When can Regression-Adjusted Control Variates Help?

Rare Events, Sobolev Embedding and Minimax Optimality

Haoxuan Chen

Institute for Computational and Mathematical Engineering (ICME)
Stanford University

Joint work with Prof. Jose Blanchet, Dr. Yiping Lu, and Prof. Lexing Ying

Stanford | Institute for Computational & Mathematical Engineering

NYU | COURANT INSTITUTE OF MATHEMATICAL SCIENCES

Agenda

Part I: Intro to Regression-Adjusted Control Variates (RACV)

- What is a RACV-based algorithm? How to interpret it as an estimator?
- Applications in various fields (numerical analysis - quadrature rule, trace estimation; statistics - causal inference; machine learning - gradient estimation)

Part II: Information Theoretic Lower Bound

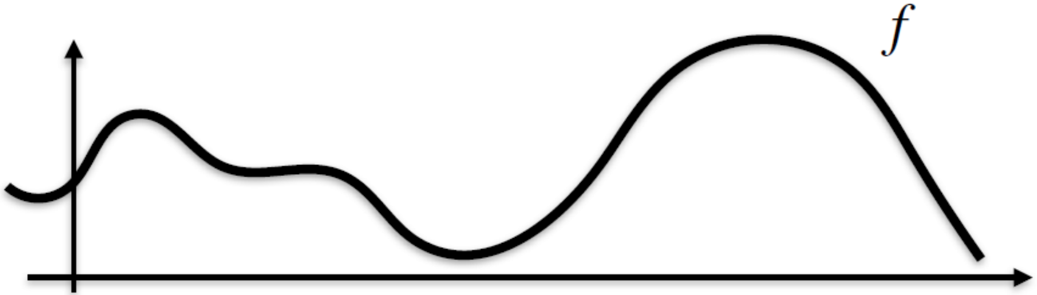
- Recap of non-parametric statistics - what is minimax optimality?
- Information Theoretic Lower Bound for Quadrature Rules

Part III: Minimax Optimal Upper Bound

- Claim: use different algorithms for functions of varying degrees of smoothness
- Proof Sketch: Use Sobolev Embedding Theorem appropriately

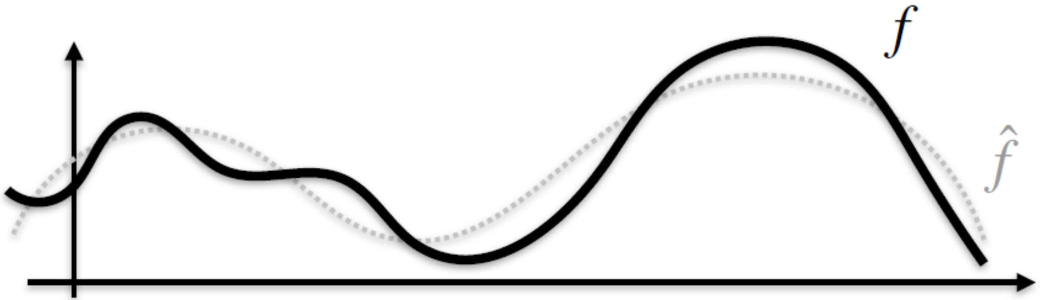
Quadrature Rule

Aim Estimate $\mathbb{E}_P f$



Quadrature Rule

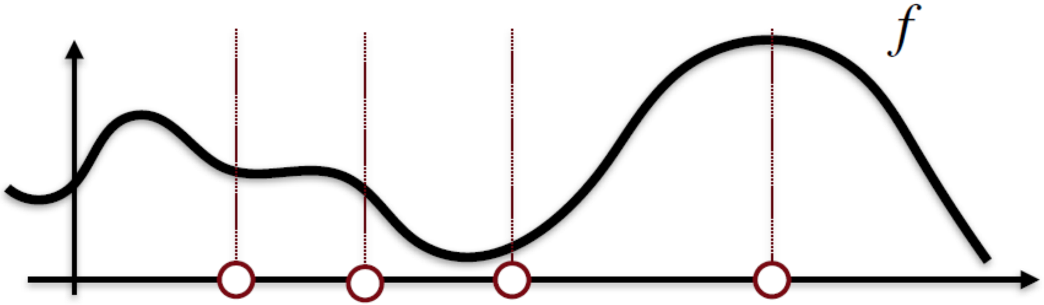
Aim Estimate $\mathbb{E}_P f \approx \mathbb{E}_P \hat{f}$



"piece-wise polynomial"

Quadrature Rule via Monte Carlo

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



Key Idea: Doubly Robust Estimation

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

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

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

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

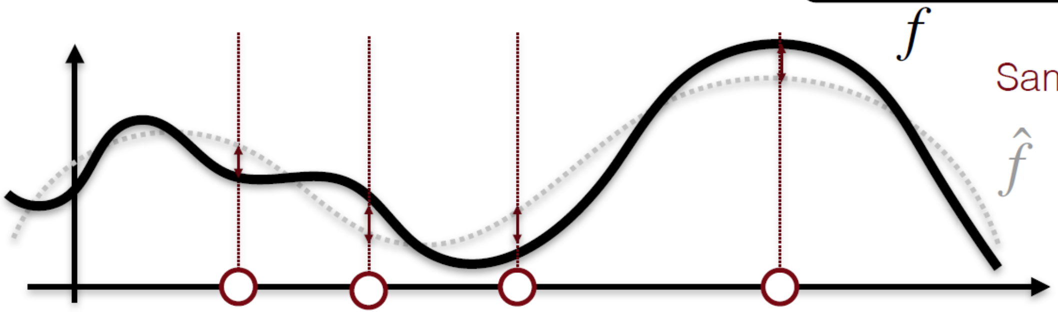
Smaller error

Quadrature Rule

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

Debiasing

"semi-"parametric



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

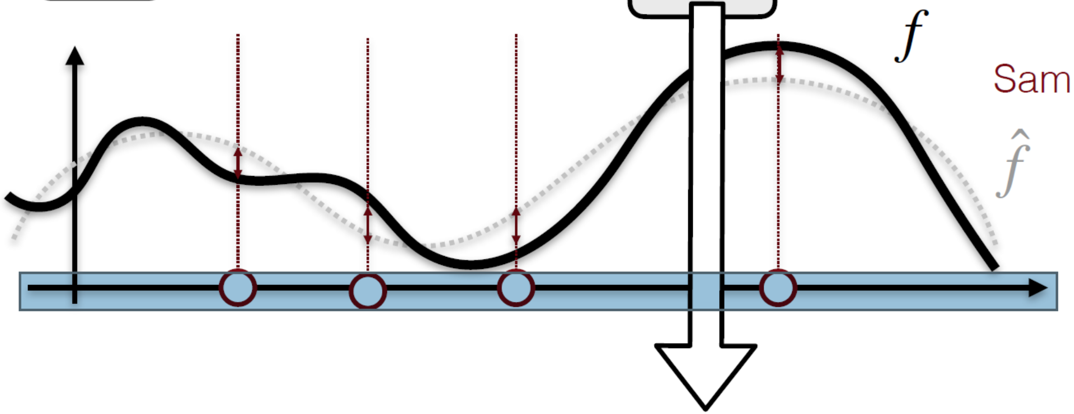
Quadrature Rule

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

Debiasing

“semi-"parametric

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



(nonparametric-)“Regression-adjusted” control variate

Modern Applications of RACV

Numerical Analysis: estimating the trace of a matrix

- Task: given matrix $A \in \mathbb{R}^{d\times d}$ and an oracle for computing matrix-vector multiplication (at most m queries), estimate $\mathrm{tr}(A)$
- Naive Monte Carlo Algorithm (Hutch): appropriately sampled $X \in \mathbb{R}^{d\times m}$

$$\mathrm{tr}(A) \approx \frac{1}{m}\sum_{i=1}^{m} x_i^T A x_i = \mathrm{tr}(X^T A X)$$

- RACV-based Algorithm: appropriately sampled orthonormal matrices $Q, G \in \mathbb{R}^{d\times \frac{m}{3}}$

$$E_1 = \mathrm{tr}(Q^T A Q),\ \ E_2 = \mathrm{tr}(G^T(I - QQ^T)A(I - QQ^T)G)$$

Estimator $E_1 + E_2 \approx \mathrm{tr}(A)$ (Hutch++)

(Lin 2017; Mewyer-Musco-Musco-Woodruff 2020)

Modern Applications of RACV

Machine Learning: estimation of gradient

- Task: estimate the gradient $\nabla_\theta \mathbb{E}_{q_\theta}[f(x)] = \mathbb{E}_{q_\theta}[f(x)\nabla_\theta \log q_\theta(x)]$ via $\{x_i\}_{i=1}^n$
- Naive Monte-Carlo estimator: $\frac{1}{n}\sum_{i=1}^n f(x_i)\nabla_\theta \log q_\theta(x_i)$
- Leave-one-out estimator:

$$\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - \frac{1}{n-1}\sum_{j\neq i} f(x_j)\right)\nabla_\theta \log q_\theta(x_i)$$

