# The Confluence of Statistics and Optimization in Signal Processing and Machine Learning

Yuejie Chi

Carnegie Mellon University

2022 CIF Townhall Meeting

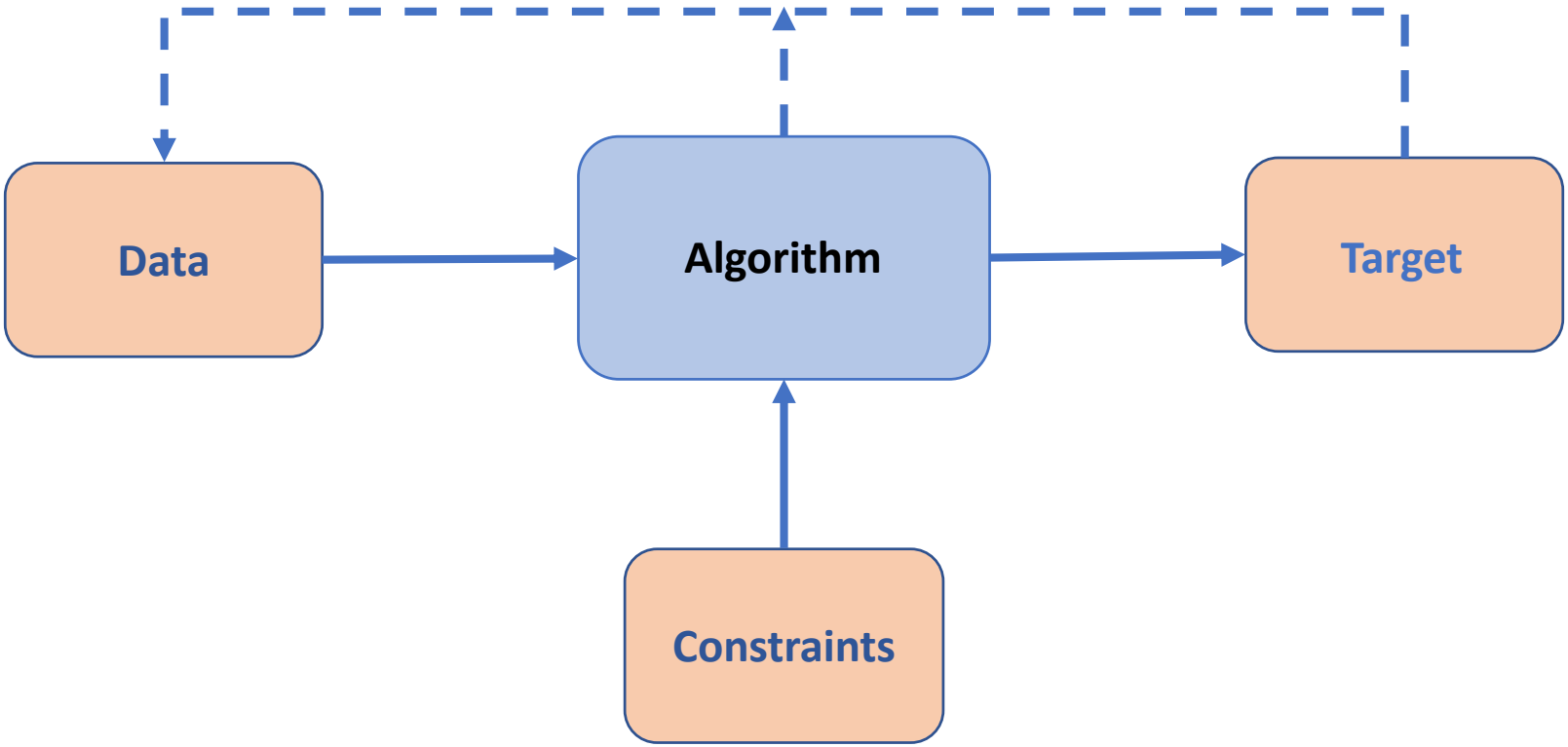
# Unprecedented opportunities and challenges

- **Massive data:**
  acquisition, storage, transmission, processing, analysis and deployment

- **Massive computing:**
  diverse and heterogeneous computation platforms, interplay and integration across the whole stack

- **Massive wishlist:**
  performance, efficiency, privacy, security, fairness, interpretability, etc

# A diagram of the pipeline




Our community is good at designing algorithms.

