

# On the Power and Limit of Scientific Machine Learning

---

CBMS conference on deep learning and numerical PDEs

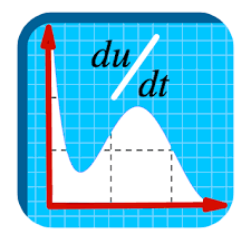
## Joint work with

Jose Blanchet, Jiajin Li, Jikai Jin, Haoxuan Chen, Lexing Ying...

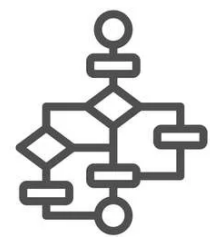
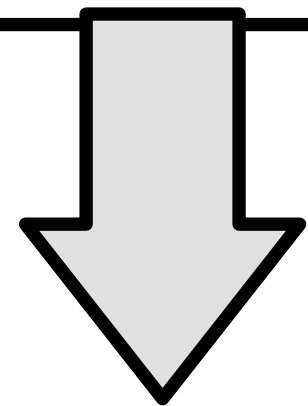
Yiping Lu [yplu@stanford.edu](mailto:yplu@stanford.edu)  
<https://2prime.github.io/>

# Two Disciplines in Science

## Structural Model



Differential equation modeling



Solving using numerical algorithms



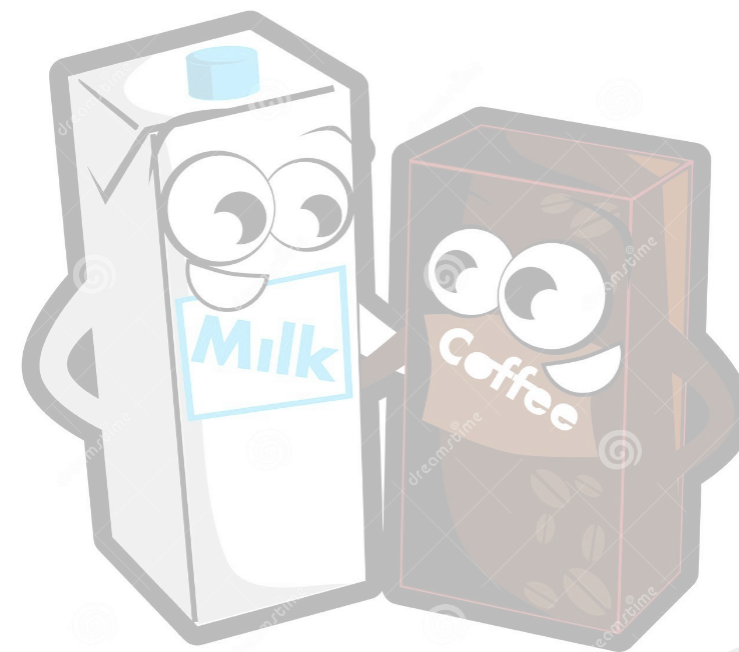
**Transparent**



**Lots of approximations  
Limits the power**

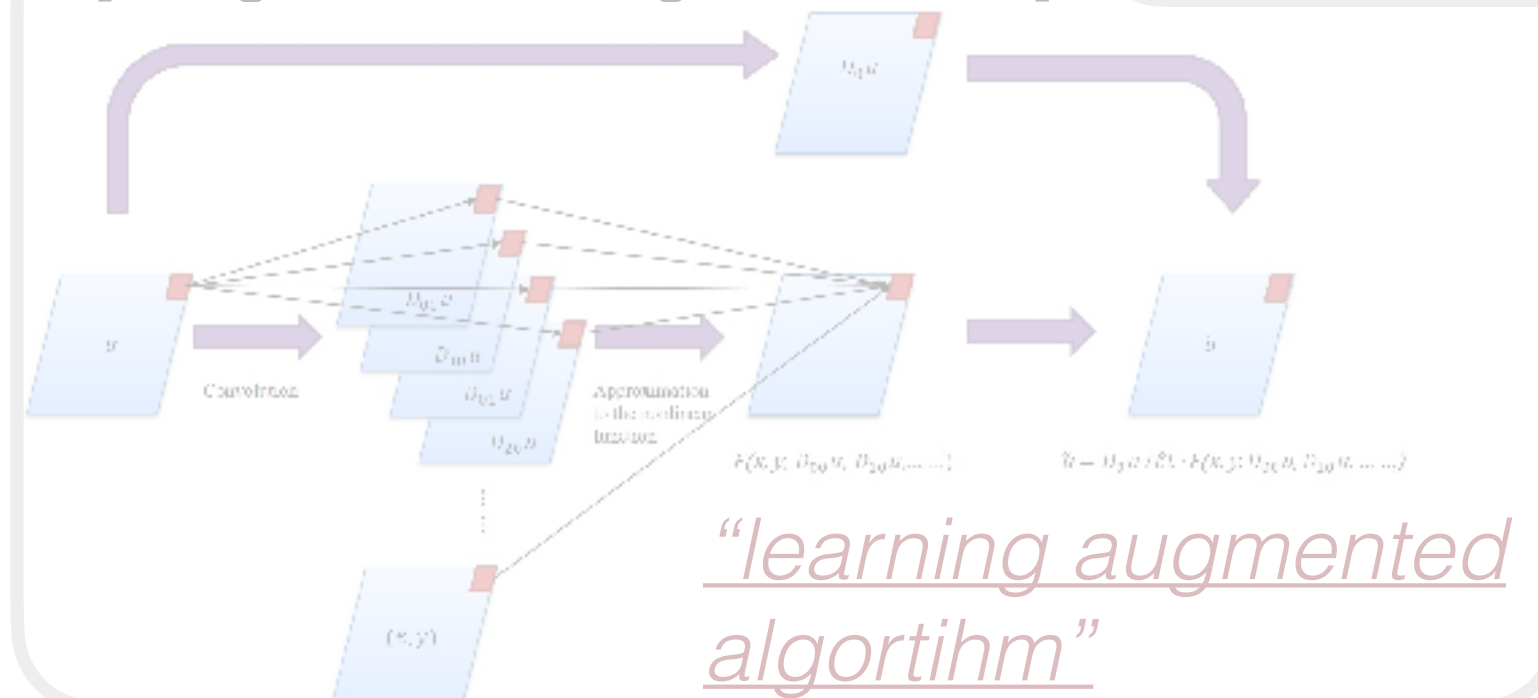


Make Useful Prediction

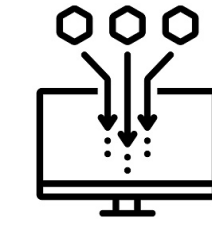


[Long-Lu-Ma-Dong ICML2018]

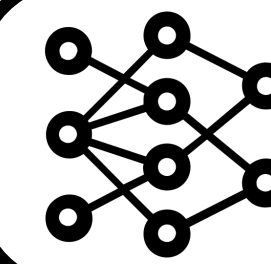
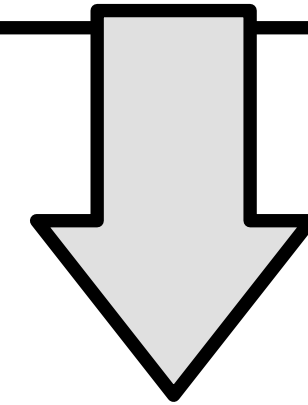
PDE-Net



## Machine Learning



Data Collecting



Machine Learning



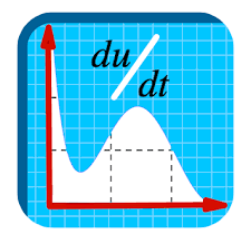
**Flexible, Accurate**



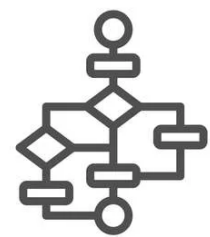
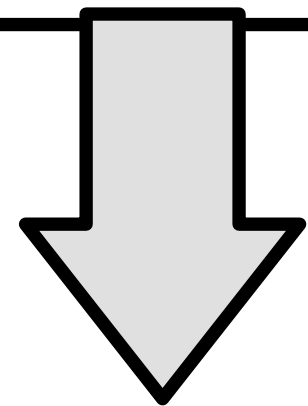
**Blackbox  
Data intensive**

# Two Disciplines in Science

## Structural Model



Differential equation modeling



Solving using numerical algorithms



**Transparent**



**Lots of approximations  
Limits the power**

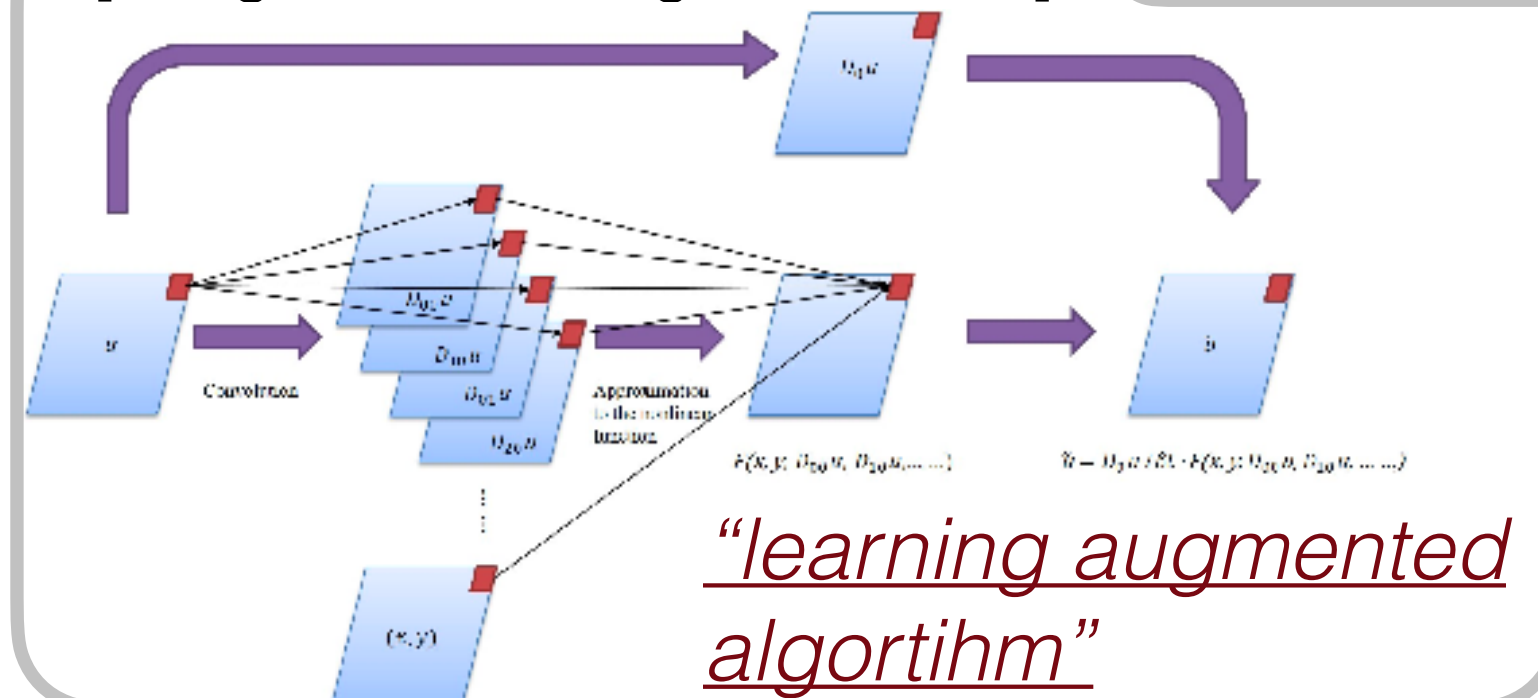


Make Useful Prediction

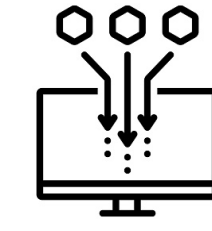


[Long-Lu-Ma-Dong ICML2018]

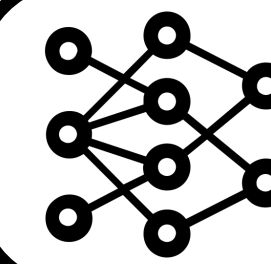
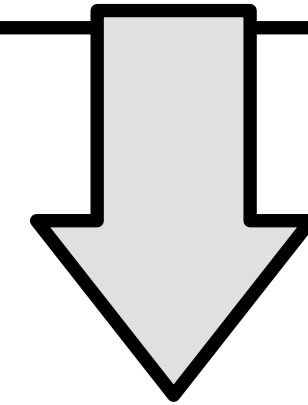
PDE-Net



## Machine Learning



Data Collecting



Machine Learning



**Flexible, Accurate**



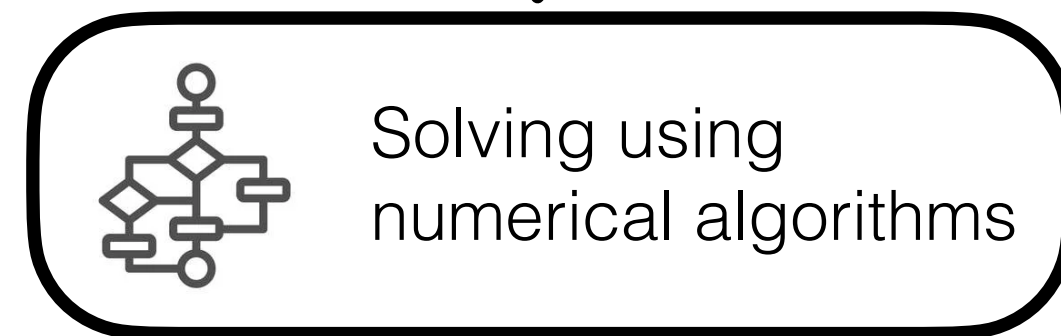
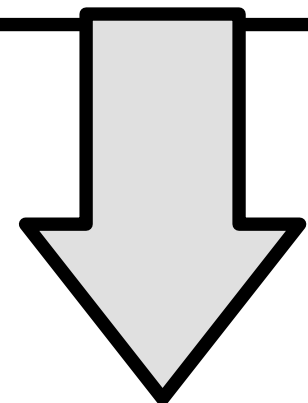
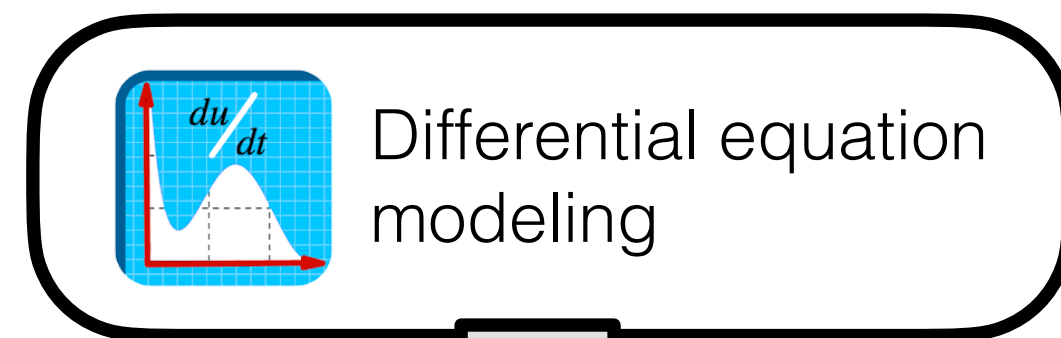
**Blackbox  
Data intensive**



# Two Disciplines in Science

Can we understand it theoretically?

## Structural Model

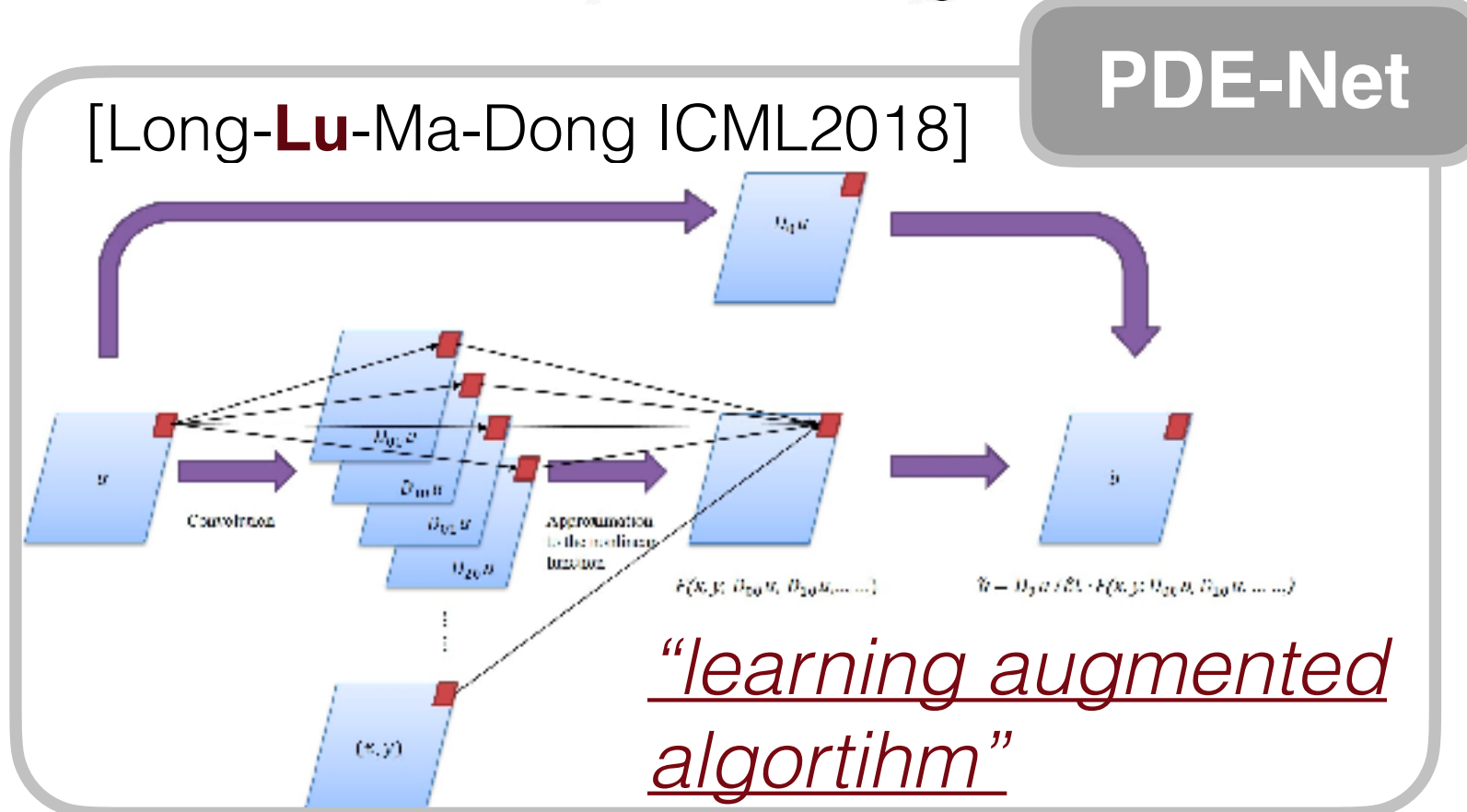


😊 **Transparent**

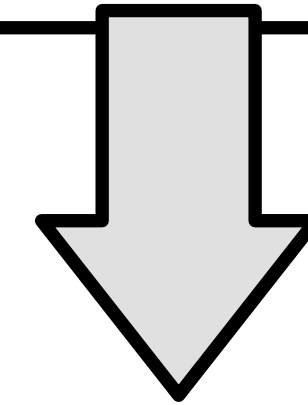
😞 **Lots of approximations  
Limits the power**



Just ML + physic data?



## Machine Learning



😊 **Flexible, Accurate**

😞 **Blackbox  
Data intensive**





# Machine Learning Research

**Aim:** fit function  $(x_i, y_i = f(x_i)), i = 1, 2, \dots, n$

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 selection 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



# Machine Learning Research

**Aim:** fit function  $(x_i, y_i = f(x_i)), i = 1, 2, \dots, n$

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 selection of  $f$ ”

Best Estimator

**Step 0** Specify your task!



What is the task of scientific machine learning?

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



# Not Just Differential Equation models

---

“Physic” Model


$$Au = f$$





# Not Just Differential Equation models

“Physic” Model

$$Au = f$$

Hamilton Jacobi Equation

Value Function

Reward Function

Kolmogorov 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]

Recover parameter  $\theta$  in model  $A_\theta$   
*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]....

[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]...



# Machine Learning Research

## Scientific

**Aim:** fit function  $(x_i, y_i = f(x_i)), i = 1, 2, \dots, n$

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

**Physical Equation**

$$Au = f$$

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

Recover parameter  $\theta$  in Model  $A_\theta$

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

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



# Machine Learning Research

## Scientific

**Aim:** fit function  $(x_i, y_i = f(x_i)), i = 1, 2, \dots, n$

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

Standard approximation and statistical exercises?

**Physical Equation**

$$Au = f$$

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

Recover parameter  $\theta$  in Model  $A_\theta$

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

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





# Machine Learning Research

## Scientific

**Aim:** fit function  $(x_i, y_i = f(x_i)), i = 1, 2, \dots, n$



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

**New insights for:**  
Operator learning  
Solving PDE  
Quadrature Rule

**Physical Equation**

$$Au = f$$

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

Recover parameter  $\theta$  in Model  $A_\theta$

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

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



# Machine Learning Research

## Scientific

**Aim:** fit function  $(x_i, y_i = f(x_i)), i = 1, 2, \dots, n$



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

**New insights for:**

- Operator learning
- Solving PDE
- Quadrature Rule

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  $\theta$  in Model  $A_\theta$

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

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



# Machine Learning Research

## Scientific

**Aim:** fit function  $(x_i, y_i = f(x_i)), i = 1, 2, \dots, n$



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

**New insights for:**

Operator learning

Solving PDE

Quadrature Rule

New technique for semi-parametric statistic via sobolev embedding

**Physical Equation**

$$Au = f$$

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

Recover parameter  $\theta$  in Model  $A_\theta$

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

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



# Optimal Quadrature Rule via ML

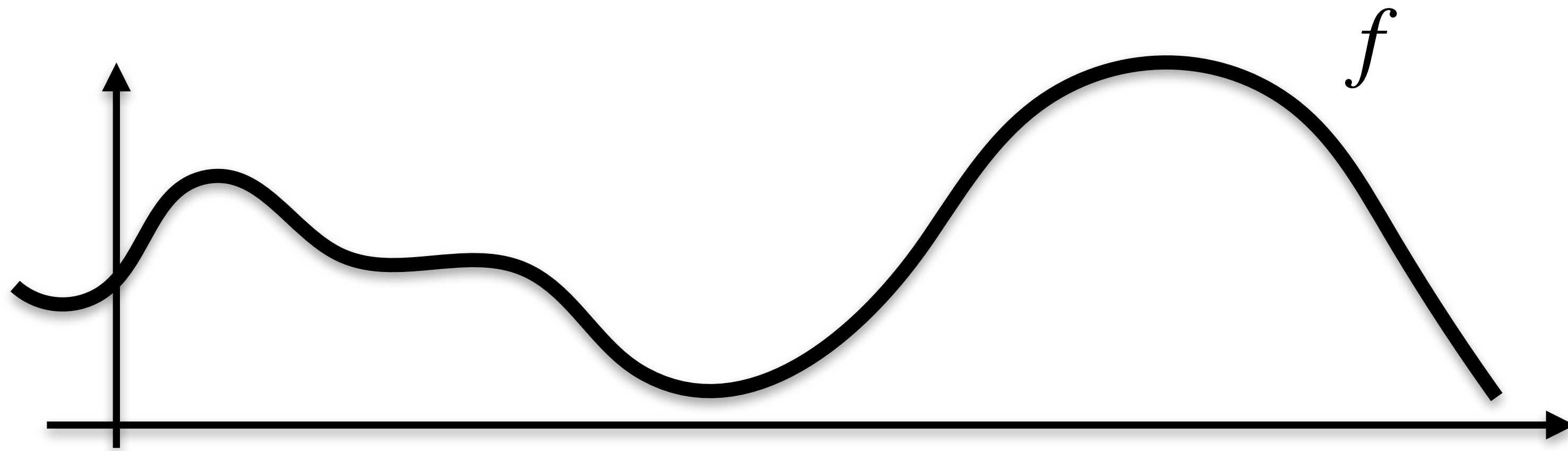
<https://arxiv.org/abs/2305.16527>



# Quadrature Rule

---

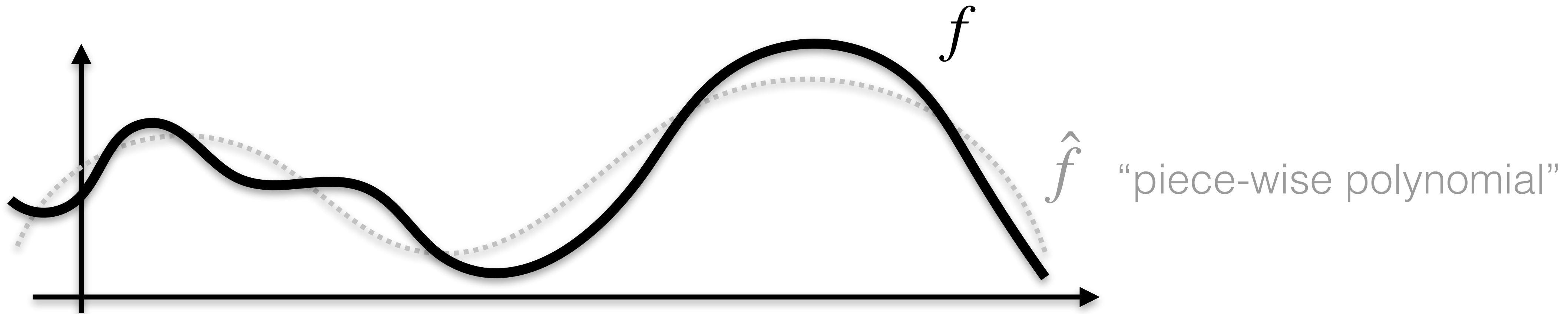
**Aim** Estimate  $\mathbb{E}_p f$



# Quadrature Rule

---

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

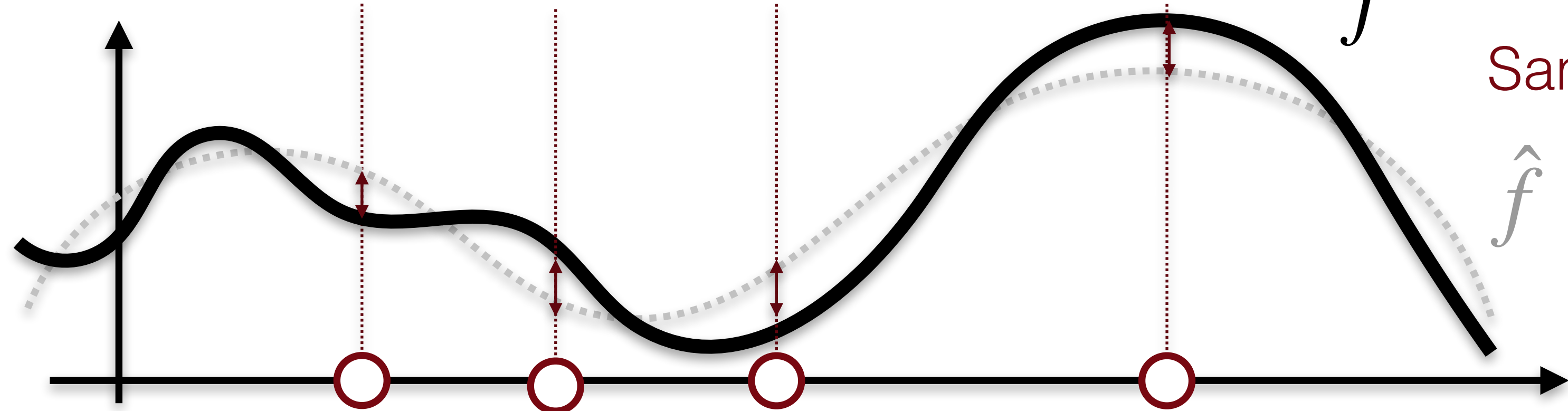


# Quadrature Rule

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



Debiasing  
"semi-"parametric



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



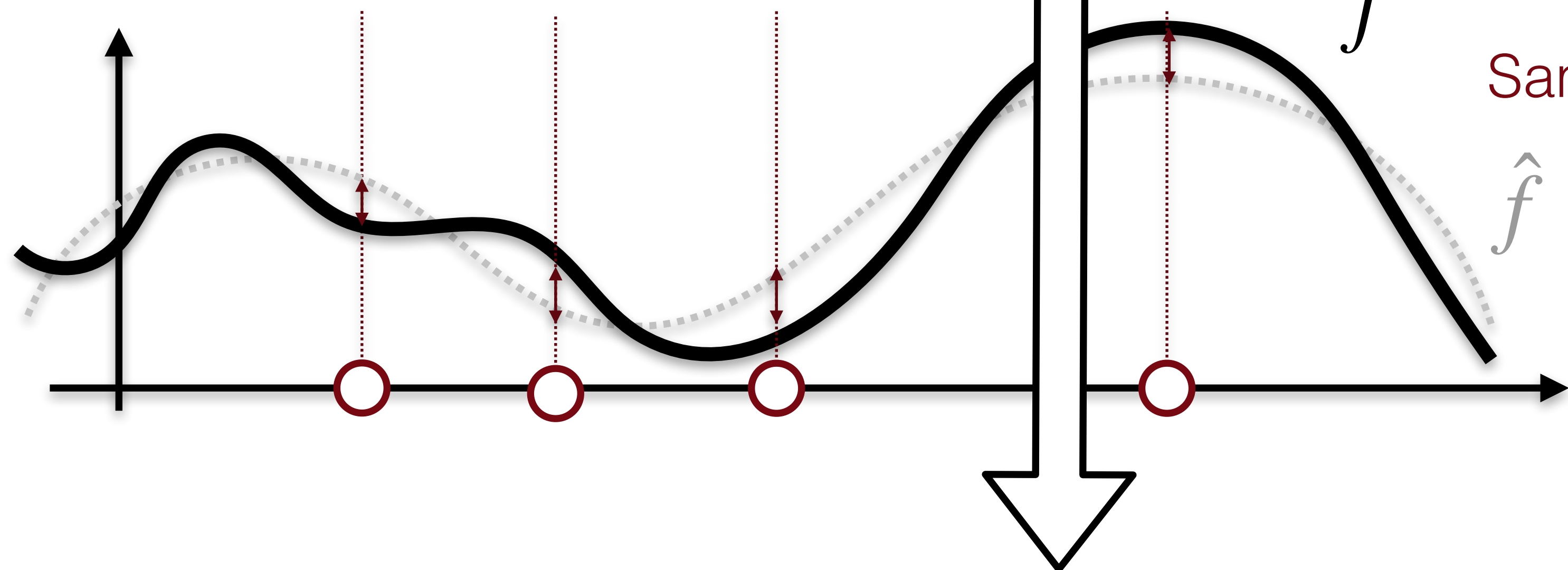
# Quadrature Rule

**Aim**

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



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 ....

