Minimax rates for Batched Stochastic Optimization

John Duchi
based on joint work with Feng Ruan and Chulhee Yun

Stanford University

Tradeoffs

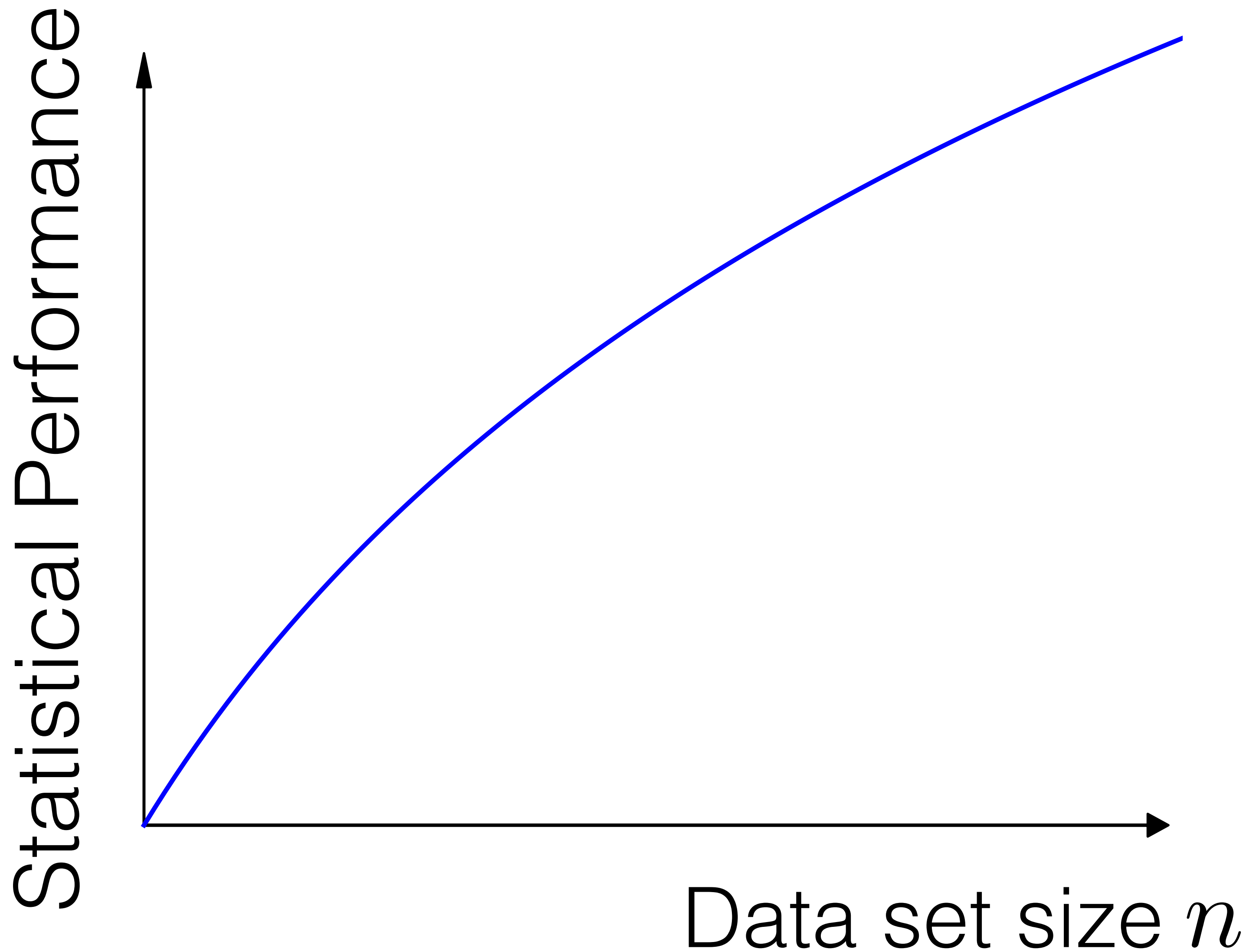
- Major problem in theoretical statistics: how do we characterize statistical optimality for problems with constraints?

Tradeoffs

- Major problem in theoretical statistics: how do we characterize statistical optimality for problems with constraints?
  - Computational [Berthet & Rigollet 13, Ma & Wu 15, Brennan et al. 18, Feldman et al. 18]
  - Privacy [Dwork et al. 06, Hardt & Talwar 09, Duchi et al. 13]
  - Robustness [Huber 81, Hardt & Moitra 13, Diakonikolas et al. 16]
  - Memory / communication [Duchi et al. 14, Braverman et al. 15, Steinhardt & Duchi 16]

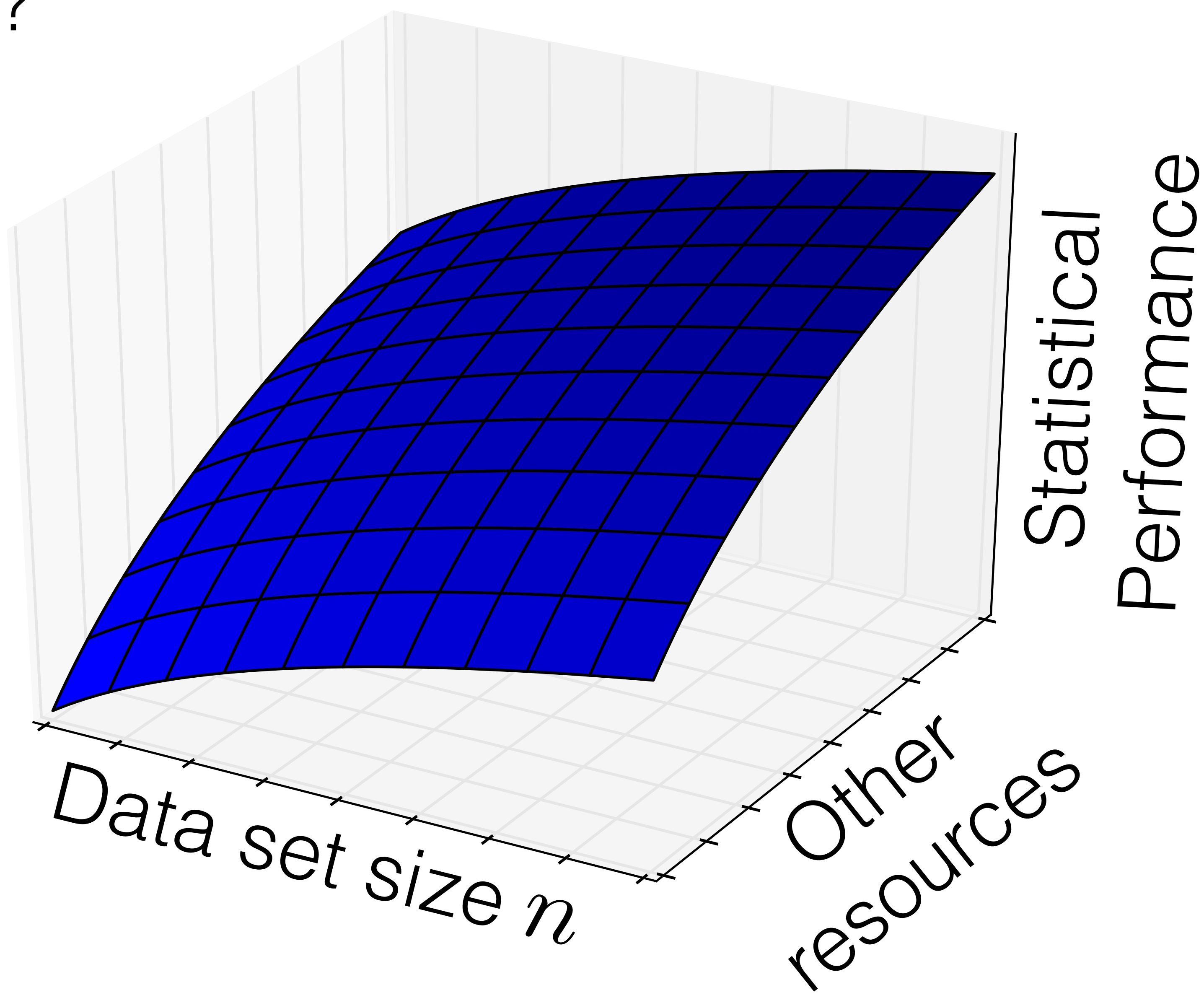
Tradeoffs

- Major problem in theoretical statistics: how do we characterize statistical optimality for problems with constraints?



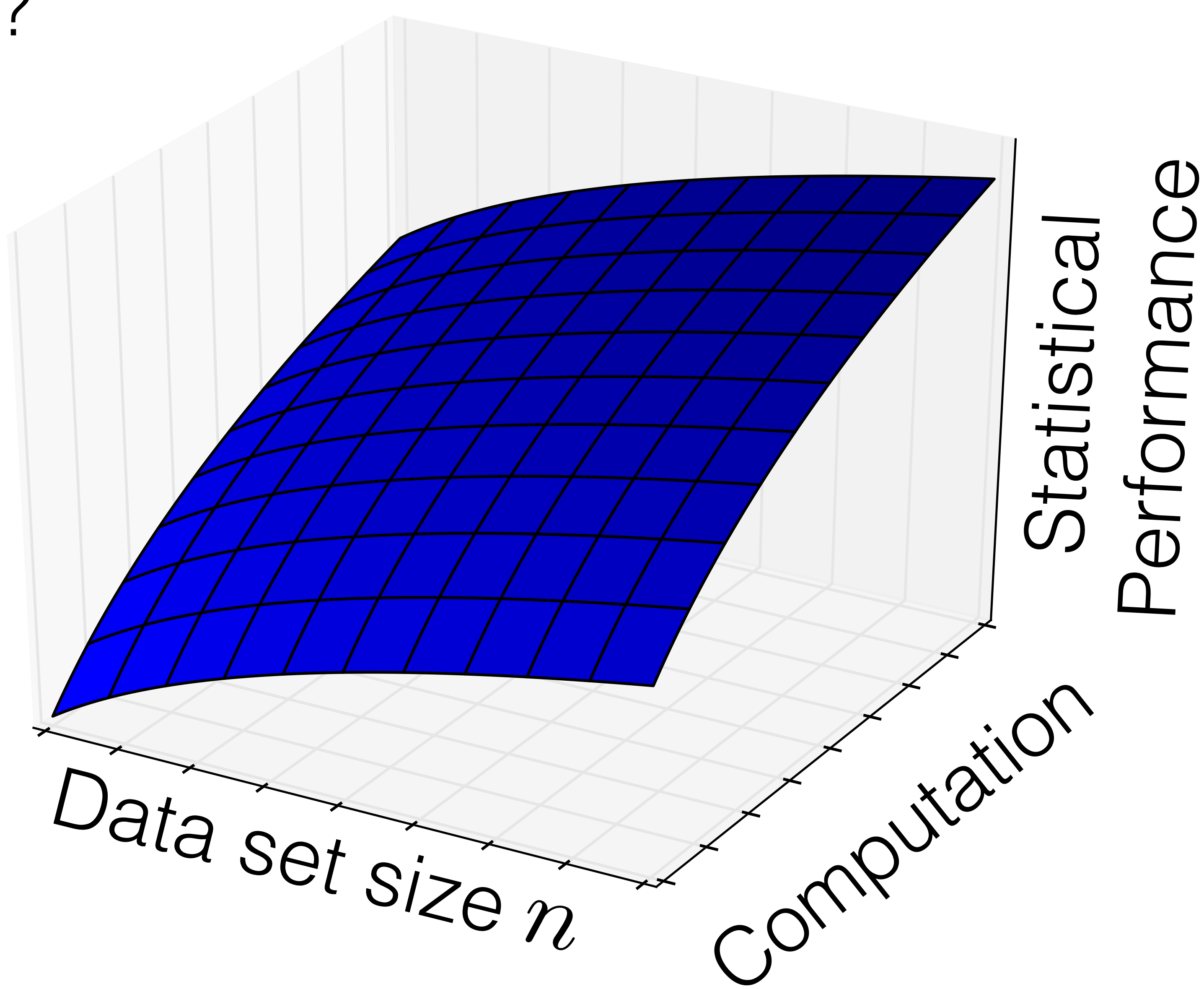
Tradeoffs

- Major problem in theoretical statistics: how do we characterize statistical optimality for problems with constraints?



