Stanford

ICME
INSTITUTE for COMPUTATIONAL & MATHEMATICAL ENGINEERING

On the Power and Limit of Scientific Machine Learning

Joint work with
Jose Blanchet, Jiajin Li, Jikai Jin, Haoxuan Chen, Lexing Ying…

Yiping Lu yplu@stanford.edu
https://2prime.github.io/

Two Disciplines in Science

Structural Model

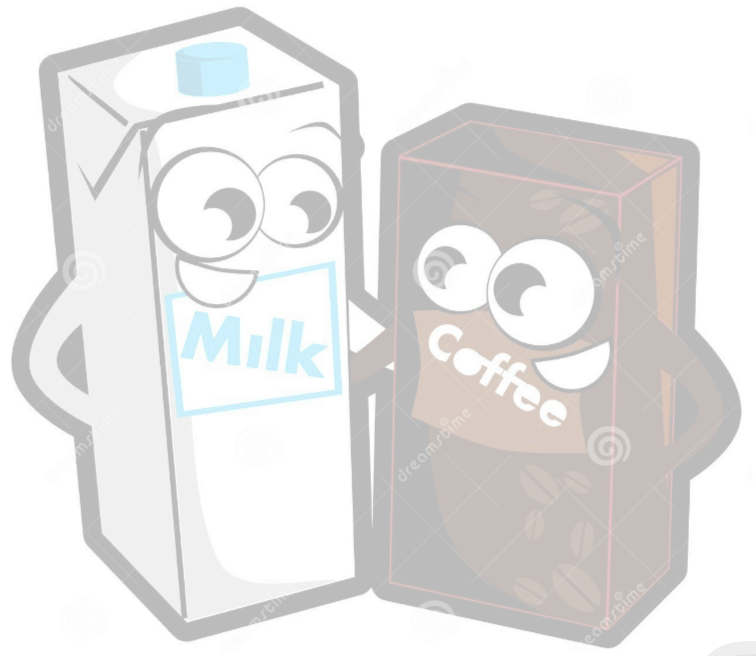
Differential equation modeling

Solving using numerical algorithms

Transparent

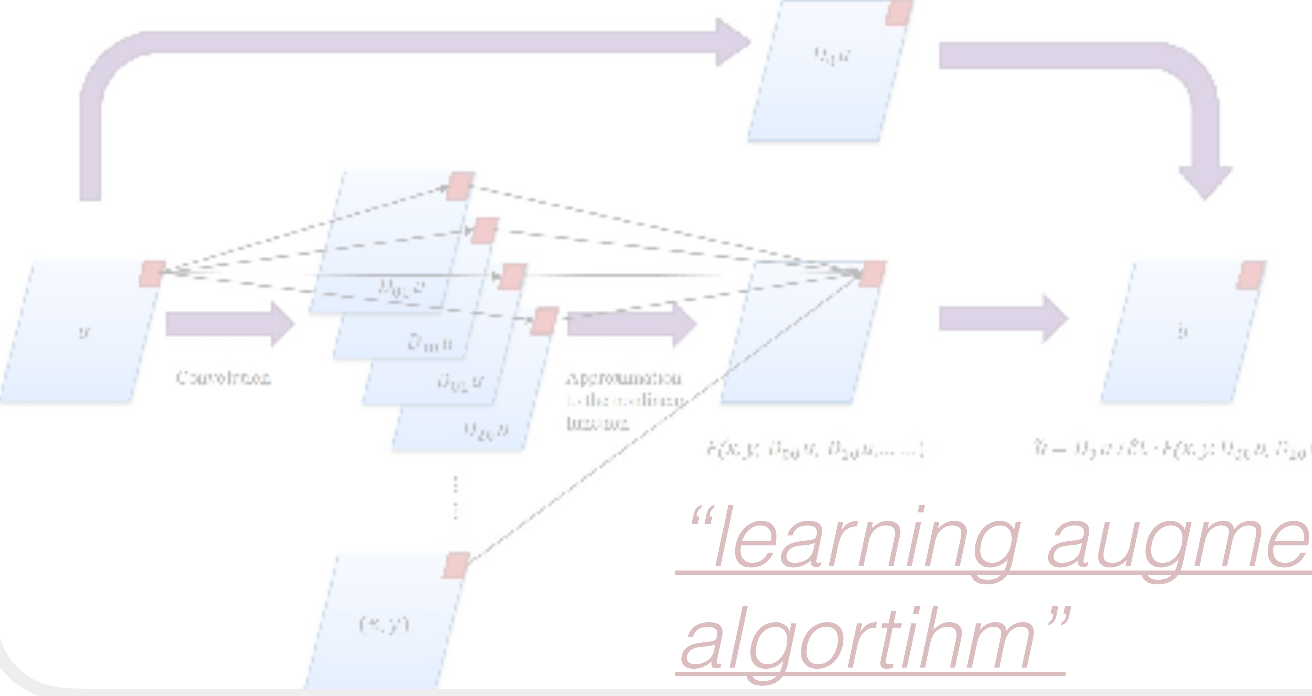
Lots of approximations
Limits the power

Make Useful Prediction



PDE-Net

[Long-Lu-Ma-Dong ICML2018]



"learning augmented algortihm"

Machine Learning

Data Collecting

Machine Learning

Flexible, Accurate

Blackbox
Data intensive

Two Disciplines in Science

Structural Model

Differential equation modeling

Solving using numerical algorithms

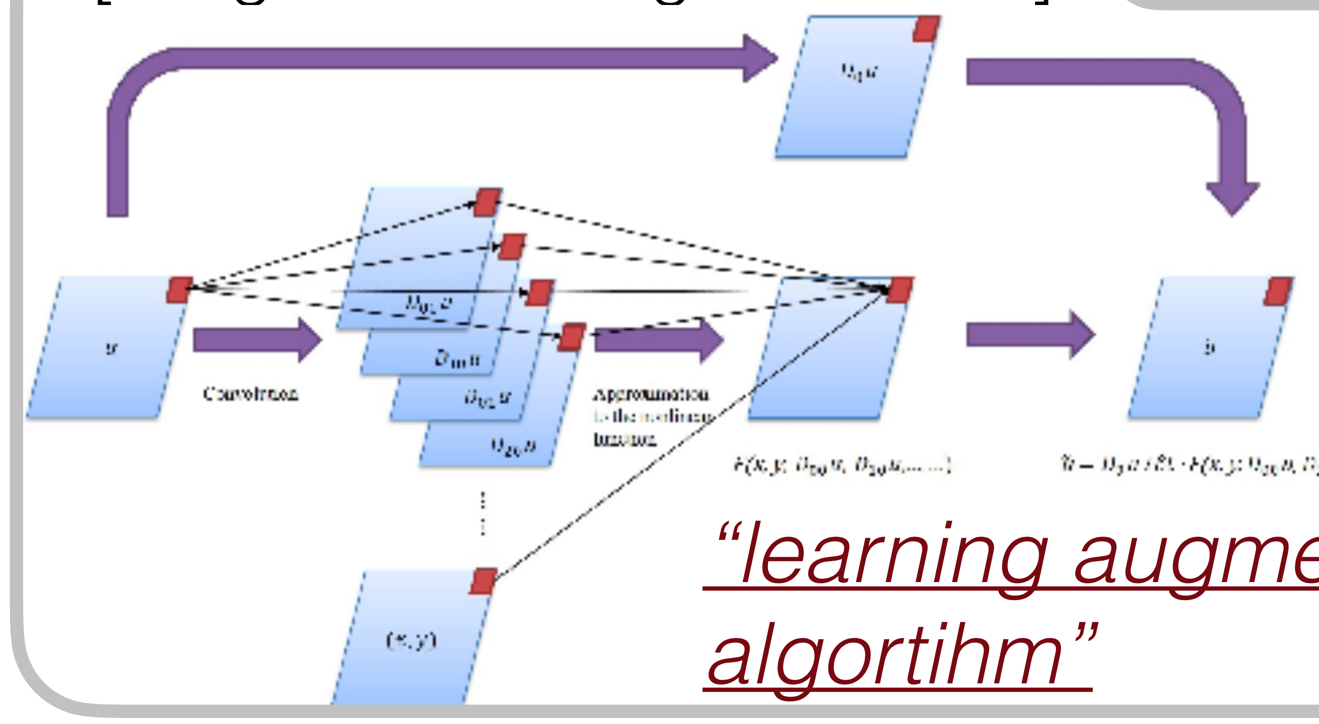
Transparent

Lots of approximations
Limits the power

Make Useful Prediction

PDE-Net

[Long-**Lu**-Ma-Dong ICML2018]



"learning augmented algortihm"

Machine Learning

Data Collecting

Machine Learning

Flexible, Accurate

Blackbox
Data intensive

Two Disciplines in Science

Can we understand it theortically?

Just ML + physic data?

Structural Model

Differential equation modeling

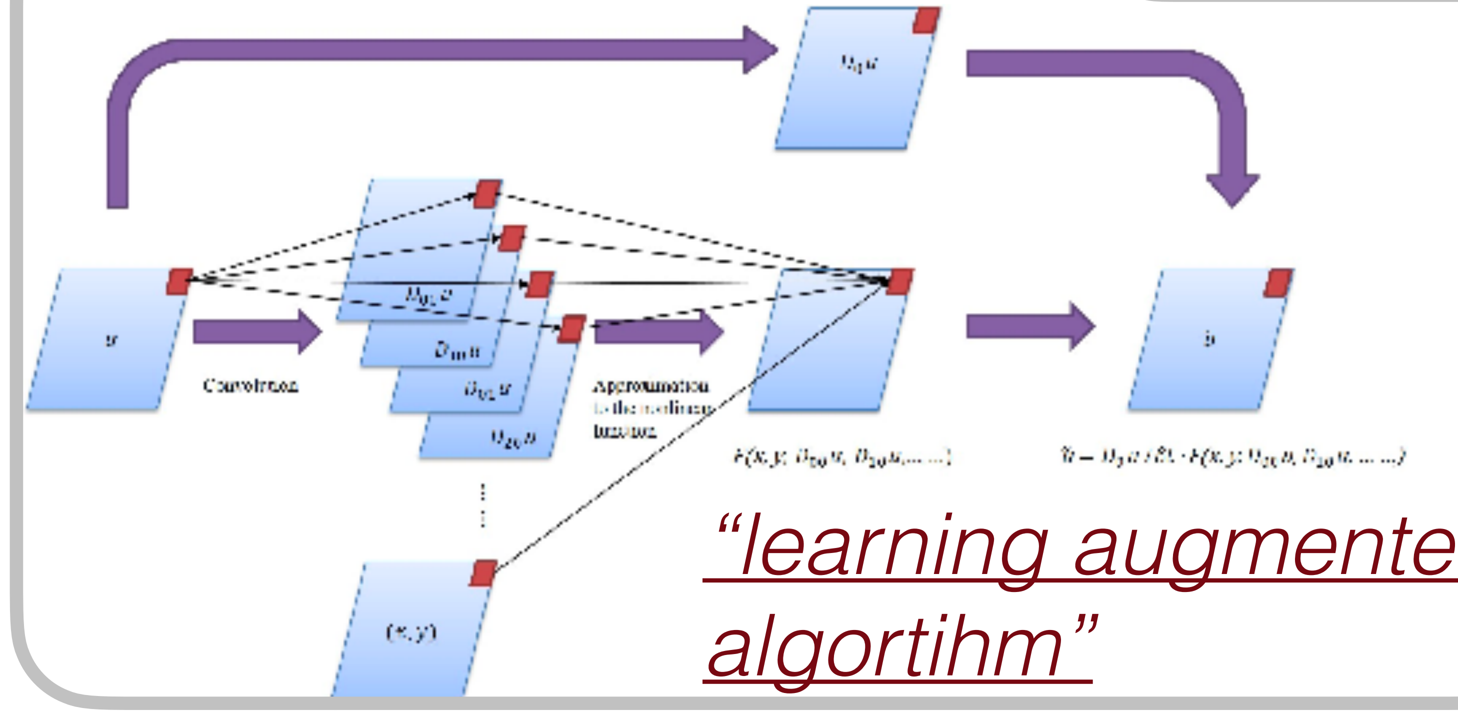
Solving using numerical algorithms

Transparent

Lots of approximations
Limits the power

PDE-Net

[Long-**Lu**-Ma-Dong ICML2018]



"learning augmented algortihm"

Machine Learning

Data Collecting

Machine Learning

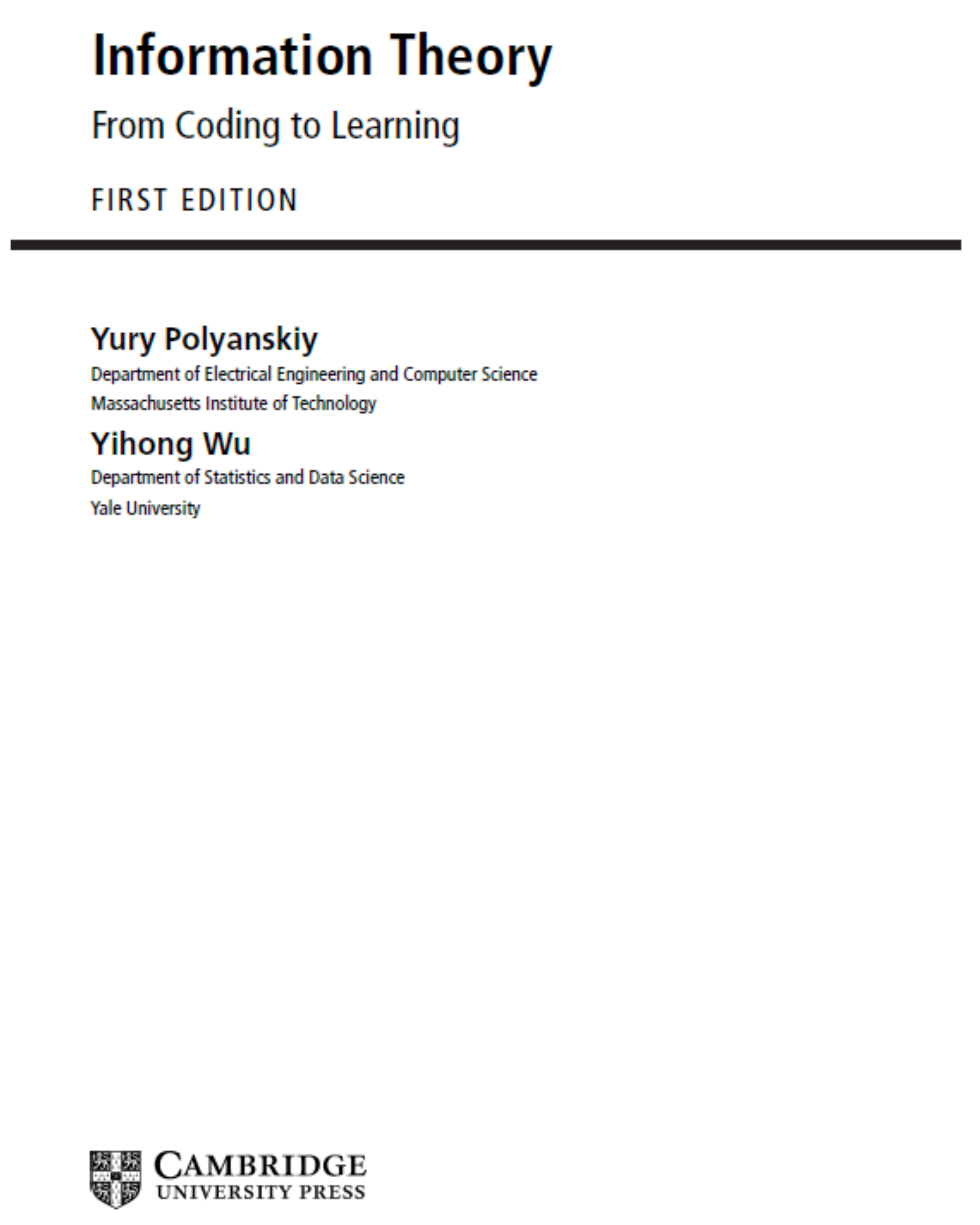
Flexible, Accurate

Blackbox
Data intensive

Machine Learning Research

Aim: fit function $(x_i, y_i = \boxed{f}(x_i)),\ i = 1,2,\cdots,n$

Specify problem set, i.e. the space of f


Information Theory
From Coding to Learning
FIRST EDITION

Yury Polyanskiy
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Yihong Wu
Department of Statistics and Data Science
Yale University

CAMBRIDGE
UNIVERSITY PRESS

Step 1 Information-Theoretical Lower Bound

Step 2 Statistical guarantee for the estimator

"Minimax Optimal" Algorithms

"worst case selection of f"

Best Estimator

Why we have a lower bound?

For all estimator $H : (\text{data})^{\otimes n} \to$ function, we have

$$\sup_{\boxed{f \in \mathcal{F}}} \mathbb{E}_{\text{data}_i \sim f} \| H(\text{data}_1, \cdots, \text{data}_n) - f \| \geq n^{\text{rate}}$$

$\|f\| < 1$

Using information theory

1. Generate similar data (in TV, KL….)
2. f_1 and f_2 have a $\boxed{\text{gap}}$

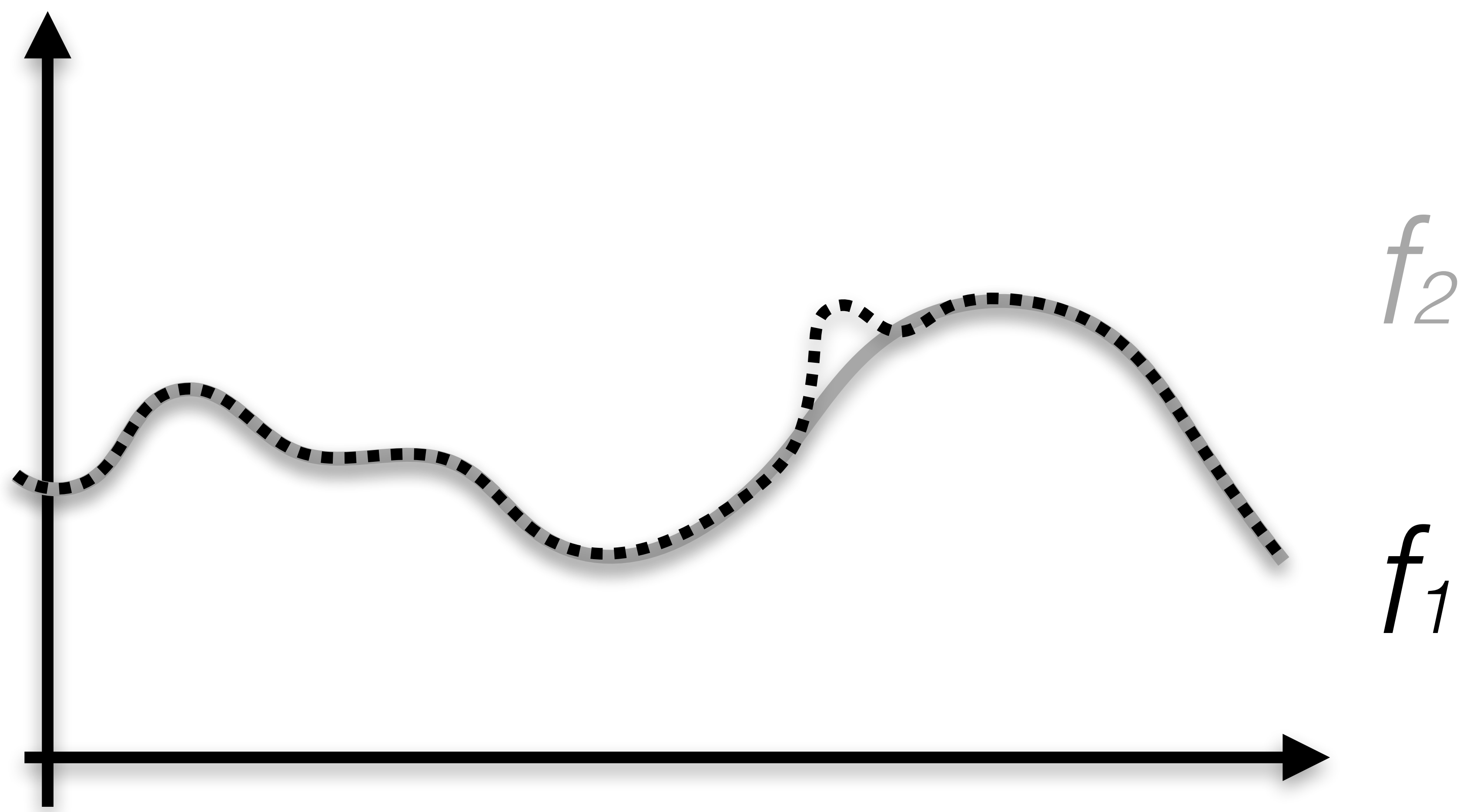
The gap is not distinguishable

Why we have a lower bound?

For all estimator $H : (\text{data})^{\otimes n} \to$ function, we have

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{\text{data}_i \sim f} \| H(\text{data}_1, \cdots, \text{data}_n) - f \| \geq n^{\text{rate}}$$

$\|f\| < 1$



Using information theory

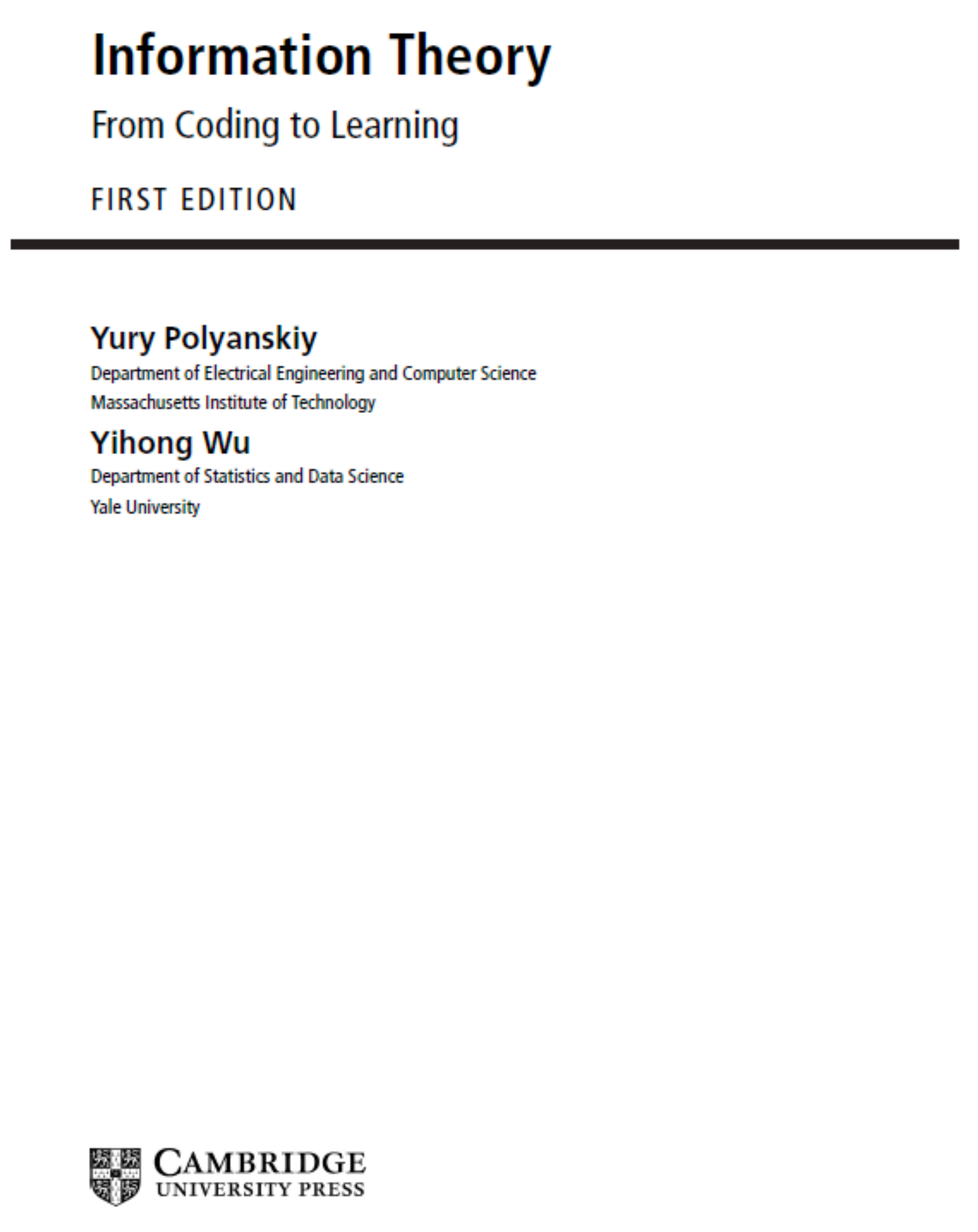
1. Generate similar data (in TV, KL…)
2. f_1 and f_2 have a $\boxed{\text{gap}}$

The gap is not distinguishable

Machine Learning Research

Aim: fit function $(x_i, y_i = \boxed{f}(x_i)),\ i = 1,2,\cdots,n$

Specify problem set, i.e. the space of f


Information Theory
From Coding to Learning
FIRST EDITION
Yury Polyanskiy
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Yihong Wu
Department of Statistics and Data Science
Yale University
CAMBRIDGE
UNIVERSITY PRESS

Step 1 Informatio heoretical Lower Bound

Step 2 Statistical rantee for the estimator

"Mini max Opt al" Algorithms

"worst case sel n of f"

Best Estimator

What is the task of scientific machine learning?

Step 0 Specify your task!

Not Just Differential Equation models

"Physic" Model

$$\boxed{\boxed{A}u = f}$$

Not Just Differential Equation models

"Physic" Model

$$Au = f$$

Hamilton Jacobi Equation

Value Function

Reward Function

Kolomoglov Equation

Committor function

Boundary Condition

Incentive Model
Super-martingale OT

Pricing policy/tax

Agent Utility Distribution

Current Research

$Au = f$

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19]

Current Research

$Au = f$

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19]

Learn from data pair $\{u_i, f_i\}$
"Operator Learning/Functional data analysis"

Methodology

[Brunton-Proctor-Kutz 16][Khoo-Lu-Ying 18] [Long-Lu-Li-Dong 18][Lu-Jin-Pang-Zhang-Karniadakis 20] [Li-Kovachki-…-Stuart-Anandkumar 20]

Theory

[Lanthaler-Mishra-Karniadakis 22] [Talwai-Shameli-Simchi-Levi 21][de Hoop-Kovachki-Nelsen-Stuart 21][Li-Meunier-Mollenhauer-Gretton 22] [Liu-Yang-Chen-Zhao-Liao 22]….

Current Research

$Au = f$

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19]

Learn from data pair $\{u_i, f_i\}$
"Operator Learning/Functional data analysis"

Methodology

[Brunton-Proctor-Kutz 16][Khoo-Lu-Ying 18] [Long-Lu-Li-Dong 18][Lu-Jin-Pang-Zhang-Karniadakis 20] [Li-Kovachki-…-Stuart-Anandkumar 20]

Theory

[Lanthaler-Mishra-Karniadakis 22] [Talwai-Shameli-Simchi-Levi 21][de Hoop-Kovachki-Nelsen-Stuart 21][Li-Meunier-Mollenhauer-Gretton 22] [Liu-Yang-Chen-Zhao-Liao 22]….

Recover parameter θ in model A_θ
E.g. Drift, Diffusion Strength

[Brunton-Proctor-Kutz 16] [Long-Lu-Dong 20] [Liang-Yang 22]…

[Nickl-Ray 20] [Nickl 20] [Baek-Farias-Georgescu-Li-Peng-Sinha-Wilde-Zheng 20] [Agrawl-Yin-Zeevi 21]…

Scientific Machine Learning Research

Aim: fit function $(x_i, y_i = \boxed{f}(x_i)),\ i = 1,2,\cdots,n$

Specify problem set, i.e. the space of f

Step 1 Information-Theoretical Lower Bound

Step 2 Statistical guarantee for the estimator

"Minimax Optimal" Algorithms

"worst case function of f"

Best Estimator

Information Theory

From Coding to Learning

FIRST EDITION

Yury Polyanskiy
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Yihong Wu
Department of Statistics and Data Science
Yale University

Cambridge University Press

Physical Equation $Au = f$

Reconstruct u with observation of f: $\{x_i, f(x_i)\}$	Recover parameter θ in Model A_θ	Learn the model A from data pair $\{u_i, f_i\}$

Scientific Machine Learning Research

Aim: fit function $(x_i, y_i = \boxed{f}(x_i)),\ i = 1,2,\cdots,n$

Specify problem set, i.e. the space of f

Step 1 Information-Theoretical Lower Bound

Step 2 Statistical guarantee for the estimator

"Minimax Optimal" Algorithms

"worst case function of f"

Best Estimator

Information Theory
From Coding to Learning
FIRST EDITION

Yury Polyanskiy
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Yihong Wu
Department of Statistics and Data Science
Yale University

CAMBRIDGE UNIVERSITY PRESS

Standard approximation and statistical exercises?

Physical Equation $Au = f$

Reconstruct u with observation of f: $\{x_i, f(x_i)\}$	Recover parameter θ in Model A_θ	Learn the model A from data pair $\{u_i, f_i\}$

Scientific Machine Learning Research

Aim: fit function $(x_i, y_i = \boxed{f}(x_i)), i = 1,2,\cdots,n$

Specify problem set i.e. the space of f

New insights for:

Operator learning
Solving PDE
Quadrature Rule

Step 1 Information-Theoretical Lower Bound

Step 2 Statistical guarantee for the estimator

"Minimax Optimal" Algorithms

"worst case function of f"

Best Estimator

Information Theory
From Coding to Learning
FIRST EDITION

Yury Polyanskiy
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Yihong Wu
Department of Statistics and Data Science
Yale University

CAMBRIDGE UNIVERSITY PRESS

Physical Equation $Au = f$

Reconstruct u with observation of f: $\{x_i, f(x_i)\}$	Recover parameter θ in Model A_θ	Learn the model A from data pair $\{u_i, f_i\}$

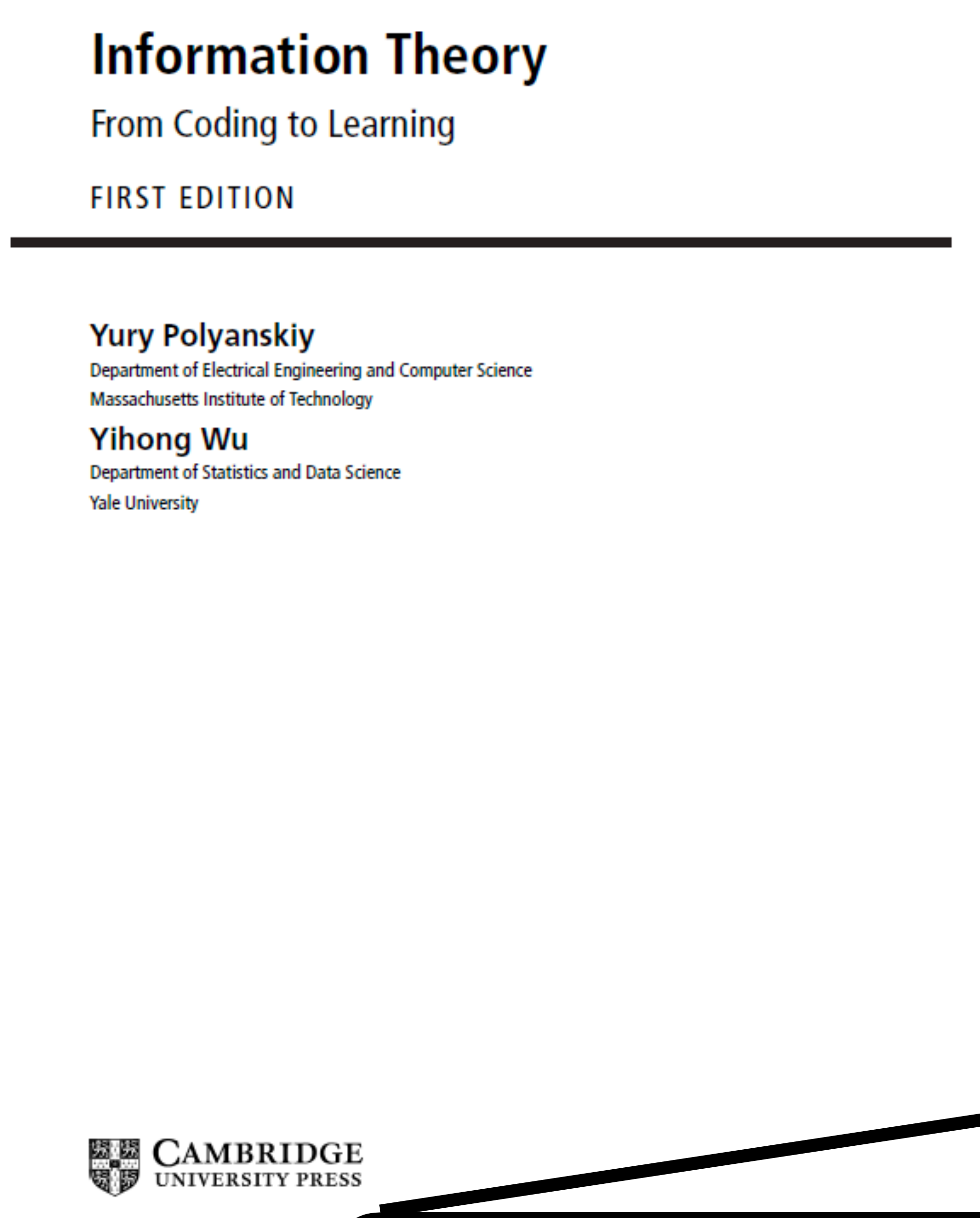
Scientific Machine Learning Research

Aim: fit function $(x_i, y_i = \boxed{f}(x_i)),\ i = 1,2,\cdots,n$

Specify problem set, i.e. the space of f

New insights for:

Operator learning
Solving PDE
Quadrature Rule


Information Theory
From Coding to Learning
FIRST EDITION
Yury Polyanskiy
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Yihong Wu
Department of Statistics and Data Science
Yale University
CAMBRIDGE UNIVERSITY PRESS

Step 1 Information-Theoretical Lower Bound

Step 2 Statistical guarantee for the estimator

"$\boxed{\text{Mini}}\boxed{\text{max}}$ Optimal" Algorithms

"worst case function of f"

Best Estimator

Fundamental difference between finite dimension and infinite dimension machine learning

Physical Equation $Au = f$

Reconstruct u with observation of f: $\{x_i, f(x_i)\}$	Recover parameter θ in Model A_θ	Learn the model A from data pair $\{u_i, f_i\}$

Scientific Machine Learning Research

Aim: fit function $(x_i, y_i = \boxed{f}(x_i)),\ i = 1,2,\cdots,n$

Specify problem set, i.e. the space of f

Step 1 Information-Theoretical Lower Bound

Step 2 Statistical guarantee for the estimator

"$\boxed{\text{Mini}}\boxed{\text{max}}$ Optimal" Algorithms

"worst case function of f"

Best Estimator

New insights for:

Operator learning

Solving PDE

Quadrature Rule

New technique for semi-parametric statistic via sobolev embedding

Information Theory

From Coding to Learning

FIRST EDITION

Yury Polyanskiy

Department of Electrical Engineering and Computer Science

Massachusetts Institute of Technology

Yihong Wu

Department of Statistics and Data Science

Yale University

CAMBRIDGE UNIVERSITY PRESS

Physical Equation $Au = f$

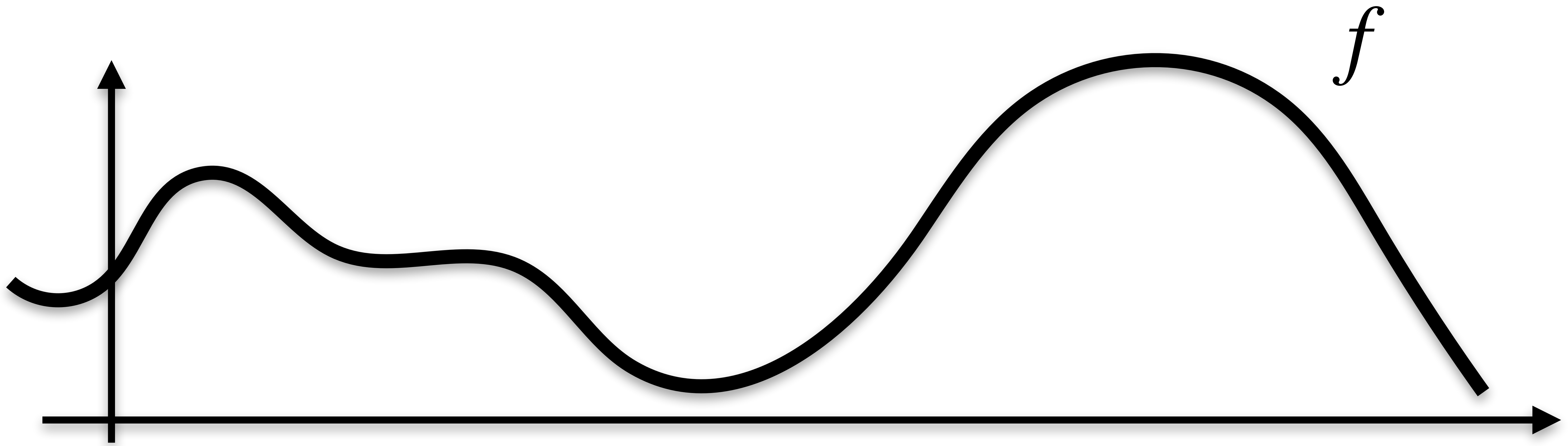
Reconstruct u with observation of f: $\{x_i, f(x_i)\}$	Recover parameter θ in Model A_θ	Learn the model A from data pair $\{u_i, f_i\}$

Optimal Quadrature Rule via ML

https://arxiv.org/abs/2305.16527

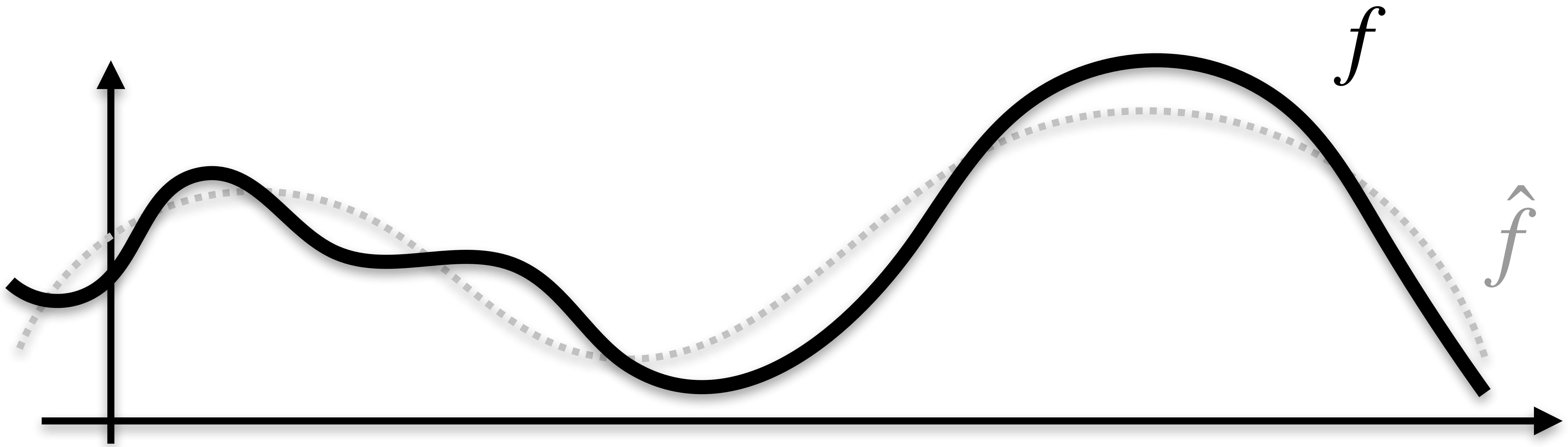
Quadrature Rule

Aim Estimate $\mathbb{E}_P f$



Quadrature Rule

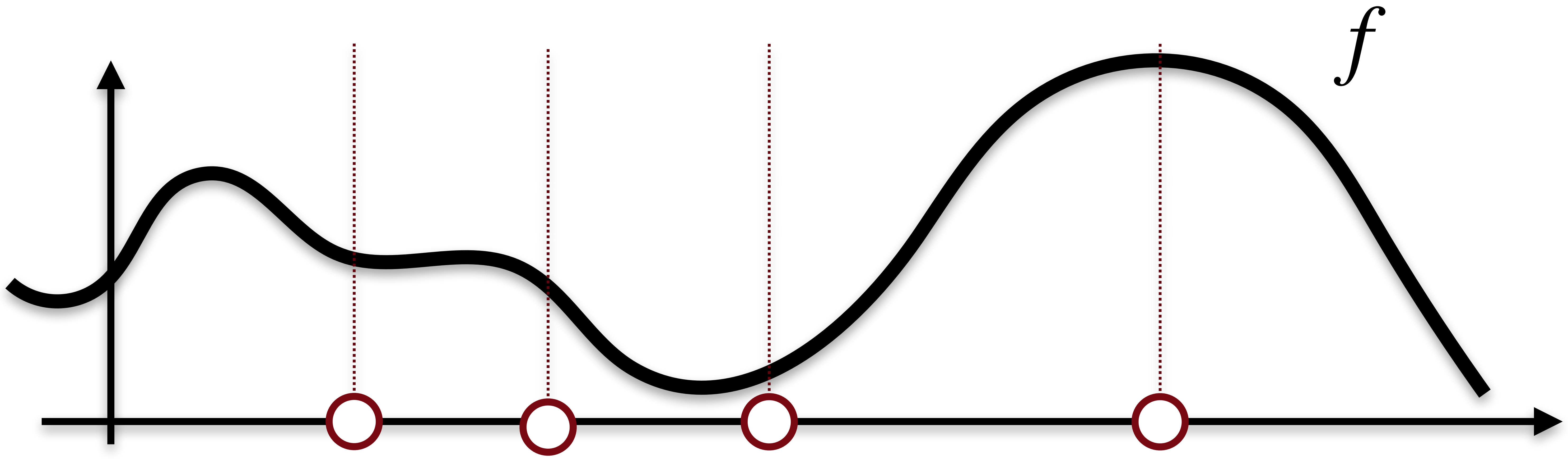
Aim Estimate $\mathbb{E}_P f \approx \mathbb{E}_P \hat{f}$



"piece-wise polynomial"

Quadrature Rule via Monte Carlo

Aim Estimate $\mathbb{E}_P f \approx \mathbb{E}_{\hat{P}} f$



Aim Estimate $\mathbb{E}_P f \approx \mathbb{E}_P \hat{f}$

$$\approx \mathbb{E}_{\hat{P}} f$$

$$\begin{aligned} xy &= x\hat{y} + x(y - \hat{y}) \\ &= \hat{x}y + y(x - \hat{x}) \end{aligned}$$

$$xy = \hat{x}y + \hat{y}x - \hat{x}\hat{y} + \boxed{(y - \hat{y})(x - \hat{x})}$$

Smaller error

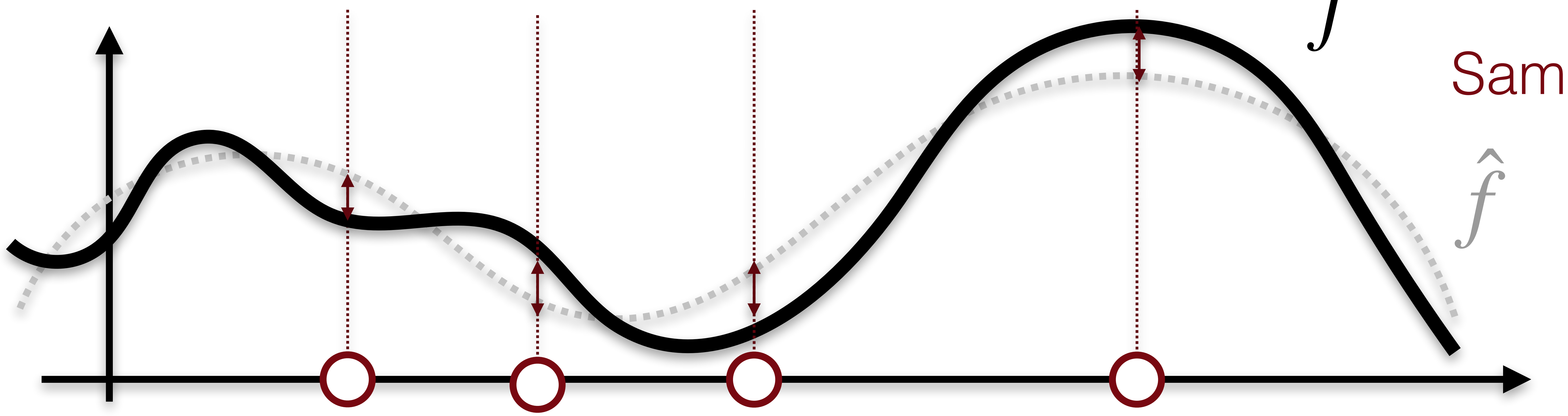
Quadrature Rule

Aim Estimate $\mathbb{E}_P f = \mathbb{E}_P \hat{f} + \boxed{\mathbb{E}_P (f - \hat{f})}$

Debiasing

“semi-"parametric

Sample extra data to know $f - \hat{f}$

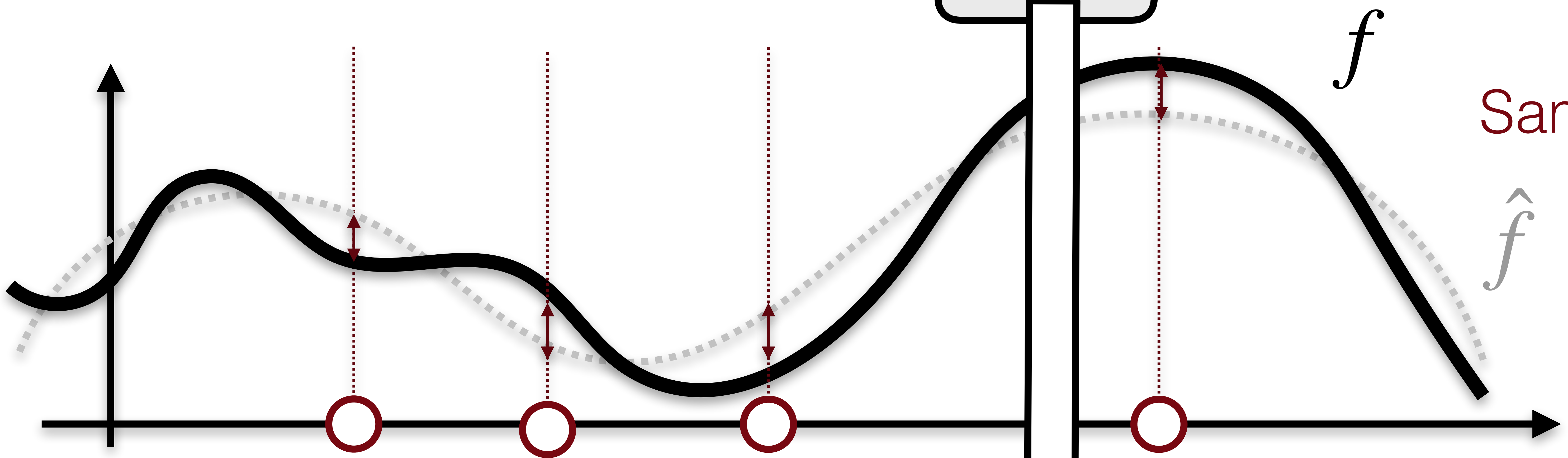


Quadrature Rule

Aim Estimate $\mathbb{E}_P f = \boxed{\mathbb{E}_P \hat{f}} + \mathbb{E}_P(f - \hat{f})$

Debiasing

“semi-"parametric



Sample extra data to know $f - \hat{f}$

(nonparametric-)“Regression-adjusted” control variate

"Modern" regression-adjusted cv

Trace estimation:

Hutch++ Lin 17 Numerische Mathematik Mewyer-Musco-Musco-Woodruff 20

Dimension Reduction:

Sobczyk and Luisier Neuips 22

Conformal Prediction:

Conformalized quantile regression Romano-Patterson-Candes Neurips 19

Gradient Estimation

Shi-Zhou-Hwang-Tisias-Mackey Neurips 22 outstanding paper

Causal Inference:

Double Robust estimation ….

"Qudrature" Rule (Today)

Bootstrapping, sketching….

Understanding this statistically…

Is this algorithm statistical optimal?

When this improves MC estimator?

Aim Estimate $\mathbb{E}_P f$

Step 1 Using half of the data to estimate $\hat{f}$

Step 2 $\mathbb{E}_P f = \mathbb{E}_P(\hat{f}) + \mathbb{E}_P \boxed{(f - \hat{f})}$

Low order term

Understanding this statistically…

Is this algorithm statistical optimal?

Why consider $q-$th moment?

When this improves MC estimator?

Why consider $W^{s,p}$?

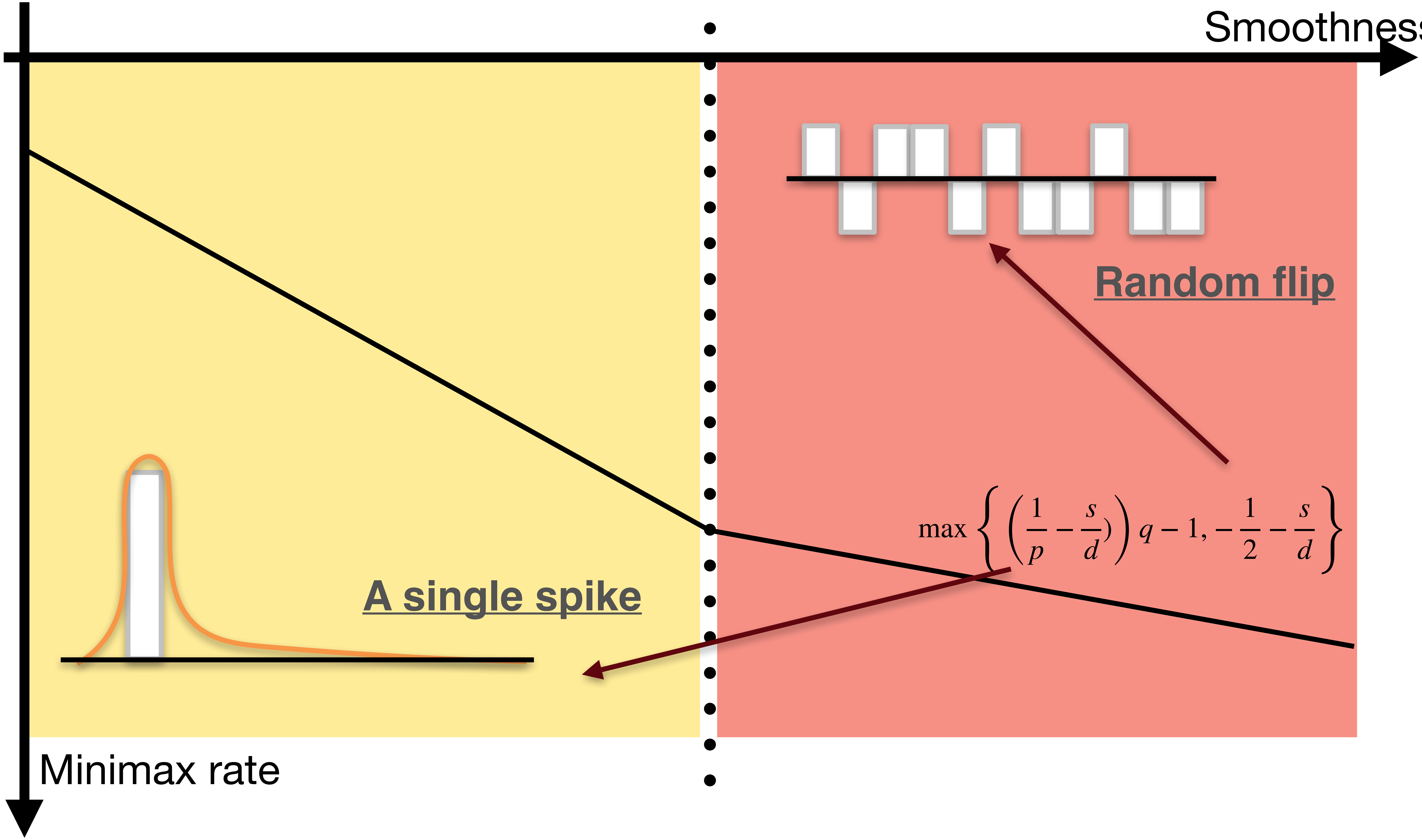
Aim Estimate $\cancel{\mathbb{E}_P f}$ $\quad \mathbb{E}_P f^q, f \in W^{s,\boxed{p}}$

Step 1 Using half of the data to estimate $\hat{f}$

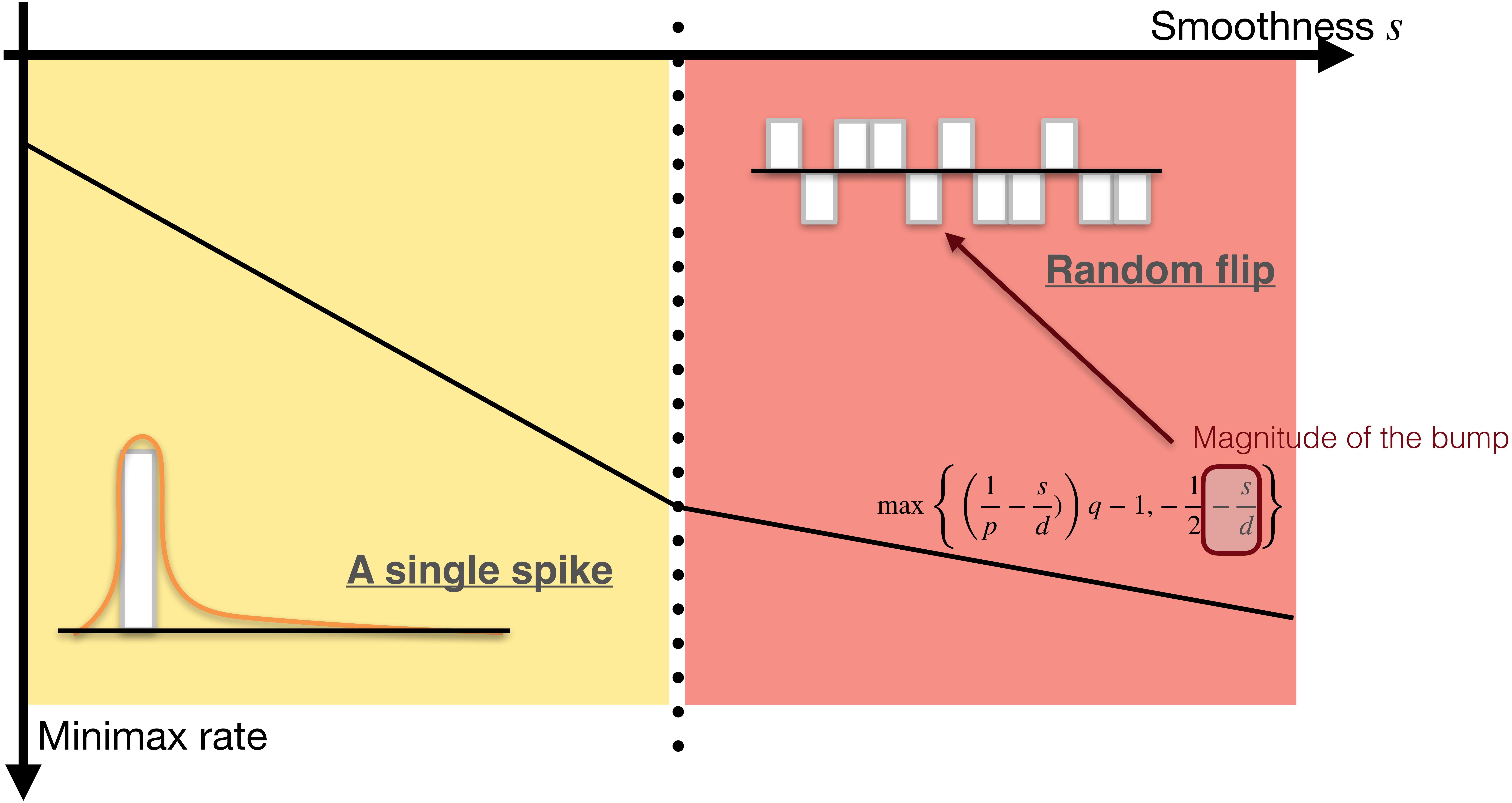
Step 2 $\mathbb{E}_P f^q = \mathbb{E}_P (\hat{f})^q + \mathbb{E}_P \boxed{(f^q - \hat{f}^q)}$

Low order term

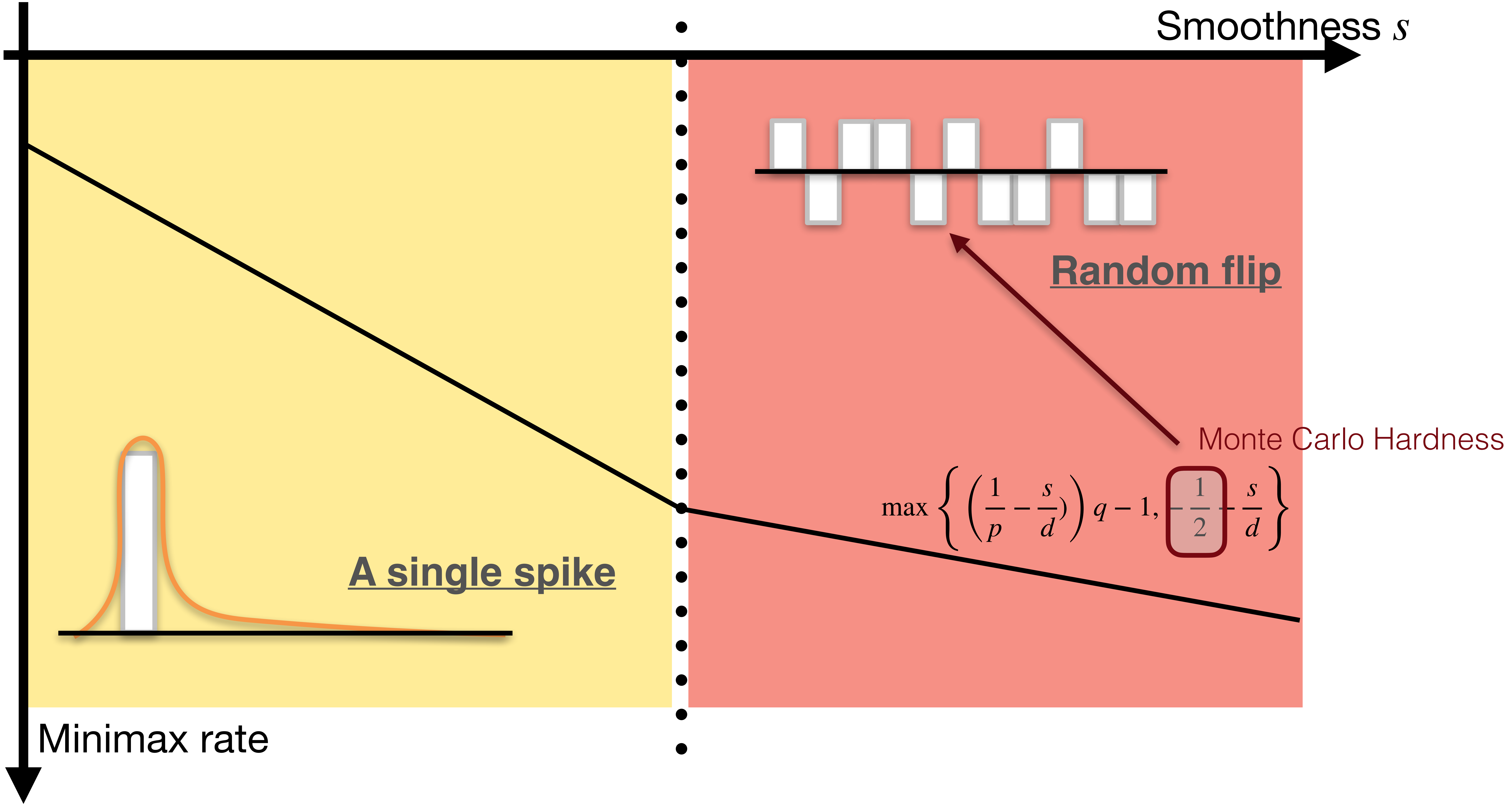
Setting the information theoretical limit



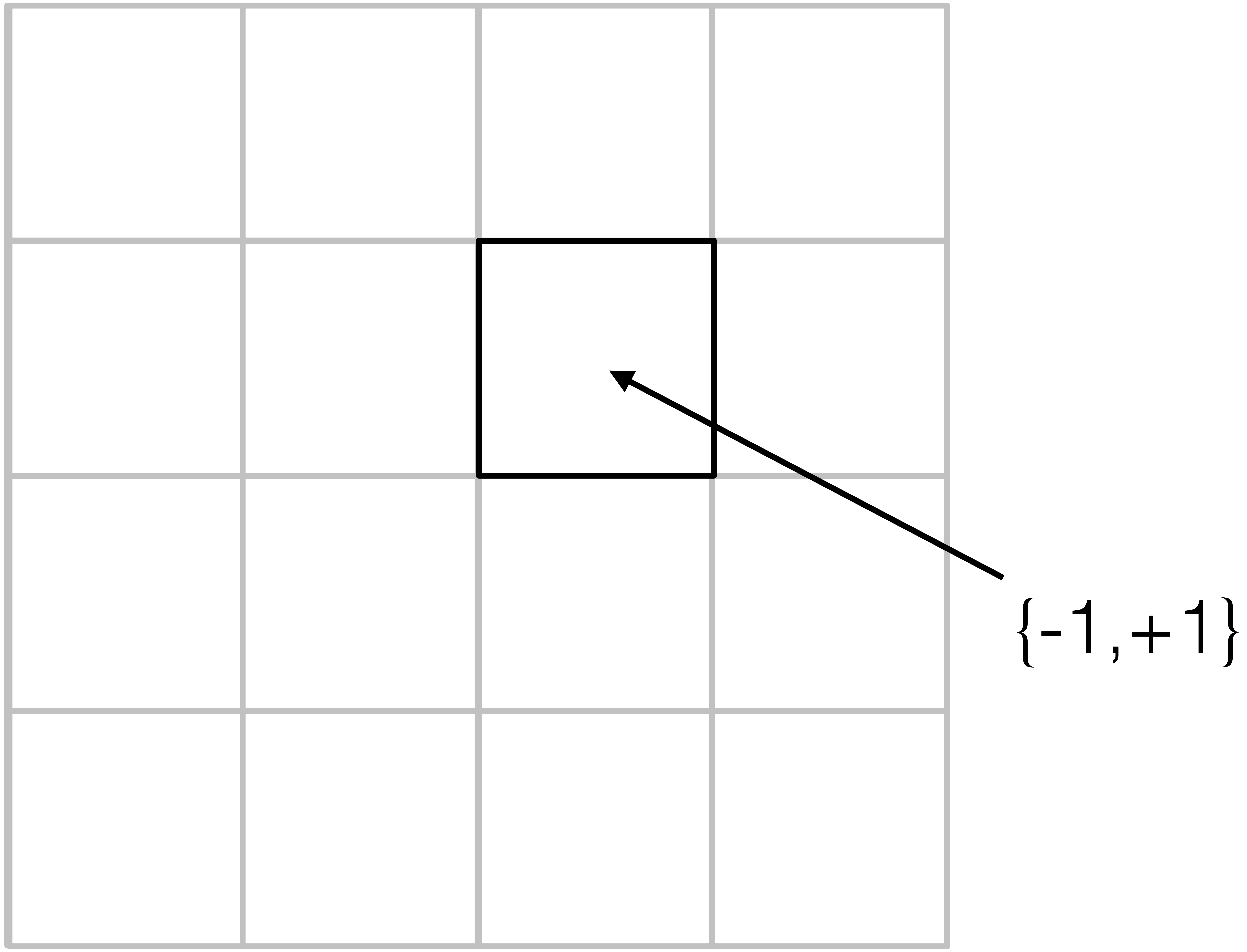
Setting the information theoretical limit



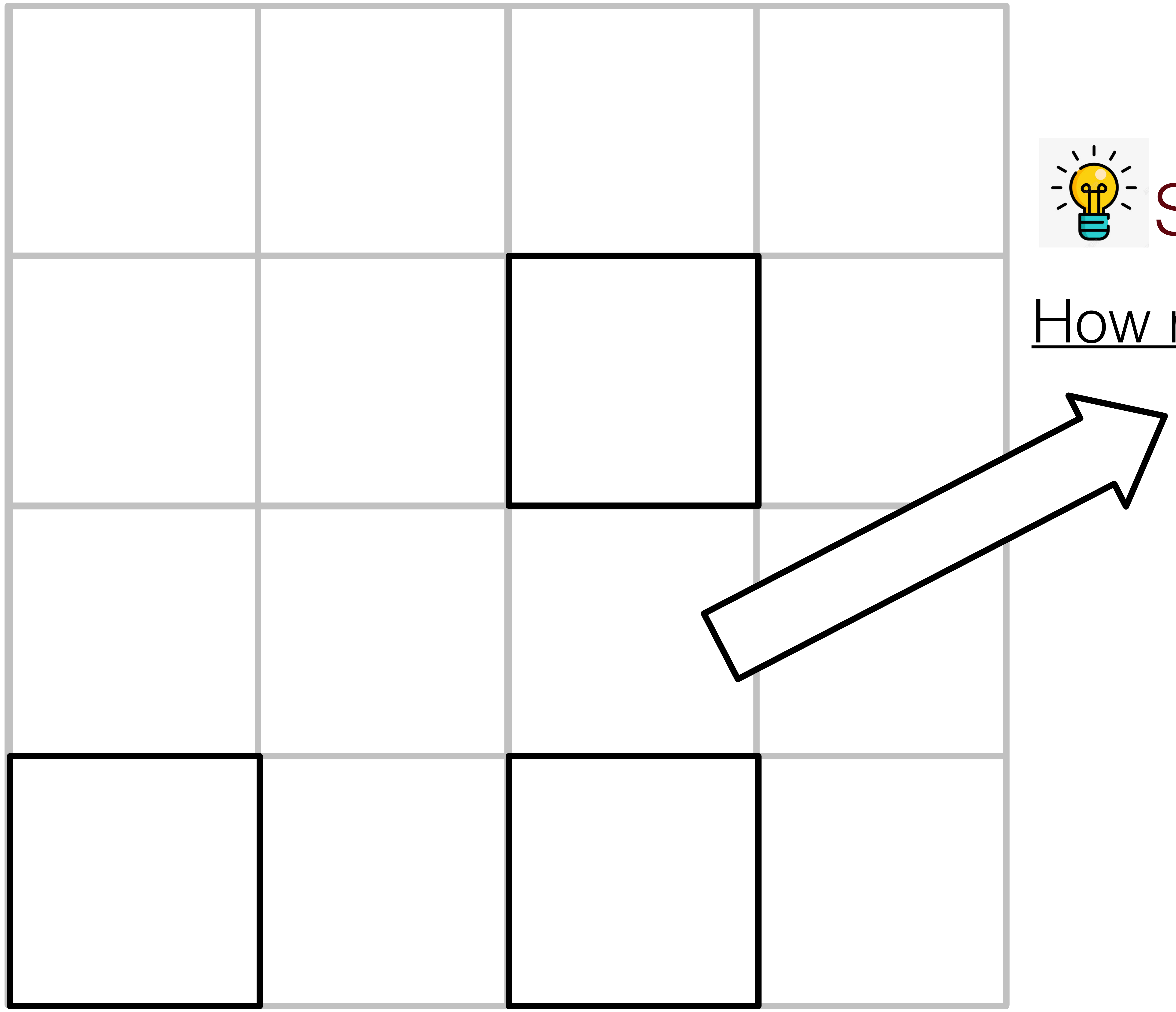
Setting the information theoretical limit



Understanding the hardness in this regime



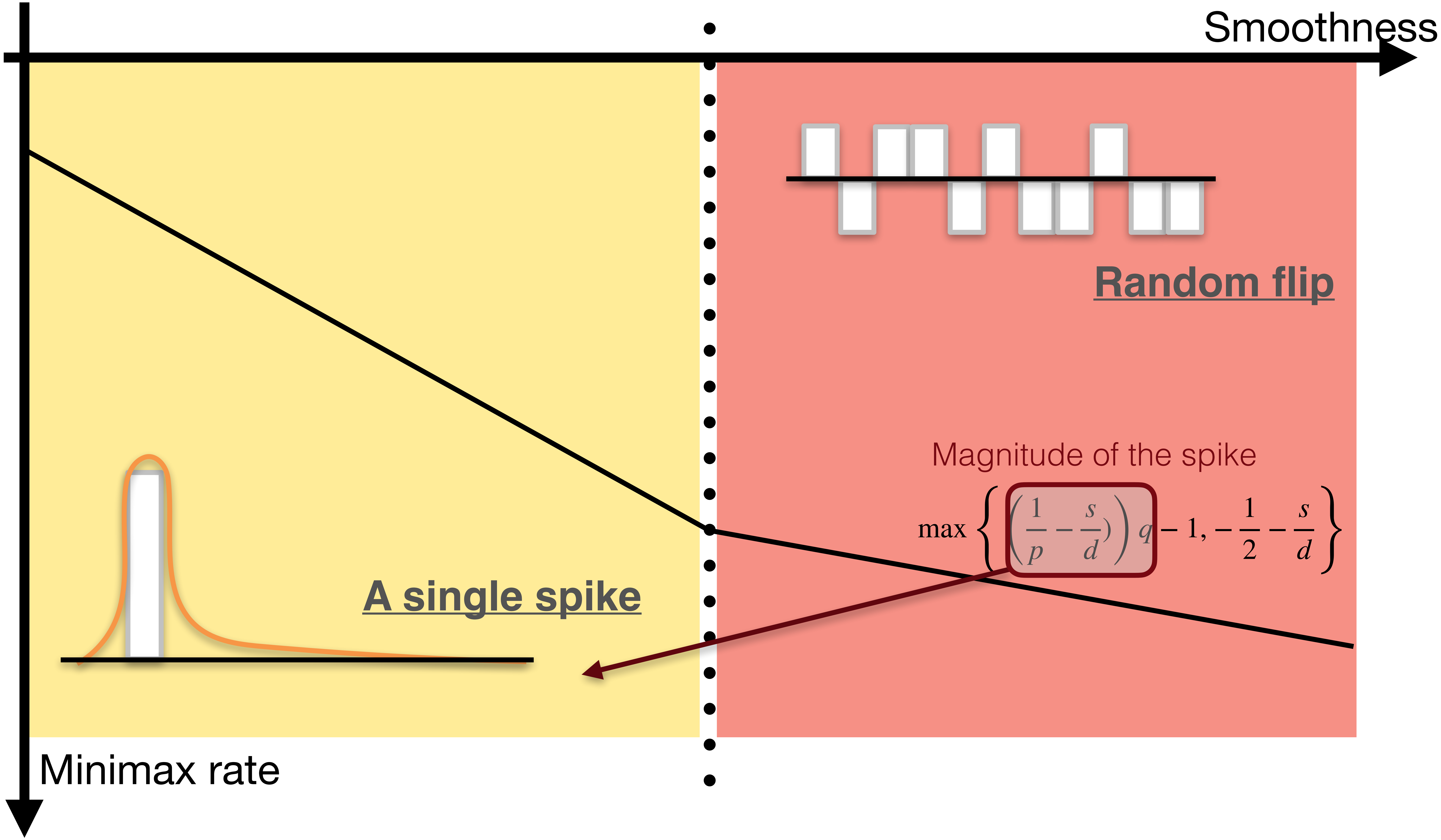
Understanding the hardness in this regime