“Quadrature” 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**

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

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  $\mathbb{E}_P f$   $\mathbb{E}_P f^q, f \in W^{s,p}$

**Step 1**

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

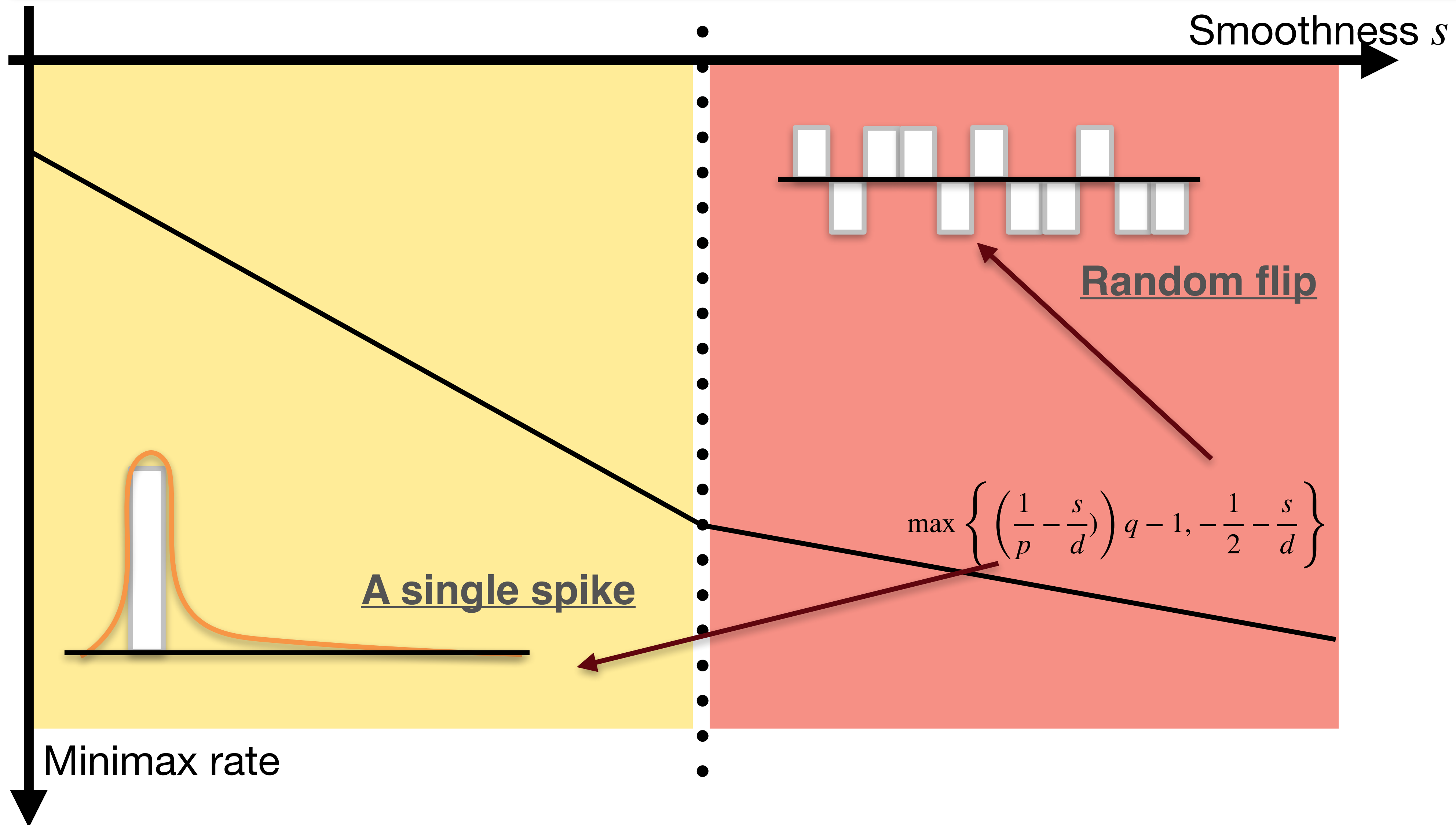
**Step 2**

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

Low order term

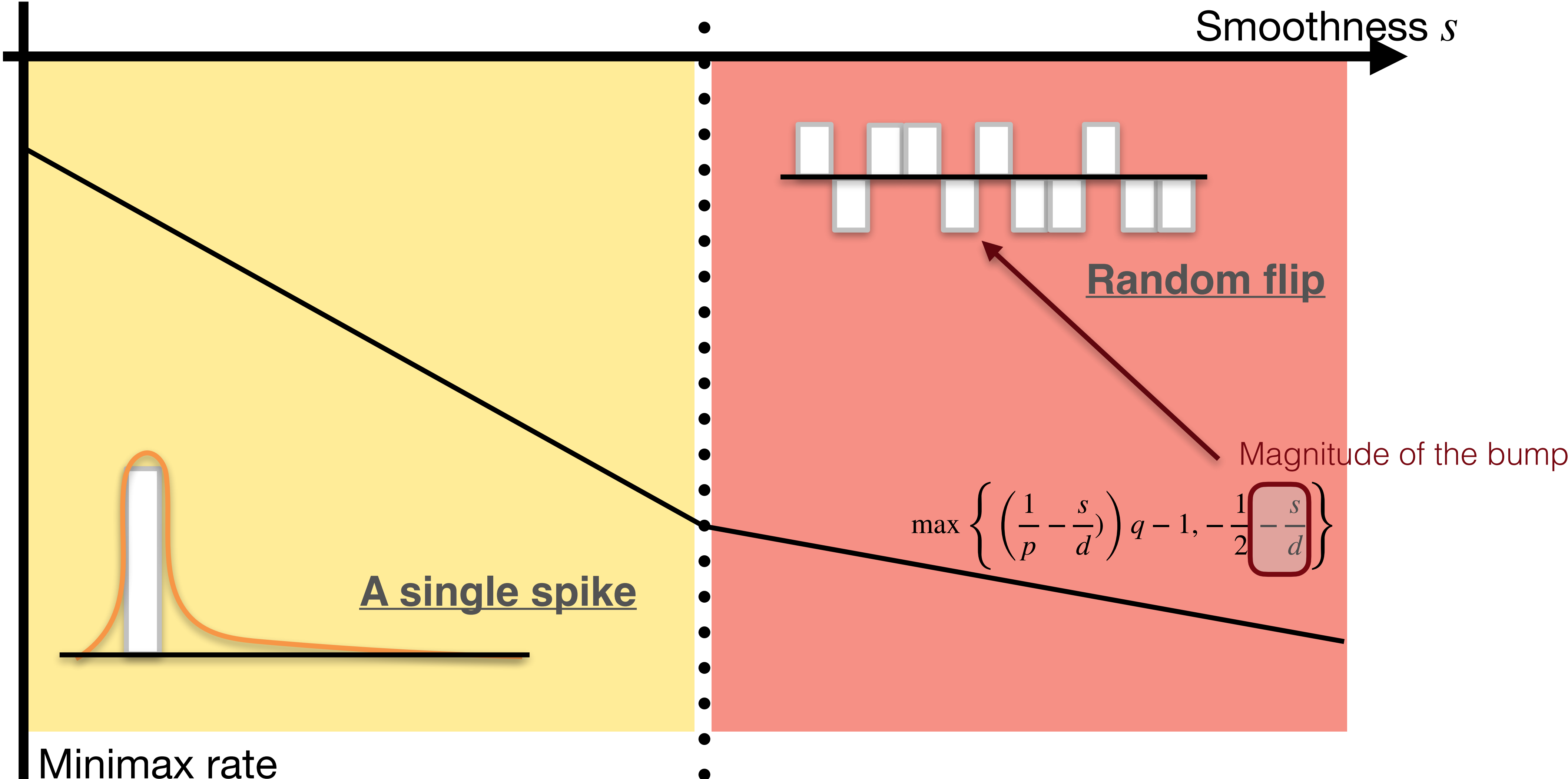


# Setting the information theoretical limit

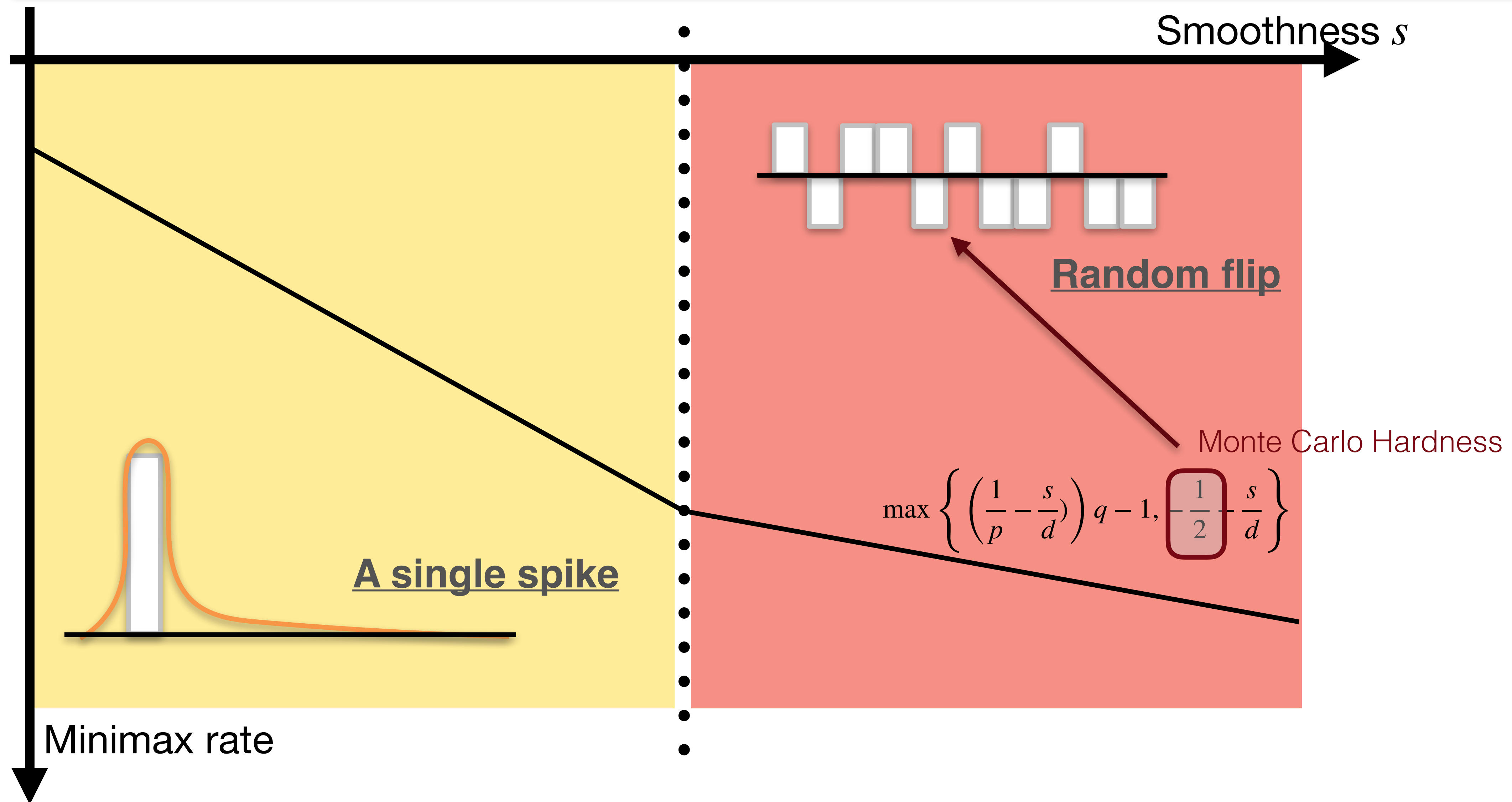




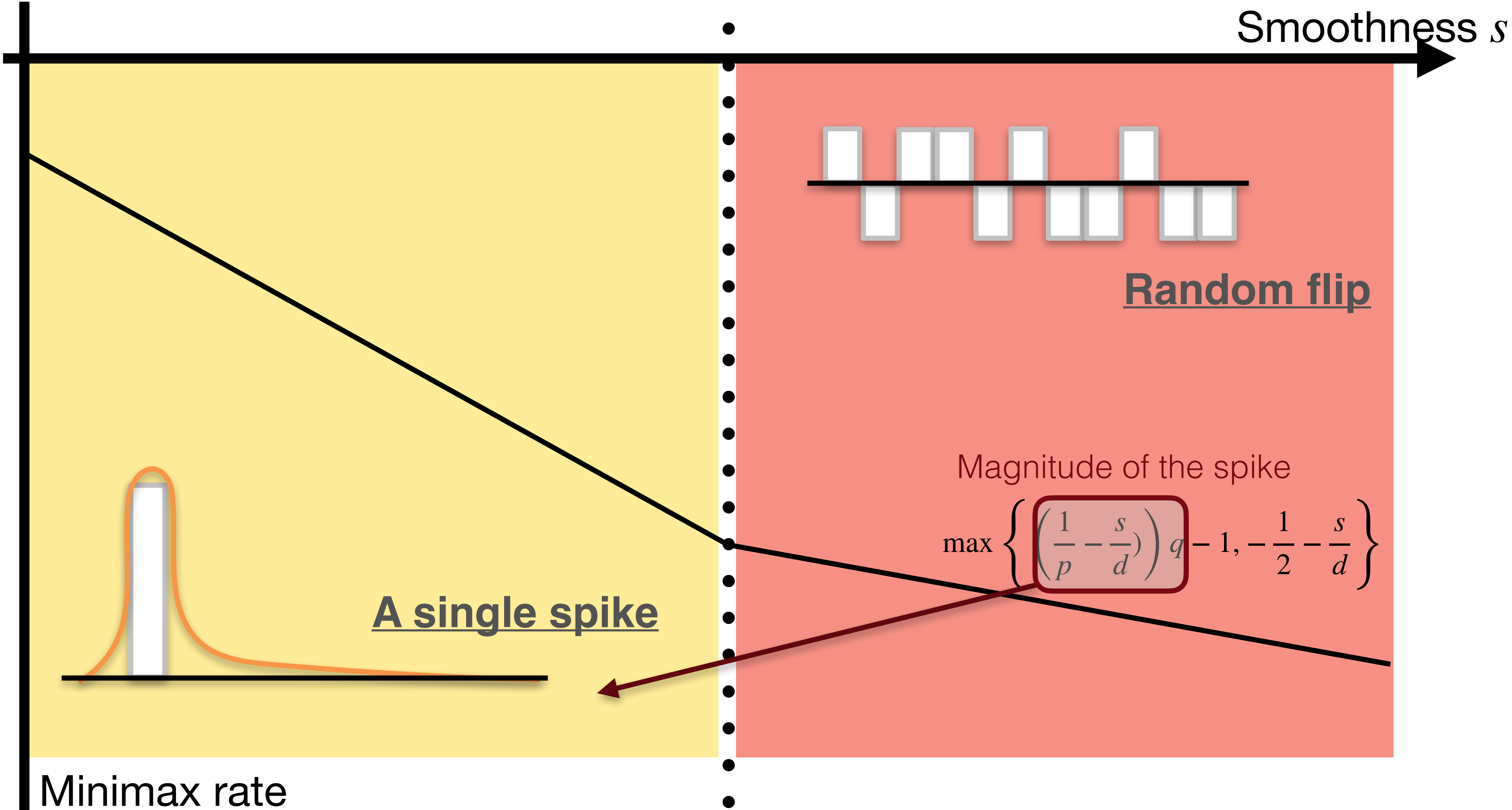
# Setting the information theoretical limit



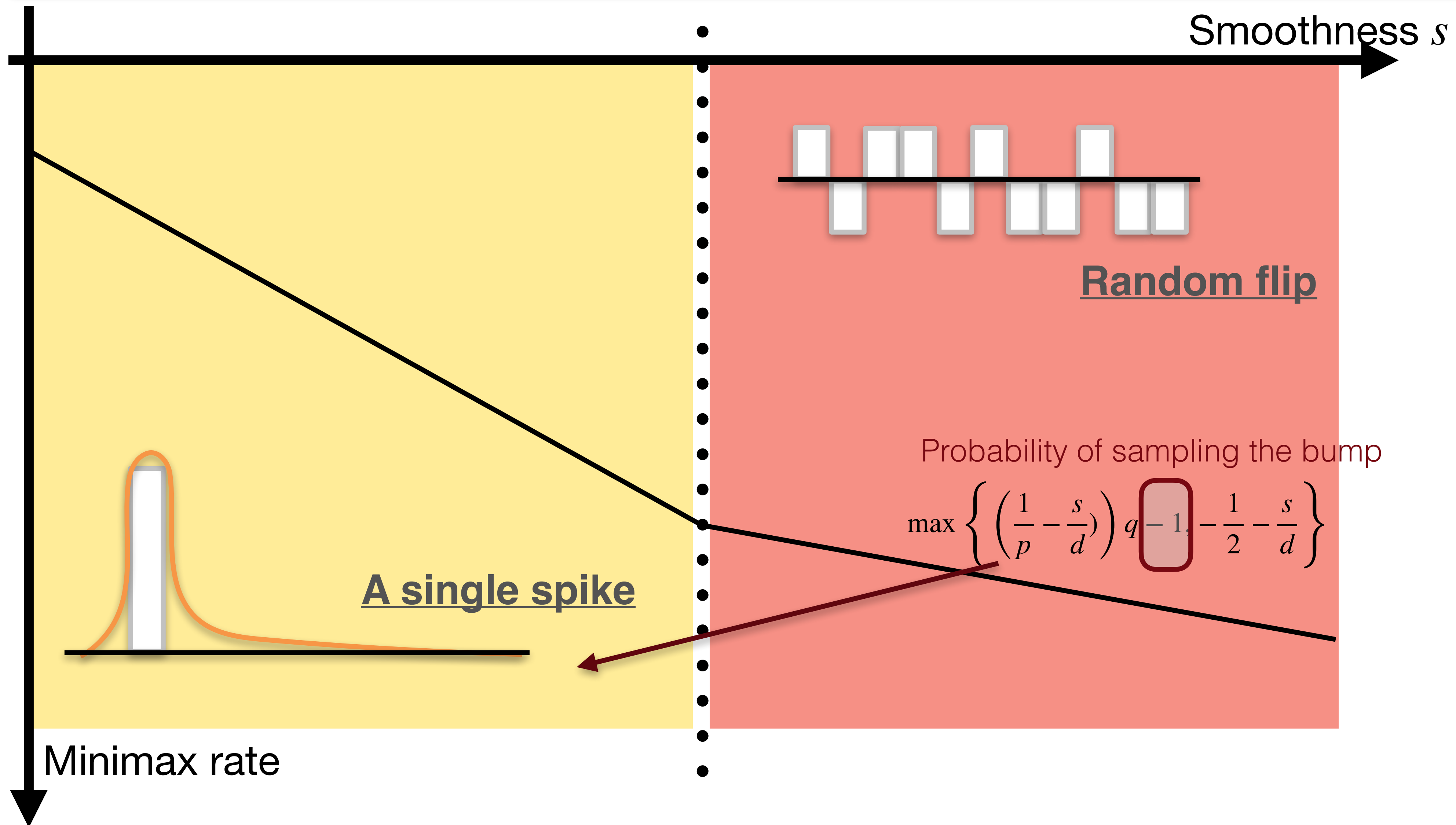
# Setting the information theoretical limit



# Setting the information theoretical limit



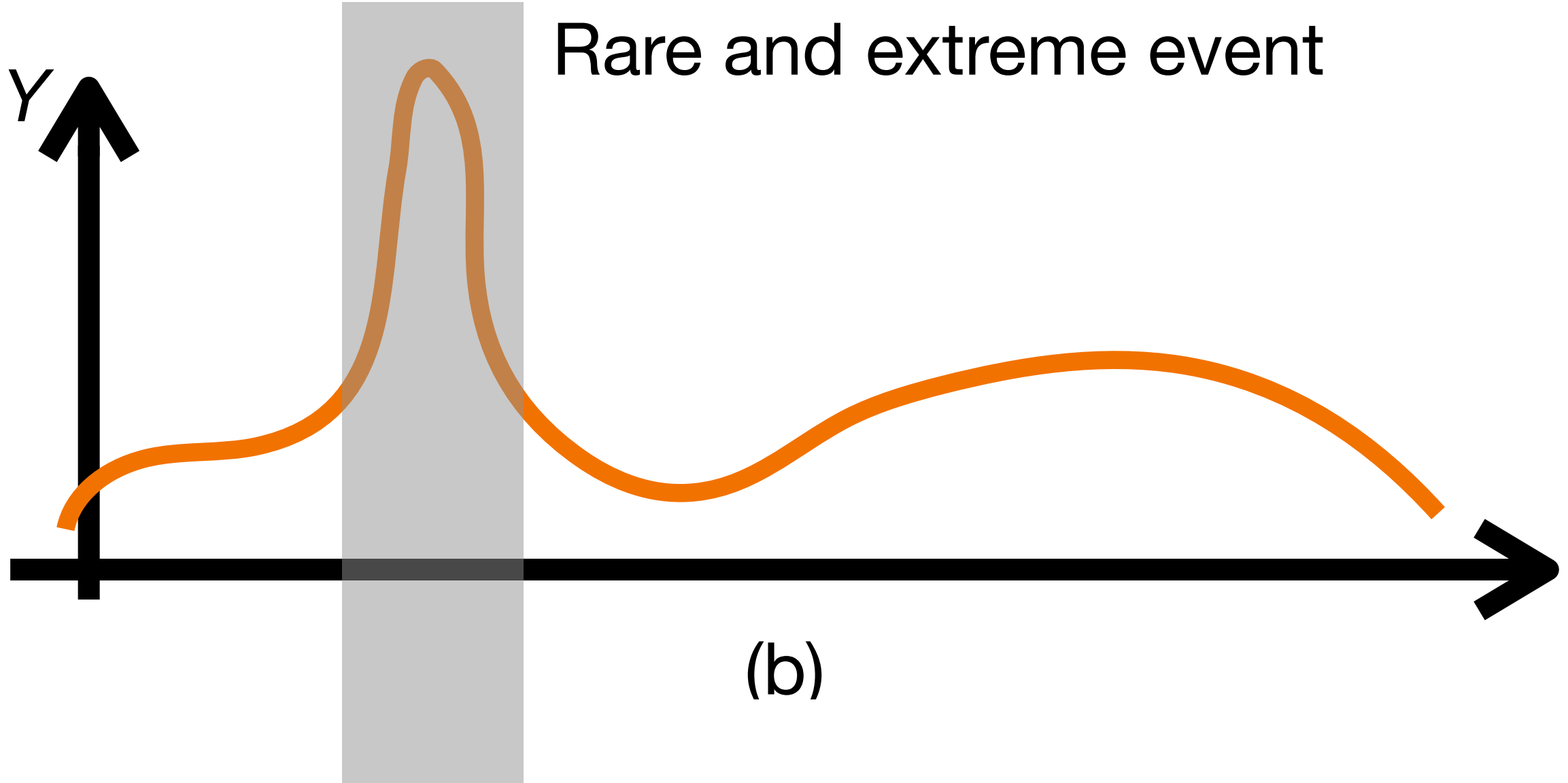
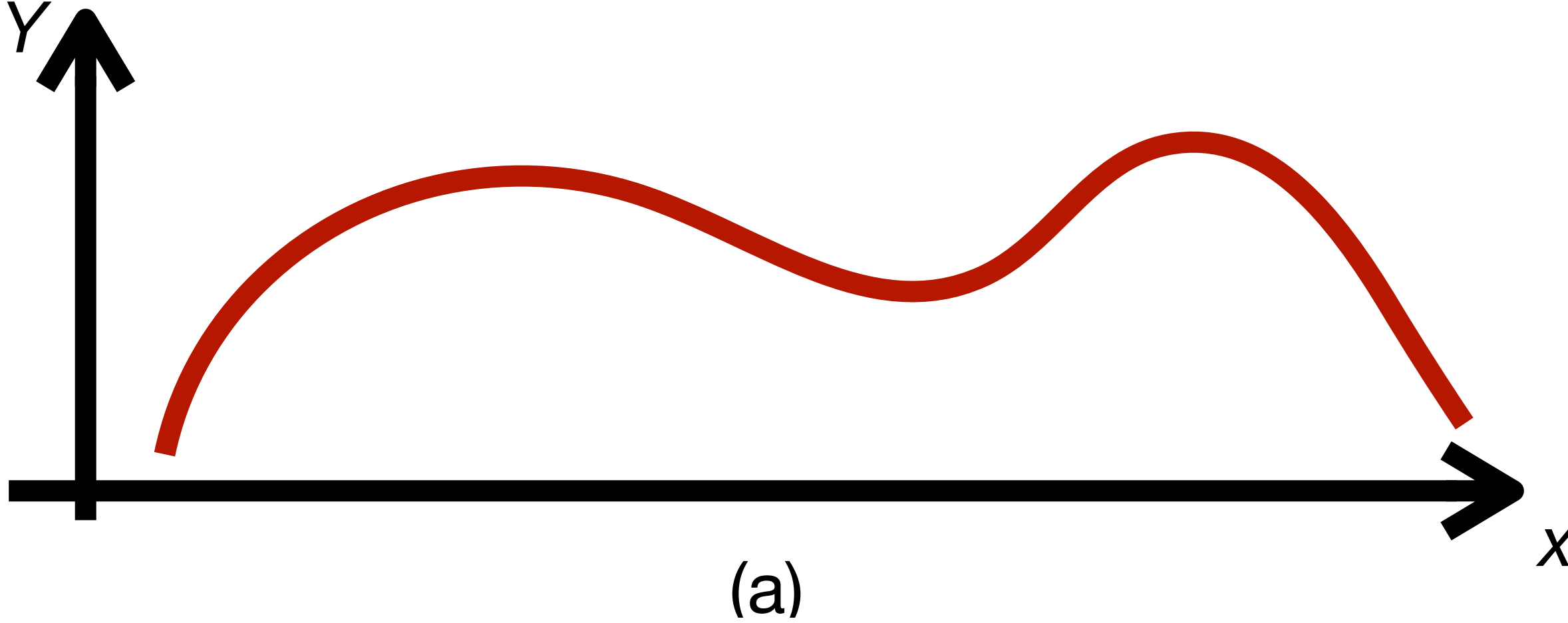
# Setting the information theoretical limit