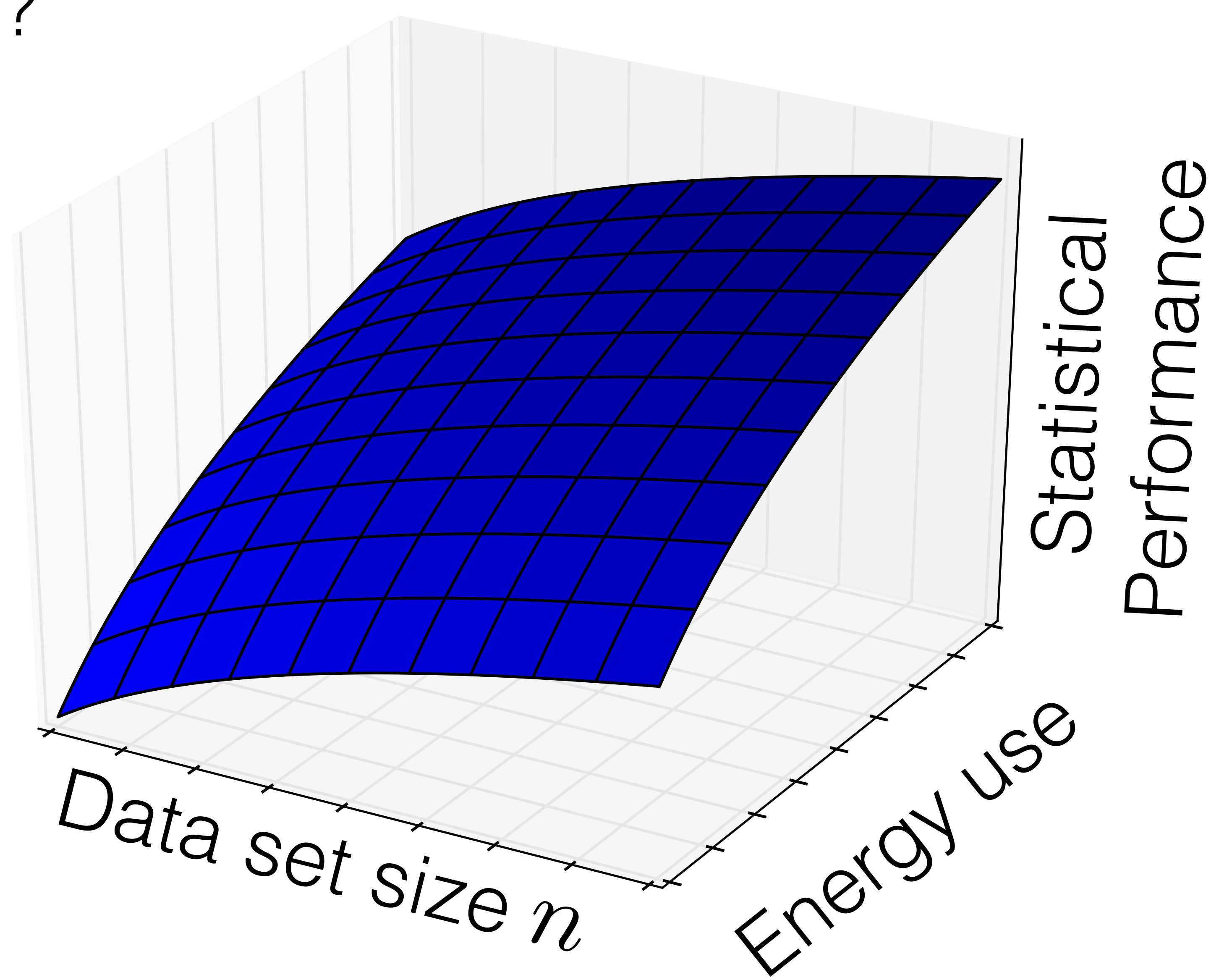
Tradeoffs

- Major problem in theoretical statistics: how do we characterize statistical optimality for problems with constraints?



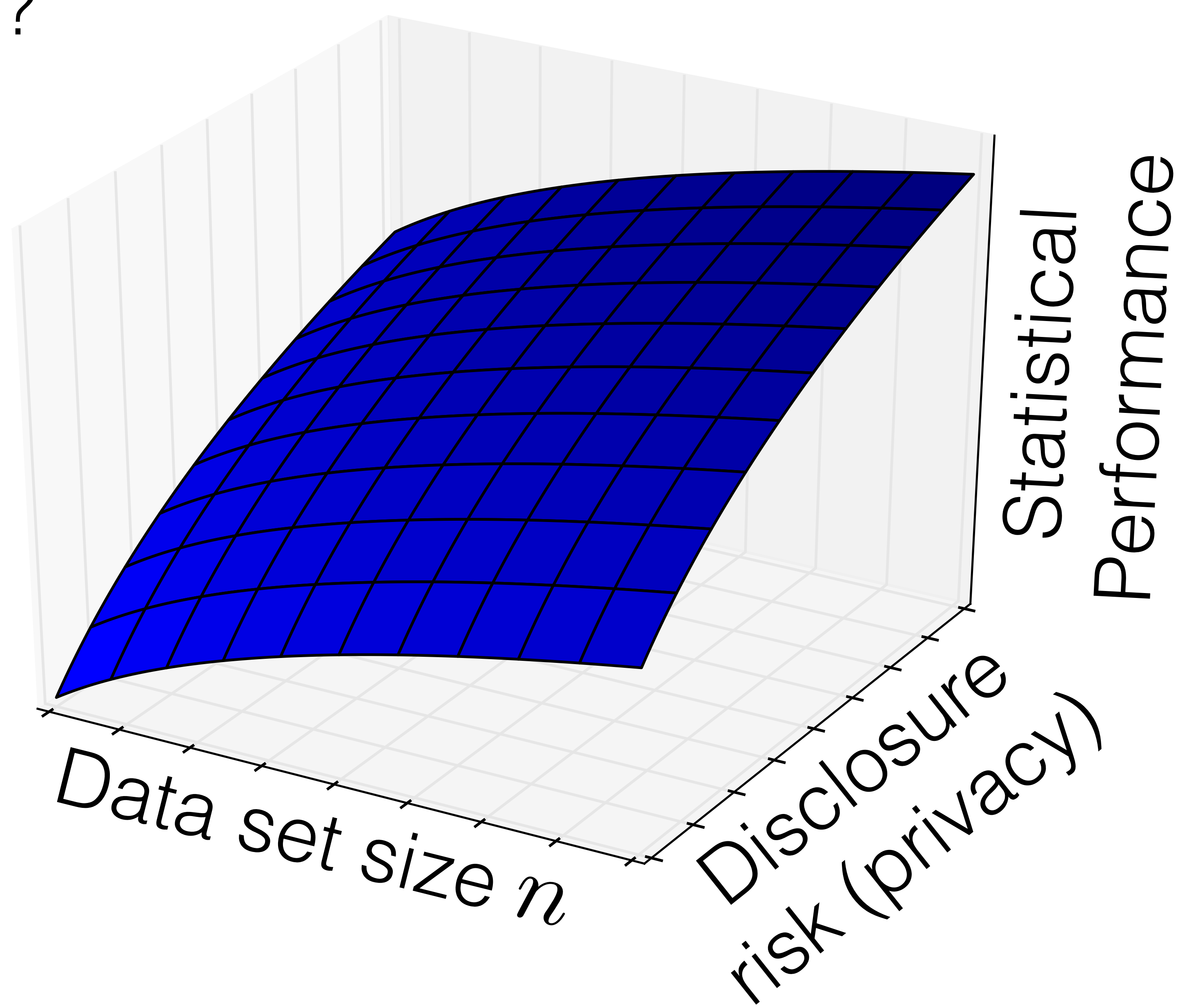
Tradeoffs

- Major problem in theoretical statistics: how do we characterize statistical optimality for problems with constraints?



Tradeoffs

- Major problem in theoretical statistics: how do we characterize statistical optimality for problems with constraints?



Problem Setting

minimize $f(x)$

where f convex

given mean-zero noisy gradient information

$$g = \nabla f(x) + \xi$$

Problem Setting

minimize $f(x)$

where f convex

given mean-zero noisy gradient information

$$g = \nabla f(x) + \xi$$

computational complexity for these problems?

Stochastic Gradient methods

Iterate (for k = 1, 2, …)

$$g_k = \nabla f(x_k) + \xi_k$$

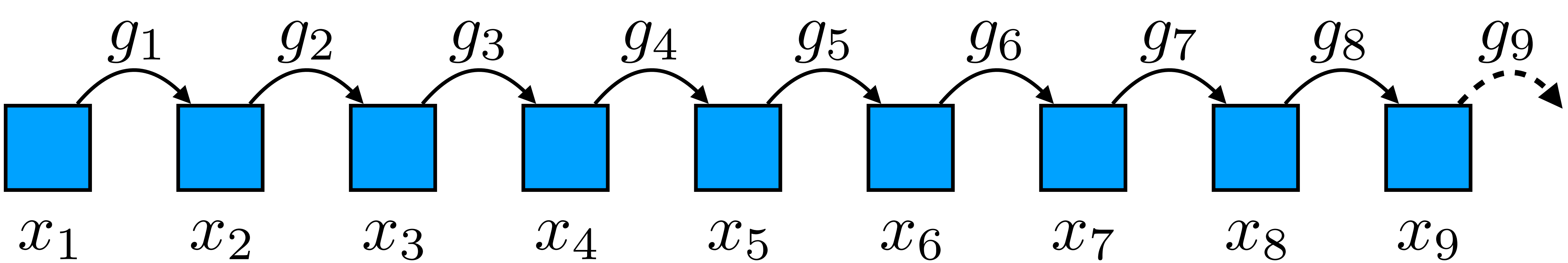
$$x_{k+1} = x_k - \alpha_k g_k$$

Stochastic Gradient methods

Iterate (for k = 1, 2, …)

$$g_k = \nabla f(x_k) + \xi_k$$

$$x_{k+1} = x_k - \alpha_k g_k$$

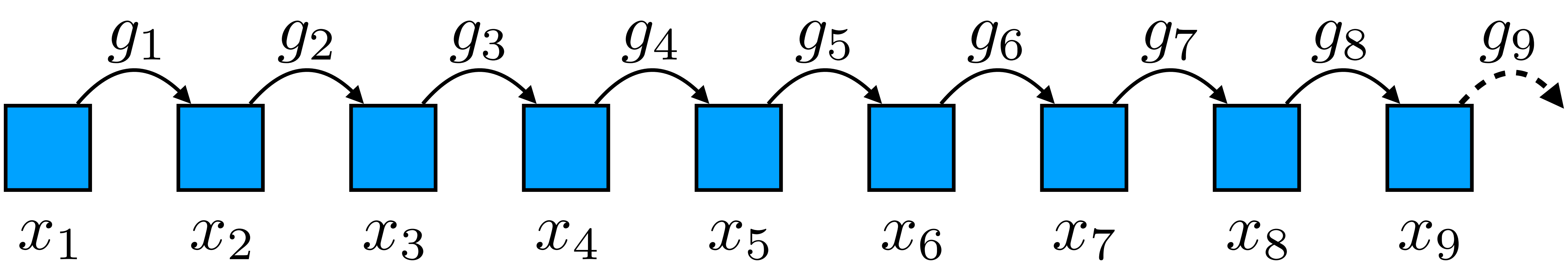


Stochastic Gradient methods

Iterate (for k = 1, 2, …)

$$g_k = \nabla f(x_k) + \xi_k$$

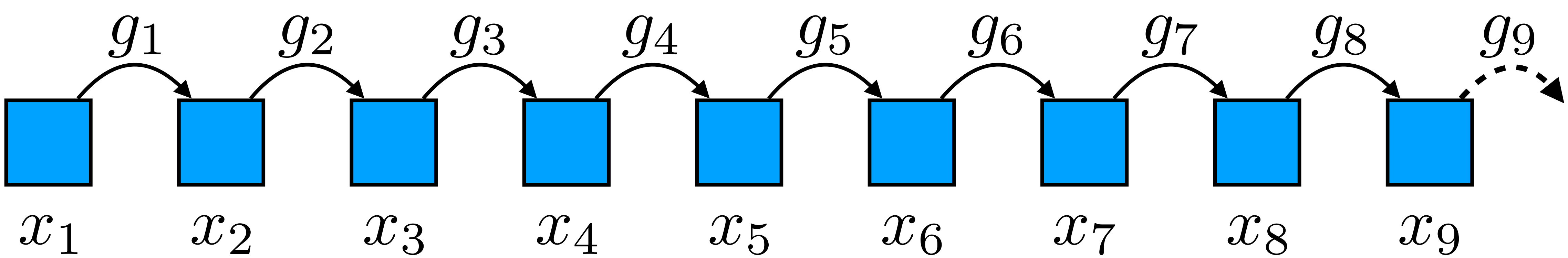
$$x_{k+1} = x_k - \alpha_k g_k$$



Theorem (Nemirovski & Yudin 83; Nemirovski et al. 09; Agarwal et al. 11)
After k iterations, we have (optimal) convergence

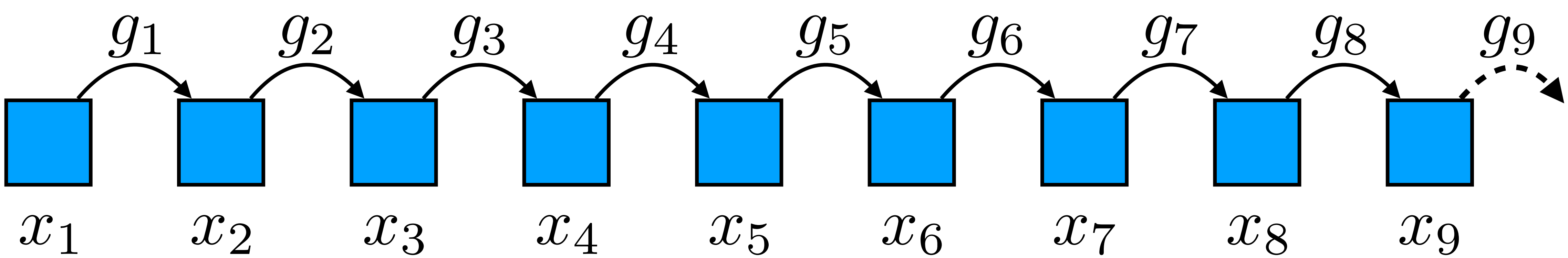
$$\mathbb{E}[f(\overline{x}_k)] - f^\star \lesssim \frac{1}{\sqrt{k}}$$

Parellelization and interactivity?