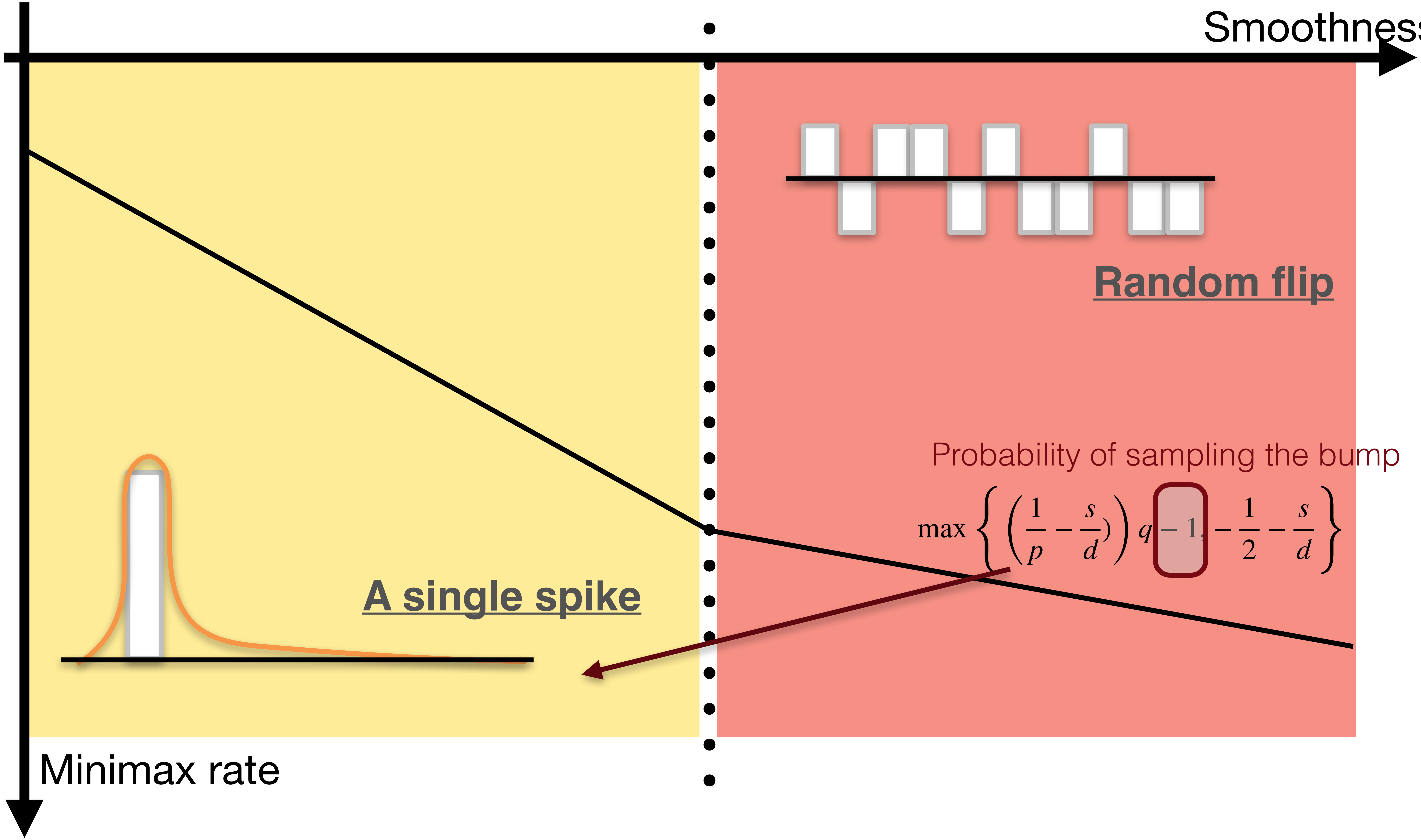
Sample without replacement

<u>How many boxes have quadrature points?</u>

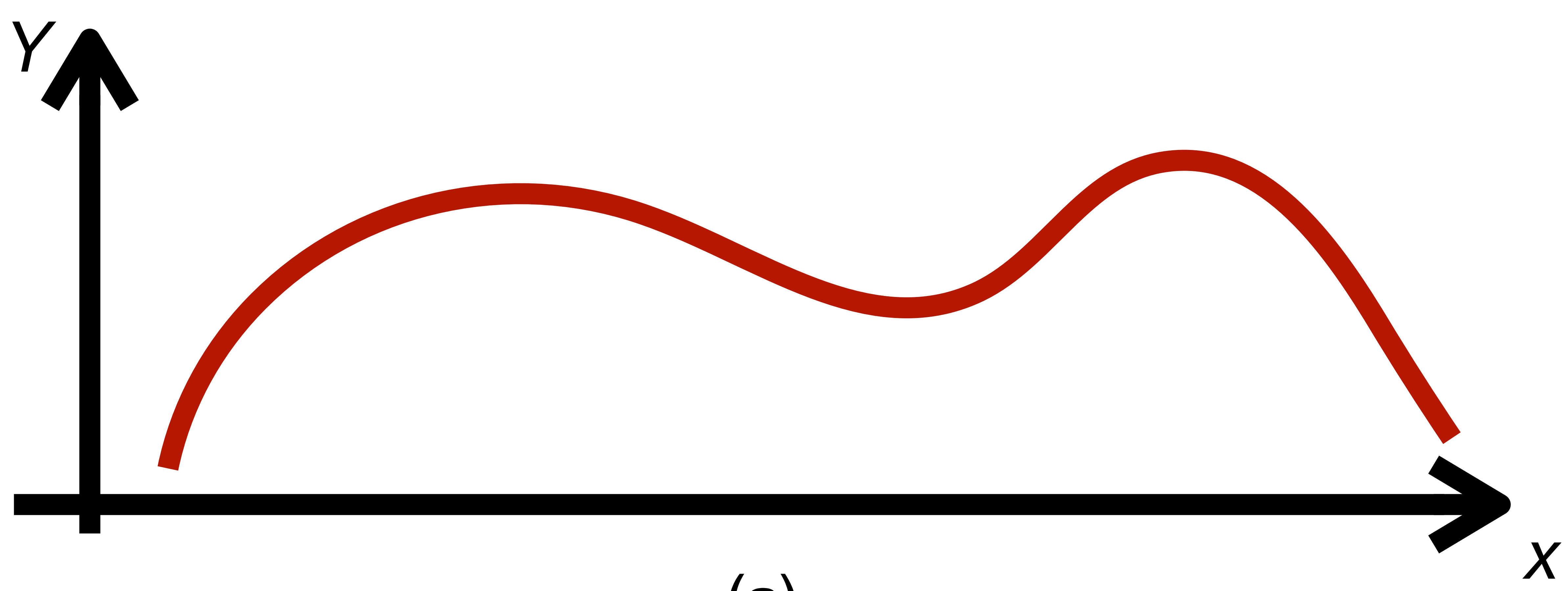
Setting the information theoretical limit

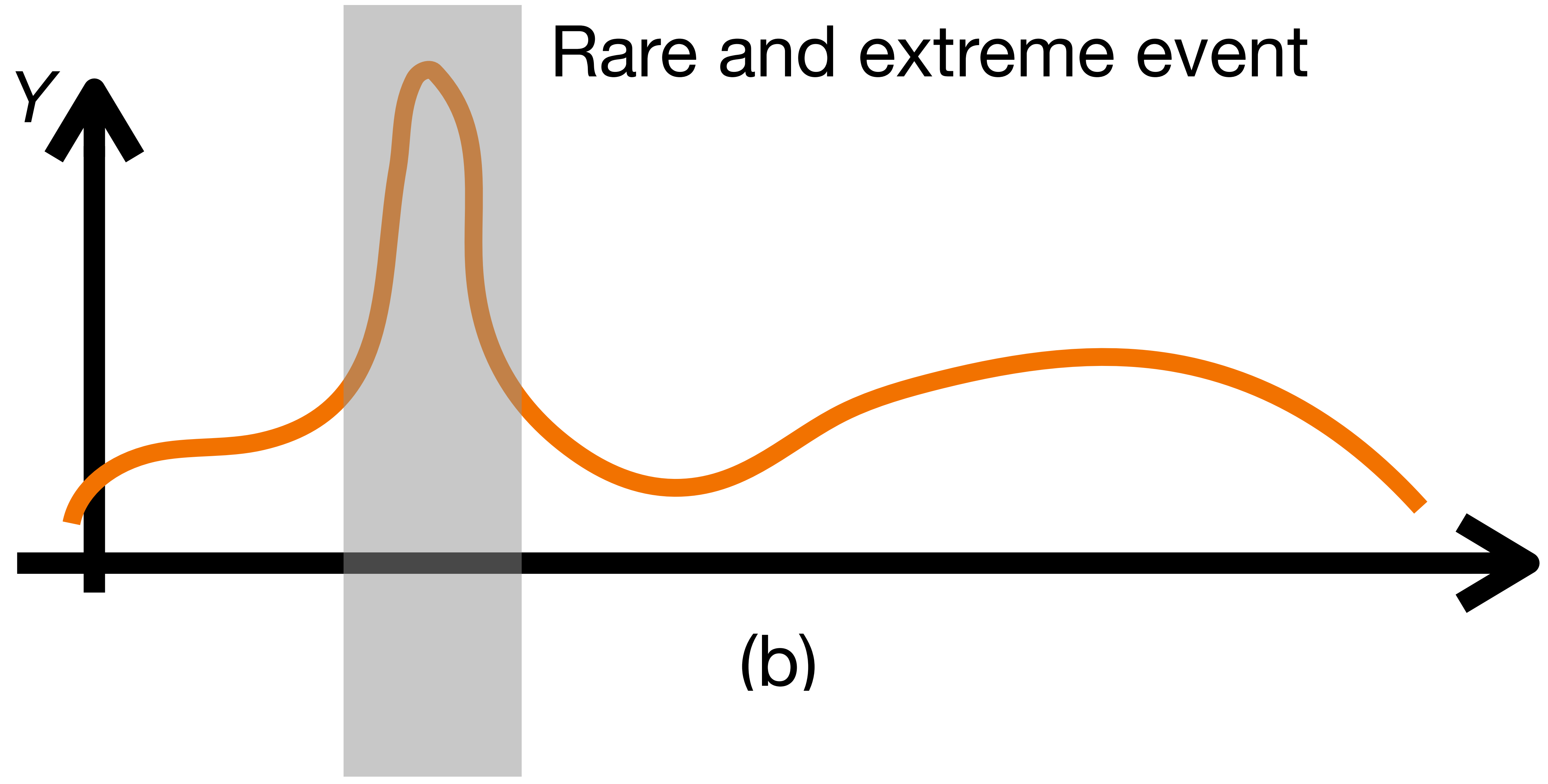


Setting the information theoretical limit

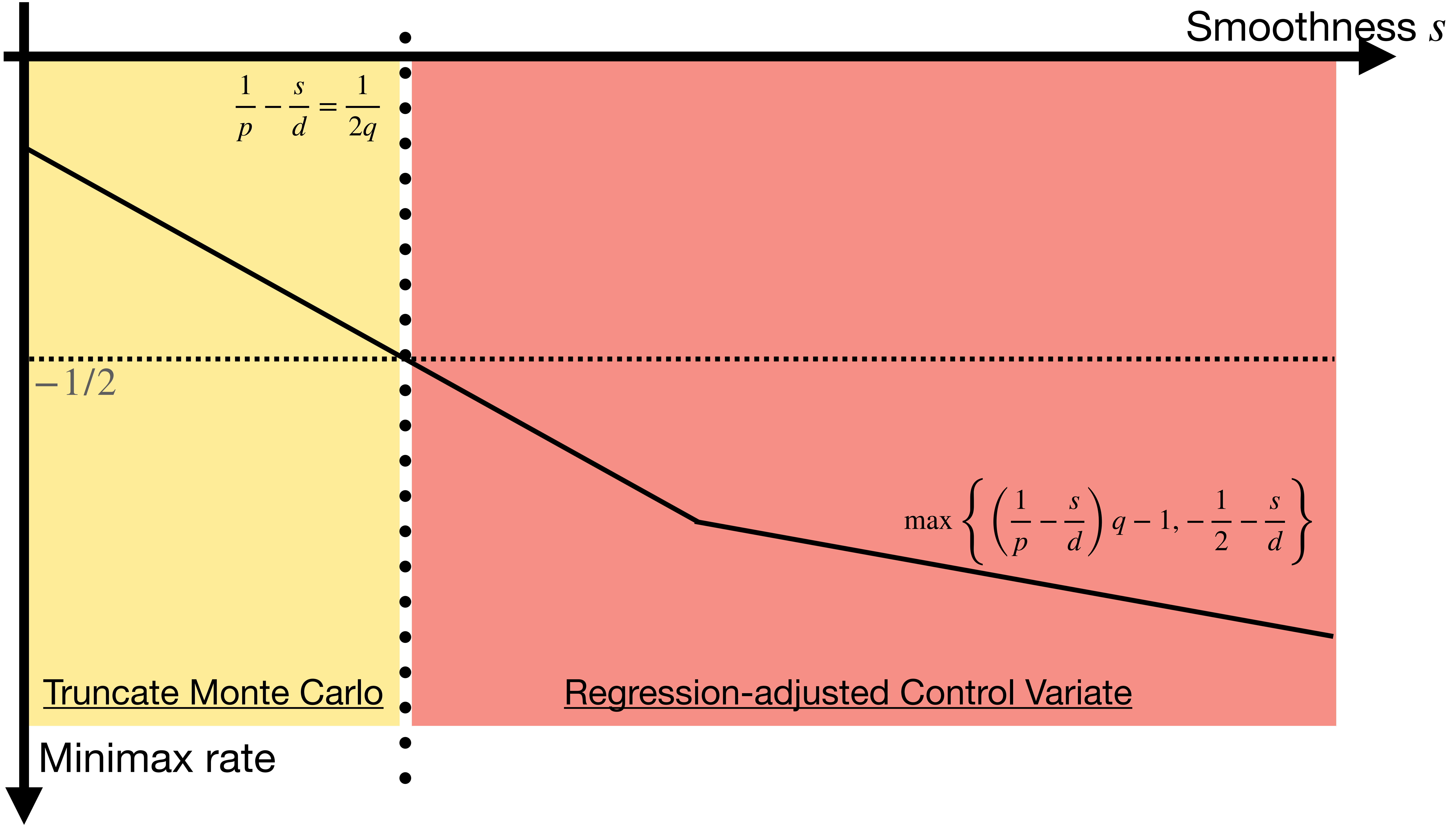


Rare Event and Smoothness…

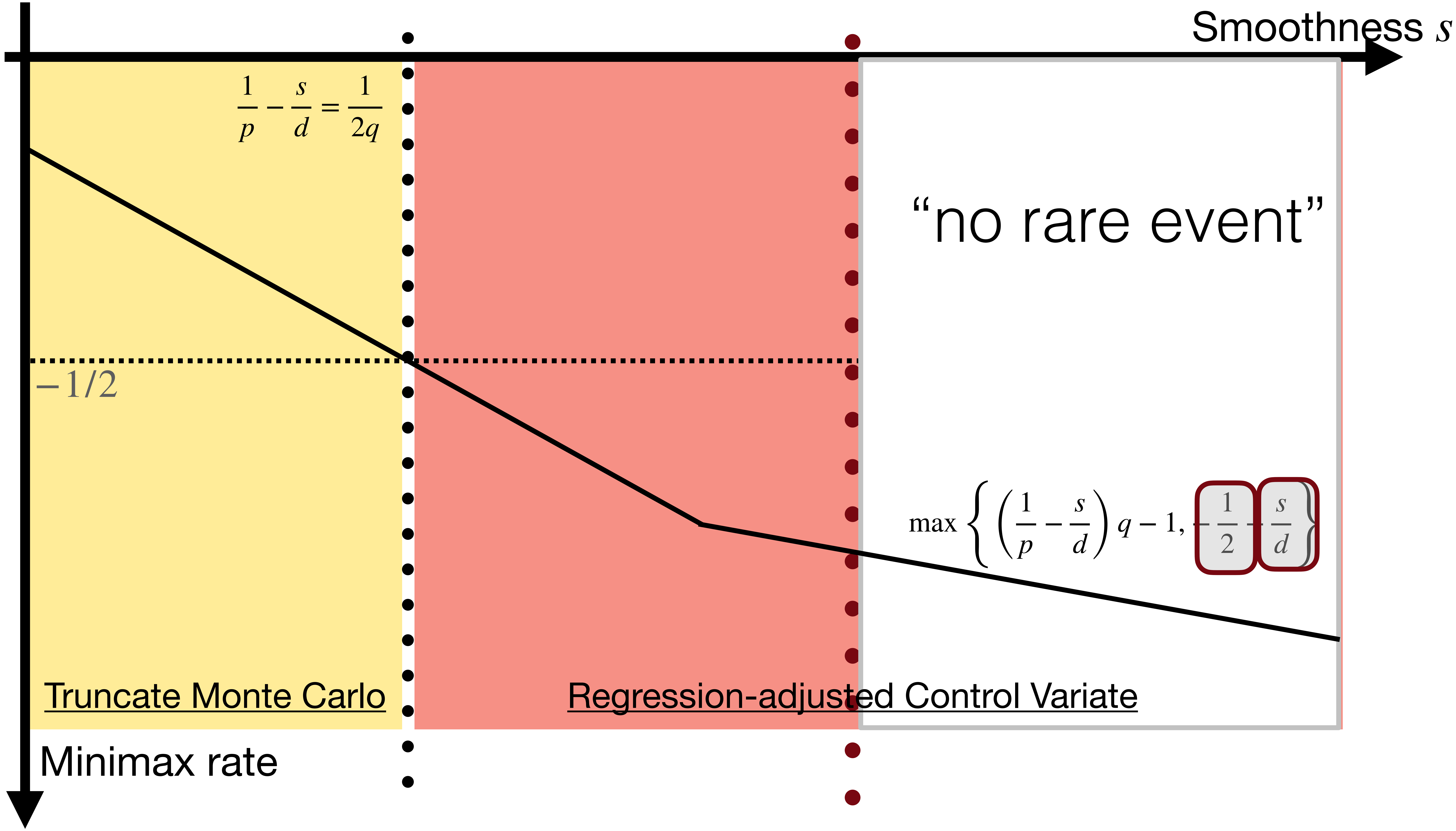




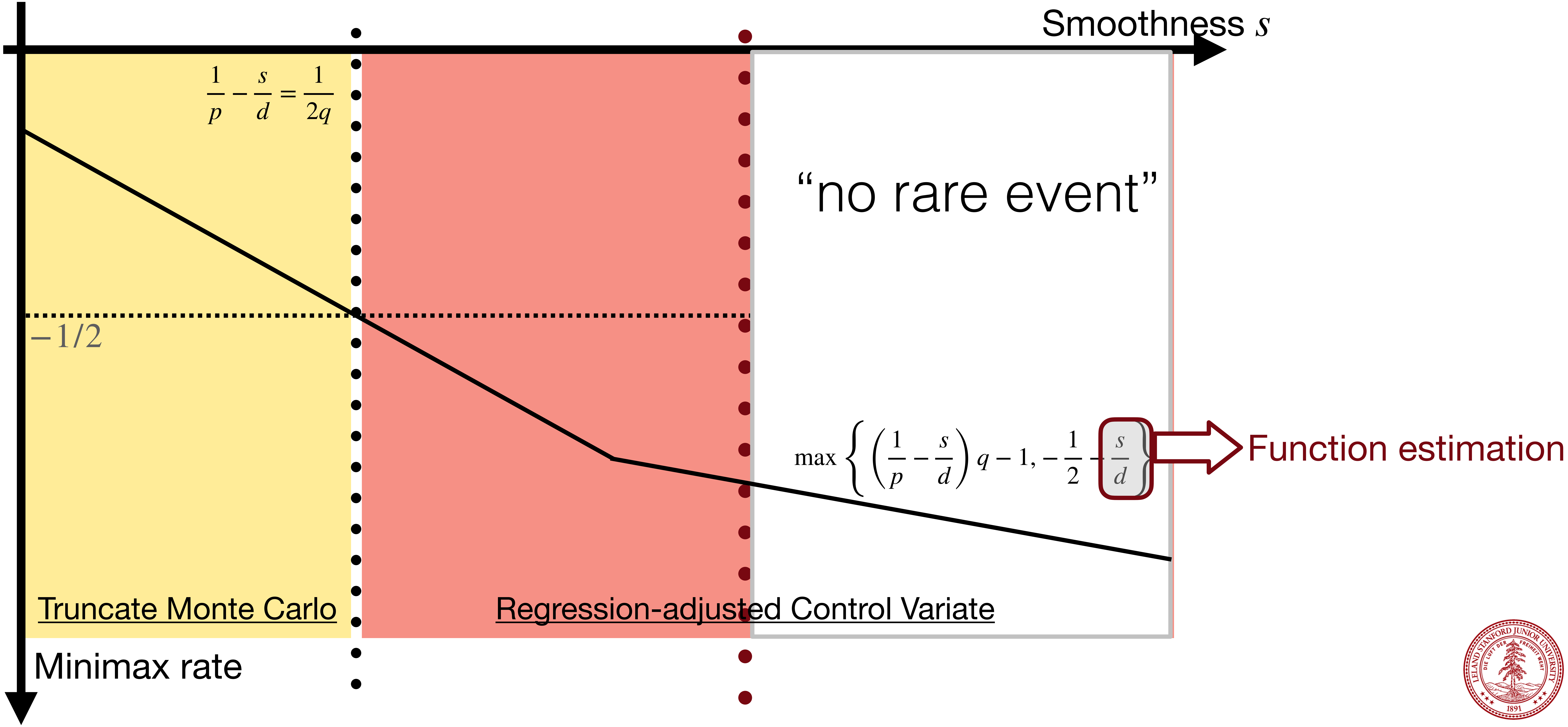
When the control variate helps



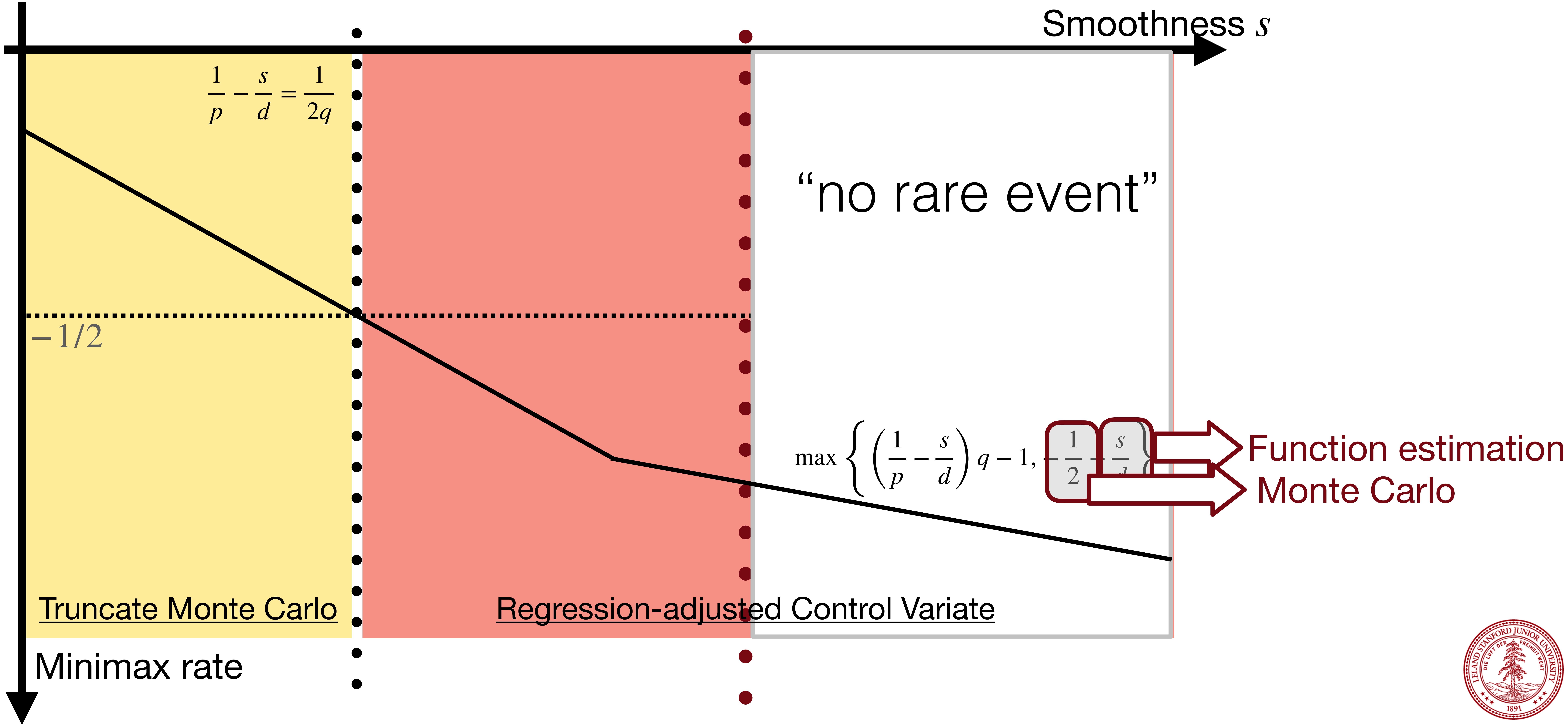
When the control variate helps



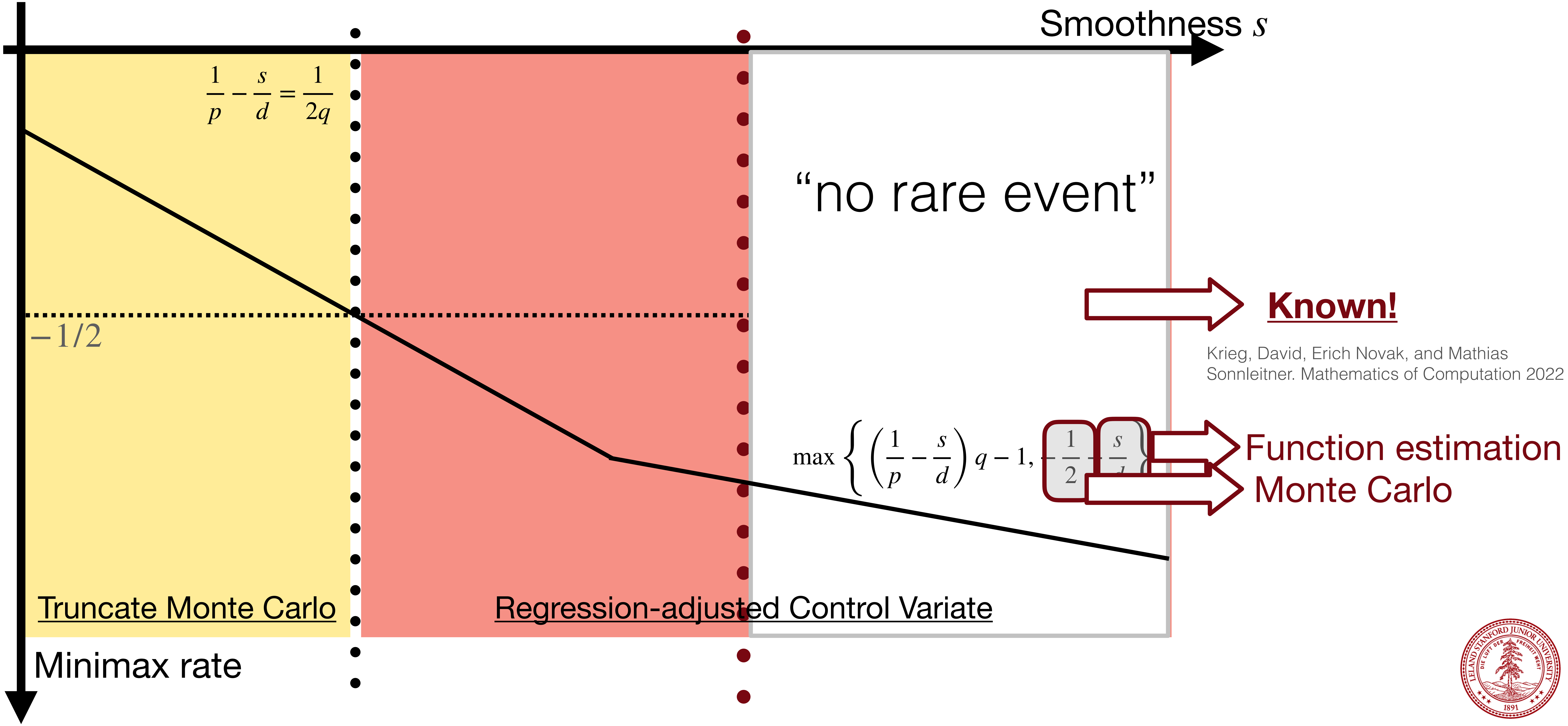
When the control variate helps



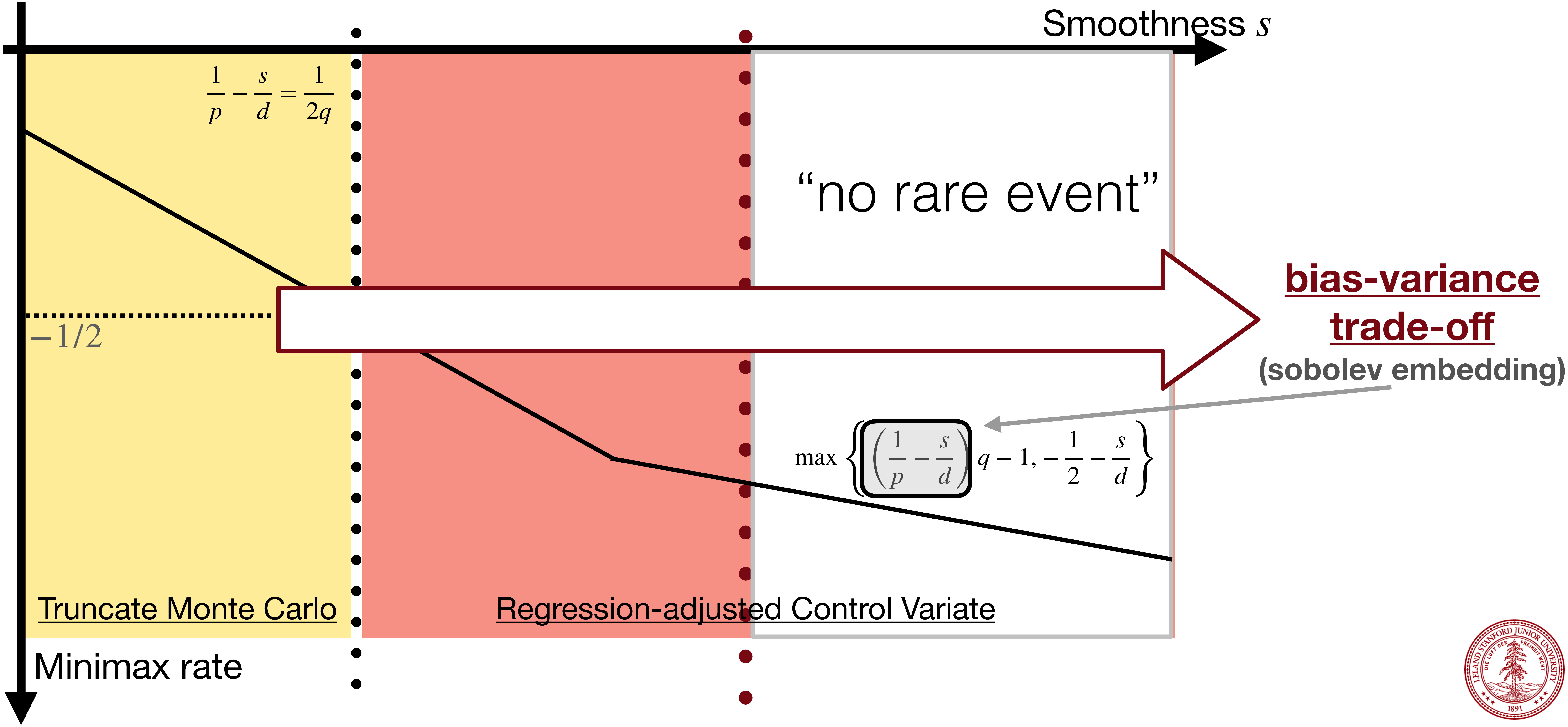
When the control variate helps



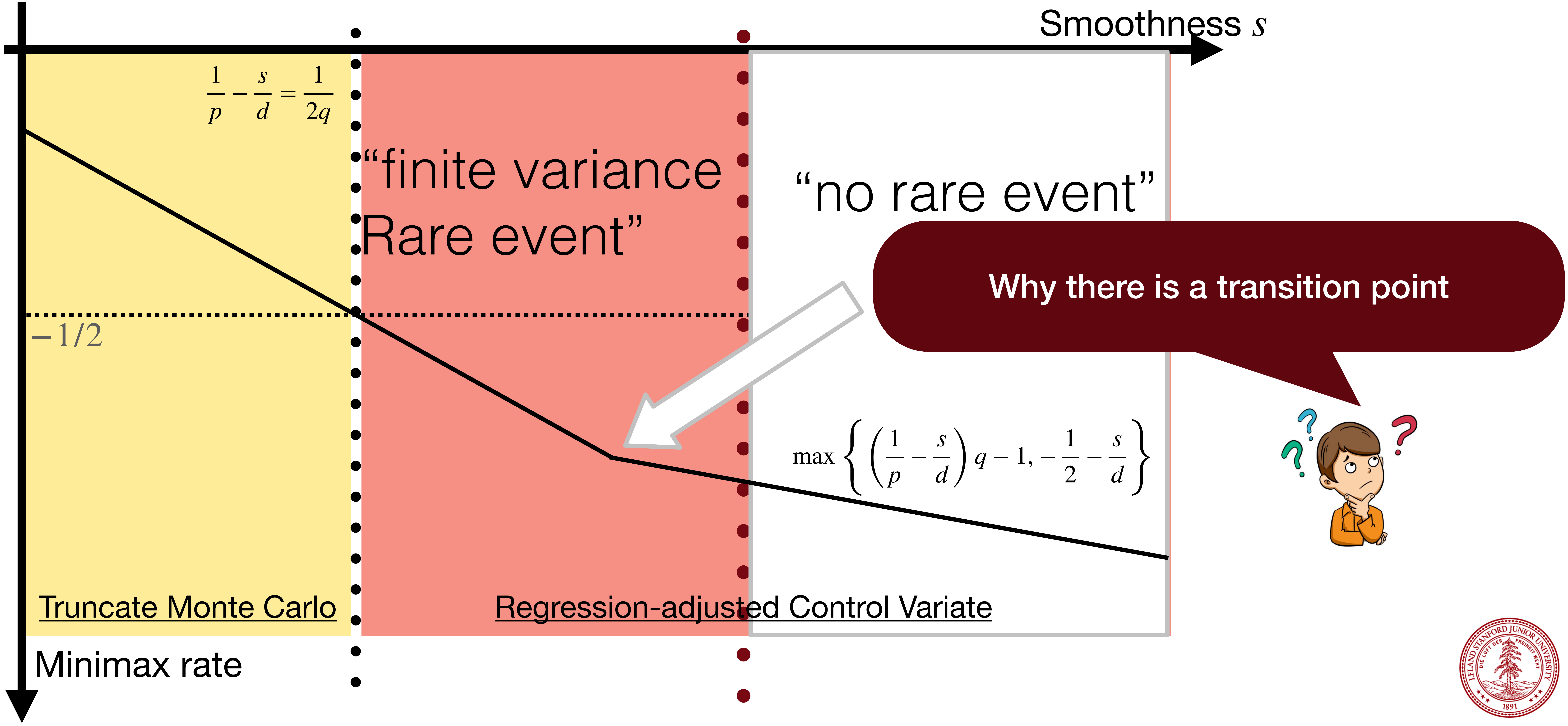
When the control variate helps



When the control variate helps



When the control variate helps



Semi-parametric efficiency…

Example Monte Carlo Estimate $\cancel{\mathbb{E}_P f}$ $\mathbb{E}_P f^q, f \in W^{s,p}$

Step 1 Using half of the data to estimate $\hat{f}$

Step 2 $\mathbb{E}_P f^q = \mathbb{E}_P(\hat{f})^q + \mathbb{E}_P \boxed{(f^q - \hat{f})^q}$

Low order term

$$\boxed{f^{q-1}(f - \hat{f})} + \boxed{(f - \hat{f})^q}$$

"influnce function" (gradient) **Error propagation**

Semi-parametric efficiency…

Example Monte Carlo Estimate $\cancel{\mathbb{E}_P f}$ $\quad \mathbb{E}_P f^q, f \in W^{s,p}$