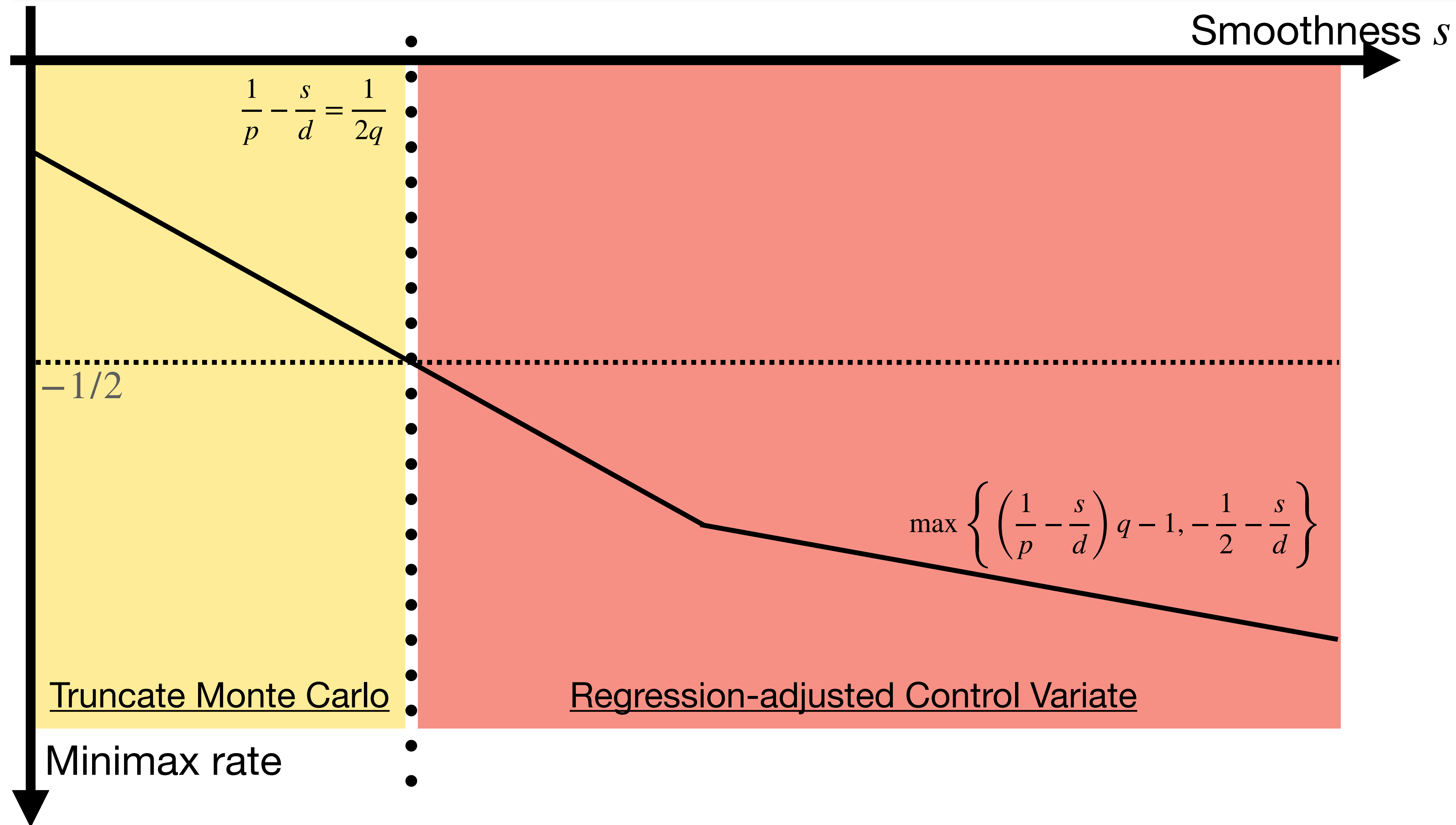


# Rare Event and Smoothness...

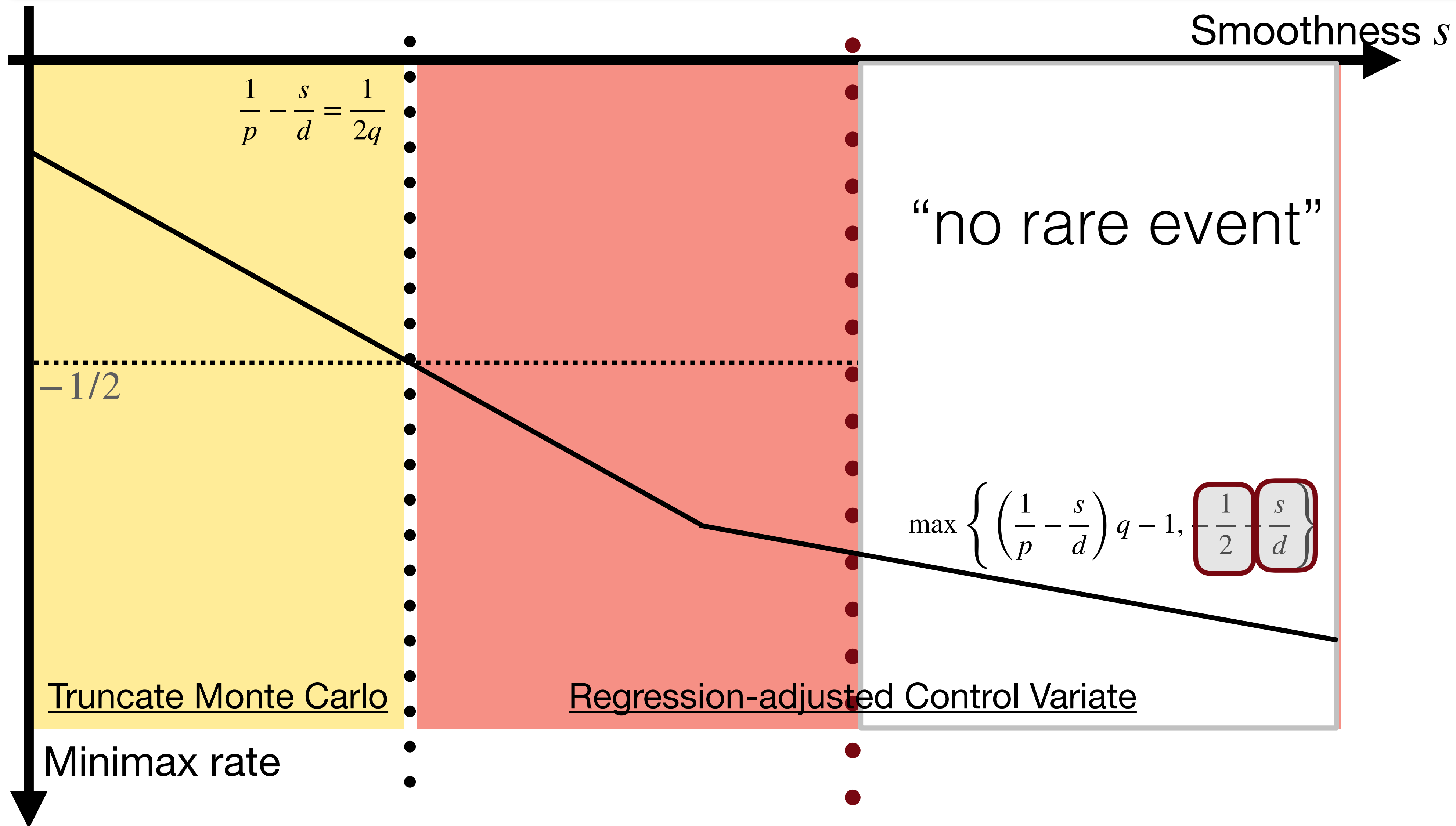
---



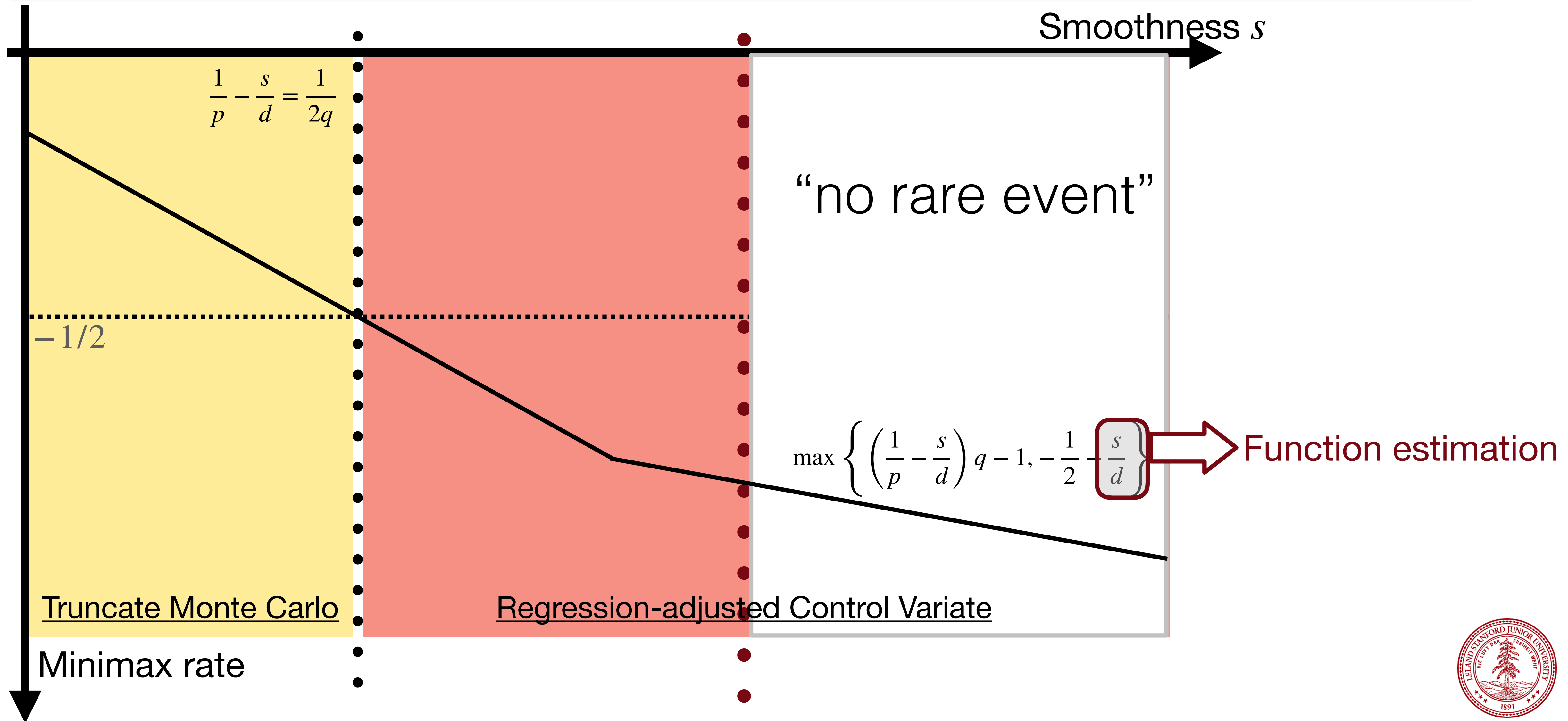
# When the control variate helps



# When the control variate helps

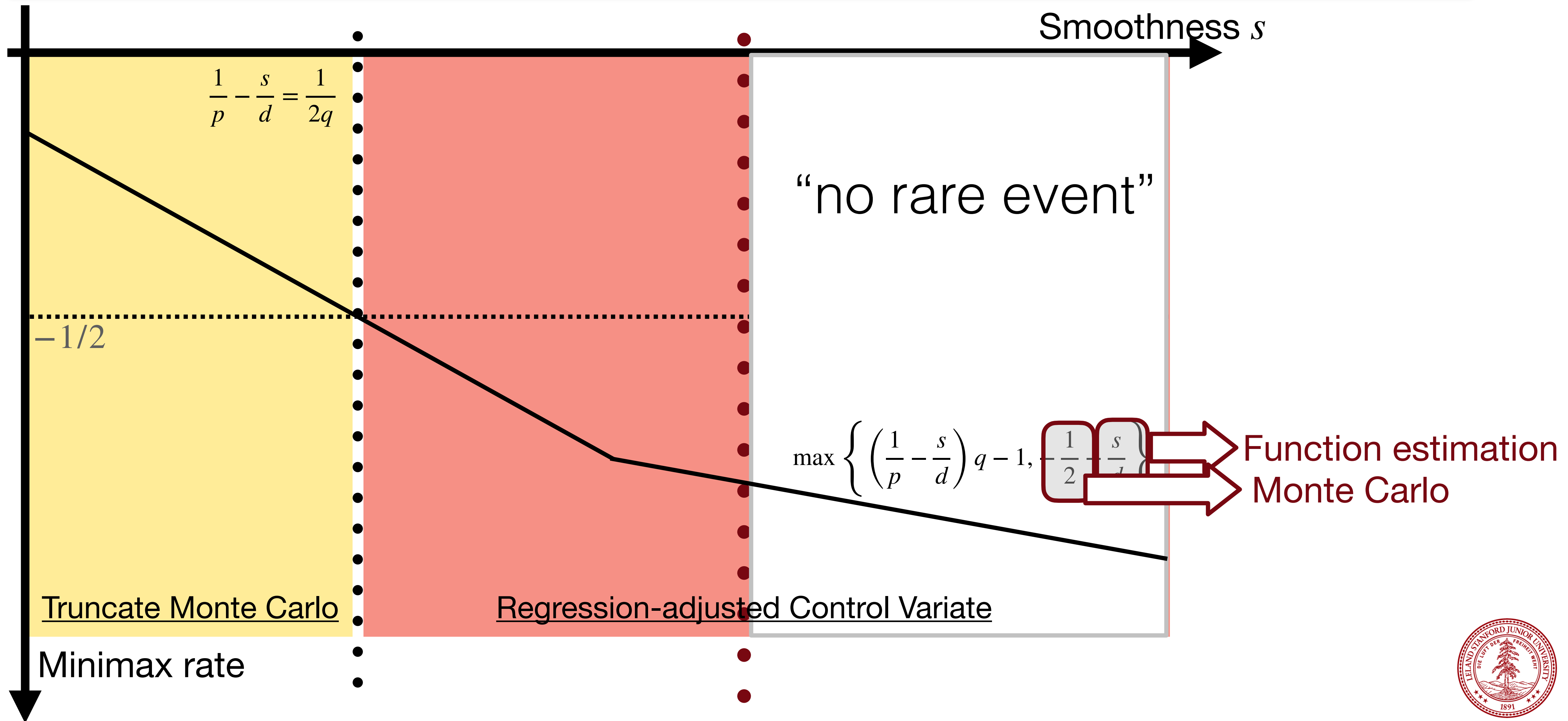


# When the control variate helps

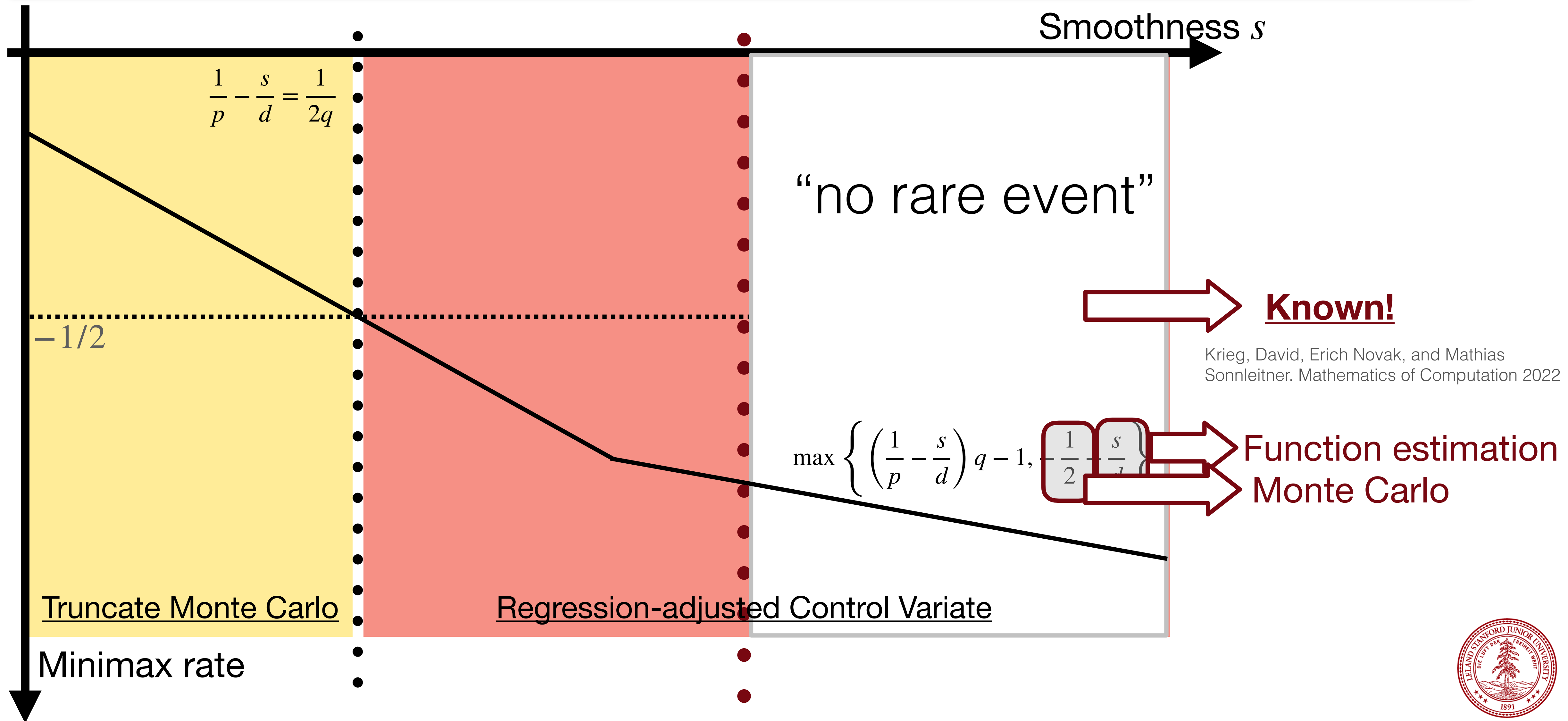




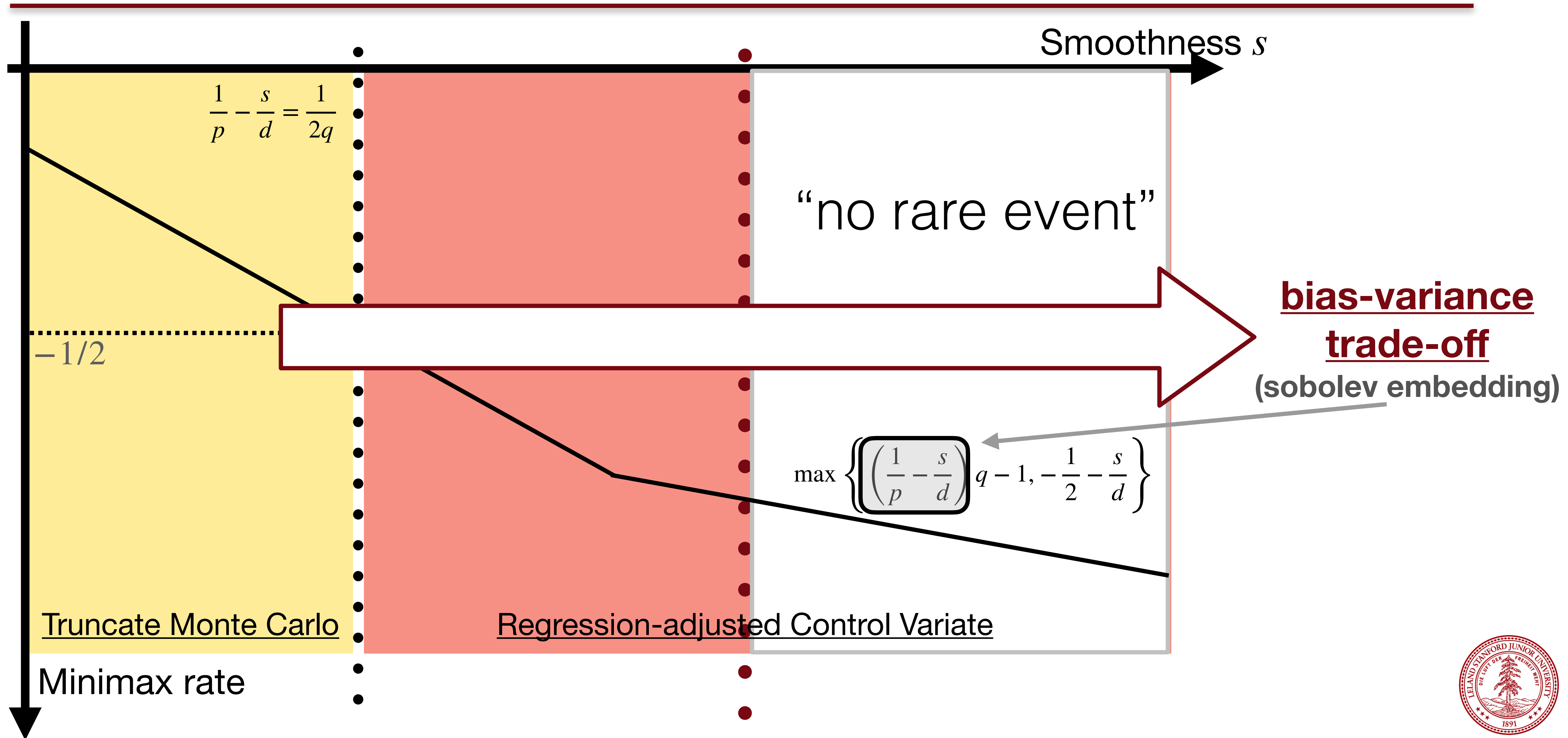
# When the control variate helps



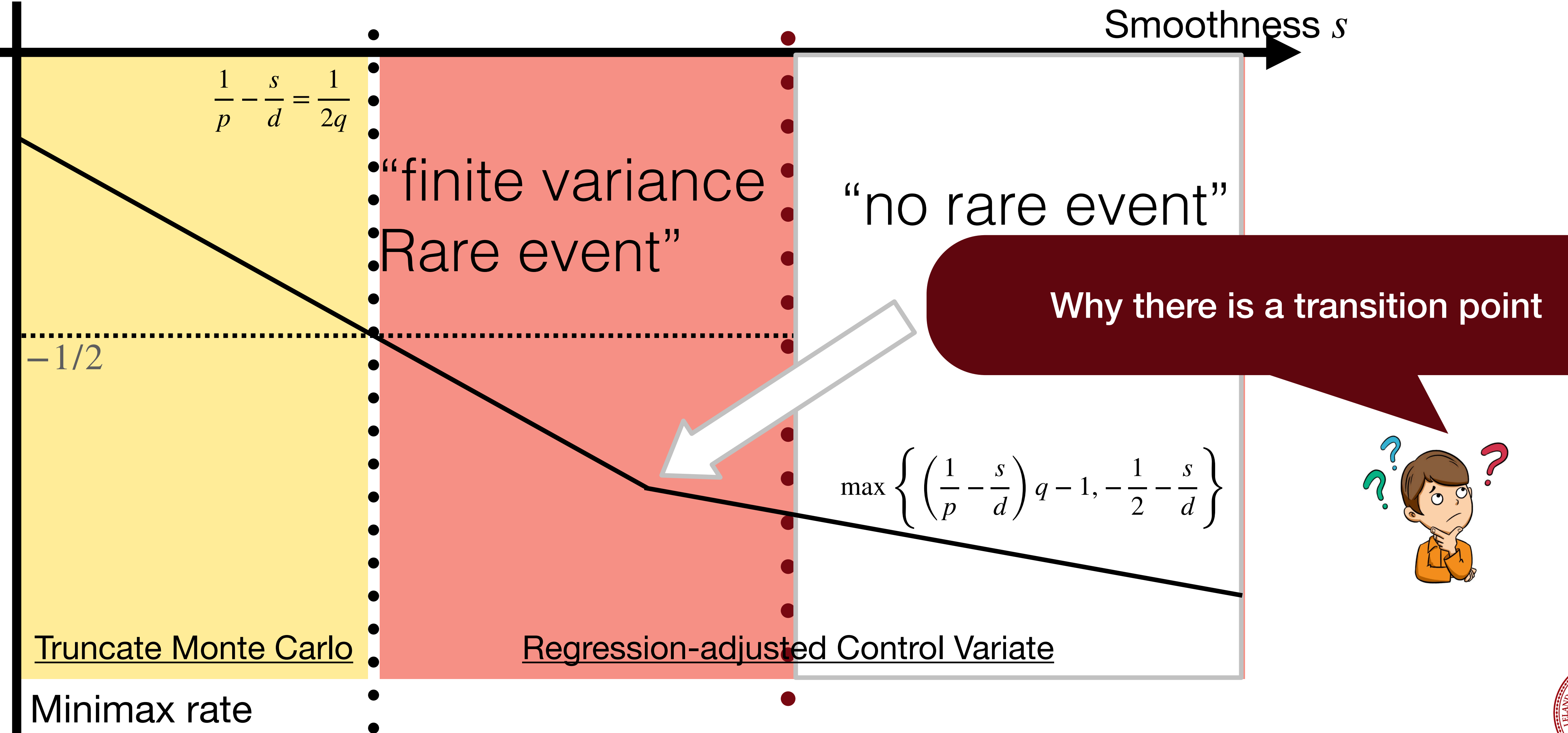
# When the control variate helps



# When the control variate helps



# When the control variate helps



# Semi-parametric efficiency...

**Example**

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

**Step 1**

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

**Step 2**

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

Low order term

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

“influnce function” (gradient)      Error propagation





# Semi-parametric efficiency...

**Example**

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

**Step 1**

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

**Step 2**

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

Low order term

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

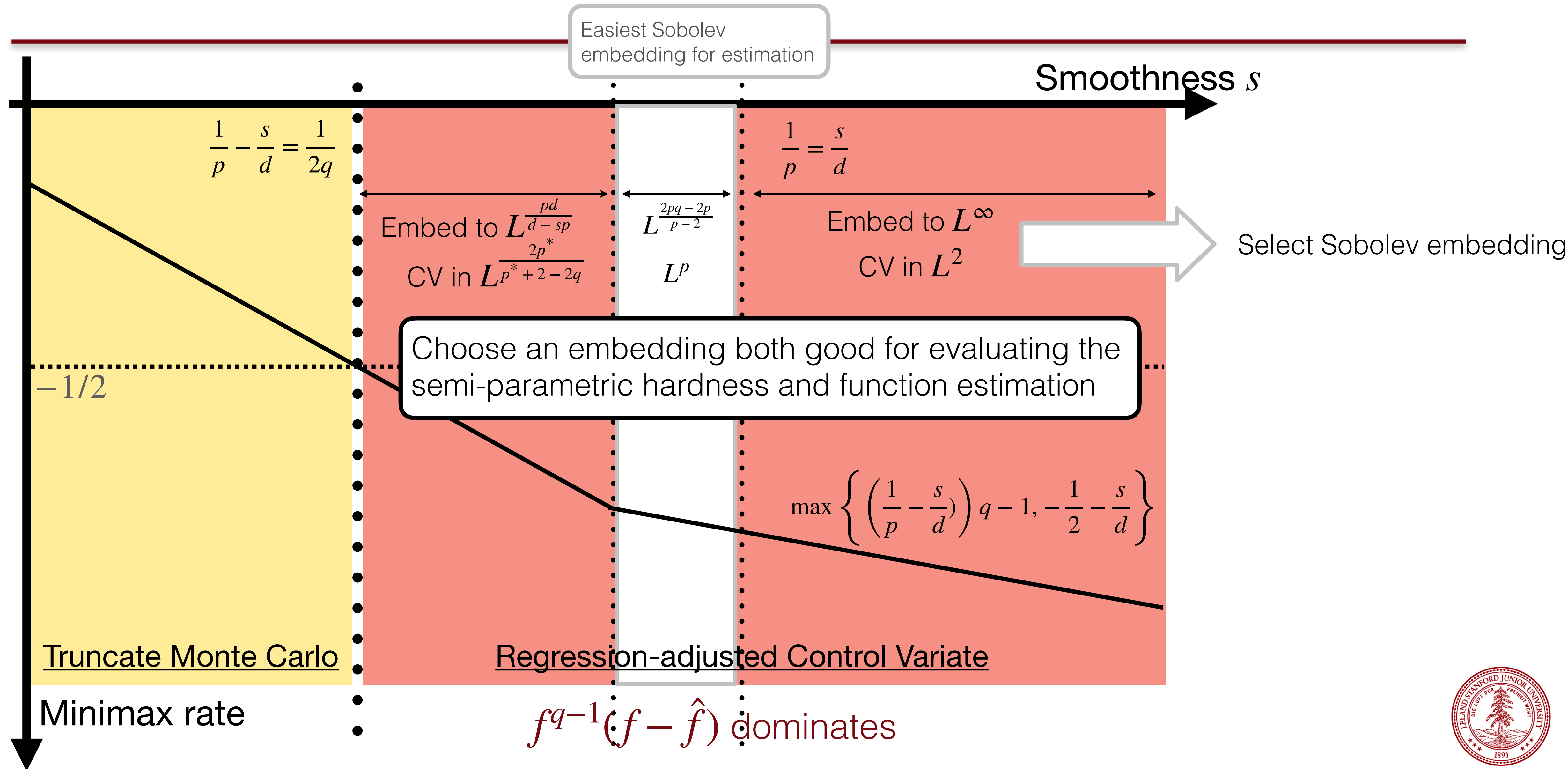
“influnce function” (gradient)