# Optimization as a common lens

Optimization has emerged as a common lens to derive algorithms

$$\min_{\boldsymbol{x} \in \mathcal{C}} \; f(\boldsymbol{x})$$

- In signal processing, the minimizer produces the signal reconstruction.
- In machine learning, the minimizer produces the ML model.

Importantly, the objective function often represents a surrogate of what we want, rather than what we really want.

# Large-scale optimization for machine learning

**Finite-sum minimization:**

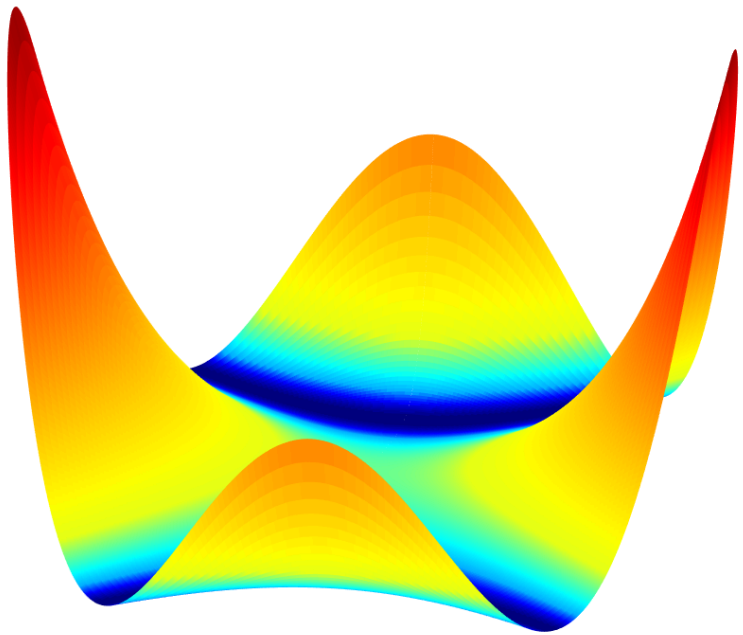
$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{f_i(\boldsymbol{x})}_{i\text{th data}}$$

Waves of efficient (near-optimal) algorithms designs for solving under various resource constraints and problem structures.

**Examples:** stochastic variance reduction algorithms (Johnson and Tong, 2013; Defazio et al 2014), communication-efficient distributed optimization (McMahan et al 2017).

# Hidden convexity in nonconvex problems

Despite nonconvex, many problems possess benign geometry that enables much stronger convergence analysis.



**Examples:** "no bad local minima" property in complete dictionary learning (Sun et al, 2016), PCA (Baldi and Hornik, 1989), linear neural network (Kawaguchi, 2016), ...

# The interplay between statistics and optimization

The ever-increasing problem dimensions call for non-asymptotic understandings of the statistical and computational complexities!

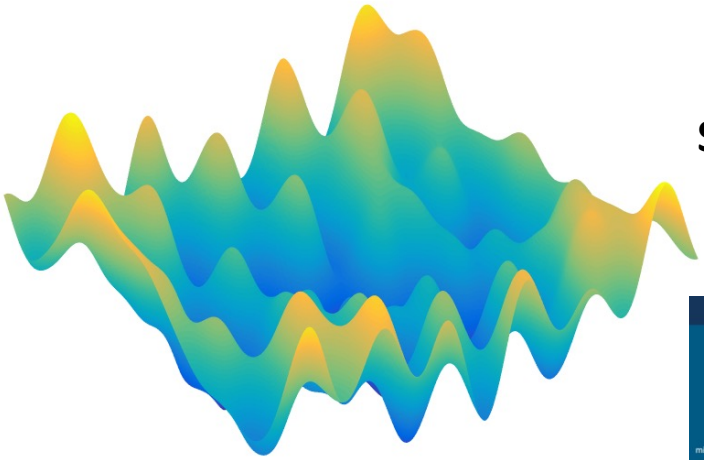
For example, one would like to pin down the performance of the iterates $\boldsymbol{x}_t$ of gradient descent rather than just the final optimizer

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\eta}{N} \sum_{i=1}^{N} \nabla f_i(\boldsymbol{x}_t)$$

- Classical optimization focuses on black-box optimization.
- Classical statistics ignores how the computation is done.

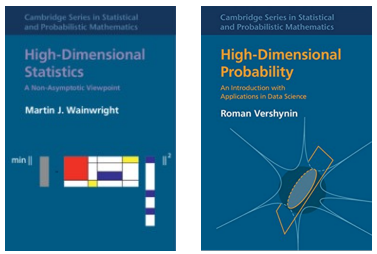
New advances at the interface of optimization and statistics are made/required to answer such questions!