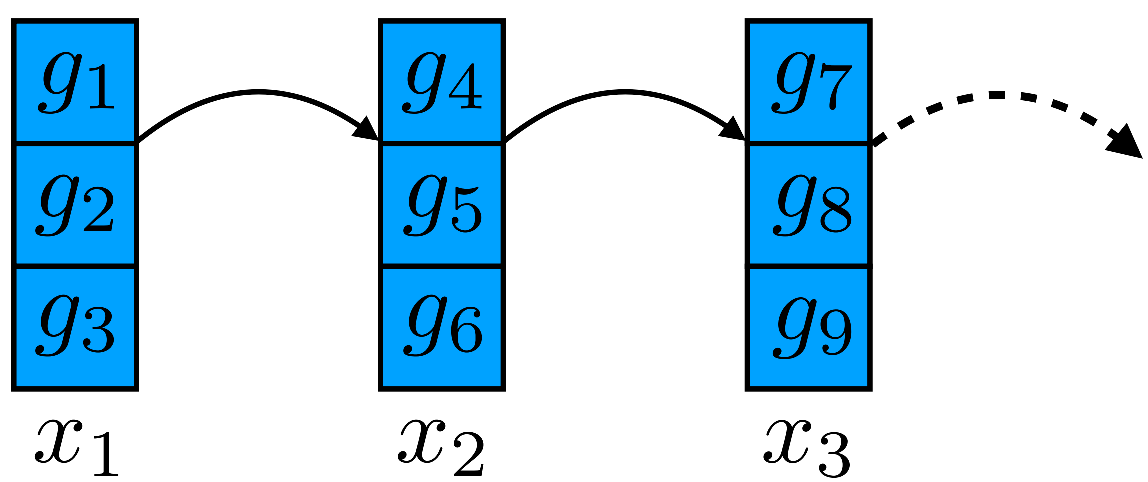


Requires many iterations, lots of interaction, no parallelism

Parellelization and interactivity?



Requires many iterations, lots of interaction, no parallelism



Trade off breadth for depth?

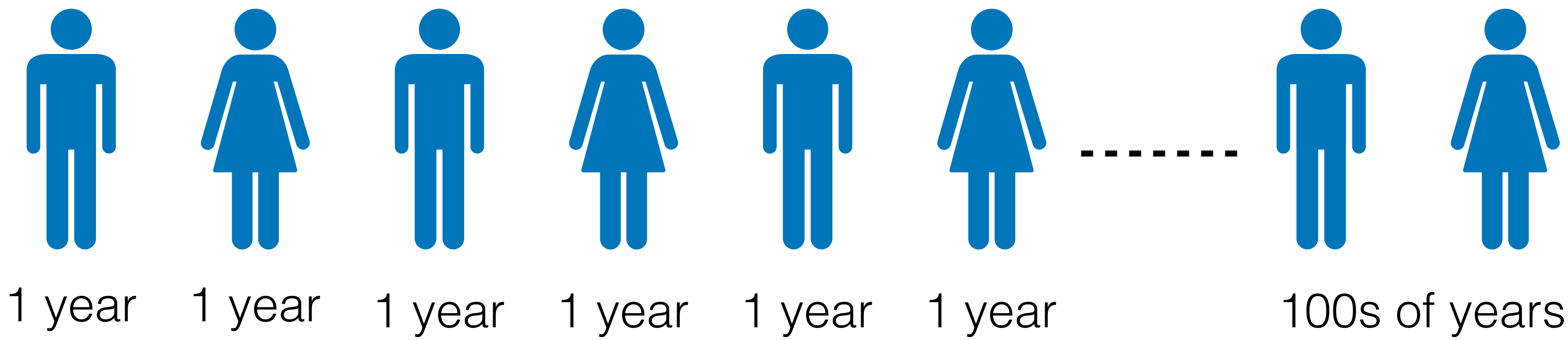
Batched optimization?

Medical trials [Perchet et al. 16, Hardwick & Stout 02, Stein 45]

Batched optimization?

Medical trials [Perchet et al. 16, Hardwick & Stout 02, Stein 45]

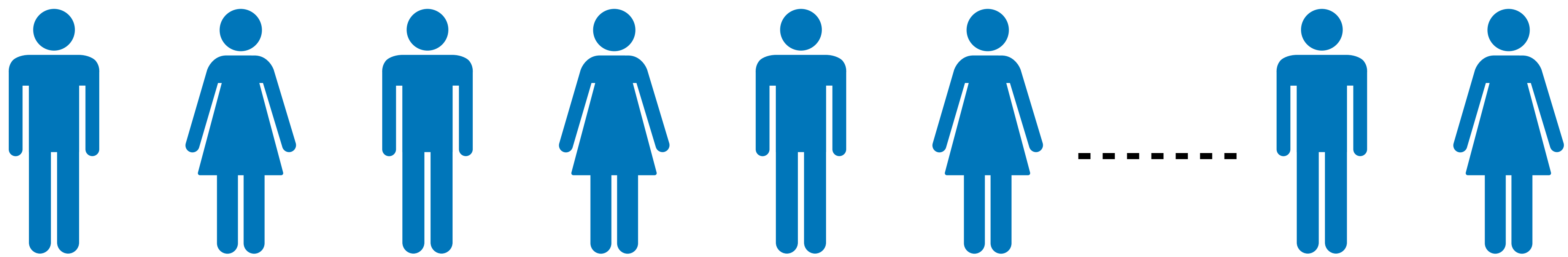
Ideal: get patient, give treatment, observe outcome



Batched optimization?

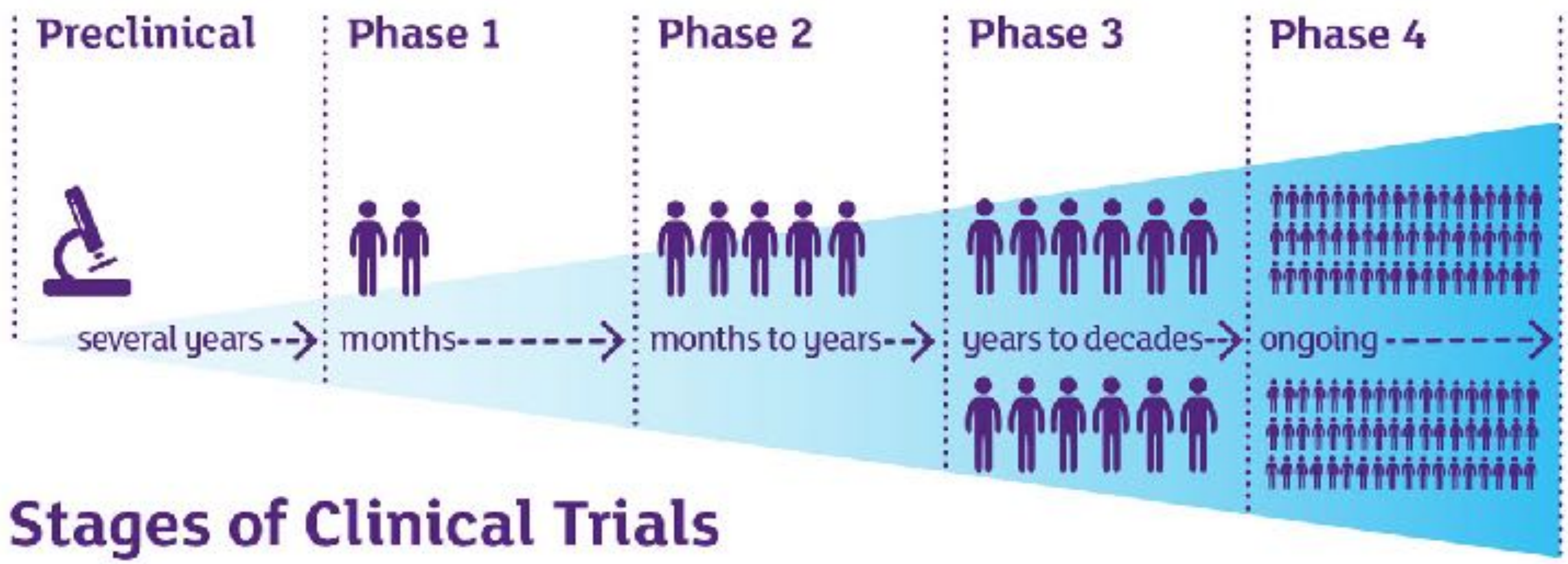
Medical trials [Perchet et al. 16, Hardwick & Stout 02, Stein 45]

Ideal: get patient, give treatment, observe outcome



1 year 1 year 1 year 1 year 1 year 1 year 100s of years

Reality:



Local Differential Privacy

Problem Statement

Problem: given M rounds of adaptation and n computations, what is the optimal error in optimization?

Problem Statement

Problem: given M rounds of adaptation and n computations, what is the optimal error in optimization?

$\mathcal{A}_{M,n}$ = Algorithms with *M* rounds of computation and *n* (noisy) gradient computations

$\mathcal{F}$ = Function class of interest

Problem Statement

Problem: given M rounds of adaptation and n computations, what is the optimal error in optimization?

$\mathcal{A}_{M,n}$ = Algorithms with *M* rounds of computation and *n* (noisy) gradient computations

$\mathcal{F}$ = Function class of interest

Study **minimax optimization** error

$$\mathbb{E}[f(\widehat{x})] - f^\star$$

Problem Statement

Problem: given M rounds of adaptation and n computations, what is the optimal error in optimization?

$\mathcal{A}_{M,n}$ = Algorithms with *M* rounds of computation and *n* (noisy) gradient computations

$\mathcal{F}$ = Function class of interest

Study **minimax optimization** error

$$\sup_{f \in \mathcal{F}} \{\mathbb{E}[f(\widehat{x})] - f^\star\}$$

Problem Statement

Problem: given M rounds of adaptation and n computations, what is the optimal error in optimization?

$\mathcal{A}_{M,n}$ = Algorithms with M rounds of computation and n (noisy) gradient computations

$\mathcal{F}$ = Function class of interest

Study **minimax optimization** error

$$\inf_{\widehat{x} \in \mathcal{A}_{M,n}} \sup_{f \in \mathcal{F}} \{\mathbb{E}[f(\widehat{x})] - f^\star\}$$

Problem Statement

Problem: given M rounds of adaptation and n computations, what is the optimal error in optimization?

$\mathcal{A}_{M,n}$ = Algorithms with *M* rounds of computation and *n* (noisy) gradient computations

$\mathcal{F}$ = Function class of interest

Study **minimax optimization** error

$$\mathfrak{M}_{M,n}(\mathcal{F}) := \inf_{\widehat{x} \in \mathcal{A}_{M,n}} \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}[f(\widehat{x})] - f^\star \right\}$$

Background and hopes

Background and hopes

- [Perchet, RCS 18] Batched Bandits. For 2 armed bandit, optimal regret achievable with M = O(log log n)

Background and hopes

- [Perchet, RCS 18] Batched Bandits. For 2 armed bandit, optimal regret achievable with M = O(log log n)

- [Nemirovski et al. 09, Ghadimi & Lan 12] Stochastic strongly convex optimization:

$$f(\widehat{x}) - f^\star \lesssim \|x_0 - x^\star\|^2 \exp(-M/\sqrt{\mathrm{Cond}(f)}) + \frac{\mathrm{Var}(\xi)}{\lambda n}$$

or

$$M \gtrsim \sqrt{\mathrm{Cond}(f)} \log n$$ rounds

Background and hopes

- [Perchet, RCS 18] Batched Bandits. For 2 armed bandit, optimal regret achievable with M = O(log log n)

- [Nemirovski et al. 09, Ghadimi & Lan 12] Stochastic strongly convex optimization:

$$f(\widehat{x}) - f^\star \lesssim \|x_0 - x^\star\|^2 \exp(-M/\sqrt{\mathrm{Cond}(f)}) + \frac{\mathrm{Var}(\xi)}{\lambda n}$$

or

$$M \gtrsim \sqrt{\mathrm{Cond}(f)} \log n \quad \text{rounds}$$

- [Smith, TU 17] To solve convex optimization, need

$$M \gtrsim \log \frac{1}{\epsilon} \quad \text{rounds}$$

Main Results

$$\mathcal{F}_{H,\lambda} := \{\lambda \text{ strongly convex}, H \text{ smooth } f\}$$

Main Results

$$\mathcal{F}_{H,\lambda} := \{\lambda \text{ strongly convex}, H \text{ smooth } f\}$$

Theorem (D., Ruan, Yun 18)

$$\mathfrak{M}_{M,n}(\mathcal{F}_{H,\lambda}) \geq C(d,n) \cdot n^{-\left(1-\left(\frac{d}{d+2}\right)^{M}\right)}$$

where $C(d,n) \gg \dfrac{1}{\mathrm{poly}(n)}$

Main Results

$$\mathcal{F}_{H,\lambda} := \{\lambda \text{ strongly convex}, H \text{ smooth } f\}$$

Theorem (D., Ruan, Yun 18)

$$\mathfrak{M}_{M,n}(\mathcal{F}_{H,\lambda}) \geq C(d,n) \cdot n^{-\left(1-\left(\frac{d}{d+2}\right)^M\right)}$$

where $C(d,n) \gg \dfrac{1}{\mathrm{poly}(n)}$

Theorem (D., Ruan, Yun 18)