Step 1 Using half of the data to estimate $\hat{f}$

Step 2 $\mathbb{E}_P f^q = \mathbb{E}_P(\hat{f})^q + \mathbb{E}_P \boxed{(f^q - \hat{f}^q)}$

Low order term

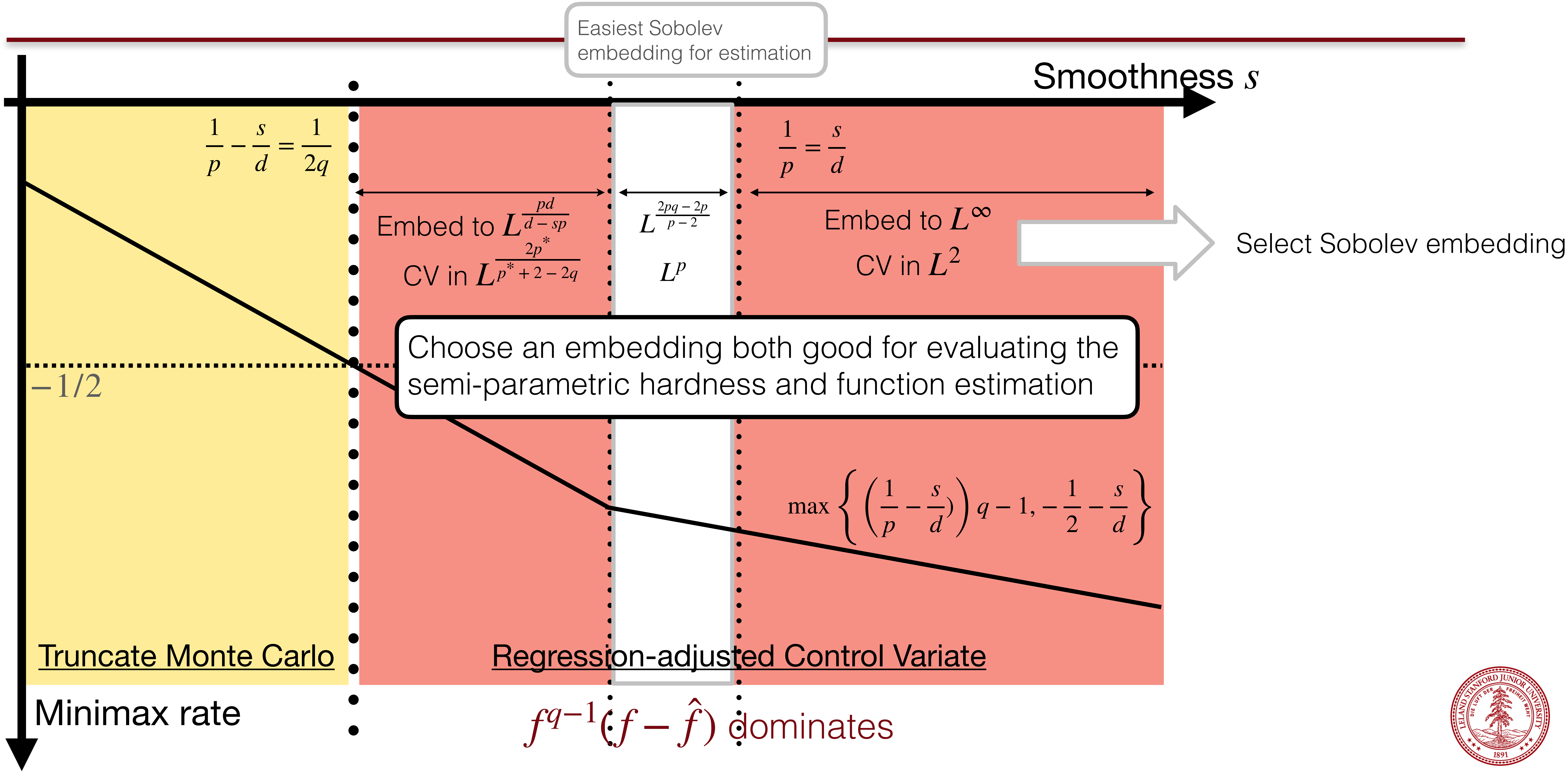
$$\boxed{f^{q-1}}\boxed{(f-\hat{f})} + (f-\hat{f})^q$$

"influnce function" (gradient)

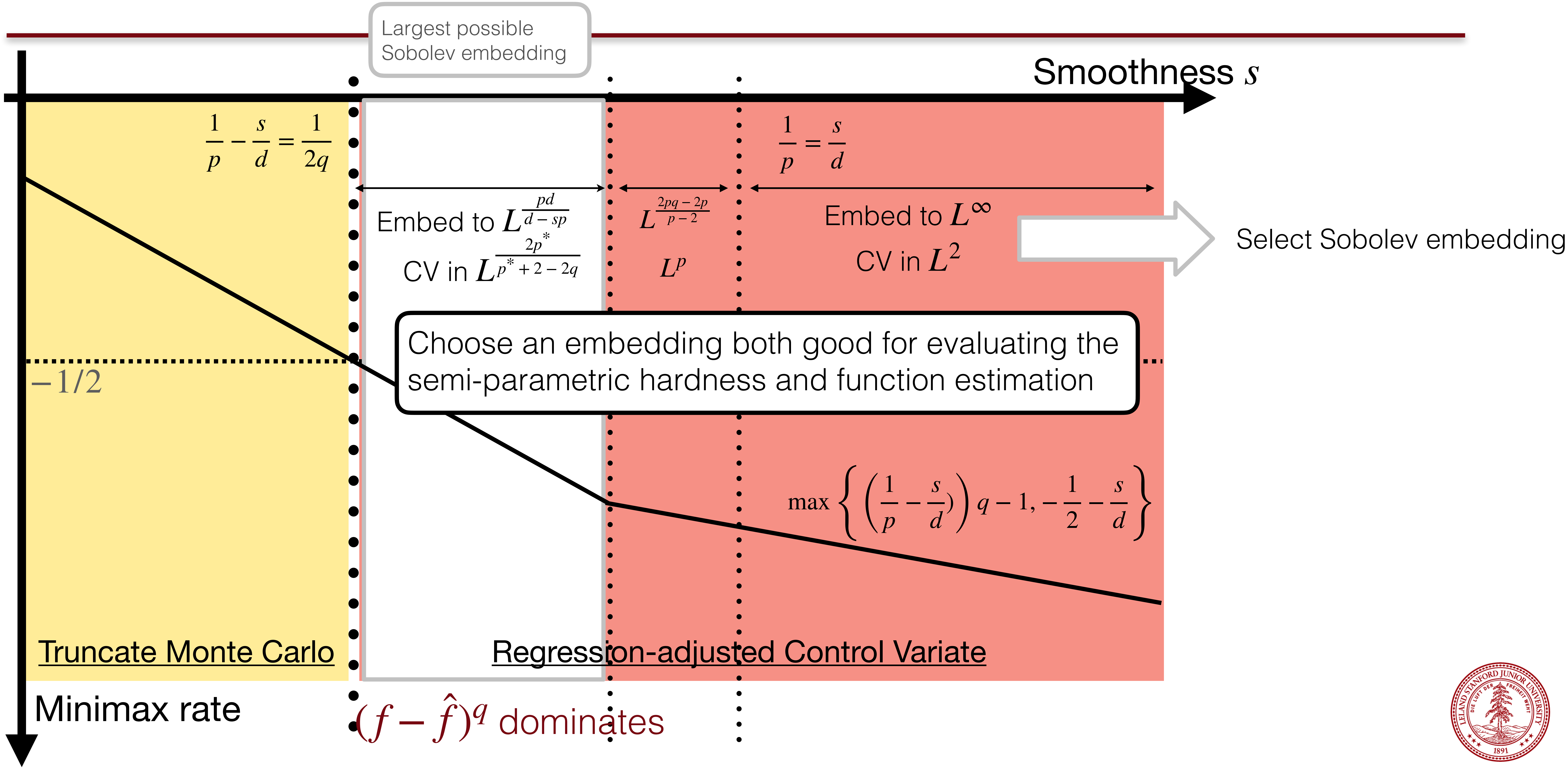
Embed f^{q-1} and $f-\hat{f}$ into "dual" space

How to select the sobolev emebedding

Tricky part of the Proof:select embedding

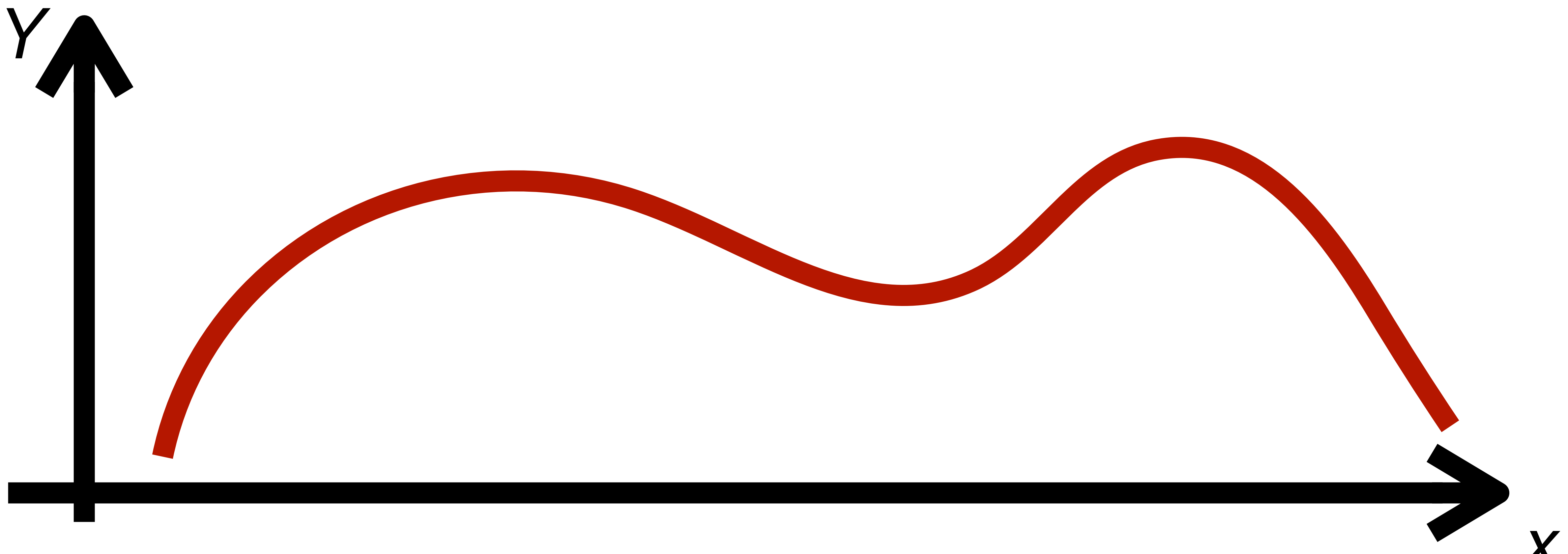


Tricky part of the Proof:select embedding



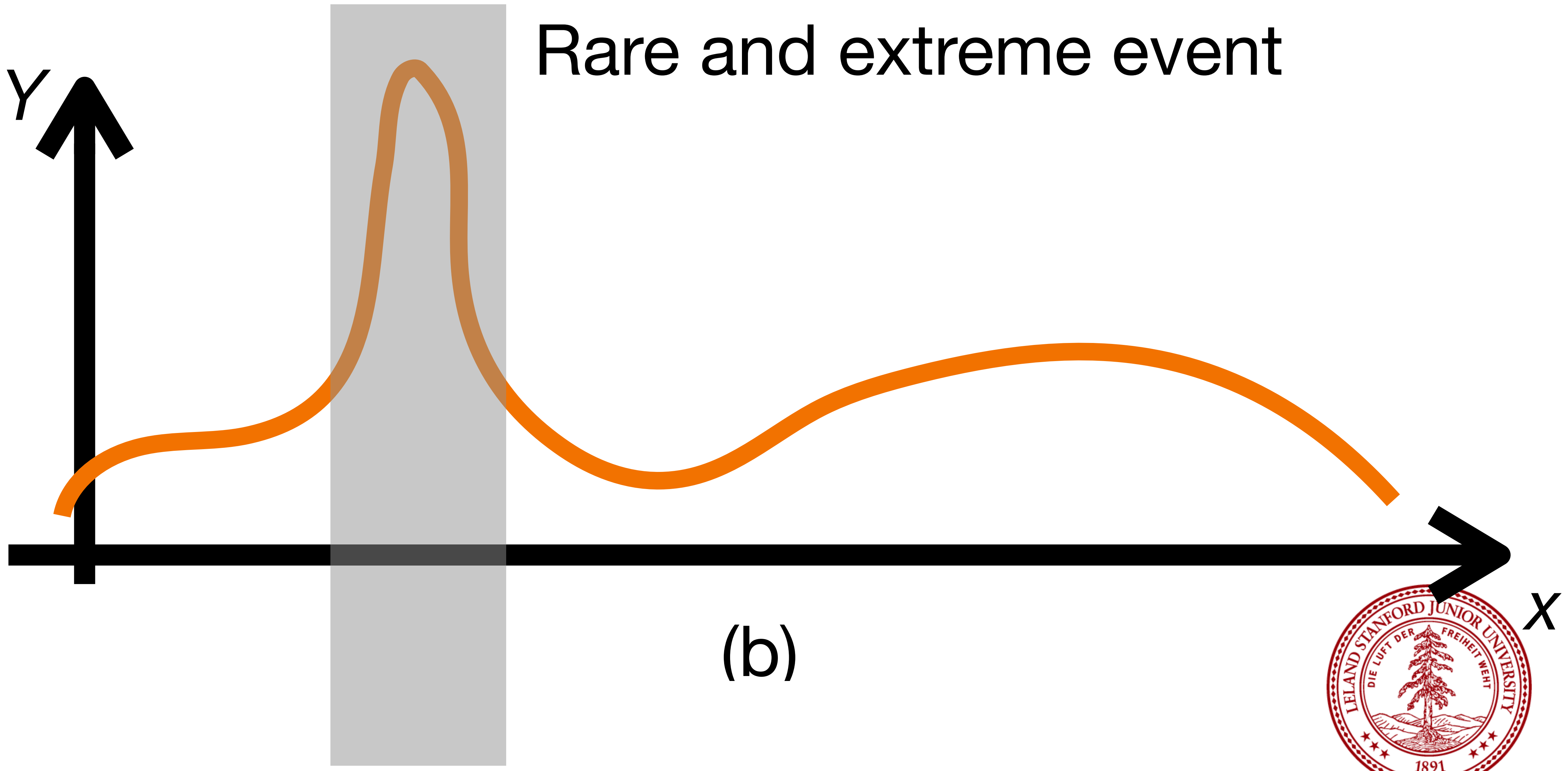
Take home message

a) Statistical optimal regression is the optimal control variate
b) It helps only if there isn't a hard to simulate (infinite variance) Rare and extreme event



(a)

q control the extremeness



(b)

Optimal Statistical PDE Solver

ICLR 2022

$$Au = f$$

Reconstruct u with observation of f: $\{x_i, f(x_i)\}$

Recover parameter θ in Model A_θ

Learn the model A from data pair $\{u_i, f_i\}$

Current Research

$Au = f$

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19]

Recover parameter θ in model A_θ
E.g. Drift, Diffusion Strength

om data pair $\{u_i, f_i\}$
or Learning/Functional data analysis"

ology

[B…unton-Proctor-Kutz 16][Khoo-Lu-Ying 18]

Is direct (plug-in) estimator optimal?

Theory

[Lanthaler-Mishra-Karniadakis 22] [Talwai-Shameli-Simchi-Levi 21][de Hoop-Kovachki-Nelsen-Stuart 21][Li-Meunier-Mollenhauer-Gretton 22] [Liu-Yang-Chen-Zhao-Liao 22]….

[Jin-Lu-Blanchet-Ying 23]

ong 20] [Liang-Yang 22]…

[Nickl-Ray 20] [Nickl 20] [Baek-Farias-Georgescu-Li-Peng-Sinha-Wilde-Zheng 20] [Agrawl-Yin-Zeevi 21]…

Current Research

$Au = f$

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19] [Rahme-Jelassi-Matt Weinberg 21]

Main Idea

Change solving the model to solving a minimization problem

Example: $\Delta u = f$

Current Research

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19] [Rahme-Jelassi-Matt Weinberg 21]

Main Idea

Change solving the model to solving a minimization problem

Example: $\Delta u = f$

1 Design a criteria of whether the model have been solved

$$\int |\nabla u(x)|^2 - 2u(x)f(x)dx$$

[DRM]

$$\int (\Delta u - f)^2 dx$$

[DGM, PINN, …]

2 Sample Average Approximation+ML

Current Research

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19] [Rahme-Jelassi-Matt Weinberg 21]

Main Idea

Change solving the model to solving a minimization problem

Example: $\Delta u = f$

1 Design a criteria of whether the model have been solved

$$\int |\nabla u(x)|^2 - 2u(x)f(x)dx$$

[DRM]

$$\int (\Delta u - f)^2 dx$$

[DGM, PINN, …]

2 Sample Average Approximation+ML

Is this process optimal for all criteria?

A Non-Parametric Statistical Framework

$\Delta u + u = f$

Output An estimation of u

"Learning with gradient information"

Input Random samples $\{(x_i, f(x_i) + \text{noise})\}_{i=1}^n$

i.i.d samples

Aim The best estimator

$$\inf_{H} \max_{f \in H^\alpha} \mathbb{E}_{\{(x_i, f(x_i)+\text{noise})\}_{i=1}^n} \| H(\{(x_i, f(x_i)+\text{noise})\}_{i=1}^n) - u \|_{H^\beta}$$

Evaluation in Sobolev norm

Uniformly good on all Sobolev functions

Estimator

A Non-Parametric Statistical Framework

Theorem (informal)

Minimax lower bound for t-order linear elliptic PDE:

Evaluation in Sobolev norm

$$\inf_{H} \max_{f\in H^\alpha} \mathbb{E}_{\{(x_i,f(x_i)+\text{noise})\}_{i=1}^n} \|H(\{(x_i,f(x_i)+\text{noise})\}_{i=1}^n) - u\|_{H^\beta} \gtrsim n^{-\frac{(\alpha-\beta)}{d+2\alpha-2t}}$$

Order of the PDE

Very similar to nonparametric rate $n^{-\frac{\alpha}{d+2\alpha}}$

A Non-Parametric Statistical Framework

Theorem (informal)

Minimax lower bound for t-order linear elliptic PDE:

Evaluation in Sobolev norm

$$\inf_{H} \max_{f\in H^\alpha} \mathbb{E}_{\{(x_i,f(x_i)+\text{noise})\}_{i=1}^n} \|H(\{(x_i,f(x_i))+\text{noise})\}_{i=1}^n) - u\|_{H^\beta} \gtrsim n^{-\frac{(\alpha-\beta)}{d+2\alpha-2t}}$$

Order of the PDE

Empirical process/fast rate generalization bound

Is PINN and DRM statistical optimal?

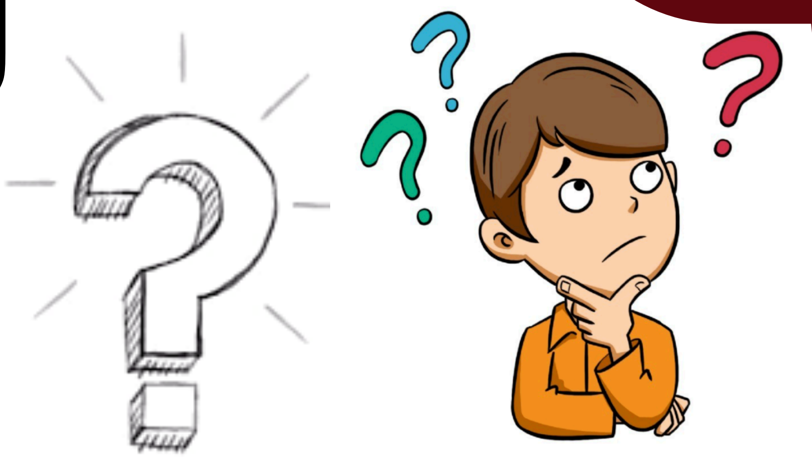
Artifact of analysis? NN ansatz? Objective?

For $\beta = 2$

PINN

For $\beta = 1$

DRM



Is Deep Ritz Optimal? A Fourier View

$\boxed{Au = f}$

Solving $\Delta u + u = f$ from random samples $\{(x_i, f(x_i) + \text{noise})\}_{i=1}^{n}$

Why not first learn f then learn u

Naive way to do this?

Naive Estimator $\hat{f} = \sum_{|z|<S} \hat{f}_z^F \phi_z$ where $\hat{f}_z^F = \sum f(x_i)\phi_z(x_i)$

Then $u = A^{-1}f = \sum_{|z|<S} \frac{1}{|z|^2 + 1} \hat{f}_z^F \phi_z$ Fourier Basis

Naive Estimator is Optimal with proper selection of S

Is Deep Ritz Optimal? A Fourier View

$\boxed{Au = f}$

Solving $\Delta u + u = f$ from random samples $\{(x_i, f(x_i) + \text{noise})\}_{i=1}^n$

Why not first learn f then learn u

Naive Estimator $\hat{f} = \sum_{|z|<S} \hat{f}_z^F \phi_z$ where $\hat{f}_z^F = \sum f(x_i)\phi_z(x_i)$

Then $u = A^{-1}f = \sum_{|z|<S} \frac{1}{|z|^2 + 1} \hat{f}_z^F \phi_z$

DRM Estimator $\hat{u} = \sum_{|z|<S} \hat{u}_z^F \phi_z$ and plug in

$$\hat{u}^F = \arg\min_{\hat{u}^F} \int \frac{1}{2} \left| \sum_{|z|<S} \hat{u}_z^F (\nabla \phi_z + \phi_z) \right|^2 - \sum_{|z|<S} \hat{u}_z^F \hat{f}_z^F$$

How is naive estimator different from DRM?

Is Deep Ritz Optimal? A Fourier View

$\boxed{Au = f}$

Solving $\Delta u + u = f$ from random samples $\{(x_i, f(x_i) + \text{noise})\}_{i=1}^n$

Why not first learn f then learn u

Naive Estimator $\hat{f} = \sum_{|z|<S} \hat{f}_z^F \phi_z$ where $\hat{f}_z^F = \sum f(x_i)\phi_z(x_i)$

Then $u = A^{-1}f = \sum_{|z|<S} \frac{1}{|z|^2+1} \hat{f}_z^F \phi_z$

DRM Estimator $\hat{u} = \sum_{|z|<S} \hat{u}_z^F \phi_z$ and plug in

$$\hat{u}^F = \arg\min_{\hat{u}^F} \int \frac{1}{2} \left| \sum_{|z|<S} \hat{u}_z^F (\nabla \phi_z + \phi_z) \right|^2 - \sum_{|z|<S} \hat{u}_z^F \hat{f}_z^F$$

Naive

$$\boxed{\hat{u}_z^F = \frac{\hat{f}_z^F}{|z|^2+1}}$$

DRM

$$\boxed{\hat{u}_z^F = (\hat{A})^{-1} \hat{f}_z^F}$$

$$\hat{A} = \left(\sum_i \nabla \phi_j(x_i) \nabla \phi_k(x_i) \right)_{j,k} + \left(\sum_i \phi_j(x_i)\phi_k(x_i) \right)_{j,k}$$

Introduce further variance

Is Deep Ritz Optimal? A Fourier View

$\boxed{Au = f}$

Solving $\Delta u + u = f$ from random samples $\{(x_i, f(x_i) + \text{noise})\}_{i=1}^n$

Why not first learn f then learn u

Naive Estimator $\hat{f} = \sum_{|z|<S} \hat{f}_z^F \phi_z$ where $\hat{f}_z^F = \sum f(x_i)\phi_z(x_i)$

Then $u = A^{-1}f = \sum_{|z|<S} \frac{1}{|z|^2+1} \hat{f}_z^F \phi_z$

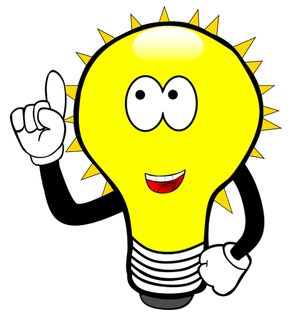
DRM Estimator $\hat{u} = \sum_{|z|<S} \hat{u}_z^F \phi_z$ and plug in

$$\hat{u}^F = \arg\min_{\hat{u}^F} \int \frac{1}{2} \left| \sum_{|z|<S} \hat{u}_z^F (\nabla \phi_z + \phi_z) \right|^2 - \sum_{|z|<S} \hat{u}_z^F \hat{f}_z^F$$

DRM discretized

$\nabla \cdot \nabla$

But not Δ



Integration by parts increase the monte-carlo variance.

Results in One Table...

Boundary condition?

	Upper Bounds		Lower Bound
Objective Function	Neural Network	Fourier Basis	
Deep Ritz	$n^{-\frac{2s-2}{d+2s-2}} \log n$	$n^{-\frac{2s-2}{d+2s-2}}$	$n^{-\frac{2s-2}{d+2s-4}}$
Modified Deep Ritz	$n^{-\frac{2s-2}{d+2s-2}} \log n$	$n^{-\frac{2s-2}{d+2s-4}}$	$n^{-\frac{2s-2}{d+2s-4}}$
PINN	$n^{-\frac{2s-4}{d+2s-4}} \log n$	$n^{-\frac{2s-4}{d+2s-4}}$	$n^{-\frac{2s-4}{d+2s-4}}$

Still open

For $\beta = 2$

PINN ✓

For $\beta = 1$

DRM

	DRM	Modified
Spectral	✗	✓
NN	✗	?

DRM or PINN

Which one optimizes faster?

DRM $\min \int |\nabla u|^2 - 2uf$

PINN $\min ||\Delta u - f||^2$

Pre-ml Experience：
Double the condition number

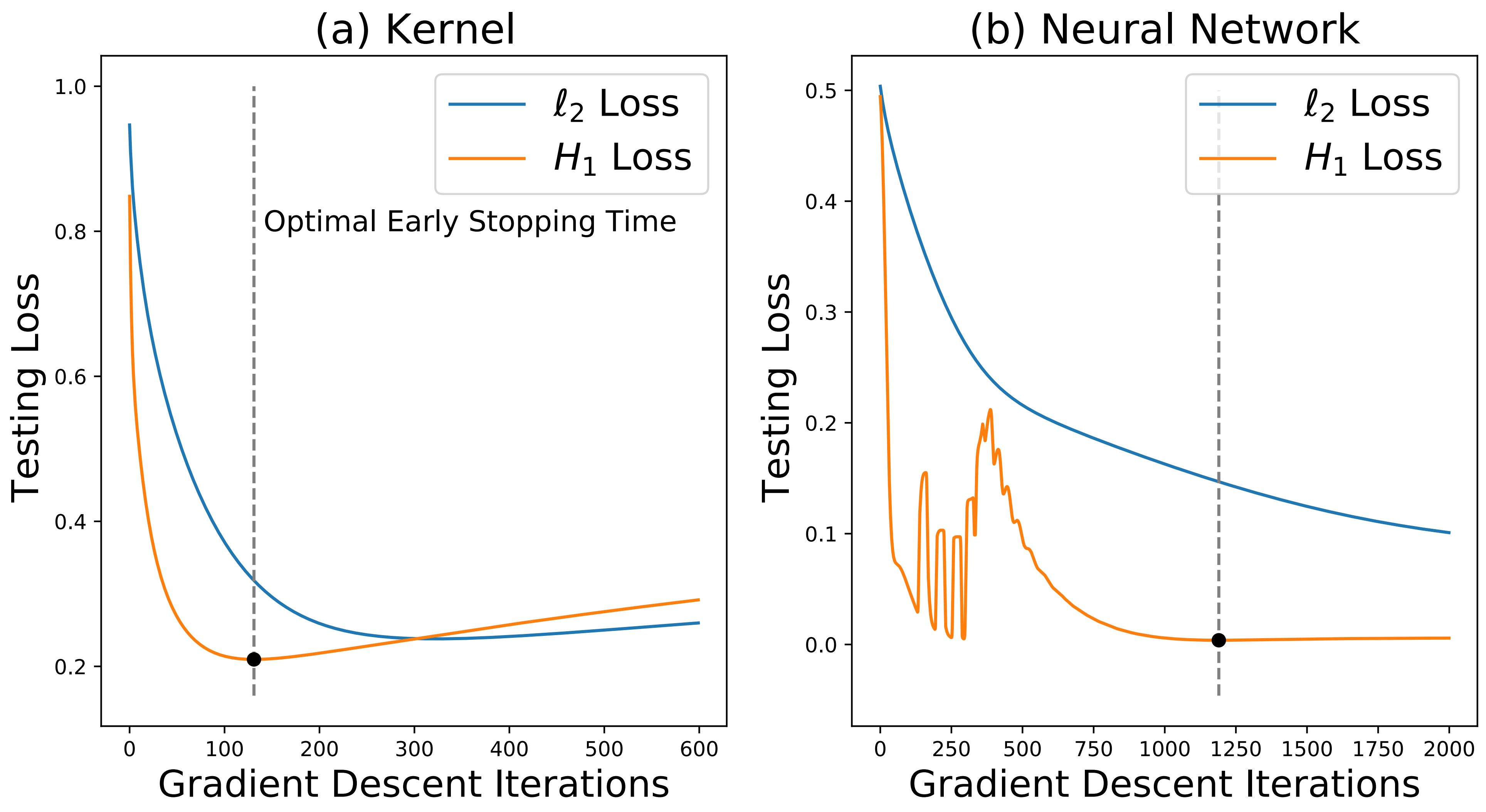
DRM or PINN

Which one optimizes faster?

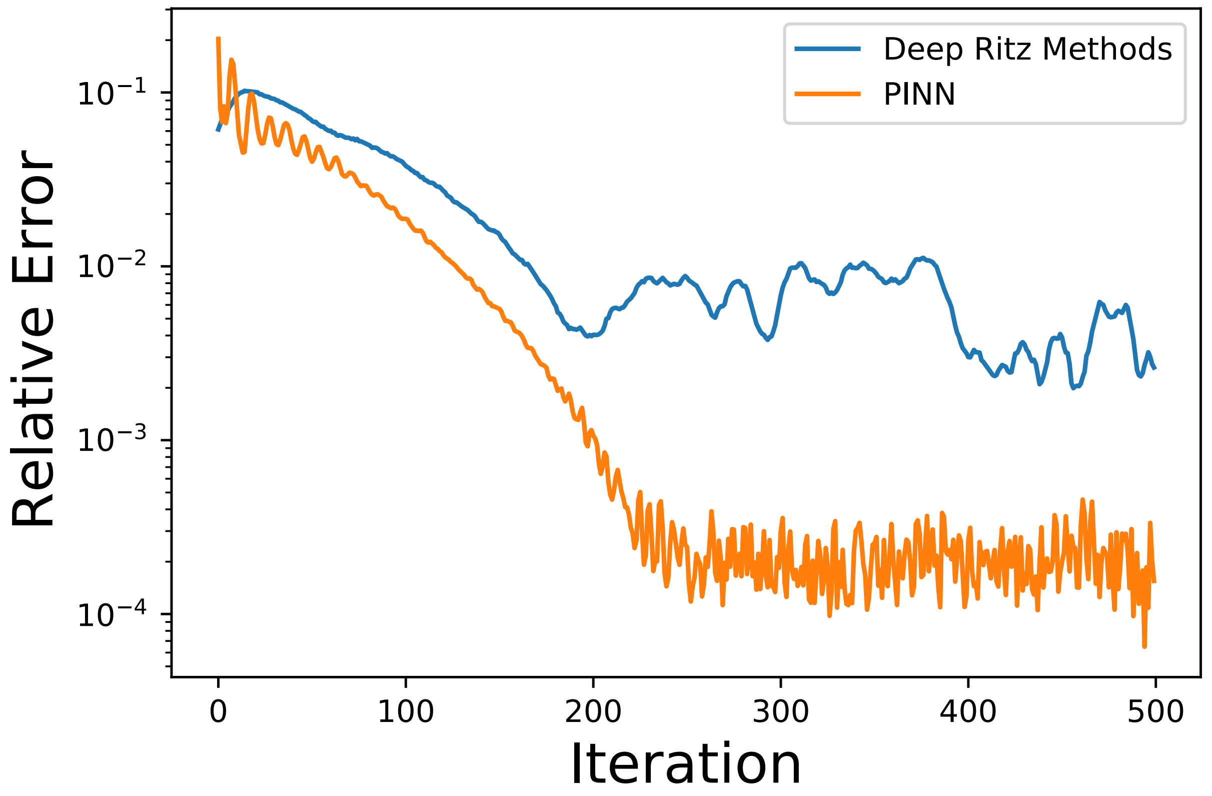
DRM $\min \int |\nabla u|^2 - 2uf$

PINN $\min \|\Delta u - f\|^2$

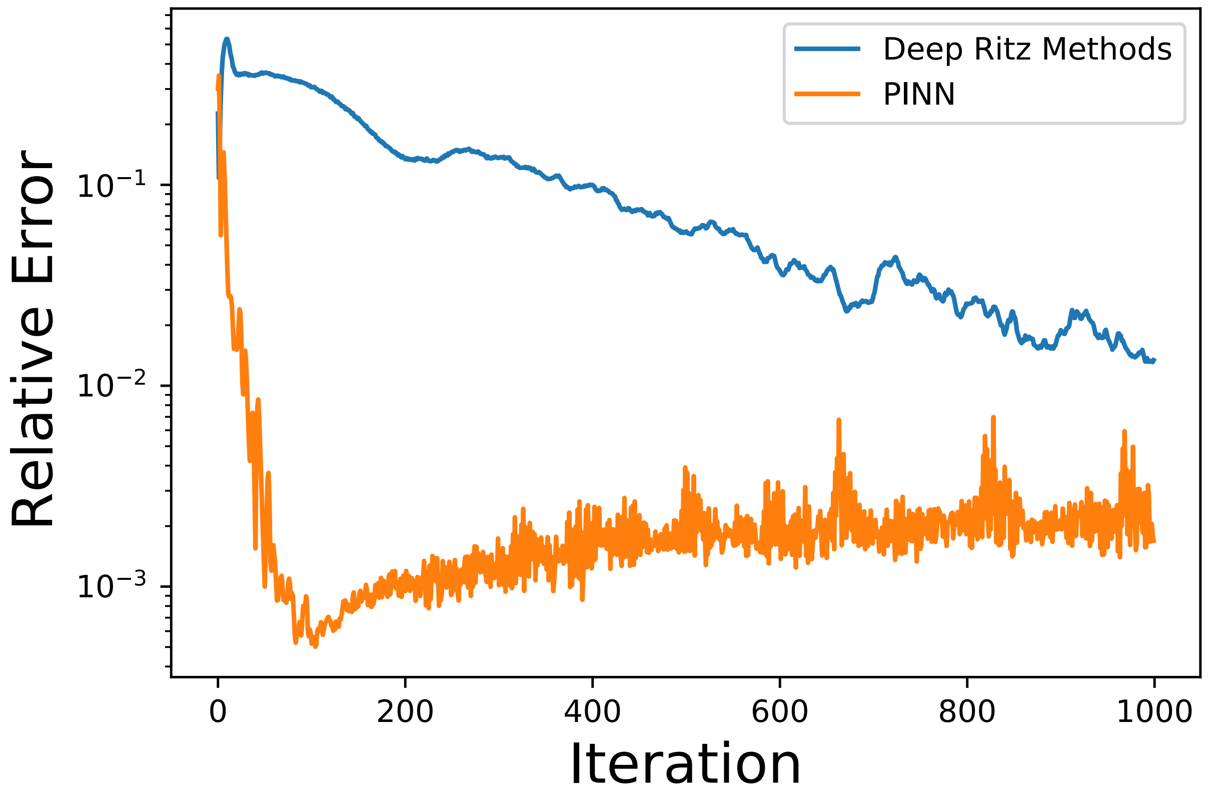
Pre-ml Experience：
Double the condition number



Sobolev Training



$f = \sin(2\pi x)$



$f = \sin(4\pi x)$

Solving $\Delta u = f$

A Kernelized Model

Machine learning is a kernelized dynamic.
Differential Operator can cancel Kernel Integral Op

Let's consider $\Delta u = f$ via minimizing $\frac{1}{2}\langle f, \mathcal{A}_1 f\rangle - \langle u, \mathcal{A}_2 f\rangle$

$$f = \langle \theta, K_x \rangle$$

- **Deep Ritz Methods**. $\mathcal{A}_1 = \Delta, \mathcal{A}_2 = Id$
- **PINN**. $\mathcal{A}_1 = \Delta^2, \mathcal{A}_2 = \Delta$

Gradient Descent

$$d\theta_t = \sum_i \langle \theta, \underbrace{\mathcal{A}_1}_{\text{Differential operator}} \underbrace{K_{x_i i} \rangle K_{x_i}}_{\text{Kernel integral operator}} - f_i \mathcal{A}_2 K_{x_i}$$

Our Result

I understand your idea,
but what's your thm?

