Beyond Online-to-Batch: Exploring the Generalization Ability of (Convex) SGD

Tomer Koren
Tel Aviv University & Google

Disclaimer

No new (faster/better/…) algorithms

No new framework/model/setting

new aspects, perspectives, discoveries

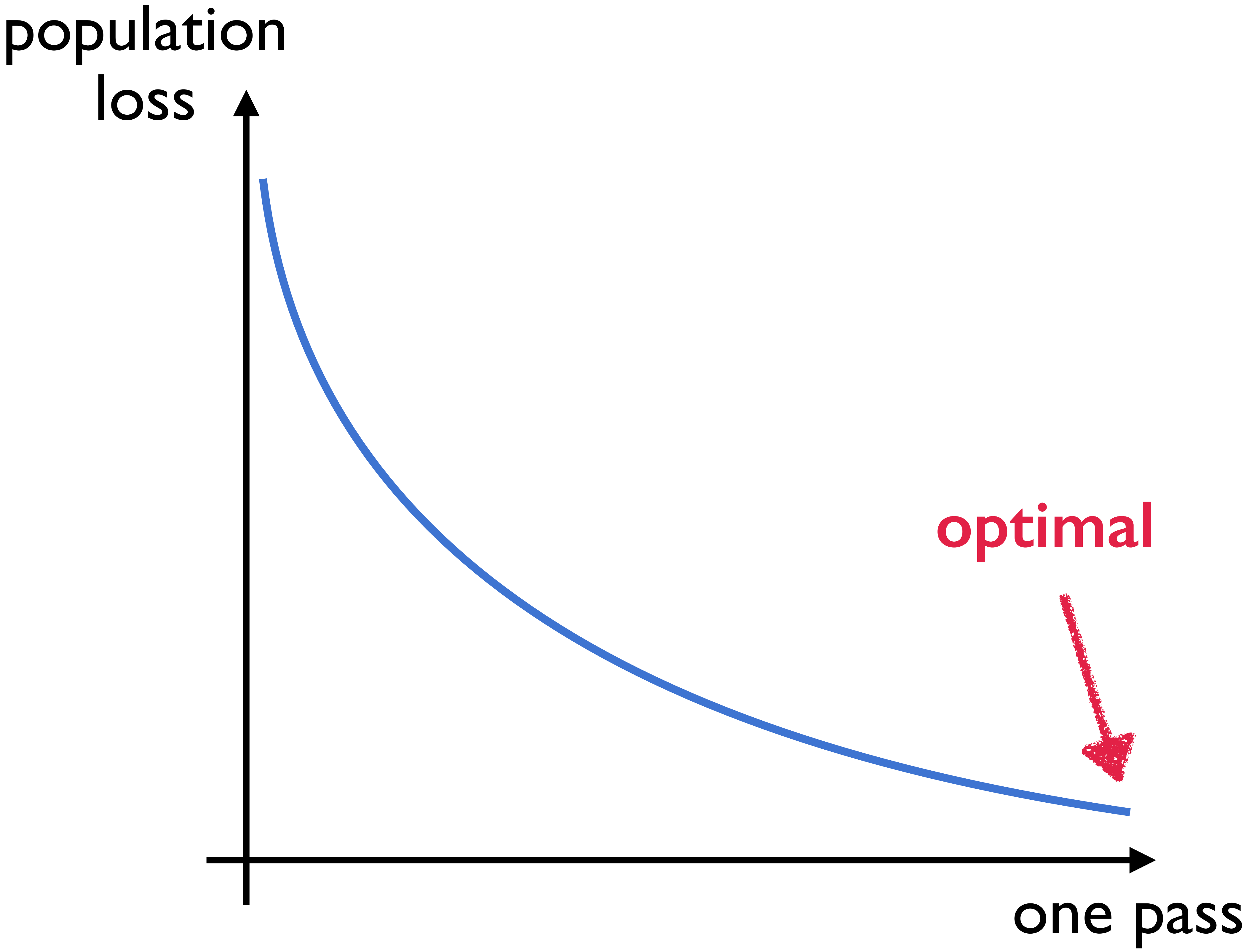
about already great algorithms

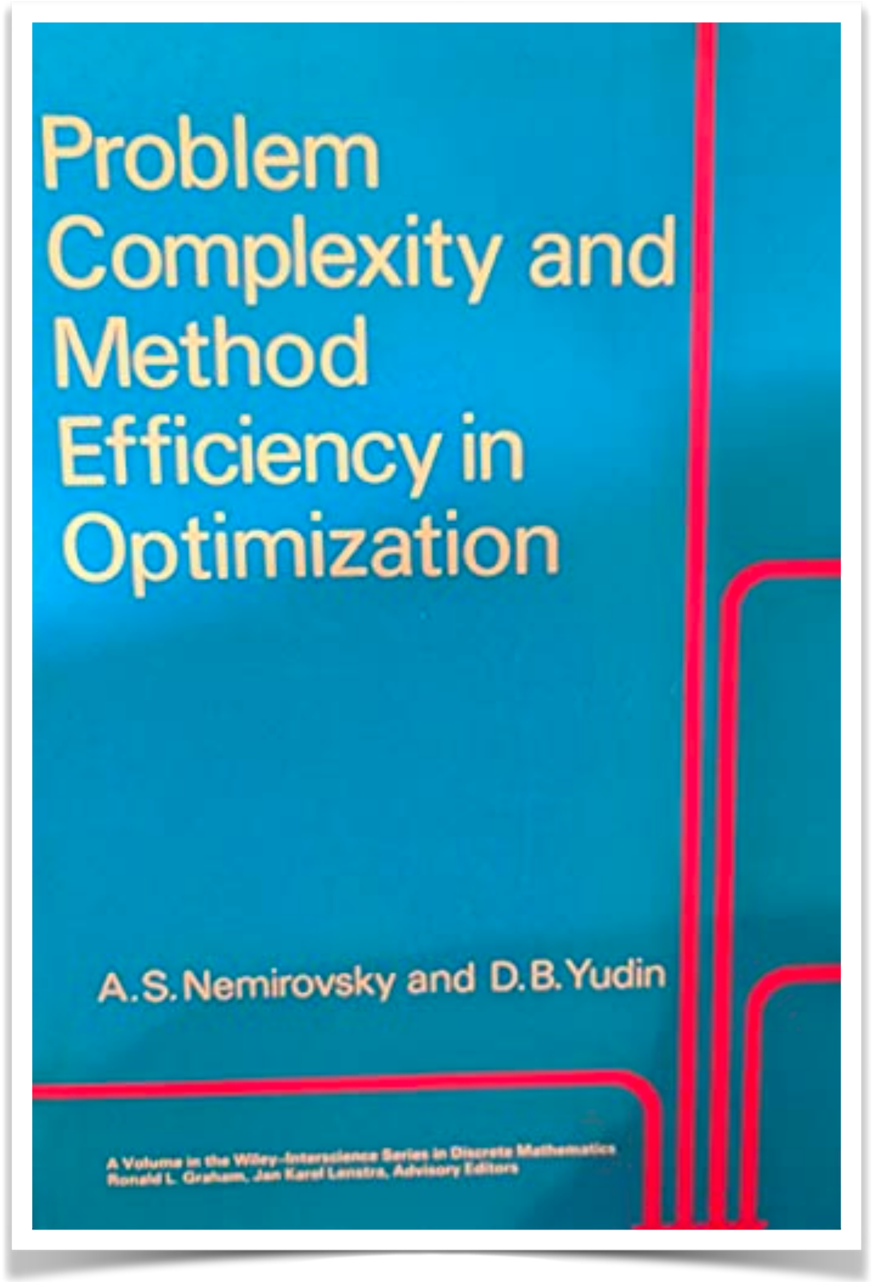
in an already great setting

Outline

Revisit a fundamental result, from statistical learning perspective:

out-of-sample performance of SGD is minimax optimal in convex optimization





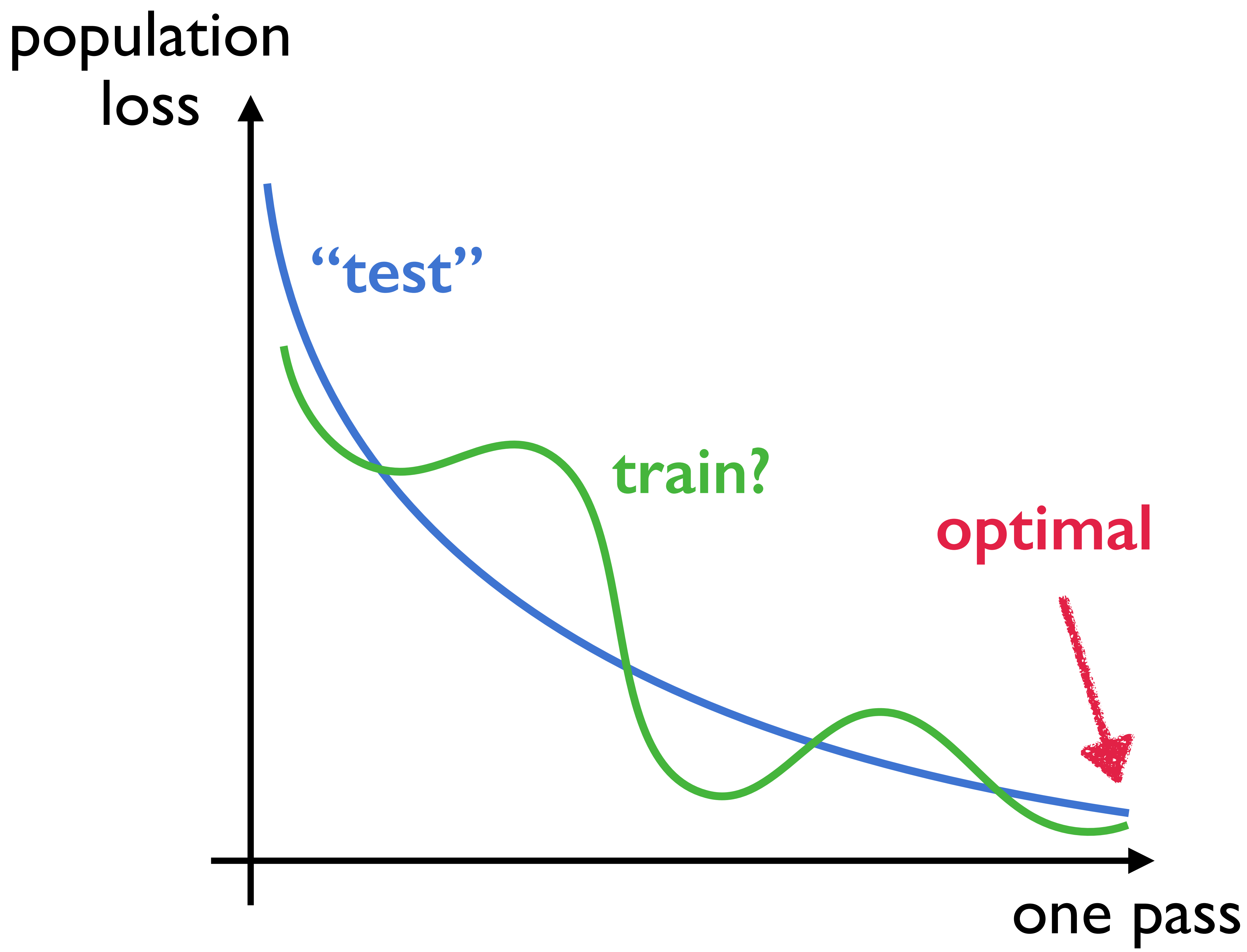
[Nemirovsky & Yudin ’83]

modern view: direct consequence of regret + online-to-batch conversion

Outline

Revisit a fundamental result, from statistical learning perspective:

out-of-sample performance of SGD is minimax optimal in convex optimization

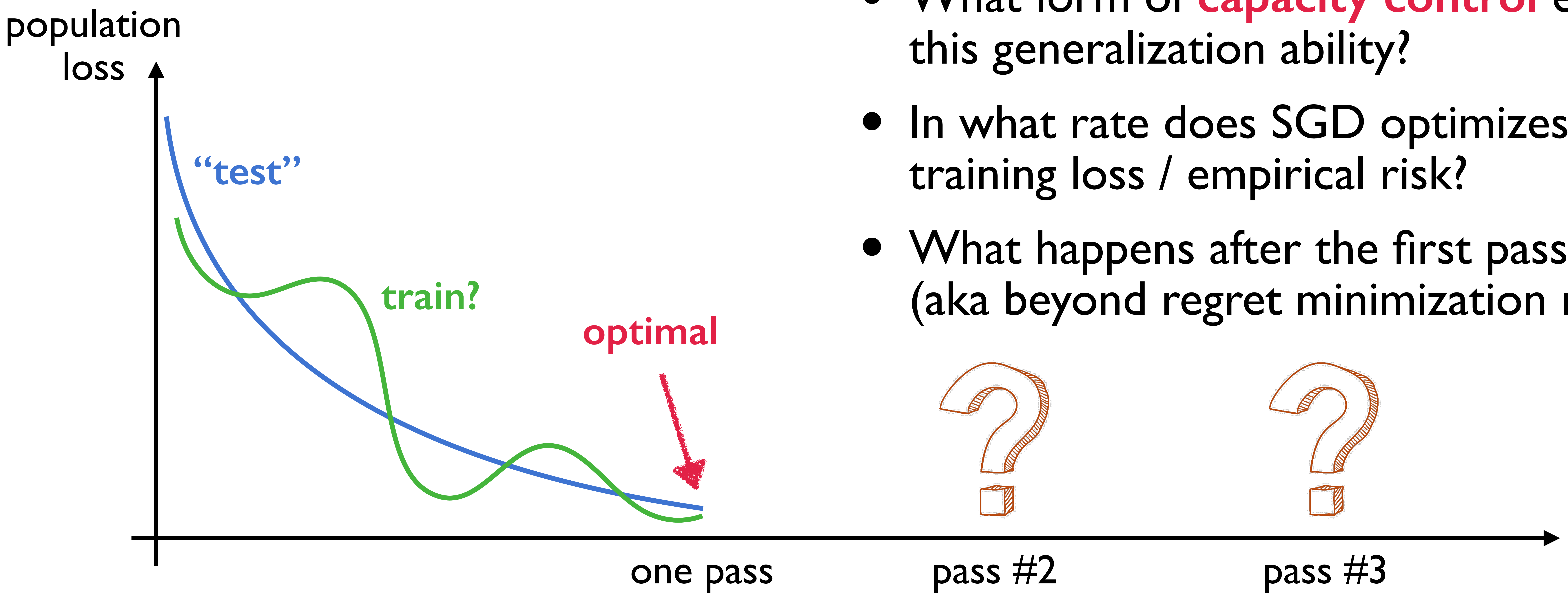


- What form of **capacity control** enables this generalization ability?
- In what rate does SGD optimizes the training loss / empirical risk?

Outline

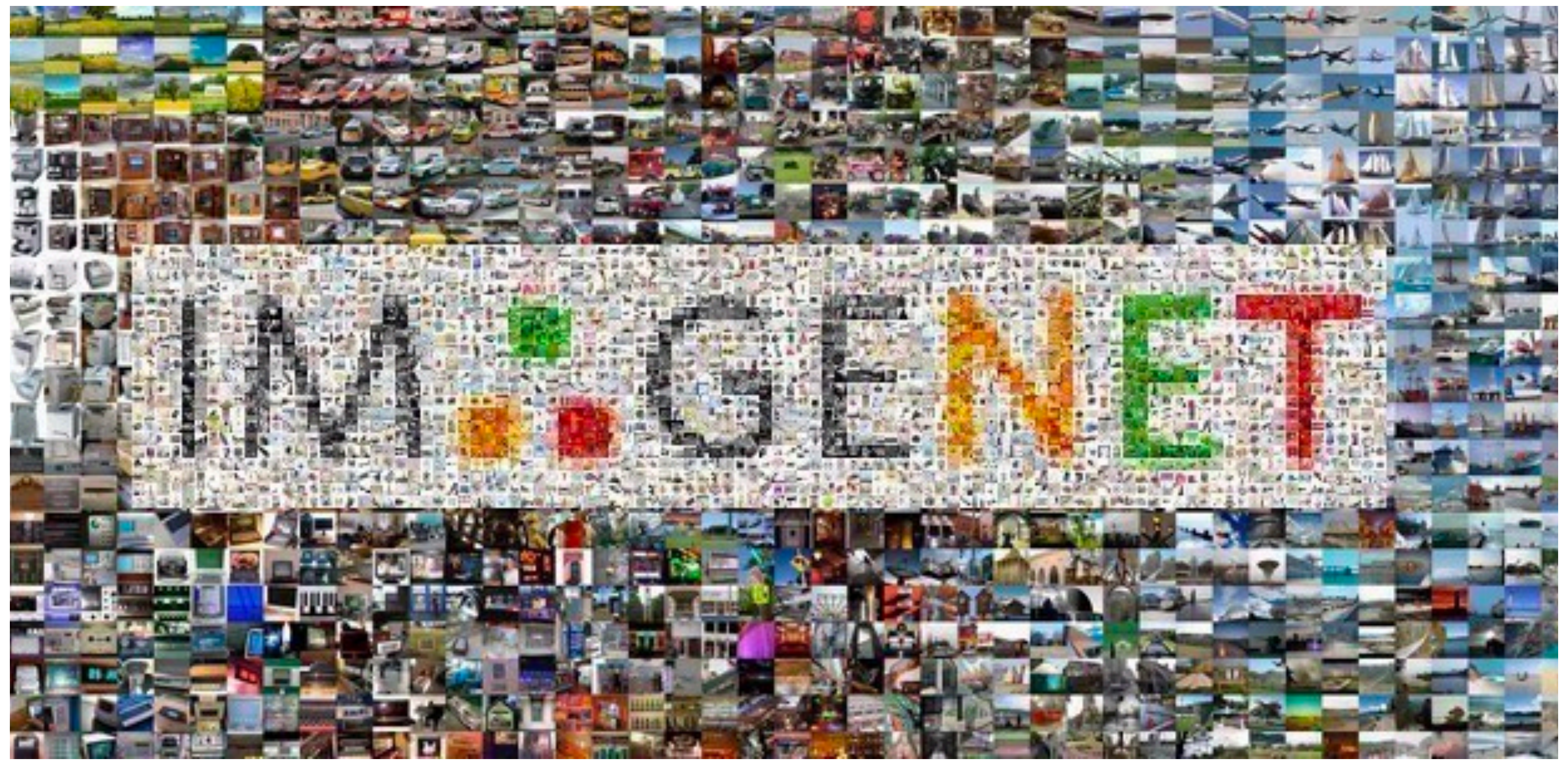
Revisit a fundamental result, from statistical learning perspective:

out-of-sample performance of SGD is minimax optimal in convex optimization

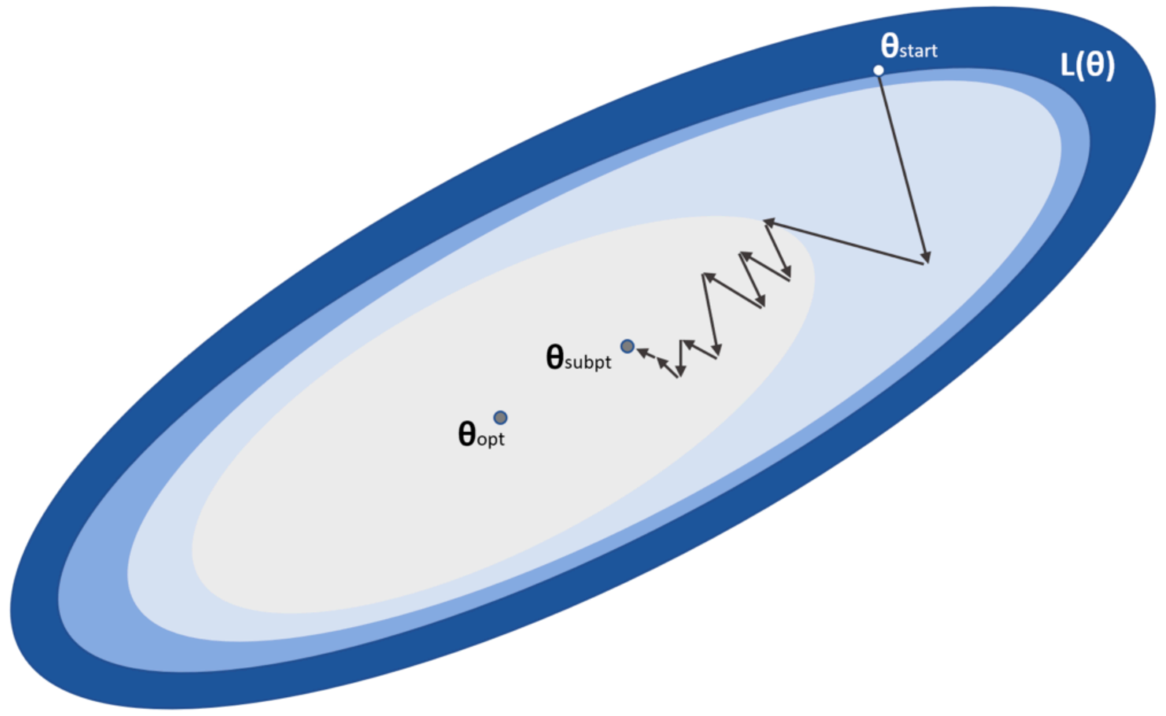


- What form of **capacity control** enables this generalization ability?
- In what rate does SGD optimizes the training loss / empirical risk?
- What happens after the first pass? (aka beyond regret minimization regime)

Generalization



Data samples
$z_1, \ldots, z_n$

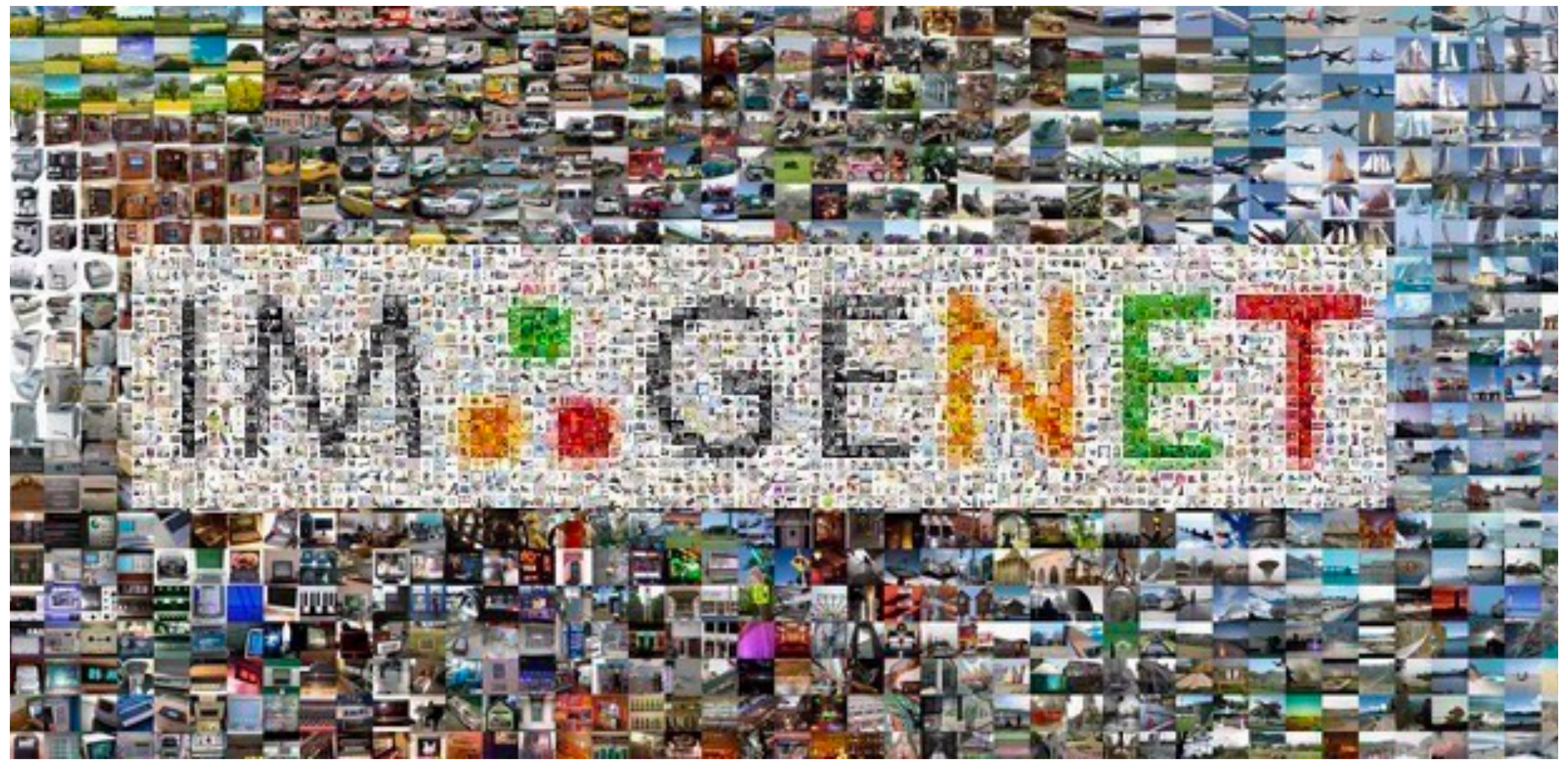


Optimization algorithm

$\hat{w}$

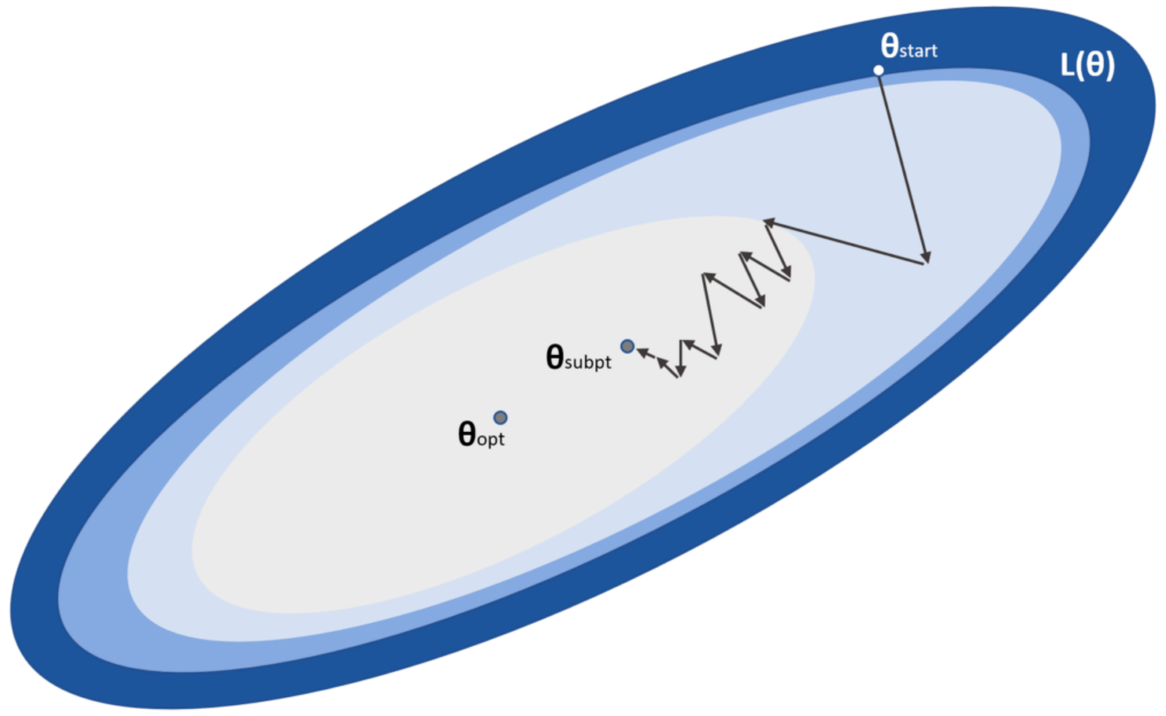
Predictive model

Generalization



Data samples

$z_1, \ldots, z_n$



Optimization algorithm

$$\hat{w}$$

Predictive model

$$\hat{F}(w) = \frac{1}{n}\sum_{i=1}^{n} f(w, z_i)$$

Empirical/training risk

generalization

$z_i \sim_{iid} D$ (unknown)

real goal

$$F(w) = \mathbb{E}_{z \sim D}\big[f(w, z)\big]$$

True/population risk

Setup: Stochastic <u>Convex</u> Optimization (SCO)

- Convex, 1-Lipschitz loss $f(w, z)$
- Goal: minimize population/true risk:

loss func.

$$F(w) = \mathbb{E}_{z \sim D}\left[f(w, z)\right]$$

data dist. (unknown)

model $\in \mathbb{R}^d$

data sample

- Access to sample $S = \{z_1, \ldots, z_n\} \sim_{iid} D$

Setup: Stochastic Convex Optimization (SCO)

- Convex, 1-Lipschitz loss $f(w, z)$
- Goal: minimize population/true risk:

loss func.

$$F(w) = \mathbb{E}_{z \sim D}\left[f(w, z)\right]$$

data dist. (unknown)

model $\in \mathbb{R}^d$

data sample

- Access to sample $S = \{z_1, \ldots, z_n\} \sim_{iid} D$

- Empirical risk:

$$\hat{F}(w) = \frac{1}{n}\sum_{i=1}^{n} f(w, z_i)$$

- Generalization gap:

$$\Delta(w) = \hat{F}(w) - F(w)$$

Generalization in SCO is (still) compelling

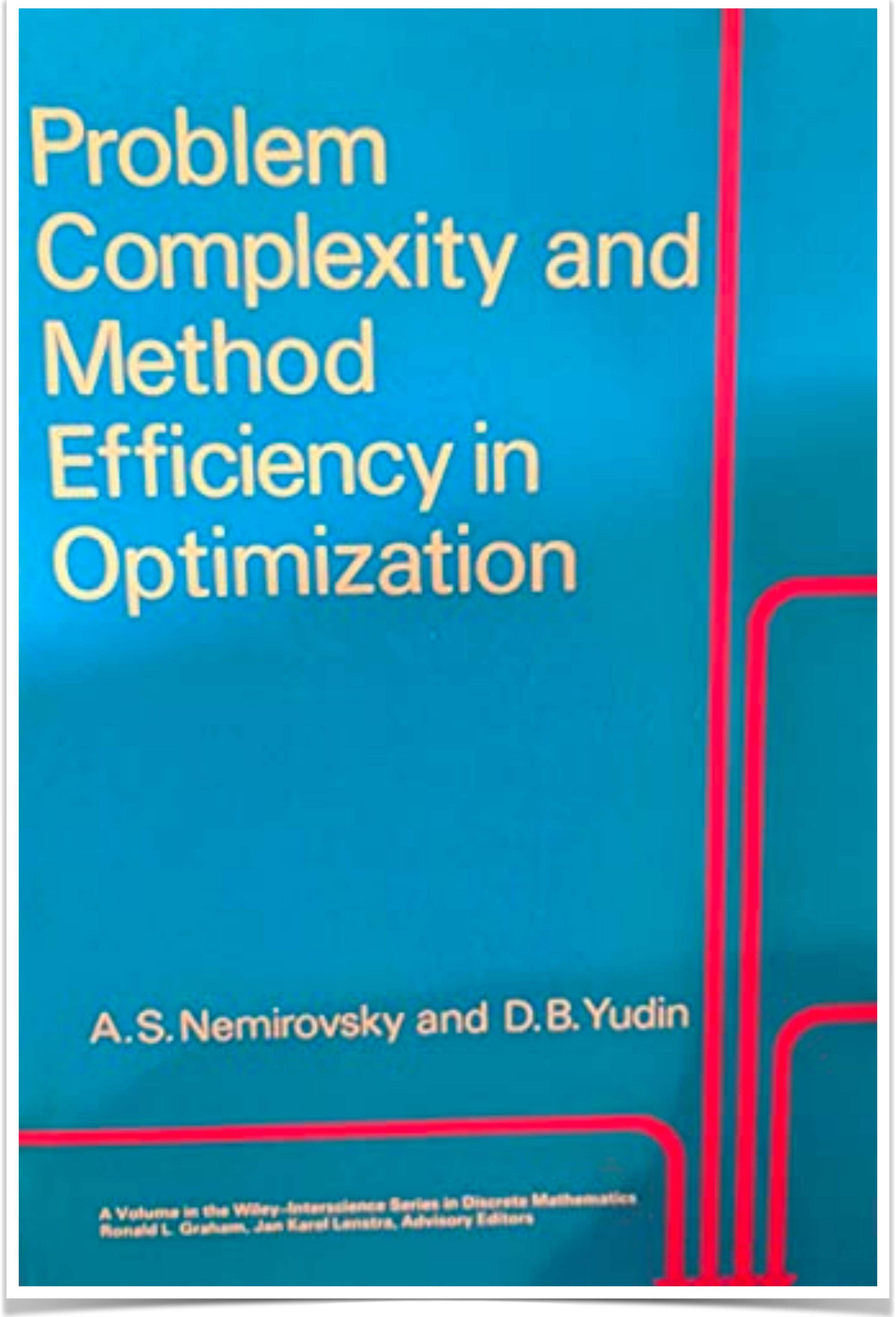
- Fundamental in optimization, extremely well studied
- Home to common, real-world algorithms (incl. SGD)
- Global optimization is "easy"—allows isolating generalization

Generalization in SCO is (still) compelling

- Fundamental in optimization, extremely well studied
- Home to common, real-world algorithms (incl. SGD)
- Global optimization is "easy"—allows isolating generalization

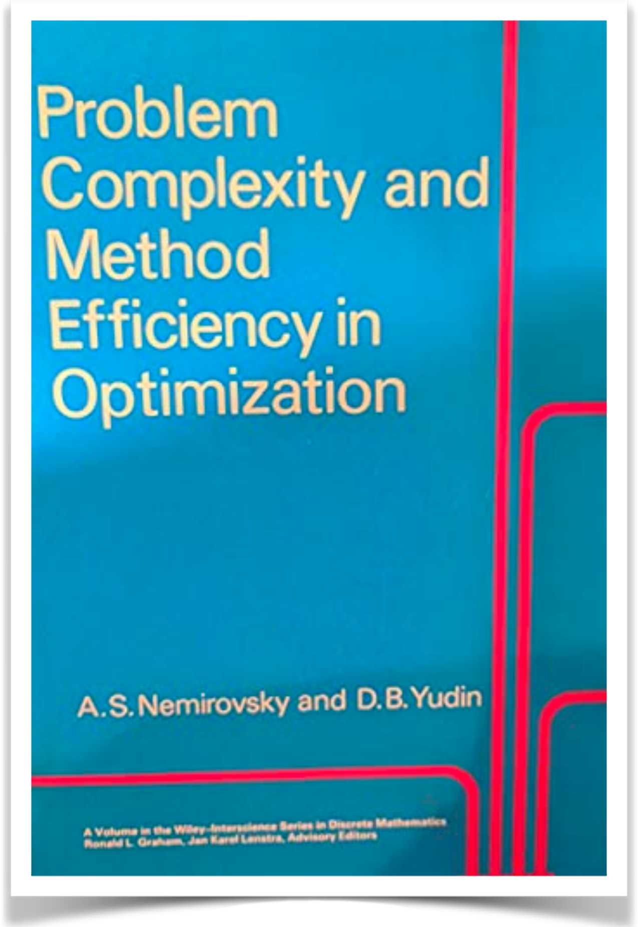
- Generalization in SCO known to be algorithm-dependent (unlike in other classical models: PAC, GLMs, …) [Shalev-Shwartz, Shamir, Srebro, Sridharan '09]
- Instructive to demonstrate interesting (generalization) phenomena in simple cases

