and R. I. Van Nice. In the article, the term "adaptive control system" is given a wide interpretation. This invention relates in a more restricted sense to adaptive control systems characterized by automatic adjustments to certain parameters in the control loop or loops to compensate for changes in plant or signal characteristics or in both plant *and* signal characteristics. At page 119 of this article, Mathias and Van Nice disclose a model reference adaptive flight control system developed and tested by H. P. Whitaker of the Massachusetts Institute of Technology. The present invention relates to improvements in adaptive control systems of this so-called "MIT" type.

These systems comprise a network of elements each having an input and an output and a number of summing points whereby branches and loops are connected to link a system input to a system output. Each element has a performance function relating its input quantity to its output quantity.

The Whitaker system makes it possible to design a control system which adjusts its own controllable parameters, so that its dynamic performance satisfies the system specifications in the presence of changing operating characteristics. The novel feature is a reference model which stores the system's specifications and permits closed-loop control of the parameters through the use of response-error functions measured during the normal operating responses of the system.

Optimum or fully adapted performances are achieved when the measured error functions have values corresponding to a specified performance index. Use of the model permits design flexibility, since the model can be made to change with the operating modes of the vehicle and can exhibit non-linear characteristics if the systems specifications required these features. If the specifications are relatively loose the system model can be crude and simple, but, on the other hand, greater control of performance, and hence greater flexibility, can be achieved by designing a model of increased sophistication. The self-improving process in these prior-art systems has employed test pulses, sampling of the error quantities, cross-correlation techniques, and in some cases a process of searching for a minimum point of an error function. With these prior-art techniques it has been possible to achieve adaptation within two or three time constants of the dominant response modes of the system.

It is an object of the instant invention to make a significant reduction in the convergence time and to eliminate much of the complexity of the previous mechanisms. Features of the invention are that adaptation is continuous, using the normal inputs of the system without any special test inputs; and the adaptive features are mechanized with simple equipment.

Another feature of the invention is the extension of the model reference principle. In addition to the model system, which is ordinarily in the form of an electrical network, this invention features in the adaptation circuit additional networks each of which has the characteristics of a filter which is the reciprocal of certain of the forward path elements, one such filter being required for each of the adaptable parameters of the control system. Other features and advantages of the invention will be apprehended from the following specifications and annexed drawings of which:

FIG. 1 is a block diagram of the model reference adaptive control system;

FIG. 2 is a mathematical block diagram of the system of FIG. 1;

FIG. 3 is a graph showing typical variation of integral-squared criterion within the adjustable parameter;

FIGS. 4*a* and 4*b* are generalized control system block diagrams represented by the single block **10** in FIG. 2;

FIG. 5 is a functional diagram illustrating a feature of the invention;

FIG. 6 is a mathematical block diagram of an aircraft roll control system to which the invention is applied by way of specific example;

FIGS. 7*a*, 7*b*, 7*c*, and 7*d* are graphs of error functions for the adaptive roll system of FIG. 6;

FIGS. 8*a*, 8*b*, 8*c*, and 8*d* are graphs of the system response of the adaptive system of FIG. 6 for supersonic flight conditions;

FIGS. 9*a*, 9*b*, 9*c*, and 9*d* are graphs of the system response for subsonic flight conditions;

FIG. 10 is a block diagram illustrating an alternative embodiment of the invention;

FIG. 11 is a block diagram illustrating another alternative embodiment of the invention.

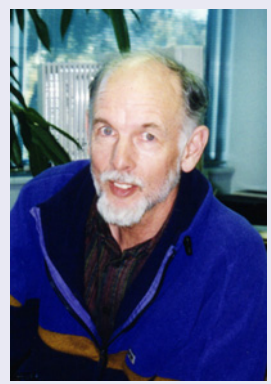
FIG. 1 is a simplified functional diagram of an adaptive system. The dotted box **10** incloses the stabilization and directional control system **11**, the aircraft **12** upon which the system **11** operates, an output quantity indicating unit **13**, and a feedback path **14**. The stabilization and directional control system **11** comprises various gyroscopes, amplifiers, and actuators which are necessary to convert the pilot's directional commands **15** to motions of the aircraft control surfaces. Equipment, perhaps involving gyroscopes, is included in the output quantity indicating unit **13**, whereby the attitude **16** of the aircraft **12** is converted into suitable electrical signals for feedback **14** to the control system **11**. A dotted box **20** encloses the elements of the adaptation unit **20** which comprise an error signal comparator **21**, a performance analyzer **22**, and an adjusting servo **23**, the output **24** of which adjusts the variable parameters of the stabilization and directional control system **11**. The performance analyzer **22** receives as inputs the Response Error Quantity from the comparator **21**, and certain intermediate signals **26** from the control system **11**, as hereinafter explained. Associated with the adaptation unit **20** is the performance reference model **30** designed to respond to the directional control commands **15** as the aircraft **12** is desired to respond. Typically the model **30** will itself be subject to reference settings **31** wherein take-off weight, changes in balance, and other such variables are accommodated. A second feedback path **34** carries the electrical indications of the aircraft attitude outputs back to the error signal comparator **21** where they may be compared to the electrical outputs **35** of the performance reference model.