Theorem (Informal)

1. The information theoretical lower bound in the kernel space matches the lower bound for learning PDE.

2. Gradient Descent with proper early stopping time selection can achieve optimal statistical rate

3. The proper early stopping time is smaller for PINN than DRM

Optimal (Linear) Operator Learning

ICLR 2023 (spotlight)

$$Au = f$$

Reconstruct u with observation of f: $\{x_i, f(x_i)\}$

Recover parameter θ in Model A_θ

Learn the model A from data pair $\{u_i, f_i\}$

(Linear) Operator Learning

Can we learn the mapping from infinite dimensional space to infinite dimensional space?

Functional data analysis!

Data are function pairs $\{u_i, f_i\}_{i=1}^{n}$

Aim Learn a mapping from function space to function space

u_i f_i

Let's first understanding the linear case!

Linear Operator itself is important still…

Learn $p(Y|X)$ via learning the linear operator

$$\boxed{p_{\text{in}}(x) \to p_{\text{out}}(y) := \int p(y|x)p_{\text{in}}(x)dx}$$

Distribution is ***infinite dimensional***

Distribution of *x* → Distribution of *y*

Linear operator

Linear Operator itself is important still...

Learn $p(Y|X)$ via learning the linear operator

$$\boxed{p_{\text{in}}(x) \to p_{\text{out}}(y) := \int p(y|x)p_{\text{in}}(x)dx}$$

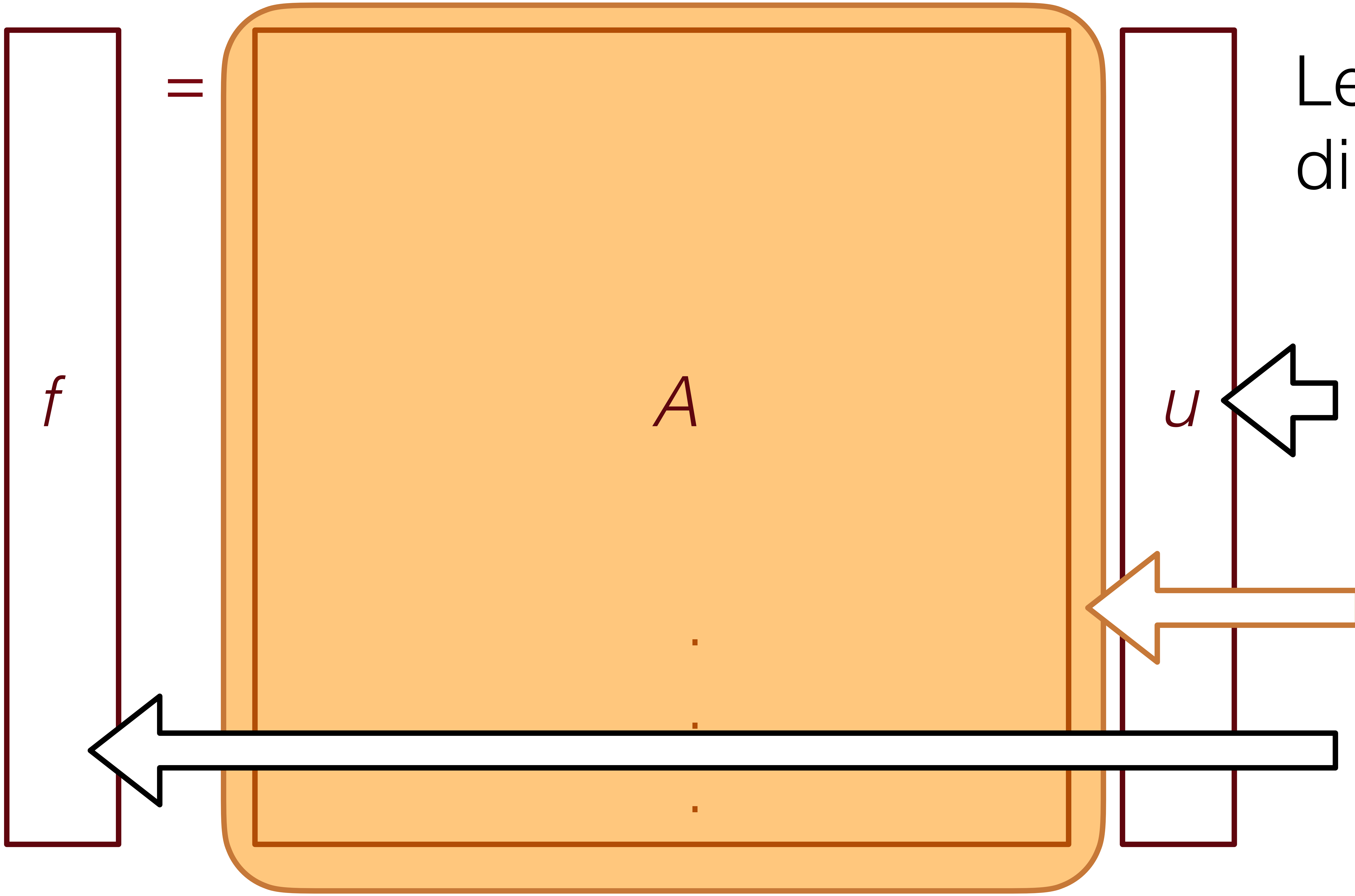
*Distribution is **infinite dimensional***

Instrumental variable regression
[Singh-Chernozhukov-Newey 2022]

Time series modeling
[Kostic-Novelli-Maurere-Ciliberto-Rosasco-Pontil 2022]

Generator/Koopman Operator/CME

Linear Operator Learning



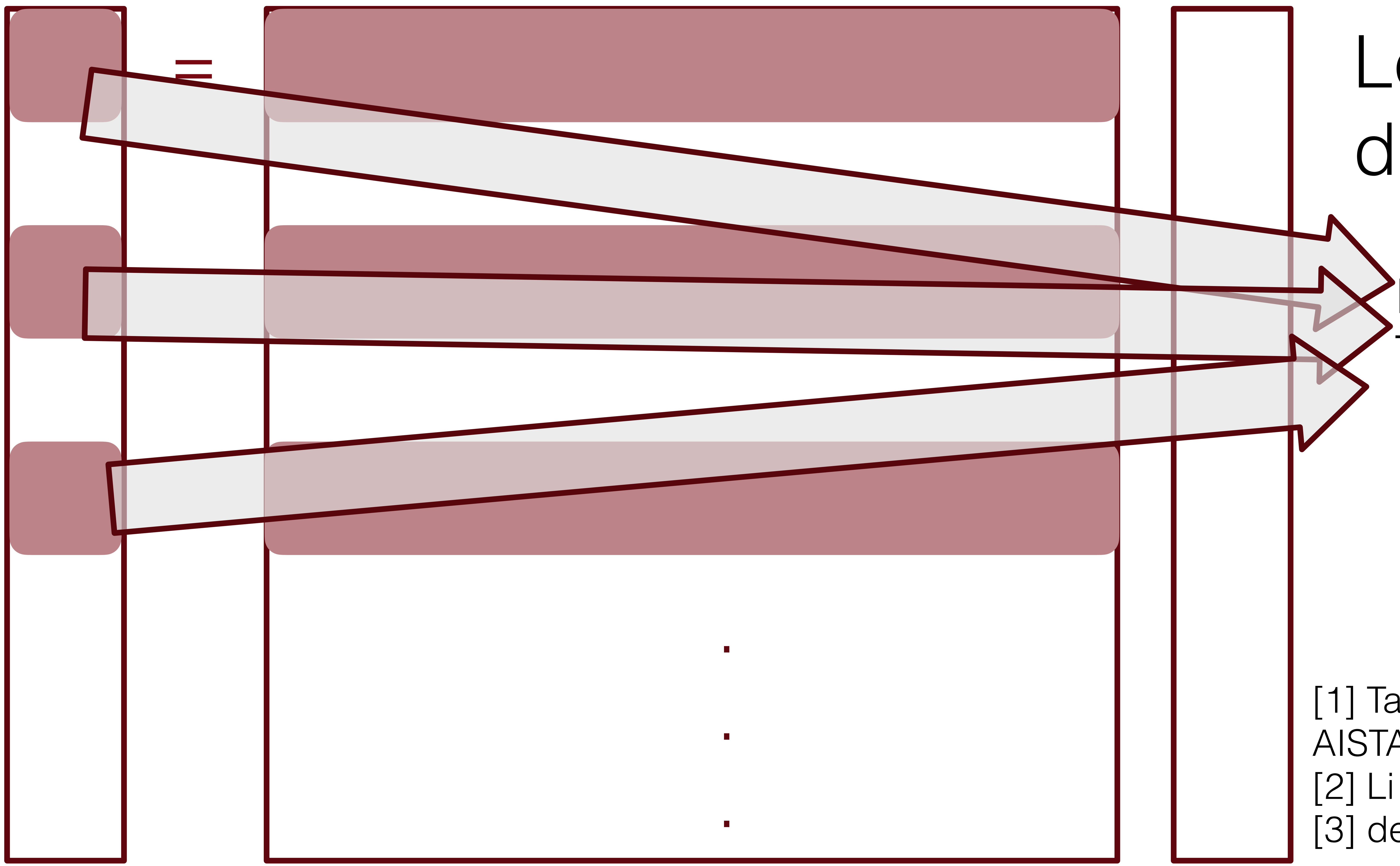
Learning "infinite-dimension" matrix

Randomly sampled

Aim to learn

Data collected

Why infinite dimensional operator is hard

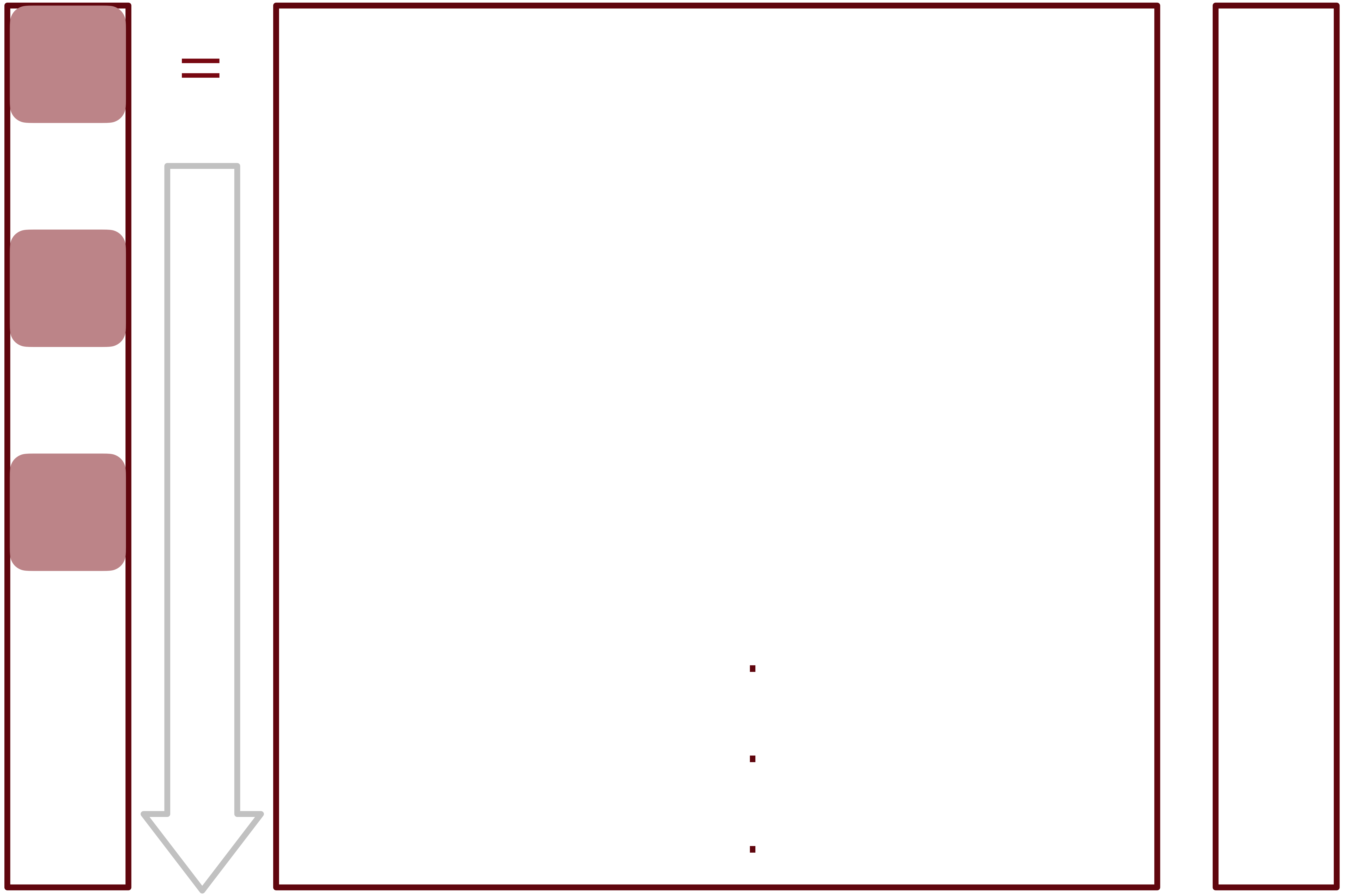


Learning “infinite-dimension” matrix

If every row have O(1) variance,
The total variance is ∞

[1] Talwai P, Shameli A, Simchi-Levi D. AISTAT 2022
[2] Li Z, Meunier D, A Gretton. Neurips 2022
[3] de Hoop M V, et al. arXiv:2108.12515

Why infinite dimensional operator is hard



Fast Variance Decay

Learning "infinite-dimension" matrix

Previous Work:
Assume Fast Eigen Decay to ensure finite variance.

[1] Talwai P, Shameli A, Simchi-Levi D. AISTAT 2022
[2] Li Z, Meunier D, A Gretton. Neurips 2022
[3] de Hoop M V, et al. arXiv:2108.12515

Why infinite dimensional operator is hard

=

Learning "infinite-

trix

Decay

ance.

Will removing the fast variance decay assumption leads to some thing different?

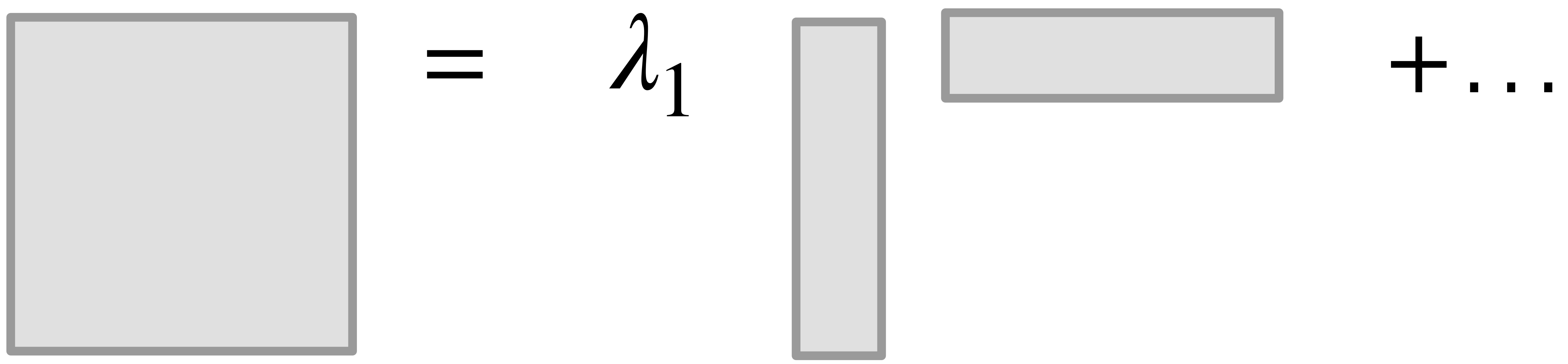
.

.

[1] Talwai P, Shameli A, Simchi-Levi D. AISTAT 2022
[2] Li Z, Meunier D, A Gretton. Neurips 2022
[3] de Hoop M V, et al. arXiv:2108.12515

Spaces we are interested

Hilbert space have finite variance as finite dimensional space



Eigen decomposition

$$K(x, y) = \sum_{n=1}^{\infty} \lambda_n e_n(u) e_n(v)$$

Eigen decay $\lambda_n \propto n^{-\frac{1}{p}}$

Ensures finite variance

Spaces we are interested

Hilbert space have finite variance as finite dimensional space

$= \lambda_1 \quad + \ldots$

Eigen decomposition

$$K(x, y) = \sum_{n=1}^{\infty} \lambda_n e_n(u) e_n(v)$$

Eigen decay $\lambda_n \propto n^{-\frac{1}{p}}$

"Kernel Sobolev space": larger than RKHS H^{β} ***Fourier expansion***

$$= a_1 \lambda_1^{\beta/2} \; e_1 \; + \; a_2 \lambda_2^{\beta/2} \; e_2 \; + \ldots$$

with $(a_i)_{i=1}^{\infty} \in \ell_2, \beta \in (0,1)$

"slower eigendecay"

Spaces we are interested

Hilbert space have finite variance as finite dimensional space

$= \lambda_1 \qquad + \ldots$

Eigen decomposition

$$K(x,y) = \sum_{n=1}^{\infty} \lambda_n e_n(u) e_n(v)$$

Eigen decay $\lambda_n \propto n^{-\frac{1}{p}}$

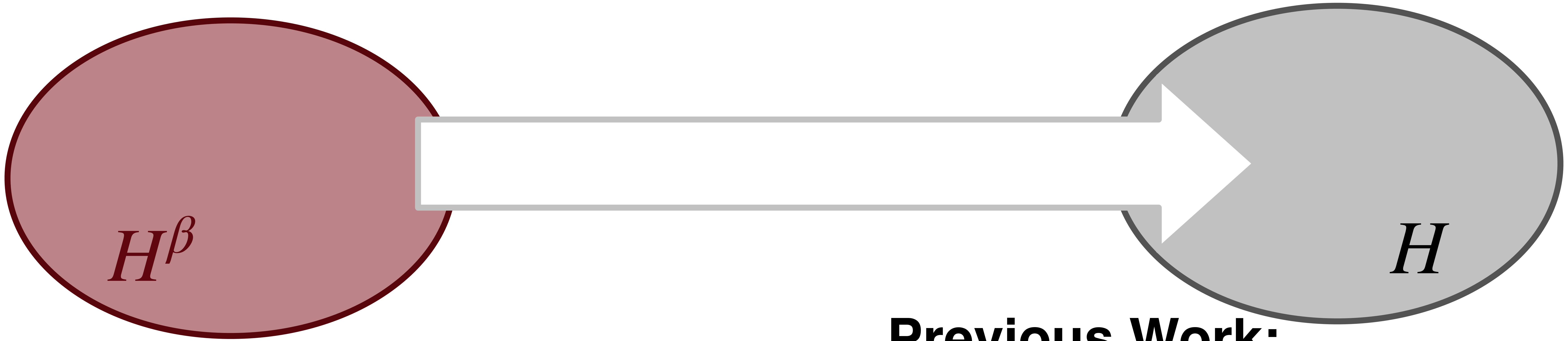
"Kernel Sobolev space": larger than RKHS H^{β} ***Fourier expansion***

$$= a_1 \lambda_1^{\beta/2} \; e_1 \; + \; a_2 \lambda_2^{\beta/2} \; e_2 \; + \ldots$$

with $(a_i)_{i=1}^{\infty} \in \ell_2, \beta \in (0,1)$

l2 $\beta = 0$ ⟷ *RKHS* $\beta = 1$

Problem Formulation



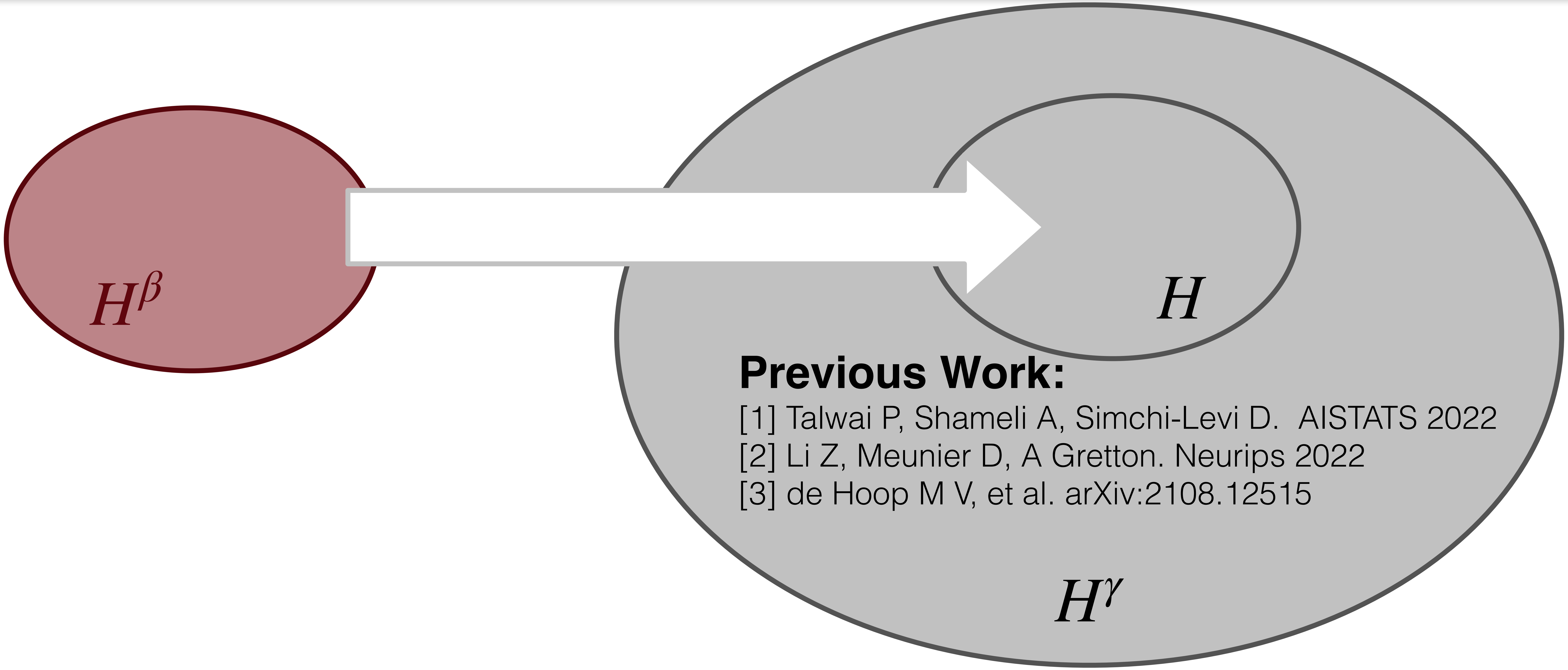
H^β is a larger space

Previous Work:

[1] Talwai P, Shameli A, Simchi-Levi D. AISTATS 2022
[2] Li Z, Meunier D, A Gretton. Neurips 2022
[3] de Hoop M V, et al. arXiv:2108.12515

Same technique as $H^\beta \to \mathbb{R}$ for ridge regression

Problem Formulation



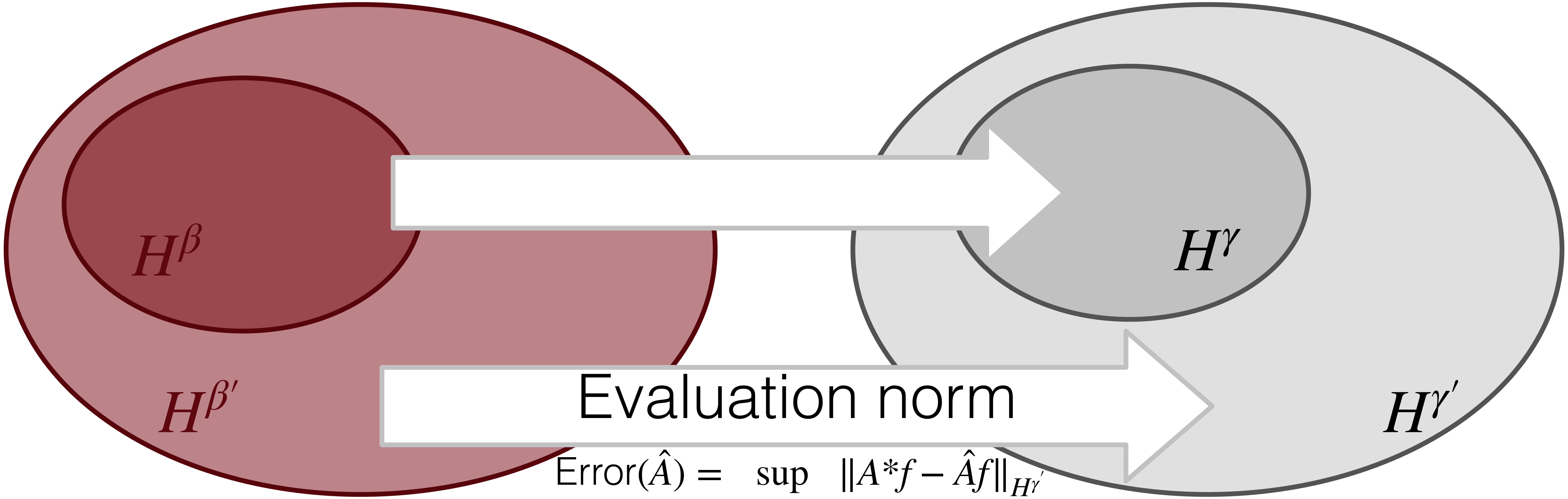
How the optimal rate depend on γ (output space complexity)?

Is the previous algorithm still Optimal?

Problem Formulation

Learn an operator A^* with bounded $\|\cdot\|_{H^\beta \to H^\gamma}$ norm
Respect to $\|\cdot\|_{H^{\beta'} \to H^{\gamma'}}$

Hilbert-schmidt norm



$$\text{Error}(\hat{A}) = \sup_{\|f\|_{H^{\beta'}} \le 1} \|A^* f - \hat{A} f\|_{H^{\gamma'}}$$

Main Result: Lower bound

Learn an operator A^* with bounded $\|\cdot\|_{H^\beta \to H^\gamma}$ norm
Respect to $\|\cdot\|_{H^{\beta'} \to H^{\gamma'}}$ Hilbert-schmidt norm

For all (randomized) estimators $\mathscr{L}$, we have

$$\sup_{\|A\|_{H^\beta \to H^\gamma} \le 1} \|\mathscr{L}(\{u_i, f_i\}_{i=1}^N) - A\|^2_{H^{\beta'} \to H^{\gamma'}} \gtrsim N^{-\min\{\frac{\beta-\beta'}{\beta+p}, \frac{\gamma-\gamma'}{\gamma}\}}$$

With ***N*** random observations

Main Result: Lower bound

Learn an operator A^* with bounded $\|\cdot\|_{H^\beta \to H^\gamma}$ norm
Respect to $\|\cdot\|_{H^{\beta'} \to H^{\gamma'}}$ Hilbert-schmidt norm

For all (randomized) estimators $\mathscr{L}$, we have

Only output function space

$$\sup_{\|A\|_{H^\beta \to H^\gamma} \le 1} \|\mathscr{L}(\{u_i, f_i\}_{i=1}^N) - A\|^2_{H^{\beta'} \to H^{\gamma'}} \gtrsim N^{-\min\{\boxed{\frac{\beta - \beta'}{\beta + p}}, \boxed{\frac{\gamma - \gamma'}{\gamma}}\}}$$

Only input function space

With ***N*** random observations

Same rate as previous work
p : Eigen-decay of RKHS

New Rate in the literature caused by infinite dimensional output

Main Result: Lower bound

Learn an operator A^* with bounded $\|\cdot\|_{H^\beta \to H^\gamma}$ norm
Respect to $\|\cdot\|_{H^{\beta'} \to H^{\gamma'}}$ Hilbert-schmidt norm

For all (randomized) estimators $\mathscr{L}$, we have

Only output function space

$$\sup_{\|A\|_{H^\beta \to H^\gamma} \le 1} \|\mathscr{L}(\{u_i, f_i\}_{i=1}^N) - A\|^2_{H^{\beta'} \to H^{\gamma'}} \gtrsim N^{-\min\{\frac{\beta-\beta'}{\beta+p}, \frac{\gamma-\gamma'}{\gamma}\}}$$

Only input function space

With ***N*** random observations

Reason we introduce the test norm

Main Result: Lower bound

Learn an operator A^* with bounded $\|\cdot\|_{H^\beta\to H^\gamma}$ norm
Respect to $\|\cdot\|_{H^{\beta'}\to H^{\gamma'}}$ Hilbert-schmidt norm

For all (randomized) estimators $\mathscr{L}$, we have

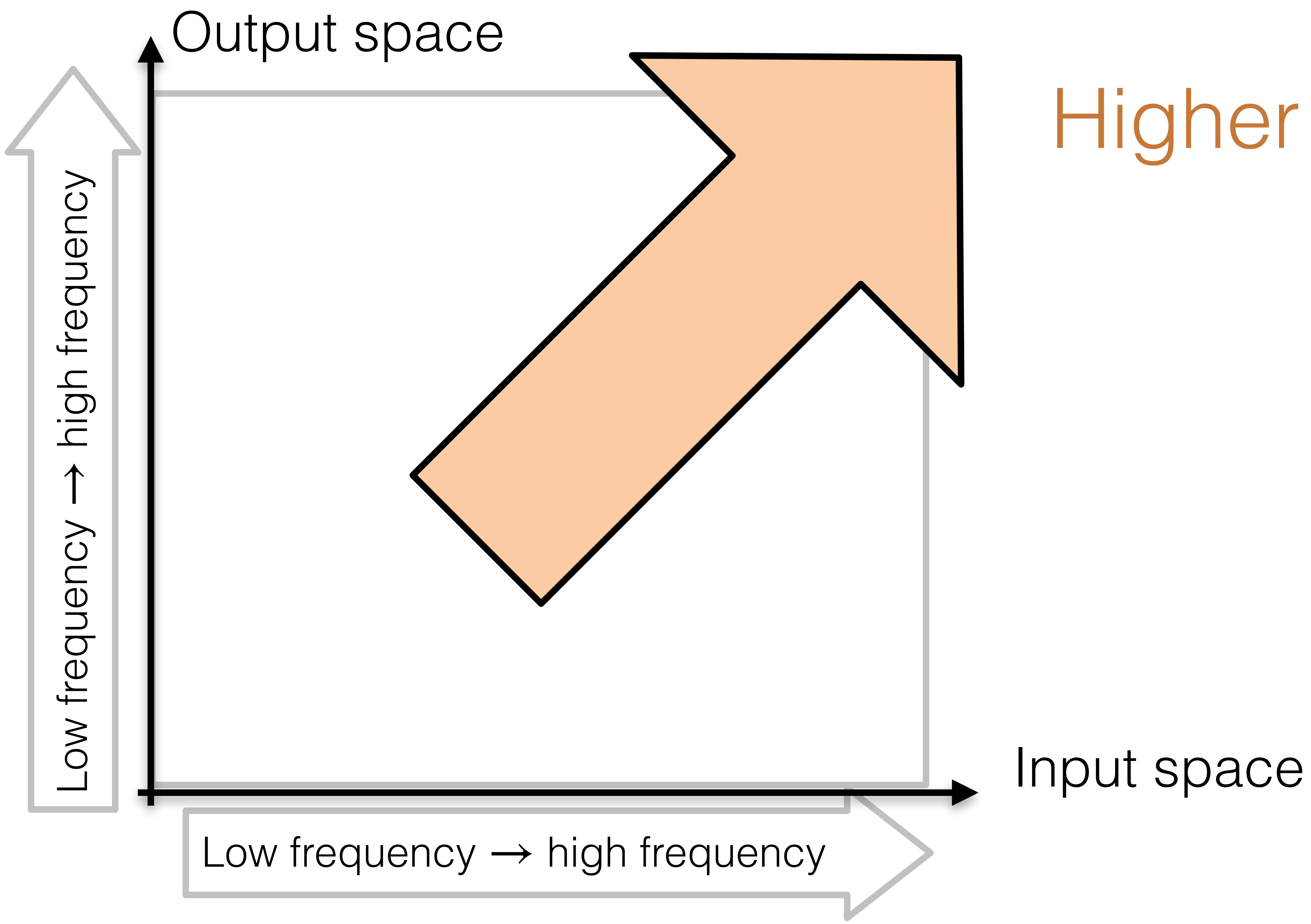
$$\sup_{\|A\|_{H^\beta\to H^\gamma}\le 1} \|\mathscr{L}(\{u_i,f_i\}_{i=1}^N)-A\|^2_{H^{\beta'}\to H^{\gamma'}} \gtrsim N^{-\min\{\frac{\beta-\beta'}{\beta+p},\frac{\gamma-\gamma'}{\gamma}\}}$$

With ***N*** random observations

A magic result, can you explain it to me in a simple way?