SGD is minimax optimal in SCO



[Nemirovsky & Yudin '83]

SGD is minimax optimal in SCO

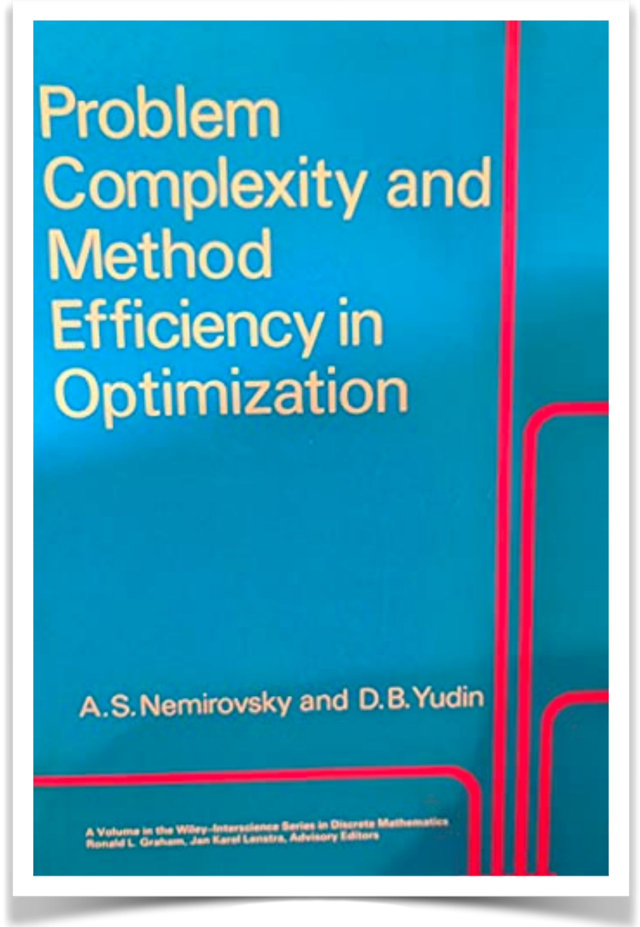


SGD
(one pass)

$$w_0 = \text{init}$$

$$w_{t+1} = w_t - \eta \nabla_w f(w_t, z_t)$$

SGD is minimax optimal in SCO



SGD
(one pass)

$$\boxed{\begin{aligned} w_0 &= \text{init} \\ w_{t+1} &= w_t - \eta \nabla_w f(w_t, z_t) \end{aligned}}$$

- **Theorem:**
  [N&Y '83]

$$\mathbb{E}\left[F(\hat{w})\right] - F^\star \lesssim \frac{1}{\eta n} + \eta \lesssim \frac{\mathbf{1}}{\sqrt{\boldsymbol{n}}}$$

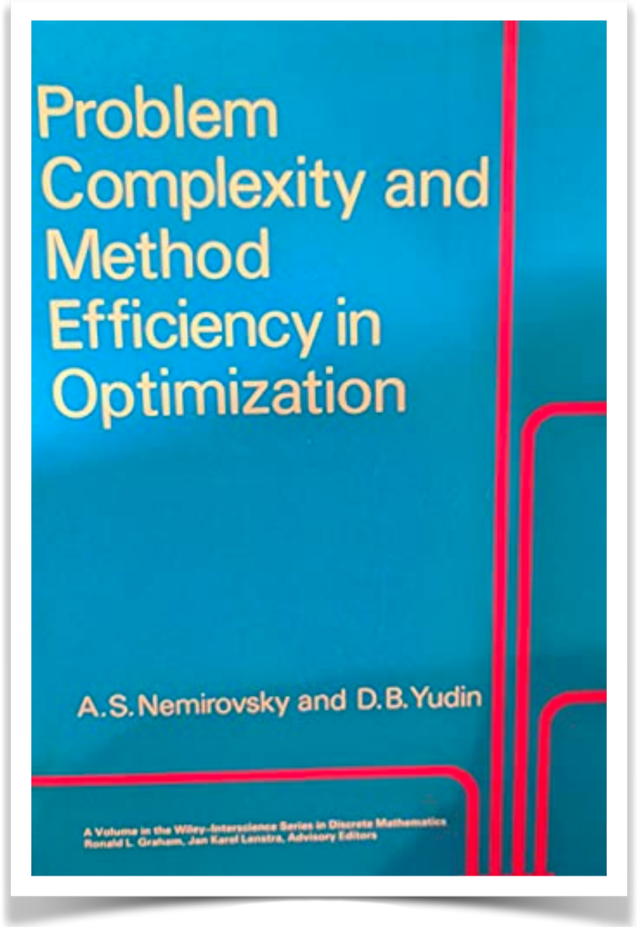
(f convex , G-Lipschitz)

$$\hat{w} = \frac{1}{n}\Sigma_{t=1}^{n} w_t$$

$$\eta \cong \frac{1}{\sqrt{n}}$$

- Bound is **minimax optimal** up to constants, **independent of dimension**
- Modern view: consequence of regret + online-to-batch conversion
- Many extensions: other geometries (mirror descent), adaptive versions, …

Proof from the book

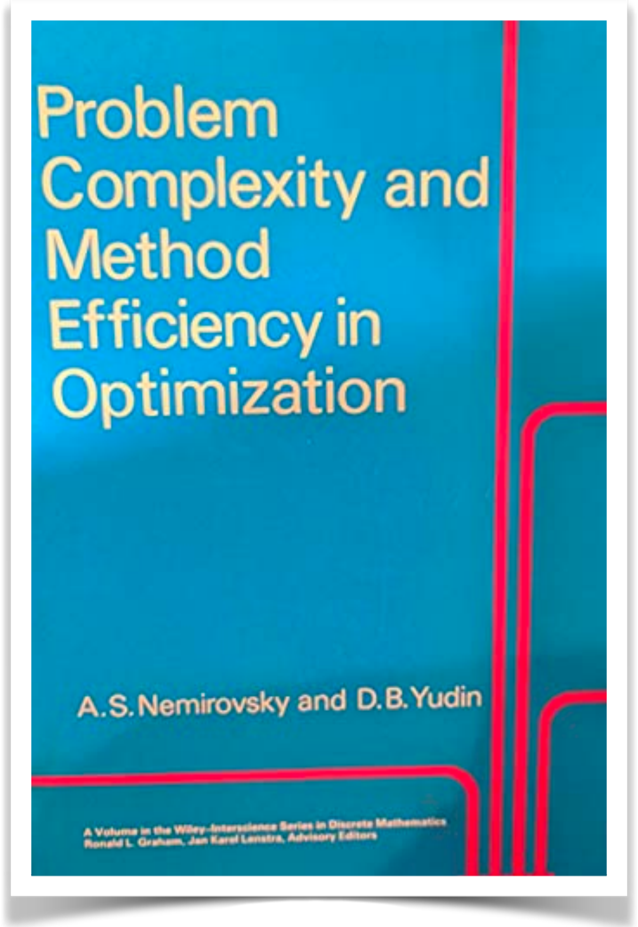


SGD

$$w_{t+1} = w_t - \eta g_t \ ; \qquad \mathbb{E}[g_t \mid w_t] = \nabla F(w_t)$$

$$g_t = \nabla_w f(w_t, z_t)$$

Proof from the book



SGD $\boxed{w_{t+1} = w_t - \eta g_t \; ; \qquad \mathbb{E}[g_t \mid w_t] = \nabla F(w_t)}$

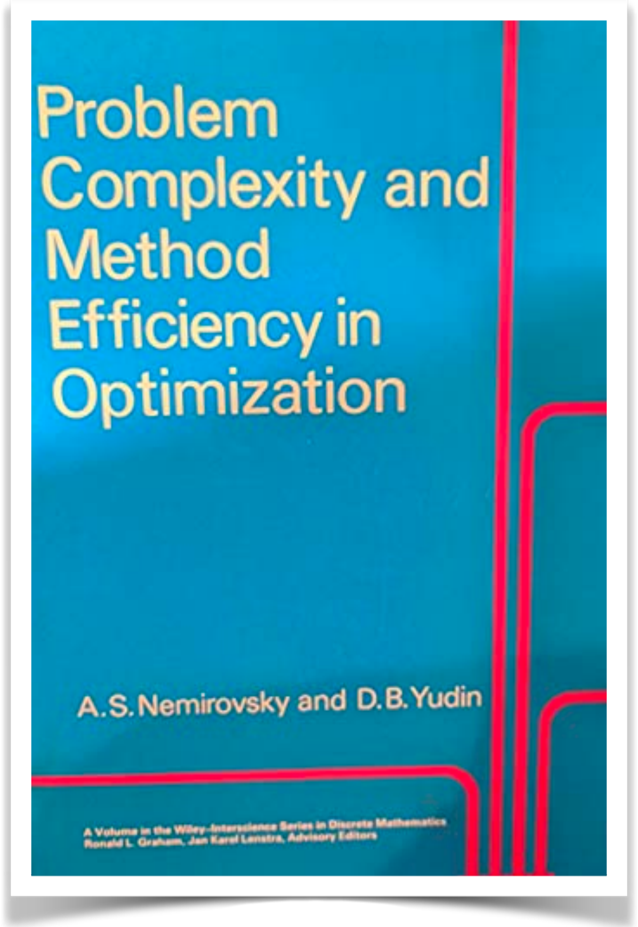
$$\Rightarrow \quad \|w_{t+1} - w^*\|^2 \le \|w_t - w^*\|^2 - 2\eta g_t \cdot (w_t - w^*) + \eta^2 \|g_t\|^2$$

$$\Rightarrow \quad g_t \cdot (w_t - w^*) \le \frac{1}{2\eta}\left(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2\right) + \frac{\eta}{2} G^2 \qquad (\|g_t\| \le G)$$

$$\Rightarrow \quad \frac{1}{n}\sum_{t=1}^{n} g_t \cdot (w_t - w^*) \le \frac{1}{2\eta n}\|w_1 - w^*\|^2 + \frac{\eta}{2} G^2$$

(online gradient descent regret bound)

Proof from the book



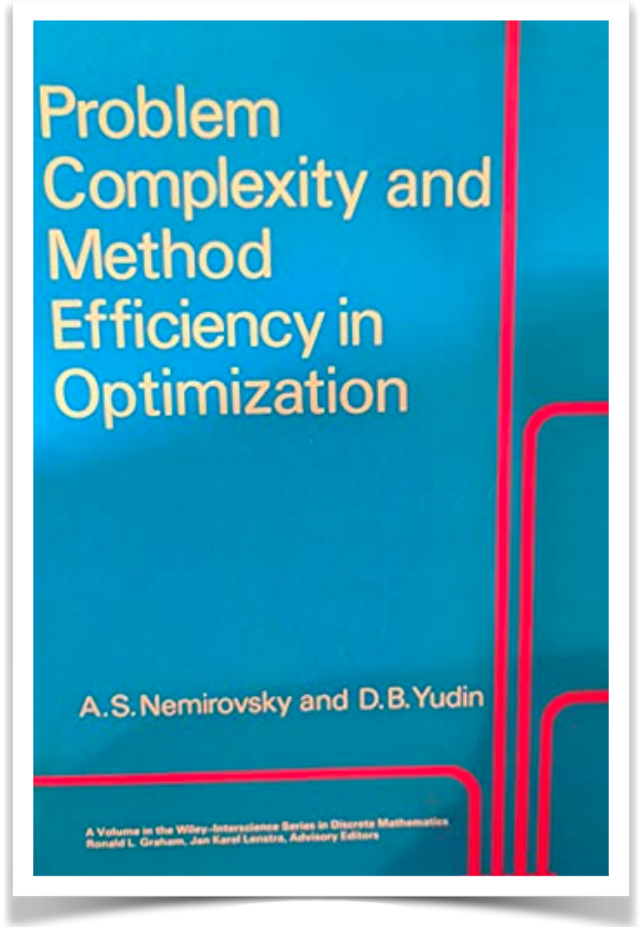
SGD

$$\boxed{w_{t+1} = w_t - \eta g_t \; ; \qquad \mathbb{E}[g_t \mid w_t] = \nabla F(w_t)}$$

$$\frac{1}{n}\sum_{t=1}^{n} g_t \cdot (w_t - w^*) \leq \frac{1}{2\eta n}\|w_1 - w^*\|^2 + \frac{\eta}{2}G^2 \cong \frac{\boldsymbol{O(1)}}{\sqrt{\boldsymbol{n}}}$$

for $\eta \cong \frac{1}{\sqrt{n}}$

Proof from the book



SGD $\boxed{w_{t+1} = w_t - \eta g_t \; ; \qquad \mathbb{E}[g_t \mid w_t] = \nabla F(w_t)}$

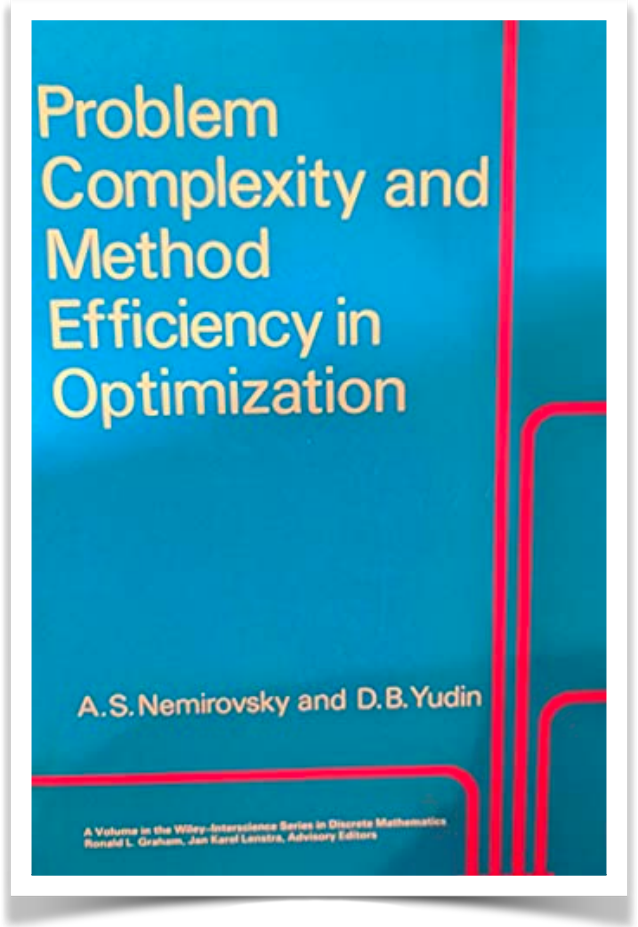
$$\frac{1}{n}\sum_{t=1}^{n} g_t \cdot (w_t - w^*) \le \frac{1}{2\eta n}\|w_1 - w^*\|^2 + \frac{\eta}{2}G^2 \cong \frac{O(1)}{\sqrt{n}} \qquad \text{for } \eta \cong \frac{1}{\sqrt{n}}$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} g_t \cdot (w_t - w^*)\right] = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[\nabla F(w_t)\cdot(w_t - w^*)\right] \ge \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[F(w_t) - F(w^*)\right]$$

$$\mathbb{E}\left[F(\overline{w})\right] - F(w^*) \le \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[F(w_t) - F(w^*)\right] \lesssim \frac{O(1)}{\sqrt{n}}$$

$$\left(\overline{w} = \tfrac{1}{n}\Sigma_{t=1}^{n} w_t\right)$$

Proof from the book



SGD $\boxed{w_{t+1} = w_t - \eta g_t \; ; \qquad \mathbb{E}[g_t \mid w_t] = \nabla F(w_t)}$

$$\frac{1}{n}\sum_{t=1}^{n} g_t \cdot (w_t - w^*) \le \frac{1}{2\eta n}\|w_1 - w^*\|^2 + \frac{\eta}{2}G^2 \cong \frac{O(1)}{\sqrt{n}} \qquad \text{for } \eta \cong \frac{1}{\sqrt{n}}$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} g_t \cdot (w_t - w^*)\right] = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[\nabla F(w_t)\cdot(w_t - w^*)\right] \ge \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[F(w_t) - F(w^*)\right]$$

$$\mathbb{E}\left[F(\overline{w})\right] - F(w^*) \le \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[F(w_t) - F(w^*)\right] \lesssim \frac{O(1)}{\sqrt{n}}$$

$$\left(\overline{w} = \tfrac{1}{n}\Sigma_{t=1}^{n} w_t\right)$$

* Empirical risk?