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

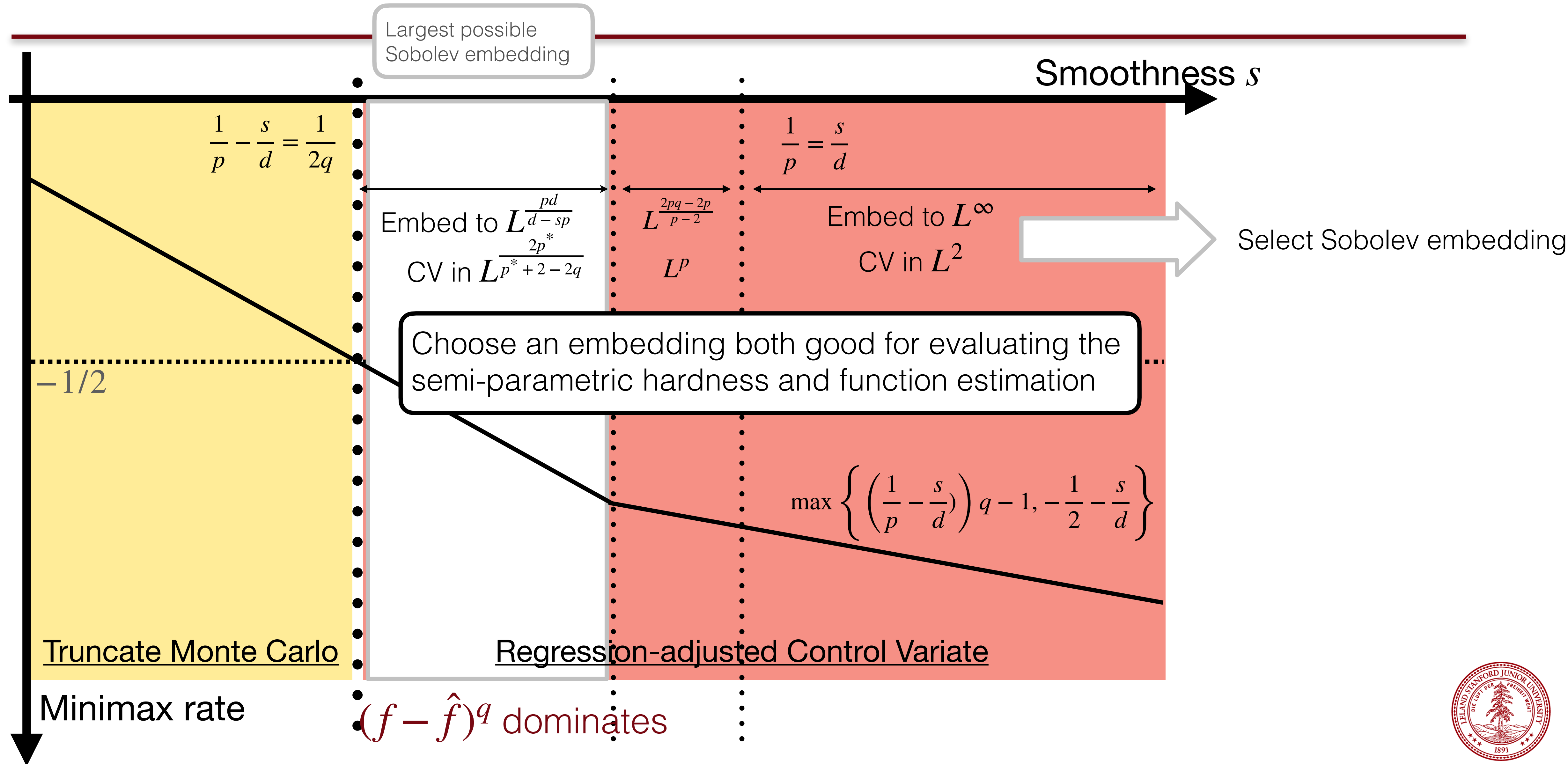
How to select the sobolev emebedding



# Tricky part of the Proof: select embedding



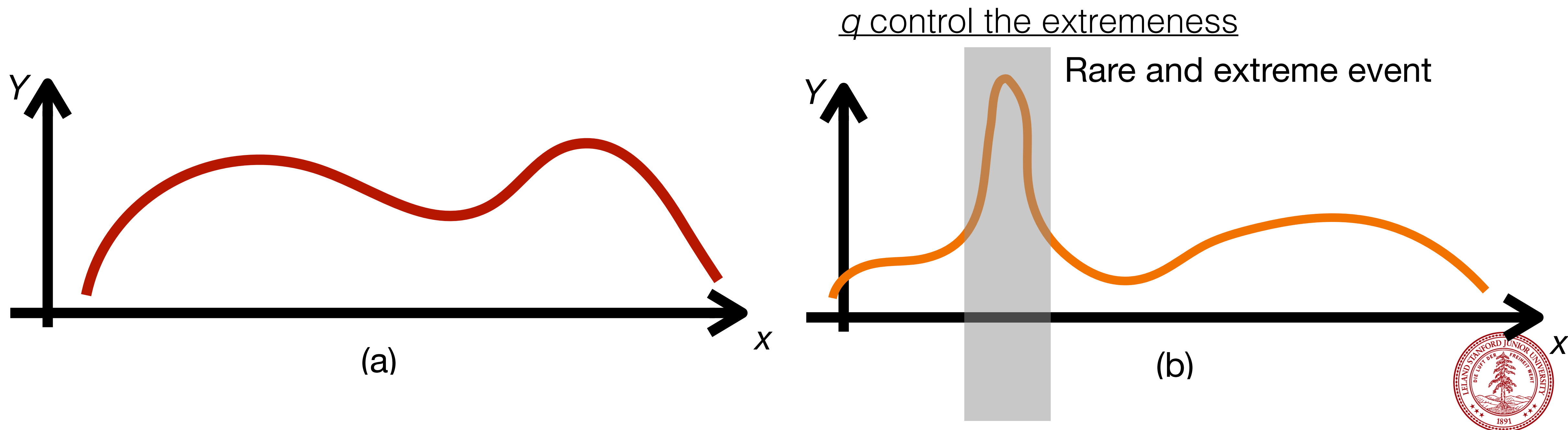
# Tricky part of the Proof: select embedding



# 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



# Optimal (Linear) Operator Learning

ICLR 2023 (spotlight)

$$Au = f$$

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

Recover parameter  $\theta$  in  
Model  $A_\theta$

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!**



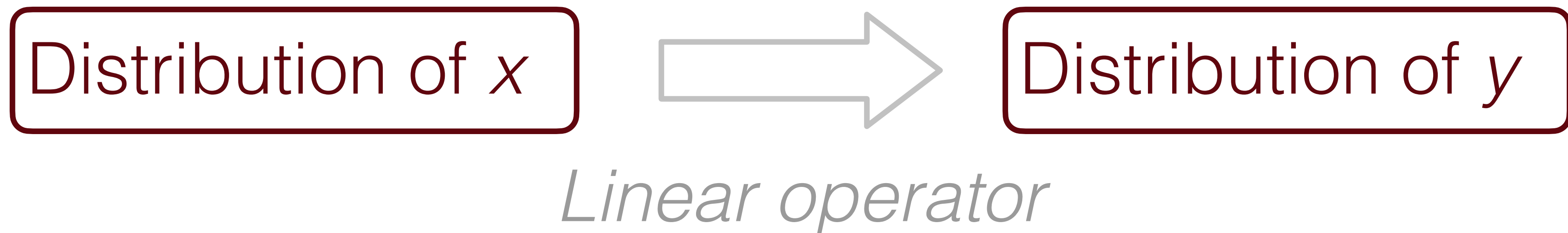
# Linear Operator itself is important still...

---

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

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

Distribution is ***infinite dimensional***



# Linear Operator itself is important still...

---

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

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

Distribution is ***infinite dimensional***

Instrumental variable regression  
[Singh-Chernozhukov-Newey 2022]

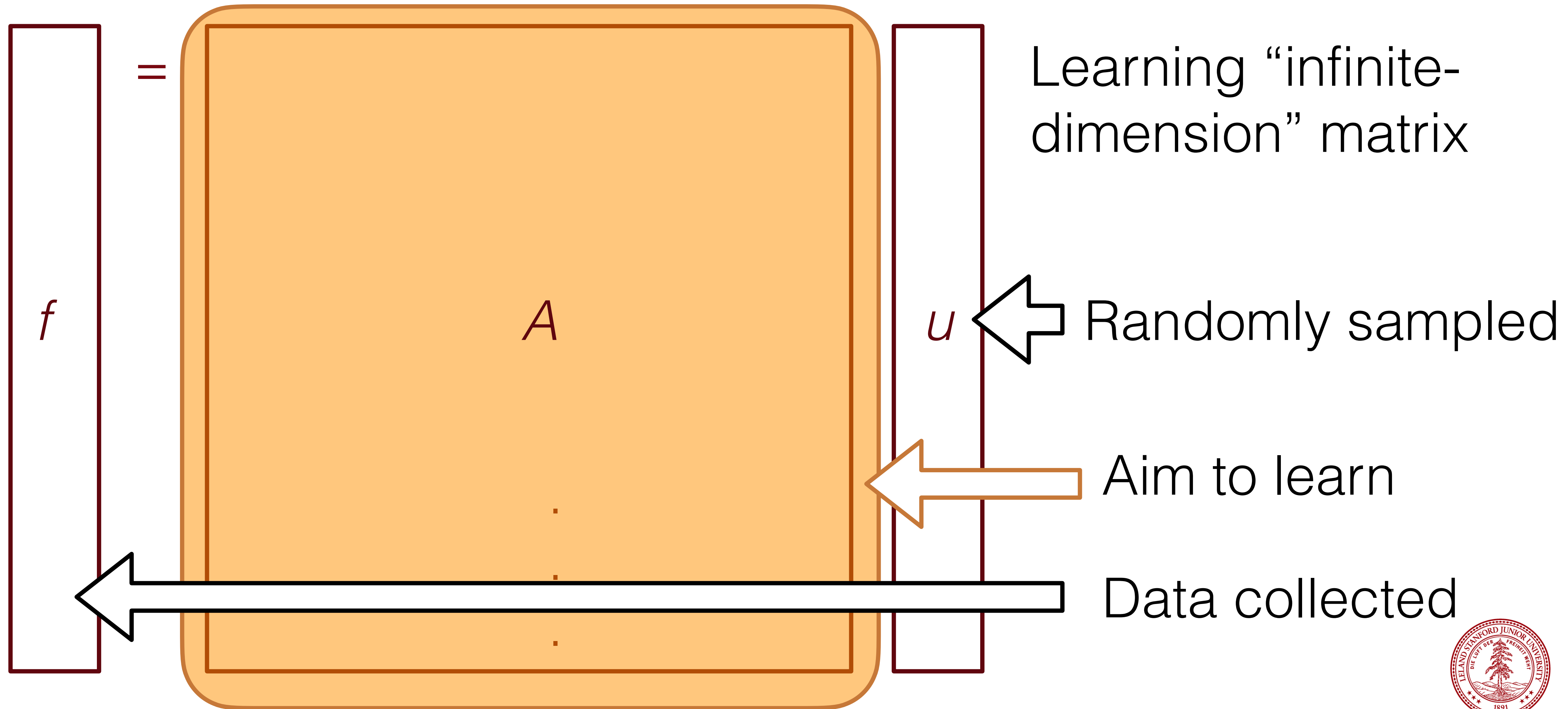
Time series modeling  
[Kostic-Novelli-Maurere-Ciliberto-Rosasco-Pontil 2022]



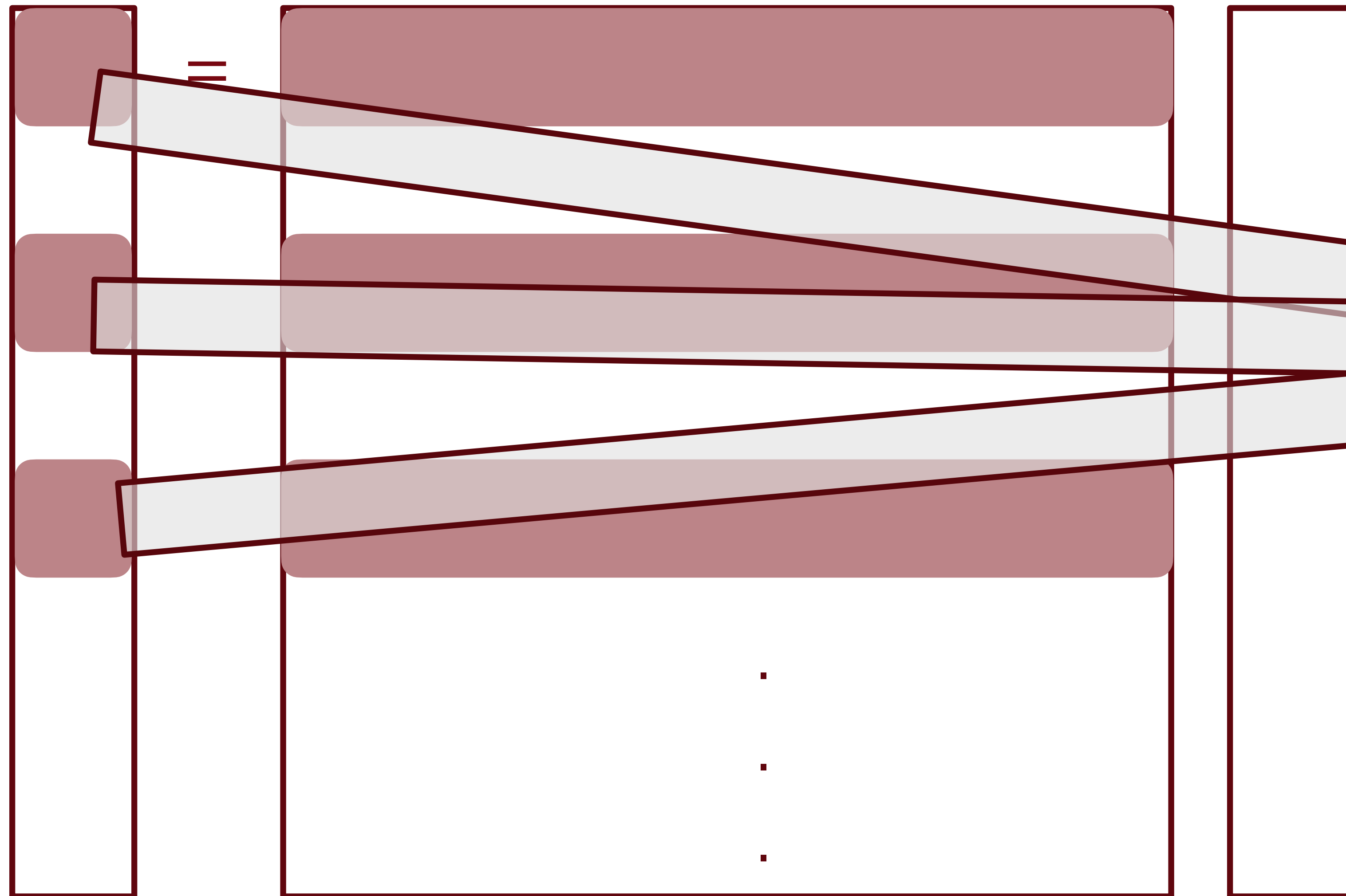
Generator/Koopman  
Operator/CME



# Linear Operator Learning



# Why infinite dimensional operator is hard



Learning “infinite-  
dimension” matrix

If every row have  $O(1)$  variance,  
The total variance is  $\infty$

[1] Talwai P, Shameli A, Simchi-Levi D.  
AISTAT 2022

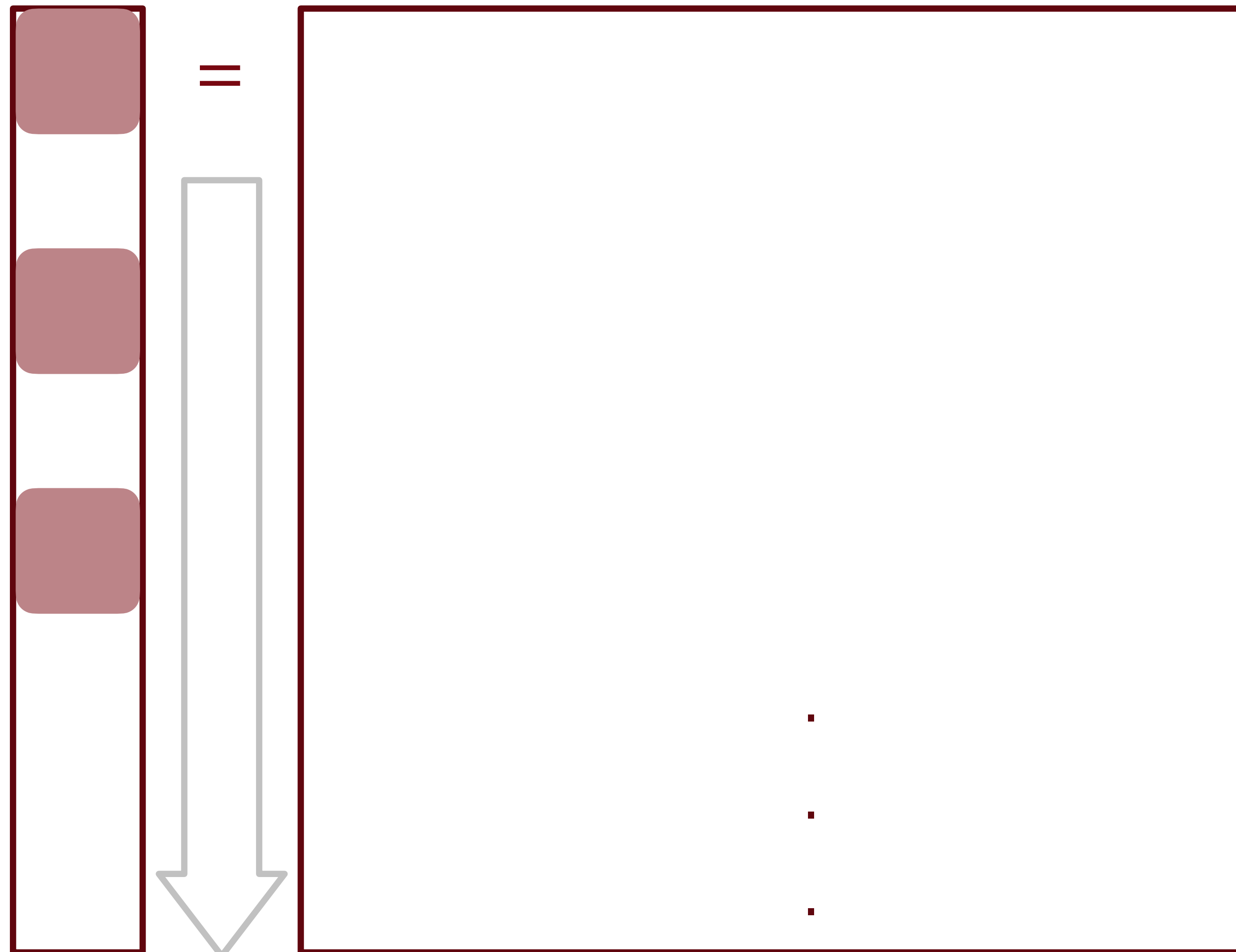
[2] Li Z, Meunier D, A Gretton. Neurips 2022

[3] de Hoop M V, et al. arXiv:2108.12515





# Why infinite dimensional operator is hard



Learning “infinite-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-  
matrix

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

*Decay  
ance.*

[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

---

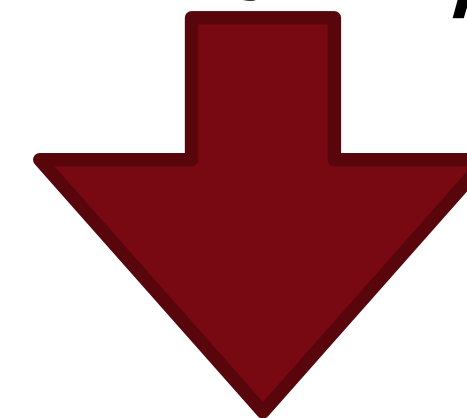
Hilbert space have finite variance as finite dimensional space

$$\square = \lambda_1 \begin{matrix} \text{vertical bar} \\ \text{horizontal bar} \end{matrix} + \dots$$

Eigen decomposition

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

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



**Ensures finite variance**

# Spaces we are interested

Hilbert space have finite variance as finite dimensional space

Eigen decomposition

$$\square = \lambda_1 \begin{matrix} \text{vertical bar} \\ \text{horizontal bar} \end{matrix} + \dots$$

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

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

“Kernel Sobolev space”: larger than RKHS  $H^\beta$  Fourier expansion

$$\text{horizontal bar} = a_1 \lambda_1^{\beta/2} \text{horizontal bar } e_1 + a_2 \lambda_2^{\beta/2} \text{horizontal bar } e_2 + \dots$$

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

Eigen decomposition

$$\square = \lambda_1 \begin{matrix} \text{vertical bar} \\ \text{horizontal bar} \end{matrix} + \dots$$

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

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

“Kernel Sobolev space”: larger than RKHS  $H^\beta$  Fourier expansion

$$\text{horizontal bar} = a_1 \lambda_1^{\beta/2} \text{horizontal bar } e_1 + a_2 \lambda_2^{\beta/2} \text{horizontal bar } e_2 + \dots$$

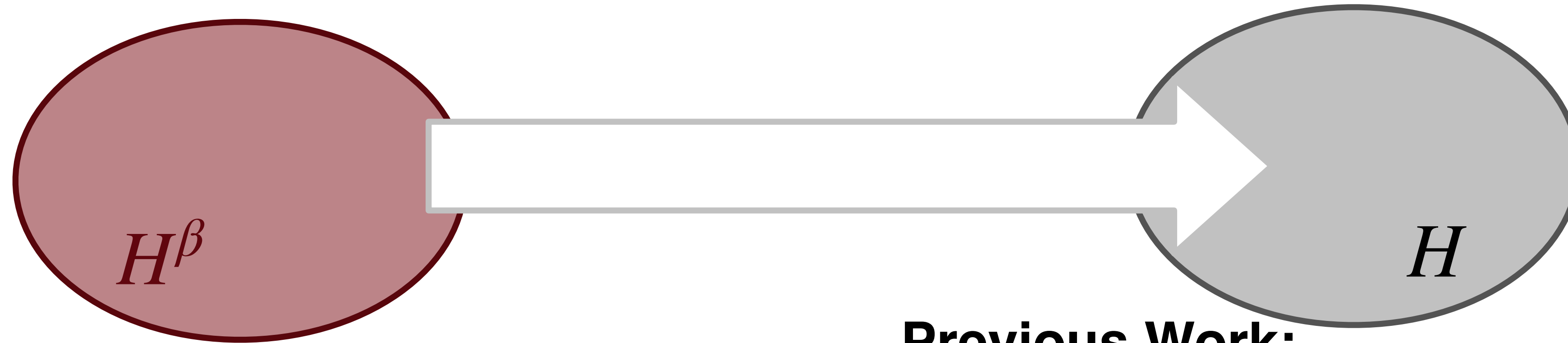
with  $(a_i)_{i=1}^{\infty} \in \ell_2, \beta \in (0, 1)$





# Problem Formulation

---



$H^\beta$  is a larger space

## Previous Work:

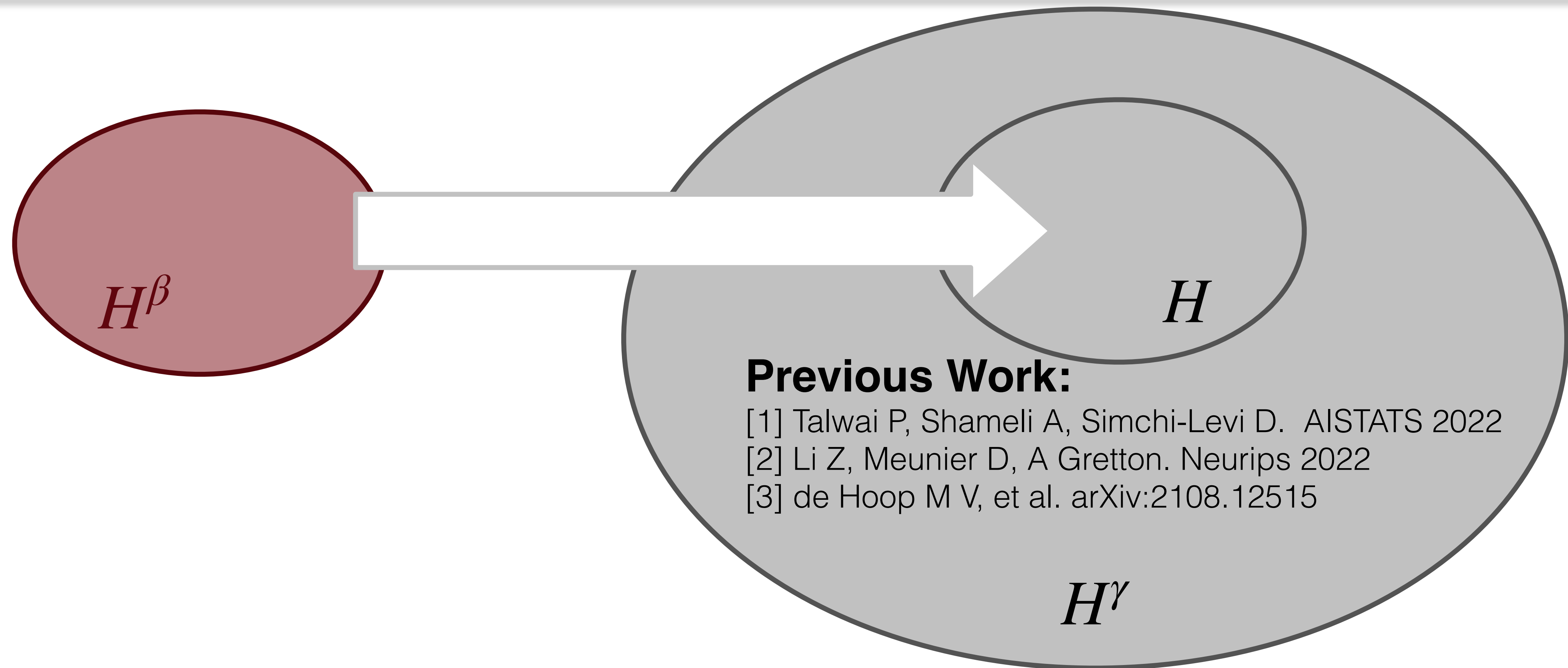
- [1] Talwai P, Shameli A, Simchi-Levi D. AISTATS 2022
- [2] Li Z, Meunier D, A Gretton. Neurips 2022
- [3] de Hoop M V, et al. arXiv:2108.12515

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



# Problem Formulation

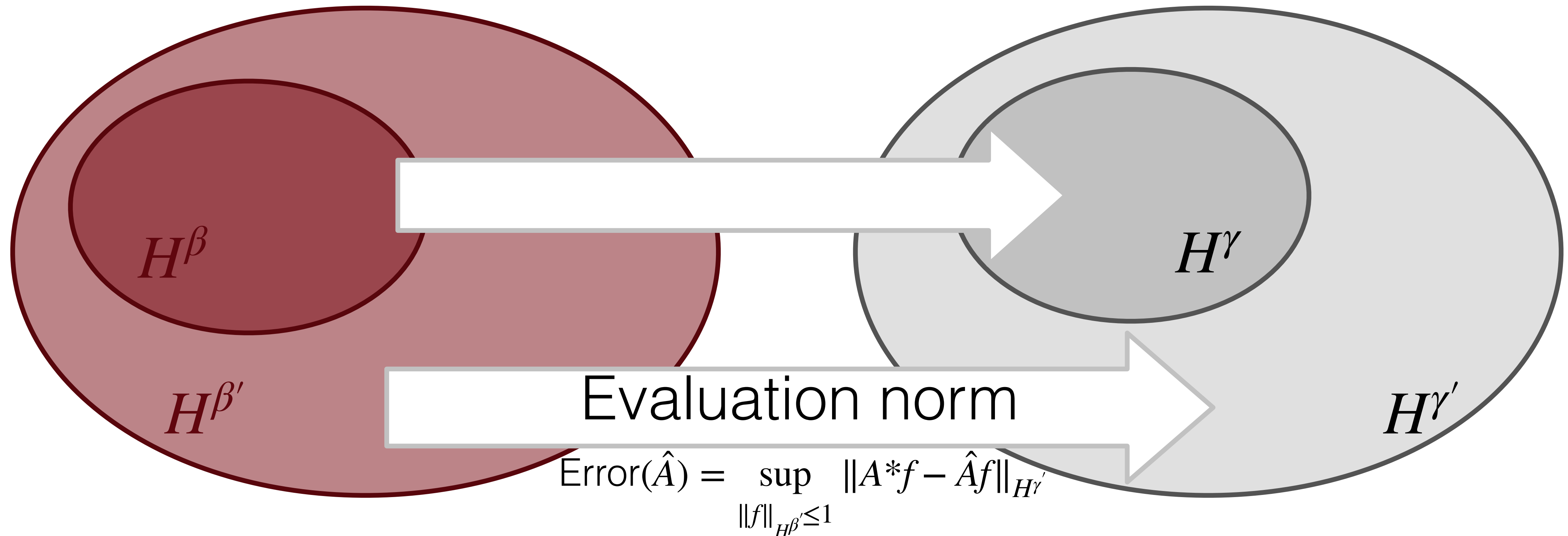
---



How the optimal rate depend on  $\gamma$  (output space complexity)?  
Is the previous algorithm still Optimal?