Consider the matrix view…

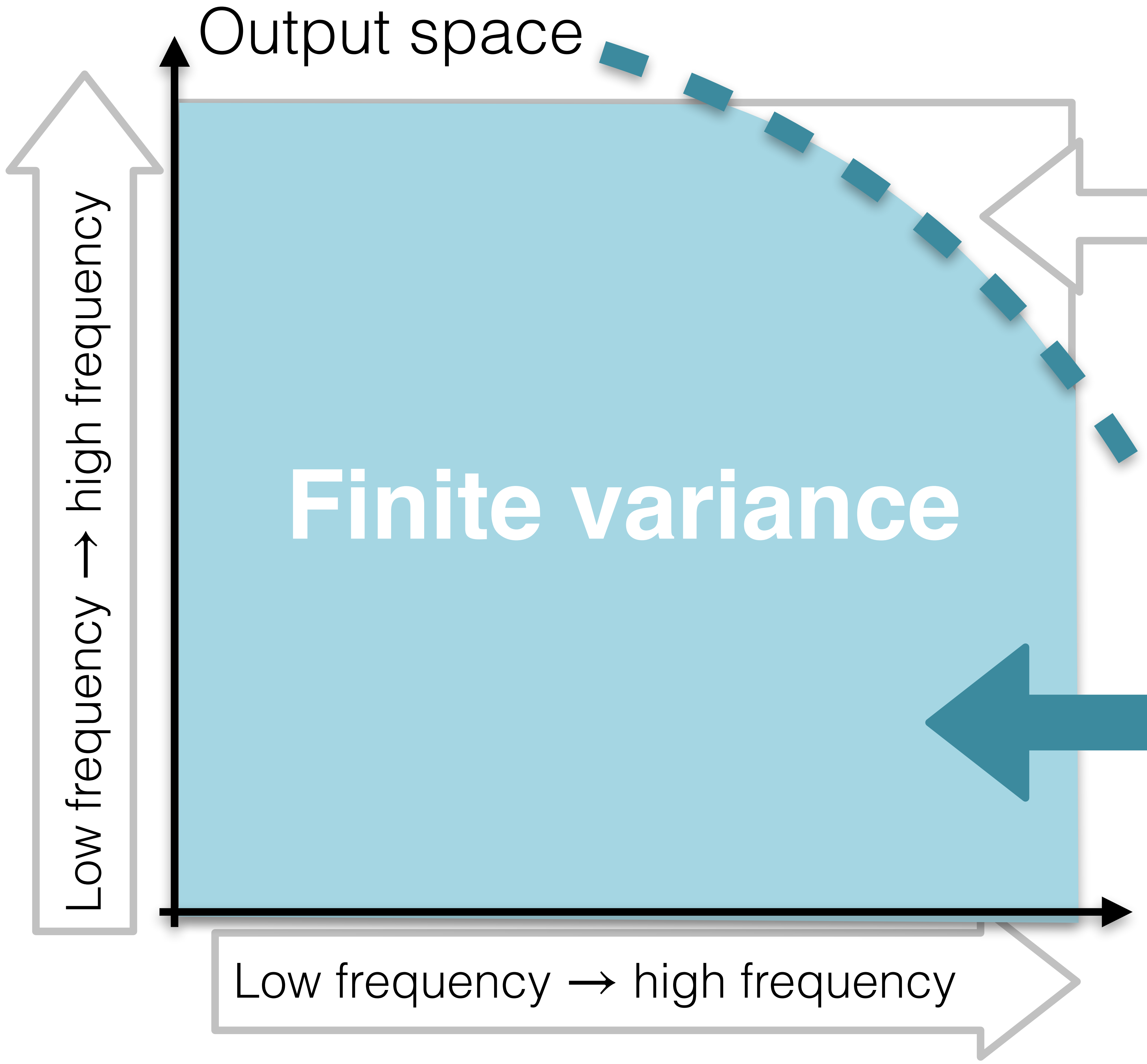
Operator is an "infinite" dimensional "matrix"



Higher Variance but Smaller Bias

Bias Variance Tradeoff

What is needed to achieve N^θ learning rate



Ignore part of the matrix

Learn part of the matrix

<ins>*"Trade off"*</ins>

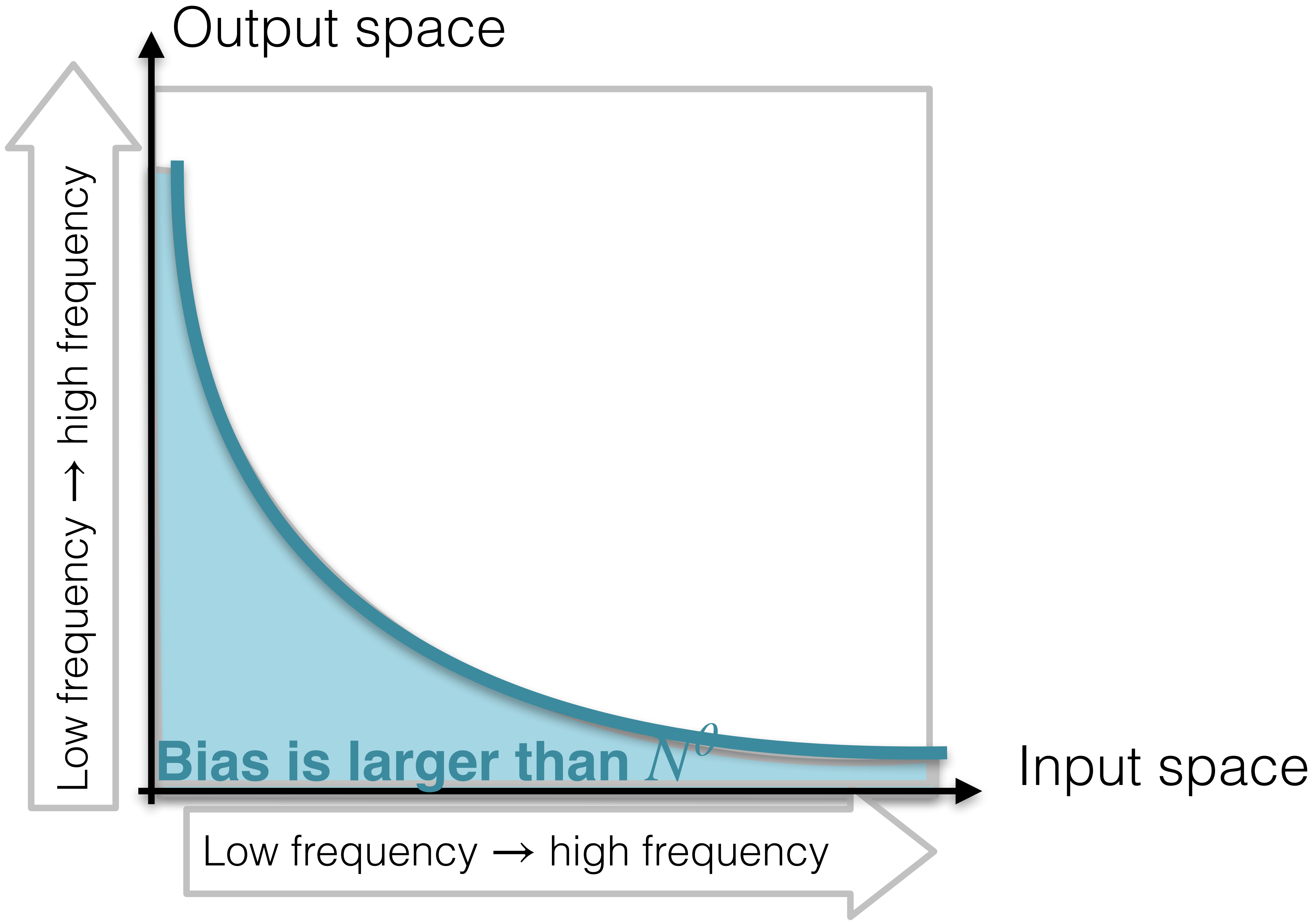
Bias
approximation error

+

Variance

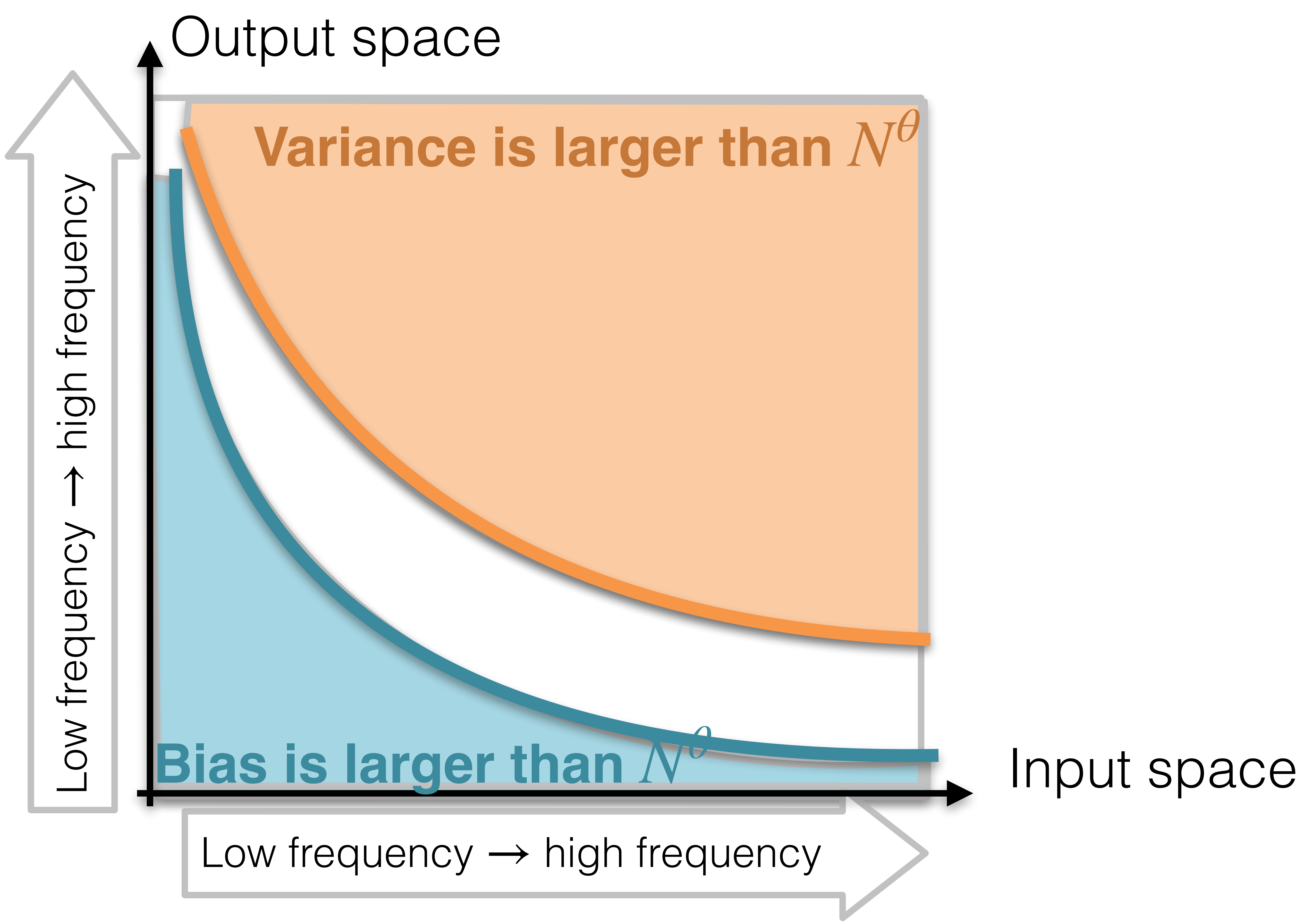
Optimal shape for Bias Variance Trade Off

What is needed to achieve N^{θ} learning rate



Optimal shape for Bias Variance Trade Off

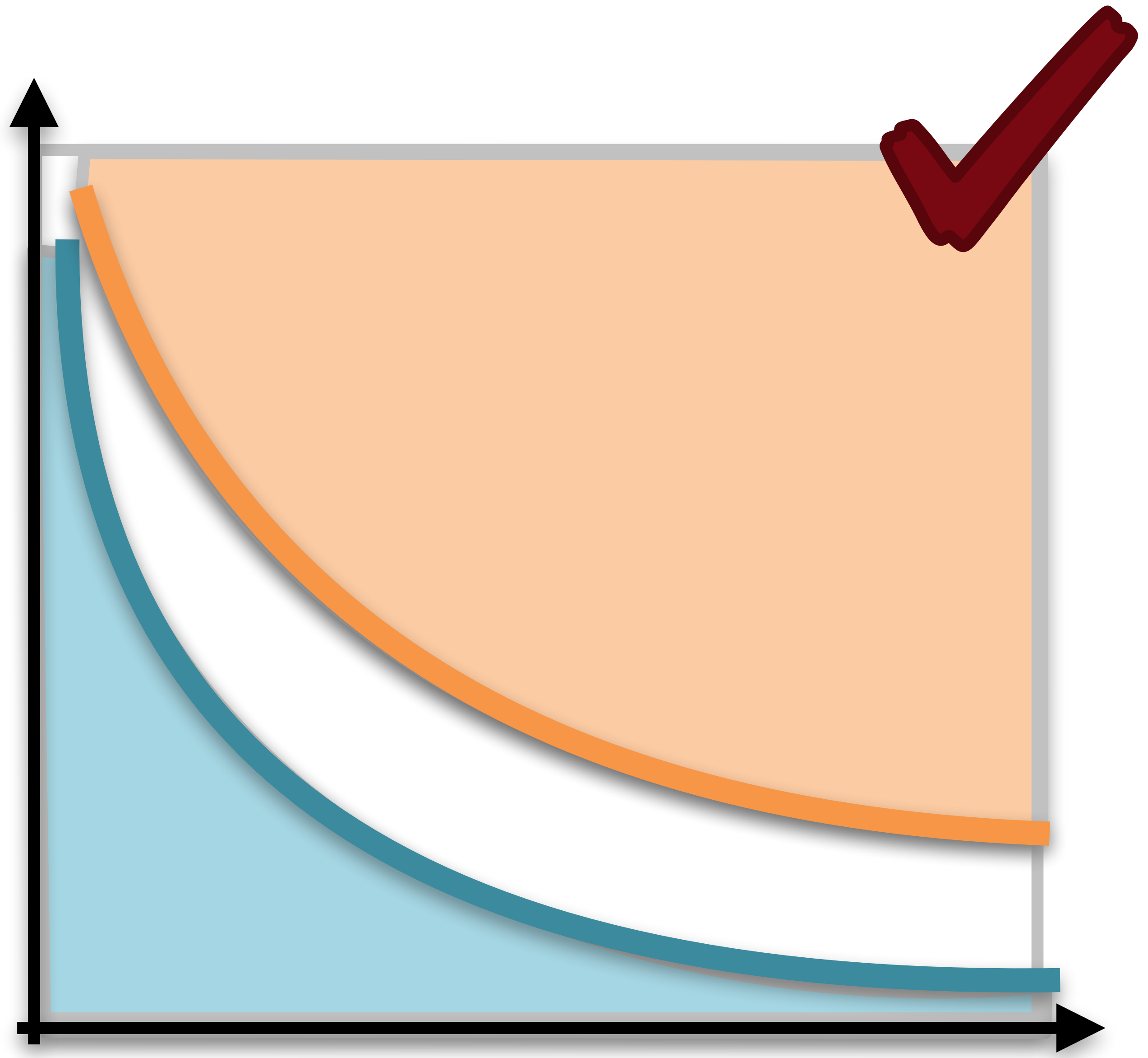
What is needed to achieve N^{θ} learning rate

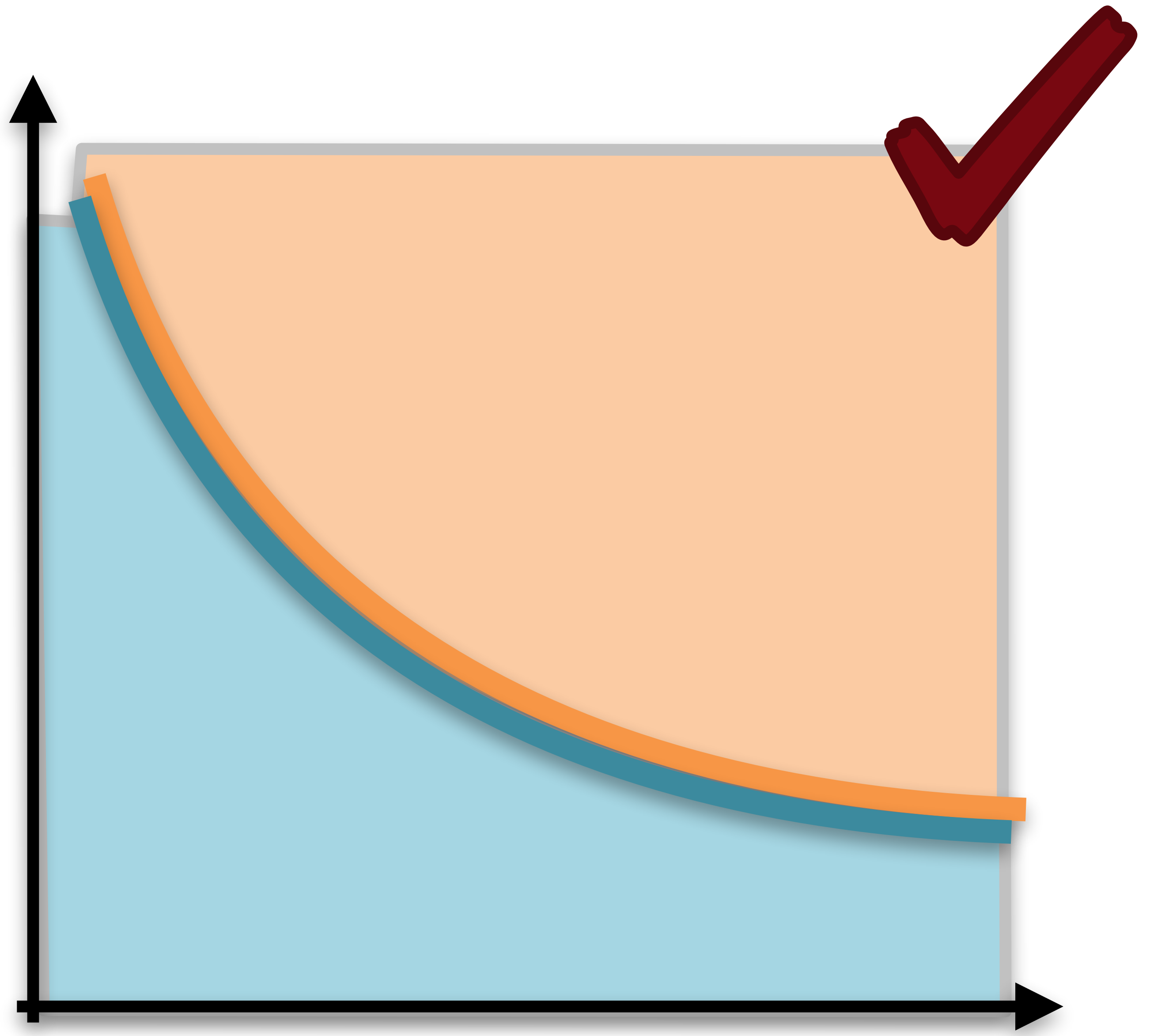


Optimal shape for Bias Variance Trade Off

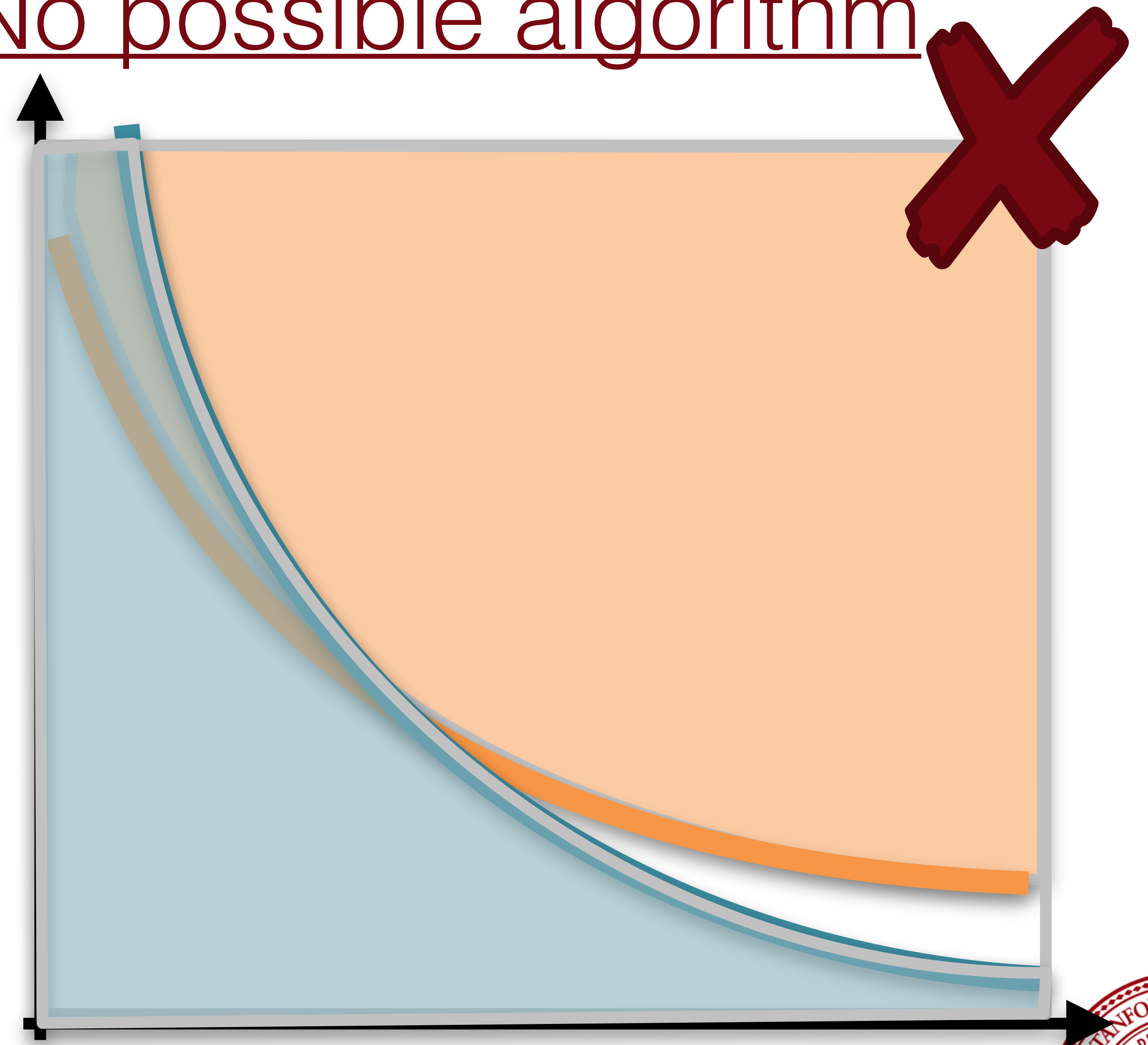
What is needed to achieve N^{θ} learning rate

When θ varies, there are three possible cases



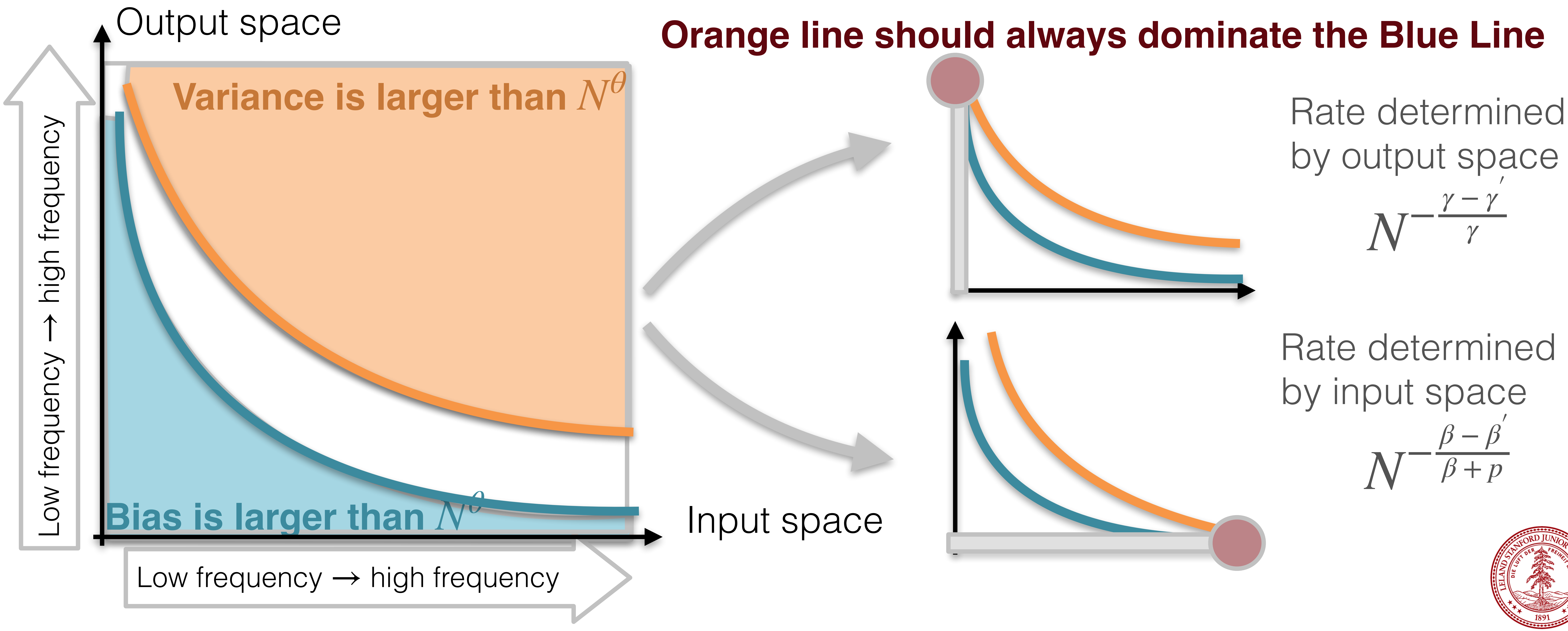


No possible algorithm



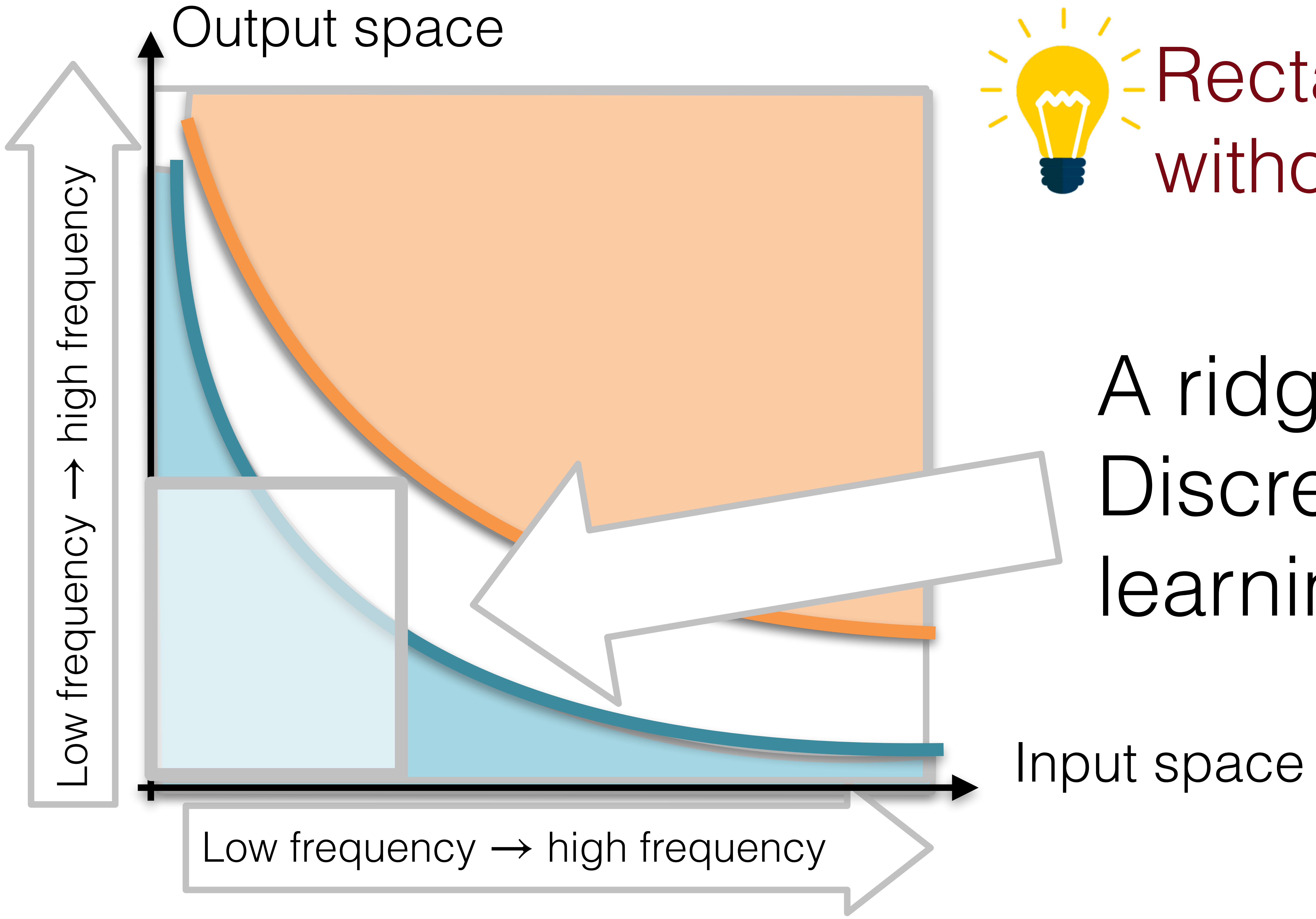
Optimal shape for Bias Variance Trade Off

What is needed to achieve N^{θ} learning rate



Optimal Algorithm

What is the OPTIMAL machine learning algorithm?

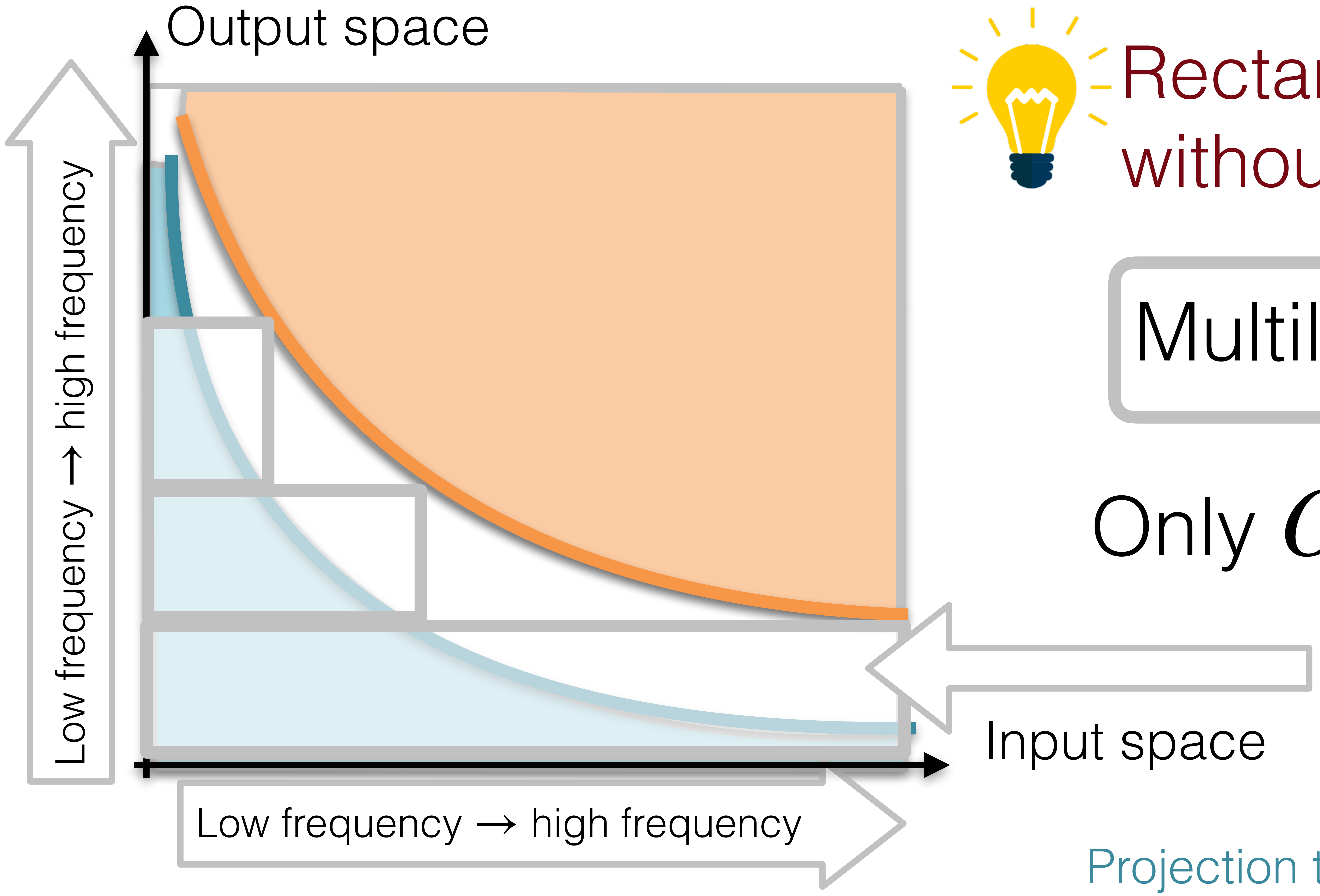


Rectangular covering the blue part without touching the orange part

A ridge-regression/ Discretization(PCA-Net) is learning a rectangular

Optimal Algorithm

What is the OPTIMAL machine learning algorithm?



Rectangular covering the blue part without touching the orange part

Multilevel Training

Only $O(\ln\ln N)$ level is needed

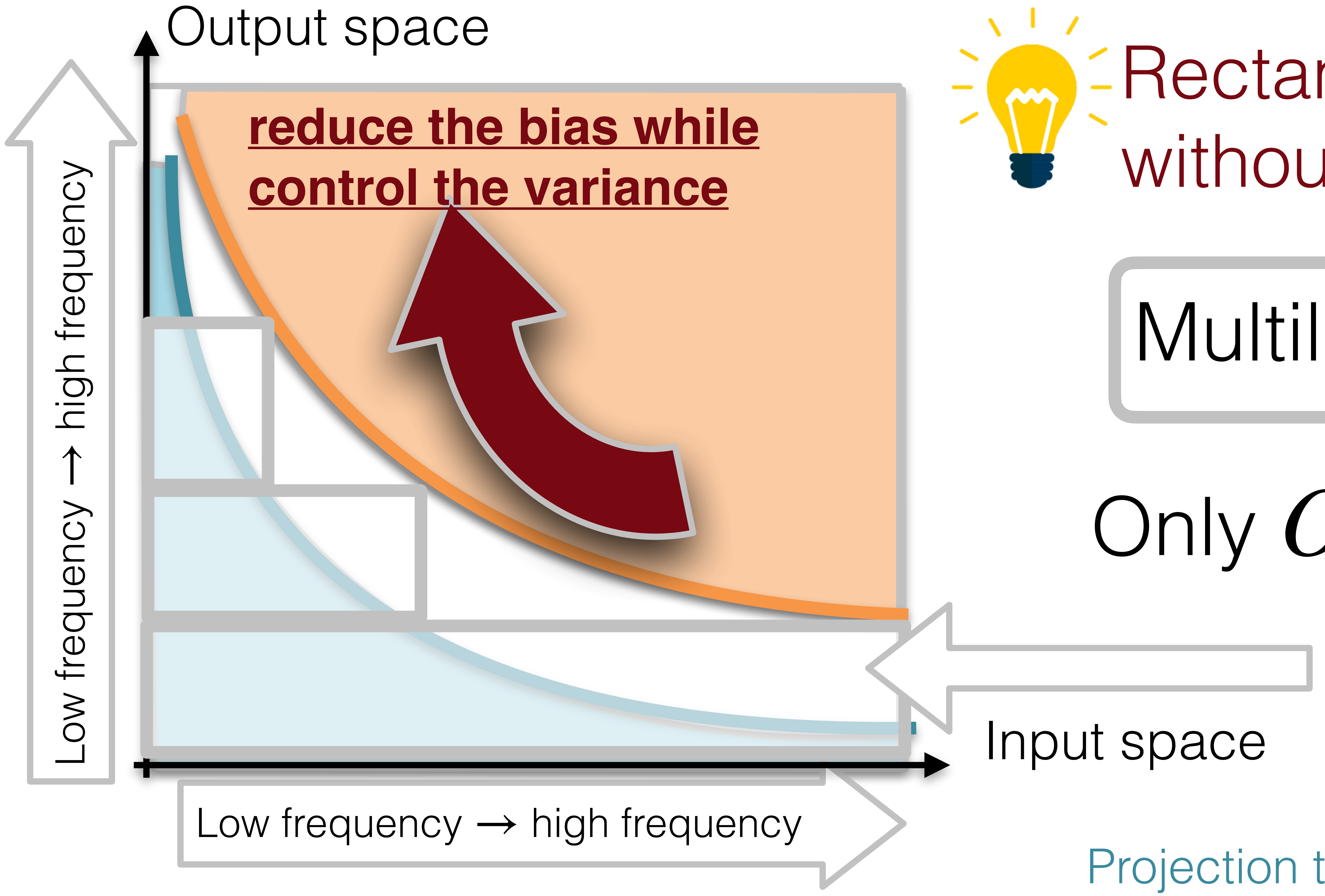
$$\sum_{j\le\gamma_i} \rho_j f_j \otimes \rho_j f_j \qquad \hat{C}_{LK}(\hat{C}_{KK}+\lambda_i^{(K)} I)^{-1}$$

Ridge regression

Projection to certain basis in output space

Optimal Algorithm

What is the OPTIMAL machine learning algorithm?