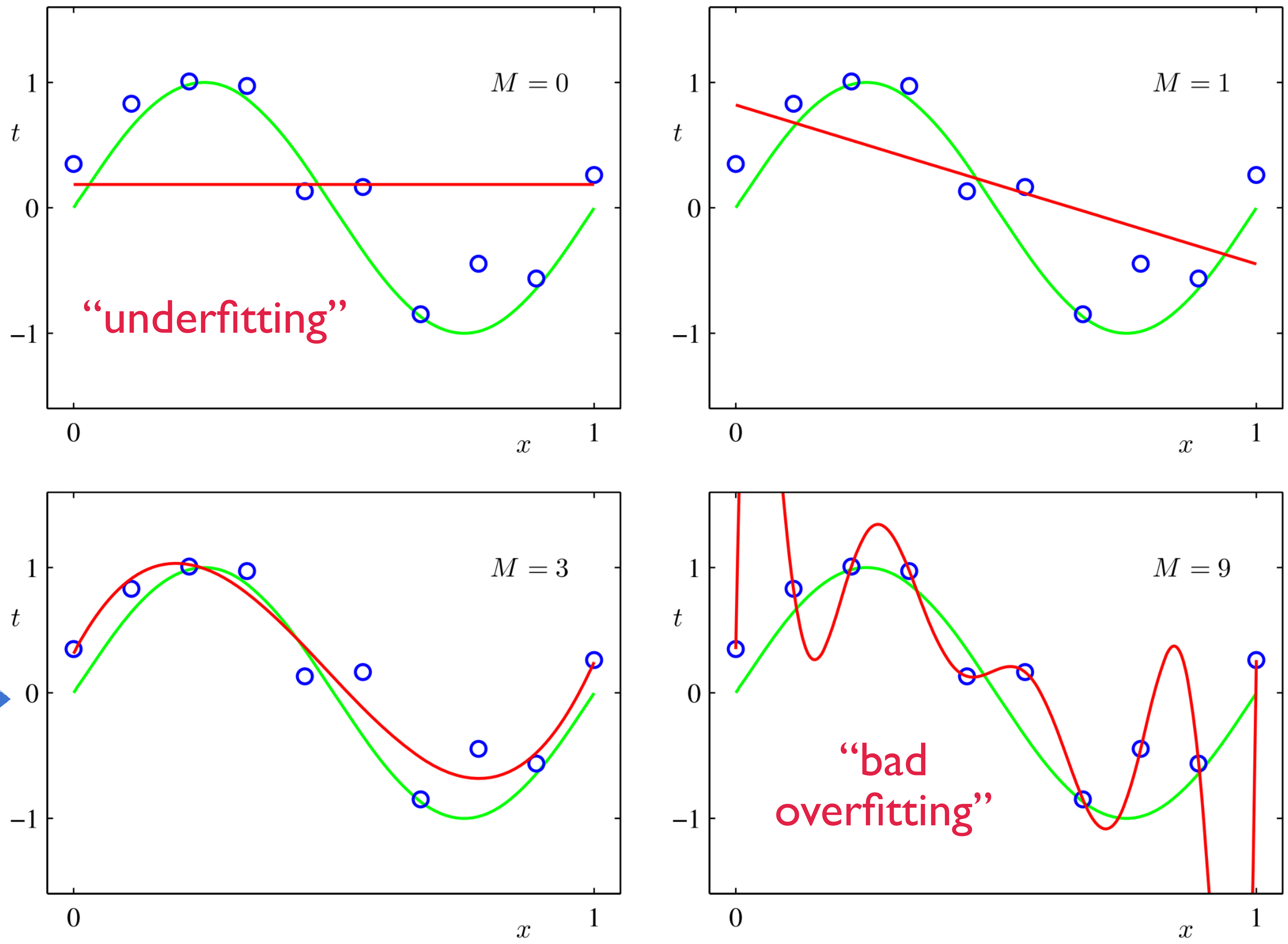
What form of capacity control is at play?

Capacity control

"with too much fitting, the model adapts itself too closely to the training data, and will not generalize well"

The Elements of Statistical Learning [Hastie-Tibshirani-Friedman '09]

controlled model capacity



uncontrolled model capacity

Capacity control $\iff$ Generalization Bounds

test-error $\leq$ train-error + |generalization-gap|

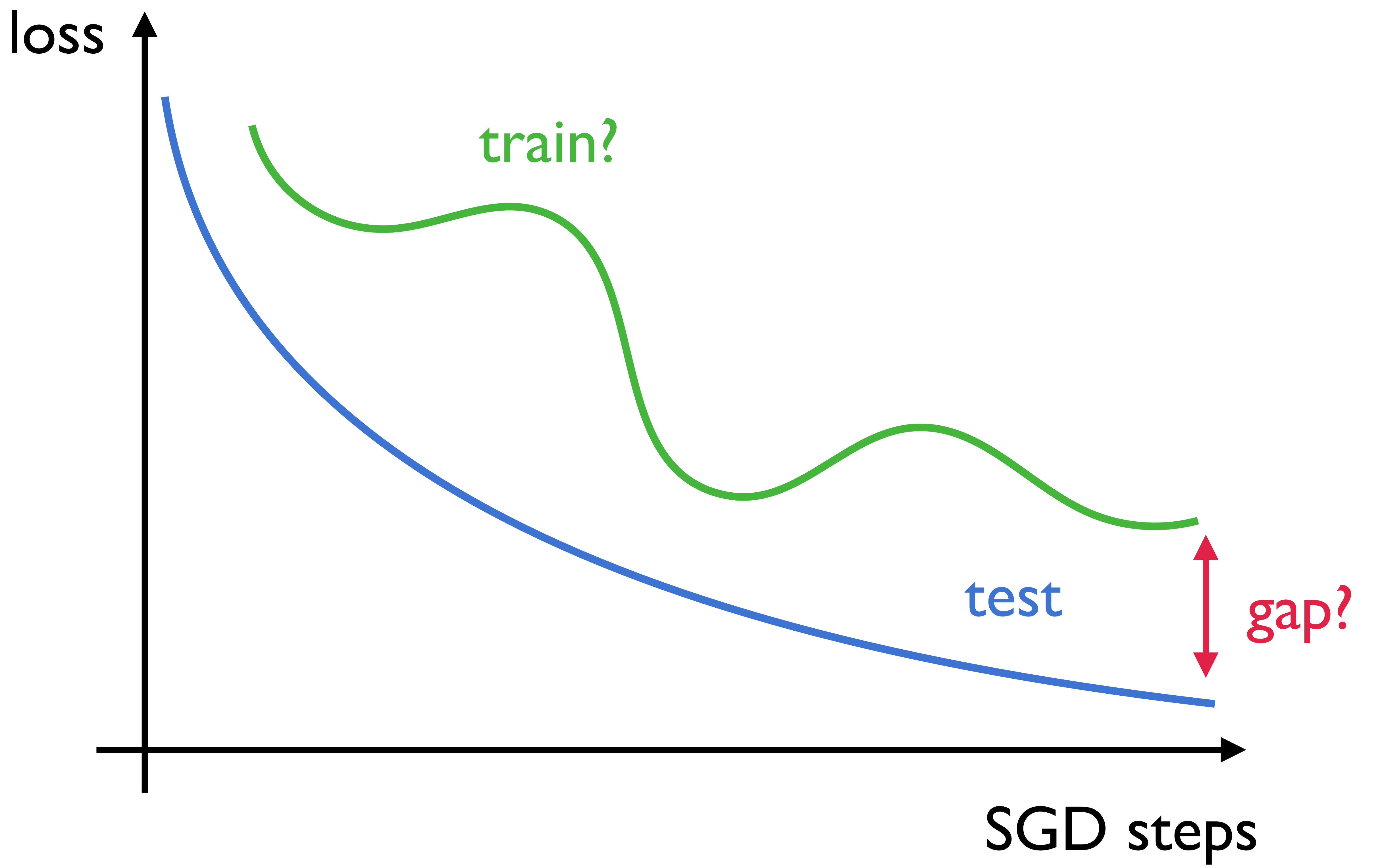
$$\mathbb{E}[F(\hat{w}) - F^\star] \leq \mathbb{E}[\hat{F}(\hat{w}) - \hat{F}^\star] + \boxed{|\mathbb{E}F(\hat{w}) - \mathbb{E}\hat{F}(\hat{w})|}$$

capacity control

- VC / Rademacher bounds [Vapnik '71, Valiant '84, Bartlett et al. '02, …]
- Algorithmic stability [Bousquet & Elisseeff '02, Hardt et al. '16, …]
- PAC-Bayes [McAllester '99; Dziugaite and Roy '18, …]
- Sample compression [Littlestone & Warmuth '86; Arora et al. '18, …]
- Information-theoretic bounds [Xu & Raginsky '17, Neu '21, …]
- …
- More recently: Implicit bias, interpolating algorithms, benign overfitting, …

"Laws of Large Numbers approach"

SGD's generalization gap?



$$\text{test-error} \leq |\text{train-error}| + |\text{gen-gap}|$$

$O(1/\sqrt{n})$? ?

SGD doesn't generalize

(one pass) SGD

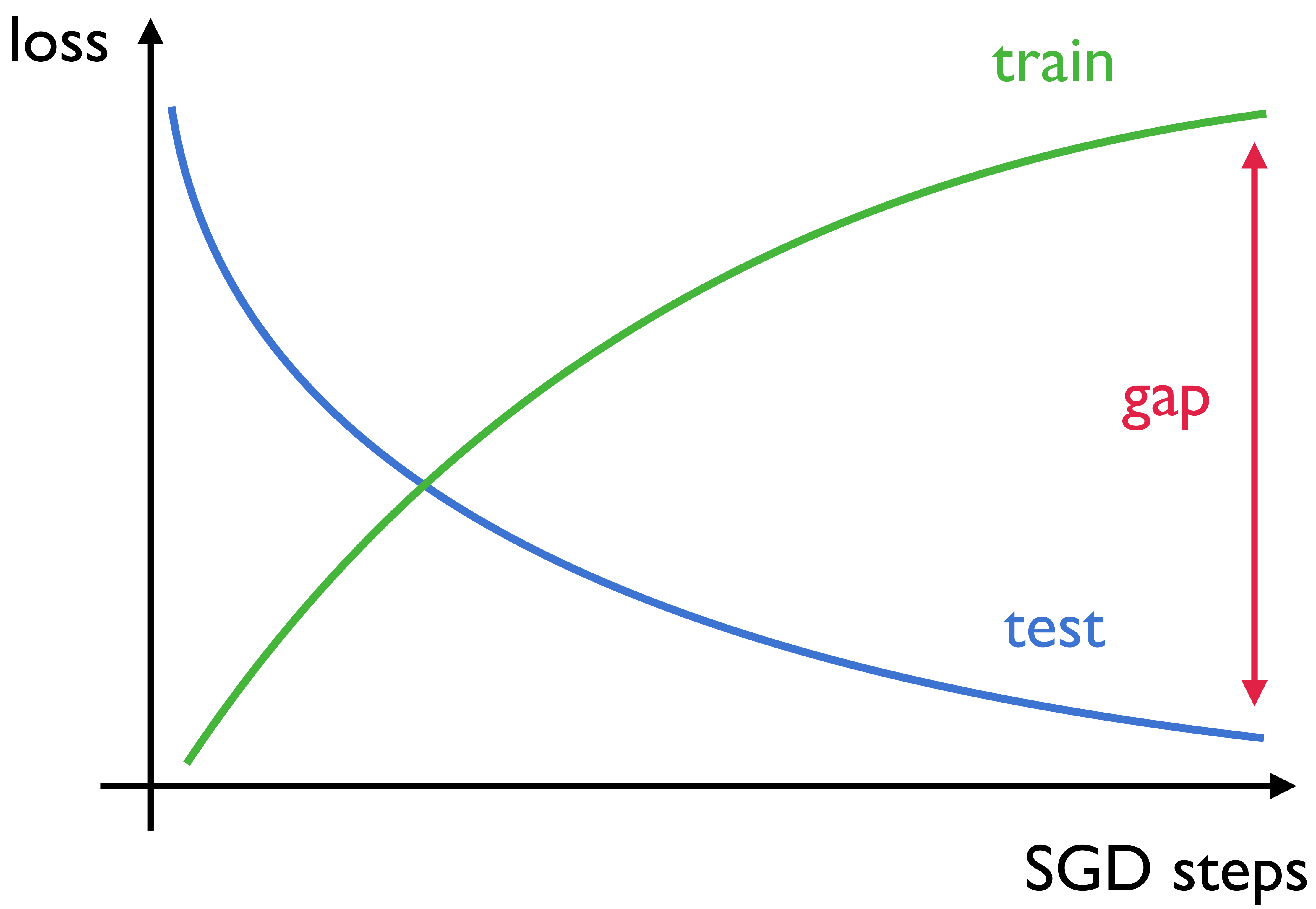
$$w_{t+1} = \Pi_W[w_t - \eta \nabla f(w_t, z_t)]$$
$$\hat{w} \quad = \frac{1}{n}\Sigma_{t=1}^{n} w_t$$

Theorem: $\exists$ dist. D and convex, 1-Lipschitz loss $f(w,z)$ over $W = \{\text{unit ball in } \mathbb{R}^d\}$ in dim $d = \tilde{\Theta}(n)$ s.t. for SGD "from the book" (one pass, $\eta \cong 1/\sqrt{n}$):

1. **True risk is (optimally) small:** $\mathbb{E}\left[F(\hat{w}) - F^{\star}\right] \lesssim \frac{1}{\sqrt{n}}$ [N&Y '83]
2. **Empirical risk is (trivially) large:** $\mathbb{E}\left[\hat{F}(\hat{w}) - \hat{F}^{\star}\right] \gtrsim 1$
3. **Gen-gap is (trivially) large:** $\mathbb{E}\left[\hat{F}(\hat{w}) - F(\hat{w})\right] \gtrsim 1$

- K, Livni, Mansour, Sherman '22; Schliserman, Sherman, K '25; Vansover-Hager, K, Livni '25

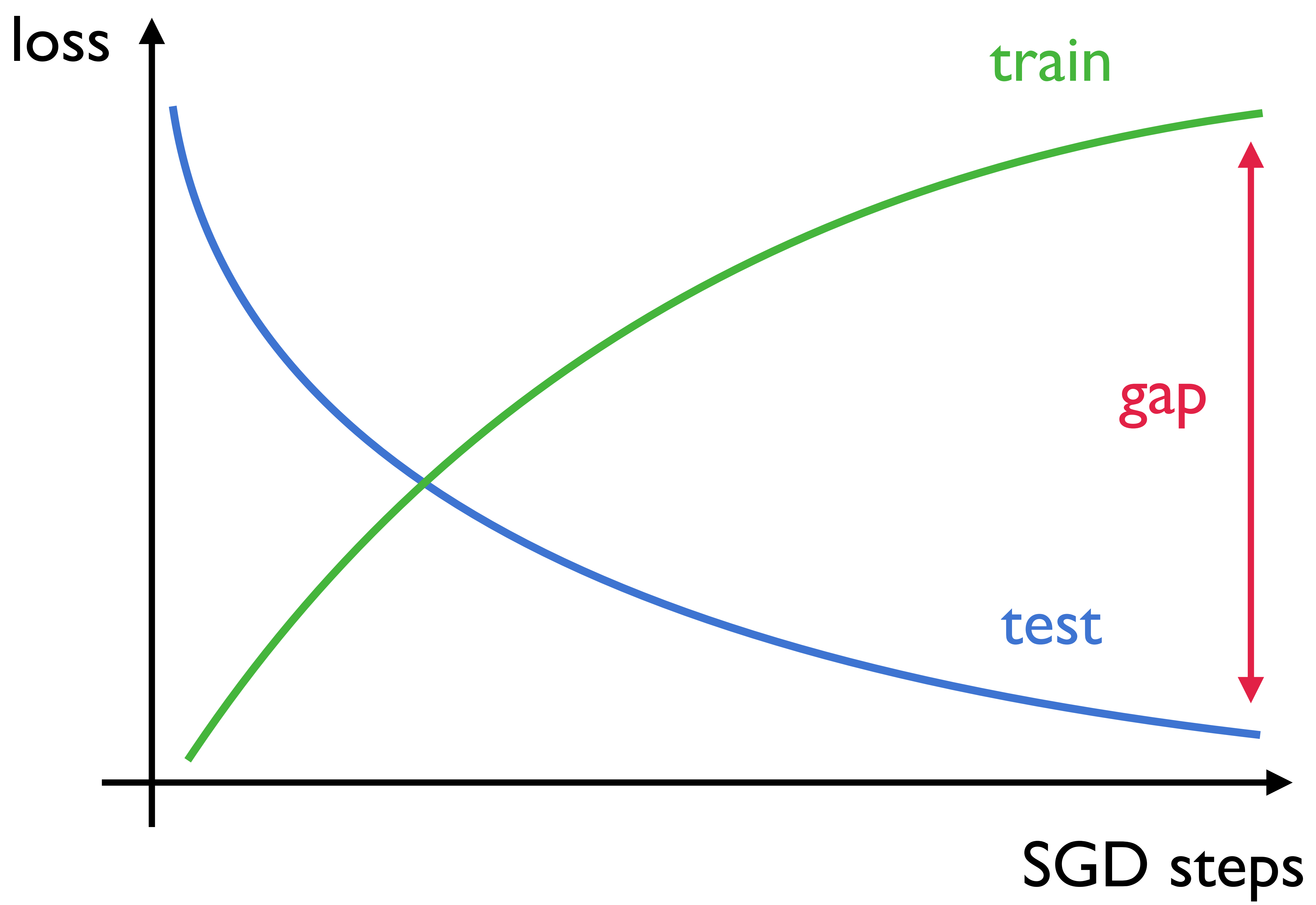
"Benign Underfitting"



$$\text{test-error} \leq |\text{train-error}| + |\text{gen-gap}|$$

$$O(1/\sqrt{n}) \qquad \Omega(1) \qquad \Omega(1)$$

"Benign Underfitting"