# Problem Formulation

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



# Main Result: Lower bound

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

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

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

With  $\mathbf{N}$  random observations



# Main Result: Lower bound

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

For all (randomized) estimators  $\mathcal{L}$ , we have Only output function space

$$\sup_{\|A\|_{H^\beta \rightarrow H^\gamma} \leq 1} \|\mathcal{L}(\{u_i, f_i\}_{i=1}^N) - A\|_{H^{\beta'} \rightarrow H^{\gamma'}}^2 \gtrsim N^{-\min\left\{\frac{\beta - \beta'}{\beta + p}, \frac{\gamma - \gamma'}{\gamma}\right\}}$$

With  $\mathbf{N}$  random observations Only input function space

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  $\|\cdot\|_{H^\beta \rightarrow H^\gamma}$  norm  
 Respect to  $\|\cdot\|_{H^{\beta'} \rightarrow H^{\gamma'}}$  Hilbert-schmidt norm

For all (randomized) estimators  $\mathcal{L}$ , we have Only output function space

$$\sup_{\|A\|_{H^\beta \rightarrow H^\gamma} \leq 1} \|\mathcal{L}(\{u_i, f_i\}_{i=1}^N) - A\|_{H^{\beta'} \rightarrow H^{\gamma'}}^2 \gtrsim N^{-\min\left\{\frac{\beta - \beta'}{\beta + p}, \frac{\gamma - \gamma'}{\gamma}\right\}}$$

With  $\mathbf{N}$  random observations Only input function space

Reason we introduce the test norm



# Main Result: Lower bound

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

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

$$\sup_{\|A\|_{H^\beta \rightarrow H^\gamma} \leq 1} \|\mathcal{L}(\{u_i, f_i\}_{i=1}^N) - A\|_{H^{\beta'} \rightarrow H^{\gamma'}}^2 \gtrsim N^{-\min\left\{\frac{\beta - \beta'}{\beta + p}, \frac{\gamma - \gamma'}{\gamma}\right\}}$$

With  $\mathbf{N}$  random observations

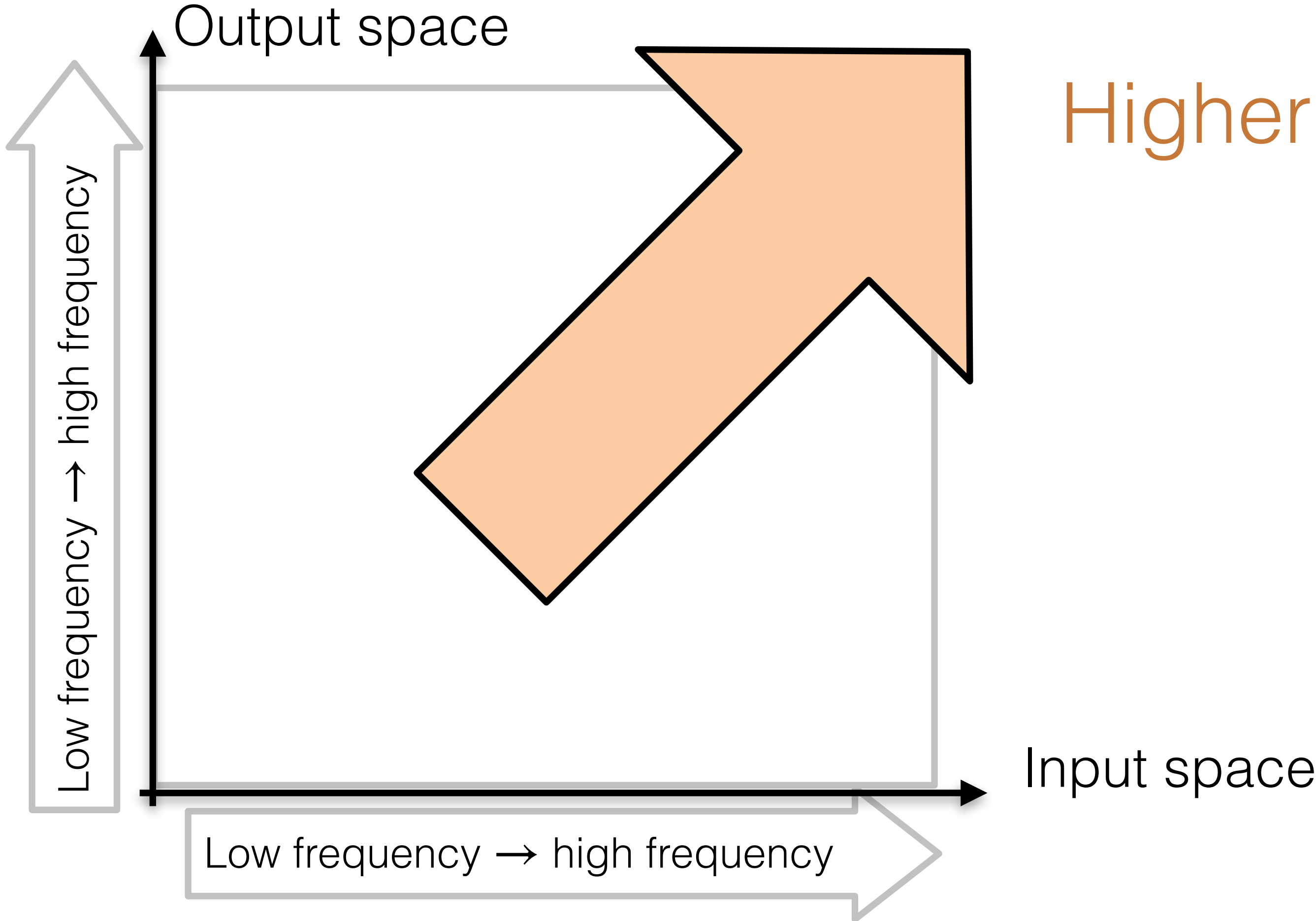


A magic result, can you explain it to me in a simple way?



# Consider the matrix view...

Operator is an “infinite” dimensional “matrix”

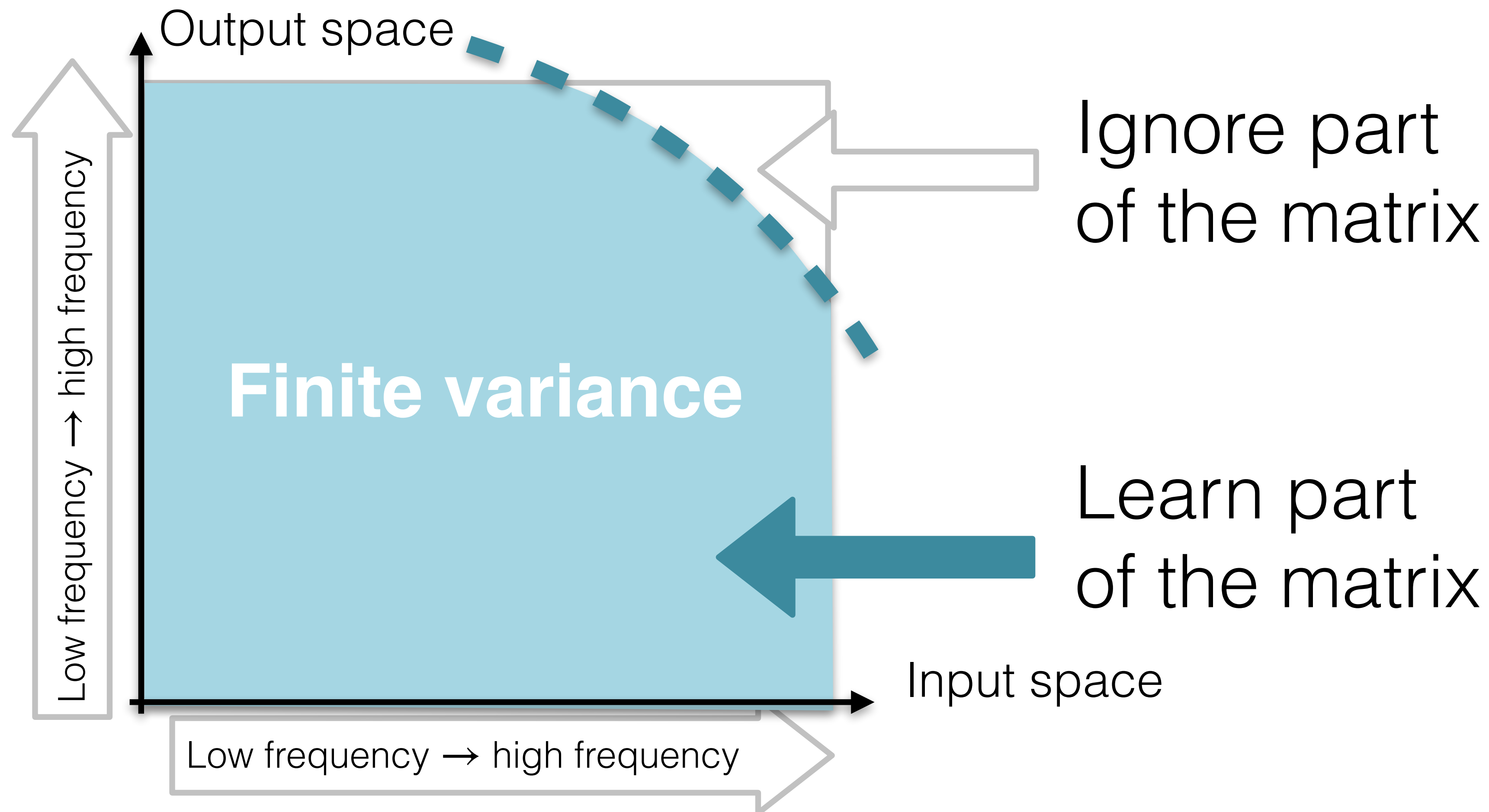


Higher Variance but Smaller Bias

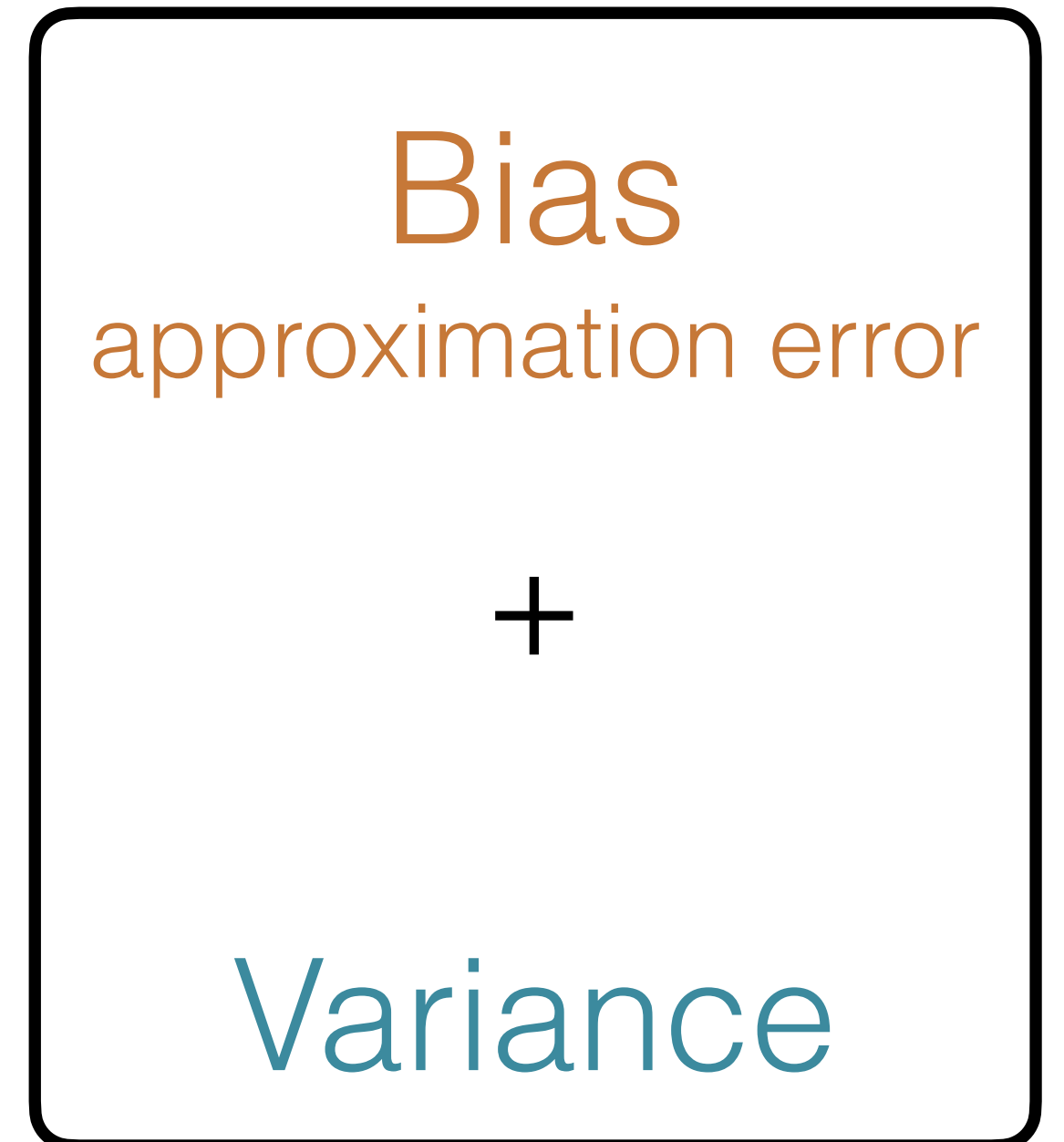


# Bias Variance Tradeoff

What is needed to achieve  $N^\theta$  learning rate

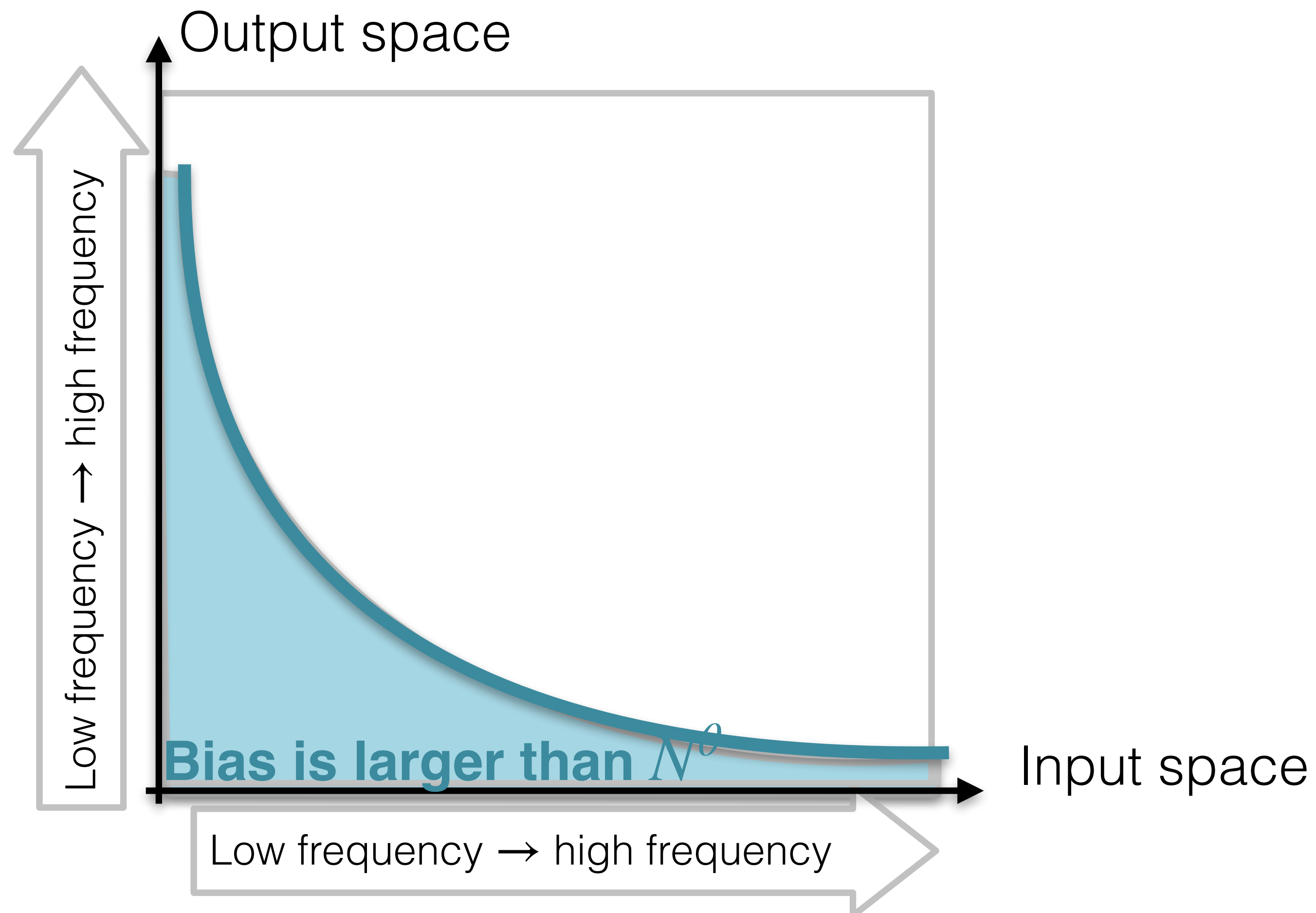


*“Trade off”*



# Optimal shape for Bias Variance Trade Off

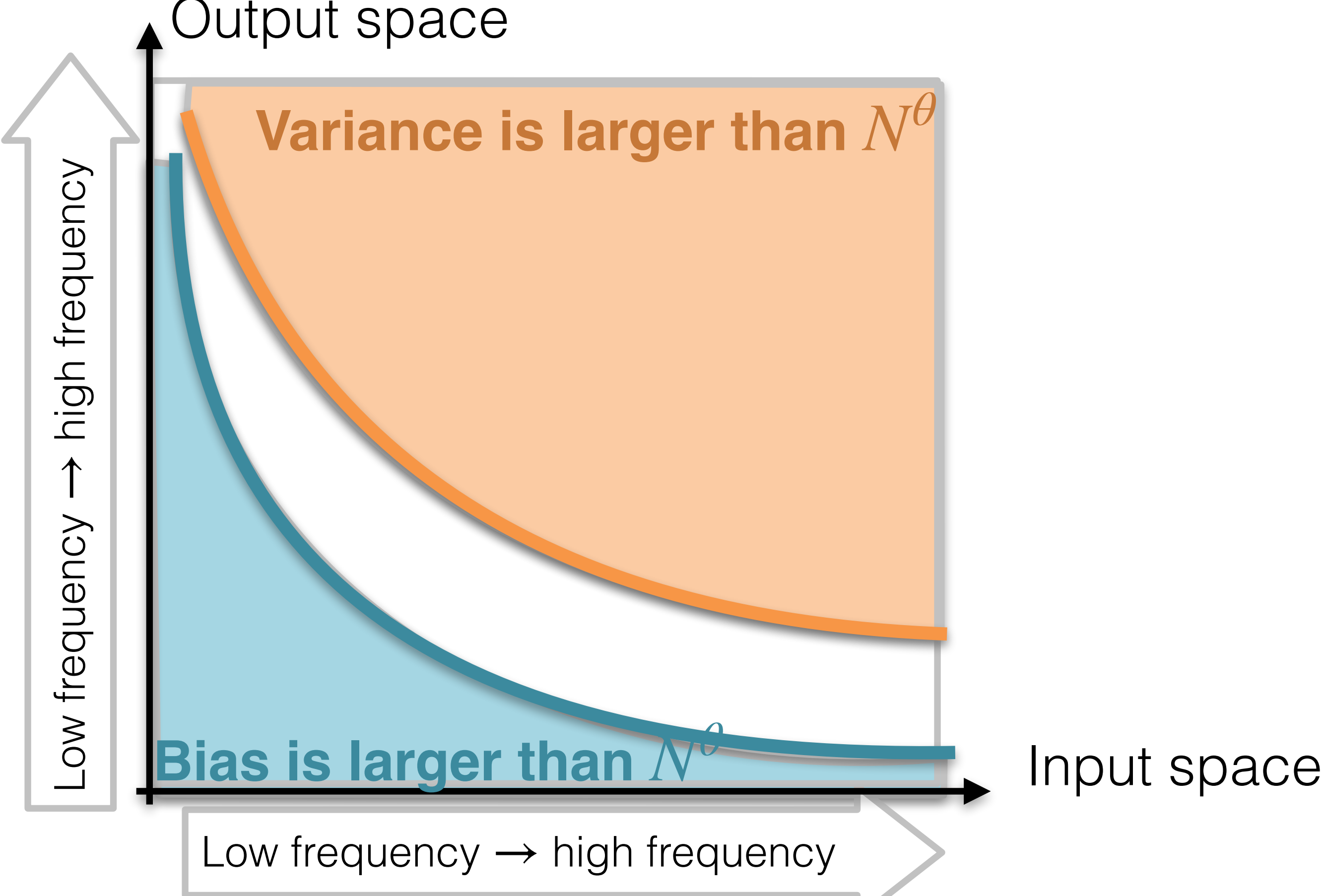
What is needed to achieve  $N^\theta$  learning rate