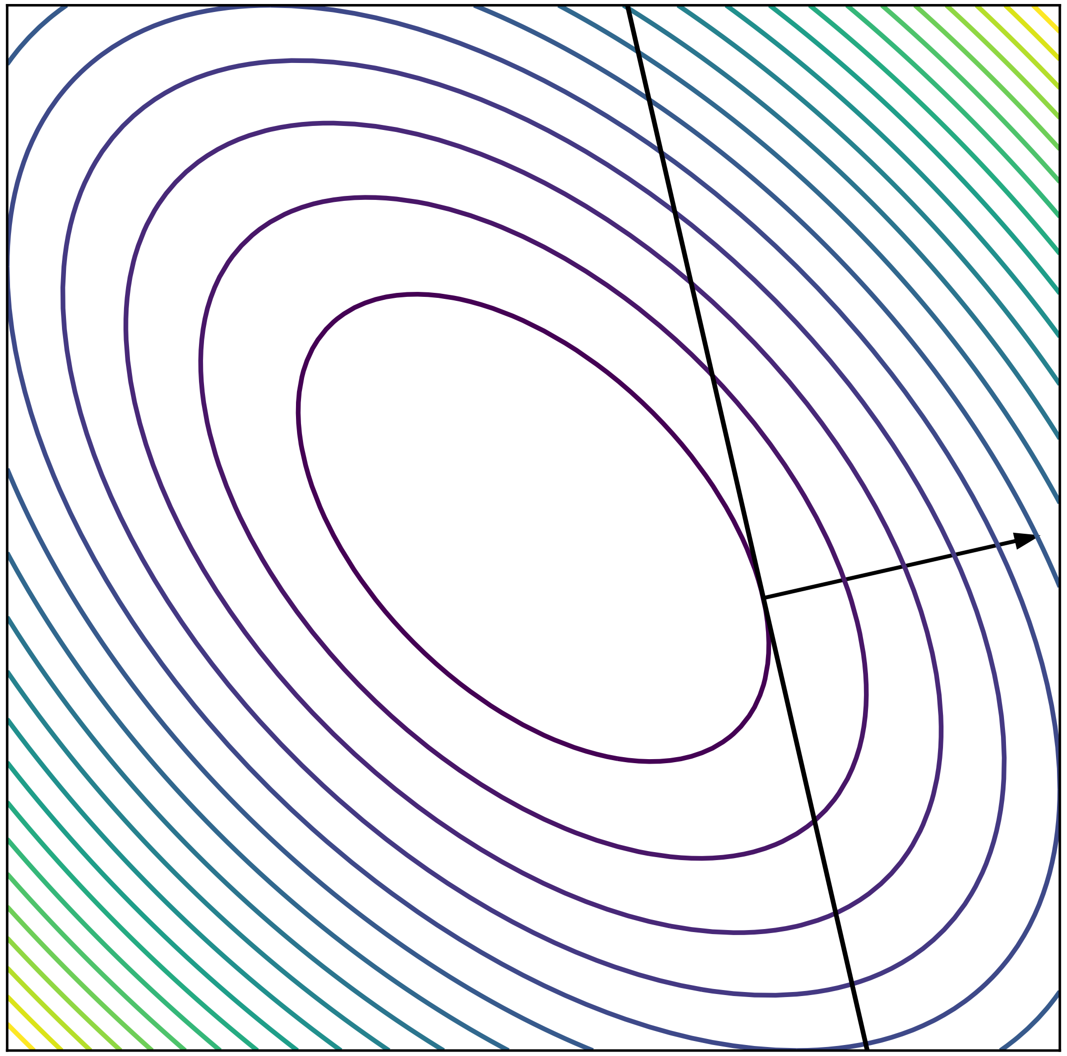
If $M \leq (d/2) \log\log n$ then there is an algorithm s.t.

$$\mathbb{P}\left(f(\widehat{x}) - f^\star \geq C n^{-\left(1-\left(\frac{d}{d+2}\right)^M\right)} \log n\right) \to 0$$

Achievability

For convex functions f,

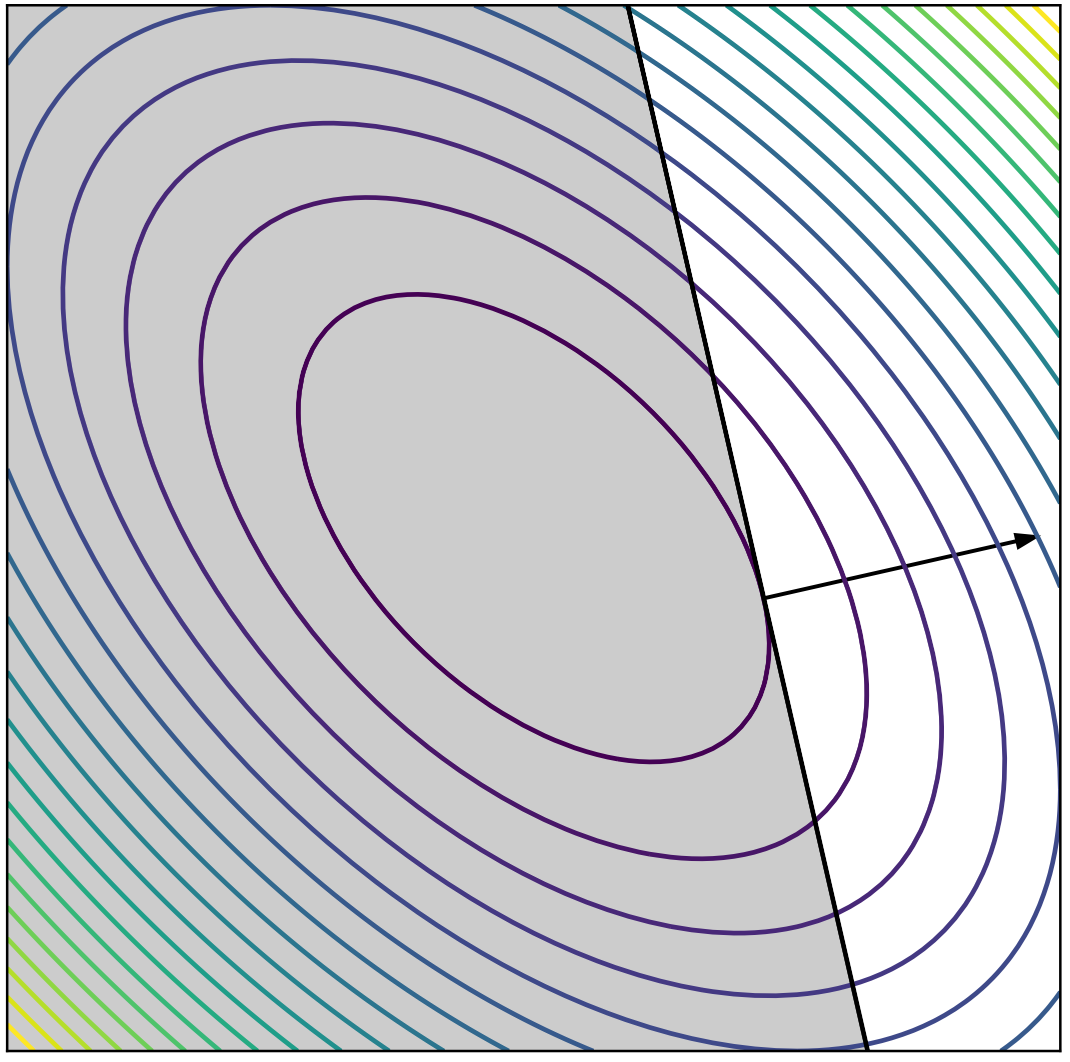
$$x^\star \in \{y \mid \langle \nabla f(x), y - x \rangle \leq 0\}$$



Achievability

For convex functions f,

$$x^\star \in \{y \mid \langle \nabla f(x), y - x \rangle \leq 0\}$$



Achievability

Maintain feasible box $\mathcal{B}_t = c_t + [-r_t, r_t]^d$ with center c_t

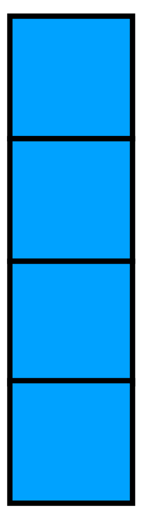
At round t, take points $x_i \in \mathcal{B}_t, i = 1, \ldots, m$

Achievability

Maintain feasible box $\mathcal{B}_t = c_t + [-r_t, r_t]^d$ with center c_t

At round t, take points $x_i \in \mathcal{B}_t, i = 1, \ldots, m$

get parallel (noisy) gradients



$$\widehat{\nabla} f(x_1)$$

$$\vdots$$

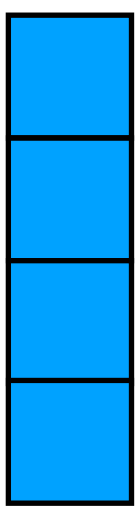
$$\widehat{\nabla} f(x_m)$$

Achievability

Maintain feasible box $\mathcal{B}_t = c_t + [-r_t, r_t]^d$ with center c_t

At round t, take points $x_i \in \mathcal{B}_t, i = 1, \ldots, m$

get parallel (noisy) gradients



$$\widehat{\nabla} f(x_1)$$

$$\vdots$$

$$\widehat{\nabla} f(x_m)$$

w.h.p.

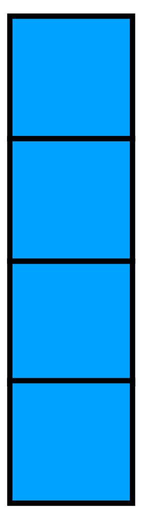
$$x^\star \in \{y \mid \langle \widehat{\nabla} f(x_i), y - x \rangle \leq \epsilon \|y - x\|\}$$

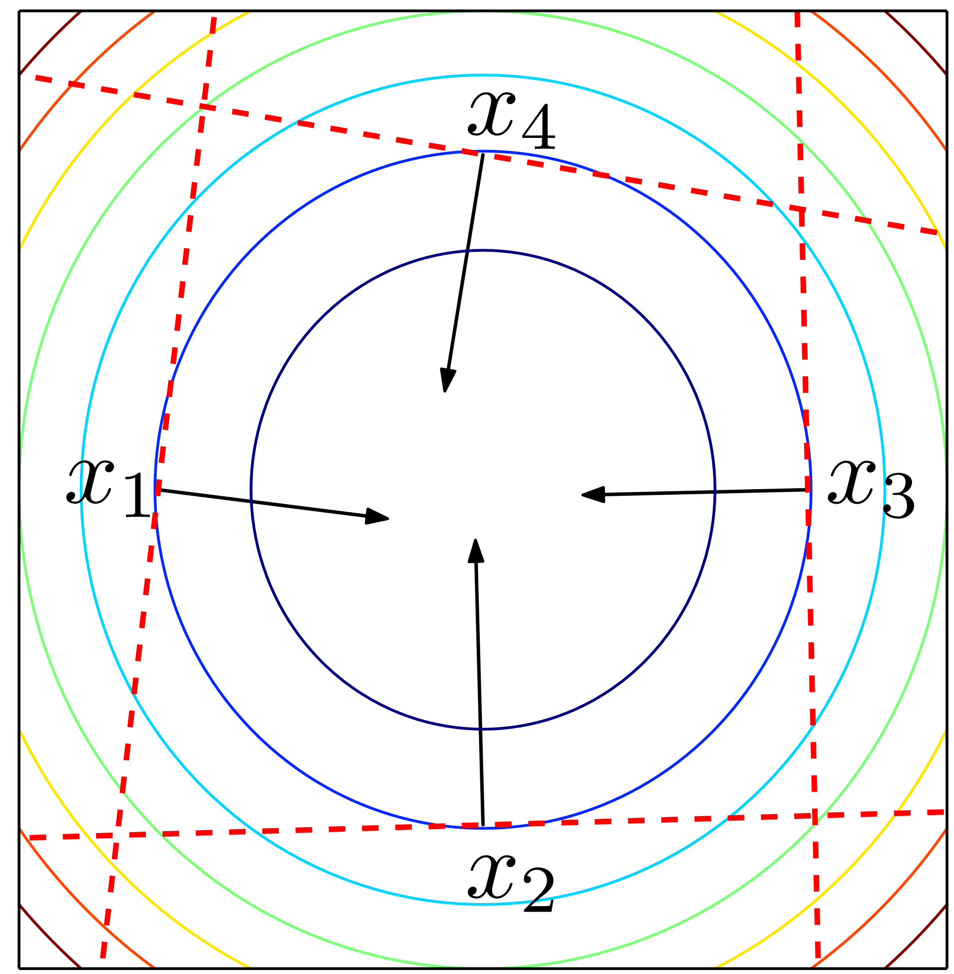
Achievability

Maintain feasible box $\mathcal{B}_t = c_t + [-r_t, r_t]^d$ with center c_t

At round t, take points $x_i \in \mathcal{B}_t, i = 1, \ldots, m$

get parallel (noisy) gradients





w.h.p.