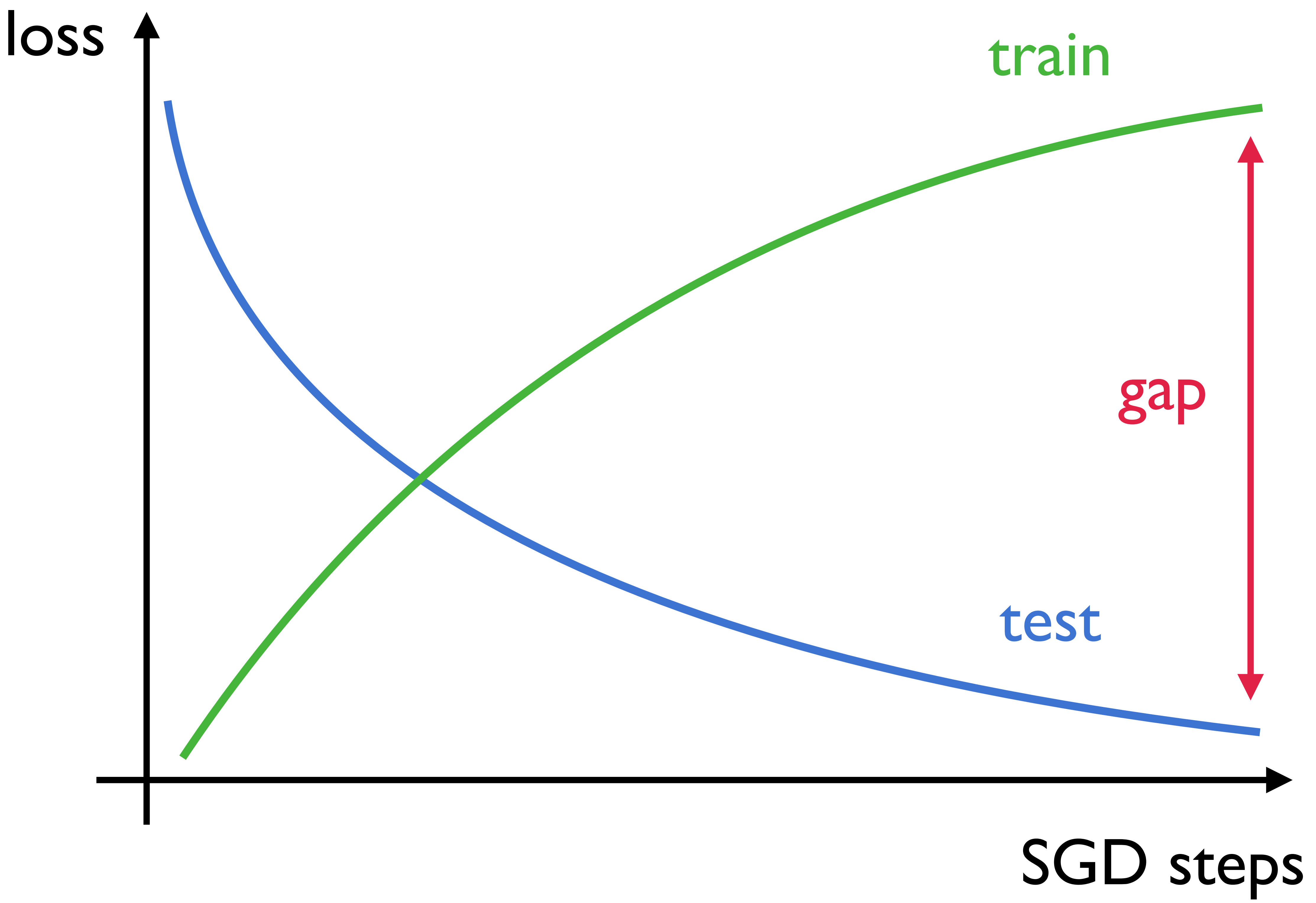


- "Laws of Large Numbers approach" fails for SGD
- Even in most basic, fundamental setup (convex optimization)

test-error $\leq$ **|train-error|** + **|gen-gap|**

$O(1/\sqrt{n})$ $\Omega(1)$ $\Omega(1)$

“Benign Underfitting”



test-error $\leq$ |train-error| + |gen-gap|

$O(1/\sqrt{n})$ $\Omega(1)$ $\Omega(1)$

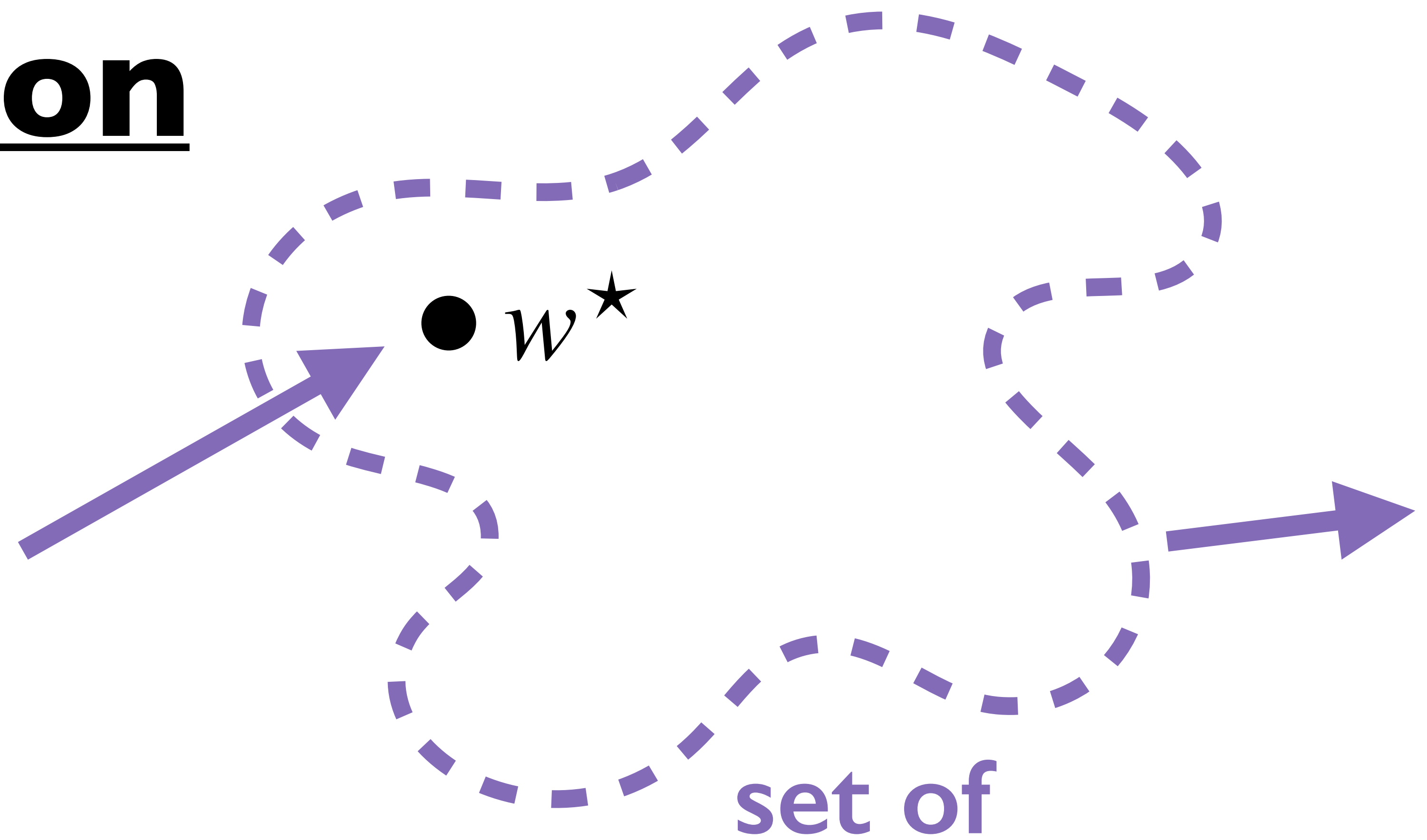
- “Laws of Large Numbers approach” fails for SGD
- Even in most basic, fundamental setup (convex optimization)
- Out-of-sample performance (only?) explained by stochastic approx. / regret analysis
- No (effective) “implicit bias”

Contemporary wisdom: "Implicit bias"

Optimization

$$\min_{w} \hat{F}(w)$$

train-loss



H_{alg}

set of "simple" models

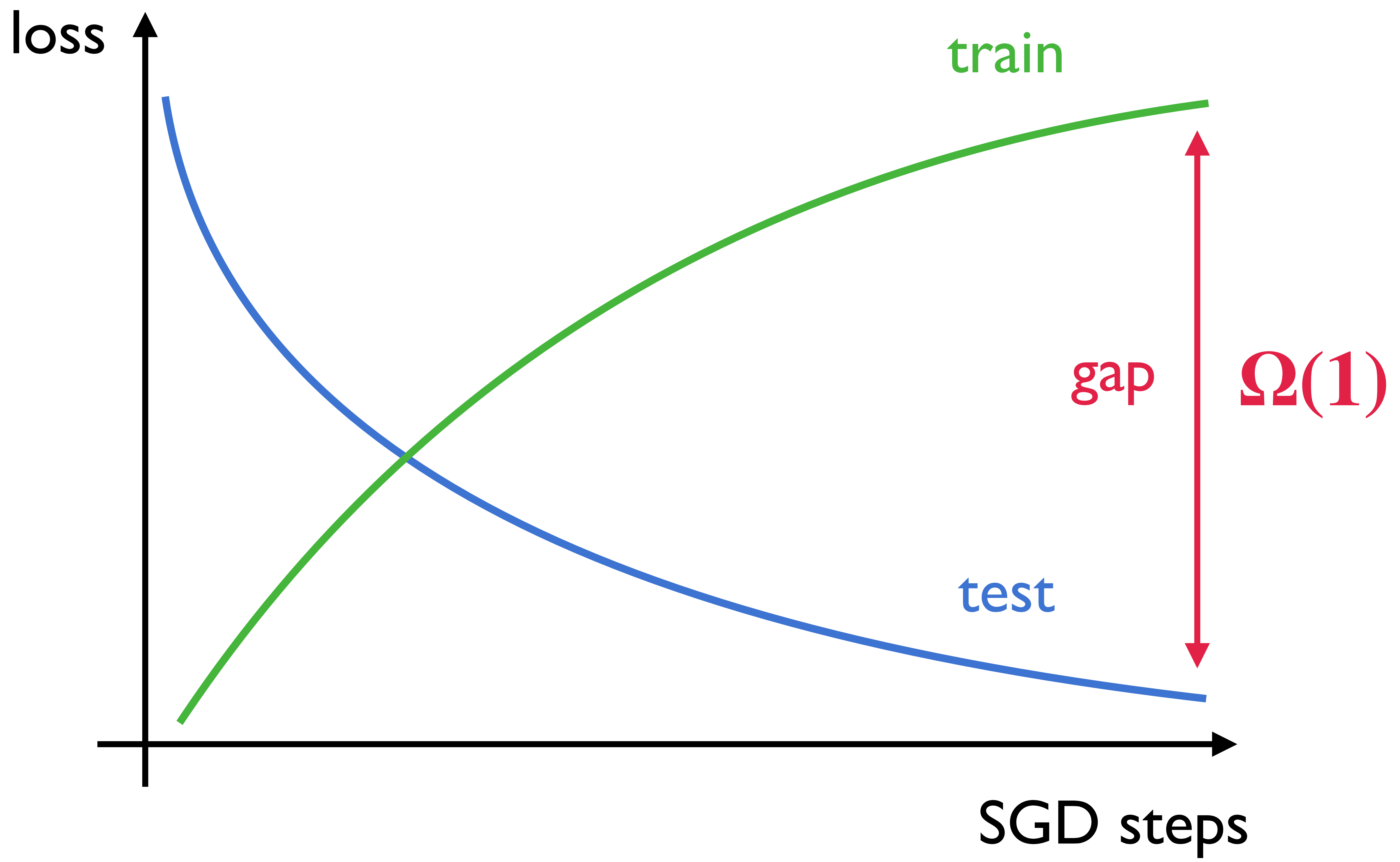
Generalization

$$\sup_{w^\star \in H_{\text{alg}}} \left| F(w^\star) - \hat{F}(w^\star) \right|$$

|gen-gap|

- **Modern belief:** common optimization algorithms are <u>implicitly biased</u> towards "simple" models, thus generalize
- "Simple" is e.g.: low norm, sparse, low-rank, short MDL, …

"Benign Underfitting"



No implicit bias
(for SGD in SCO)

$$\left| F(w) - \hat{F}(w) \right|$$

is large at
SGD solution

More generally

(one pass) SGD

$$w_{t+1} = \Pi_W[w_t - \eta \nabla f(w_t, z_t)]$$
$$\hat{w} \quad = \frac{1}{n}\Sigma_{t=1}^n w_t$$

Theorem: $\forall\ n, \eta > 0$, $\exists$ dist. D and convex, 1-Lipschitz $f(w, z)$
over $W = \{$unit ball in $\mathbb{R}^d\}$ in dim $d = \tilde{\Theta}(n)$
s.t. for SGD with any $\eta > 0$:

1. **Empirical risk is large:** $\mathbb{E}\left[\hat{F}(\hat{w}) - \hat{F}^\star\right] \gtrsim \min\left\{\eta\sqrt{n} + \frac{1}{\eta\sqrt{n}}, 1\right\}$

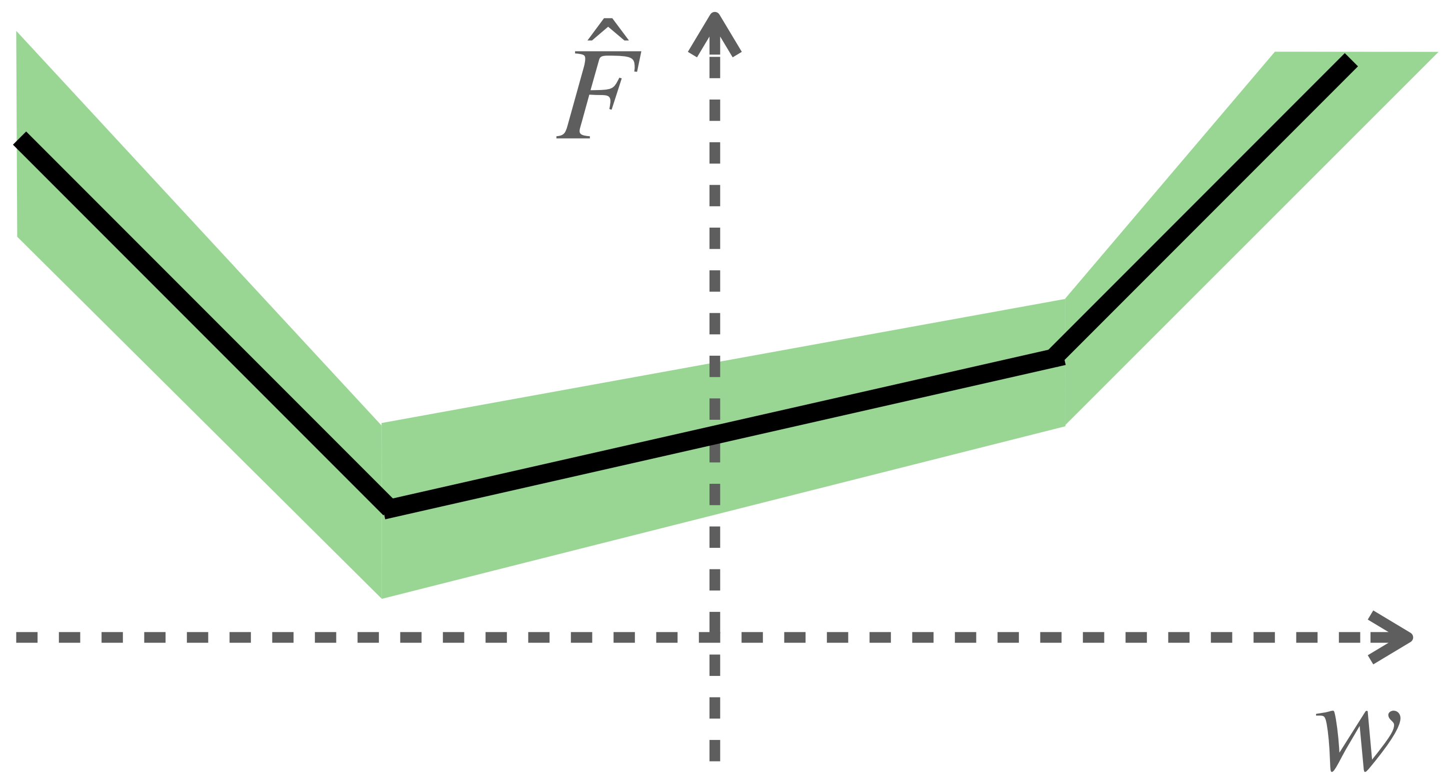
2. **Gen-gap is large:** $\mathbb{E}\left[\hat{F}(\hat{w}) - F(\hat{w})\right] \gtrsim \min\left\{\eta\sqrt{n} + \frac{1}{\eta\sqrt{n}}, 1\right\}$