- RACV-enhanced leave-one-out estimator: $E_1 + E_2$

$$E_1 = \frac{2}{n}\sum_{i=1}^{\frac{n}{2}}\left((f(x_i) - b(x_i)) - \frac{1}{\frac{n}{2}-1}\sum_{1\leqslant j\neq i\leqslant \frac{n}{2}}(f(x_j) - b(x_j))\right)\nabla_\theta \log q_\theta(x_i)$$

$$E_2 = \frac{2}{n}\sum_{i=\frac{n}{2}+1}^{n}\left(b(x_i) - \frac{1}{\frac{n}{2}-1}\sum_{\frac{n}{2}<j\neq i\leqslant n} b(x_j)\right)\nabla_\theta \log q_\theta(x_i)$$

(Shi-Zhou-Hwang-Titsias-Mackey 2022)

Modern Applications of RACV

Other applications include:

- Conformal Prediction (Romano-Patterson-Candes 2019)
- Causal Inference (Jordan-Wang-Zhou 2022; Angelopoulos-Bates-Fannjiang-Jordan-Zrnic 2023)

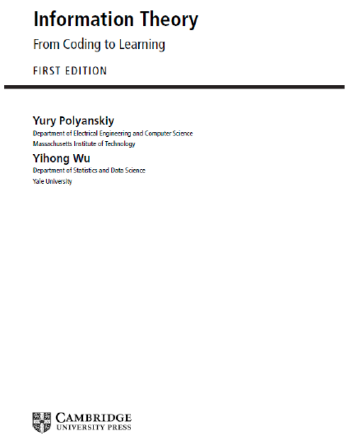
Unifying Framework

- Step I: Get a "coarse" estimator via a subset of the data
- Step II: Obtain a "finer" estimator via calibration on the remaining dataset (estimating the bias)

Machine Learning Research

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

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 selection of f"

Best Estimator

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 gap

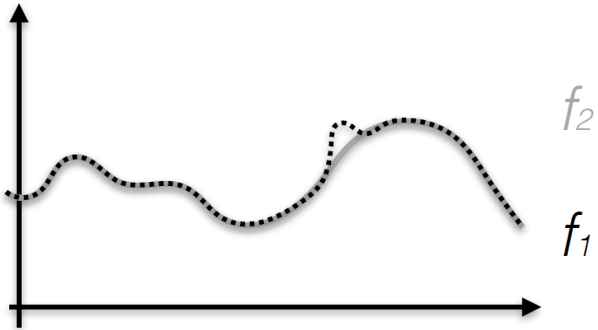
The gap is not distinguishable

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 gap

The gap is not distinguishable

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 $W^{s,p}$?

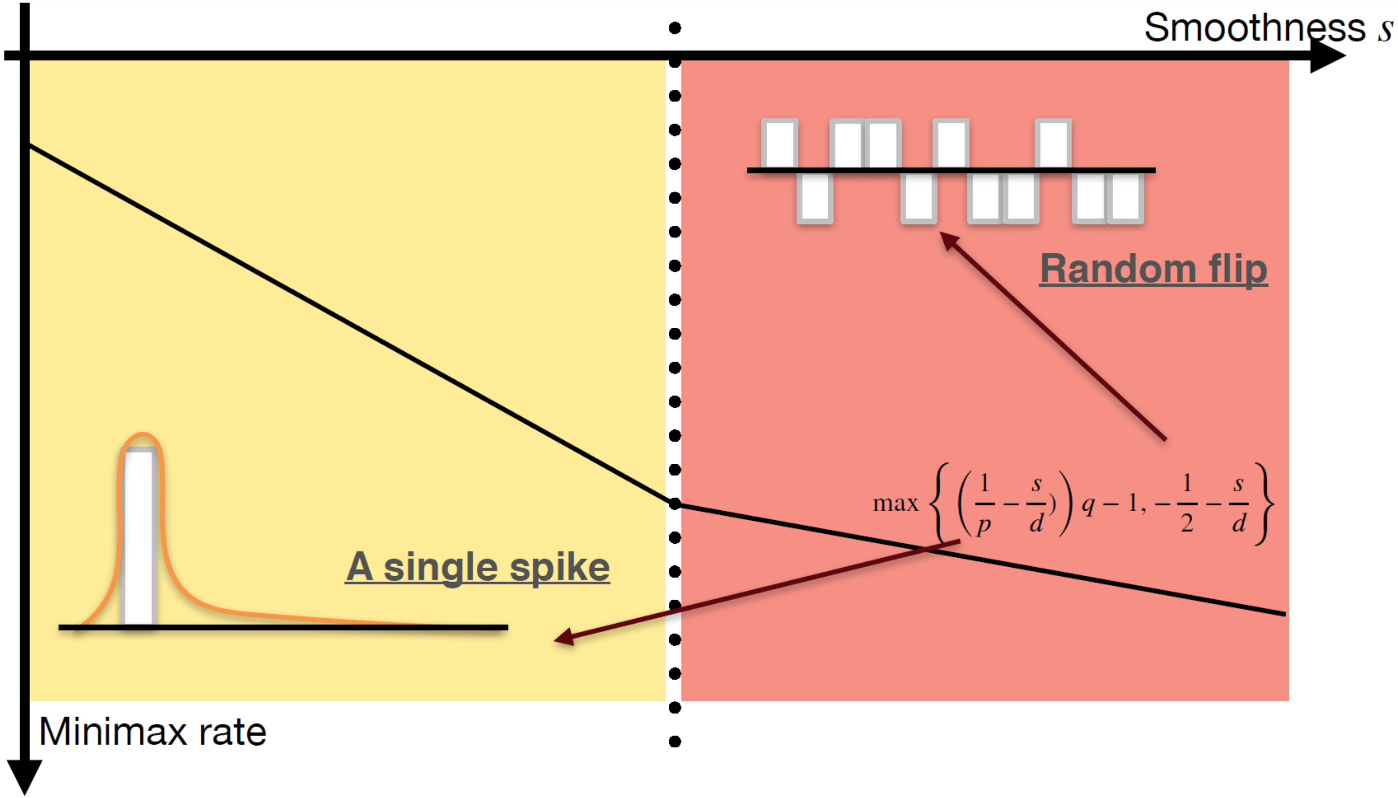
Aim Estimate ~~$\mathbb{E}_P f$~~ $\mathbb{E}_P f^q, f \in W^{s,p}$

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

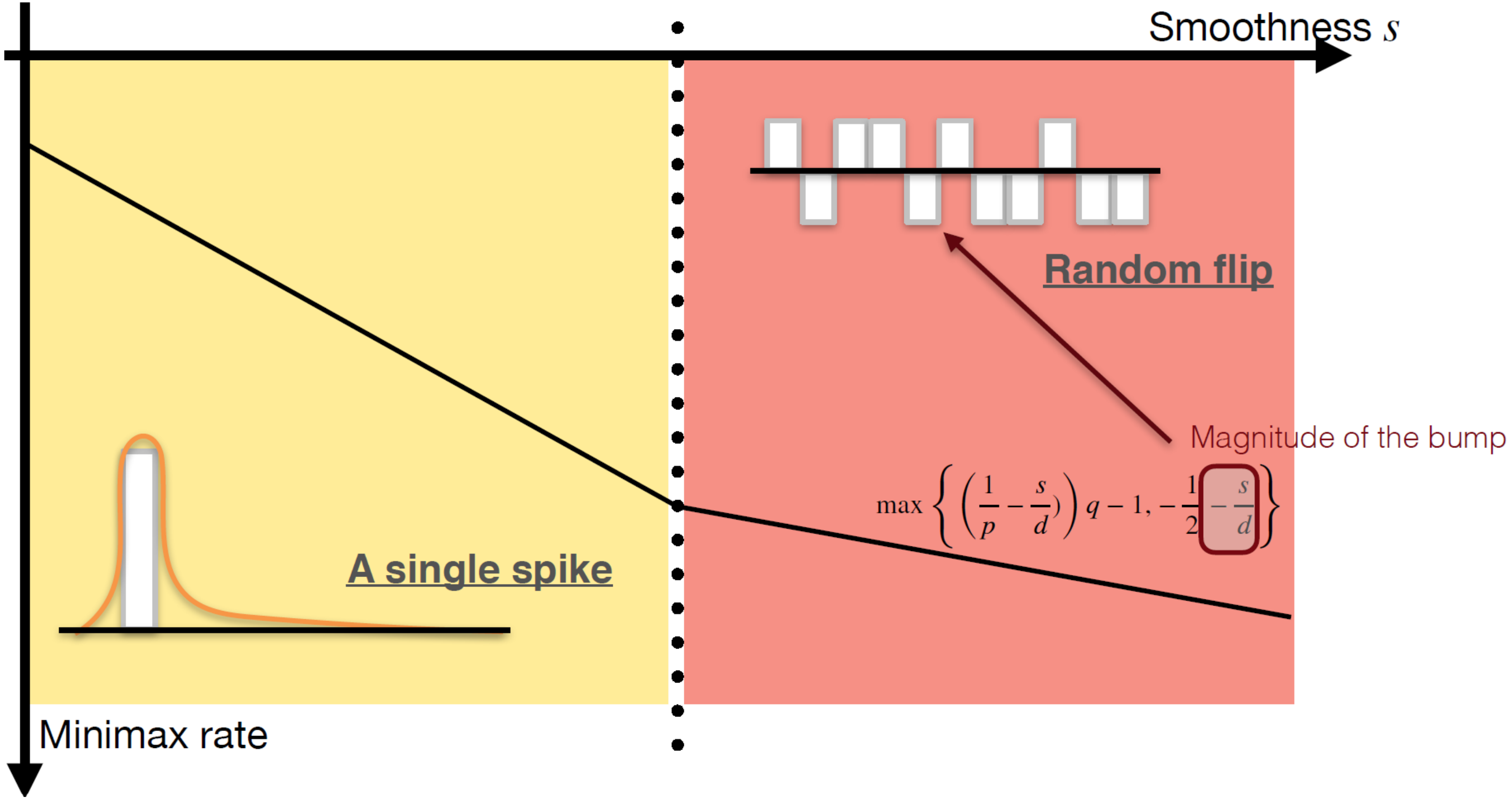
Step 2 $\mathbb{E}_P f^q = \mathbb{E}_P (\hat{f})^q + \mathbb{E}_P (f^q - \hat{f}^q)$