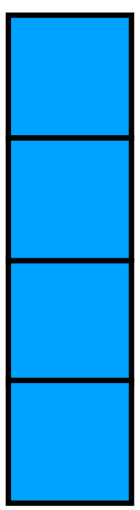
$$x^\star \in \{y \mid \langle \widehat{\nabla} f(x_i), y - x \rangle \leq \epsilon \|y - x\|\}$$

Achievability

Maintain feasible box $\mathcal{B}_t = c_t + [-r_t, r_t]^d$ with center c_t

At round t, take points $x_i \in \mathcal{B}_t, i = 1, \ldots, m$

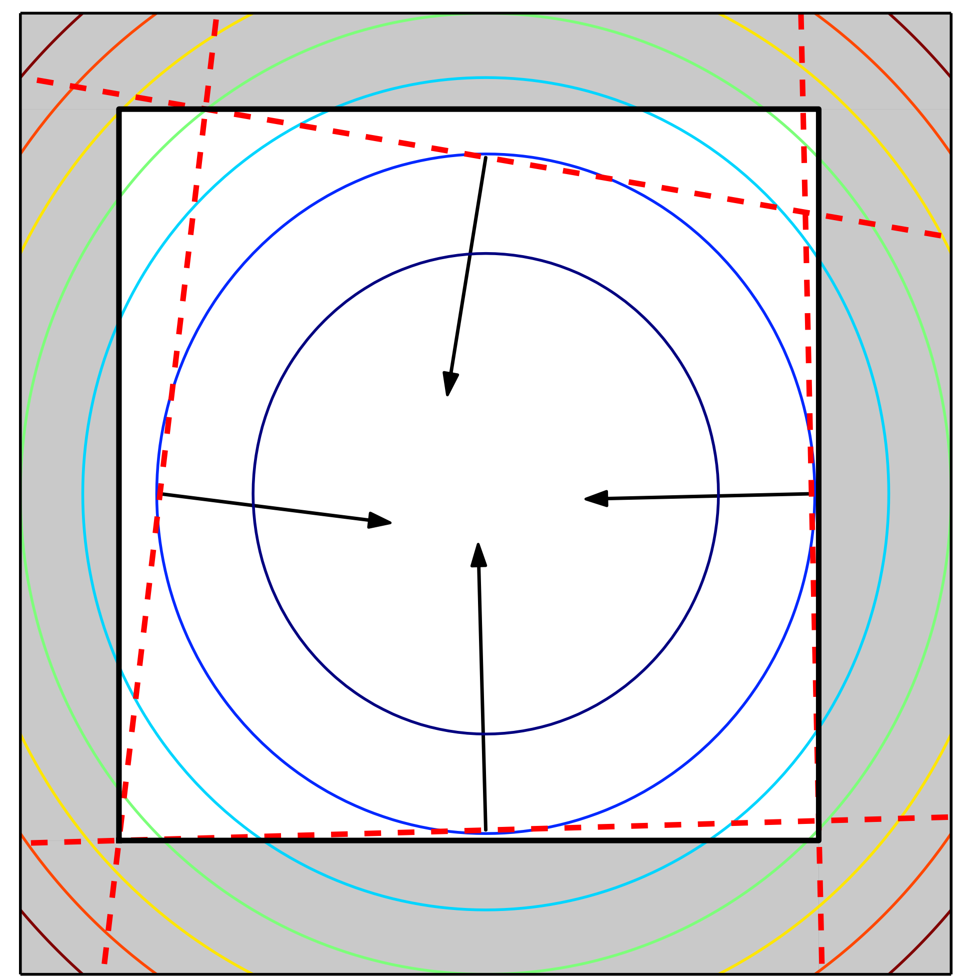
get parallel (noisy) gradients



$$\widehat{\nabla} f(x_1)$$

$$\vdots$$

$$\widehat{\nabla} f(x_m)$$



w.h.p.

$$x^\star \in \{y \mid \langle \widehat{\nabla} f(x_i), y - x \rangle \leq \epsilon \|y - x\|\}$$

Recursive shrinking

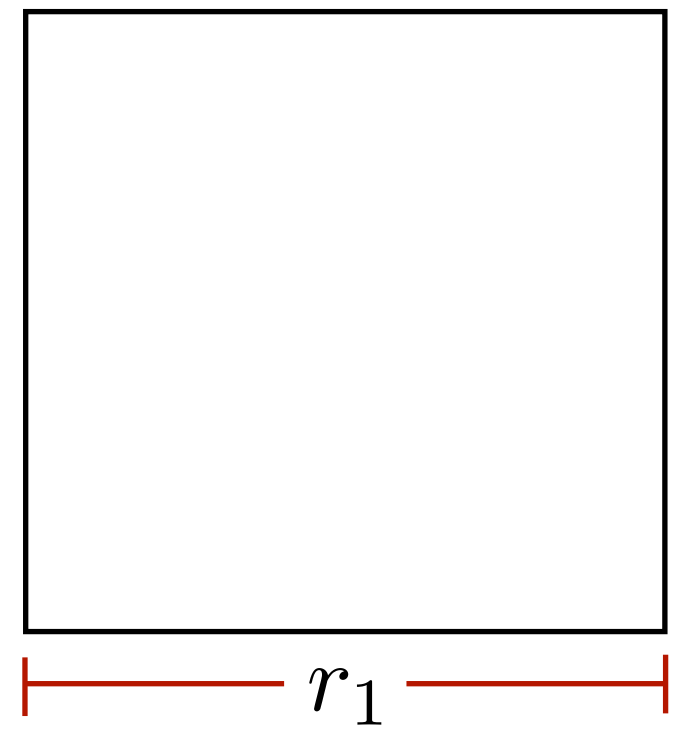
Box radius decreases as

$$r_t \leq \nu r_{t-1}^{\beta}$$

Recursive shrinking

Box radius decreases as

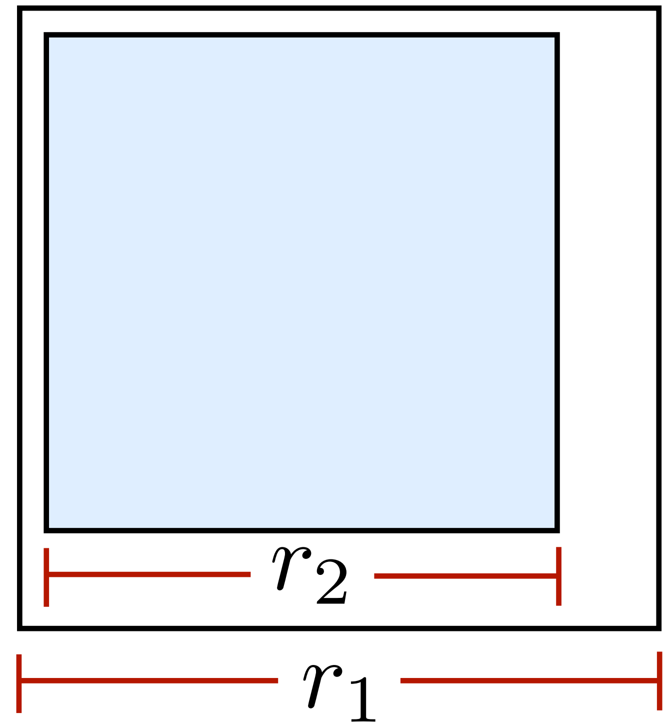
$$r_t \le \nu r_{t-1}^{\beta}$$



Recursive shrinking

Box radius decreases as

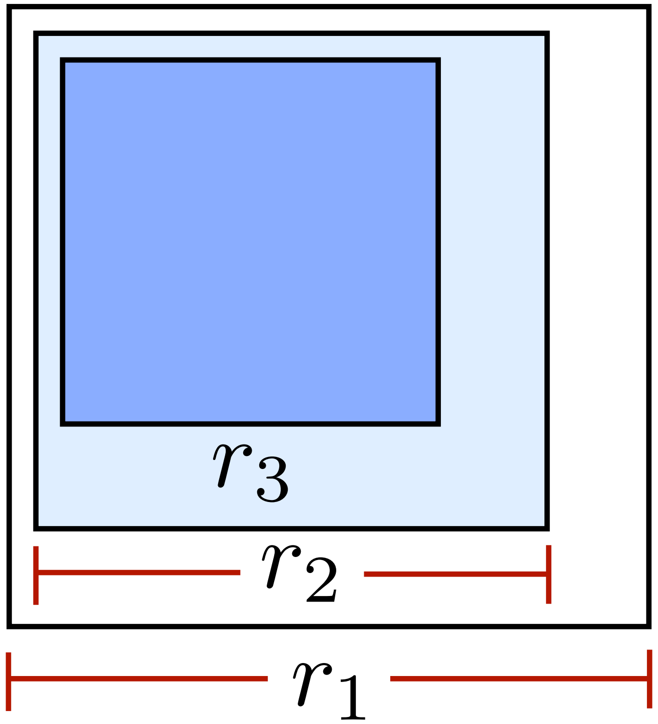
$$r_t \leq \nu r_{t-1}^{\beta}$$



Recursive shrinking

Box radius decreases as

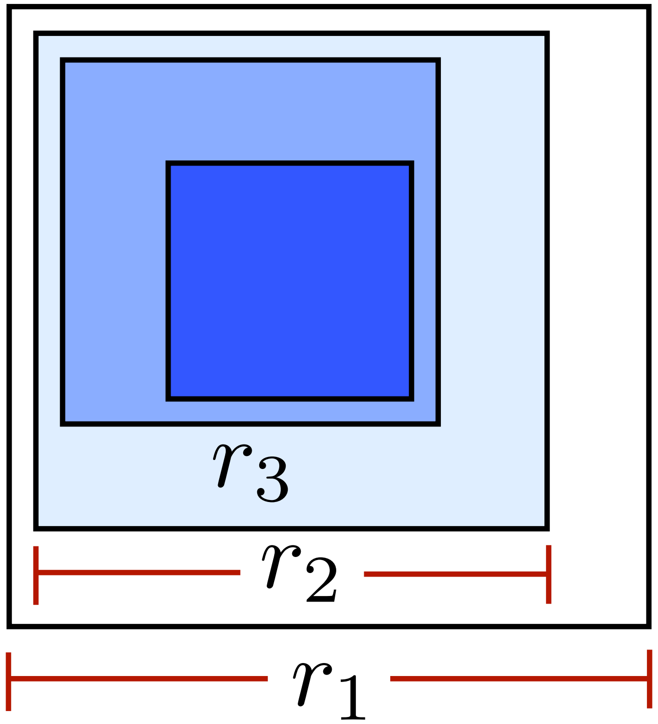
$$r_t \leq \nu r_{t-1}^{\beta}$$



Recursive shrinking

Box radius decreases as

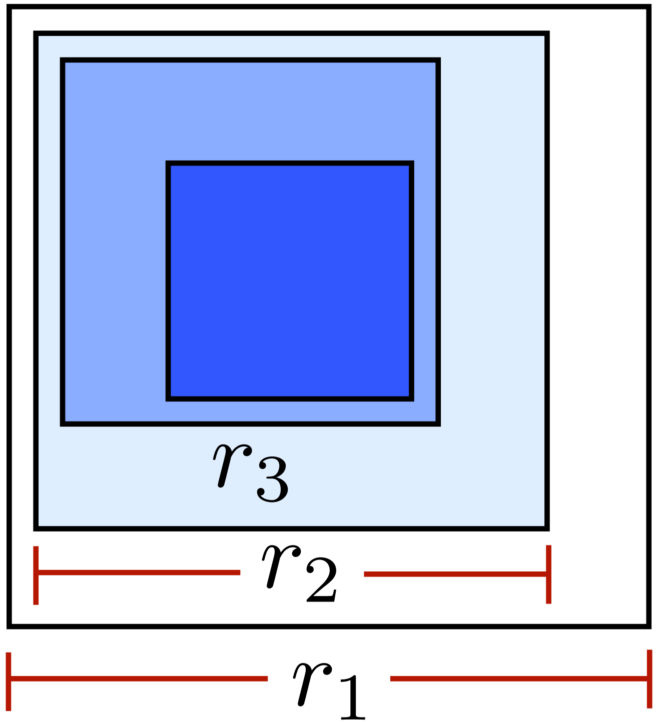
$$r_t \leq \nu r_{t-1}^{\beta}$$



Recursive shrinking

Box radius decreases as

$$r_t \leq \nu r_{t-1}^{\beta}$$

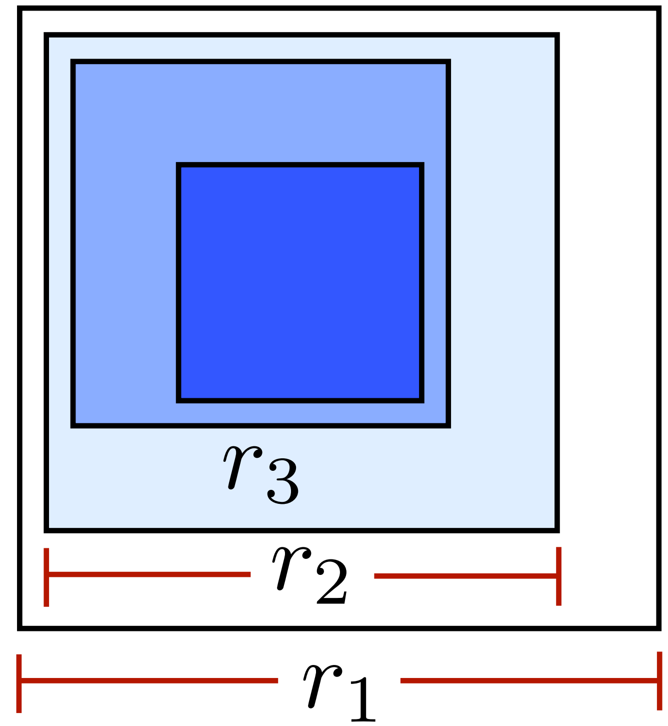


or, recursively

$$r_t \leq \nu r_{t-1}^{\beta} \leq \nu^{1+\beta} r_{t-2}^{\beta^2} \leq \cdots$$

$$\leq \nu^{\sum_{j=0}^{t-1} \beta^j} r_0^{\beta^t} \approx \nu^{\frac{\beta^t - 1}{\beta - 1}}$$