Rectangular covering the blue part without touching the orange part

Multilevel Training

Only $O(\ln \ln N)$ level is needed

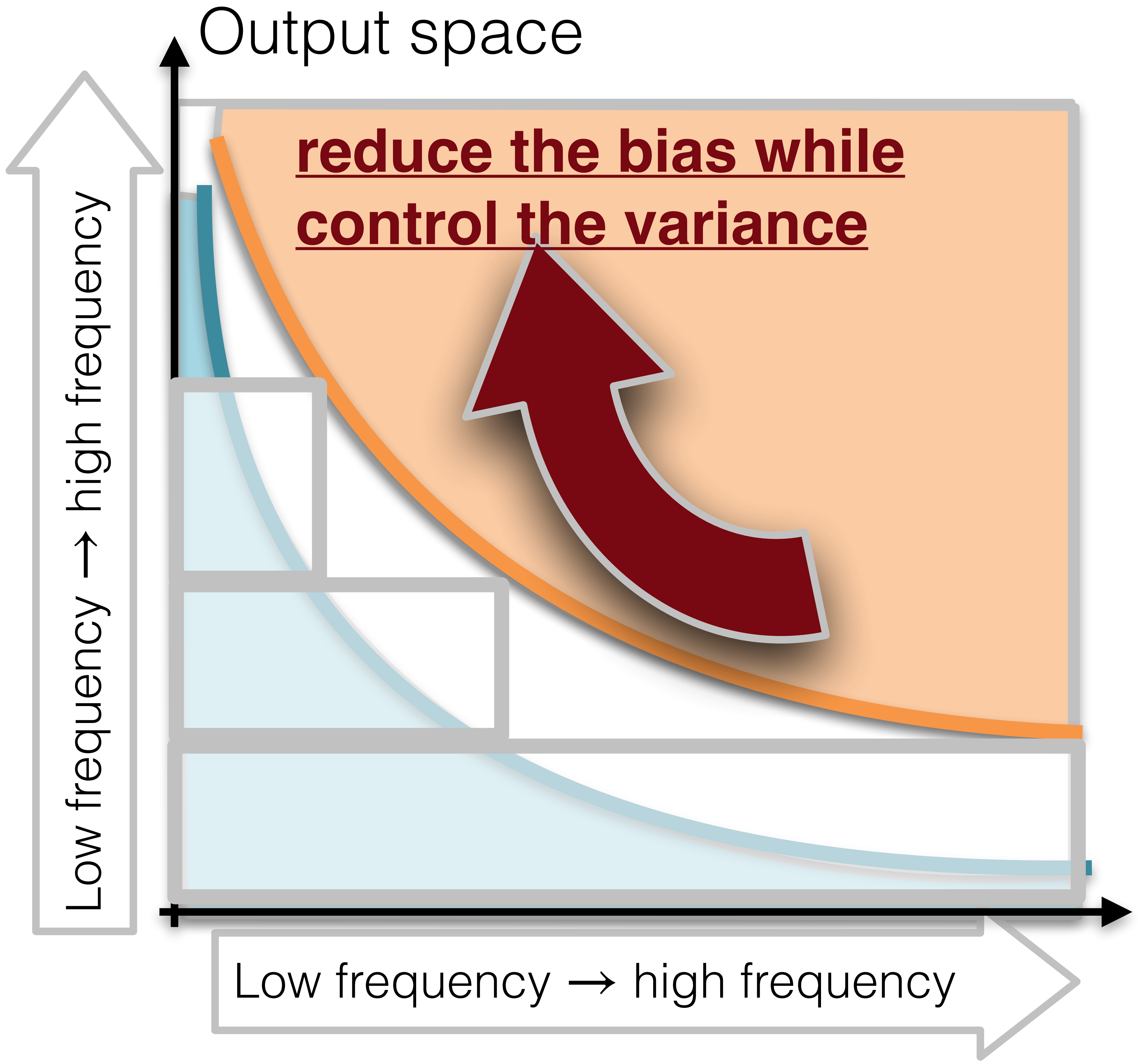
$\sum_{j \le \gamma_i} \rho_j f_j \otimes \rho_j f_j$

Projection to certain basis in output space

$\hat{C}_{LK}(\hat{C}_{KK} + \lambda_i^{(K)} I)^{-1}$

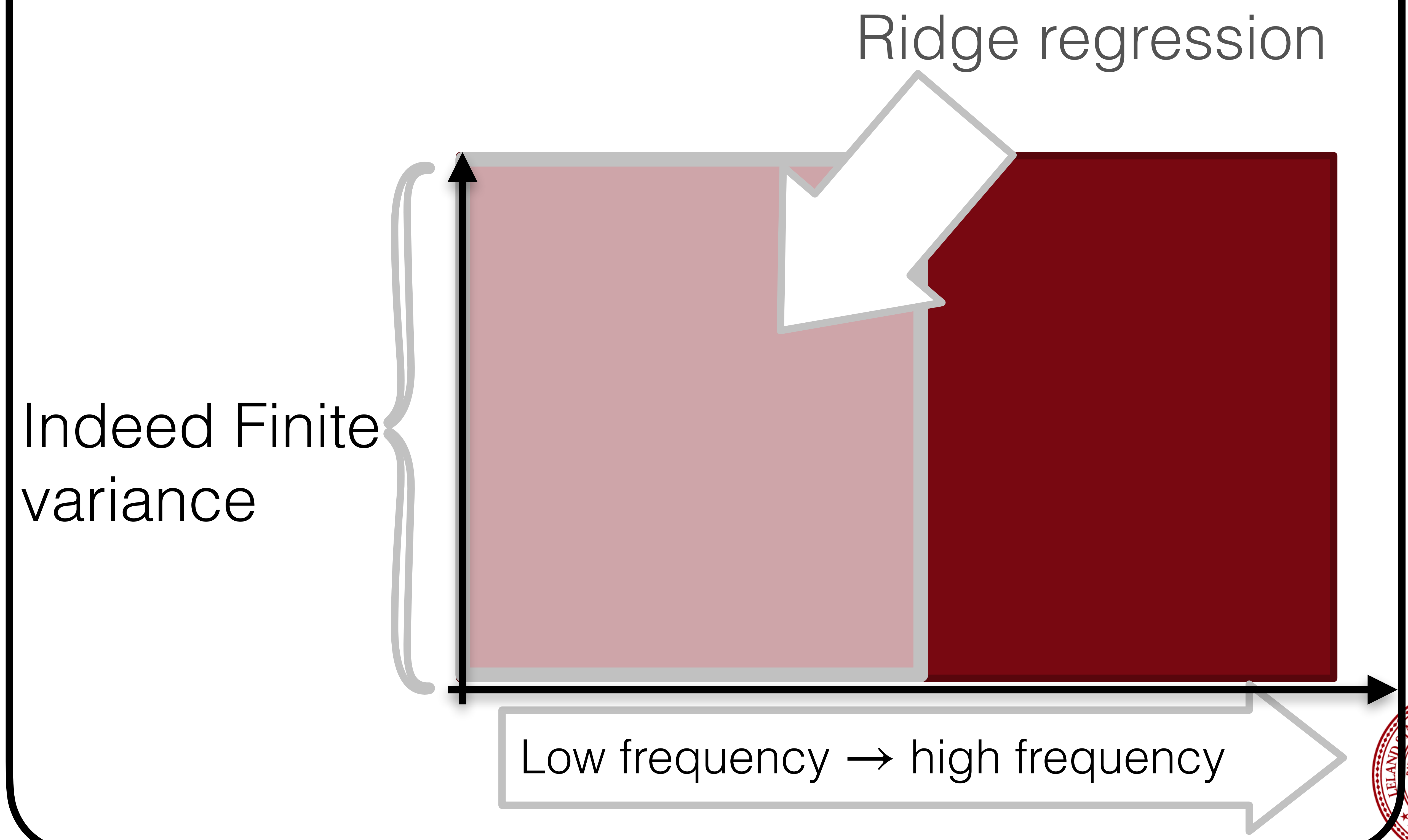
Ridge regression

Optimal Algorithm Changed…



Previous Works

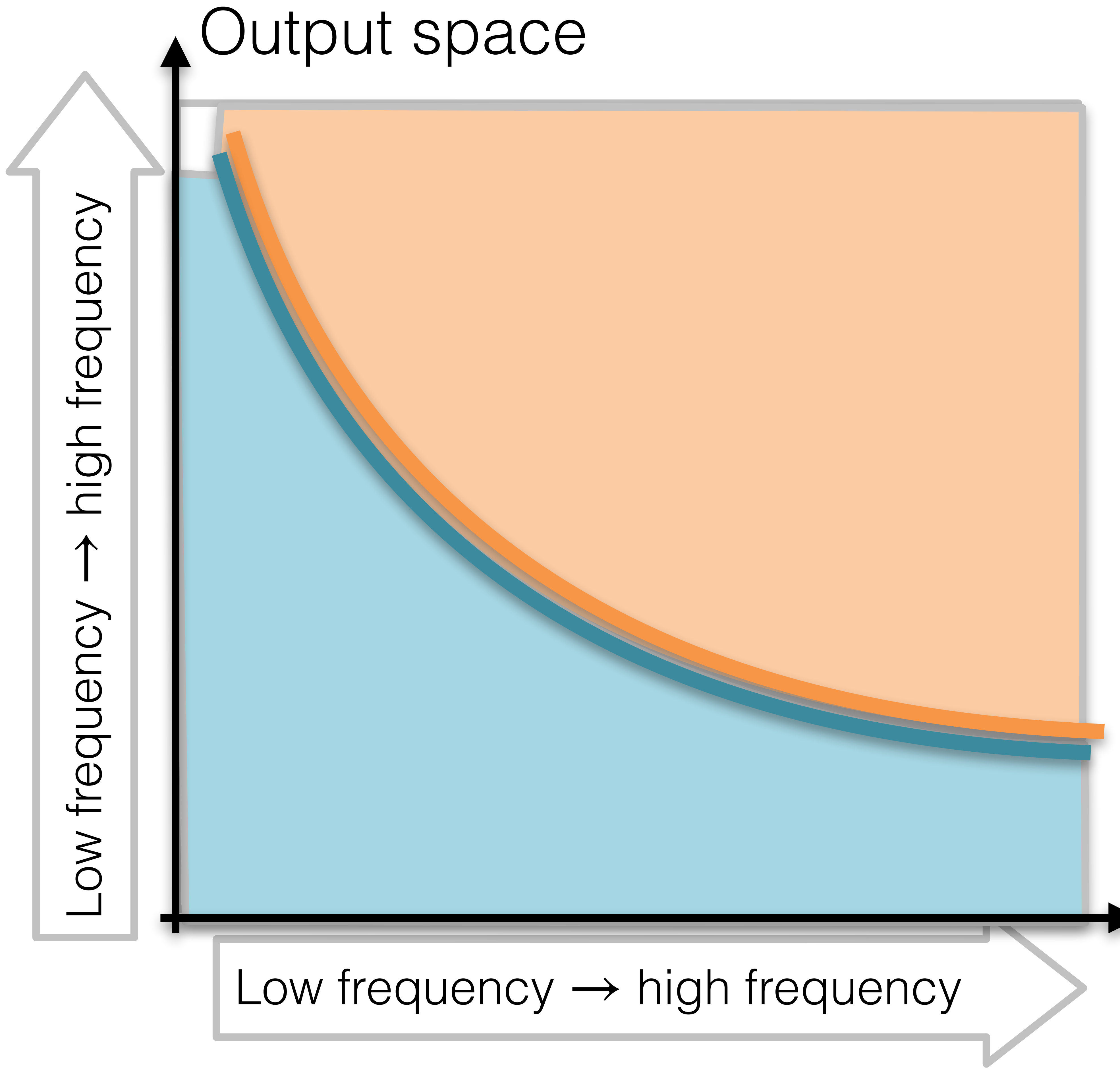
[1] Talwai P, Shameli A, Simchi-Levi D. AISTATS 2022
[2] Li Z, Meunier D, A Gretton. Neurips 2022
[3] de Hoop M V, et al. arXiv:2108.12515



Optimal Algorithm

Multilevel Training

What is the OPTIMAL machine learning algorithm?



What if the two lines coincide?

Output space Learning rate

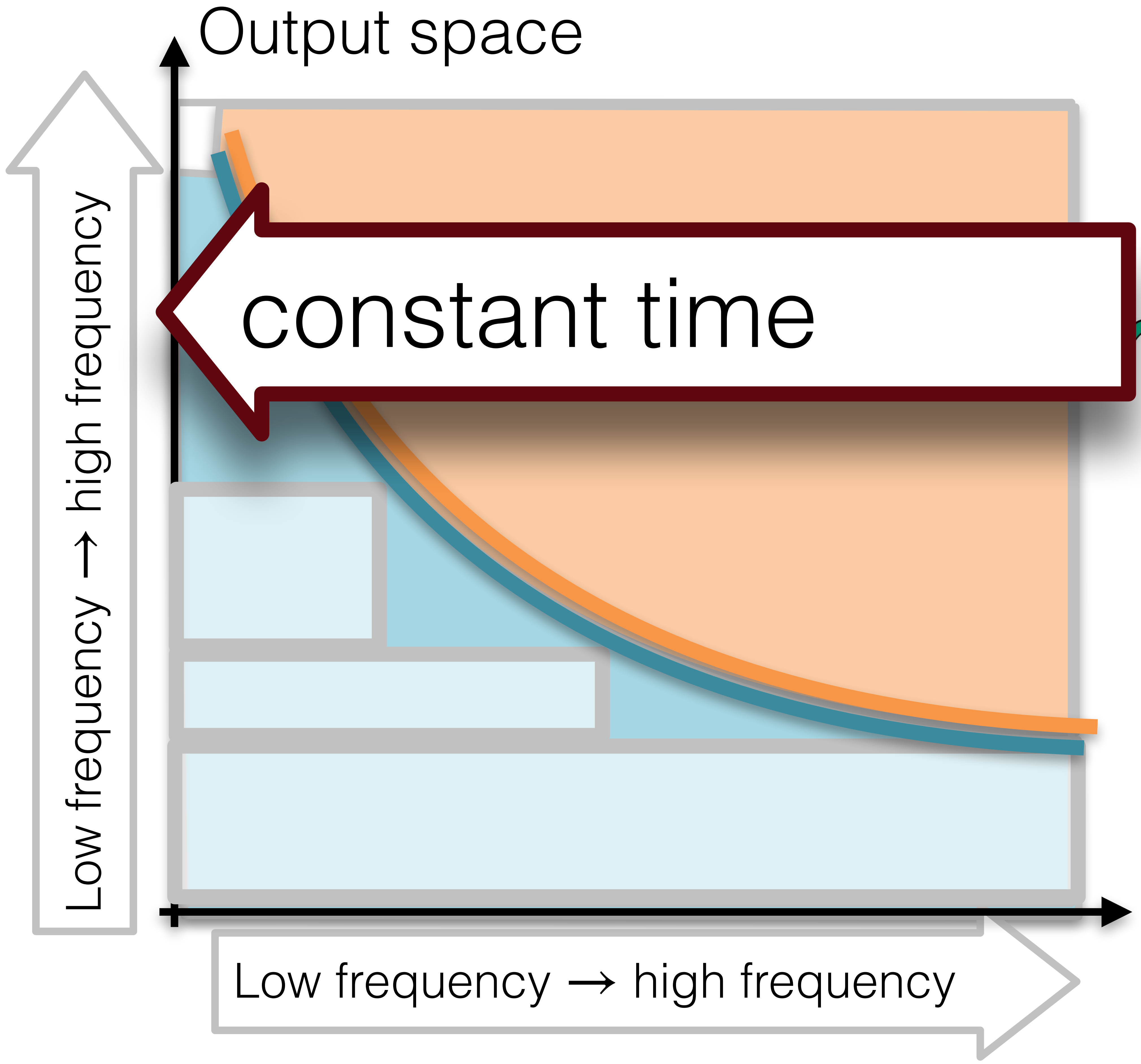
Input space learning rate

$$\frac{\gamma - \gamma'}{\gamma} = \frac{\beta - \beta'}{\beta + p}$$

Optimal Algorithm

Multilevel Training

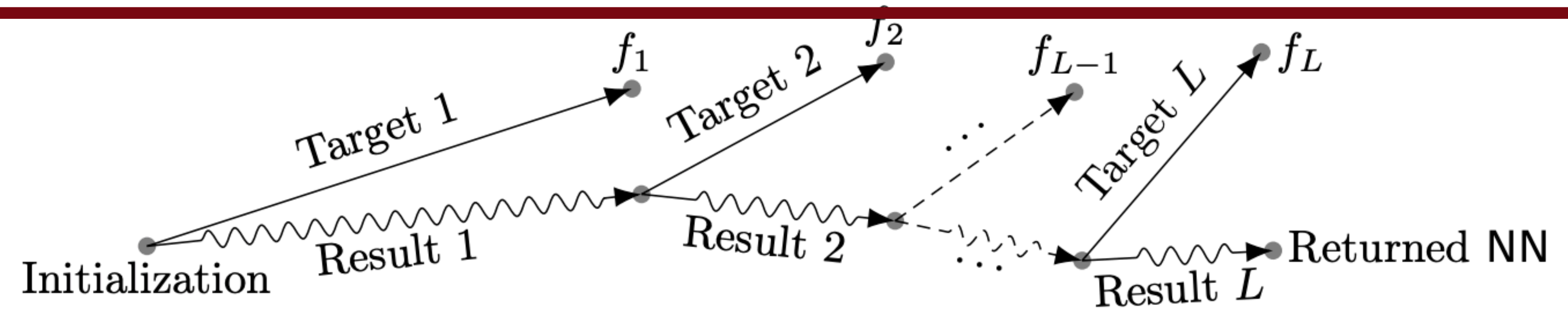
What is the OPTIMAL machine learning algorithm?



What if the two lines coincide?

Only $O(\ln N)$ level is needed

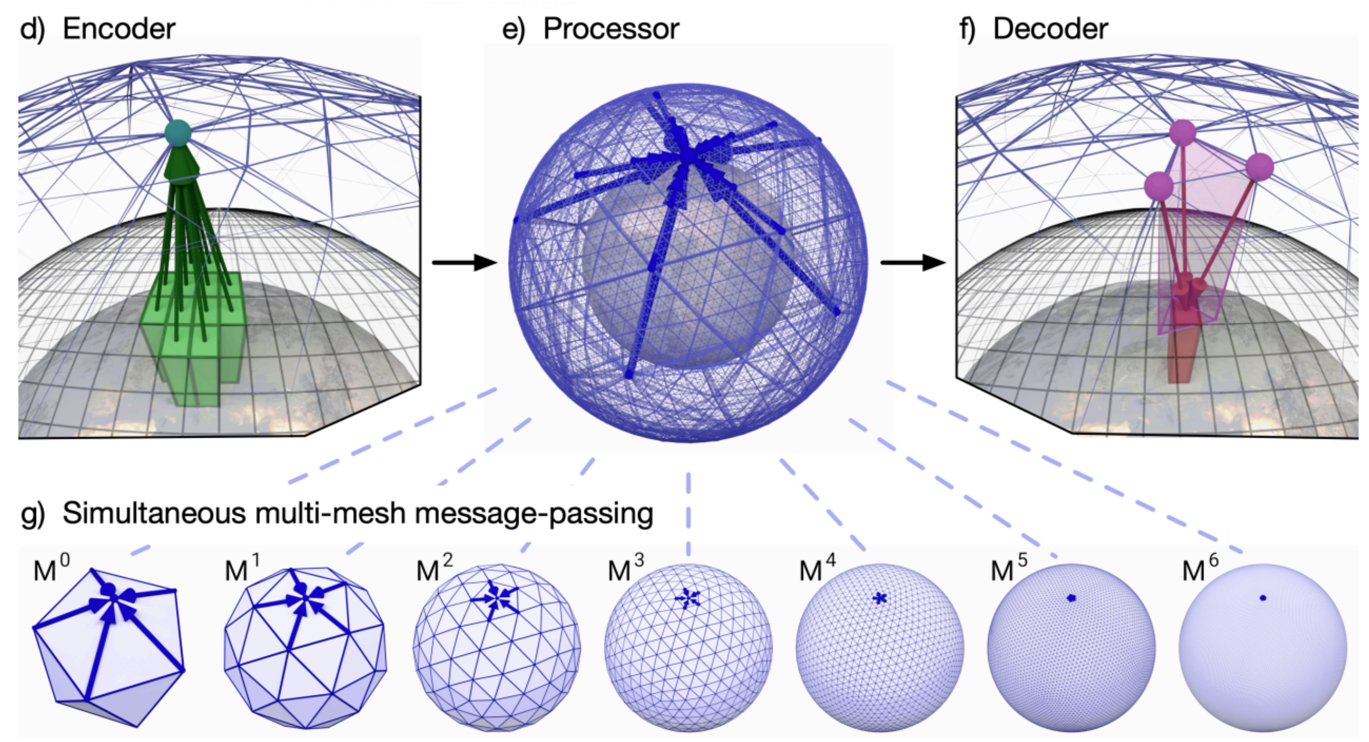
Matches Empirical Using



Coarse grid → Fine grid

Fast reconstruction of hierarchical matrix/ Green function ***Linear Case***
[Lin-Lu-Ying 11][Boullé-Kim-Shi-Townsend 22] [Schäfer-Owhadi 21]…

Multi-level Machine Learning
[Lye-Mishra-Molinaro 21][Li-Fan-Ying 21]



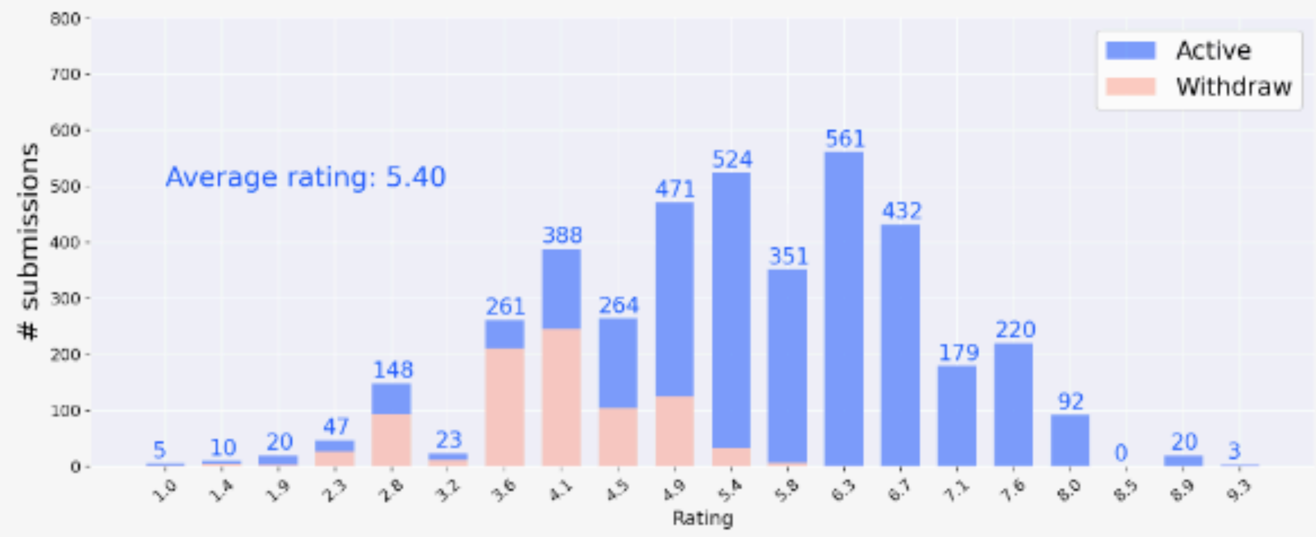
GraphCast: Learning skillful medium-range global weather forecasting

Remi Lam*,1, Alvaro Sanchez-Gonzalez*,1, Matthew Willson*,1, Peter Wirnsberger*,1, Meire Fortunato*,1, Alexander Pritzel*,1, Suman Ravuri[1], Timo Ewalds[1], Ferran Alet[1], Zach Eaton-Rosen[1], Weihua Hu[1], Alexander Merose[2], Stephan Hoyer[2], George Holland[1], Jacklynn Stott[1], Oriol Vinyals[1], Shakir Mohamed[1] and Peter Battaglia[1]
*equal contribution, [1]DeepMind, [2]Google

https://arxiv.org/pdf/2212.12794.pdf

ICLR Statistics

👉 R7 : ratings @2022-12-17 | Rating distribution:



👉 R6 : ratings @2022-12-11 | Rating distribution.
👉 R5 : ratings @2022-12-04 | Rating distribution.
👉 R4 : ratings @2022-11-28 | Rating distribution.
👉 R3 : ratings @2022-11-21 | Rating distribution.
👉 R2 : ratings @2022-11-17 | Rating distribution.
👉 R1 : ratings @2022-11-05 | Rating distribution.
👉 ΔR : R7-R1.
👉 ICLR 2022 statistics.

Ranked top 4/4126 in all ICLR 2023 submissions

All Submissions | Statistics

# (40419)	Title	R1	R7	R7-std	ΔR	Ratings
1	Universal Few-shot Learning of Dense Prediction Tasks with Visual Token Matching	8.00	9.33	0.94	1.33	10, 8, 6 10, 8, 10
2	Emergence of Maps in the Memories of Blind Navigation Agents	8.50	9.00	1.00	0.50	8, 8, 8, 10 8, 8, 10, 10
3	Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning	8.25	9.00	1.00	0.75	8, 10, 10, 5 8, 10, 10, 8
4	Minimax Optimal Kernel Operator Learning via Multilevel Training	7.40	8.80	0.98	1.40	10, 5, 8, 8, 6 10, 8, 8, 8, 10

Take home message

Learning in infinite dimensional space is hard due to the infinite variance

The hardness of learning a linear operator is determined by the harder part between the input and output space
(In some cases, infinite variance will not leads to slower rate)

Single level ML leads to sub-optimal rate, multi-level is needed.
(Matches empirical use)

Current Research

$Au = f$

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19]

Recover parameter θ in model A_θ
E.g. Drift, Diffusion Strength

Learn from data pair $\{u_i, f_i\}$
"Operator Learning/Functional data analysis"

Methodology

[Brunton-Proctor-Kutz 16][Khoo-Lu-Ying 18] [Long-Lu-Li-Dong 18][Lu-Jin-Pang-Zhang-Karniadakis 20] [Li-Kovachki-…-Stuart-Anandkumar 20]

Theory

[Lanthaler-Mishra-Karniadakis 22] [Talwai-Shameli-Simchi-Levi 21][de Hoop-Kovachki-Nelsen-Stuart 21][Li-Meunier-Mollenhauer-Gretton 22] [Liu-Yang-Chen-Zhao-Liao 22]….

[Jin-Lu-Blanchet-Ying 23]

ong 20] [Liang-Yang 22]…

[Nickl-Ray 20] [Nickl 20] [Baek-Farias-Georgescu-Li-Peng-Sinha-Wilde-Zheng 20] [Agrawl-Yin-Zeevi 21]…

Current Research

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19] [Rahme-Jelassi-Matt Weinberg 21]

Main Idea

Change solving the model to solving a minimization problem

Example: $\Delta u = f$

$$\int |\nabla u(x)|^2 - 2u(x)f(x)dx \qquad \int (\Delta u - f)^2 dx$$

sub-optimal ***optimal***

[**Lu**-Chen-Lu-Ying-Blanchet ICLR22]
Direct Sample Average Approximation is not optimal for all criteria.

Minimax Lower Bound+"Fast rate generalization bound"

Current Research

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology
[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control a

[Guo-Hu-Xu
Zariphopou

Auction

[Duetting-F
Ravindranath 19] [Rahme-Jelassi-Matt Weinberg 21]

DRM discretized
$\nabla \cdot \nabla$
But not Δ

Main Idea

Change solving the model to solving a minimization problem

Example: $\Delta u = f$

$$\int |\nabla u(x)|^2 - 2u(x)f(x)dx$$

sub-optimal

$$\int (\Delta u - f)^2 dx$$

optimal

[**Lu**-Chen-Lu-Ying-Blanchet ICLR22]
Direct Sample Average Approximation is not optimal for all criteria.

Minimax Lower Bound+"Fast rate generalization bound"

Current Research

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19] [Rahme-Jelassi-Matt Weinberg 21]

Main Idea

Change solving the model to solving a minimization problem

Example: $\Delta u = f$

$$\int |\nabla u(x)|^2 - 2u(x)f(x)dx \qquad \int (\Delta u - f)^2 dx$$

“implicit Sobolev acceleration”

Faster

[**Lu**-Blanchet-Ying Neurips22] analysis the optimization dynamic.

Using sobolev norm as loss function can accelerate optimization

Current Research

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19] [Rahme-Jelassi-Matt Weinberg 21]

Main Idea

Change solving the model to solving a minimization problem

Example: $\Delta u = f$

$$\int |\nabla u(x)|^2 - 2u(x)f(x)dx \qquad \int (\Delta u - f)^2 dx$$

Pre-ml Experience：
Double the condition number

Current Research

Reconstruct the solution u
With observation of f: $\{x_i, f(x_i)\}$

Methodology

[Han-Jentzen-E 18] [Yu-E 18] [Raissi-Perdikaris-Karniadakis 19] [Sirignano-Spiliopoulos 18] [Chen-Hosseini-Owhadi-Stuart 21] [Zang-Bao-Ye-Zhou 20]…

Control and MFG

[Guo-Hu-Xu-Zhang 19][Wang-Zariphopoulou-Zhou 21][Dai-Gluzman 22]

Auction

[Duetting-Feng-Narasimhan-Parkes-Ravindranath 19] [Rahme-Jelassi-Matt Weinberg 21]

Main Idea

Change solving the model to solving a minimization problem

Example: $\Delta u = f$

$$\int |\nabla u(x)|^2 - 2u(x)f(x)dx \qquad \int (\Delta u - f)^2 dx$$

$$u = \langle \theta, K_x \rangle$$

"Differential operator preconditions the kernel integral operator"

Research Overview

yplu@stanford.edu

$$Au = f$$

Reconstruct u with observation of f: $\{x_i, f(x_i)\}$

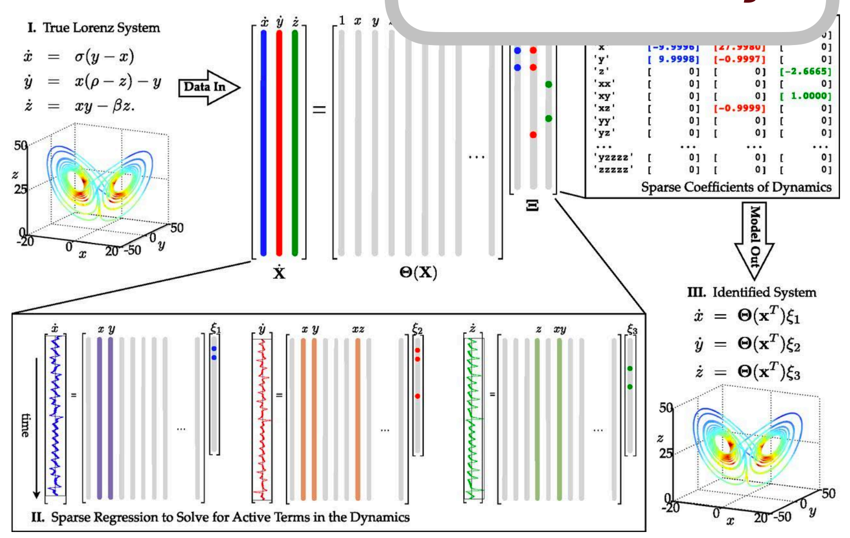
Recover parameter θ in Model A_θ

Learn the model A from data pair $\{u_i, f_i\}$

Interaction between model and data

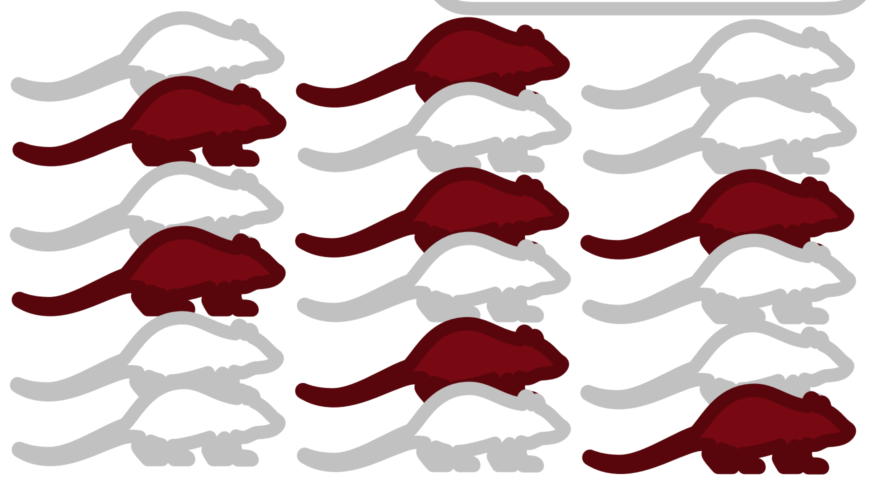
Rough Modeling

Consistency



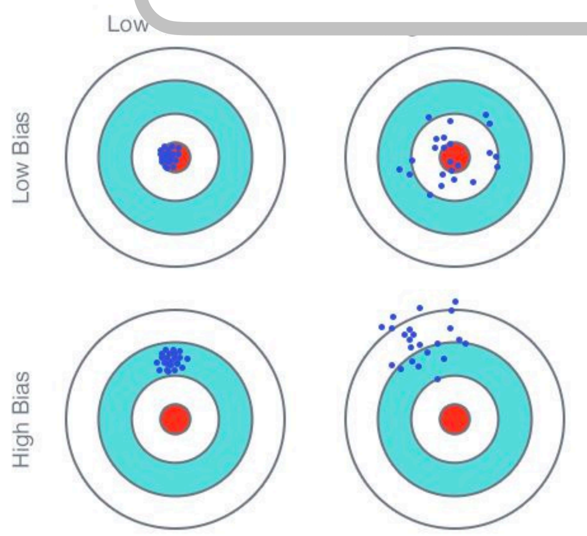
Experiment Design

Computation



Model Learning

Convergence rate



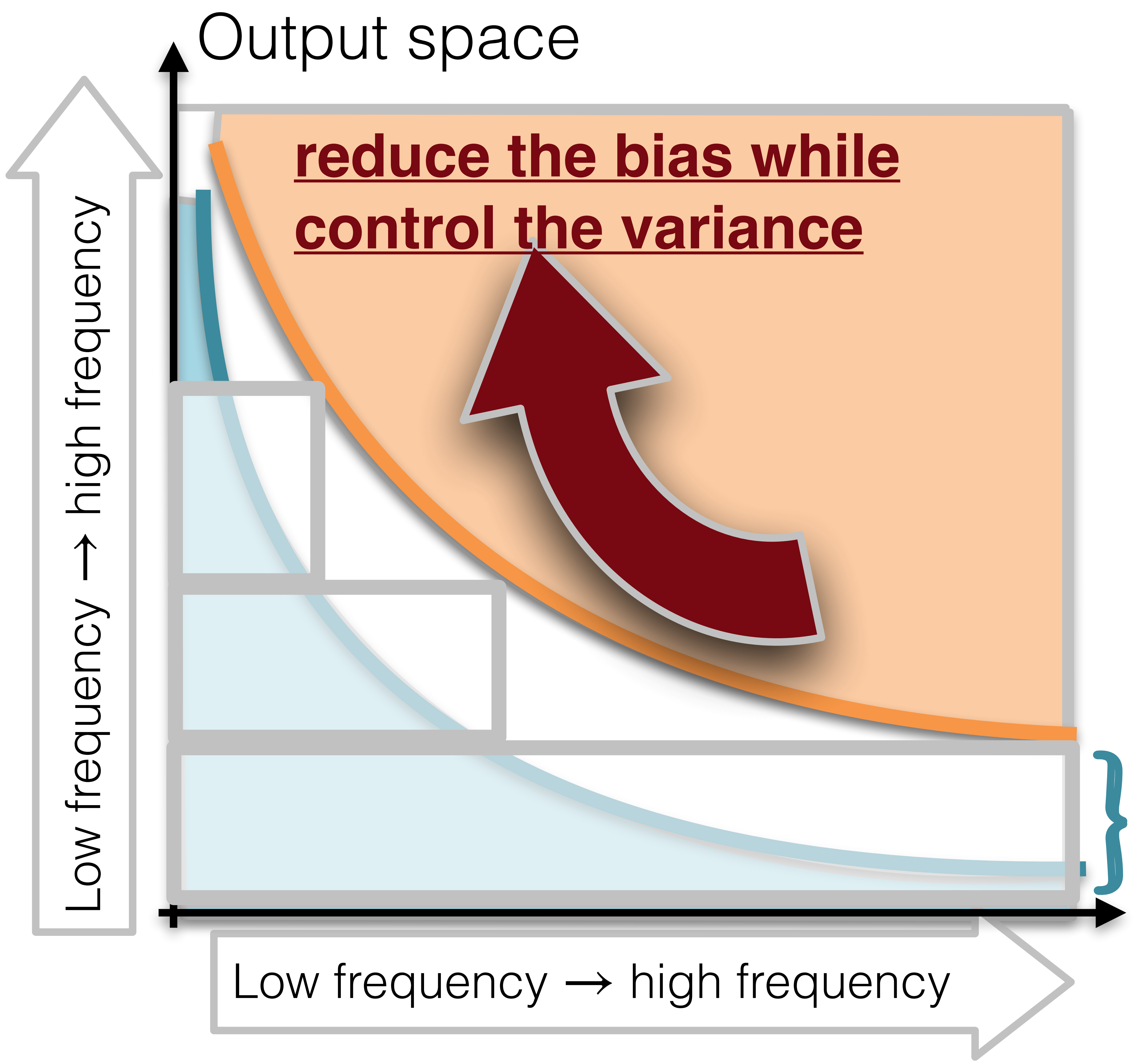
Uncertainty Quantification

Inference

Contact: yplu@stanford.edu

Optimal Algorithm

What is the OPTIMAL machine learning algorithm?