Low order term

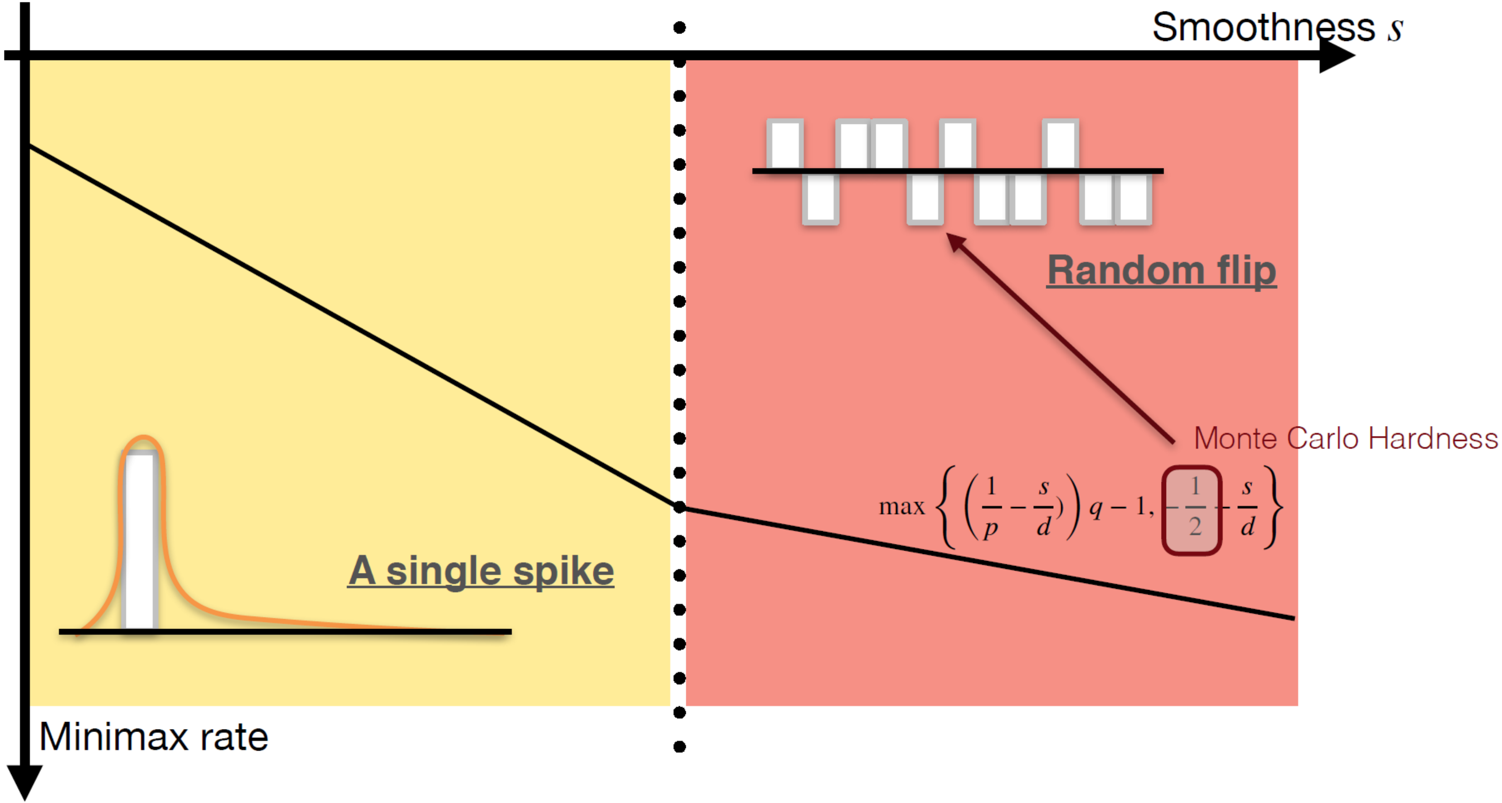
Setting the information theoretical limit



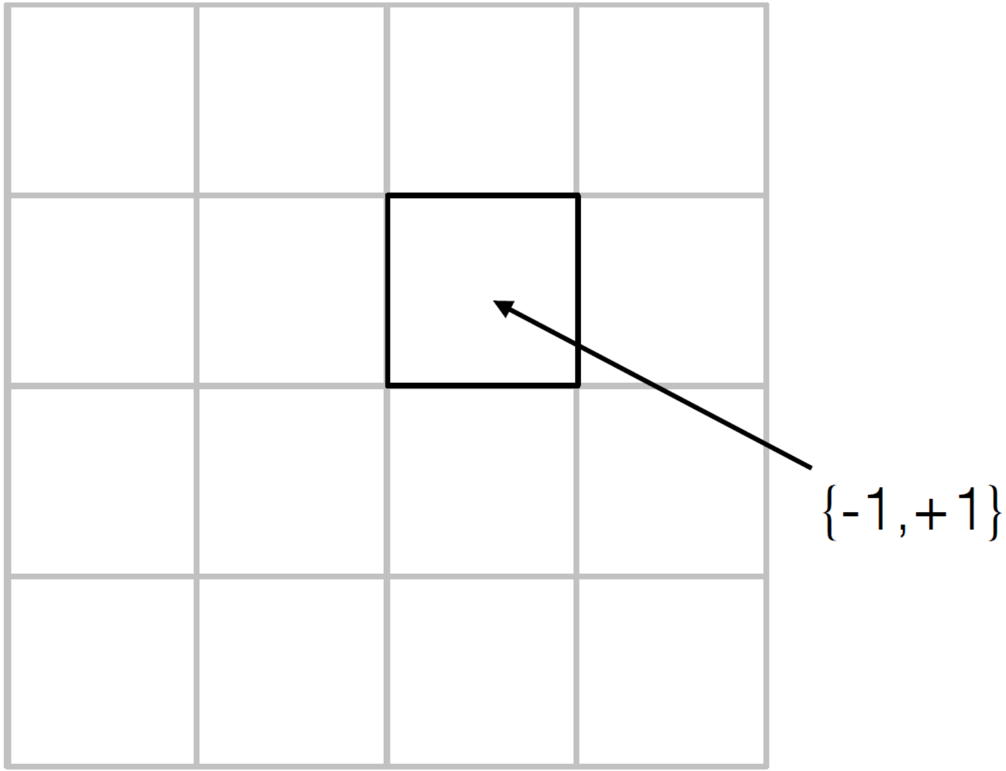
Setting the information theoretical limit



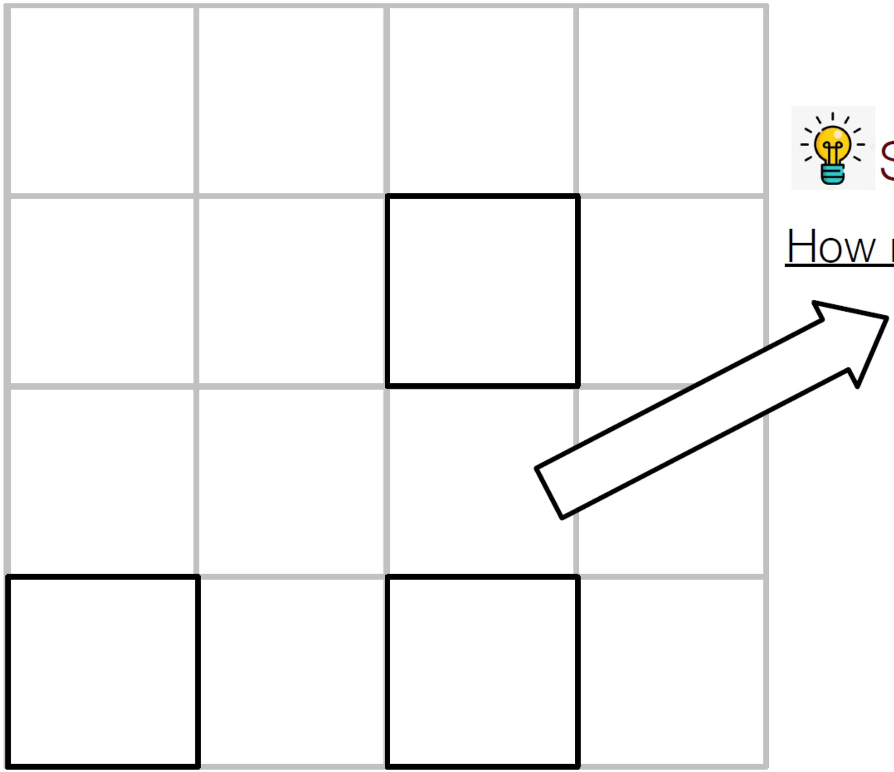
Setting the information theoretical limit