In these systems there are two different kinds of error quantities. One is the true system error quantity, [*e*],

Whitaker's "MIT Rule" circa 1961

$$\theta_{k+1} = \theta_k - \gamma \frac{d\hat{J}(\theta_k)}{d\theta}; \; J = \int (y_r(t) - y_p(t))^2 \, dt.$$

Dept of Systems Engineering, ANU



Brian Anderson

John Moore

Deep learning as nonparametric statistical methodology

Deep compositions of nonlinear functions, trained on data with SGD

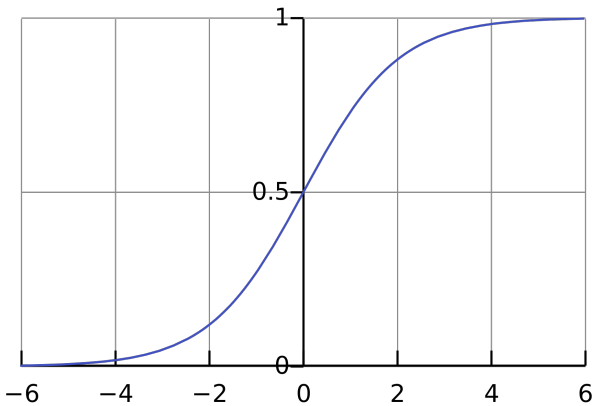
$$h = h_L \circ h_{L-1} \circ \cdots \circ h_1,$$
$$h_i : x \mapsto \sigma(W_i x)$$

Deep learning as nonparametric statistical methodology

Deep compositions of nonlinear functions, trained on data with SGD

$$h = h_L \circ h_{L-1} \circ \cdots \circ h_1,$$
$$h_i : x \mapsto \sigma(W_i x)$$

with, e.g., $\sigma(v)_i = \dfrac{1}{1 + \exp(-v_i)}$,

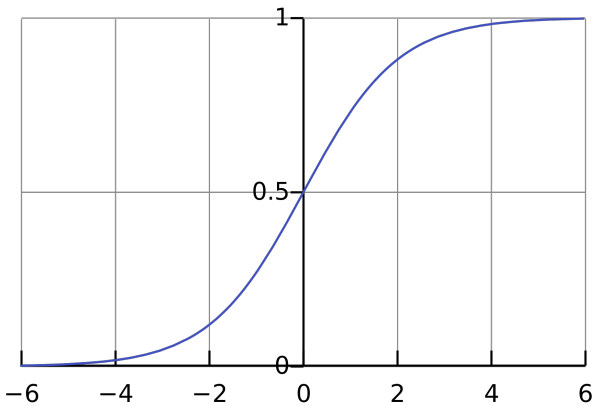


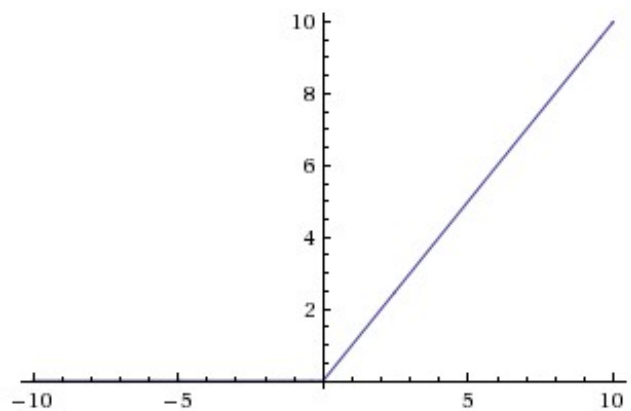
Deep learning as nonparametric statistical methodology

Deep compositions of nonlinear functions, trained on data with SGD

$$h = h_L \circ h_{L-1} \circ \cdots \circ h_1,$$
$$h_i : x \mapsto \sigma(W_i x)$$

with, e.g., $\sigma(v)_i = \dfrac{1}{1 + \exp(-v_i)}$, $\sigma(v)_i = \max\{0, v_i\}$,





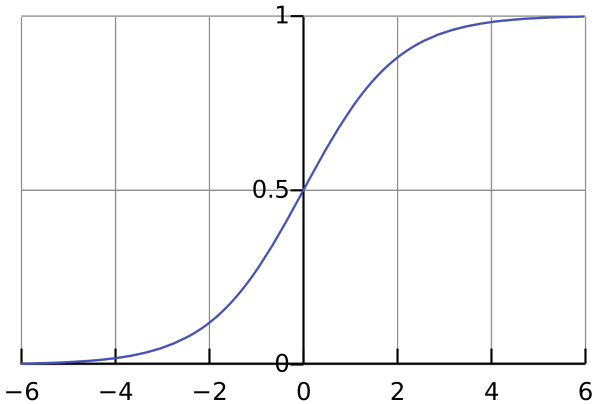
Deep learning as nonparametric statistical methodology

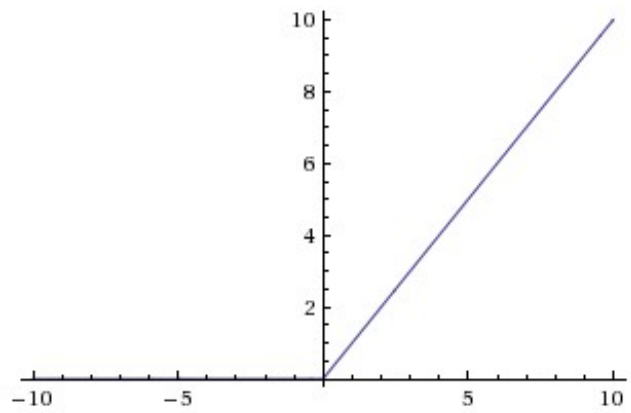
Deep compositions of nonlinear functions, trained on data with SGD

$$h = h_L \circ h_{L-1} \circ \cdots \circ h_1,$$
$$h_i : x \mapsto \sigma(W_i x)$$

with, e.g., $\sigma(v)_i = \dfrac{1}{1+\exp(-v_i)}$, $\sigma(v)_i = \max\{0, v_i\}$,

or "resnets," "max pooling," "attention," ...





Nonparametric Statistical Learning Methodology

Approximation

What kinds of functions can these compositions approximate well?

Nonparametric Statistical Learning Methodology

Approximation

What kinds of functions can these compositions approximate well?