Recursive shrinking



Box radius decreases as

$$r_t \leq \nu r_{t-1}^{\beta}$$

or, recursively

$$r_t \leq \nu r_{t-1}^{\beta} \leq \nu^{1+\beta} r_{t-2}^{\beta^2} \leq \cdots$$

$$\leq \nu^{\sum_{j=0}^{t-1} \beta^j} r_0^{\beta^t} \approx \nu^{\frac{\beta^t - 1}{\beta - 1}}$$

for us, dimension d

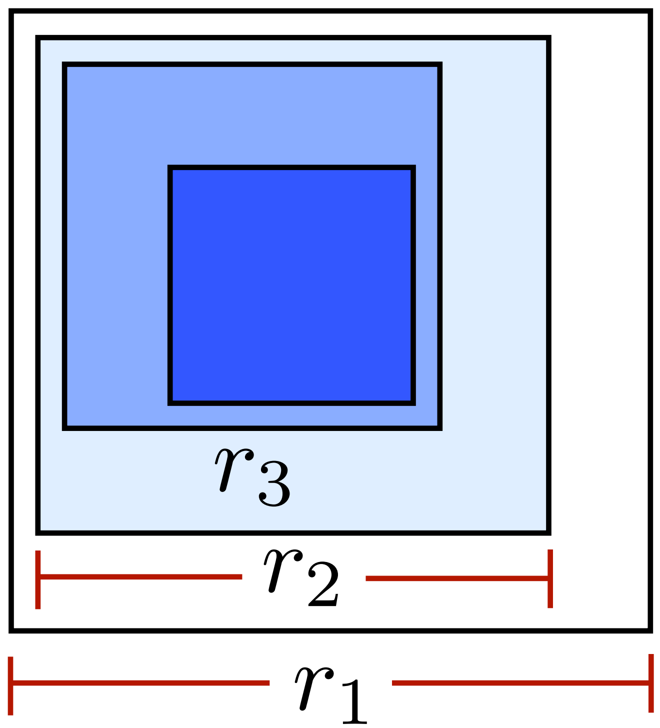
$$\beta = \frac{d}{d+2} \quad \nu = n^{-\frac{1}{d+2}}$$

Recursive shrinking

Box radius decreases as $r_t \leq \nu r_{t-1}^{\beta}$

with $\beta = \frac{d}{d+2}$ or, recursively

$$r_t \leq \nu^{\frac{\beta^t - 1}{\beta - 1}}$$



Recursive shrinking

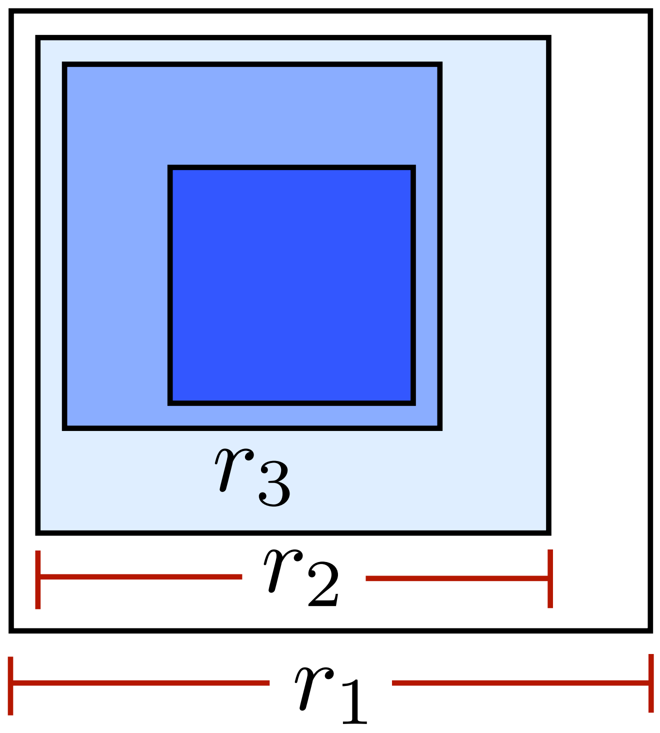
Box radius decreases as $r_t \leq \nu r_{t-1}^{\beta}$

with $\beta = \frac{d}{d+2}$ or, recursively

$$r_t \leq \nu^{\frac{\beta^t - 1}{\beta - 1}}$$

and

$$\nu^{\frac{\beta^t - 1}{\beta - 1}} \lesssim \frac{1}{n} \quad \text{iff} \quad \beta^t \lesssim \frac{1}{\log n}$$



Recursive shrinking

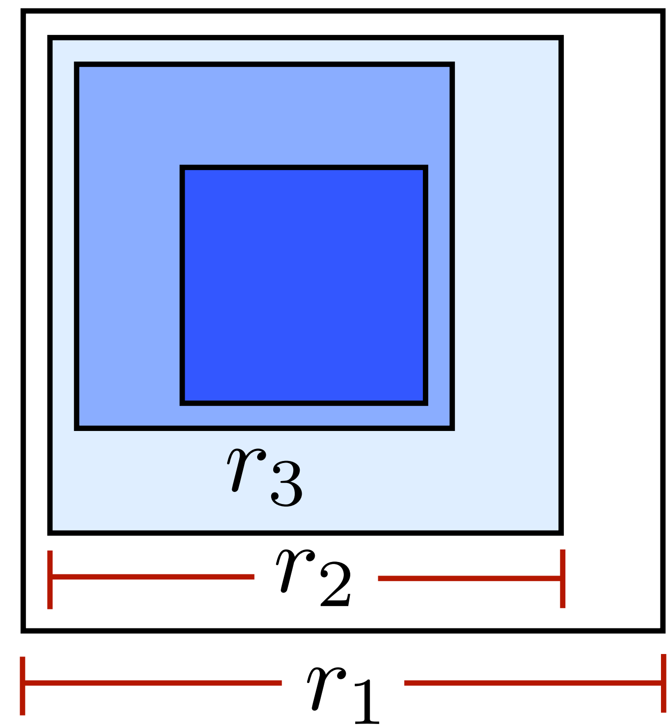
Box radius decreases as $r_t \leq \nu r_{t-1}^{\beta}$

with $\beta = \frac{d}{d+2}$ or, recursively

$$r_t \leq \nu^{\frac{\beta^t - 1}{\beta - 1}}$$

and

$$\nu^{\frac{\beta^t - 1}{\beta - 1}} \lesssim \frac{1}{n} \quad \text{iff} \quad \beta^t \lesssim \frac{1}{\log n}$$



Solution: $$t \gtrsim \frac{\log \log n}{\log 1/\beta} = \frac{\log \log n}{\log(1 + d/2)}$$

Main Results

$$\mathcal{F}_{H,\lambda} := \{\lambda \text{ strongly convex}, H \text{ smooth } f\}$$

Theorem (D., Ruan, Yun 18)

$$\mathfrak{M}_{M,n}(\mathcal{F}_{H,\lambda}) \geq C(d,n) \cdot n^{-\left(1-\left(\frac{d}{d+2}\right)^M\right)}$$

where $C(d,n) \gg \dfrac{1}{\mathrm{poly}(n)}$

Theorem (D., Ruan, Yun 18)

If $M \leq (d/2) \log\log n$ then there is an algorithm s.t.

$$\mathbb{P}\left(f(\widehat{x}) - f^\star \geq C n^{-\left(1-\left(\frac{d}{d+2}\right)^M\right)} \log n\right) \to 0$$

Lower bound

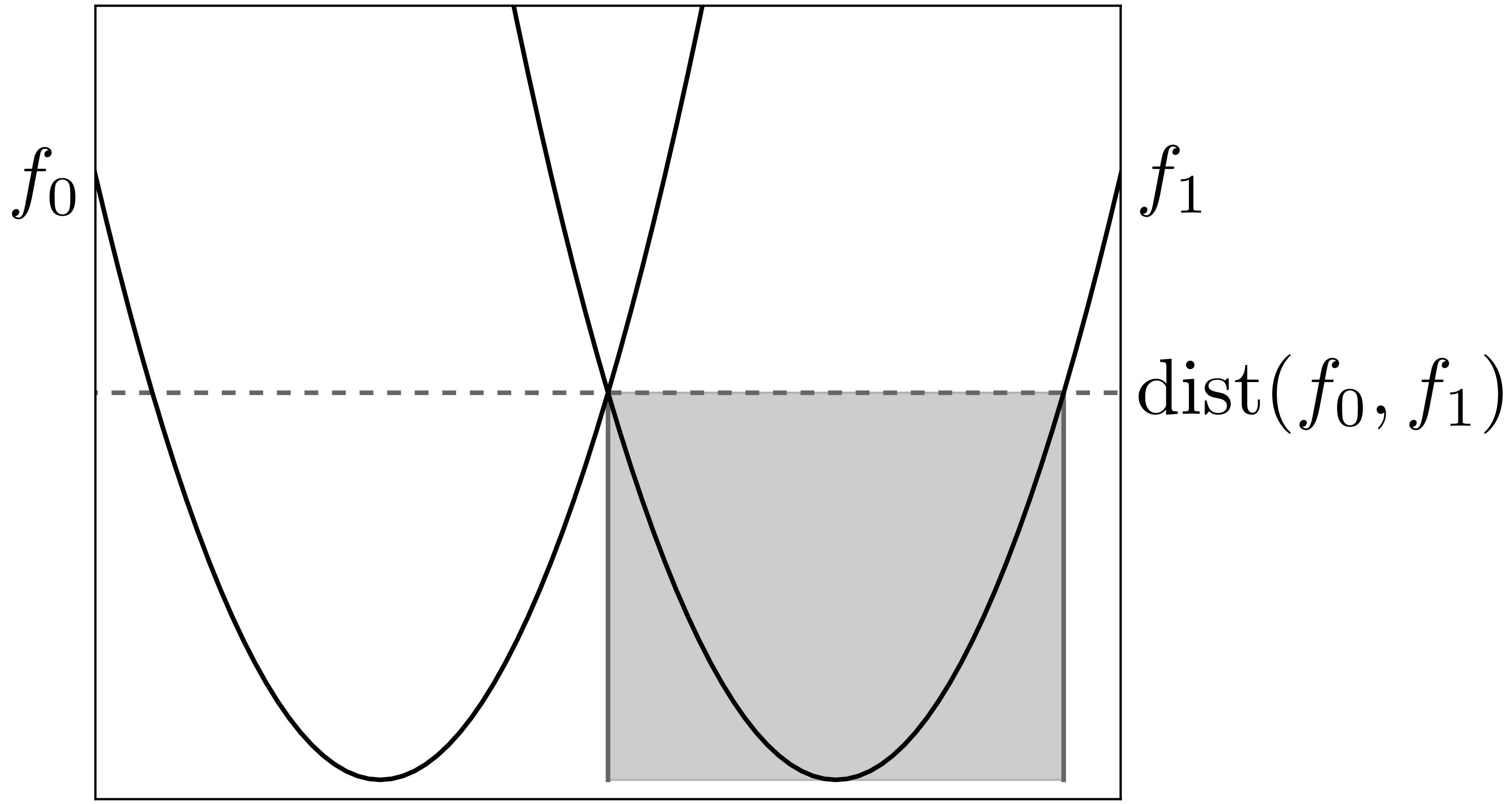
How do we prove lower bounds?

1. Two functions, where optimizing one means *not* optimizing the other [Agarwal BRW 13, Duchi 17]
2. Make information available to algorithm to distinguish them small

Lower bound

How do we prove lower bounds?

1. Two functions, where optimizing one means *not* optimizing the other [Agarwal BRW 13, Duchi 17]
2. Make information available to algorithm to distinguish them small



Lower bound

How do we prove lower bounds?

1. Two functions, where optimizing one means *not* optimizing the other [Agarwal BRW 13, Duchi 17]
2. Make information available to algorithm to distinguish them small

Lower bound

How do we prove lower bounds?

1. Two functions, where optimizing one means *not* optimizing the other [Agarwal BRW 13, Duchi 17]
2. Make information available to algorithm to distinguish them small

$$\inf_{\widehat{x}} \max_{v \in \{0,1\}} \mathbb{E}\left[f_v(\widehat{x}) - f_v^\star\right]$$

$$\geq \frac{\mathrm{dist}(f_0, f_1)}{2} \inf_{\mathrm{Alg}\ \mathsf{A}} \mathbb{P}\left(\mathsf{A} \text{ distinguishes } f_0, f_1\right)$$

Lower bound

How do we prove lower bounds?

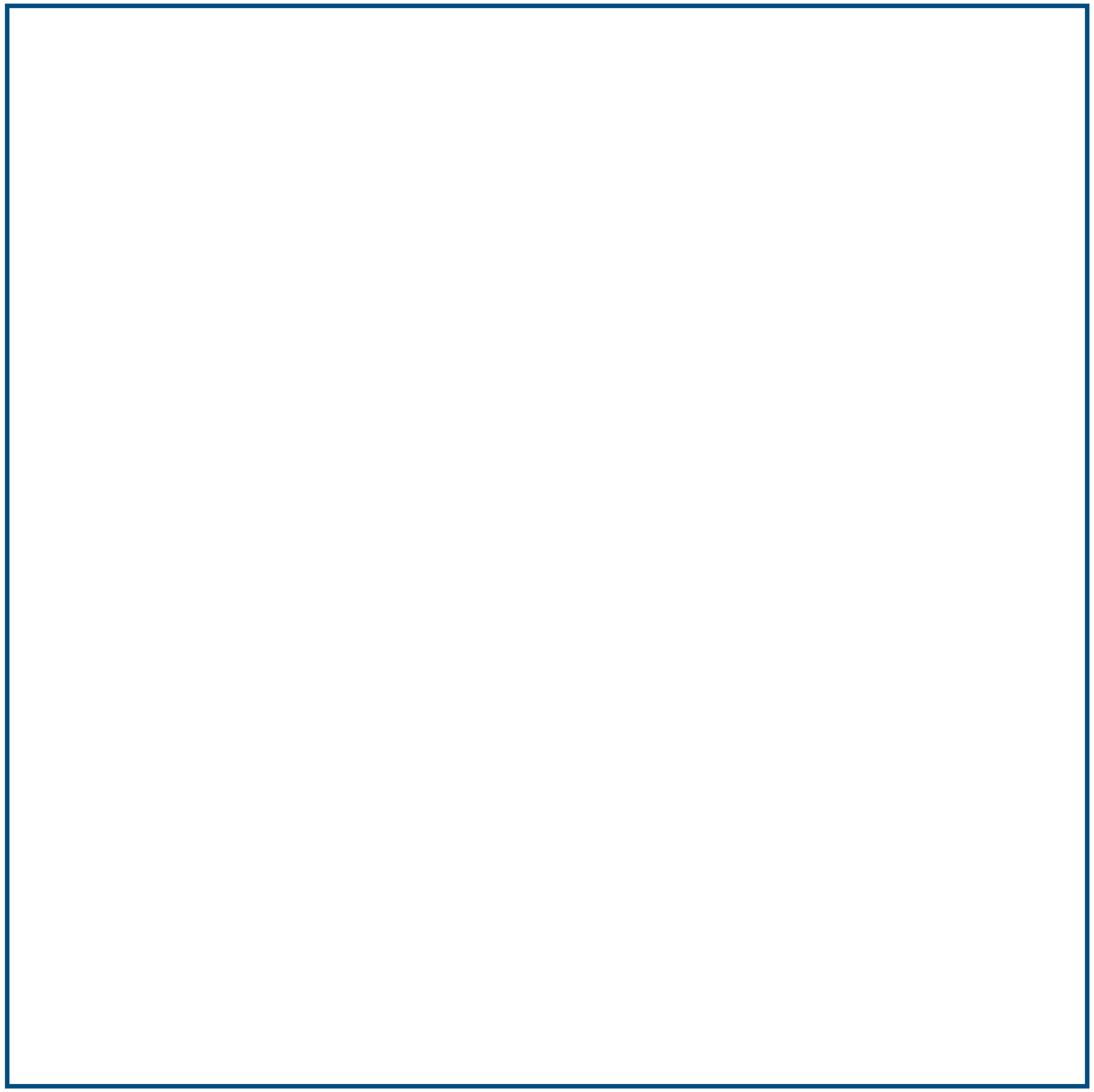
1. Two functions, where optimizing one means *not* optimizing the other [Agarwal BRW 13, Duchi 17]
2. Make information available to algorithm to distinguish them small