# Optimal shape for Bias Variance Trade Off

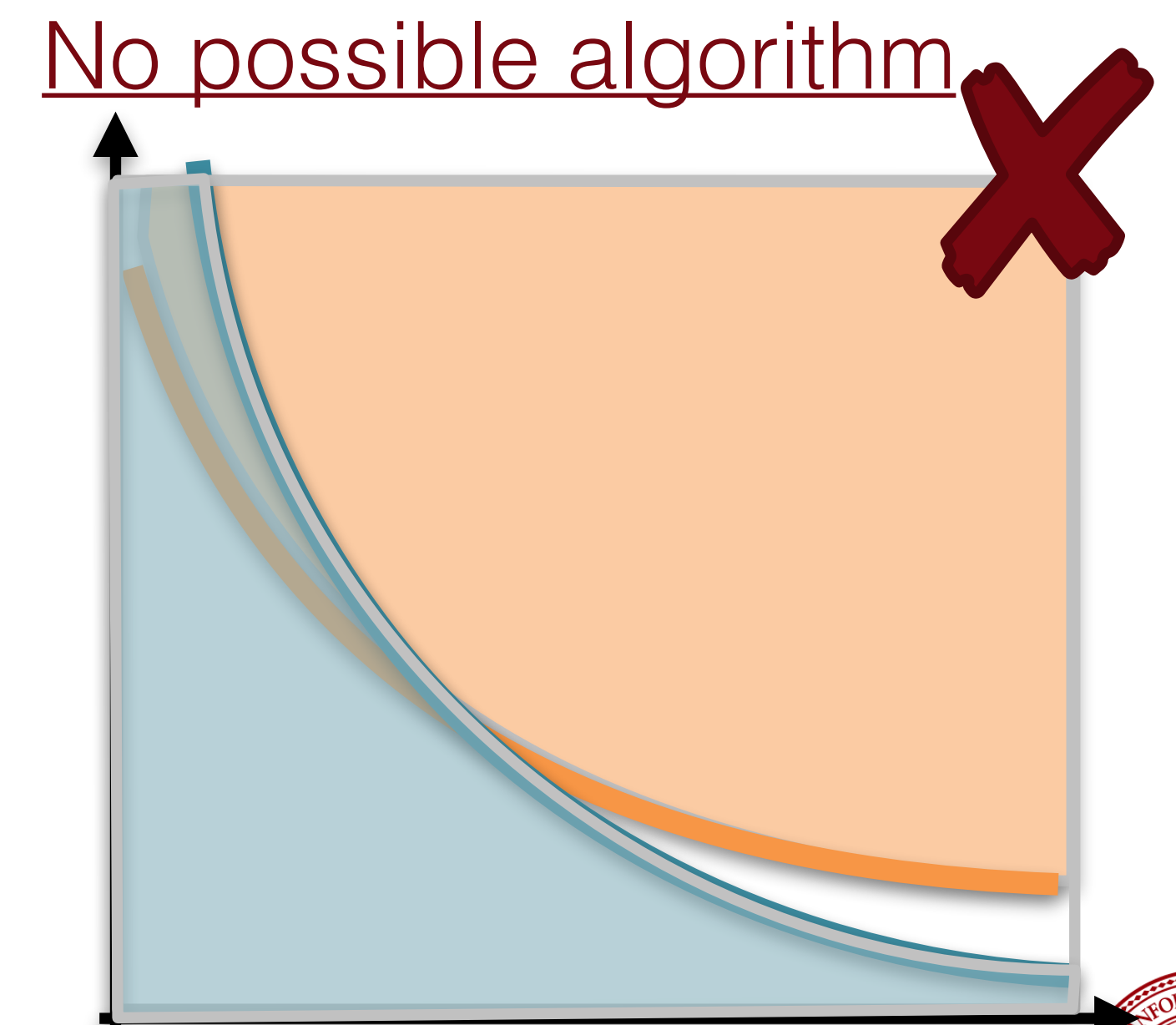
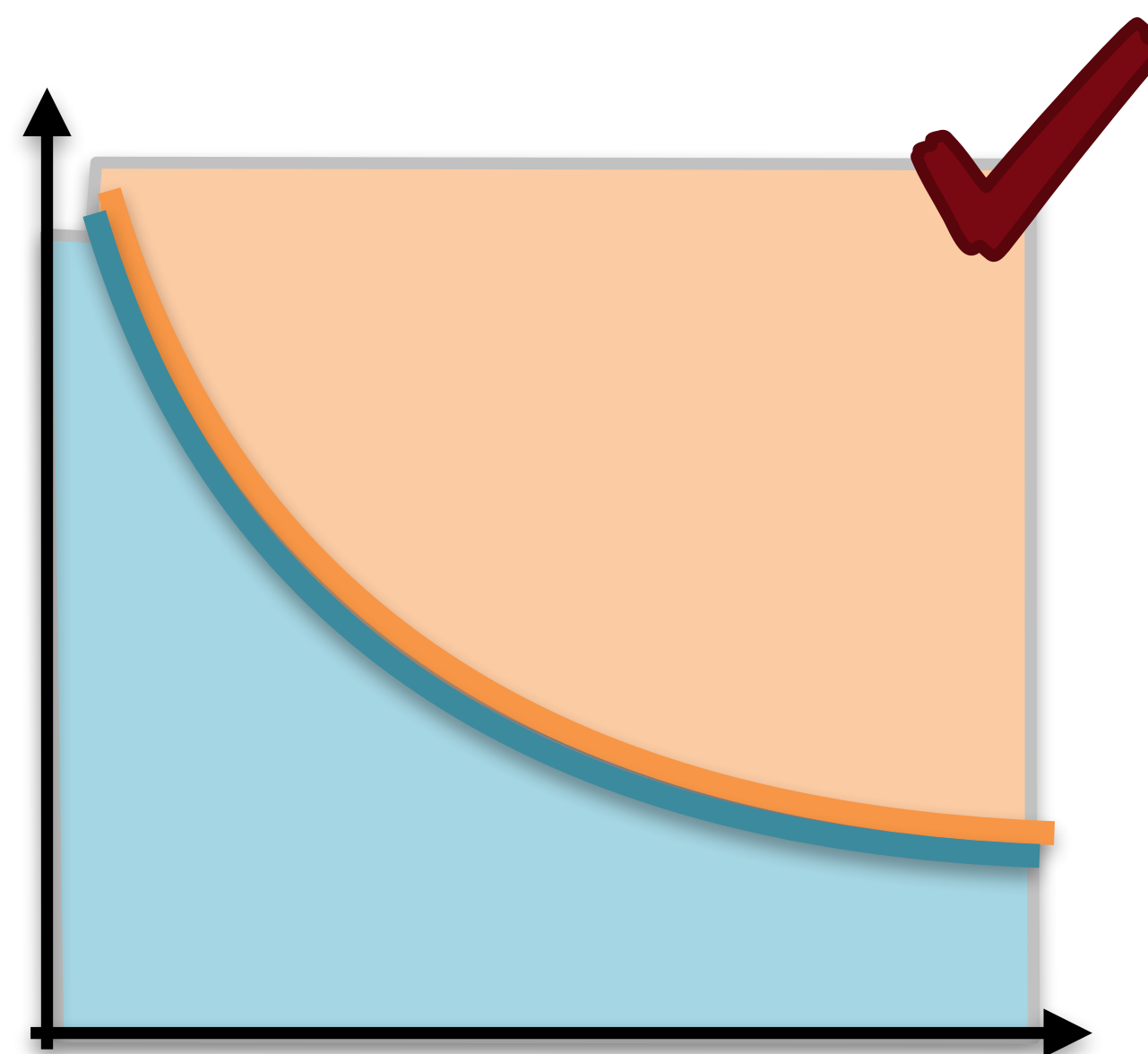
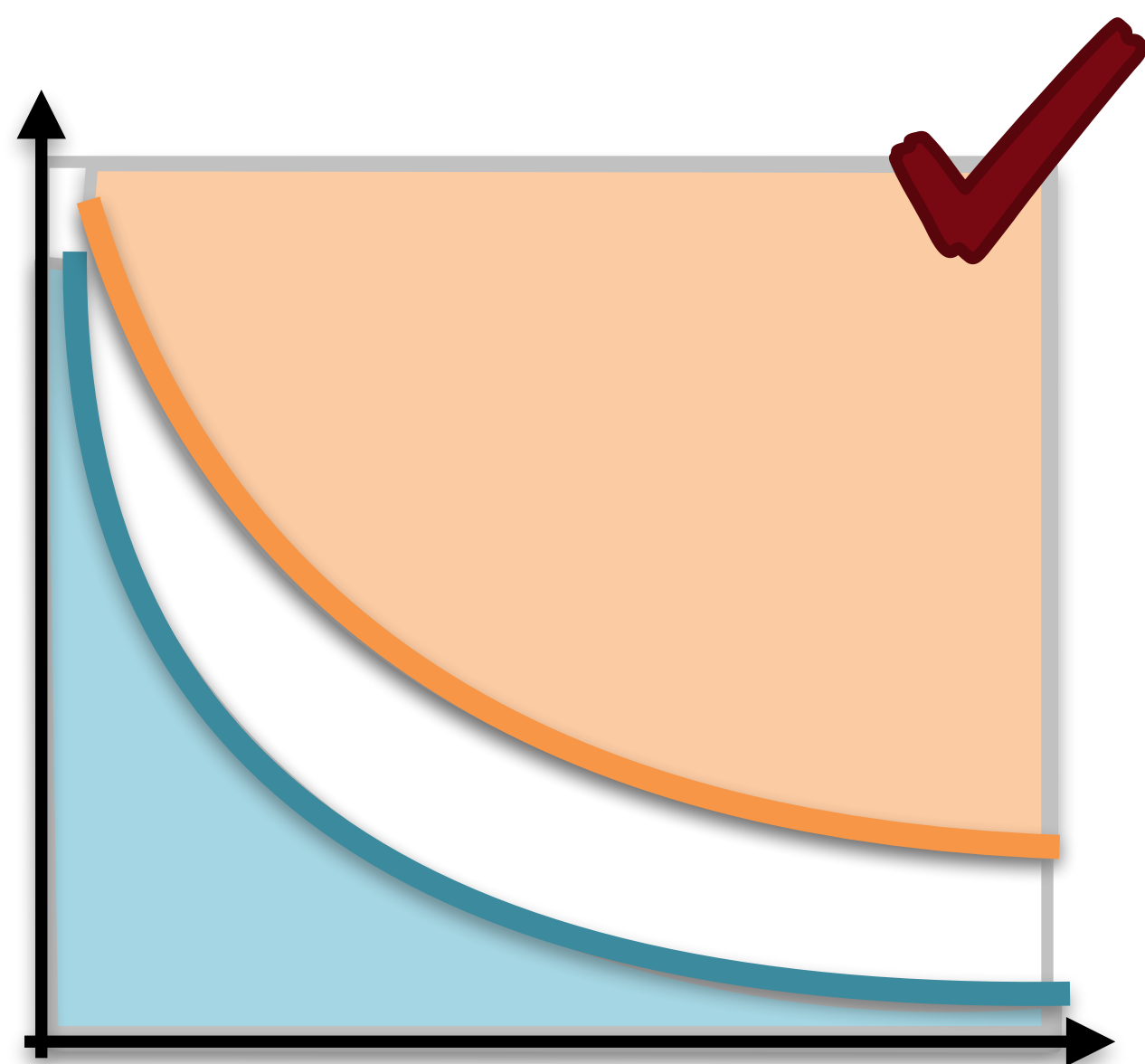
What is needed to achieve  $N^\theta$  learning rate



# Optimal shape for Bias Variance Trade Off

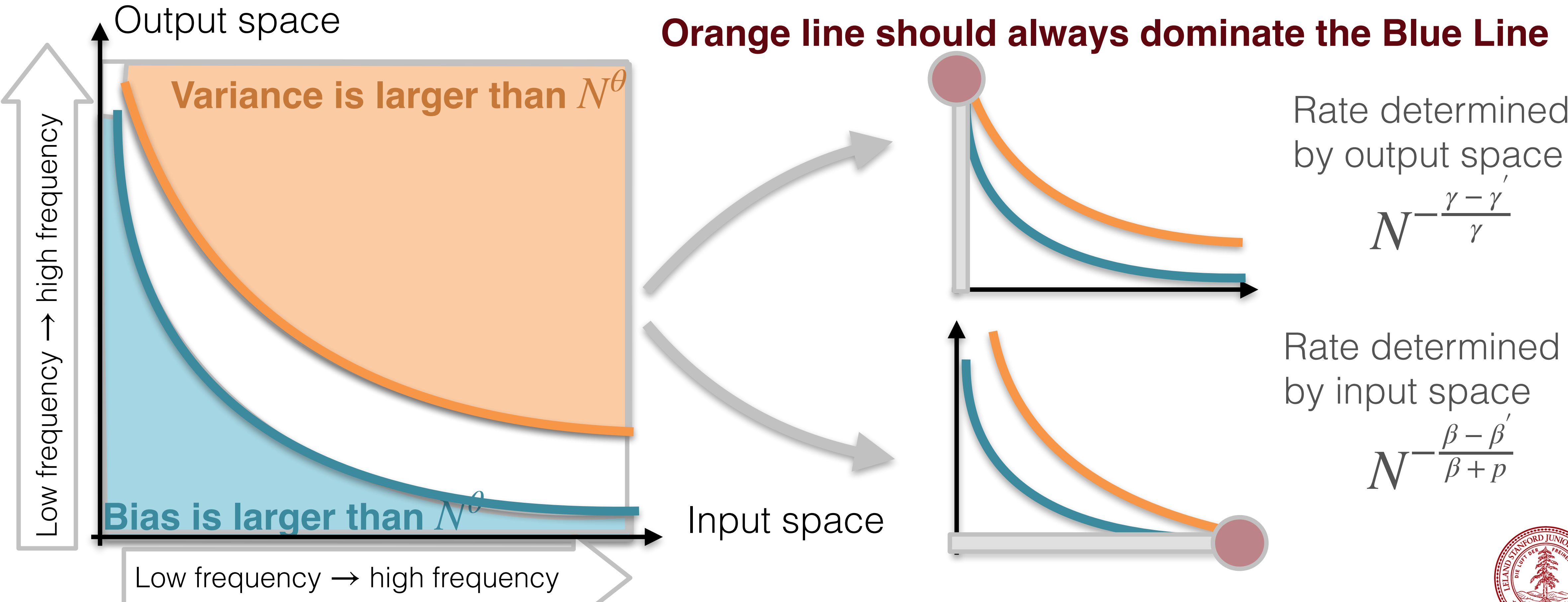
What is needed to achieve  $N^\theta$  learning rate

When  $\theta$  varies, there are three possible cases



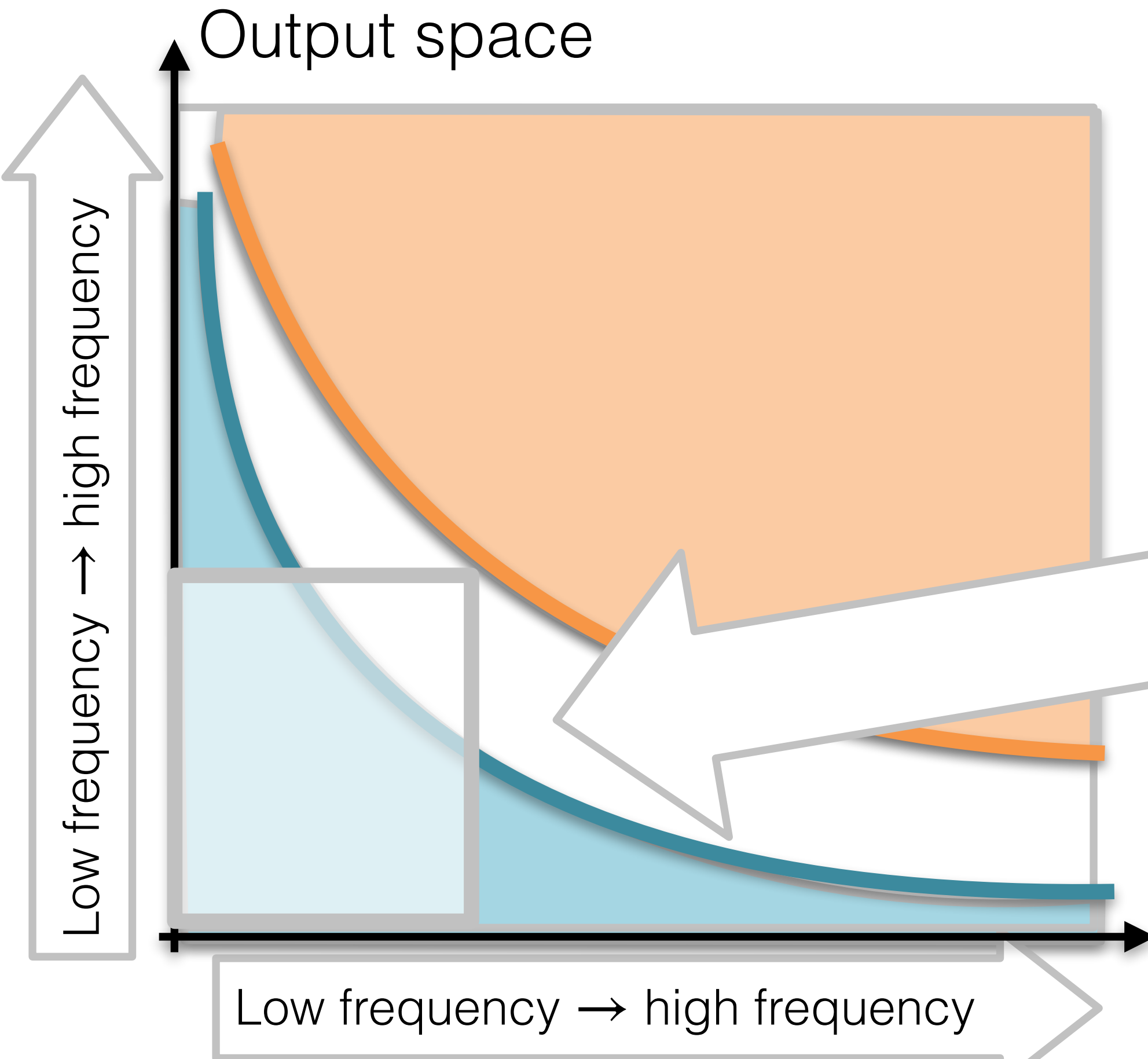
# Optimal shape for Bias Variance Trade Off

What is needed to achieve  $N^\theta$  learning rate



# Optimal Algorithm

What is the **OPTIMAL** machine learning algorithm?



Rectangular covering the blue part without touching the orange part

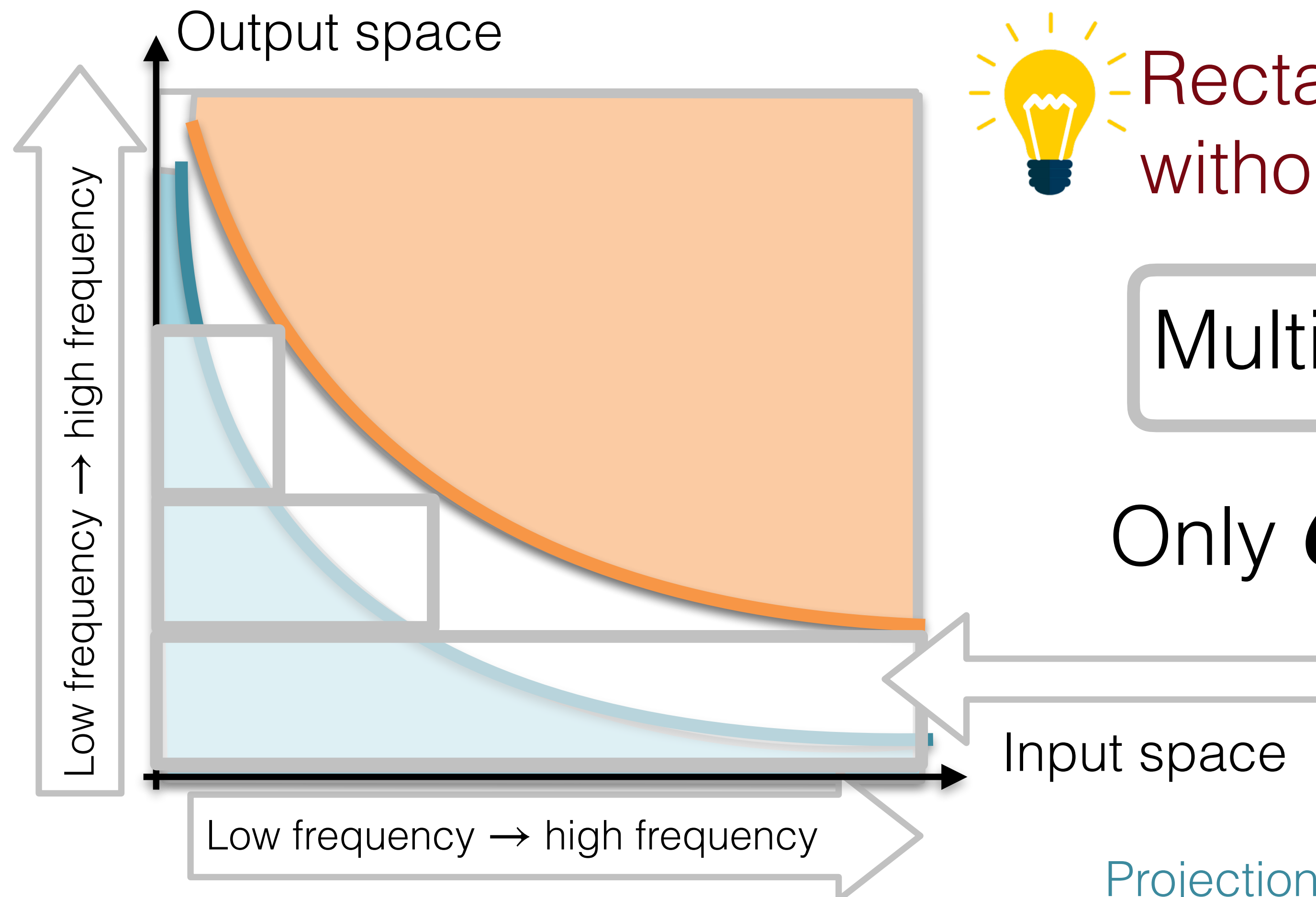
A ridge-regression/  
Discretization(PCA-Net) is learning a rectangular





# 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

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

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

Ridge regression

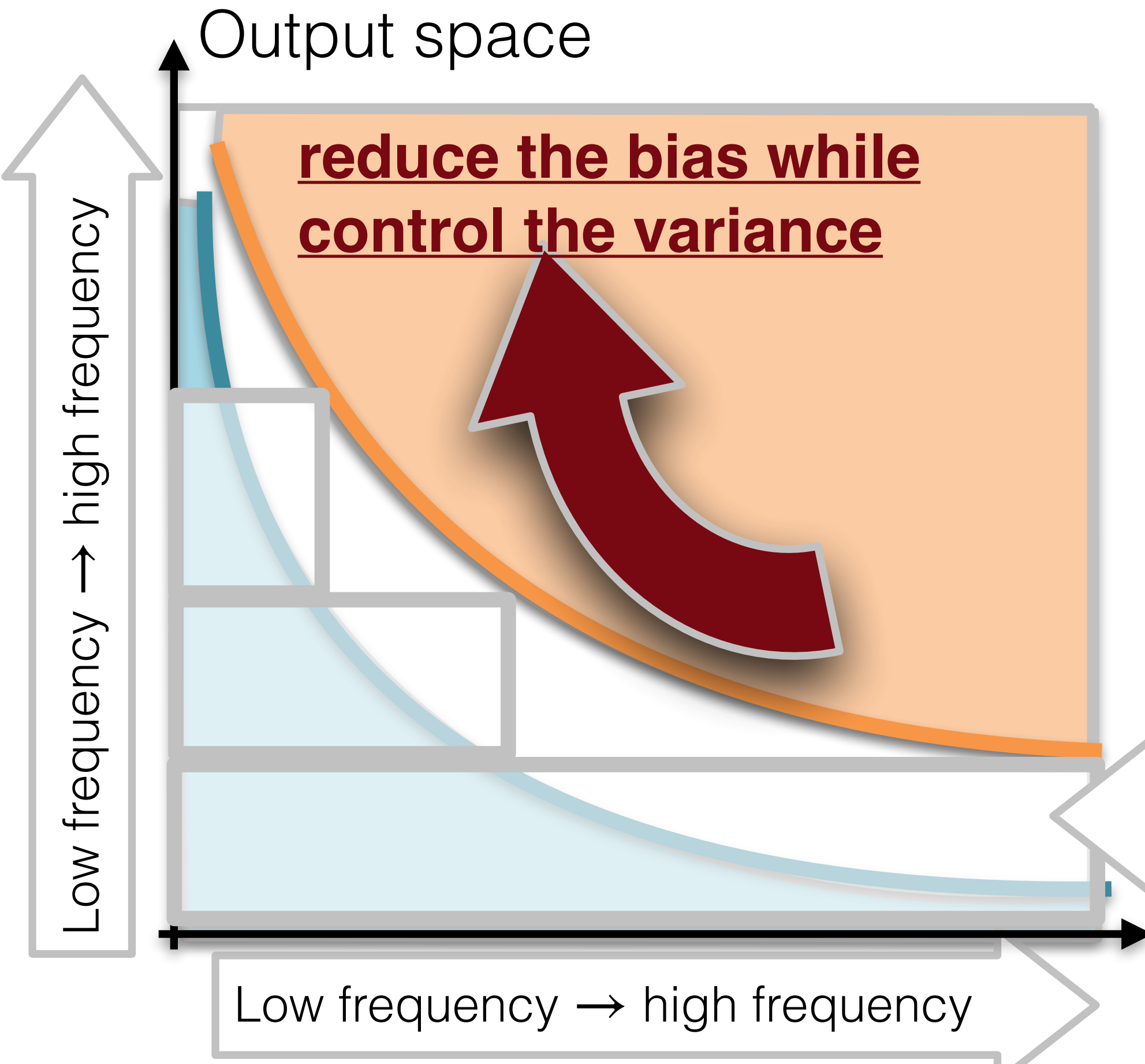
Projection to certain basis in output space





# 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

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

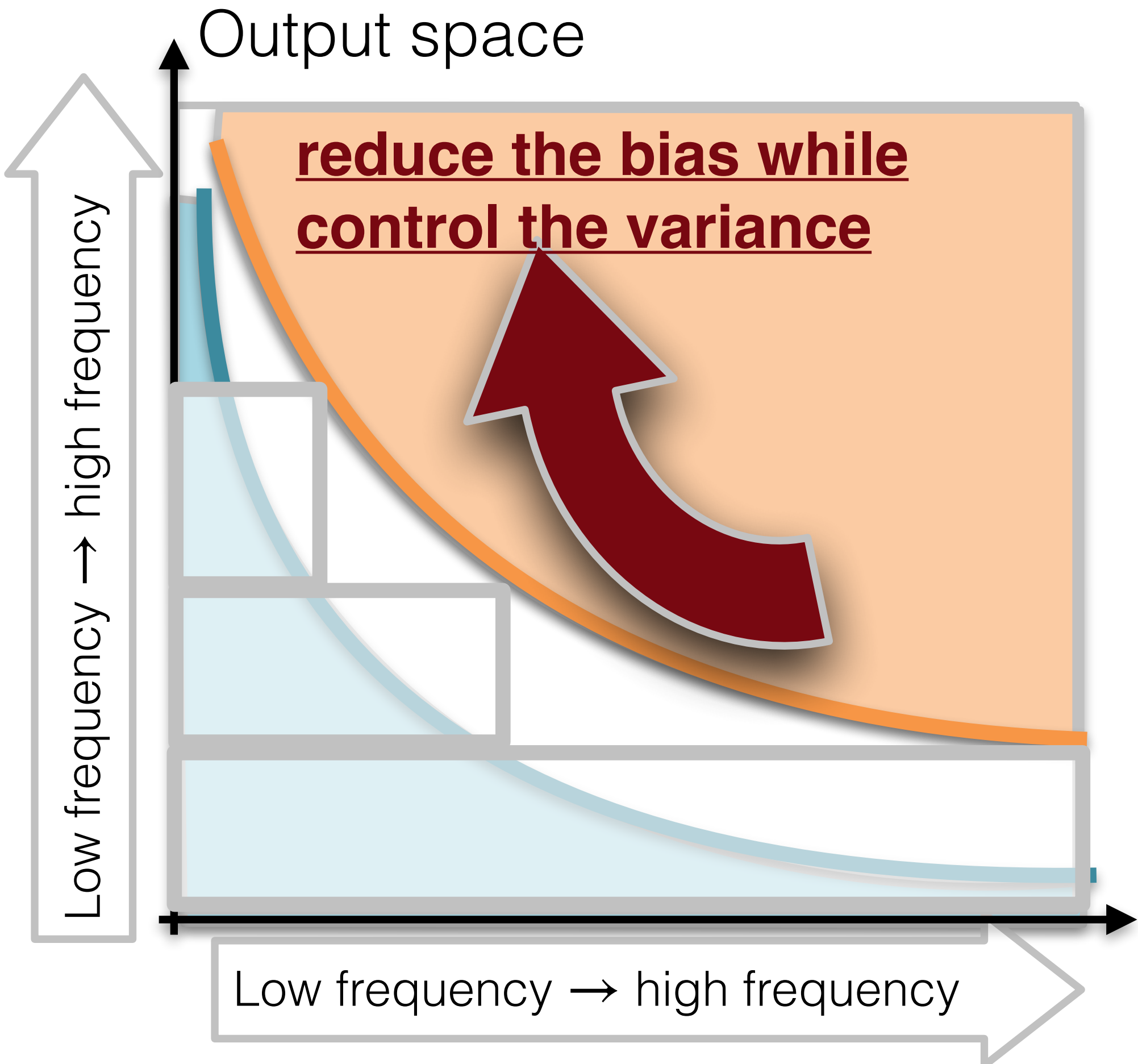
$$\hat{C}_{LK} (\hat{C}_{KK} + \lambda_i^{(K)} I)^{-1}$$