Estimation

How can we effectively trade off complexity with sample size requirements?

Nonparametric Statistical Learning Methodology

Approximation

What kinds of functions can these compositions approximate well?

Estimation

How can we effectively trade off complexity with sample size requirements?

Optimization

How can we *efficiently* find a prediction rule that fits the data well?
(More precisely, how can we efficiently find a prediction rule that gives a good balance of complexity and fit to the data?)

Nonparametric Statistical Learning Methodology

Approximation

What kinds of functions can these compositions approximate well?

Estimation

How can we effectively trade off complexity with sample size requirements?

Optimization

How can we *efficiently* find a prediction rule that fits the data well?
(More precisely, how can we efficiently find a prediction rule that gives a good balance of complexity and fit to the data?)

Deep learning appears to give favorable trade-offs between these competing issues.

Nonparametric Statistical Learning Methodology

Approximation

What kinds of functions can these compositions approximate well?

Estimation

How can we effectively trade off complexity with sample size requirements?

Optimization

How can we *efficiently* find a prediction rule that fits the data well?
(More precisely, how can we efficiently find a prediction rule that gives a good balance of complexity and fit to the data?)

Deep learning appears to give favorable trade-offs between these competing issues.
We don't understand why.

Nonparametric Statistical Learning: Estimation

Estimation

Typically, we aim for a trade-off between

- Fit to the training data,
- Complexity $\Omega(f)$ of a prediction rule

Nonparametric Statistical Learning: Estimation

Estimation

Typically, we aim for a trade-off between

- Fit to the training data,

$$\hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell_i(f)$$

- Complexity $\Omega(f)$ of a prediction rule

Nonparametric Statistical Learning: Estimation

Estimation

Typically, we aim for a trade-off between

- Fit to the training data, e.g.,

$$\hat{R}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(f) = \frac{1}{n}\sum_{i=1}^{n} (f(x_i) - y_i)^2 \,,$$

- Complexity $\Omega(f)$ of a prediction rule

Nonparametric Statistical Learning: Estimation

Estimation

Typically, we aim for a trade-off between

- Fit to the training data, e.g.,

$$\hat{R}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(f) = \frac{1}{n}\sum_{i=1}^{n} (f(x_i) - y_i)^2 \,,$$

- Complexity $\Omega(f)$ of a prediction rule, e.g.,
  - Number of parameters
  - Norm of parameter vector
  - Norm of function in a reproducing kernel Hilbert space,
  - Bandwidth of smoothing kernel,
  - ...

Nonparametric Statistical Learning: Estimation

Estimation

Typically, we aim for a trade-off between

- Fit to the training data, e.g.,

$$\hat{R}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(f) = \frac{1}{n}\sum_{i=1}^{n} (f(x_i) - y_i)^2 \,,$$

- Complexity $\Omega(f)$ of a prediction rule, e.g.,
  - Number of parameters
  - Norm of parameter vector
  - Norm of function in a reproducing kernel Hilbert space,
  - Bandwidth of smoothing kernel,
  - ...

This is especially important for nonparametric methods, that is, those for which the number of parameters grows with the sample size.

Estimators for Inverse Problems: Convex Regularization

Found Comput Math (2012) 12:805–849
DOI 10.1007/s10208-012-9135-7

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

The Convex Geometry of Linear Inverse Problems

Venkat Chandrasekaran · Benjamin Recht · Pablo A. Parrilo · Alan S. Willsky

Received: 2 December 2010 / Revised: 25 February 2012 / Accepted: 3 July 2012 /
Published online: 16 October 2012

Abstract In applications throughout science and engineering one is often faced with the challenge of solving an ill-posed inverse problem, where the number of available measurements is smaller than the dimension of the model to be estimated. However in many practical situations of interest, models are constrained structurally so that they only have a few degrees of freedom relative to their ambient dimension. This paper provides a general framework to convert notions of simplicity into convex penalty functions, resulting in convex optimization solutions to linear, underdetermined inverse problems. The class of simple models considered includes those formed as the sum of a few atoms from some (possibly infinite) elementary atomic set; exam-

Communicated by Emmanuel Candès.

V. Chandrasekaran (✉)
Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA
e-mail: venkatc@caltech.edu

B. Recht
Computer Sciences Department, University of Wisconsin, Madison, WI 53706, USA
e-mail: brecht@cs.wisc.edu

P.A. Parrilo · A.S. Willsky
Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

P.A. Parrilo
e-mail: parrilo@mit.edu

A.S. Willsky
e-mail: willsky@mit.edu

Estimators for Inverse Problems: Convex Regularization

Found Comput Math (2012) 12:805–849
DOI 10.1007/s10208-012-9135-7

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

The Convex Geometry of Linear Inverse Problems

Venkat Chandrasekaran · Benjamin Recht · Pablo A. Parrilo · Alan S. Willsky

Received: 2 December 2010 / Revised: 25 February 2012 / Accepted: 3 July 2012 /
Published online: 16 October 2012

Abstract In applications throughout science and engineering one is often faced with the challenge of solving an ill-posed inverse problem, where the number of available measurements is smaller than the dimension of the model to be estimated. However in many practical situations of interest, models are constrained structurally so that they only have a few degrees of freedom relative to their ambient dimension. This paper provides a general framework to convert notions of simplicity into convex penalty functions, resulting in convex optimization solutions to linear, underdetermined inverse problems. The class of simple models considered includes those formed as the sum of a few atoms from some (possibly infinite) elementary atomic set; exam-

Communicated by Emmanuel Candès.

V. Chandrasekaran (✉)
Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA
e-mail: venkatc@caltech.edu

B. Recht
Computer Sciences Department, University of Wisconsin, Madison, WI 53706, USA
e-mail: brecht@cs.wisc.edu

P.A. Parrilo · A.S. Willsky
Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

P.A. Parrilo
e-mail: parrilo@mit.edu

A.S. Willsky
e-mail: willsky@mit.edu

Estimating simple structure

e.g., sparse linear regression,
low rank tensors,
rankings from surveys, ...