$$\inf_{\widehat{x}} \max_{v \in \{0,1\}} \mathbb{E}\left[f_v(\widehat{x}) - f_v^\star\right]$$

$$\geq \frac{\mathrm{dist}(f_0, f_1)}{2} \underbrace{\inf_{\text{Alg } \mathsf{A}} \mathbb{P}\left(\mathsf{A} \text{ distinguishes } f_0, f_1\right)}_{=1-\|P_0 - P_1\|_{\rm TV}}$$

Lower bound: recursive packing

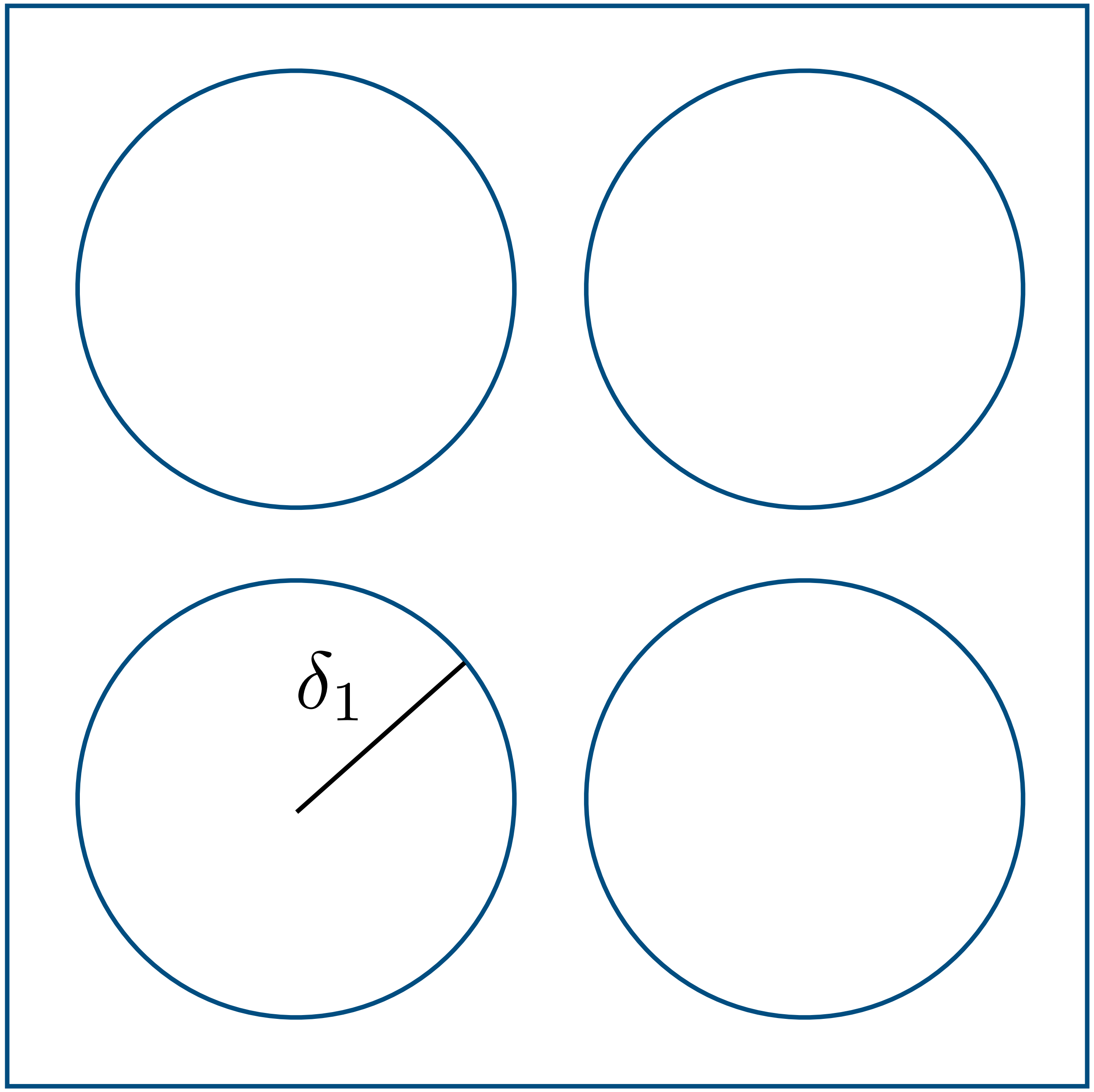
$$\frac{1}{2} \geq \delta_1 \geq \delta_2 \geq \cdots \geq \delta_M$$