Understanding the hardness in this regime



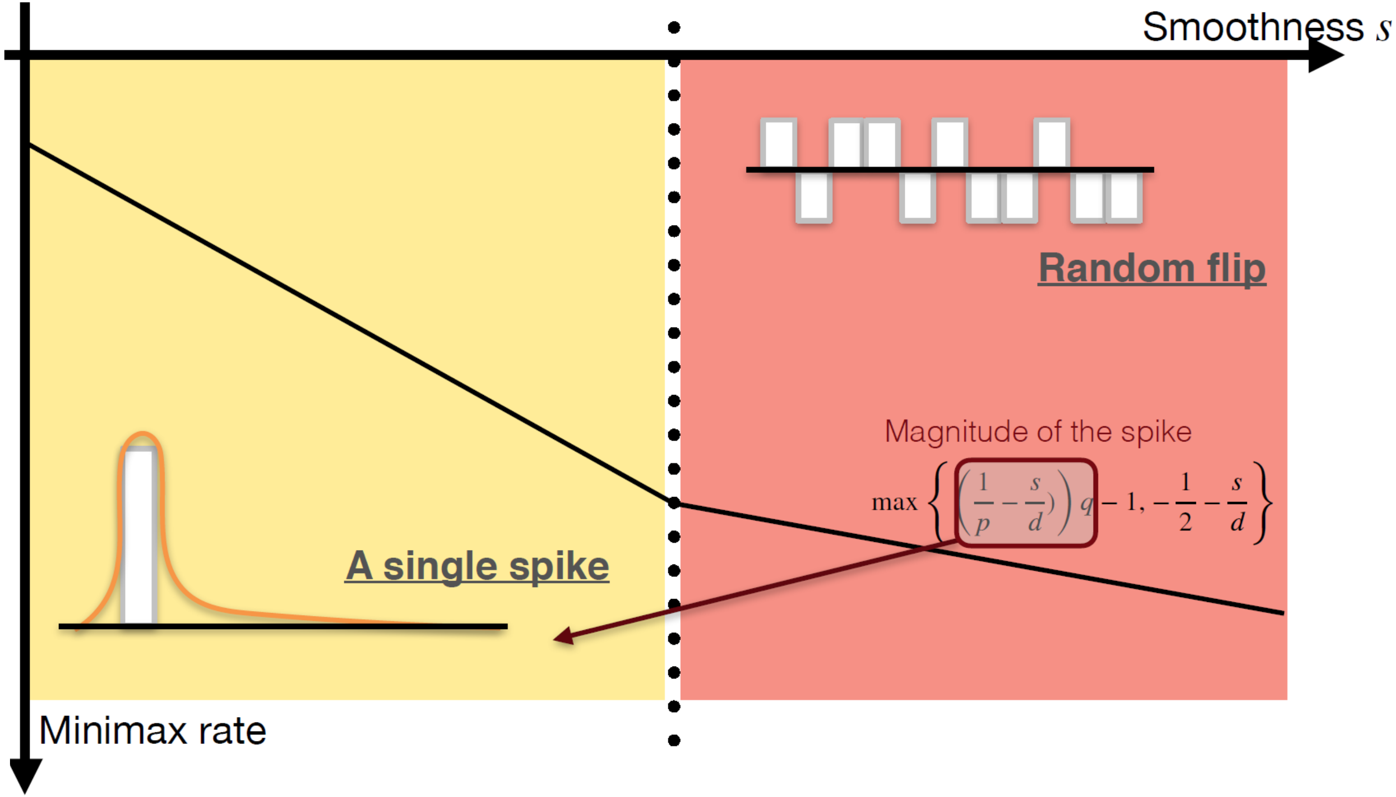
Understanding the hardness in this regime



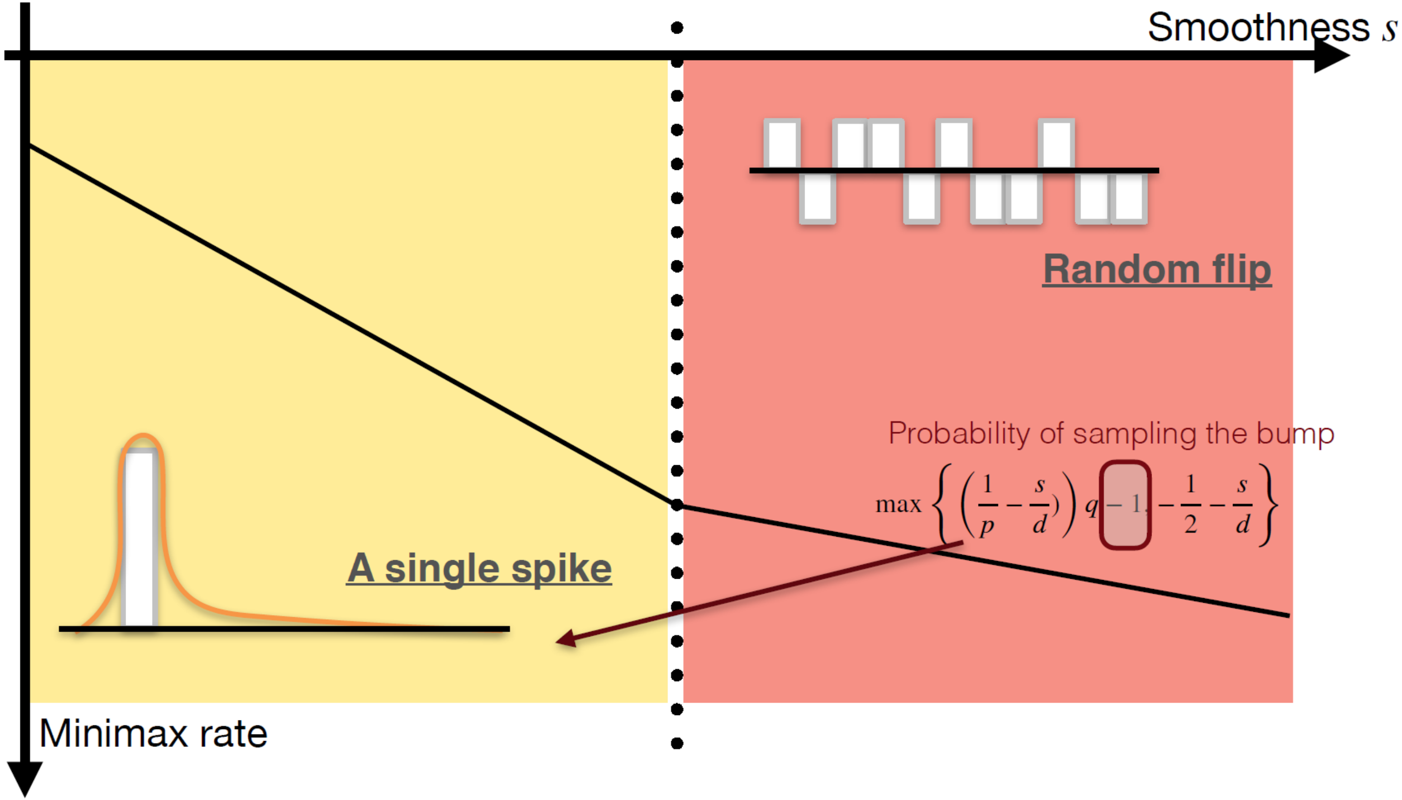
💡 Sample without replacement

How many boxes have quadrature points?

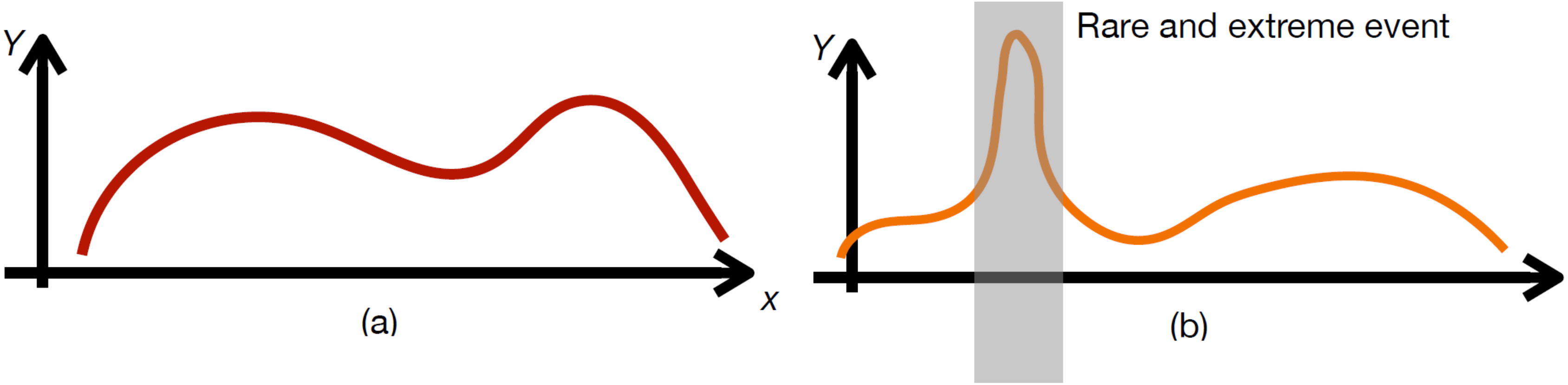
Setting the information theoretical limit