Estimators for Inverse Problems: Convex Regularization

Found Comput Math (2012) 12:805–849
DOI 10.1007/s10208-012-9135-7

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

The Convex Geometry of Linear Inverse Problems

Venkat Chandrasekaran · Benjamin Recht · Pablo A. Parrilo · Alan S. Willsky

Received: 2 December 2010 / Revised: 25 February 2012 / Accepted: 3 July 2012 /
Published online: 16 October 2012

Abstract In applications throughout science and engineering one is often faced with the challenge of solving an ill-posed inverse problem, where the number of available measurements is smaller than the dimension of the model to be estimated. However in many practical situations of interest, models are constrained structurally so that they only have a few degrees of freedom relative to their ambient dimension. This paper provides a general framework to convert notions of simplicity into convex penalty functions, resulting in convex optimization solutions to linear, underdetermined inverse problems. The class of simple models considered includes those formed as the sum of a few atoms from some (possibly infinite) elementary atomic set; exam-

Communicated by Emmanuel Candès.

V. Chandrasekaran (✉)
Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA
e-mail: venkatc@caltech.edu

B. Recht
Computer Sciences Department, University of Wisconsin, Madison, WI 53706, USA
e-mail: brecht@cs.wisc.edu

P.A. Parrilo · A.S. Willsky
Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

P.A. Parrilo
e-mail: parrilo@mit.edu

A.S. Willsky
e-mail: willsky@mit.edu

Estimating simple structure

e.g., sparse linear regression,
low rank tensors,
rankings from surveys, ...

Minimum norm fit: $\Omega(f) = \|f\|_{\mathcal{A}}$

$$\begin{aligned} \min \quad & \Omega(f) \\ \text{s.t.} \quad & \hat{R}(f) \leq \epsilon. \end{aligned}$$

Nonparametric Statistical Learning: Estimation

Regularization: balancing $\hat{R}$ and Ω

$$\min_f \; a\Omega(f) + \hat{R}(f),$$

Nonparametric Statistical Learning: Estimation

Regularization: balancing $\hat{R}$ and Ω

$$\min_f \; a\Omega(f) + \hat{R}(f), \quad \text{or} \quad \min \; \hat{R}(f) \;\; \text{s.t.} \;\; \Omega(f) \leq b,$$

Nonparametric Statistical Learning: Estimation

Regularization: balancing $\hat{R}$ and Ω

$$\min_f \; a\Omega(f) + \hat{R}(f), \quad \text{or} \quad \min \; \hat{R}(f) \;\; \text{s.t.} \;\; \Omega(f) \leq b, \quad \text{or} \quad \min \; \Omega(f) \;\; \text{s.t.} \;\; \hat{R}(f) \leq c.$$

Nonparametric Statistical Learning: Estimation

Regularization: balancing $\hat{R}$ and Ω

$$\min_f \; a\Omega(f) + \hat{R}(f), \quad \text{or} \quad \min \; \hat{R}(f) \;\; \text{s.t.} \;\; \Omega(f) \le b, \quad \text{or} \quad \min \; \Omega(f) \;\; \text{s.t.} \;\; \hat{R}(f) \le c.$$

Classical analysis: uniform convergence

$$\sup \left\{ R(f) - \hat{R}(f) : \Omega(f) \le b \right\} \le \epsilon,$$

$$R(f) := \mathbb{E}\ell(f) \qquad \frac{1}{n}\sum_{i=1}^{n} \ell_i(f) =: \hat{R}(f).$$

Nonparametric Statistical Learning: Estimation

Regularization: balancing $\hat{R}$ and Ω

$$\min_f \; a\Omega(f) + \hat{R}(f), \quad \text{or} \quad \min \; \hat{R}(f) \;\; \text{s.t.} \;\; \Omega(f) \le b, \quad \text{or} \quad \min \; \Omega(f) \;\; \text{s.t.} \;\; \hat{R}(f) \le c.$$

Classical analysis: uniform convergence

$$\sup \left\{ R(f) - \hat{R}(f) : \Omega(f) \le b \right\} \le \epsilon,$$

$$\text{so that } R(\hat{f}) := \mathbb{E}\ell(\hat{f}) \approx \frac{1}{n} \sum_{i=1}^{n} \ell_i(\hat{f}) =: \hat{R}(\hat{f}).$$

Nonparametric Statistical Learning: Estimation

Regularization: balancing $\hat{R}$ and Ω

$$\min_f \; a\Omega(f) + \hat{R}(f), \quad \text{or} \quad \min \; \hat{R}(f) \;\; \text{s.t.} \;\; \Omega(f) \le b, \quad \text{or} \quad \min \; \Omega(f) \;\; \text{s.t.} \;\; \hat{R}(f) \le c.$$

Classical analysis: uniform convergence

$$\sup \left\{ R(f) - \hat{R}(f) : \Omega(f) \le b \right\} \le \epsilon,$$

$$\text{so that } R(\hat{f}) := \mathbb{E}\ell(\hat{f}) \approx \frac{1}{n} \sum_{i=1}^{n} \ell_i(\hat{f}) =: \hat{R}(\hat{f}).$$

To exploit uniform convergence, we should consider $c \ge R^* - \epsilon$, where $R^* = \min_f R(f)$.