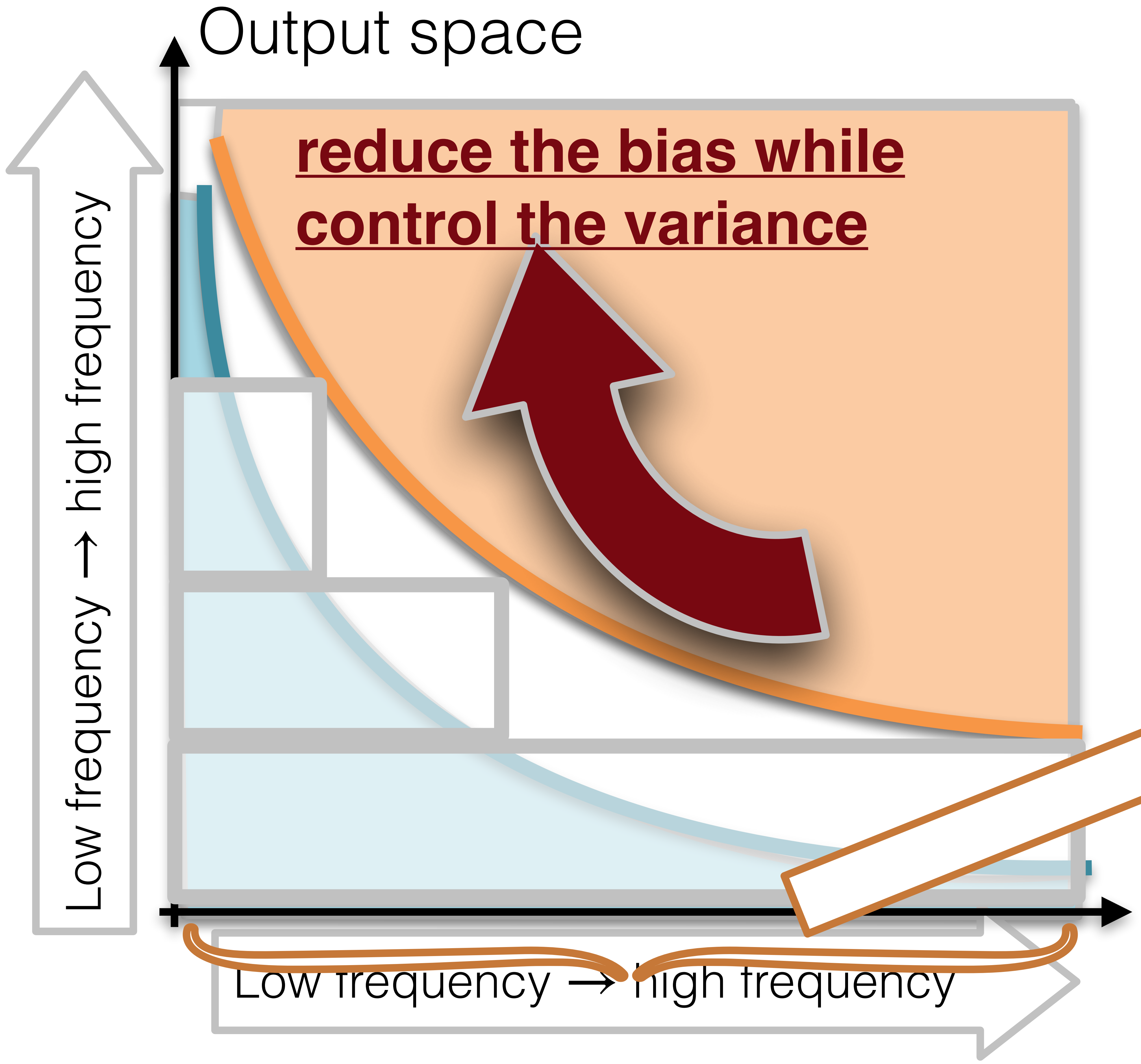
$$\hat{A}_{\mathtt{ml}} = \sum_{i=0}^{L_N} \left(\boxed{\sum_{\gamma_{i-1} \leqslant j < \gamma_i} \rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j} \right) \boxed{\hat{C}_{LK} \left(\hat{C}_{KK} + \lambda_i^{(K)} I \right)^{-1}} .$$

Ridge regression

Projection to certain basis in output space

Optimal Algorithm

What is the OPTIMAL machine learning algorithm?



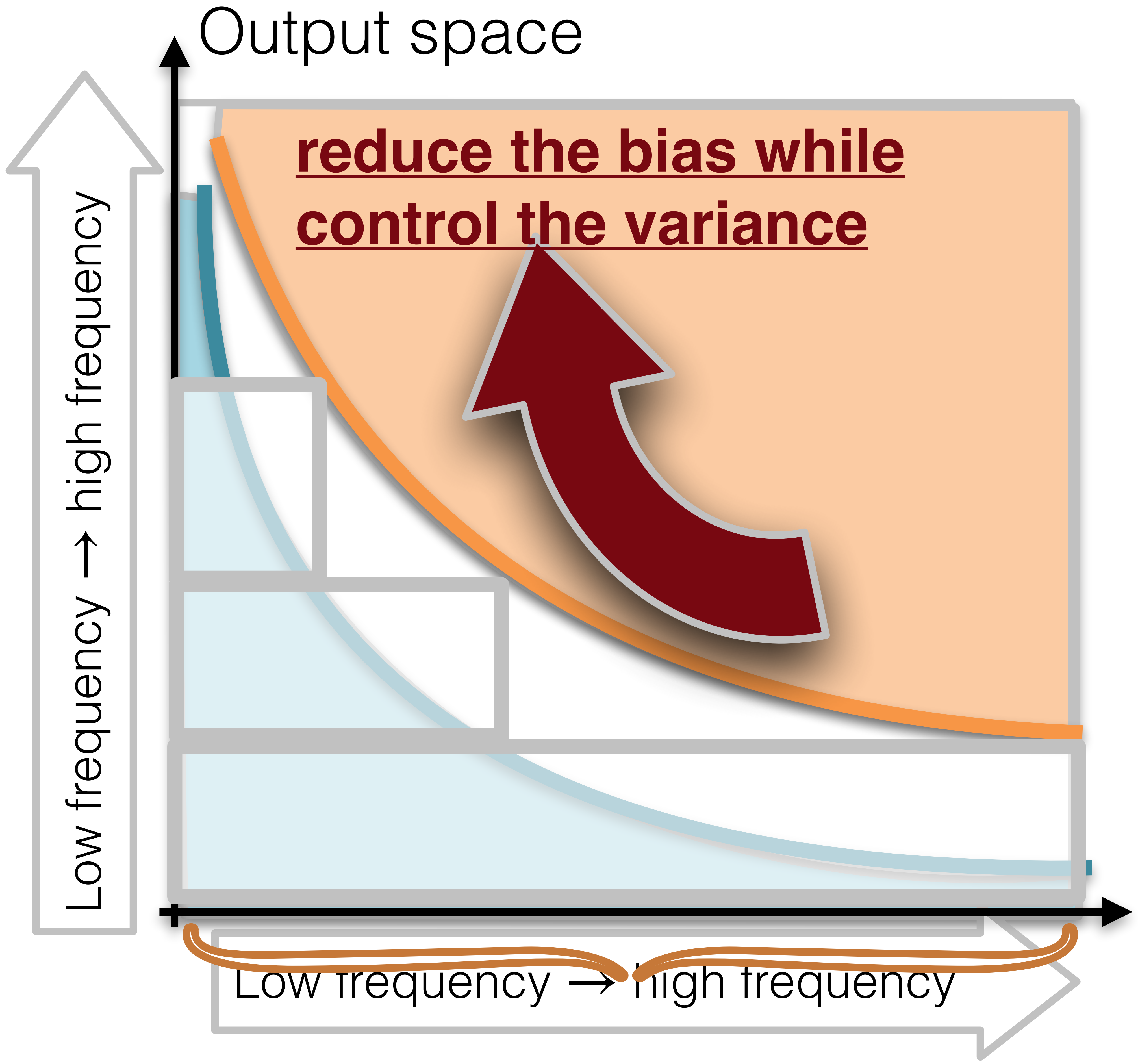
$$\hat{A}_{\mathtt{ml}} = \sum_{i=0}^{L_N} \left(\boxed{\sum_{\gamma_{i-1} \leqslant j < \gamma_i} \rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j} \right) \boxed{\hat{C}_{LK} \left(\hat{C}_{KK} + \lambda_i^{(K)} I \right)^{-1}}.$$

Ridge regression

Projection to certain basis in output space

Optimal Algorithm

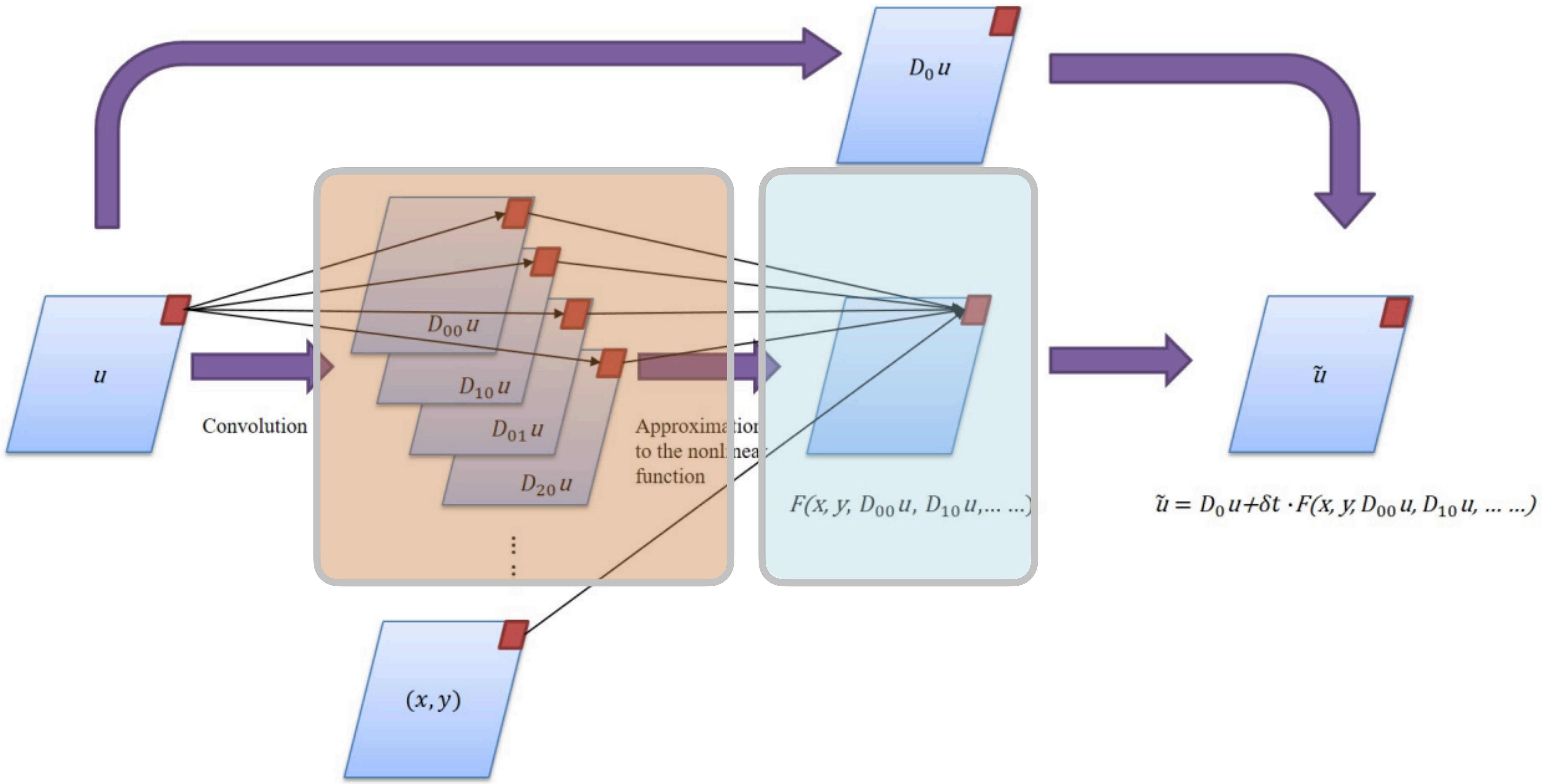
What is the OPTIMAL machine learning algorithm?



$$\hat{A}_{\mathtt{ml}} = \boxed{\sum_{i=0}^{L_N}} \left(\boxed{\sum_{\gamma_{i-1} \leqslant j < \gamma_i} \rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j} \right) \boxed{\hat{C}_{LK} \left(\hat{C}_{KK} + \lambda_i^{(K)} I \right)^{-1}}.$$

Ensemble different levels

Algorithmic Literature Overview



$$\frac{\partial u(x,t)}{\partial t} = F(\boxed{u, \nabla_x u, \nabla_x^2 u, \cdots})$$

Convolutional kernel

"Finite-difference"

$u_x = u * [-1,1]$

Neural Network

Definition 2.1 (Order of Sum Rules). *For a filter q, we say q to have sum rules of order $\alpha = (\alpha_1, \alpha_2)$, where $\alpha \in \mathbb{Z}_+^2$, provided that*

$$\sum_{k \in \mathbb{Z}^2} k^\beta q[k] = 0 \tag{2}$$

for all $\beta = (\beta_1, \beta_2) \in \mathbb{Z}_+^2$ with $|\beta| := \beta_1 + \beta_2 < |\alpha|$ and for all $\beta \in \mathbb{Z}_+^2$ with $|\beta| = |\alpha|$ but $\beta \neq \alpha$. If (2) holds for

Long Z, Lu Y, Ma X, et al. Pde-net: Learning pdes from data
International Conference on Machine Learning. PMLR, 2018: 3208-3216.

Open Problems: Nonlinear-Operator-Learning

Standard non-parametric rate: $n^{-\frac{2s}{\boxed{d}+2s}}$

"dimension"

$d = \infty$

the k-nearest-neighbour estimator (Kudraszow & Vieu, 2013). The development of functional nonparametric regression has been hindered by a theoretical barrier, which is formulated in Mas (2012) and linked to the small ball probability problem (Delaigle & Hall, 2010). Essentially, in a rather general setting, the minimax rate of nonparametric regression on a generic functional space is slower than any polynomial of the sample size, which differs markedly from the polynomial minimax rates for many functional parametric regression procedures, see, e.g., Hall & Keilegom (2007), and Yuan & Cai (2010) for functional linear regression. These endeavours in functional nonparametric regression do not exploit the intrinsic structure that is common in practice. For instance, Chen & Müller (2012) suggested that functional data often have a low-dimensional manifold structure which can be utilized for more efficient representation. In this article, we exploit the nonlinear low-dimensional structure for functional nonparametric regression.

Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness PDF

Sho Okumoto, Taiji Suzuki

28 Sept 2021 (modified: 15 Mar 2022) ICLR 2022 Spotlight Readers: Everyone Show Bibtex Show Revisions

A Non-Parametric Statistical Framework

$\Delta u + u = f$

Output An estimation of u

"Learning with gradient information"

i.i.d samples

Input Random samples $\{(x_i, f(x_i) + \text{noise})\}_{i=1}^n$

Aim The best estimator

Evaluation in Sobolev norm

$$\inf_H \max_{f \in H^\alpha} \mathbb{E}_{\{(x_i, f(x_i)+\text{noise})\}_{i=1}^n} \| H(\{(x_i, f(x_i)+\text{noise})\}_{i=1}^n) - u \|_{H^\beta}$$

Uniformly good on all Sobolev functions   Estimator

A Non-Parametric Statistical Framework

Theorem (informal)

Minimax lower bound for t-order linear elliptic PDE:

Evaluation in Sobolev norm

$$\inf_{H} \max_{f \in H^\alpha} \mathbb{E}_{\{(x_i, f(x_i)+\text{noise})\}_{i=1}^n} \|H(\{(x_i, f(x_i)) + \text{noise}\}_{i=1}^n) - u\|_{H^\beta} \gtrsim n^{-\frac{(\alpha-\beta)}{d+2\alpha-2t}}$$

Order of the PDE

Very similar to nonparametric rate $n^{-\frac{\alpha}{d+2\alpha}}$

A Non-Parametric Statistical Framework

<u>Theorem (informal)</u>

Minimax lower bound for t-order linear elliptic PDE:

Evaluation in Sobolev norm

$$\inf_{H} \max_{f \in H^\alpha} \mathbb{E}_{\{(x_i, f(x_i)+\text{noise})\}_{i=1}^n} \| H(\{(x_i, f(x_i)) + \text{noise}\}_{i=1}^n) - u \|_{H^\beta} \gtrsim n^{-\frac{(\alpha-\beta)}{d+2\alpha-2t}}$$

Order of the PDE

<u>Empirical process/fast rate generalization bound</u>

Is PINN and DRM statistical optimal?

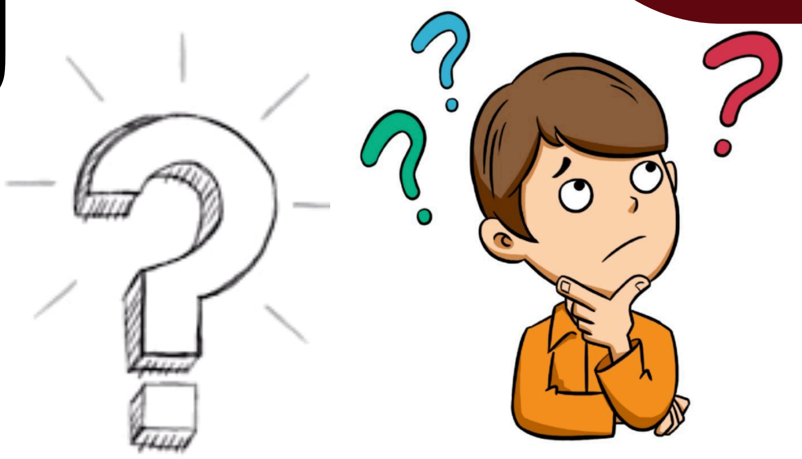
Artifact of analysis?
NN ansatz? Objective?

For $\beta = 2$

PINN

For $\beta = 1$

DRM



Is Deep Ritz Optimal? A Fourier View

$\boxed{Au = f}$

Solving $\Delta u + u = f$ from random samples $\{(x_i, f(x_i) + \text{noise})\}_{i=1}^{n}$

Why not first learn f then learn u

Naive way to do this?

<u>Naive Estimator</u> $\hat{f} = \sum_{|z|<S} \hat{f}_z^F \phi_z$ where $\hat{f}_z^F = \sum f(x_i)\phi_z(x_i)$

Then $u = A^{-1}f = \sum_{|z|<S} \frac{1}{|z|^2 + 1} \hat{f}_z^F \phi_z$ <u>Fourier Basis</u>

<u>Naive Estimator</u> is Optimal with proper selection of S

Is Deep Ritz Optimal? A Fourier View

$\boxed{Au = f}$

Solving $\Delta u + u = f$ from random samples $\{(x_i, f(x_i) + \text{noise})\}_{i=1}^n$

Why not first learn f then learn u

Naive Estimator $\hat f = \sum_{|z|<S} \hat f_z^F \phi_z$ where $\hat f_z^F = \sum f(x_i)\phi_z(x_i)$

Then $u = A^{-1} f = \sum_{|z|<S} \frac{1}{|z|^2 + 1} \hat f_z^F \phi_z$

DRM Estimator $\hat u = \sum_{|z|<S} \hat u_z^F \phi_z$ and plug in

$$\hat u^F = \arg\min_{\hat u^F} \int \frac{1}{2} \left| \sum_{|z|<S} \hat u_z^F (\nabla \phi_z + \phi_z) \right|^2 - \sum_{|z|<S} \hat u_z^F \hat f_z^F$$

How is naive estimator different from DRM?

Is Deep Ritz Optimal? A Fourier View

$\boxed{Au = f}$

Solving $\Delta u + u = f$ from random samples $\{(x_i, f(x_i) + \text{noise})\}_{i=1}^n$

Why not first learn f then learn u

Naive Estimator $\hat{f} = \sum_{|z|<S} \hat{f}_z^F \phi_z$ where $\hat{f}_z^F = \sum f(x_i)\phi_z(x_i)$

Then $u = A^{-1}f = \sum_{|z|<S} \frac{1}{|z|^2+1} \hat{f}_z^F \phi_z$

Naive

$$\boxed{\hat{u}_z^F = \frac{\hat{f}_z^F}{|z|^2+1}}$$

DRM

$$\boxed{\hat{u}_z^F = (\hat{A})^{-1} \hat{f}_z^F}$$

$$\hat{A} = \left(\sum_i \nabla \phi_j(x_i) \nabla \phi_k(x_i)\right)_{j,k} + \left(\sum_i \phi_j(x_i)\phi_k(x_i)\right)_{j,k}$$

Introduce further variance

DRM Estimator $\hat{u} = \sum_{|z|<S} \hat{u}_z^F \phi_z$ and plug in

$$\hat{u}^F = \arg\min_{\hat{u}^F} \int \frac{1}{2} \left| \sum_{|z|<S} \hat{u}_z^F (\nabla \phi_z + \phi_z) \right|^2 - \sum_{|z|<S} \hat{u}_z^F \hat{f}_z^F$$

Is Deep Ritz Optimal? A Fourier View

$\boxed{Au = f}$

Solving $\Delta u + u = f$ from random samples $\{(x_i, f(x_i) + \text{noise})\}_{i=1}^n$

Why not first learn f then learn u

Naive Estimator $\hat{f} = \sum_{|z|<S} \hat{f}_z^F \phi_z$ where $\hat{f}_z^F = \sum f(x_i)\phi_z(x_i)$

Then $u = A^{-1}f = \sum_{|z|<S} \frac{1}{|z|^2+1} \hat{f}_z^F \phi_z$

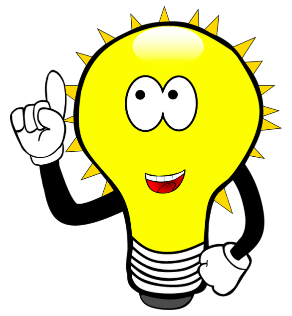
DRM Estimator $\hat{u} = \sum_{|z|<S} \hat{u}_z^F \phi_z$ and plug in

$$\hat{u}^F = \arg\min_{\hat{u}^F} \int \frac{1}{2} \left| \sum_{|z|<S} \hat{u}_z^F (\nabla \phi_z + \phi_z) \right|^2 - \sum_{|z|<S} \hat{u}_z^F \hat{f}_z^F$$

DRM discretized

$\nabla \cdot \nabla$

But not Δ



Integration by parts increase the monte-carlo variance.

Results in One Table...

Boundary condition?

Upper Bounds			Lower Bound
Objective Function	Neural Network	Fourier Basis	
Deep Ritz	$n^{-\frac{2s-2}{d+2s-2}} \log n$	$n^{-\frac{2s-2}{d+2s-2}}$	$n^{-\frac{2s-2}{d+2s-4}}$
Modified Deep Ritz	$n^{-\frac{2s-2}{d+2s-2}} \log n$	$n^{-\frac{2s-2}{d+2s-4}}$	$n^{-\frac{2s-2}{d+2s-4}}$
PINN	$n^{-\frac{2s-4}{d+2s-4}} \log n$	$n^{-\frac{2s-4}{d+2s-4}}$	$n^{-\frac{2s-4}{d+2s-4}}$

Still open

For $\beta = 2$

PINN ✓

For $\beta = 1$

DRM

	DRM	Modified
Spectral	✗	✓
NN	✗	?

DRM or PINN

Which one optimizes faster?

DRM $\min \int |\nabla u|^2 - 2uf$

PINN $\min ||\Delta u - f||^2$

Pre-ml Experience：

Double the condition number

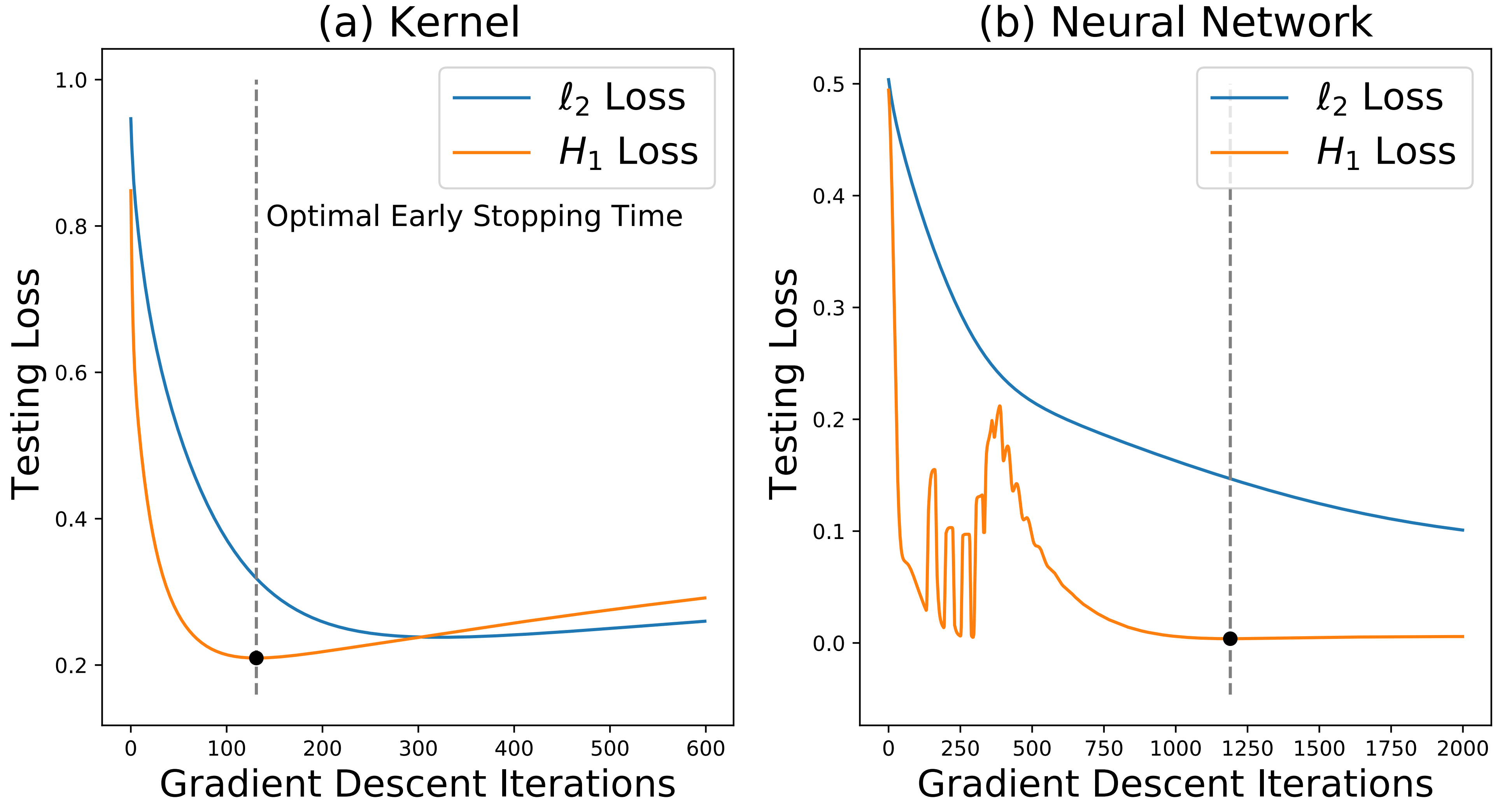
DRM or PINN

Which one optimizes faster?

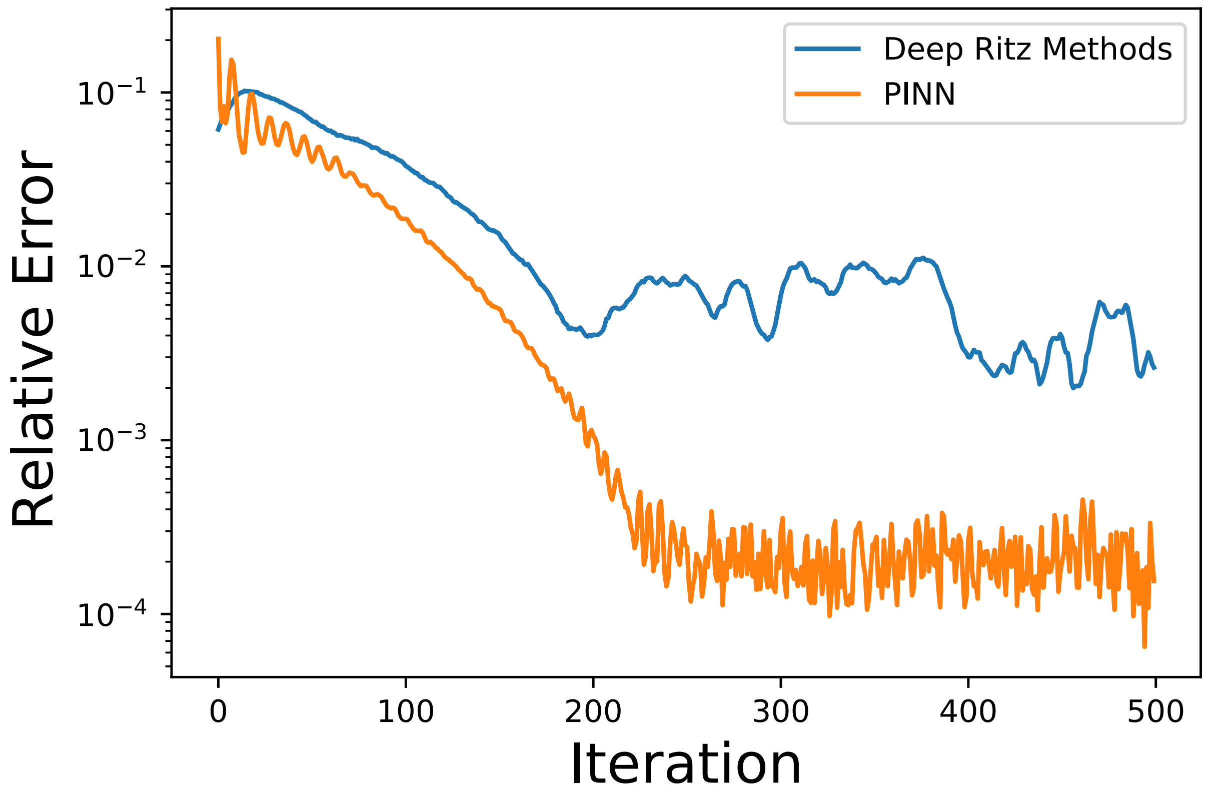
DRM $\min \int |\nabla u|^2 - 2uf$

PINN $\min \|\Delta u - f\|^2$

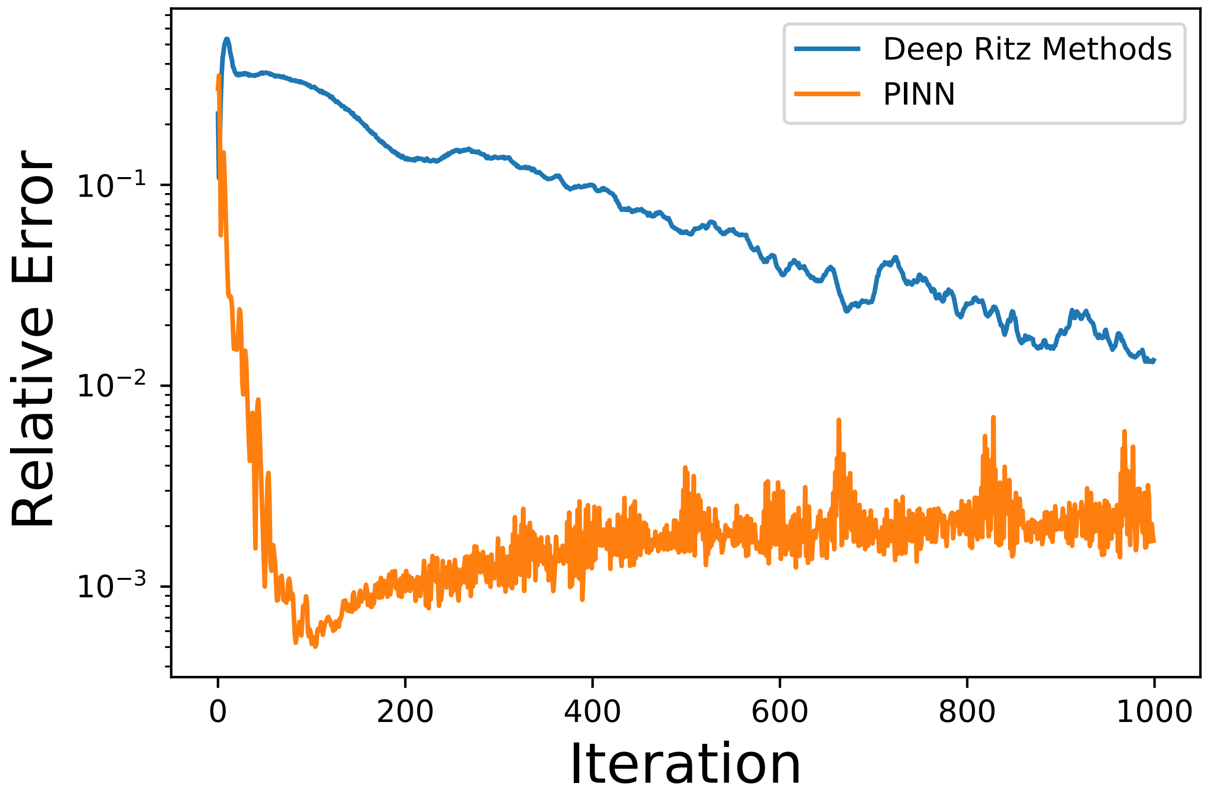
Pre-ml Experience：
Double the condition number



Sobolev Training



$f = \sin(2\pi x)$



$f = \sin(4\pi x)$

Solving $\Delta u = f$

A Kernelized Model

Machine learning is a kernelized dynamic.
Differential Operator can cancel Kernel Integral Op

Let's consider $\Delta u = f$ via minimizing $\boxed{\frac{1}{2}\langle f, \mathcal{A}_1 f\rangle - \langle u, \mathcal{A}_2 f\rangle}$

$$f = \langle \theta, K_x \rangle$$

- **Deep Ritz Methods**. $\mathcal{A}_1 = \Delta, \mathcal{A}_2 = Id$
- **PINN**. $\mathcal{A}_1 = \Delta^2, \mathcal{A}_2 = \Delta$

Gradient Descent

$$d\theta_t = \sum_i \Big\langle \theta, \underbrace{\mathcal{A}_1}_{\text{Differential operator}} \underbrace{K_{x_i i} \Big\rangle K_{x_i}}_{\text{Kernel integral operator}} - f_i \mathcal{A}_2 K_{x_i}$$

Our Result

I understand your idea, but what's your thm?

Theorem (Informal)

1. The information theoretical lower bound in the kernel space matches the lower bound for learning PDE.

2. Gradient Descent with proper early stopping time selection can achieve optimal statistical rate

3. The proper early stopping time is smaller for PINN than DRM