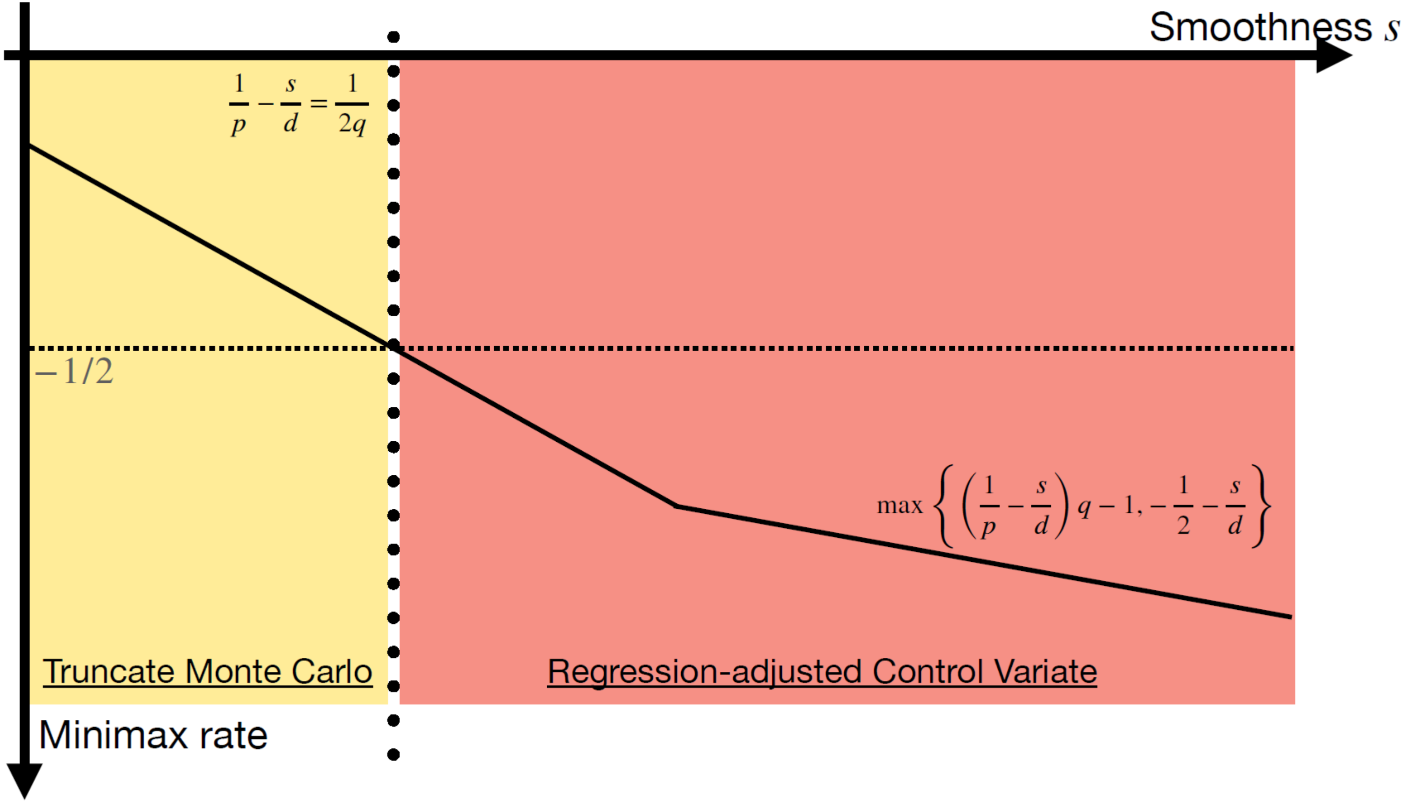
Setting the information theoretical limit