Nonparametric Statistical Learning: Estimation

Regularization: balancing $\hat{R}$ and Ω

$$\min_f \; a\Omega(f) + \hat{R}(f), \quad \text{or} \quad \min \; \hat{R}(f) \;\; \text{s.t.} \;\; \Omega(f) \le b, \quad \text{or} \quad \min \; \Omega(f) \;\; \text{s.t.} \;\; \hat{R}(f) \le c.$$

Classical analysis: uniform convergence

$$\sup \left\{ R(f) - \hat{R}(f) : \Omega(f) \le b \right\} \le \epsilon,$$

$$\text{so that } R(\hat{f}) := \mathbb{E}\ell(\hat{f}) \approx \frac{1}{n} \sum_{i=1}^{n} \ell_i(\hat{f}) =: \hat{R}(\hat{f}).$$

To exploit uniform convergence, we should consider $c \ge R^* - \epsilon$, where $R^* = \min_f R(f)$.

Overfitting regime

$$\hat{R}(f) \ll R^*$$

Deep Learning Surprises 1: Benign Overfitting

Benign Overfitting

- Deep networks can achieve zero training error (for *regression* loss)

Deep Learning Surprises 1: Benign Overfitting

Benign Overfitting

- Deep networks can achieve zero training error (for *regression* loss)
- ... with near state-of-the-art performance

Deep Learning Surprises 1: Benign Overfitting

Benign Overfitting

- Deep networks can achieve zero training error (for *regression* loss)
- ... with near state-of-the-art performance
- ... even for noisy problems ($R^* \gg 0$).

Deep Learning Surprises 1: Benign Overfitting

Benign Overfitting

- Deep networks can achieve zero training error (for *regression* loss)
- ... with near state-of-the-art performance
- ... even for noisy problems ($R^* \gg 0$).
- No tradeoff between fit to training data and complexity!

Deep Learning Surprises 1: Benign Overfitting

Benign Overfitting

- Deep networks can achieve zero training error (for *regression* loss)
- ... with near state-of-the-art performance
- ... even for noisy problems ($R^* \gg 0$).
- No tradeoff between fit to training data and complexity!
- Deep networks seem to operate in the overfitting regime ($\hat{R}(f) \ll R^*$) but still predict accurately.

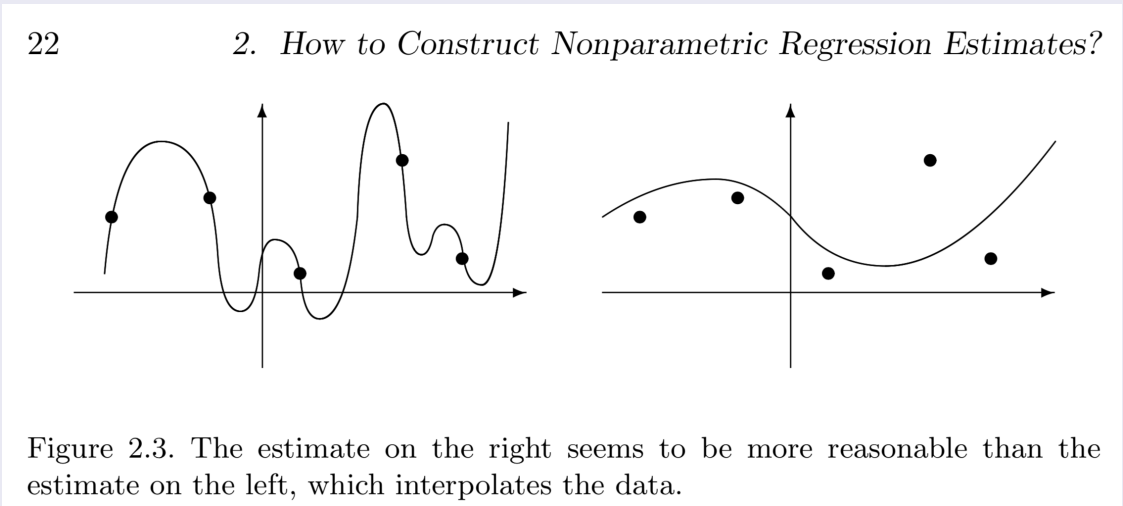
Deep Learning Surprises 1: Benign Overfitting

Benign Overfitting

- Deep networks can achieve zero training error (for *regression* loss)
- ... with near state-of-the-art performance
- ... even for noisy problems ($R^* \gg 0$).
- No tradeoff between fit to training data and complexity!
- Deep networks seem to operate in the overfitting regime ($\hat{R}(f) \ll R^*$) but still predict accurately.
- A new statistical phenomenon: *benign overfitting*.

Statistical Wisdom and Overfitting

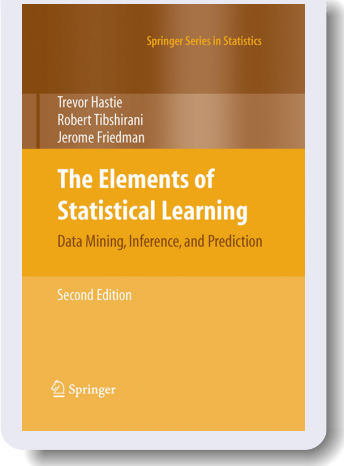
"... interpolating fits... [are] unlikely to predict future data well at all."

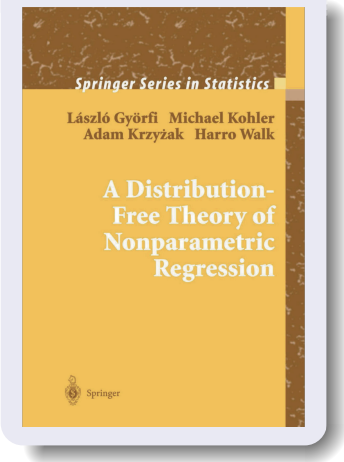


22 2. How to Construct Nonparametric Regression Estimates?

Figure 2.3. The estimate on the right seems to be more reasonable than the estimate on the left, which interpolates the data.

"... a function which interpolates the data ... is not a reasonable estimate."