Ridge regression

Projection to certain basis in output space

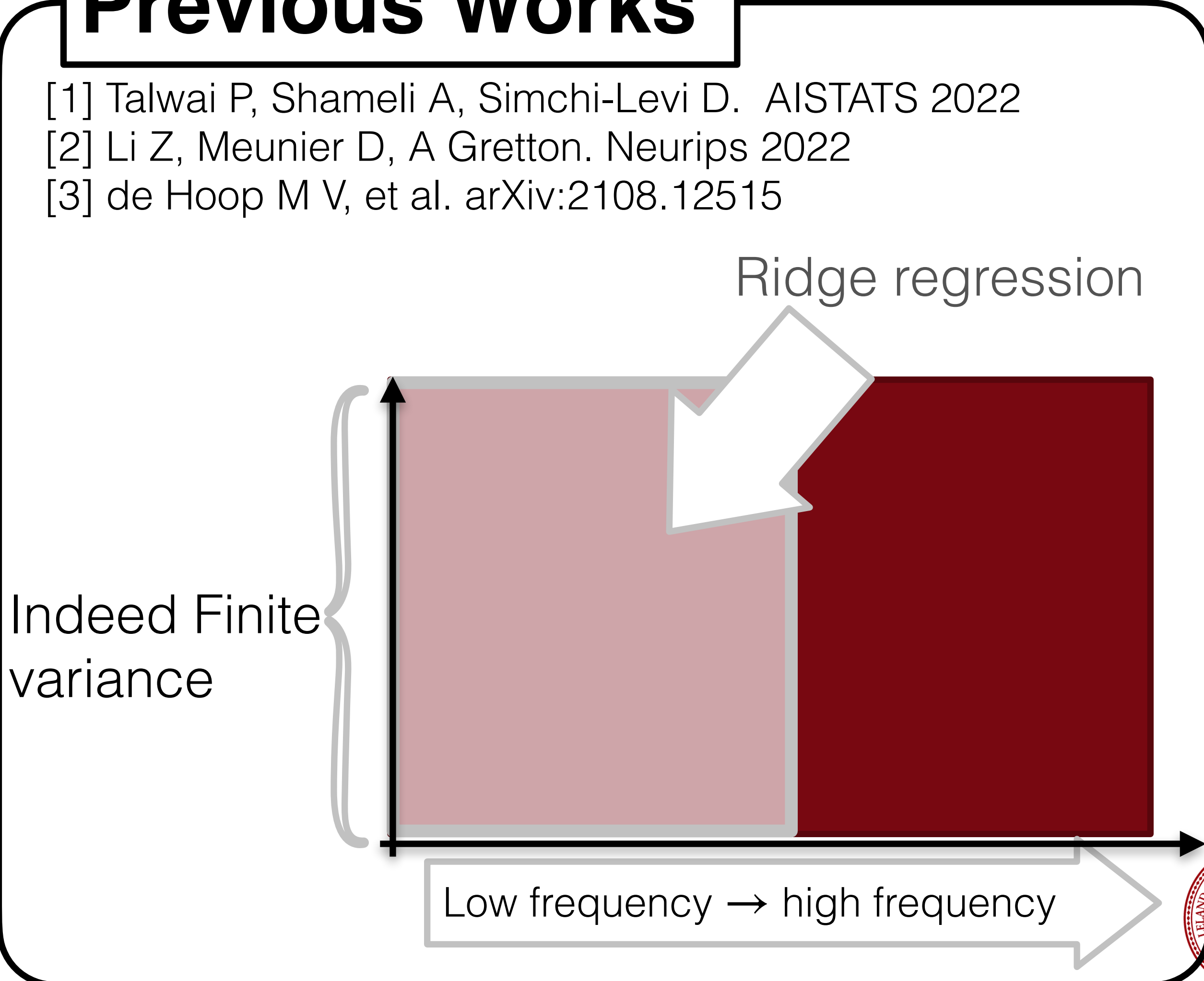


# Optimal Algorithm Changed...



## Previous Works

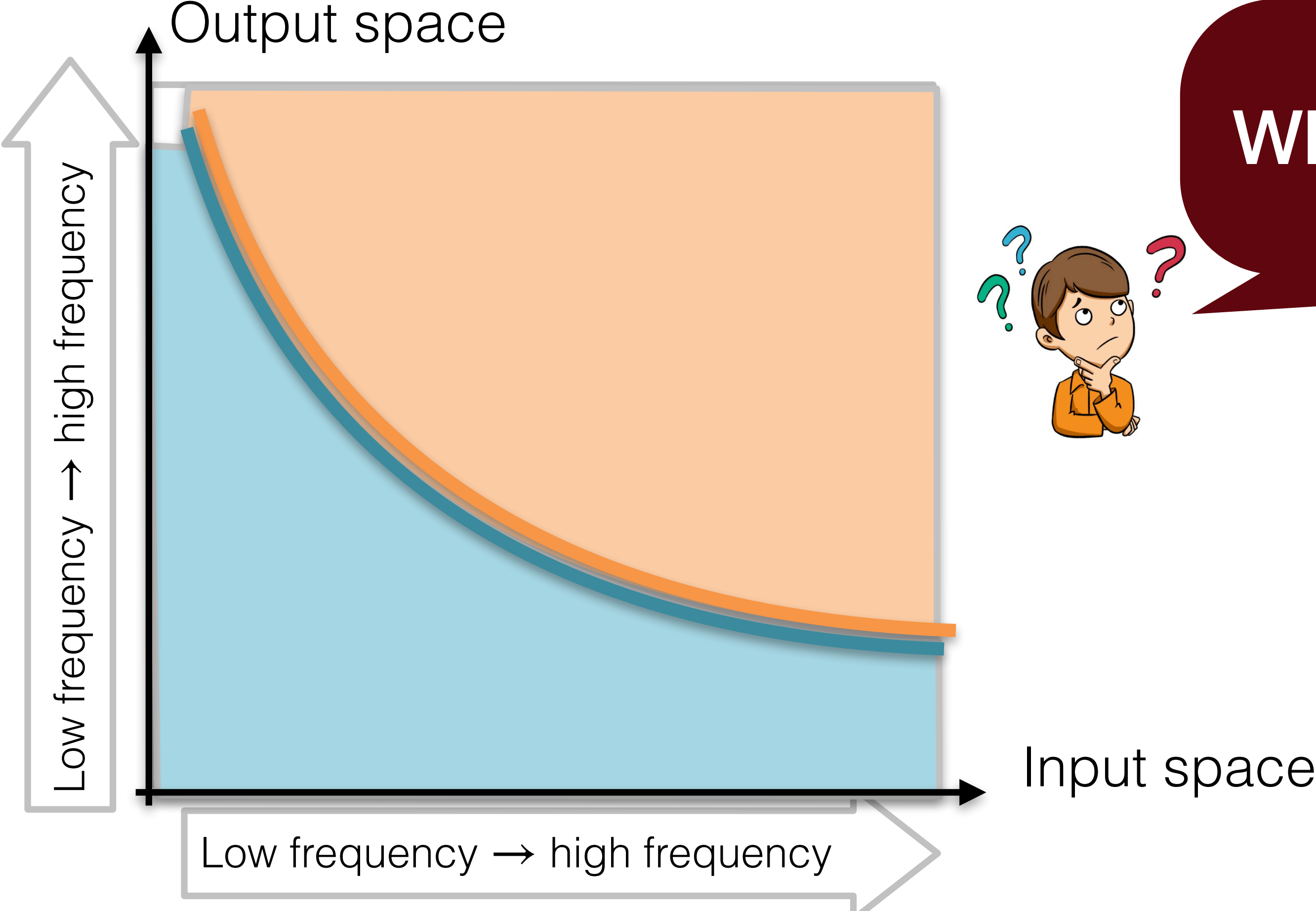
- [1] Talwai P, Shameli A, Simchi-Levi D. AISTATS 2022
- [2] Li Z, Meunier D, A Gretton. Neurips 2022
- [3] de Hoop M V, et al. arXiv:2108.12515



# Optimal Algorithm

Multilevel Training

What is the **OPTIMAL** machine learning algorithm?



What if the two lines coincide?

Output space  
Learning rate

$$\frac{\gamma - \gamma'}{\gamma}$$

=

Input space  
learning rate

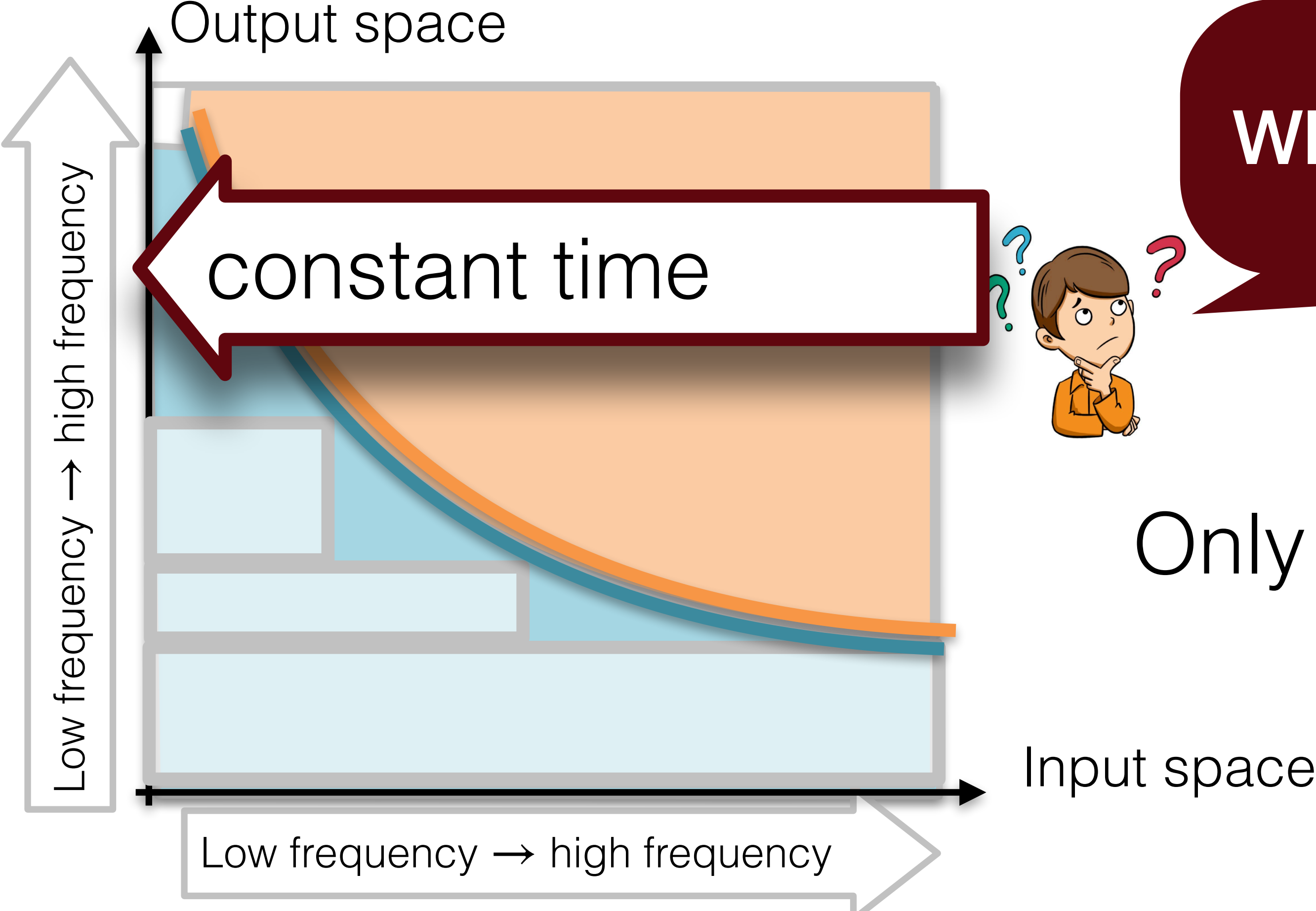
$$\frac{\beta - \beta'}{\beta + p}$$



# Optimal Algorithm

Multilevel Training

What is the **OPTIMAL** machine learning algorithm?



What if the two lines coincide?

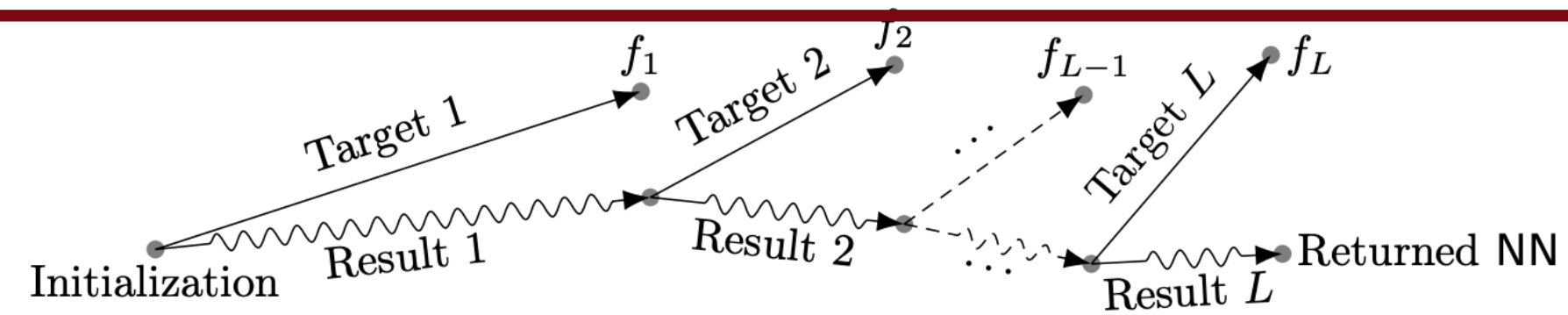


Only  $O(\ln N)$  level is needed





# Matches Empirical Using



Coarse grid

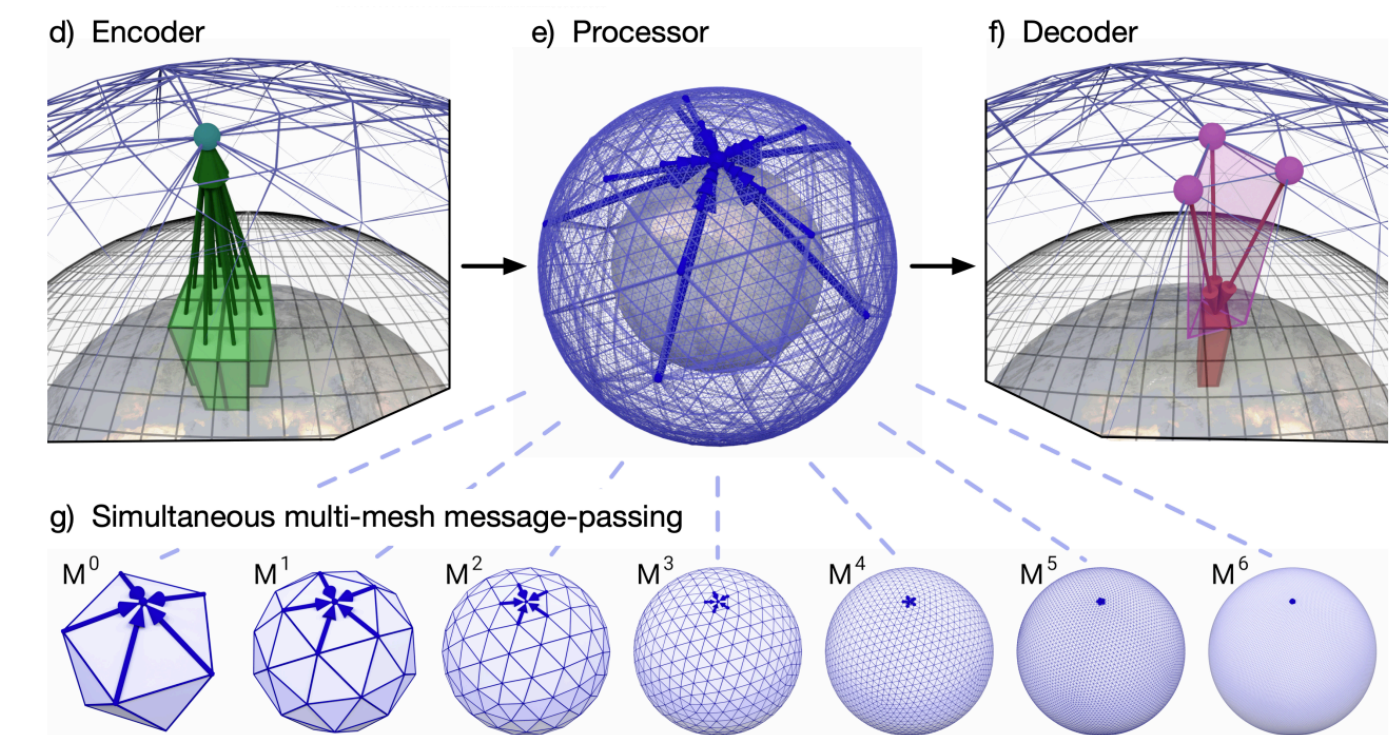
Fine grid

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<sup>\*1</sup>, Alvaro Sanchez-Gonzalez<sup>\*1</sup>, Matthew Willson<sup>\*1</sup>, Peter Wirsberger<sup>\*1</sup>, Meire Fortunato<sup>\*1</sup>,  
Alexander Pritzel<sup>\*1</sup>, Suman Ravuri<sup>1</sup>, Timo Ewalds<sup>1</sup>, Ferran Alet<sup>1</sup>, Zach Eaton-Rosen<sup>1</sup>, Weihua Hu<sup>1</sup>,  
Alexander Merose<sup>2</sup>, Stephan Hoyer<sup>2</sup>, George Holland<sup>1</sup>, Jacklynn Stott<sup>1</sup>, Oriol Vinyals<sup>1</sup>, Shakir Mohamed<sup>1</sup>  
and Peter Battaglia<sup>1</sup>

<sup>\*</sup>equal contribution, <sup>1</sup>DeepMind, <sup>2</sup>Google

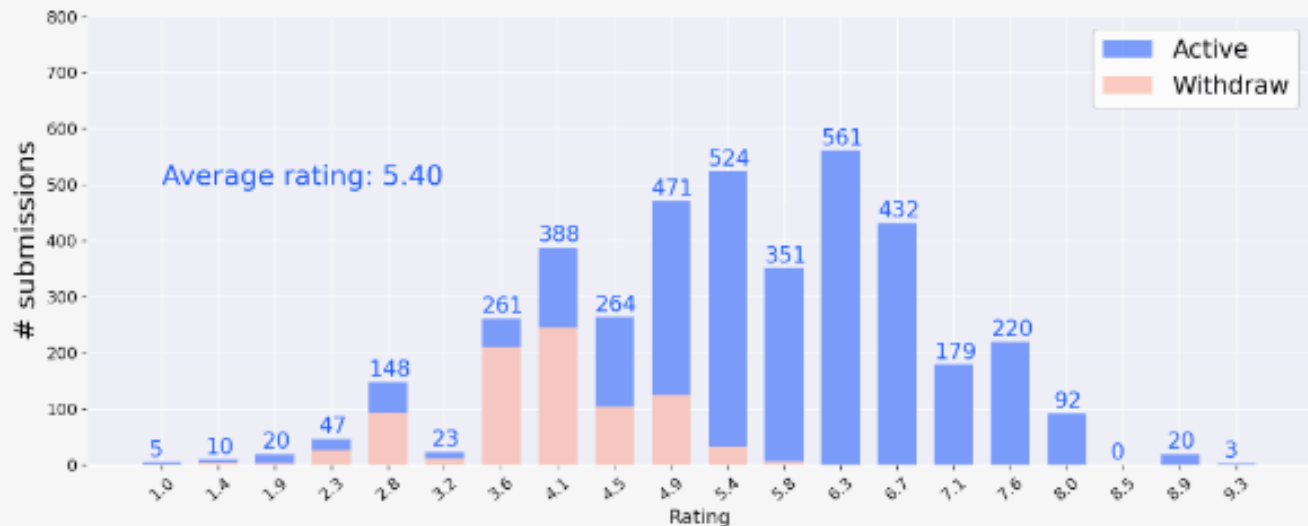
<https://arxiv.org/pdf/2212.12794.pdf>





# 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	<a href="#">Universal Few-shot Learning of Dense Prediction Tasks with Visual Token Matching</a>	8.00	9.33	0.94	1.33	10, 8, 6 10, 8, 10
2	<a href="#">Emergence of Maps in the Memories of Blind Navigation Agents</a>	8.50	9.00	1.00	0.50	8, 8, 8, 10 8, 8, 10, 10
3	<a href="#">Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning</a>	8.25	9.00	1.00	0.75	8, 10, 10, 5 8, 10, 10, 8
4	<a href="#">Minimax Optimal Kernel Operator Learning via Multilevel Training</a>	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 lead 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  $\theta$  in model  $A_\theta$   
*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]**

[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]

Recover parameter  $\theta$  in model  $A_\theta$   
*E.g. Drift, Diffusion Strength*



From data pair  $\{u_i, f_i\}$   
or Learning/Functional data analysis”

Methodology

[Buntin-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]**

[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?



# 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:  $\Delta u = f$

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

**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 Auction*  
 [Guo-Hu-Xu-Zariphopou...]

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

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

Main Idea

Change solving the model to solving a minimization problem

Example:  $\Delta u = f$

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

**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:  $\Delta u = f$

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

“implicit Sobolev acceleration”

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

**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]

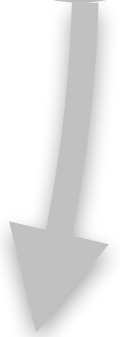
## Main Idea

Change solving the model to solving a minimization problem

Example:  $\Delta u = f$

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

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



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:  $\Delta u = f$

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



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

“Differential operator preconditions the kernel integral operator”



# Research Overview

yplu@stanford.edu

$$Au = f$$

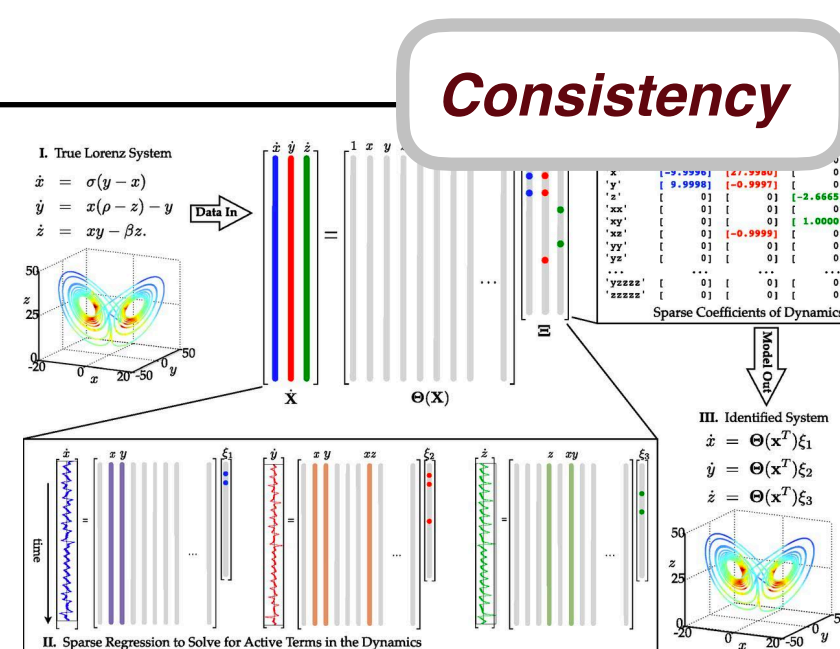
Reconstruct  $u$  with observation of  $f$ :  $\{x_i, f(x_i)\}$

Recover parameter  $\theta$  in Model  $A_\theta$

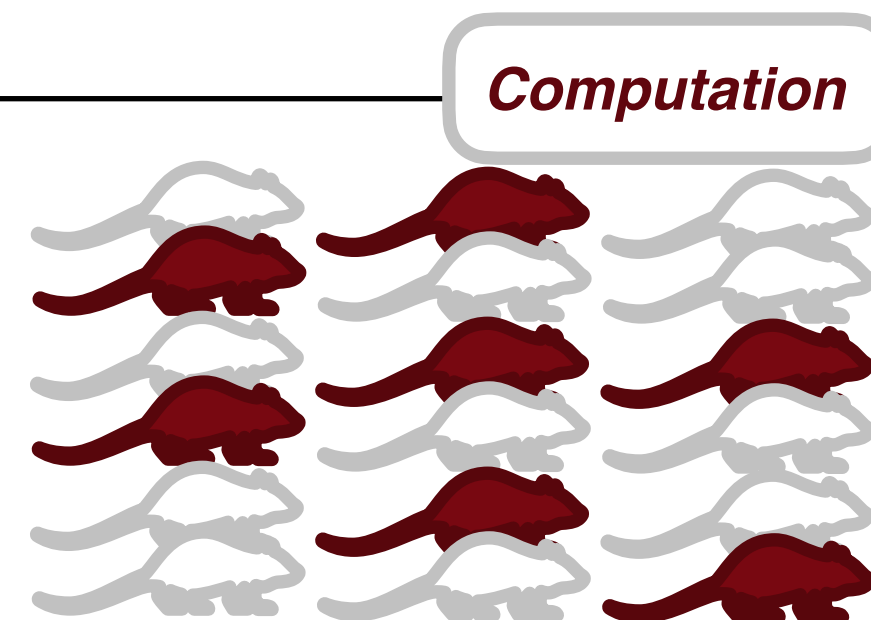
Learn the model  $A$  from data pair  $\{u_i, f_i\}$

## Interaction between model and data

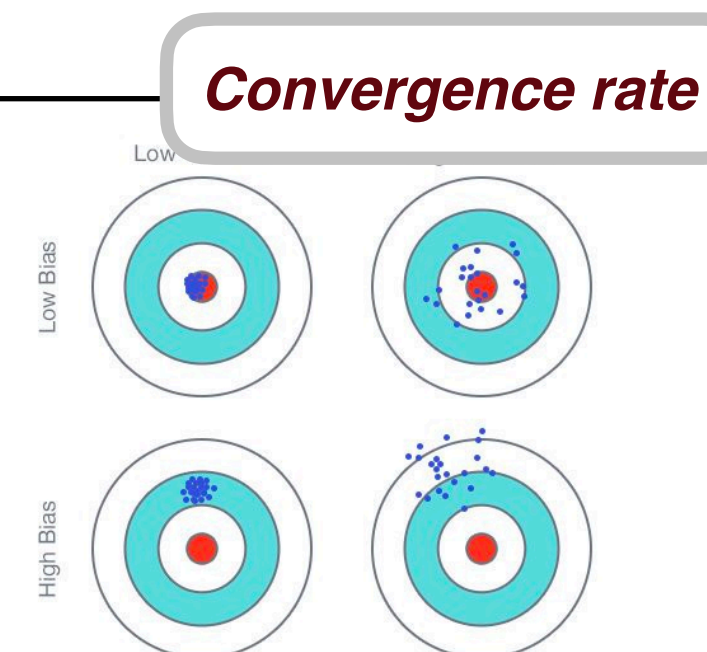
Rough Modeling



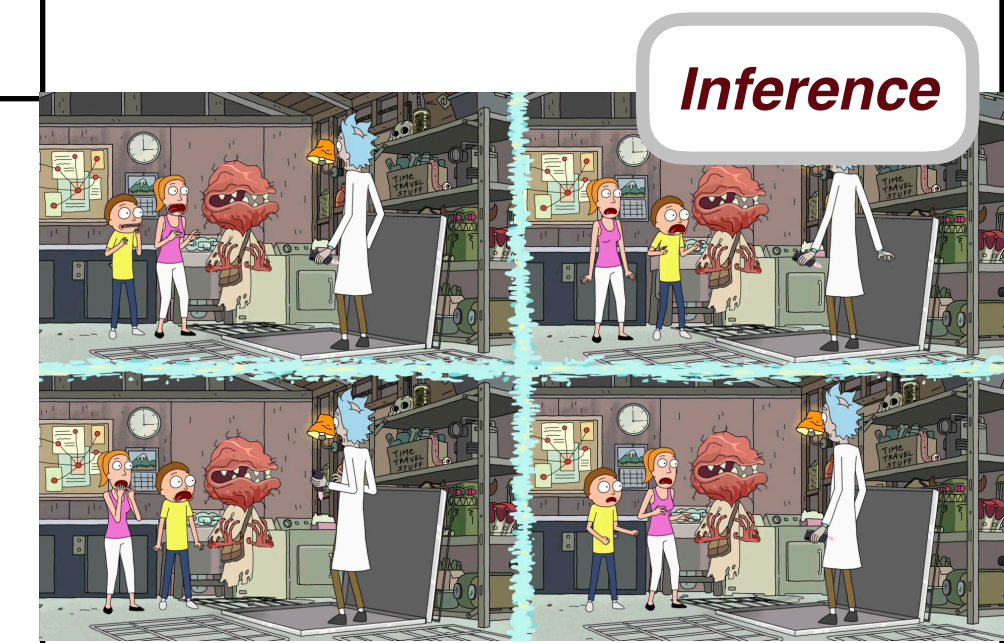
Experiment Design



Model Learning



Uncertainty Quantification





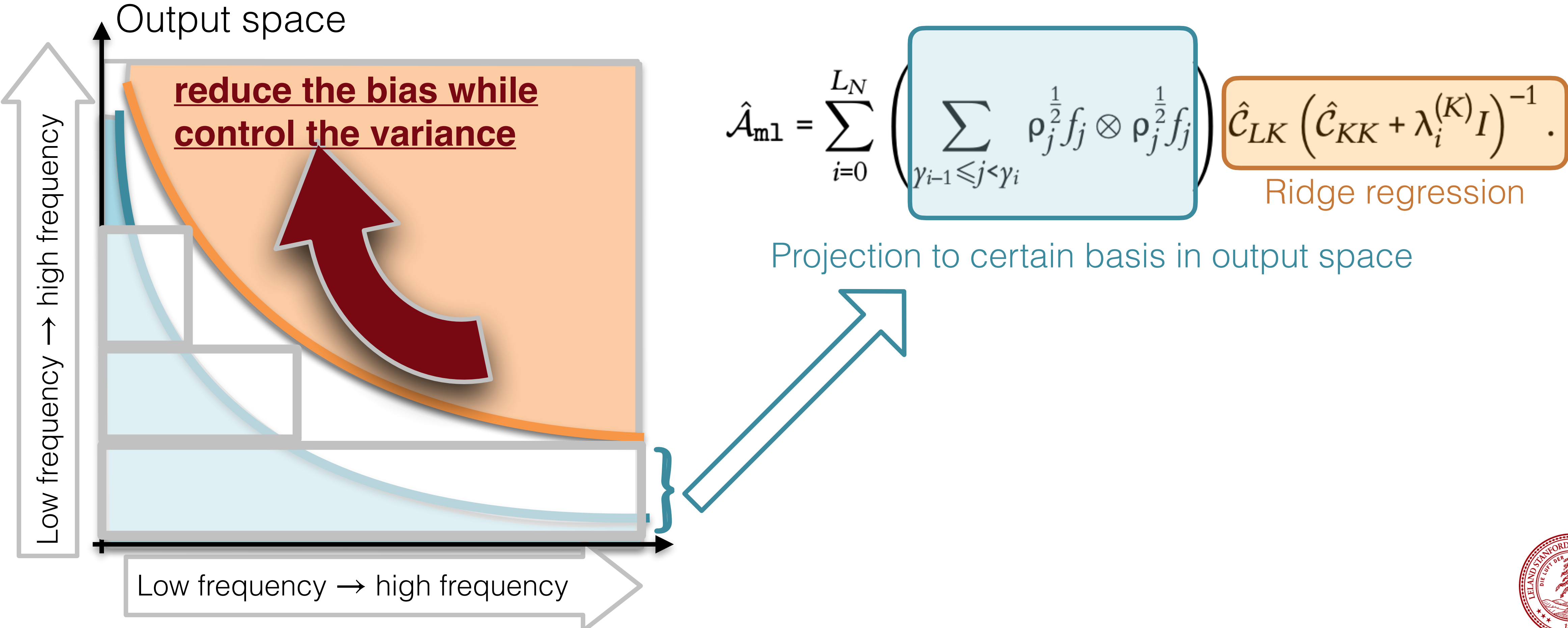


Contact: [yplu@stanford.edu](mailto:yplu@stanford.edu)