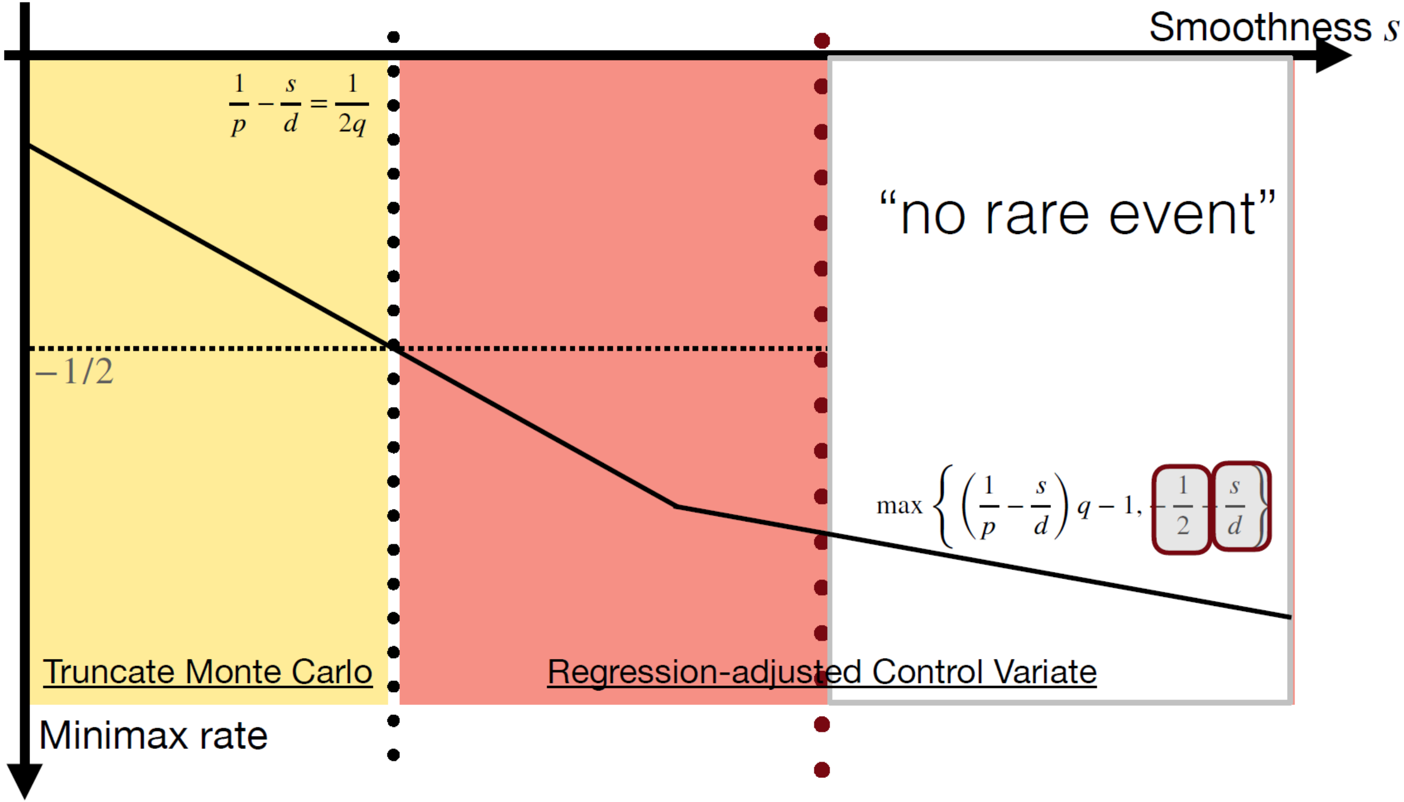
Rare Event and Smoothness…



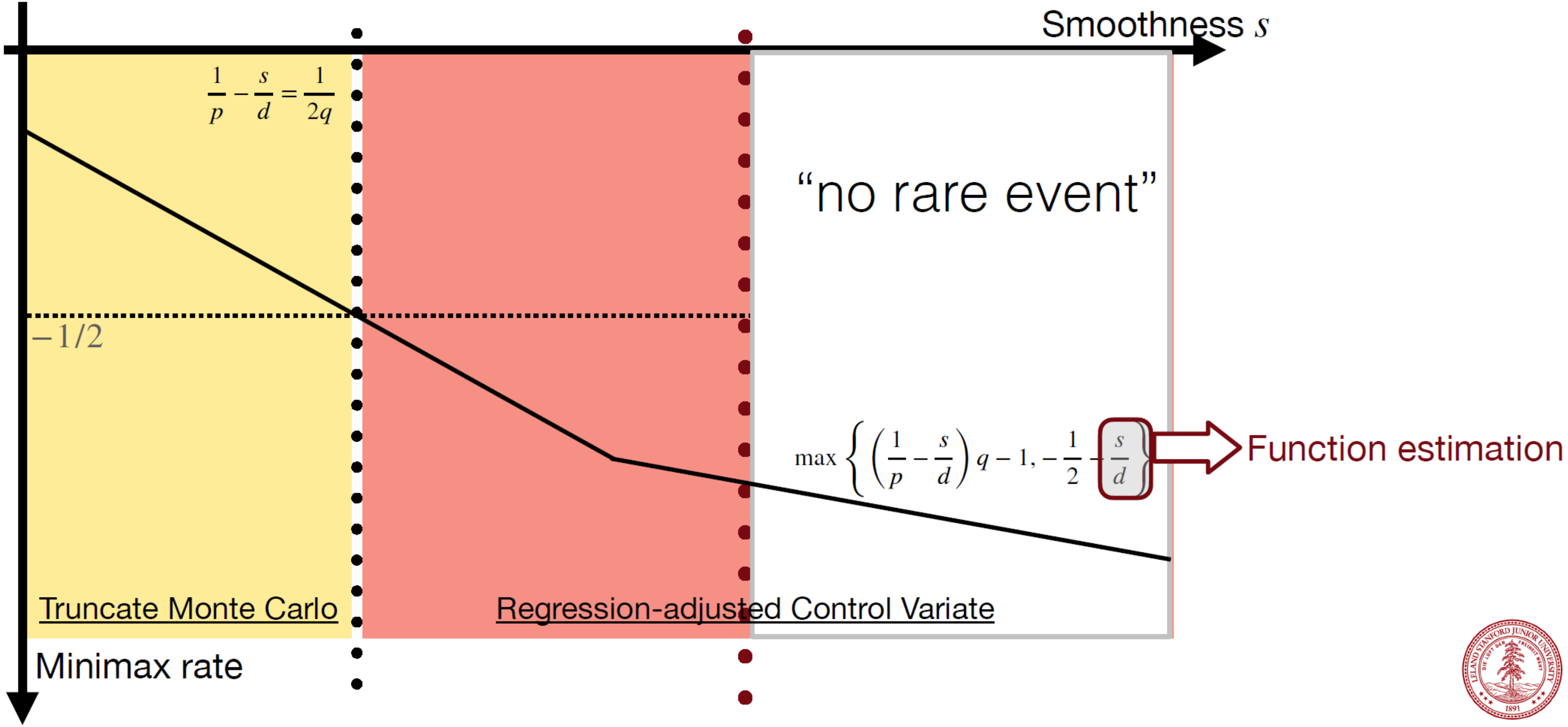
When the control variate helps



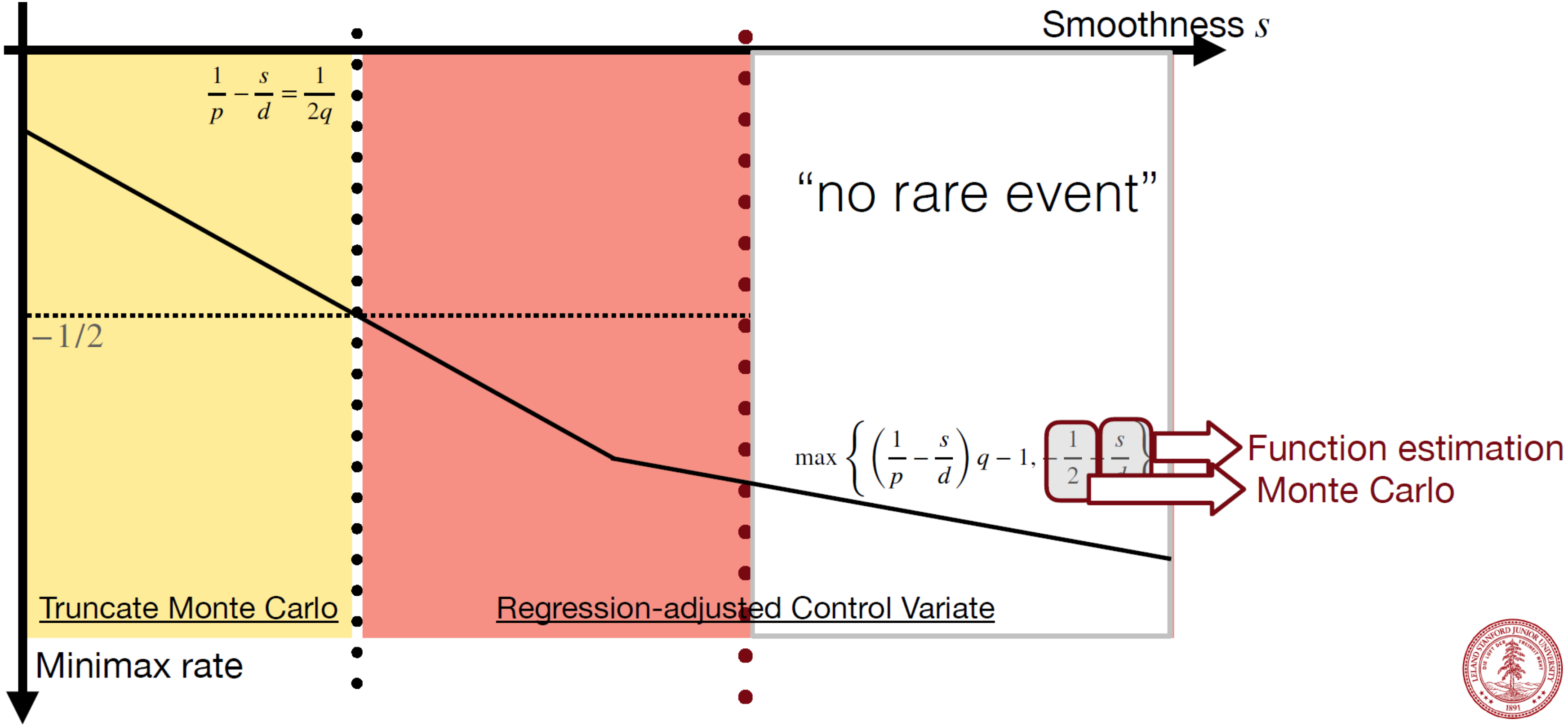
When the control variate helps



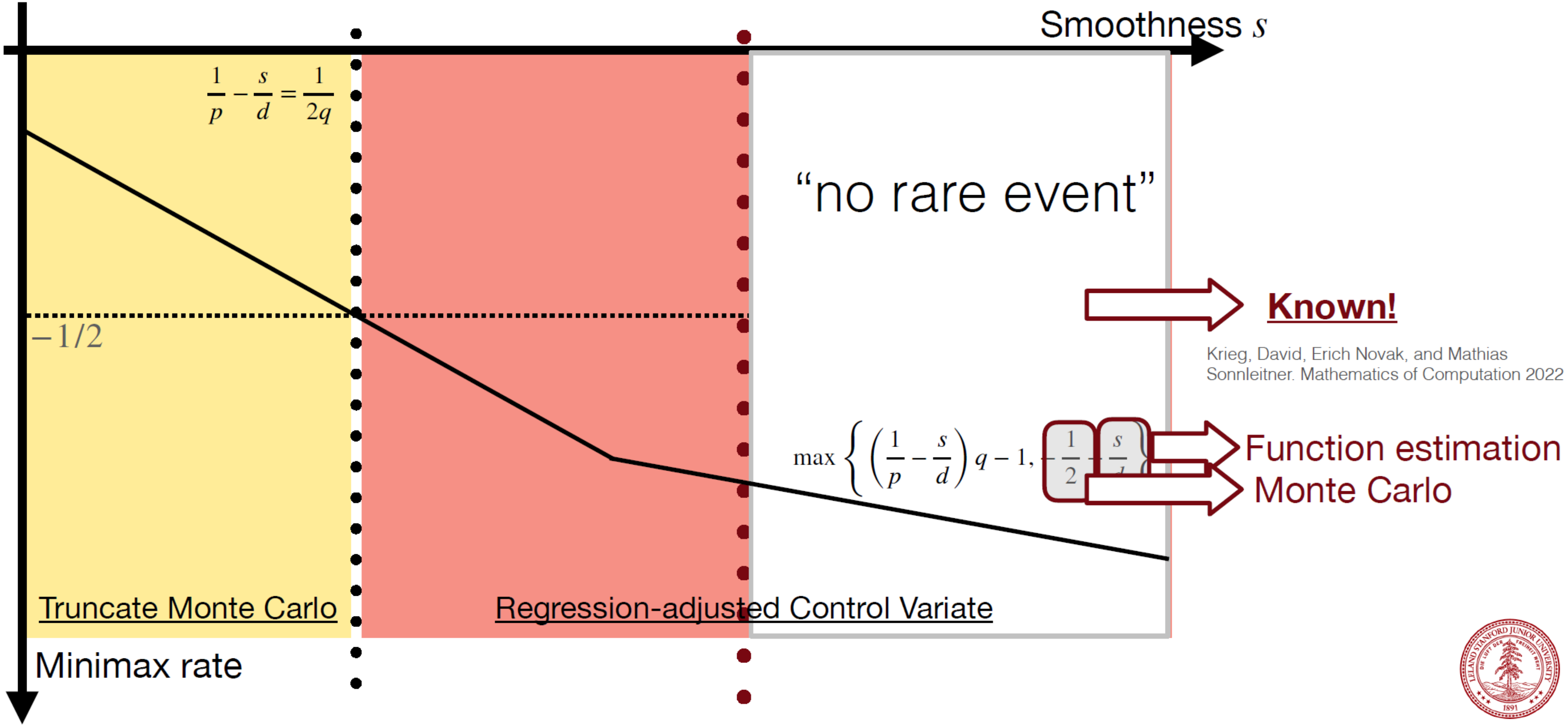
When the control variate helps



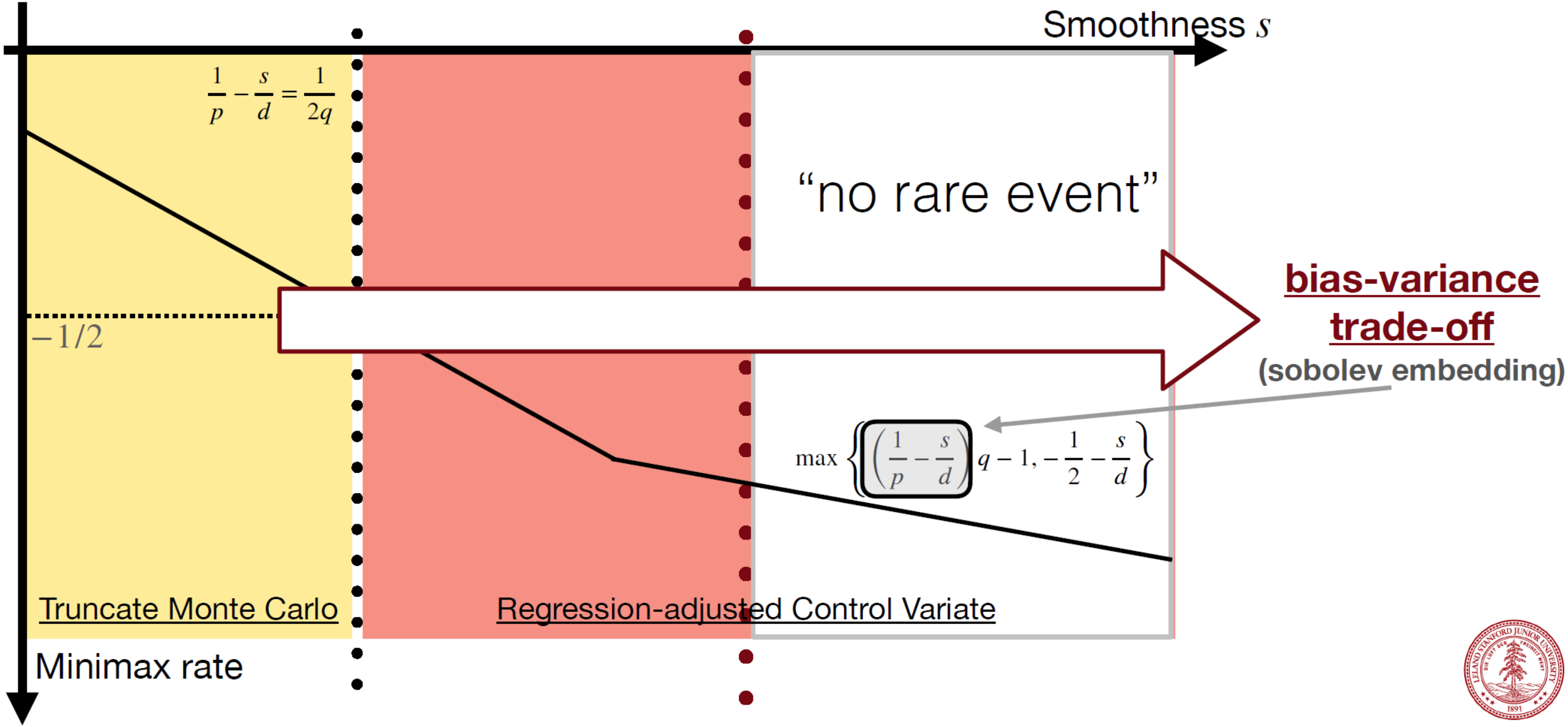
When the control variate helps



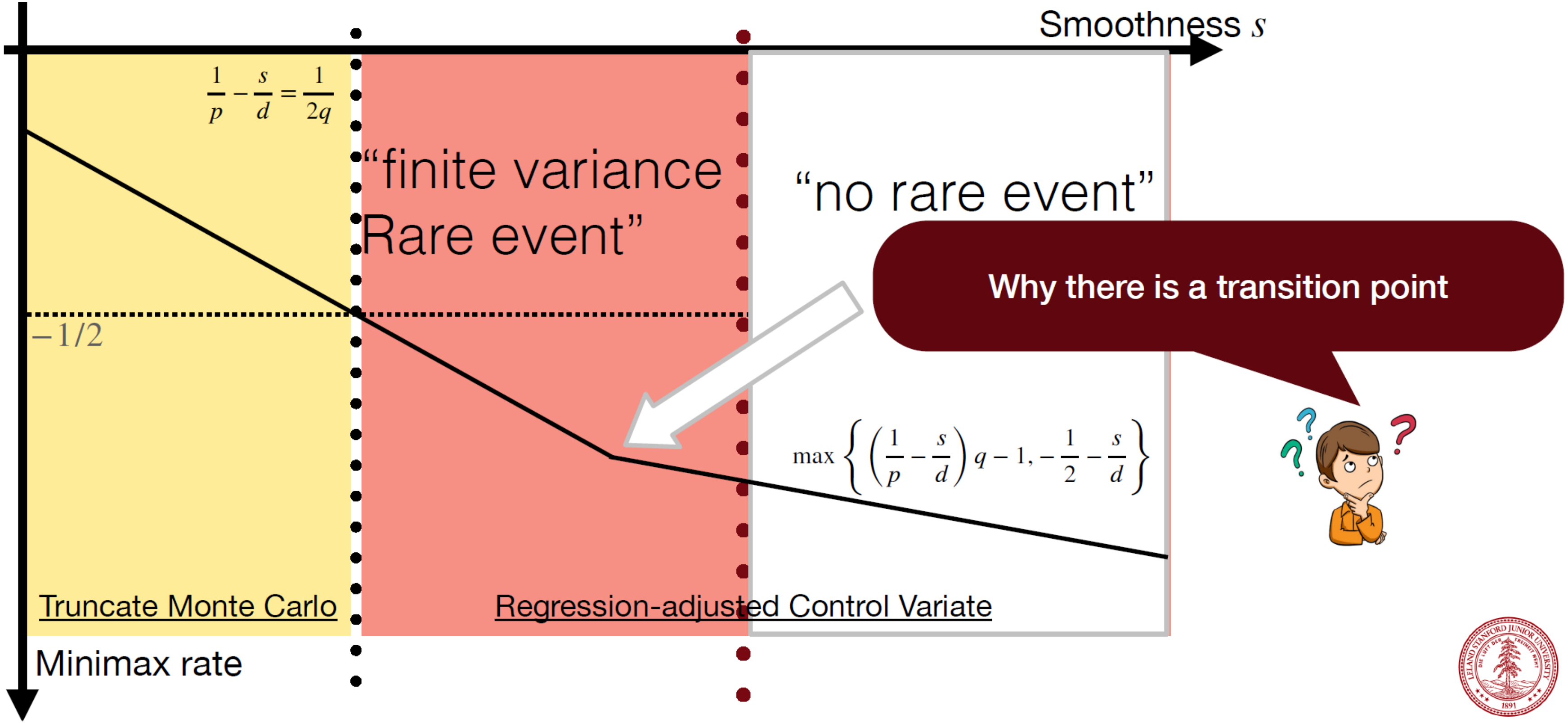
When the control variate helps



When the control variate helps



When the control variate helps



Technical Tools

Sobolev Embedding Theorem (simplified version)