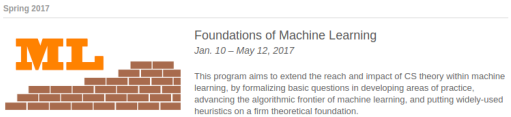
see also (B. and Rakhlin, Simons Institute, May 2019)

Deep Learning Surprises 1: Benign Overfitting

A new statistical phenomenon:
good prediction with very small training error for regression loss

- Statistical wisdom says a prediction rule should not fit too well.
- But deep networks are trained to fit noisy data perfectly, and they predict well.

Deep Learning Surprises 1: Benign Overfitting



Progress in Benign Overfitting

- Simplicial interpolation ($\approx$ nearest neighbor) (Belkin, Hsu, Mitra, 2018)
- Nadaraya-Watson estimator with singular kernels

  (Belkin, Hsu, Mitra, 2018; Belkin, Rakhlin, Tsybakov, 2018)
- Random matrix theory asymptotics ($d \asymp n$) for linear regression, random nonlinear features

  (Hastie, Montanari, Rosset, Tibshirani, 2019; Mei, Montanari, 2019; Belkin, Hsu and Xu, 2019)
- Certain reproducing kernel Hilbert spaces

  (Liang and Rakhlin, 2018; Rakhlin and Zhai, 2018; Liang, Rakhlin, Zhai, 2019)
- Minimum norm linear regression: tight upper and lower bounds for finite sample, arbitrary dimension (B., Long, Lugosi, Tsigler, 2019)

Deep Learning Surprises 1: Benign Overfitting



(B., Long, Lugosi, Tsigler, 2019)

Characterizing benign overfitting in linear regression

For $\ell(f) = (f(x) - y)^2$,

Deep Learning Surprises 1: Benign Overfitting



(B., Long, Lugosi, Tsigler, 2019)

Characterizing benign overfitting in linear regression

For $\ell(f) = (f(x) - y)^2$, $\Omega(x \mapsto \langle x, \theta \rangle) = \|\theta\|_2$,

Deep Learning Surprises 1: Benign Overfitting



(B., Long, Lugosi, Tsigler, 2019)

Characterizing benign overfitting in linear regression

For $\ell(f) = (f(x) - y)^2$, $\Omega(x \mapsto \langle x, \theta \rangle) = \|\theta\|_2$, and $\begin{pmatrix} x \\ y \end{pmatrix} = \Phi z$ where Φ is a bounded linear operator and z has subgaussian, independent entries,

Deep Learning Surprises 1: Benign Overfitting



(B., Long, Lugosi, Tsigler, 2019)

Characterizing benign overfitting in linear regression

For $\ell(f) = (f(x) - y)^2$, $\Omega\left(x \mapsto \langle x, \theta \rangle\right) = \|\theta\|_2$, and $\begin{pmatrix} x \\ y \end{pmatrix} = \Phi z$ where Φ is a bounded linear operator and z has subgaussian, independent entries,

$$c_1 \left(\frac{d^*}{n} + \frac{n}{R_{d^*}} + \phi\left(\frac{1}{n}\right) \right) \leq \mathbb{E}R(\hat{f}) - R^* \leq c_2 \left(\frac{d^*}{n} + \frac{n}{R_{d^*}} + \frac{1}{\sqrt{n}} \right),$$

where $d^* = \min\{d : r_d \geq c_3 n\}$, r_d and R_d are effective ranks of the covariance of x in the subspace orthogonal to the d highest variance directions, and ϕ is increasing.

Deep Learning Surprises 1: Benign Overfitting



(B., Long, Lugosi, Tsigler, 2019)

Characterizing benign overfitting in linear regression

For $\ell(f) = (f(x) - y)^2$, $\Omega(x \mapsto \langle x, \theta \rangle) = \|\theta\|_2$, and $\begin{pmatrix} x \\ y \end{pmatrix} = \Phi z$ where Φ is a bounded linear operator and z has subgaussian, independent entries,

$$c_1 \left(\frac{d^*}{n} + \frac{n}{R_{d^*}} + \phi\left(\frac{1}{n}\right) \right) \leq \mathbb{E}R(\hat{f}) - R^* \leq c_2 \left(\frac{d^*}{n} + \frac{n}{R_{d^*}} + \frac{1}{\sqrt{n}} \right),$$

where $d^* = \min\{d : r_d \geq c_3 n\}$, r_d and R_d are effective ranks of the covariance of x in the subspace orthogonal to the d highest variance directions, and ϕ is increasing.

That is, benign overfitting occurs iff there is a subspace where the covariance has small magnitude, high dimension, and low eccentricity.

Deep Learning Surprises 1: Benign Overfitting

Progress in Benign Overfitting

- Simplicial interpolation ($\approx$ nearest neighbor) (Belkin, Hsu, Mitra, 2018)
- Nadaraya-Watson estimator with singular kernels

  (Belkin, Hsu, Mitra, 2018; Belkin, Rakhlin, Tsybakov, 2018)
- Random matrix theory asymptotics ($d \asymp n$) for linear regression, random nonlinear features

  (Hastie, Montanari, Rosset, Tibshirani, 2019; Mei, Montanari, 2019; Belkin, Hsu and Xu, 2019)
- Certain reproducing kernel Hilbert spaces

  (Liang and Rakhlin, 2018; Rakhlin and Zhai, 2018; Liang, Rakhlin, Zhai, 2019)
- Minimum norm linear regression: tight upper and lower bounds for finite sample, arbitrary dimension (B., Long, Lugosi, Tsigler, 2019)

Benign Overfitting in Deep Networks?