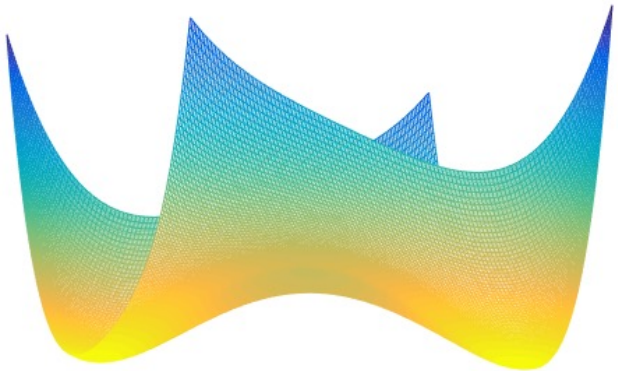
# Statistics meets optimization



worst case

**Statistical model**



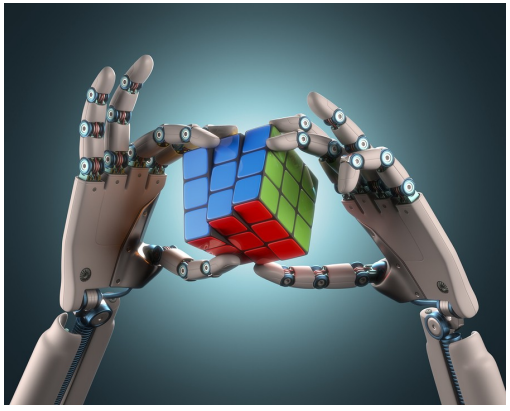



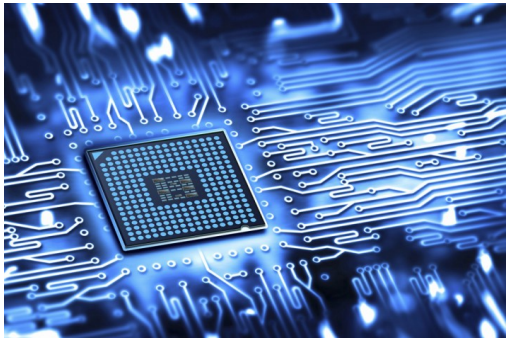
average case

finite-sum $\Longrightarrow$ statistical concentration $\Longrightarrow$ benign optimization

(Mei et al, 2018; Ma et al, 2019)

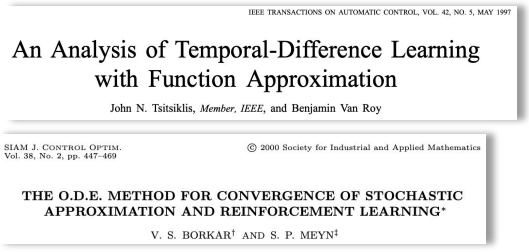
# Reinforcement Learning





*The empirical successes of RL create excitements in enabling next-generation artificial intelligence.*

# From asymptotic to non-asymptotic analyses


IEEE TRANSACTIONS ON AUTOMATIC CONTROL, VOL. 42, NO. 5, MAY 1997

An Analysis of Temporal-Difference Learning with Function Approximation

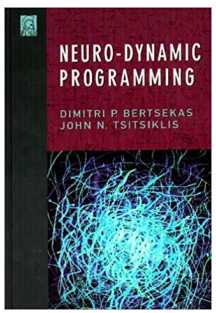
John N. Tsitsiklis, *Member, IEEE*, and Benjamin Van Roy

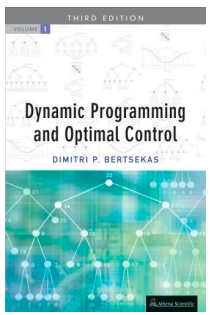
SIAM J. CONTROL OPTIM. Vol. 38, No. 2, pp. 447–469

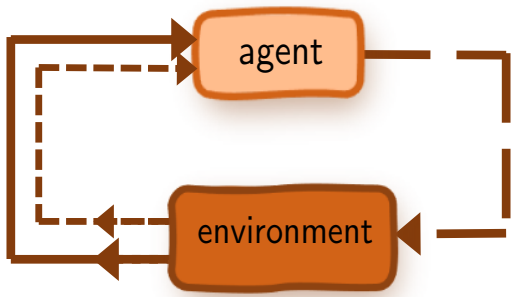
© 2000 Society for Industrial and Applied Mathematics