Fix $s, t \in \mathbb{N}_0$ and $p, q \in \mathbb{R}$ satisfying $s > t$, $p < d$ and $1 \leqslant p < q \leqslant \infty$, then we have:

(I) Inclusion between Sobolev spaces: $\frac{1}{p} - \frac{s}{d} = \frac{1}{q} - \frac{t}{d} \Rightarrow W^{s,p}(\mathbb{R}^d) \subseteq W^{t,q}(\mathbb{R}^d)$

(II) Special case when $t = 0$: $\frac{1}{p} - \frac{s}{d} \leqslant \frac{1}{q} \Rightarrow W^{s,p}(\mathbb{R}^d) \subseteq L^q(\mathbb{R}^d)$

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

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

Low order term

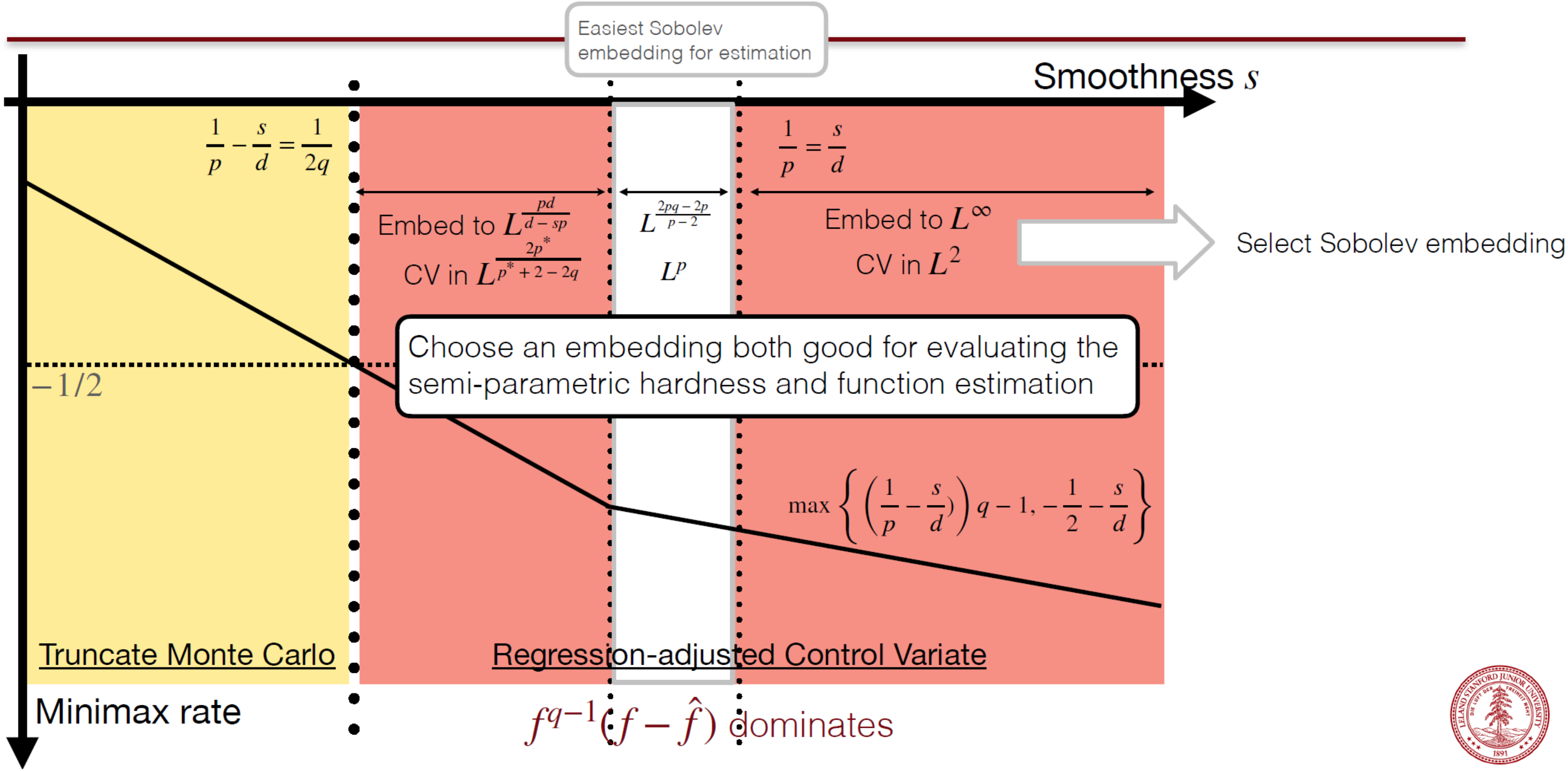
$f^{q-1}(f - \hat{f}) + (f - \hat{f})^q$