Machine Learning: Computation versus Statistics

1. Deep learning mysteries
   - Benign overfitting
   - **Implicit regularization**
2. Computational complexity of estimation

Deep Learning Surprises 2: Implicit Regularization

Regularization in the overfitting regime ($c \ll R^*$)

$$\begin{aligned} \min \ & \Omega(f) \\ \text{s.t.} \ & \hat{R}(f) \leq c. \end{aligned}$$

Deep Learning Surprises 2: Implicit Regularization

Regularization in the overfitting regime ($c \ll R^*$)

$$\begin{aligned} \min \ & \Omega(f) \\ \text{s.t.} \ & \hat{R}(f) \leq c. \end{aligned}$$

Implicit Regularization

- Stochastic gradient descent finds deep networks satisfying the (overfitting) constraint, and these deep networks predict accurately.

Deep Learning Surprises 2: Implicit Regularization

Regularization in the overfitting regime ($c \ll R^*$)

$$\begin{aligned} \min \ & \Omega(f) \\ \text{s.t.} \ & \hat{R}(f) \leq c. \end{aligned}$$

Implicit Regularization

- Stochastic gradient descent finds deep networks satisfying the (overfitting) constraint, and these deep networks predict accurately.
- What is the regularizer Ω?

Deep Learning Surprises 2: Implicit Regularization

Regularization in the overfitting regime ($c \ll R^*$)

$$\begin{aligned} \min \ & \Omega(f) \\ \text{s.t.} \ & \hat{R}(f) \leq c. \end{aligned}$$

Implicit Regularization

- Stochastic gradient descent finds deep networks satisfying the (overfitting) constraint, and these deep networks predict accurately.
- What is the regularizer Ω?
- The boundaries between the key issues of *optimization*, *estimation*, *and approximation* are blurred.

Deep Learning Surprises 2: Implicit Regularization

Progress in Implicit Regularization

- Linear. $f : x \mapsto \langle \theta, x \rangle$: $\Omega(f) = \|\theta - \theta_0\|$.
- Polynomial. θ_i replaced by θ_i^α: $\Omega(f)$ like a Huber norm.

(Gunasekar, Woodworth, Bhojanapalli, Neyshabur, Srebro, 2017)

- Logistic regression (Soudry, Hoffer, Srebro, 2017)
- Linear convolutional: $\Omega(f)$ penalizes norm of Fourier transform.

(Gunasekar, Lee, Soudry, Srebro, 2018)

Deep Learning Surprises 2: Implicit Regularization

Progress in Implicit Regularization

- Linear. $f : x \mapsto \langle \theta, x \rangle$: $\Omega(f) = \|\theta - \theta_0\|$.
- Polynomial. θ_i replaced by θ_i^α: $\Omega(f)$ like a Huber norm.

(Gunasekar, Woodworth, Bhojanapalli, Neyshabur, Srebro, 2017)

- Logistic regression (Soudry, Hoffer, Srebro, 2017)
- Linear convolutional: $\Omega(f)$ penalizes norm of Fourier transform.

(Gunasekar, Lee, Soudry, Srebro, 2018)

Implicit Regularization in Deep Networks?

Machine Learning: Computation versus Statistics

1. Deep learning mysteries
   - Benign overfitting
   - Implicit regularization
2. **Computational complexity of estimation**

Computational complexity of estimation

Mean estimation

Consider estimating the mean μ of a distribution in $\mathbb{R}^d$ from an n-sample

Computational complexity of estimation

Mean estimation

Consider estimating the mean μ of a distribution in $\mathbb{R}^d$ from an n-sample: $\Pr\left(\|\hat{\mu} - \mu\| \geq \epsilon\right) \leq \delta$.

Computational complexity of estimation

Mean estimation

Consider estimating the mean μ of a distribution in $\mathbb{R}^d$ from an n-sample: $\Pr\left(\|\hat{\mu} - \mu\| \geq \epsilon\right) \leq \delta$.

- What is the computational complexity?

Computational complexity of estimation

Mean estimation

Consider estimating the mean μ of a distribution in $\mathbb{R}^d$ from an n-sample: $\Pr\left(\|\hat{\mu} - \mu\| \geq \epsilon\right) \leq \delta$.

- What is the computational complexity?
- With subgaussian data, the empirical mean suffices.

Computational complexity of estimation

Mean estimation

Consider estimating the mean μ of a distribution in $\mathbb{R}^d$ from an n-sample: $\Pr\left(\|\hat{\mu} - \mu\| \geq \epsilon\right) \leq \delta$.

- What is the computational complexity?
- With subgaussian data, the empirical mean suffices.

  $\epsilon = O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$ (optimal).

Computational complexity of estimation

Mean estimation

Consider estimating the mean μ of a distribution in $\mathbb{R}^d$ from an n-sample: $\Pr\left(\|\hat{\mu} - \mu\| \geq \epsilon\right) \leq \delta$.

- What is the computational complexity?
- With subgaussian data, the empirical mean suffices.

  $\epsilon = O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$ (optimal).
- With weaker assumptions? e.g., just a second moment?

Computational complexity of estimation

Mean estimation

Consider estimating the mean μ of a distribution in $\mathbb{R}^d$ from an n-sample: $\Pr\left(\|\hat{\mu} - \mu\| \geq \epsilon\right) \leq \delta$.

- What is the computational complexity?
- With subgaussian data, the empirical mean suffices.

  $\epsilon = O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$ (optimal).
- With weaker assumptions? e.g., just a second moment?
  The empirical mean is sensitive to large outliers.

Computational complexity of estimation

Mean estimation

Consider estimating the mean μ of a distribution in $\mathbb{R}^d$ from an n-sample: $\Pr\left(\|\hat{\mu} - \mu\| \geq \epsilon\right) \leq \delta$.

- What is the computational complexity?
- With subgaussian data, the empirical mean suffices.