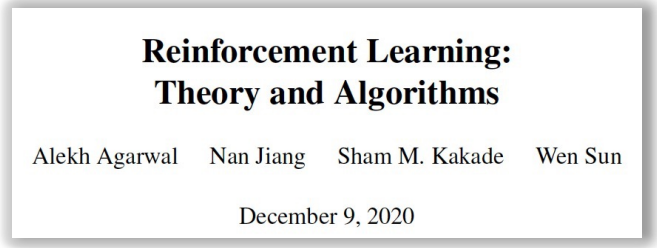
THE O.D.E. METHOD FOR CONVERGENCE OF STOCHASTIC APPROXIMATION AND REINFORCEMENT LEARNING*

V. S. BORKAR† AND S. P. MEYN‡



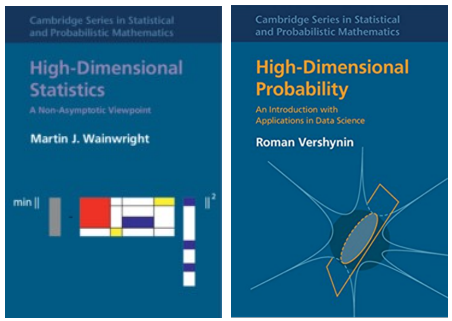





**Reinforcement Learning: Theory and Algorithms**

Alekh Agarwal Nan Jiang Sham M. Kakade Wen Sun

December 9, 2020







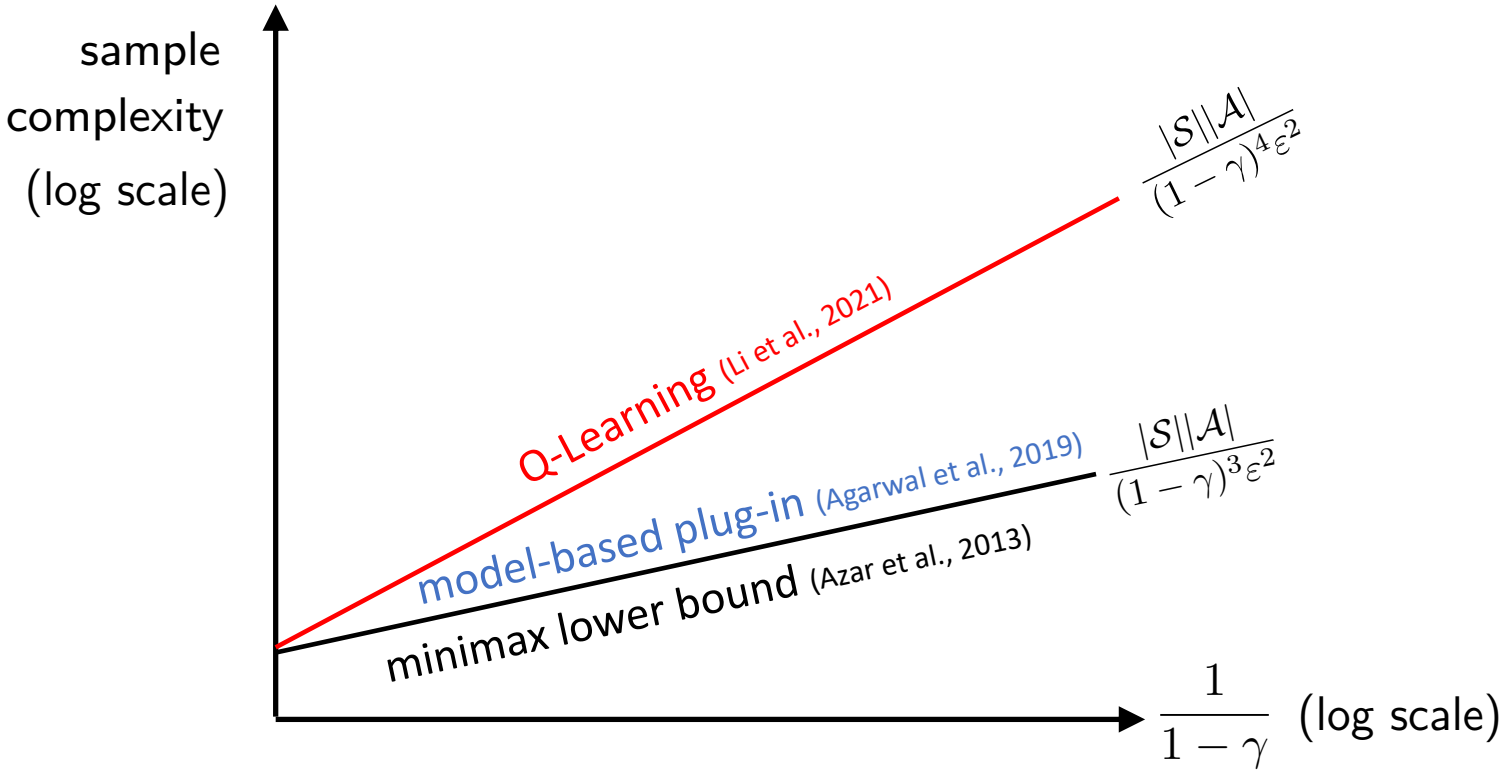
asymptotic analysis

finite-time & finite-sample analysis

1989

2020

# Statistical complexities of tabular RL

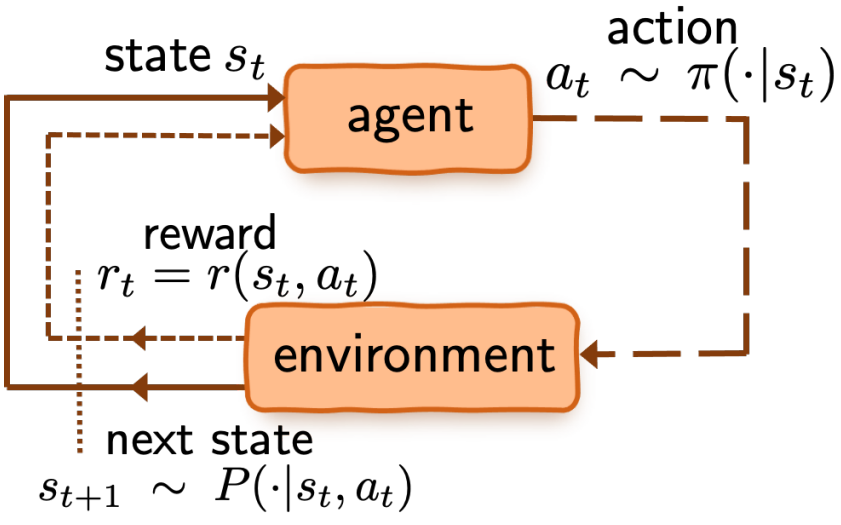


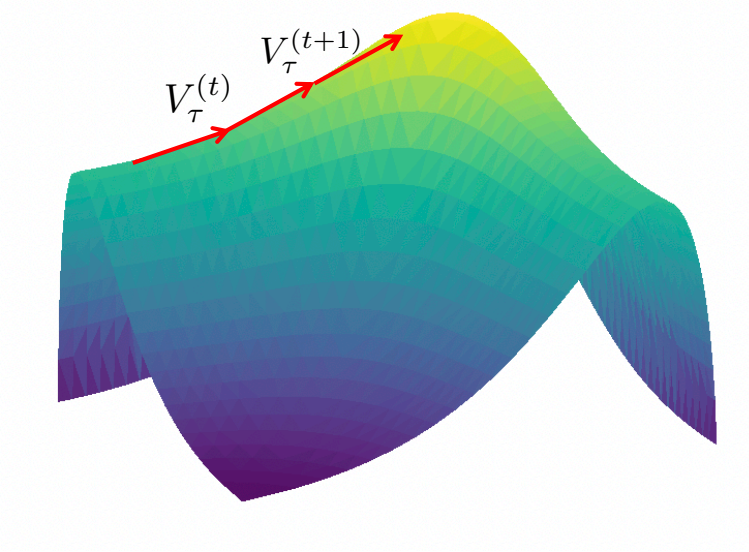

$\mathcal{S}$: state space, $\mathcal{A}$: action space, $\gamma$ discount factor,

$\epsilon$: entrywise accuracy

# Computational complexity of policy gradient methods

Little understanding on the global convergence of policy gradient methods until very recently, e.g. (Fazel et al., 2018; Bhandari and Russo, 2019; Agarwal et al., 2019; Mei et al. 2020; Cen et al. 2021), and many more.




Leveraging the problem structure to identify benign geometry for optimization.

# Summary

The confluence of high-dimensional statistics and large-scale optimization revamps the status quo in signal processing and machine learning.

- exploitation of finite-sum structures and hidden convexity
- fine-grained performance for computation outcomes
- new finite-sample and finite-time guarantees

**Look-ahead:** still a looooong way to close the theory-practice gap.

# Thank you for listening!

https://users.ece.cmu.edu/~yuejiec/