"influnce function" (gradient)

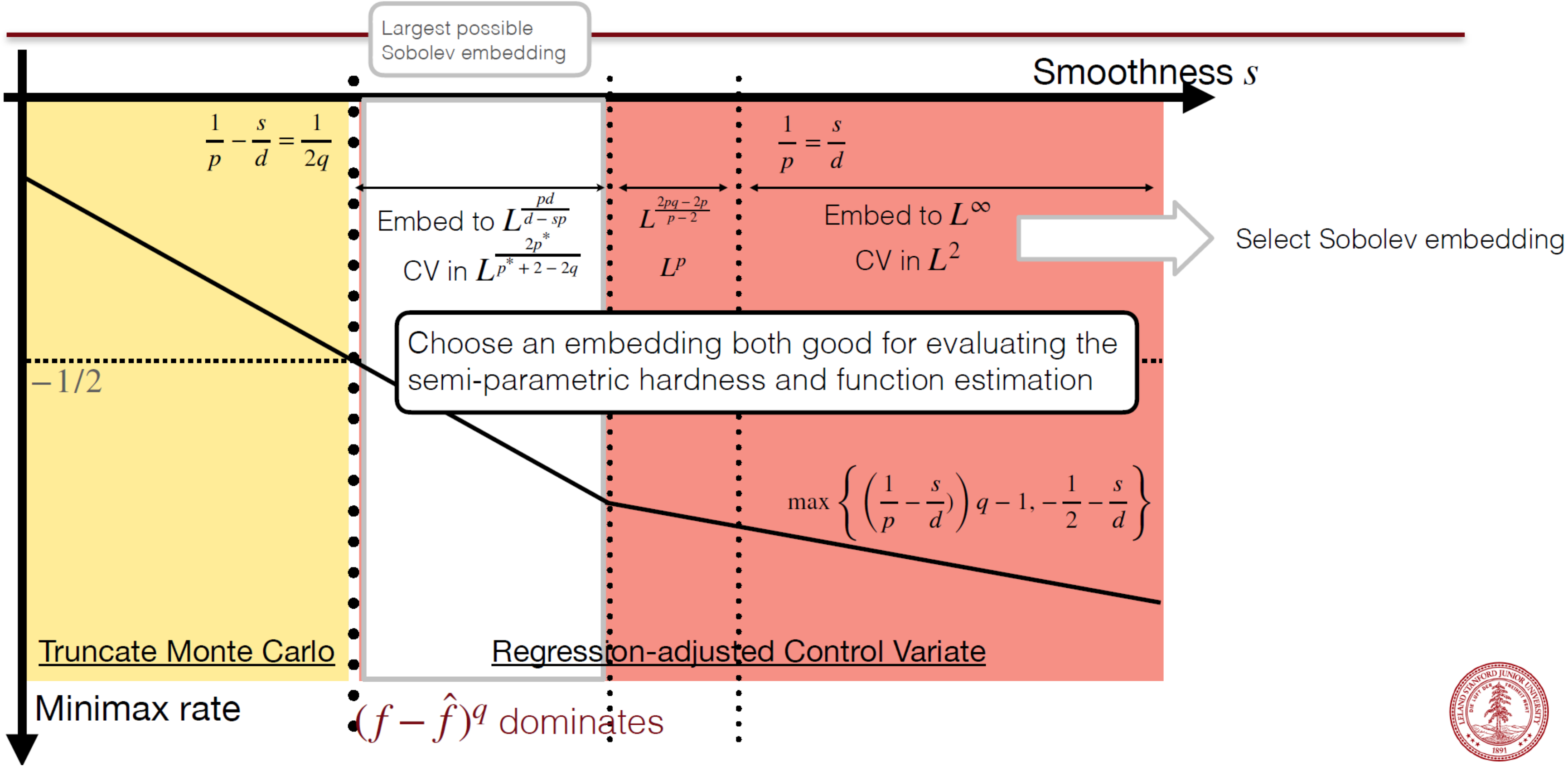
Embed f^{q-1} and $f - \hat{f}$ into "dual" space

How to select the sobolev emebedding

Tricky part of the Proof:select embedding

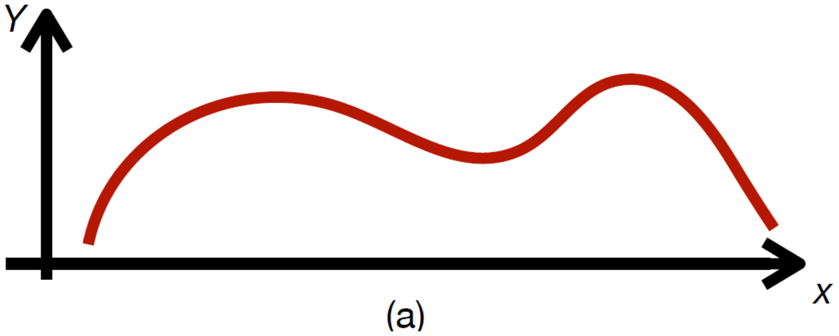


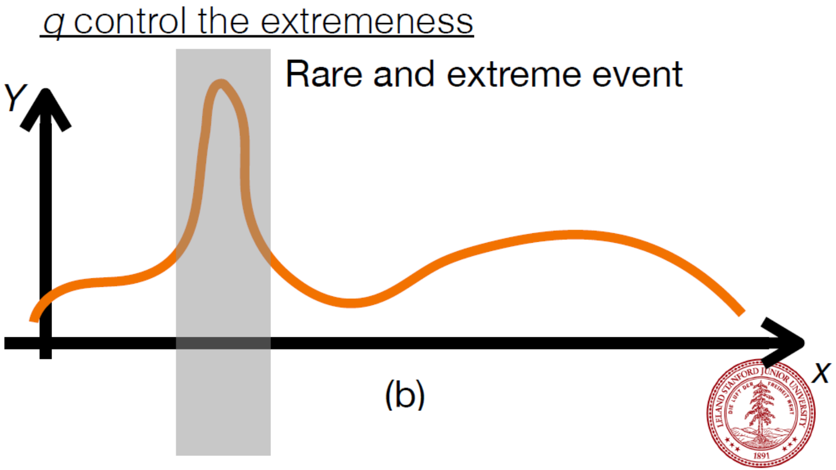
Tricky part of the Proof:select embedding



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





References

- Lin, Lin. "Randomized estimation of spectral densities of large matrices made accurate." Numerische Mathematik 136 (2017): 183-213.
- Romano, Yaniv, Evan Patterson, and Emmanuel Candes. "Conformalized quantile regression." Advances in neural information processing systems 32 (2019).
- Meyer, Raphael A., et al. "Hutch++: Optimal stochastic trace estimation." Symposium on Simplicity in Algorithms (SOSA). Society for Industrial and Applied Mathematics, 2021.
- Jordan, Michael, Yixin Wang, and Angela Zhou. "Empirical Gateaux Derivatives for Causal Inference." Advances in Neural Information Processing Systems 35 (2022): 8512-8525.
- Shi, Jiaxin, et al. "Gradient estimation with discrete Stein operators." Advances in neural information processing systems 35 (2022): 25829-25841.
- Krieg, David, Erich Novak, and Mathias Sonnleitner. "Recovery of Sobolev functions restricted to iid sampling." Mathematics of Computation 91.338 (2022): 2715-2738.
- Angelopoulos, Anastasios N., et al. "Prediction-powered inference." arXiv preprint arXiv:2301.09633 (2023).

Thank You all for Listening! Questions?