☞ E.g. implies that gen-gap is trivially large unless $n = \Omega(d)$

☞ Dim dependence is ~optimal: when $d = o(n)$ uniform convergence kicks in

Proof ideas

- Step #1: "turn off" **uniform convergence**

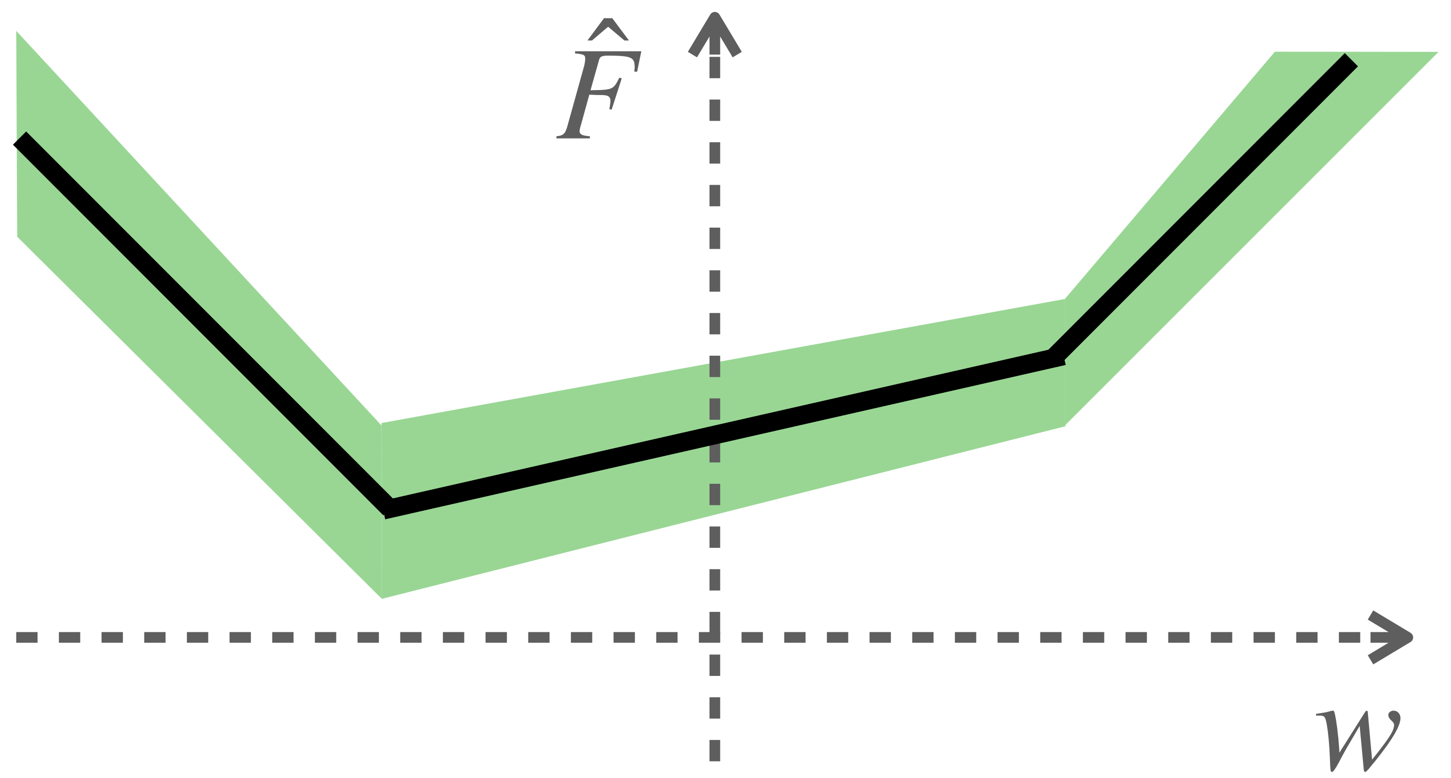


$$\sup_{w \in W} |\hat{F}(w) - F(w)| \lesssim \sqrt{\frac{d}{n}}$$

(laws of large numbers)

Proof ideas

- Step #1: "turn off" uniform convergence



$$\sup_{w \in W} |\hat{F}(w) - F(w)| \lesssim \sqrt{\frac{d}{n}}$$

(laws of large numbers)

- UC rate is $O\left(\sqrt{d/n}\right)$
  $\implies$ work in dimension $d = \Omega(n)$
- Show that $\exists\, w^{\text{bad}} \in W$
  with large generalization gap

Turn off uniform convergence

	$i = 1$								$i = d$
z_1	0	0	1	0	…	…	1	0	1
z_2	1	1	1	0	…	…	1	1	0
	⋮	⋮	⋮	⋮			⋮	⋮	⋮
z_n	1	0	1	0	…	…	1	1	1

0/1
w.p. $\frac{1}{2}$

[Shalev-Shwartz, Shamir, Srebro, Sridharan '09]

Turn off uniform convergence

w.p. $\geq \frac{1}{2}$

if $d \geq 2^n$

	$i = 1$								$i = d$
z_1	0	0	1	0	...	...	1	0	1
z_2	1	1	1	0	...	...	1	1	0
	:	:	:	:			1	:	:
							1		
							1		
							1		
z_n	1	0	1	0	...	...	1	1	1
							k		

0/1

w.p. $\frac{1}{2}$

[Shalev-Shwartz, Shamir, Srebro, Sridharan '09]

Turn off uniform convergence

z_1	0	0	1	0	…	…	1	0	1
z_2	1	1	1	0	…	…	1	1	0
	:	:	:	:			1	:	:
							1		
							1		
							1		
z_n	1	0	1	0	…	…	1	1	1
							k		

$$f(w,z) = \sum_{i=1}^{d} z(i) w^2(i)$$

convex, Lipschitz
(over unit ball)

[Shalev-Shwartz, Shamir, Srebro, Sridharan ’09]

Turn off uniform convergence

z_1	0	0	1	0	…	…	1	0	1
z_2	1	1	1	0	…	…	1	1	0
	:	:	:	:			1	:	:
							1		
							1		
							1		
z_n	1	0	1	0	…	…	1	1	1
							k		

$$f(w,z) = \sum_{i=1}^{d} z(i)w^2(i)$$

convex, Lipschitz
(over unit ball)

$$\hat{F}(e_k) = 1$$

[Shalev-Shwartz, Shamir, Srebro, Sridharan ’09]

Turn off uniform convergence

z_1	0	0	1	0	…	…	1	0	1
z_2	1	1	1	0	…	…	1	1	0
	⋮	⋮	⋮	⋮			1	⋮	⋮
							1		
							1		
							1		
z_n	1	0	1	0	…	…	1	1	1
							k		

$$f(w, z) = \sum_{i=1}^{d} z(i) w^2(i)$$

convex, Lipschitz
(over unit ball)

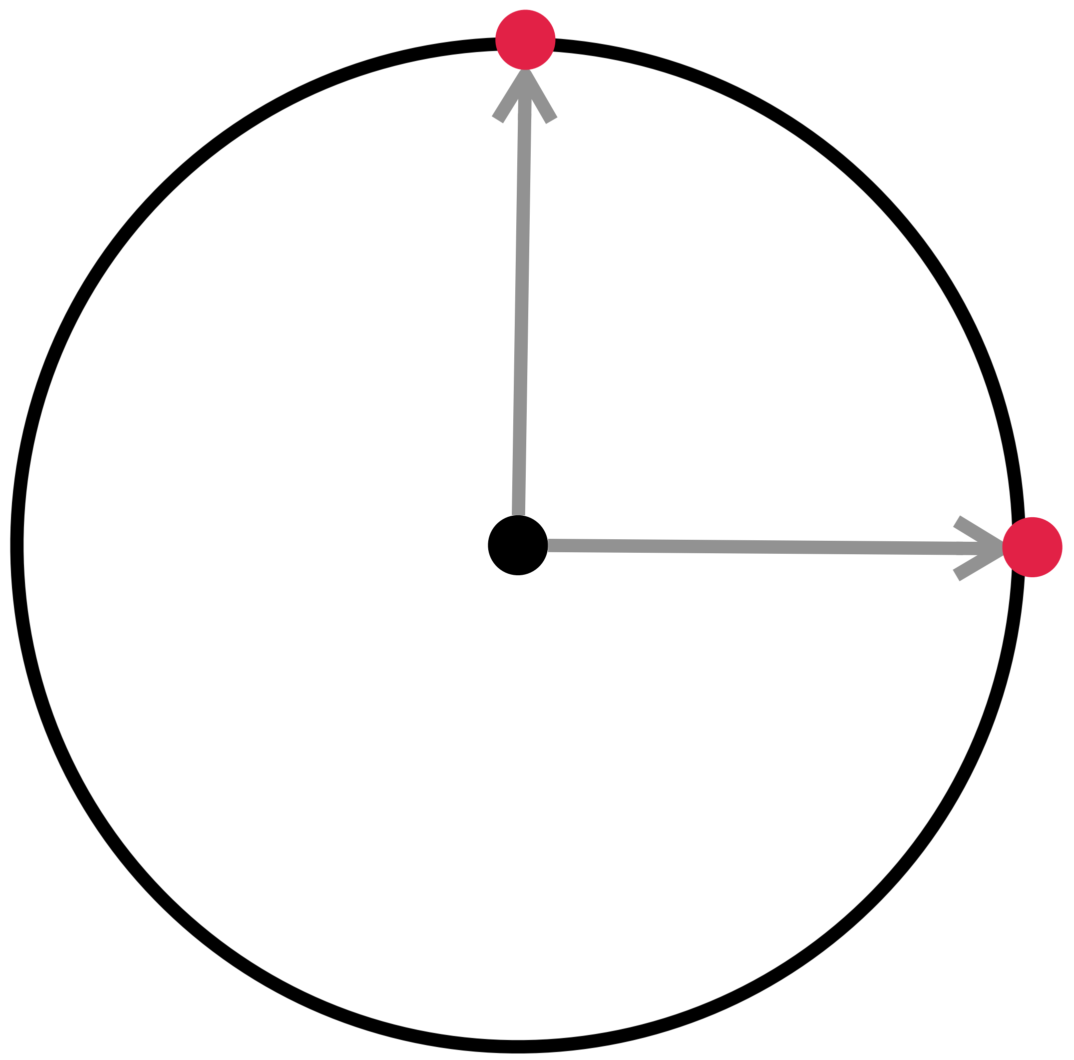
$\hat{F}(e_k) = 1$

$F(e_k) = \frac{1}{2}$

[Shalev-Shwartz, Shamir, Srebro, Sridharan ’09]

Turn off uniform convergence

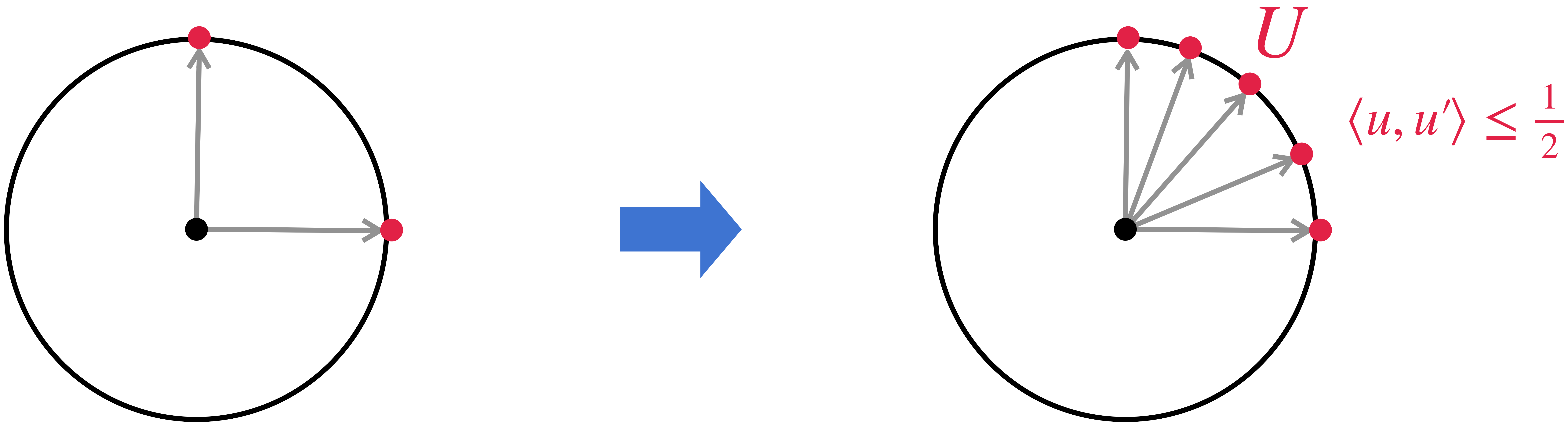
- Construction of f, D requires dimension $d \geq 2^{\Theta(n)}$...



2^n orthogonal directions in $d = 2^n$

Turn off uniform convergence

- Construction of f, D requires dimension $d \geq 2^{\Theta(n)}$…



2^n orthogonal
directions in $d = 2^n$

$\exp(n)$ **nearly** orthogonal
directions in $d = O(n)$

[Feldman '16]

Proof ideas

- Step #2: "turn off" algorithmic stability

Algorithmic stability
[Bousquet & Elisseeff '02]

learning algorithm is δ-stable

if replacing one sample in S

➜ δ change in output ($\hat{w}$)

Roughly:

δ-stability ➜ $O(\delta)$ gen-gap

Proof ideas

- Step #2: "turn off" algorithmic stability

If f is sufficiently smooth (with $\beta \leq 1/\eta$)

➜ SGD is η-stable [Hardt, Recht, Singer '16]

Algorithmic stability
[Bousquet & Elisseeff '02]

learning algorithm is δ-stable

if replacing one sample in S

➜ δ change in output ($\hat{w}$)

Roughly:

δ-stability ➜ $O(\delta)$ gen-gap

Proof ideas

- Step #2: "turn off" algorithmic stability

If f is sufficiently smooth (with $\beta \leq 1/\eta$)

➜ SGD is η-**stable** **[Hardt, Recht, Singer '16]**

f should be highly **non-smooth** around initialization

Algorithmic stability
[Bousquet & Elisseeff '02]

learning algorithm is δ-stable if replacing one sample in S ➜ δ change in output ($\hat{w}$)

Roughly:

δ-stability ➜ $O(\delta)$ gen-gap

Proof ideas

- Step #2: "turn off" algorithmic stability

If f is sufficiently smooth (with $\beta \leq 1/\eta$)

➜ SGD is **η-stable** [Hardt, Recht, Singer '16]

Algorithmic stability

[Bousquet & Elisseeff '02]

learning algorithm is δ-stable

if replacing one sample in S

➜ δ change in output ($\hat{w}$)

Roughly:

δ-stability ➜ $O(\delta)$ gen-gap

f should be highly **non-smooth** around initialization

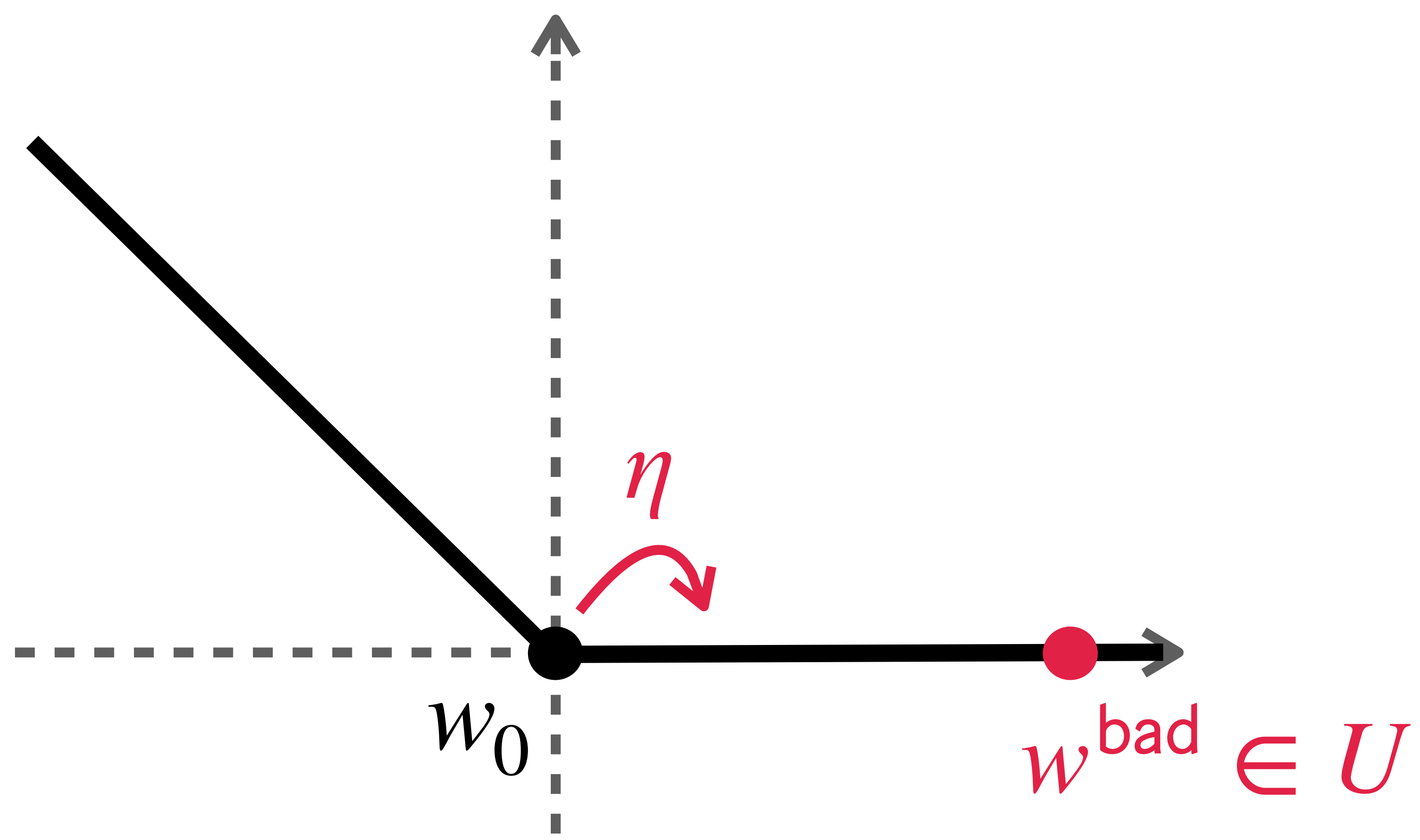
Allows for large SGD steps, potentially towards w^{bad}