# Optimal Algorithm

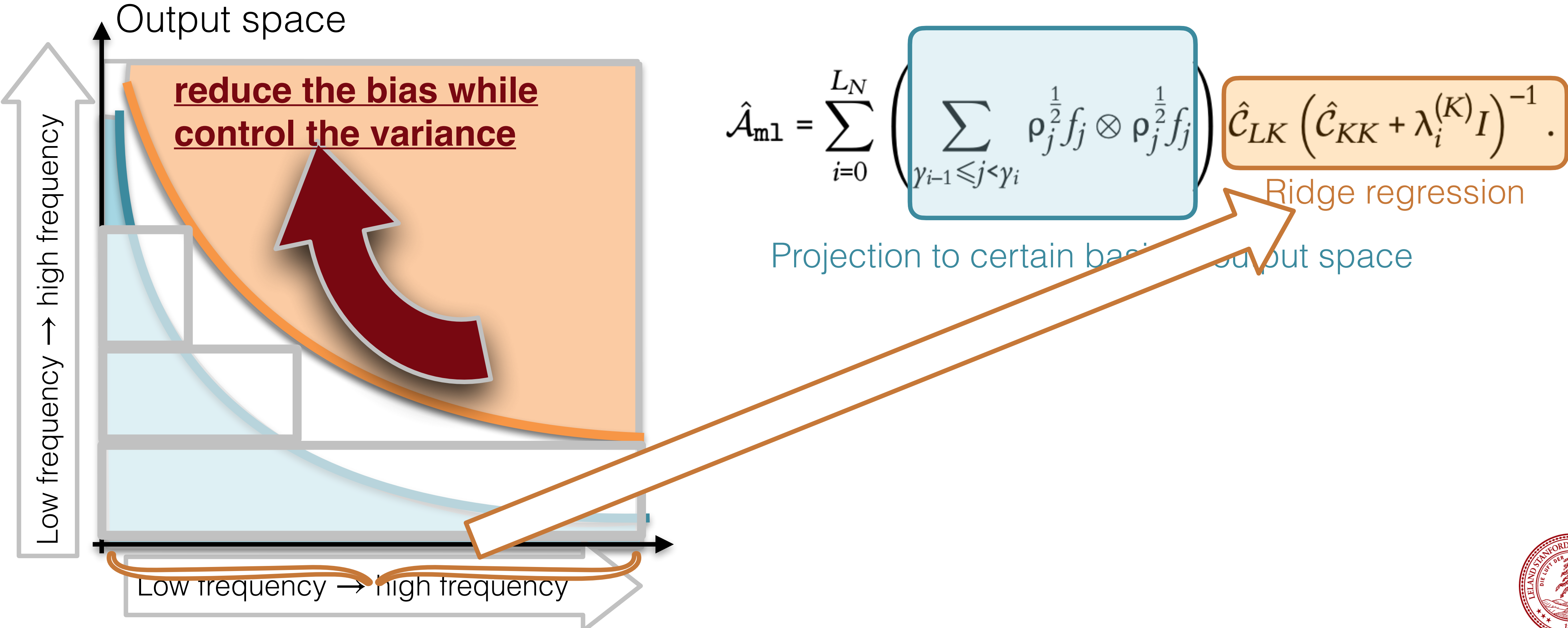
What is the **OPTIMAL** machine learning algorithm?





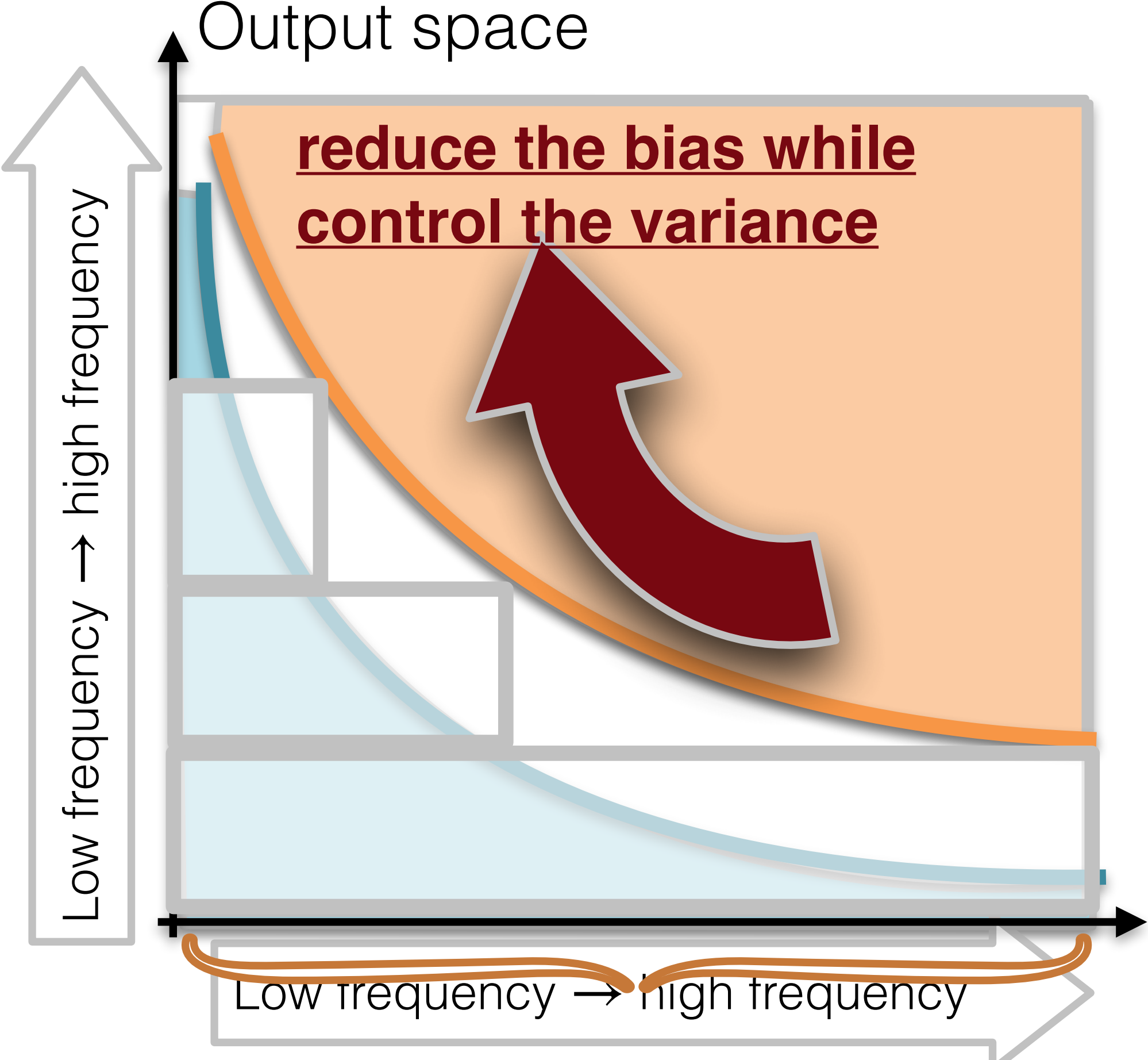
# Optimal Algorithm

What is the **OPTIMAL** machine learning algorithm?



# Optimal Algorithm

What is the **OPTIMAL** machine learning algorithm?

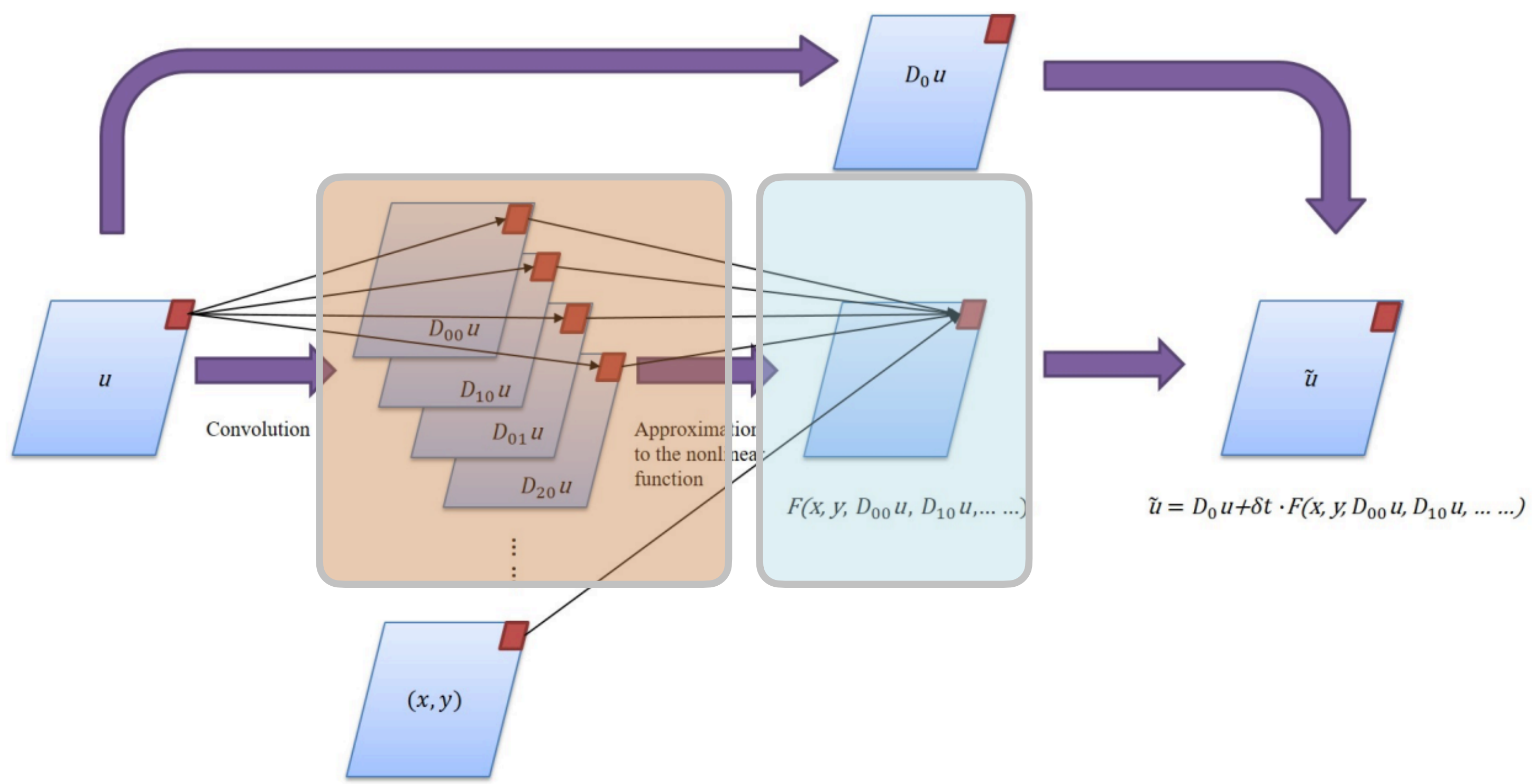


$$\hat{A}_{m1} = \sum_{i=0}^{L_N} \left( \sum_{\gamma_{i-1} \leq j < \gamma_i} \rho_j^{\frac{1}{2}} f_j \otimes \rho_j^{\frac{1}{2}} f_j \right) \hat{C}_{LK} \left( \hat{C}_{KK} + \lambda_i^{(K)} I \right)^{-1}$$

Ensemble different levels



# Algorithmic Literature Overview



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

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:  $n^{-\frac{2s}{d+2s}}$  “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

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

## 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?

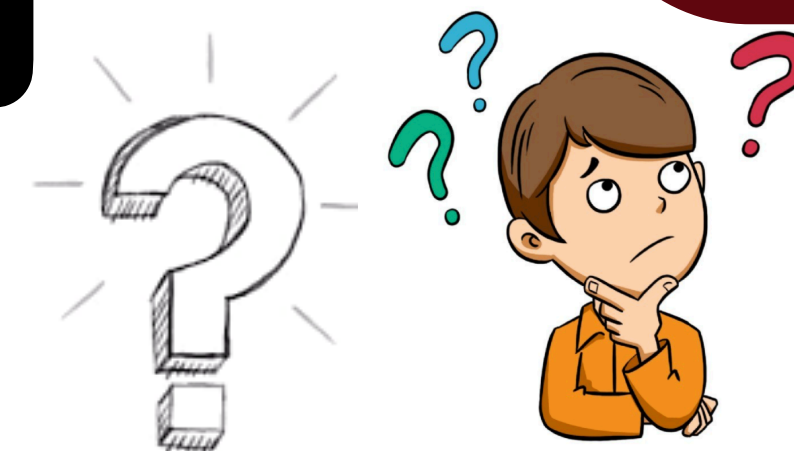
For  $\beta = 2$

PINN



For  $\beta = 1$

DRM



Artifact of analysis?  
NN ansatz? Objective?



# Is Deep Ritz Optimal? A Fourier View

$$Au = f$$

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



Naive way to do this?

Naive Estimator is **Optimal** with proper selection of  $S$

# Is Deep Ritz Optimal? A Fourier View

$$Au = f$$

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



How is naive estimator different from DRM?

**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

$$Au = f$$

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

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

**Naive**

**DRM**

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

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

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

$$\hat{A} = \begin{pmatrix} \sum_i \nabla \phi_j(x_i) \nabla \phi_k(x_i) \\ \sum_i \phi_j(x_i) \phi_k(x_i) \end{pmatrix}_{j,k} +$$

**Introduce further variance**



# Is Deep Ritz Optimal? A Fourier View

$$Au = f$$

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

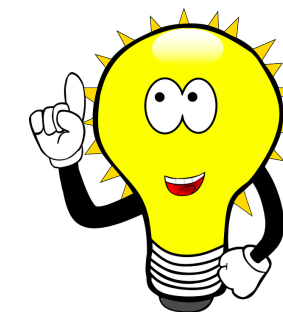
**DRM Estimator**  $\hat{u} = \sum_{|z| < S} \hat{u}_z^F \phi_z$  and plug in

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

DRM discretized

$$\nabla \cdot \nabla$$

But not  $\Delta$



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	X	✓
NN	X	?



# DRM or PINN

Which one optimizes faster?



$$\text{DRM } \min \int |\nabla u|^2 - 2uf$$
$$\text{PINN } \min \|\Delta u - f\|^2$$

Pre-ml Experience:  
Double the condition number



# DRM or PINN

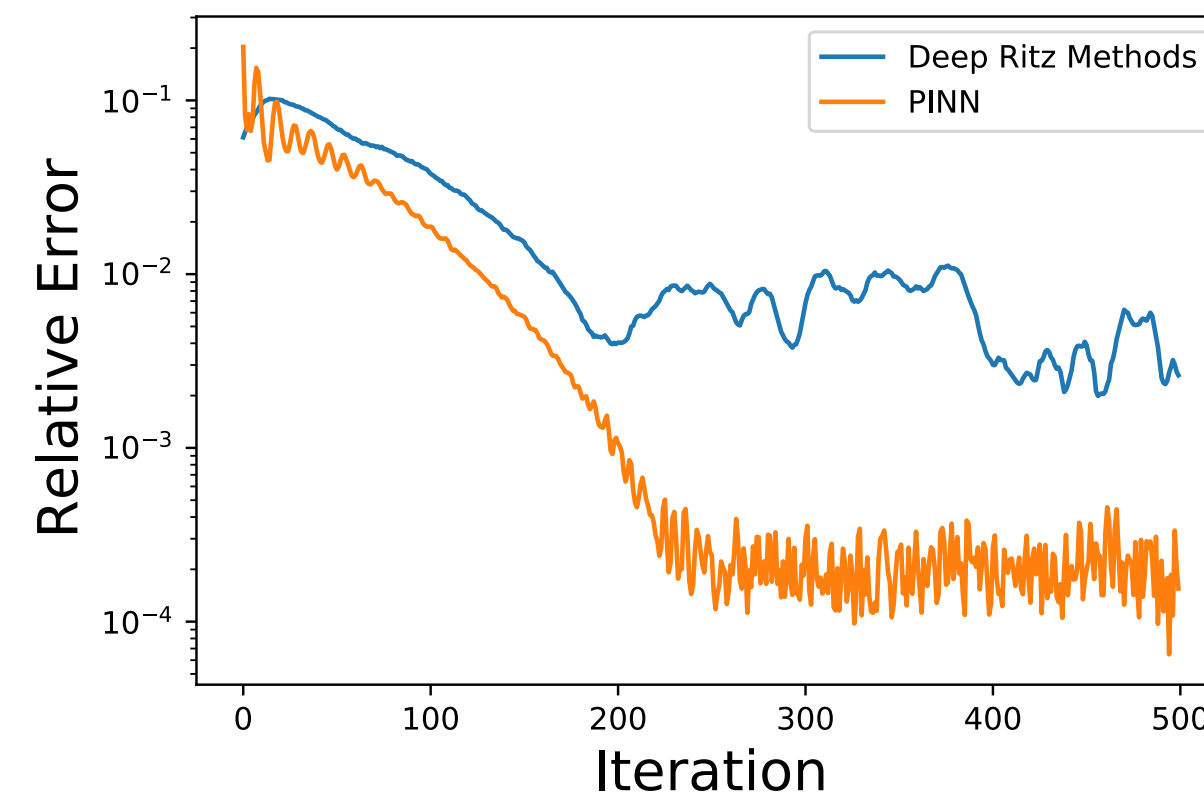
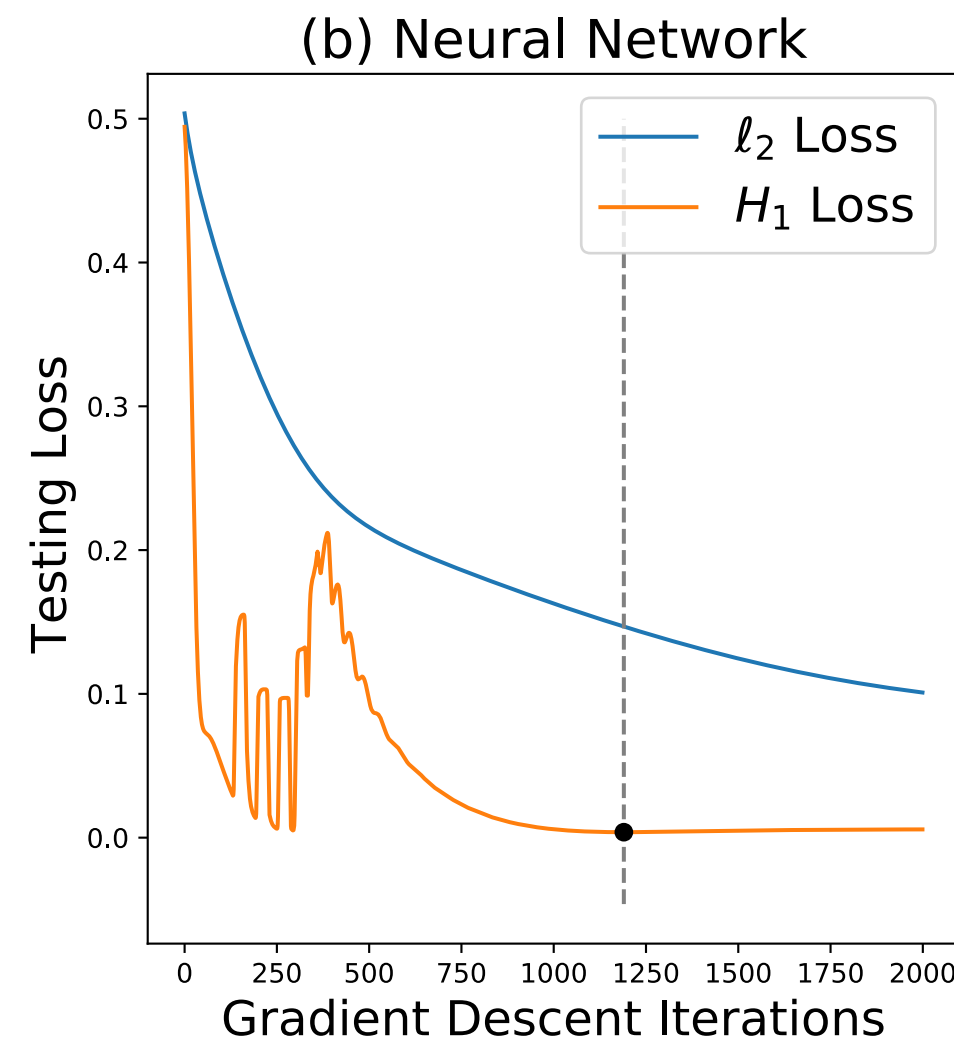
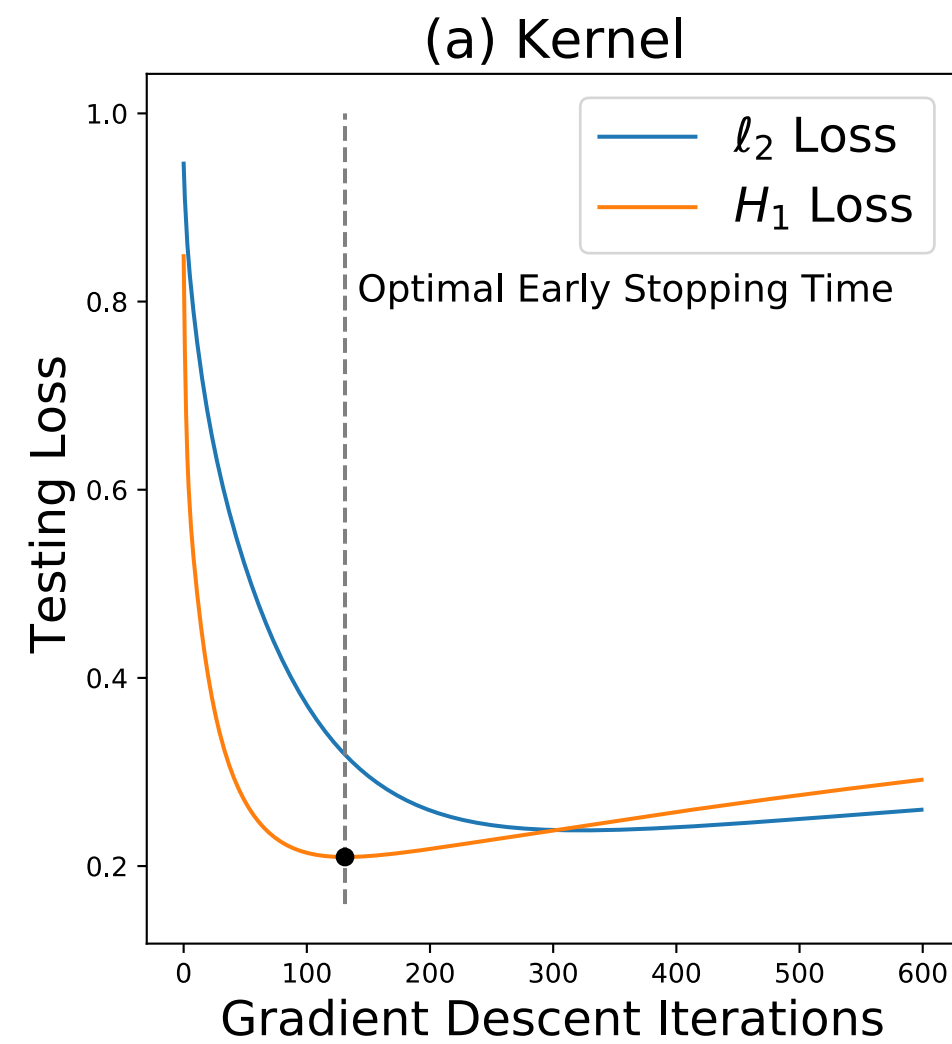
Which one optimizes faster?



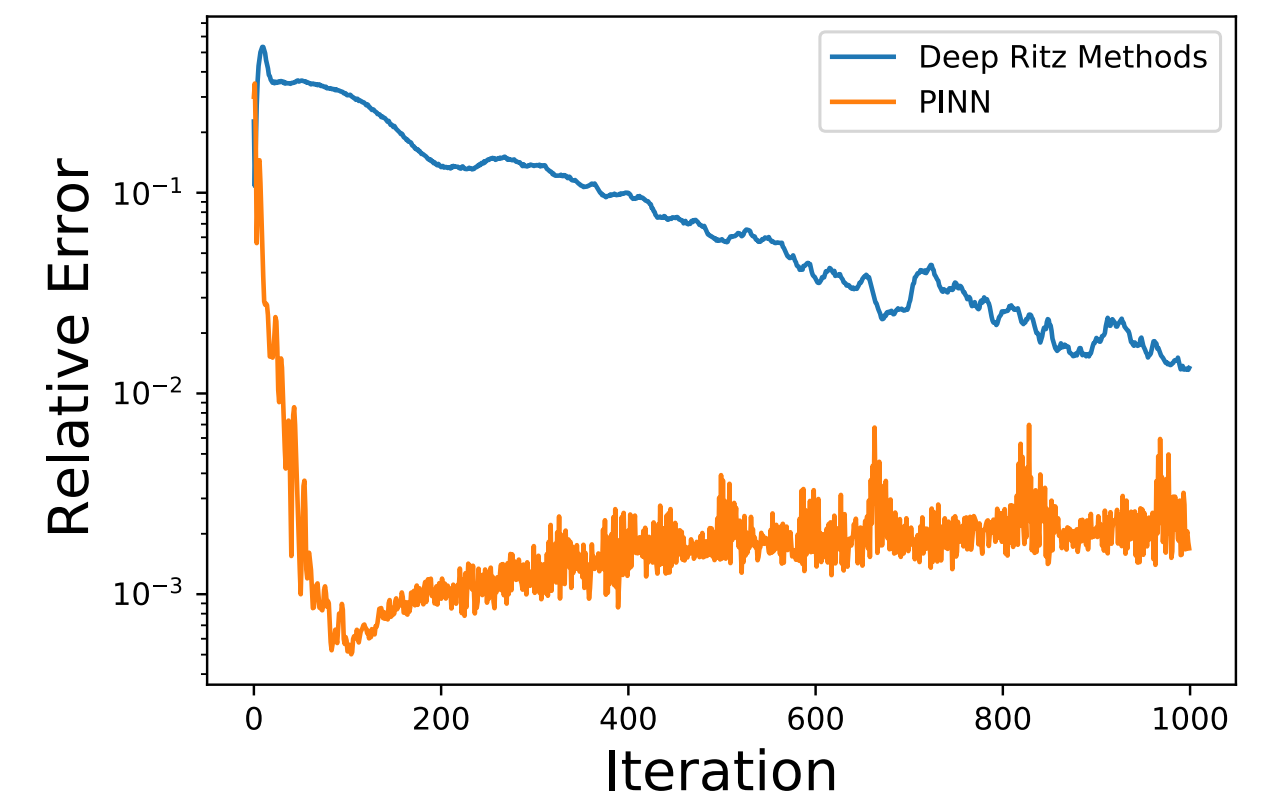
**DRM**  $\min \int |\nabla u|^2 - 2uf$

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

Pre-ml Experience:  
Double the condition number



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



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

Sobolev Training

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  $\frac{1}{2} \langle f, \mathcal{A}_1 f \rangle - \langle u, \mathcal{A}_2 f \rangle$

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

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

Gradient Descent

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

Differential operator    Kernel integral operator



# 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