Lower bound: recursive packing

$$\frac{1}{2} \geq \delta_1 \geq \delta_2 \geq \cdots \geq \delta_M$$

$\mathcal{U}^{(1)} = \delta_1$ packing of initial set

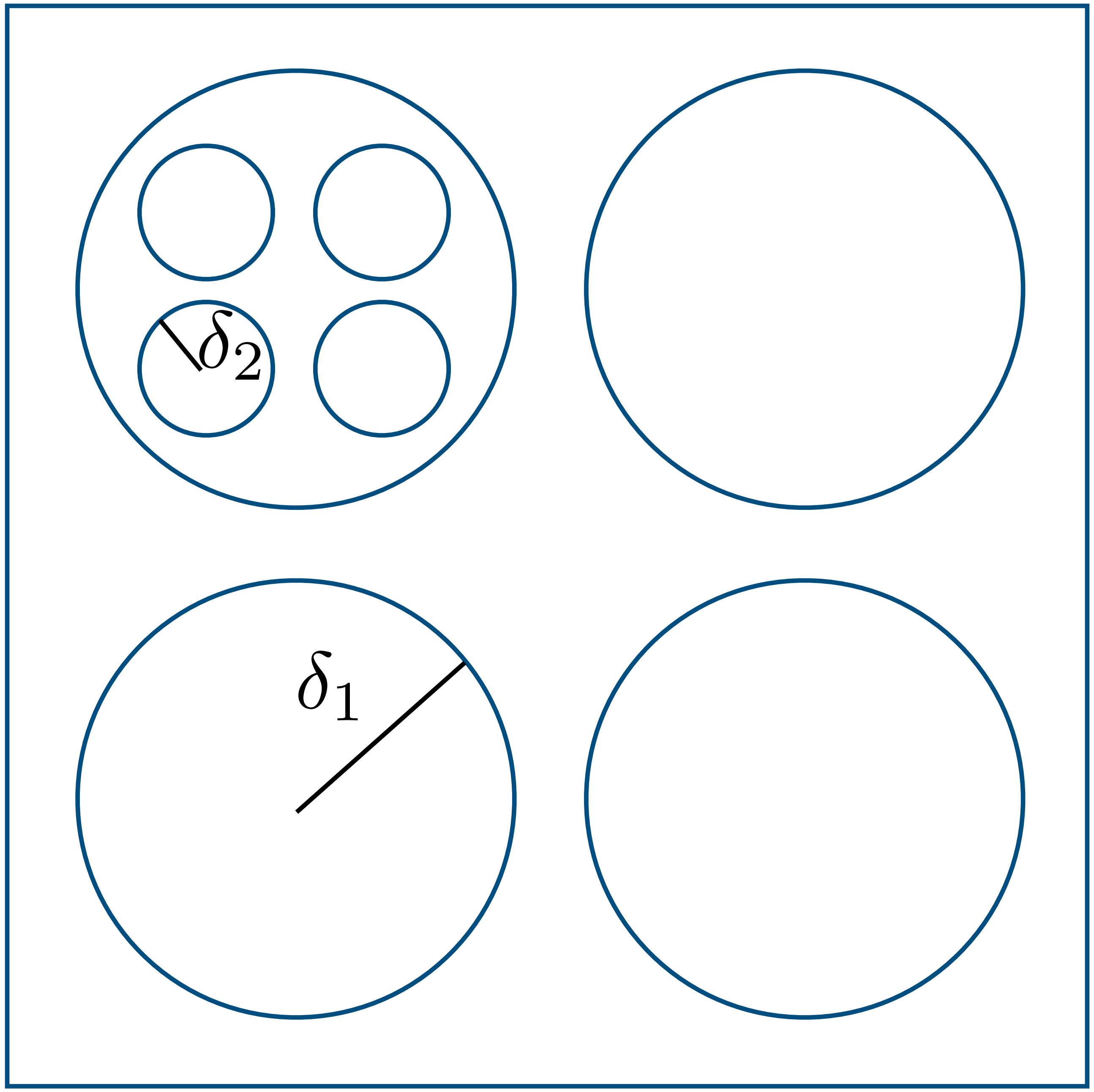


Lower bound: recursive packing

$$\frac{1}{2} \geq \delta_1 \geq \delta_2 \geq \cdots \geq \delta_M$$

$\mathcal{U}^{(1)} = \delta_1$ packing of initial set

$\mathcal{U}_u^{(t)} = 2\delta_t$ packing of ball centered at u

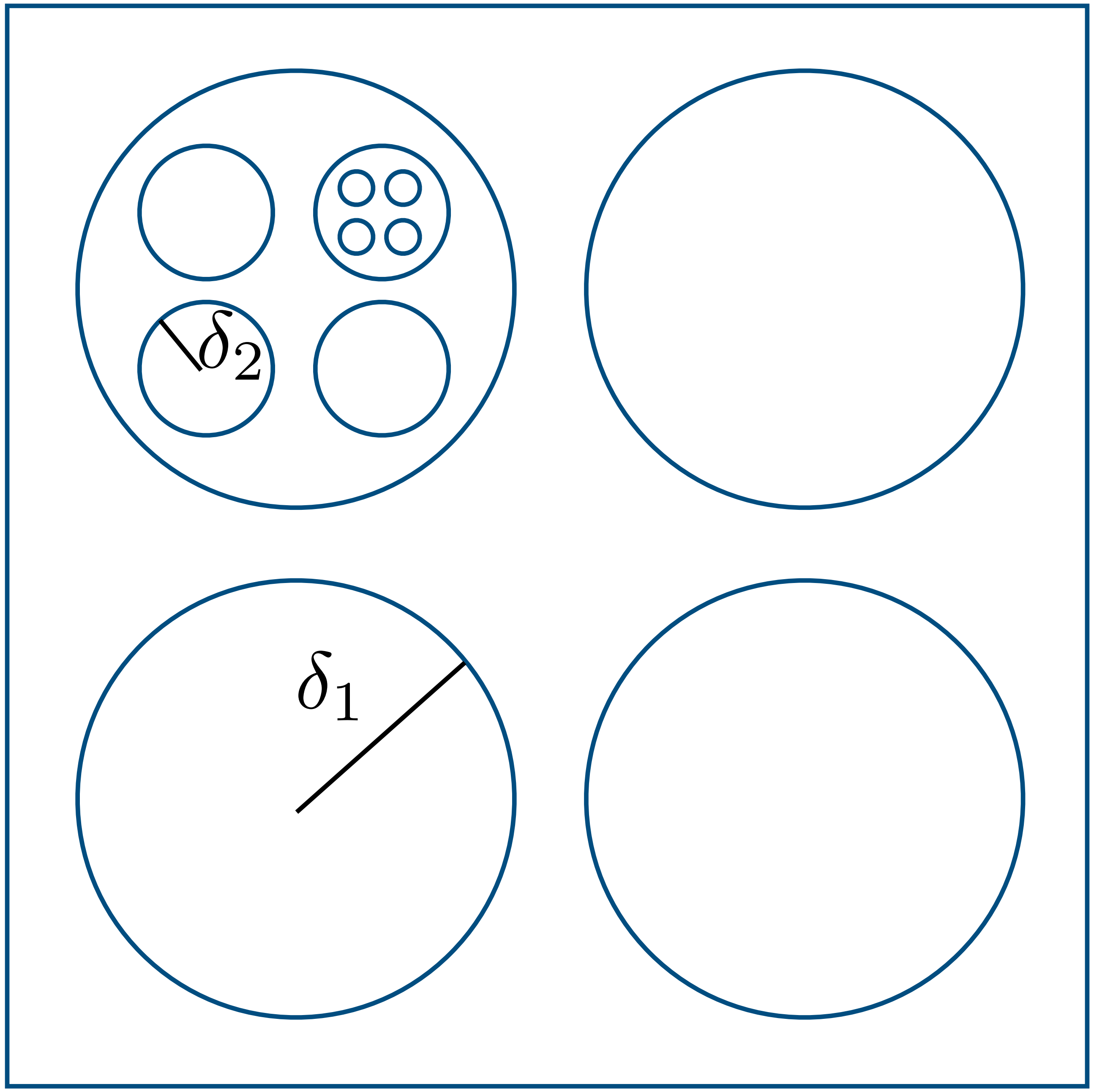


Lower bound: recursive packing

$$\frac{1}{2} \geq \delta_1 \geq \delta_2 \geq \cdots \geq \delta_M$$

$\mathcal{U}^{(1)} = \delta_1$ packing of initial set

$\mathcal{U}_u^{(t)} = 2\delta_t$ packing of ball centered at u



Lower bound: recursive packing

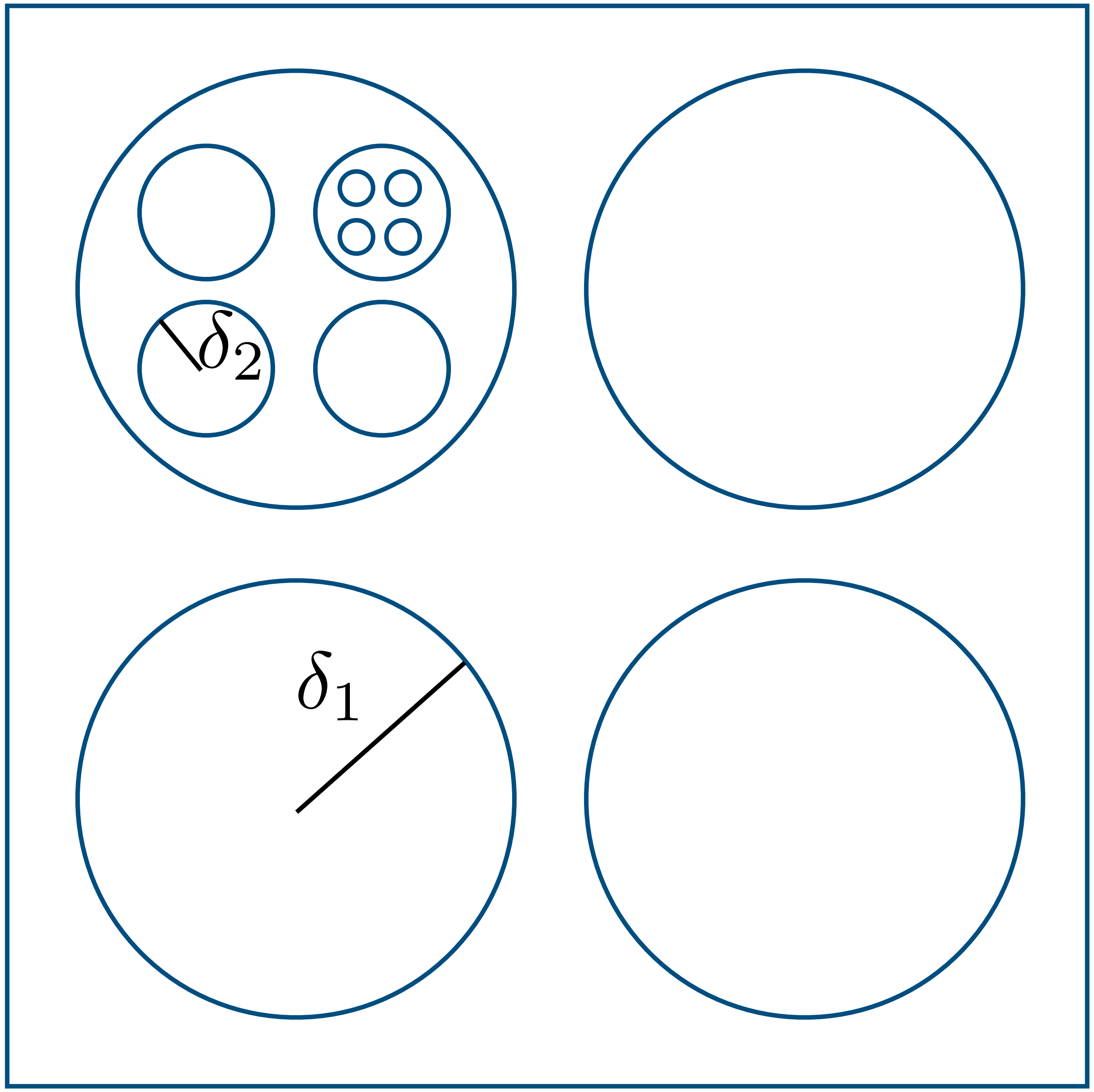
$$\frac{1}{2} \geq \delta_1 \geq \delta_2 \geq \cdots \geq \delta_M$$

$\mathcal{U}^{(1)} = \delta_1$ packing of initial set

$\mathcal{U}_u^{(t)} = 2\delta_t$ packing of ball centered at u

Idea:

1. define functions recursively on balls
2. optimization means identifying ball



Function constructions

$$\frac{1}{2} \geq \delta_1 \geq \delta_2 \geq \cdots \geq \delta_M$$

- Index functions by path down $u_{1:M} = (u_1, \ldots, u_M)$

Function constructions

$$\frac{1}{2} \geq \delta_1 \geq \delta_2 \geq \cdots \geq \delta_M$$

- Index functions by path down $u_{1:M} = (u_1, \ldots, u_M)$

$$f^{(1)}_{u_{1:M}}(x) = f^{(0)}_{u_{1:M}}(x) \qquad x \notin u_M + \delta_M \mathbb{B}$$

Function constructions

$$\frac{1}{2} \geq \delta_1 \geq \delta_2 \geq \cdots \geq \delta_M$$

- Index functions by path down $u_{1:M} = (u_1, \ldots, u_M)$

$$f^{(1)}_{u_{1:M}}(x) = f^{(0)}_{u_{1:M}}(x) \qquad x \notin u_M + \delta_M \mathbb{B}$$

$$f^{(\pm 1)}_{u_{1:M}}(x) = f^{(\pm 1)}_{u_{1:t}, \tilde{u}_{t+1:M}}(x) \qquad x \notin u_t + \delta_t \mathbb{B}$$

Function constructions

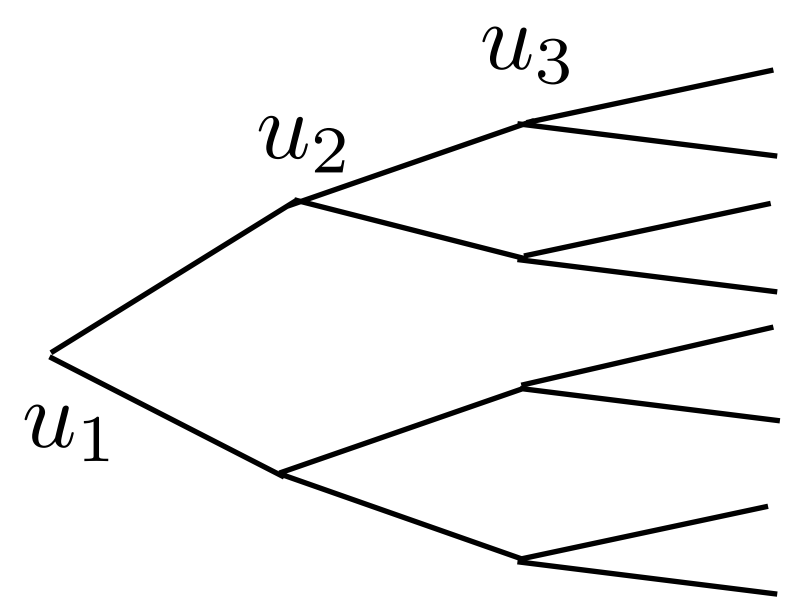
$$\frac{1}{2} \geq \delta_1 \geq \delta_2 \geq \cdots \geq \delta_M$$

- Index functions by path down $u_{1:M} = (u_1, \ldots, u_M)$

$$f^{(1)}_{u_{1:M}}(x) = f^{(0)}_{u_{1:M}}(x) \qquad x \notin u_M + \delta_M \mathbb{B}$$

$$f^{(\pm 1)}_{u_{1:M}}(x) = f^{(\pm 1)}_{u_{1:t}, \tilde{u}_{t+1:M}}(x) \qquad x \notin u_t + \delta_t \mathbb{B}$$

Optimizing well means identifying sequence defining function

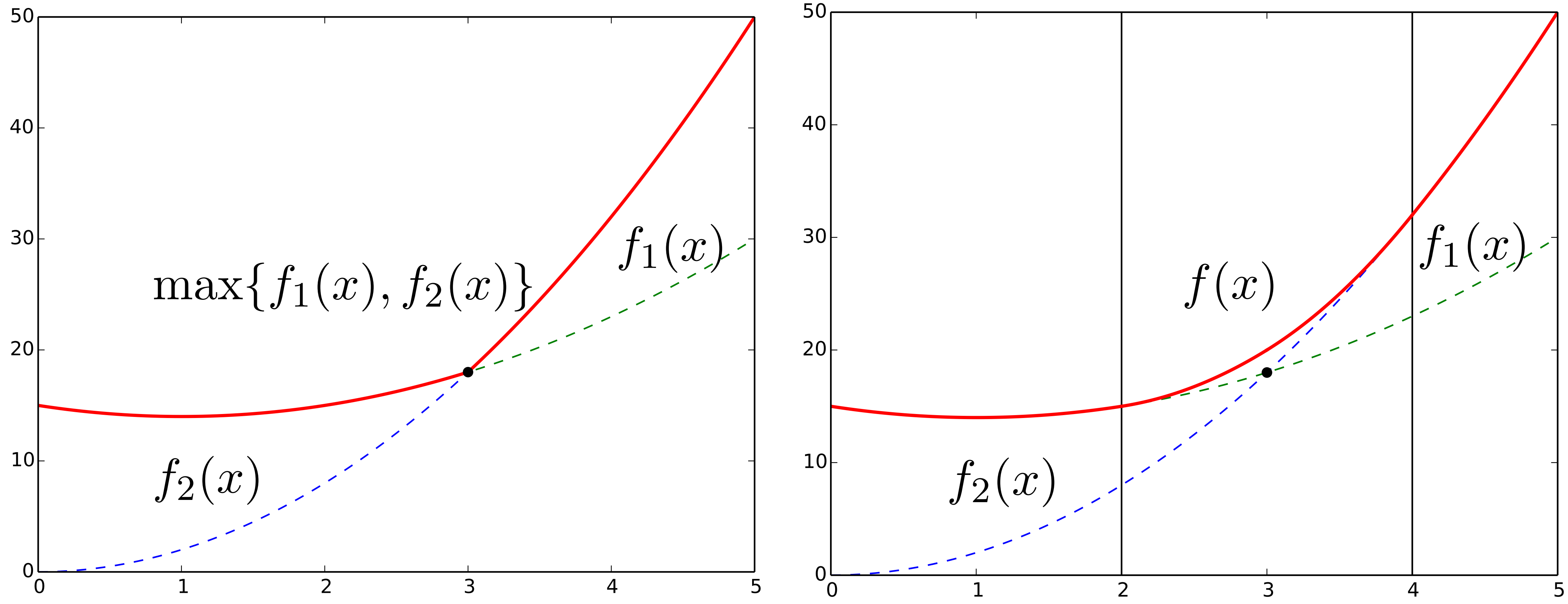


Function construction

$$f_{u_1}(x) = \frac{1}{2}\|x - u_1\|^2$$

Recurse:

$$f_{u_{1:t}}(x) = \mathrm{SmoothMax}\{f_{u_{1:t-1}}(x), \|x - u_t\|^2 + b_t\}$$



Information recursion

At each round t (of M):

$$D_{\mathrm{KL}}(\nabla f_{u_{1:M}}(x) + \xi \| \nabla f_{u_{1:t}, \tilde{u}_{t+1:M}}(x) + \xi) \lesssim \delta_t^2$$

Information recursion

At each round t (of M):

$$D_{\mathrm{KL}}(\nabla f_{u_{1:M}}(x) + \xi \| \nabla f_{u_{1:t}, \tilde{u}_{t+1:M}}(x) + \xi) \lesssim \delta_t^2$$

Choose radius for “constant” information per round:

$$\delta_t^2 \frac{n}{\#\text{ in packing}} \approx 1$$

Information recursion

At each round t (of M):

$$D_{\mathrm{KL}}(\nabla f_{u_{1:M}}(x) + \xi \| \nabla f_{u_{1:t}, \tilde{u}_{t+1:M}}(x) + \xi) \lesssim \delta_t^2$$

Choose radius for "constant" information per round:

$$\delta_t^2 \frac{n \delta_t^d}{\delta_{t-1}^d} \approx 1$$

Information recursion

At each round t (of M):

$$D_{\mathrm{KL}}(\nabla f_{u_{1:M}}(x) + \xi \| \nabla f_{u_{1:t}, \tilde{u}_{t+1:M}}(x) + \xi) \lesssim \delta_t^2$$

Choose radius for “constant” information per round:

$$\delta_t^2 \frac{n \delta_t^d}{\delta_{t-1}^d} \approx 1$$

Solution for lower bound:

$$\delta_M = n^{-\frac{1}{d+2}} \delta_{M-1}^{\frac{d}{d+2}} = n^{-\frac{1}{2}\left(1 - \left(\frac{d}{d+2}\right)^M\right)}$$