Turn off algorithmic stability

$$h(w) = \max_{u \in U} \langle w, u \rangle$$

variant of

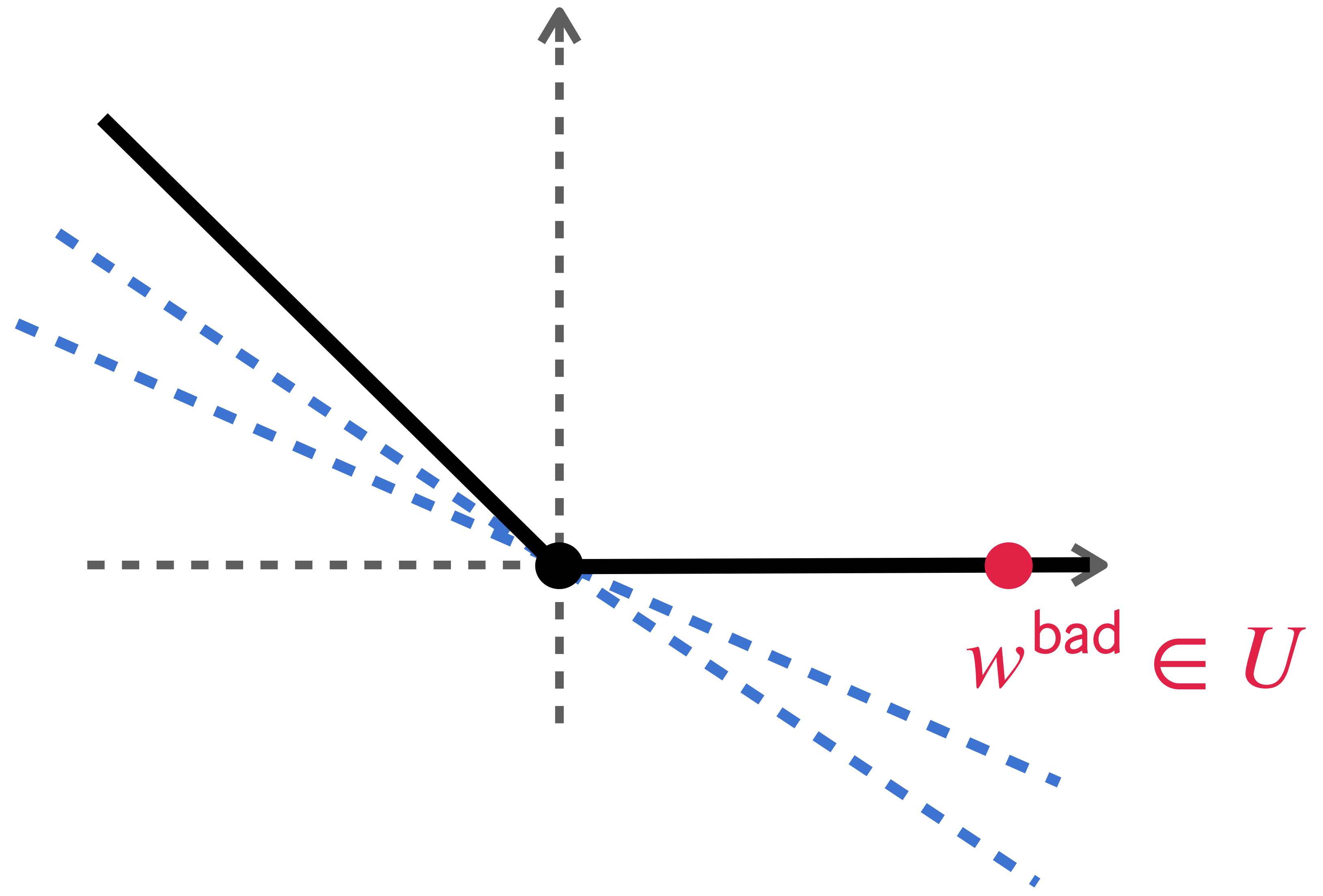
"Nemirovski's function"



- Large subgradients at init due to non-smoothness
- Many subgradients, few of them aligned with w^{bad} …

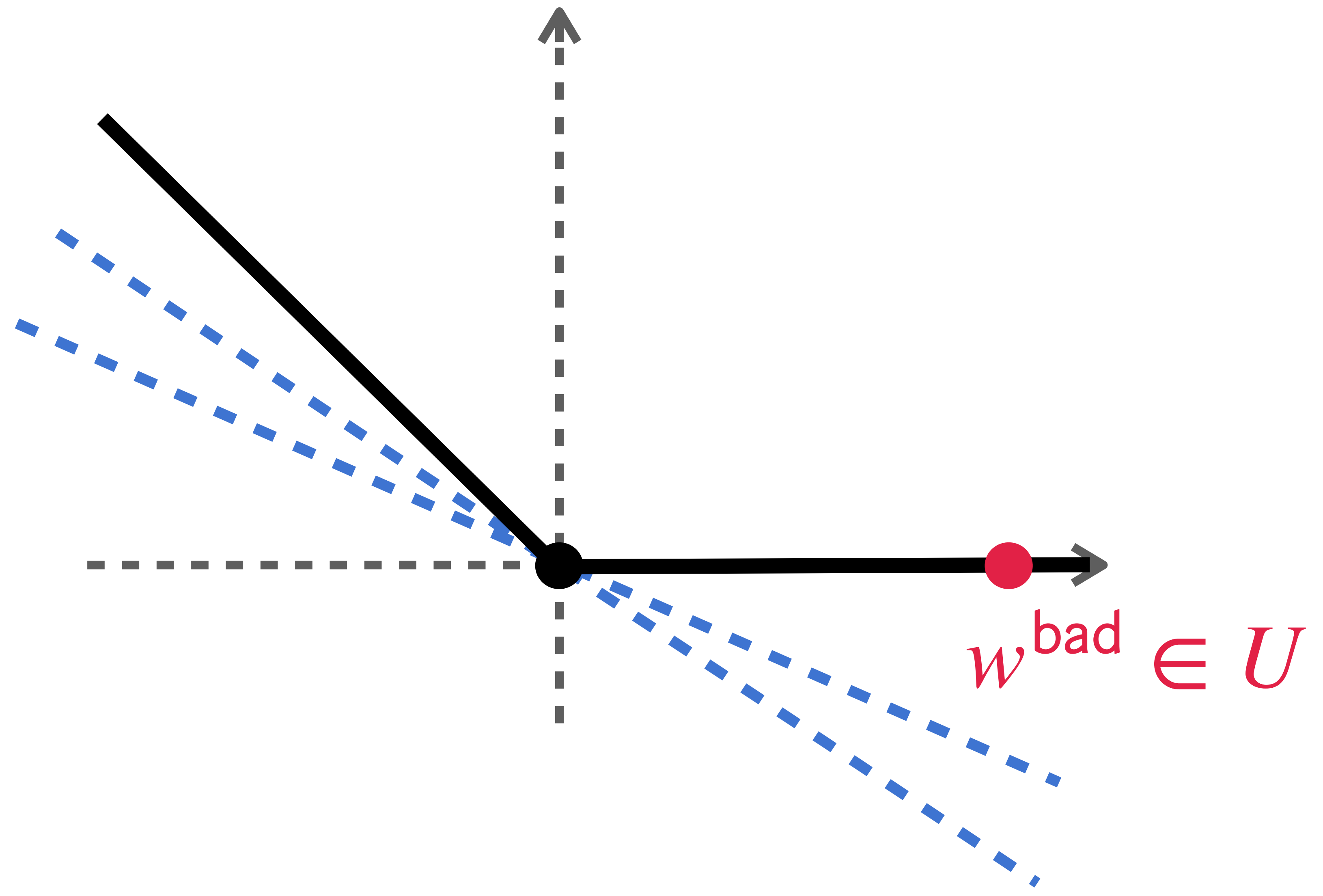
Proof ideas

- Step #3: use instability to steer SGD towards w^{bad}



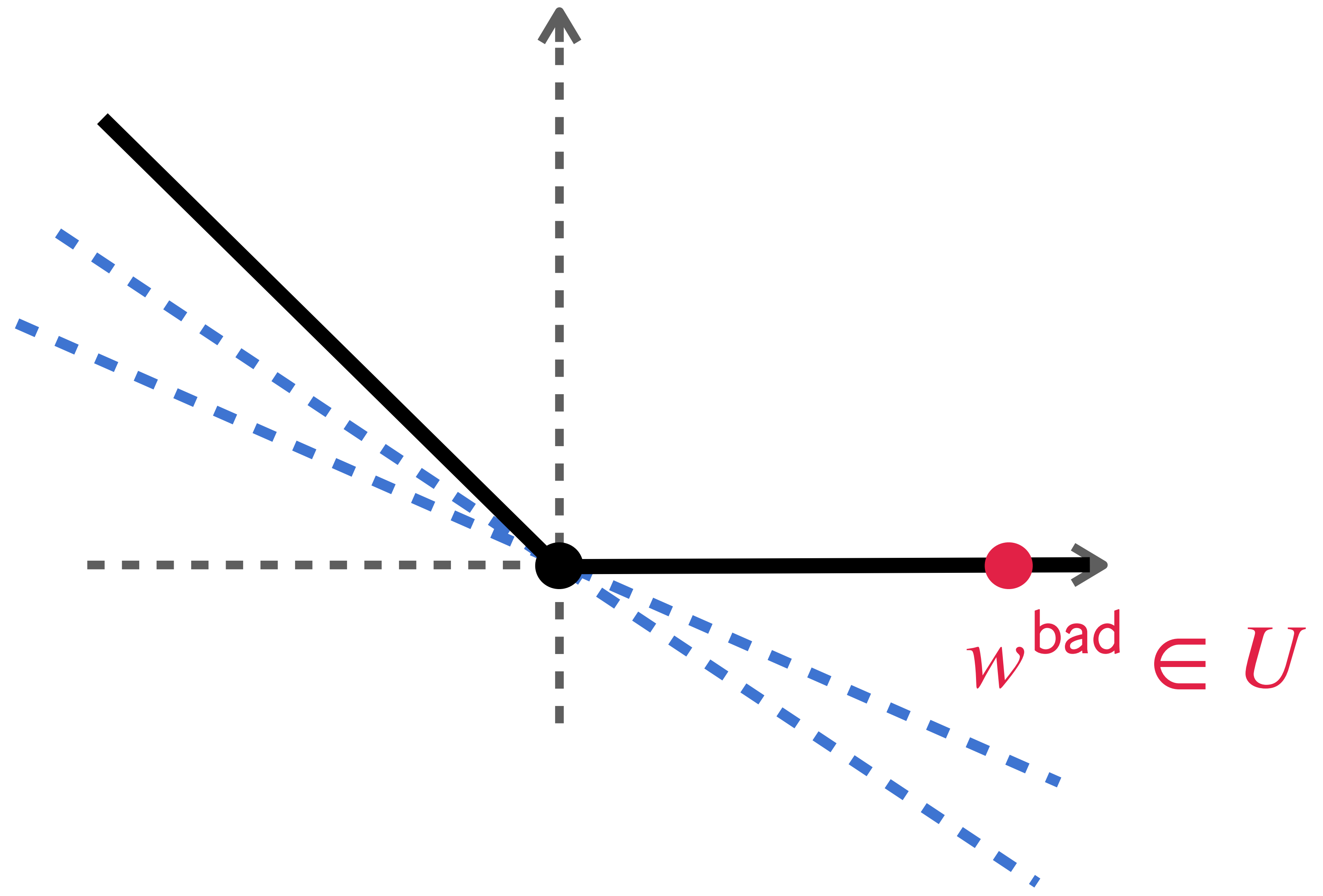
Proof ideas

- Step #3: use instability to steer SGD towards w^{bad}
  - "Cheat" with **sample dependent** subgradient oracle



Proof ideas

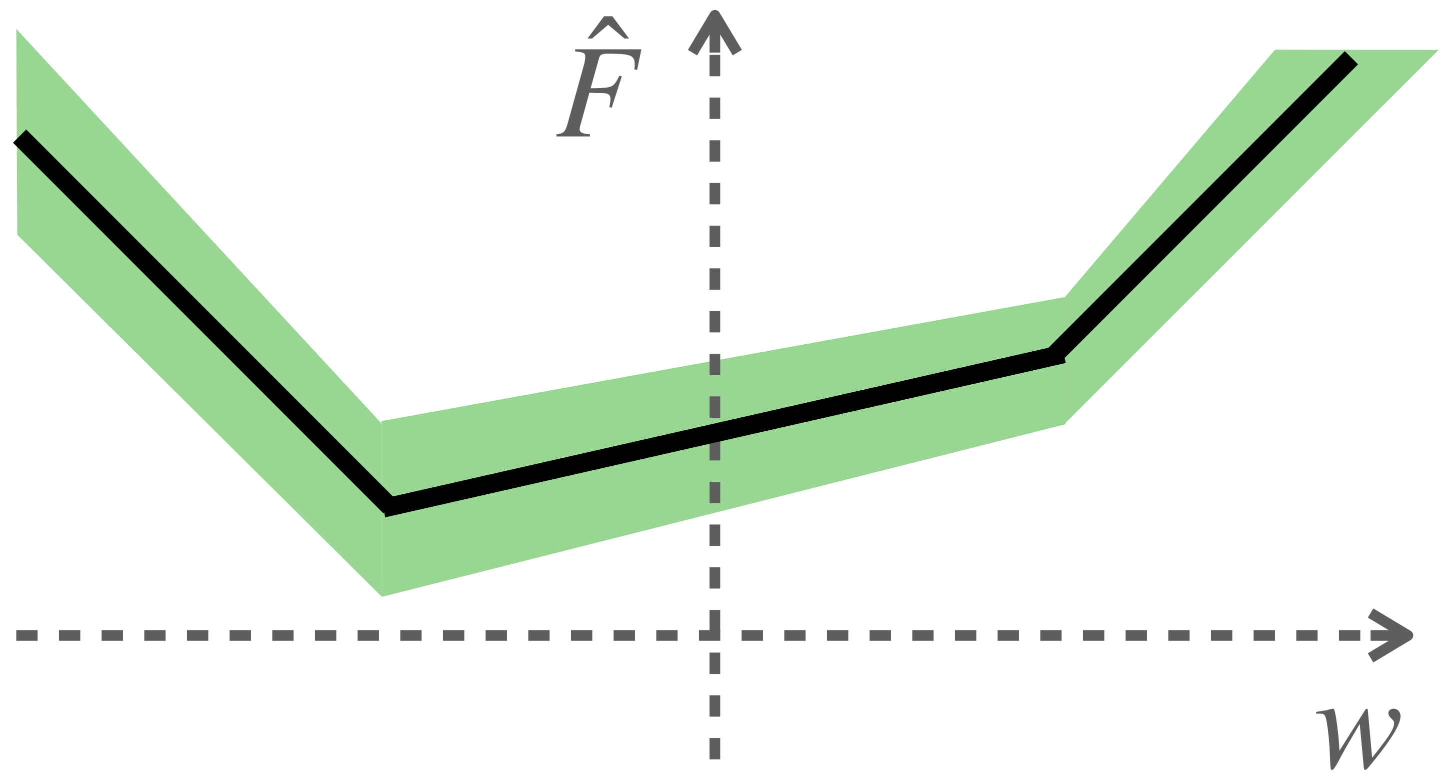
- Step #3: use instability to steer SGD towards w^{bad}
  - "Cheat" with **sample dependent** subgradient oracle
  - Cheat can be removed by **memorizing** samples into SGD iterate
  - Construction can be made differentiable (unique subgradient at every point)



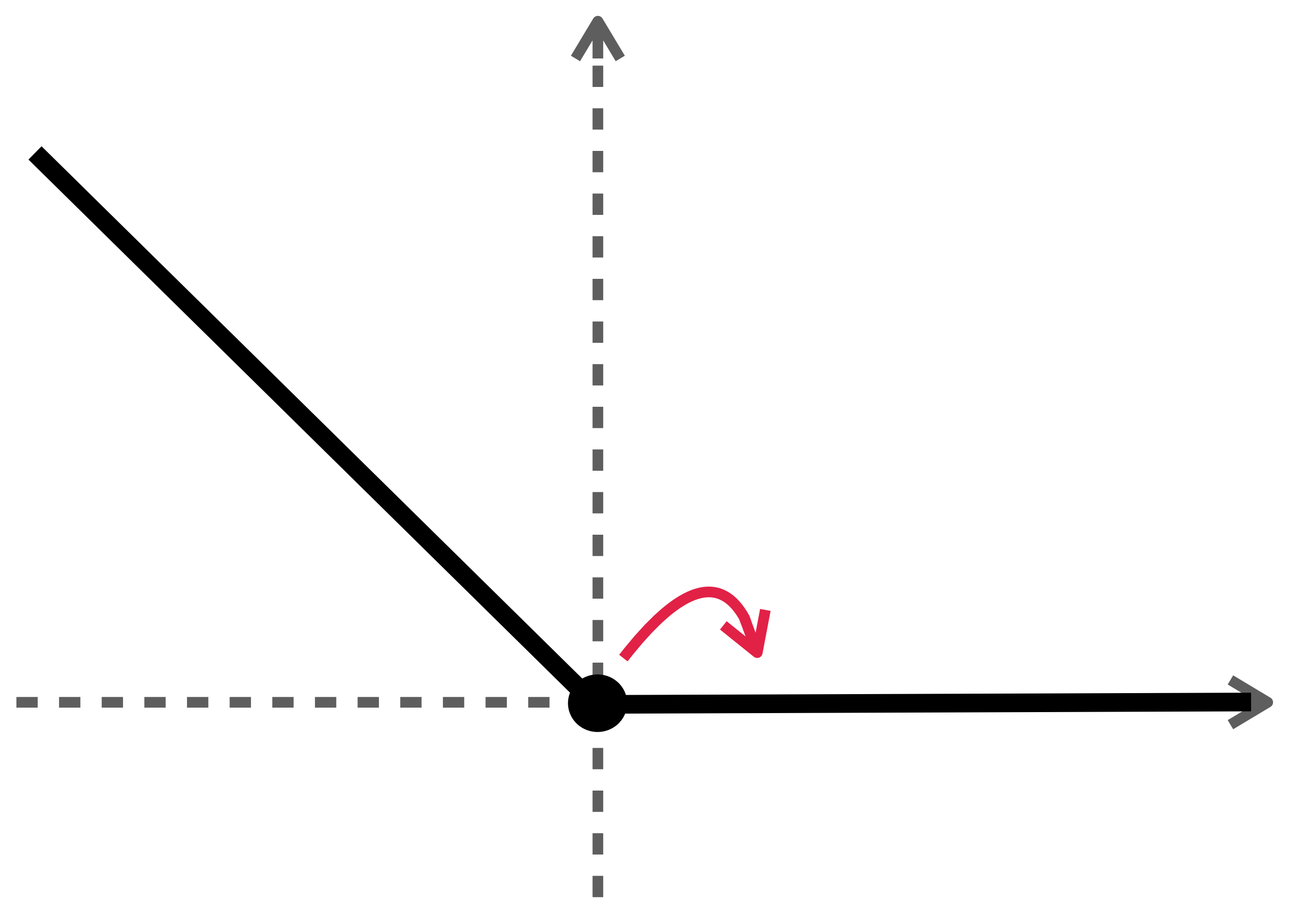
- Similar idea used implicitly in prior work [Amir, K, Livni '21], [K, Livni, Mansour, Sherman '21], [Schliserman, Sherman, K '24], …
- Formalized nicely as a reduction by [Livni '24]

Proof takeaways

- Two generalization mechanisms at play: **uniform convergence** and **algorithmic stability**



- Once "turned off", overfitting/underfitting can occur
- Regret remains controlled, but doesn't control generalization (gap)!