  $\epsilon = O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$ (optimal).
- With weaker assumptions? e.g., just a second moment?
  The empirical mean is sensitive to large outliers.
- In one dimension, we can compute a median of means.

Computational complexity of estimation

Mean estimation

Consider estimating the mean μ of a distribution in $\mathbb{R}^d$ from an n-sample: $\Pr\left(\|\hat{\mu} - \mu\| \geq \epsilon\right) \leq \delta$.

- What is the computational complexity?
- With subgaussian data, the empirical mean suffices.

  $\epsilon = O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$ (optimal).
- With weaker assumptions? e.g., just a second moment?
  The empirical mean is sensitive to large outliers.
- In one dimension, we can compute a median of means.

  $\epsilon = O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$ (optimal).

Computational complexity of estimation

Mean estimation

Consider estimating the mean μ of a distribution in $\mathbb{R}^d$ from an n-sample: $\Pr\left(\|\hat{\mu} - \mu\| \geq \epsilon\right) \leq \delta$.

- What is the computational complexity?
- With subgaussian data, the empirical mean suffices.

  $\epsilon = O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$ (optimal).
- With weaker assumptions? e.g., just a second moment?
  The empirical mean is sensitive to large outliers.
- In one dimension, we can compute a median of means.

  $\epsilon = O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$ (optimal).
- In $\mathbb{R}^d$?

Computational complexity of estimation

Mean estimation with bounded second moments

Computational complexity of estimation

Mean estimation with bounded second moments

- The optimal sample complexity for estimating the mean with heavy-tailed (i.e., just second moments) data in $\mathbb{R}^d$ is the same as with subgaussian data. (Lugosi and Mendelson, 2017)

Computational complexity of estimation

Mean estimation with bounded second moments

- The optimal sample complexity for estimating the mean with heavy-tailed (i.e., just second moments) data in $\mathbb{R}^d$ is the same as with subgaussian data. (Lugosi and Mendelson, 2017)
- Sum-of-squares machinery provides an efficient estimator.

(Hopkins, 2018)

Computational complexity of estimation

Mean estimation with bounded second moments

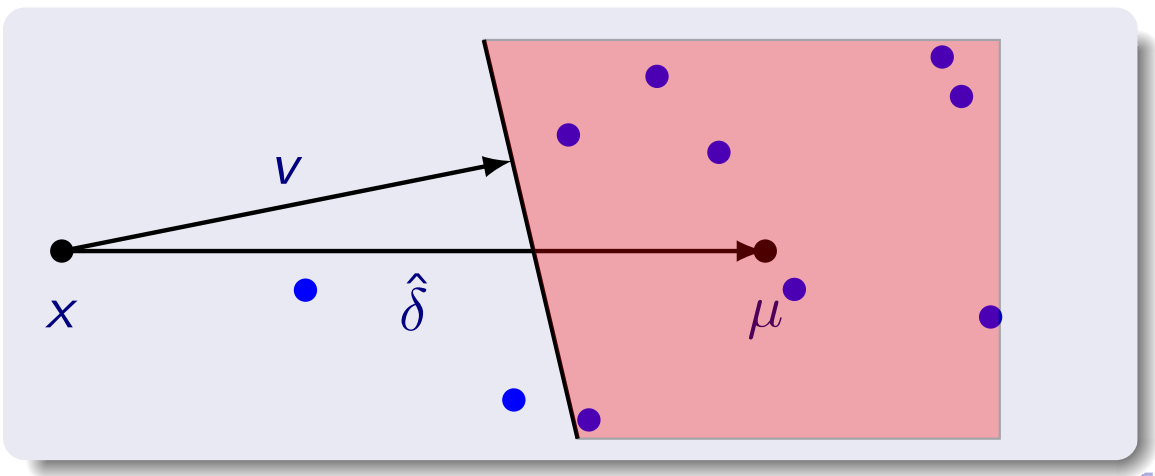
- The optimal sample complexity for estimating the mean with heavy-tailed (i.e., just second moments) data in $\mathbb{R}^d$ is the same as with subgaussian data. (Lugosi and Mendelson, 2017)
- Sum-of-squares machinery provides an efficient estimator. But $O(n^{24})$. (Hopkins, 2018)

Computational complexity of estimation

Mean estimation with bounded second moments

- The optimal sample complexity for estimating the mean with heavy-tailed (i.e., just second moments) data in $\mathbb{R}^d$ is the same as with subgaussian data. (Lugosi and Mendelson, 2017)
- Sum-of-squares machinery provides an efficient estimator. But $O(n^{24})$. (Hopkins, 2018)
- A simple descent-based method has run-time $O(n^4 + n^2 d)$.

Fast Mean Estimation with Sub-Gaussian Rates. (Cherapanamjeri, Flammarion, B., 2019.



Computational complexity of estimation

Many open problems

Computational complexity of estimation

Many open problems

- Hardness of estimation?

Computational complexity of estimation

Many open problems

- Hardness of estimation?
- Information theoretic boundary versus computational boundary?

Computational complexity of estimation

Many open problems

- Hardness of estimation?
- Information theoretic boundary versus computational boundary?
- (Fine-grained) reductions between estimation problems?

Computational complexity of estimation

Many open problems

- Hardness of estimation?
- Information theoretic boundary versus computational boundary?
- (Fine-grained) reductions between estimation problems?
- Canonical hard estimation problems?
  (What is the SAT of estimation?)

c.f., e.g., (Berthet and Rigollet, 2013), (Brennan and Bresler, 2019)

Machine Learning: Computation versus Statistics

1. Deep learning mysteries
   - Benign overfitting
   - Implicit regularization
2. Computational complexity of estimation

Machine Learning: Computation versus Statistics

1. Deep learning mysteries
   - Benign overfitting
   - Implicit regularization
2. Computational complexity of estimation

Congratulations!