- Construction induces "memorization" of training samples into SGD iterate
- Memorizing (say) half of sample suffices

What happens after the first pass?

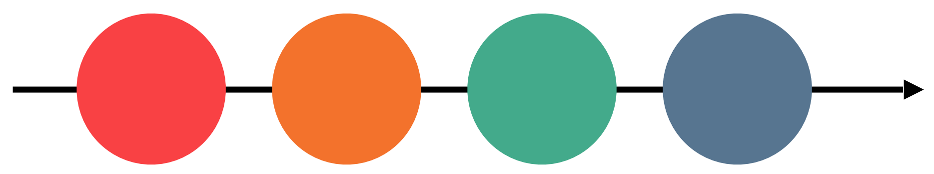
(aka beyond online-to-batch regime)

Multi-pass SGD

One-pass SGD

Optimal rate in SCO

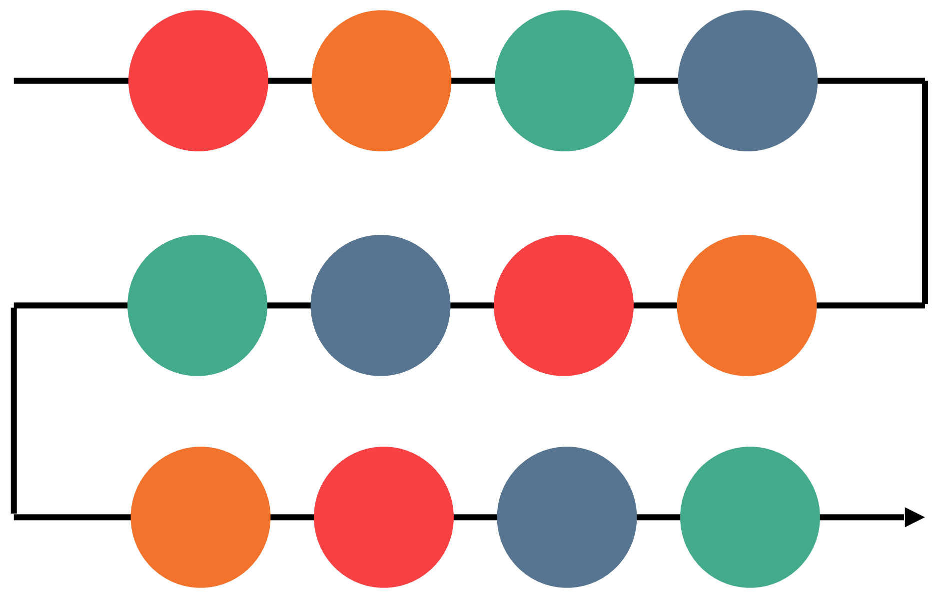
[Nemirovsky & Yudin ’83]



Multi-pass SGD

Common in practice

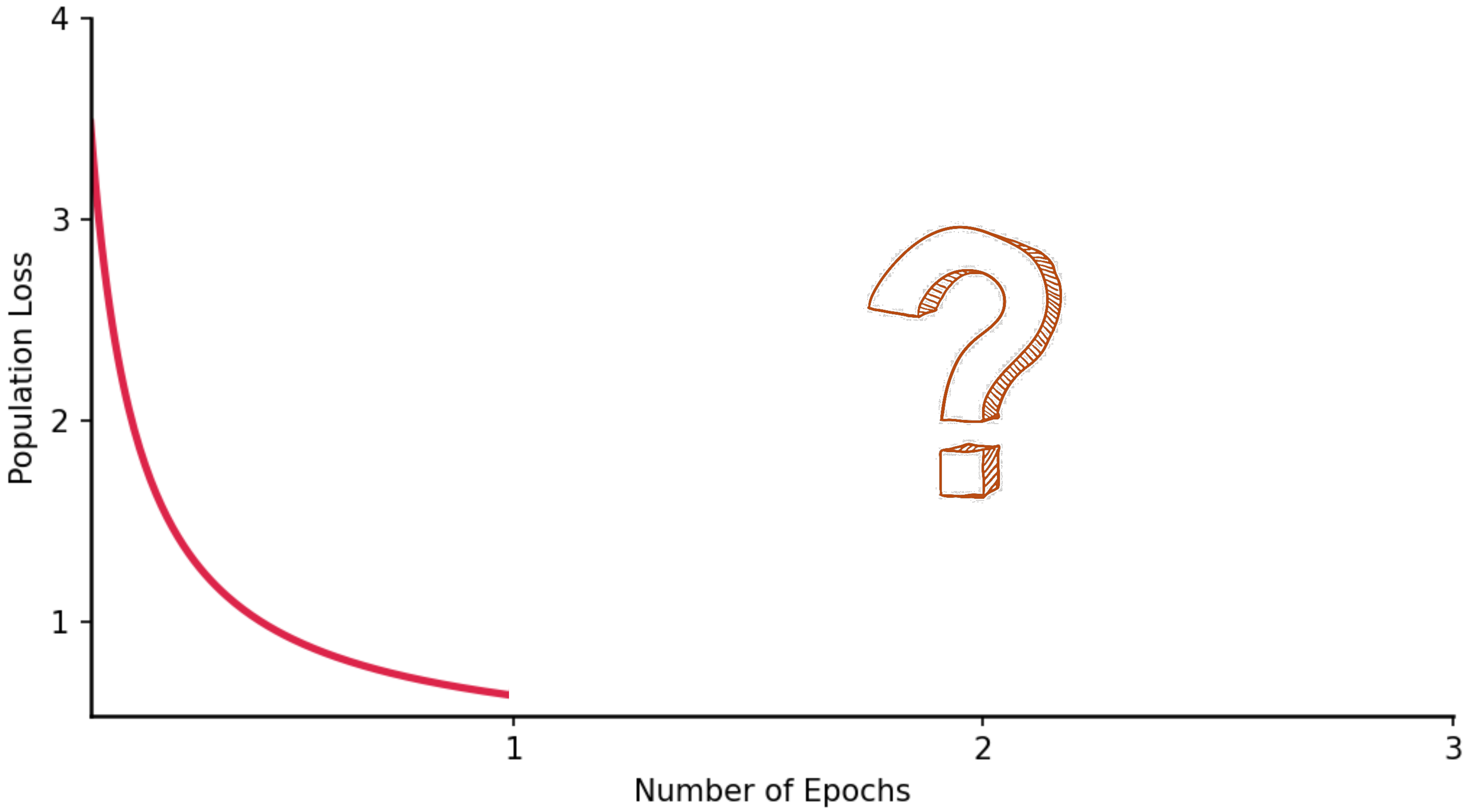
Less understood



Multi-Pass SGD in SCO

with standard step size of $\eta = 1/\sqrt{n}$, optimal after one pass.

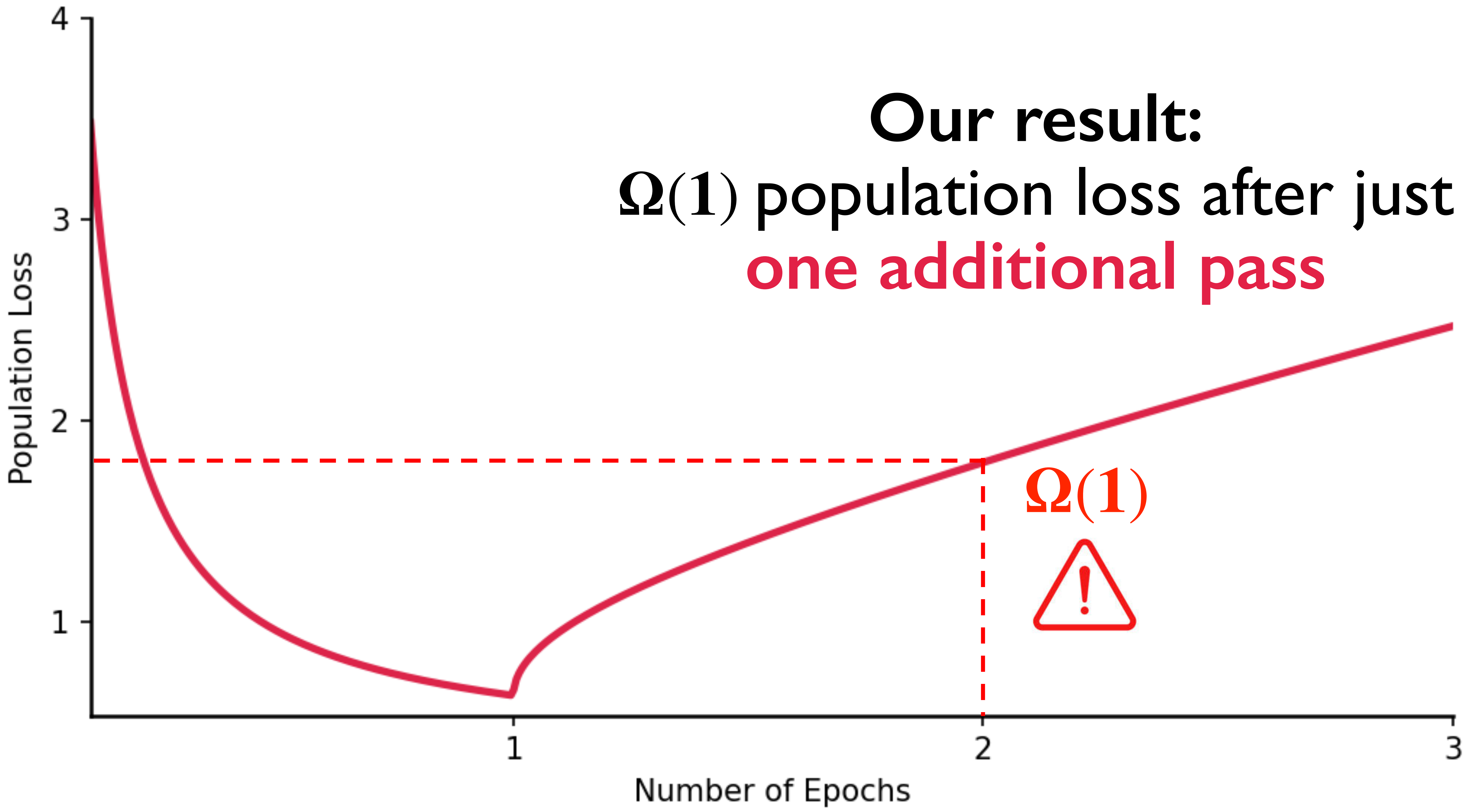
What happens if we keep training for more epochs?



Multi-Pass SGD in SCO

with standard step size of $\eta = 1/\sqrt{n}$, optimal after one pass.

What happens if we keep training for more epochs?



Our result:

$\Omega(1)$ population loss after just **one additional pass**

Tight bounds for Multi-Pass SGD

Multi-pass SGD (without-replacement):

$\Omega\left(\eta\sqrt{T} + \frac{\eta T}{n} + \frac{1}{\eta T}\right)$ population loss from the (end of the) 2nd epoch onward

With-replacement SGD:

$\Omega\left(\eta\sqrt{T} + \frac{\eta T}{n} + \frac{1}{\eta T}\right)$ population loss after $\Theta(n \log n)$ steps (i.e. $\log n$ "passes")

(thanks to coupon collector)

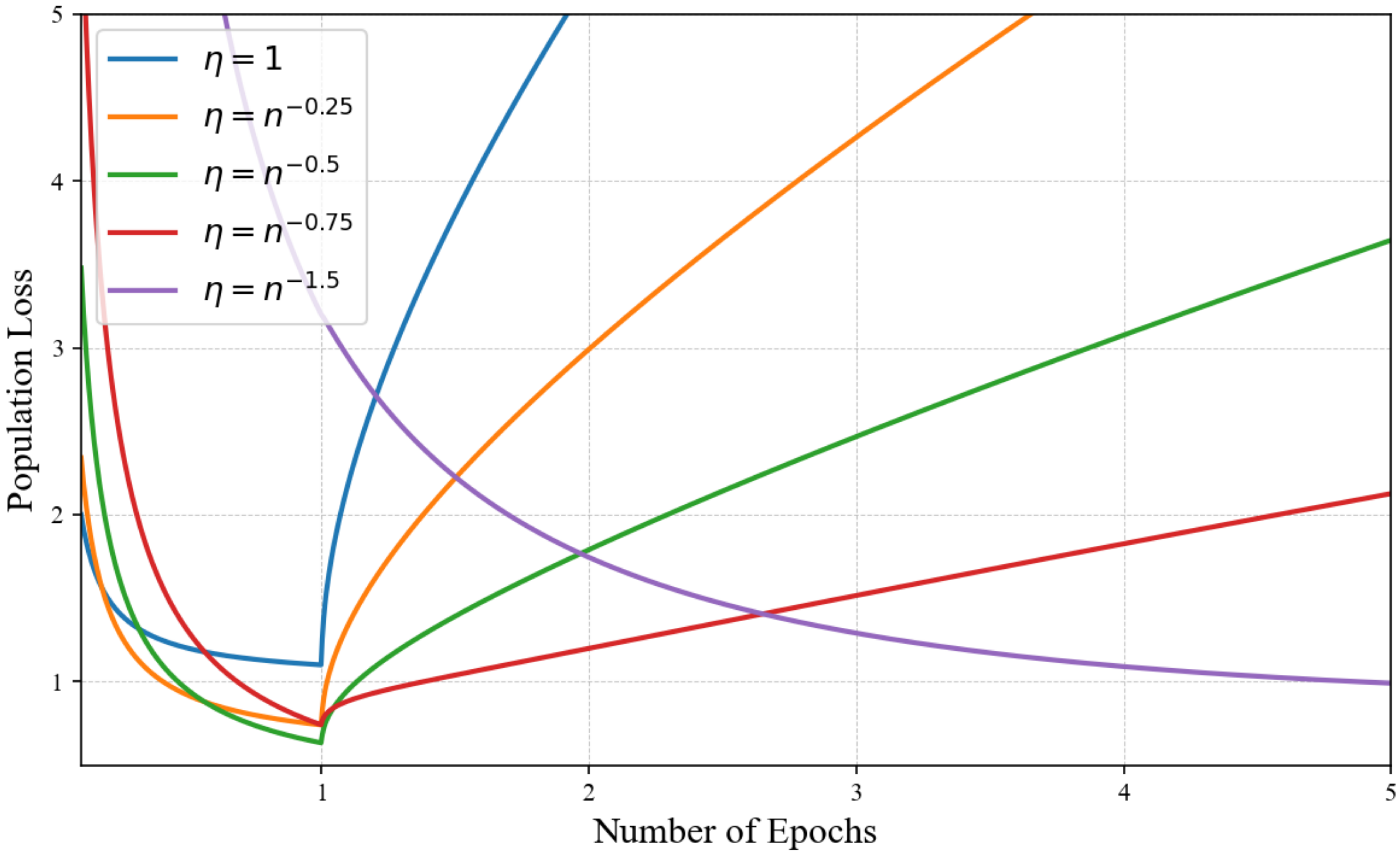
Matching upper bounds

Via algorithmic stability arguments

[Vansover-Hager, K, Livni '25]

Tight bounds for Multi-Pass SGD

for different stepsizes η

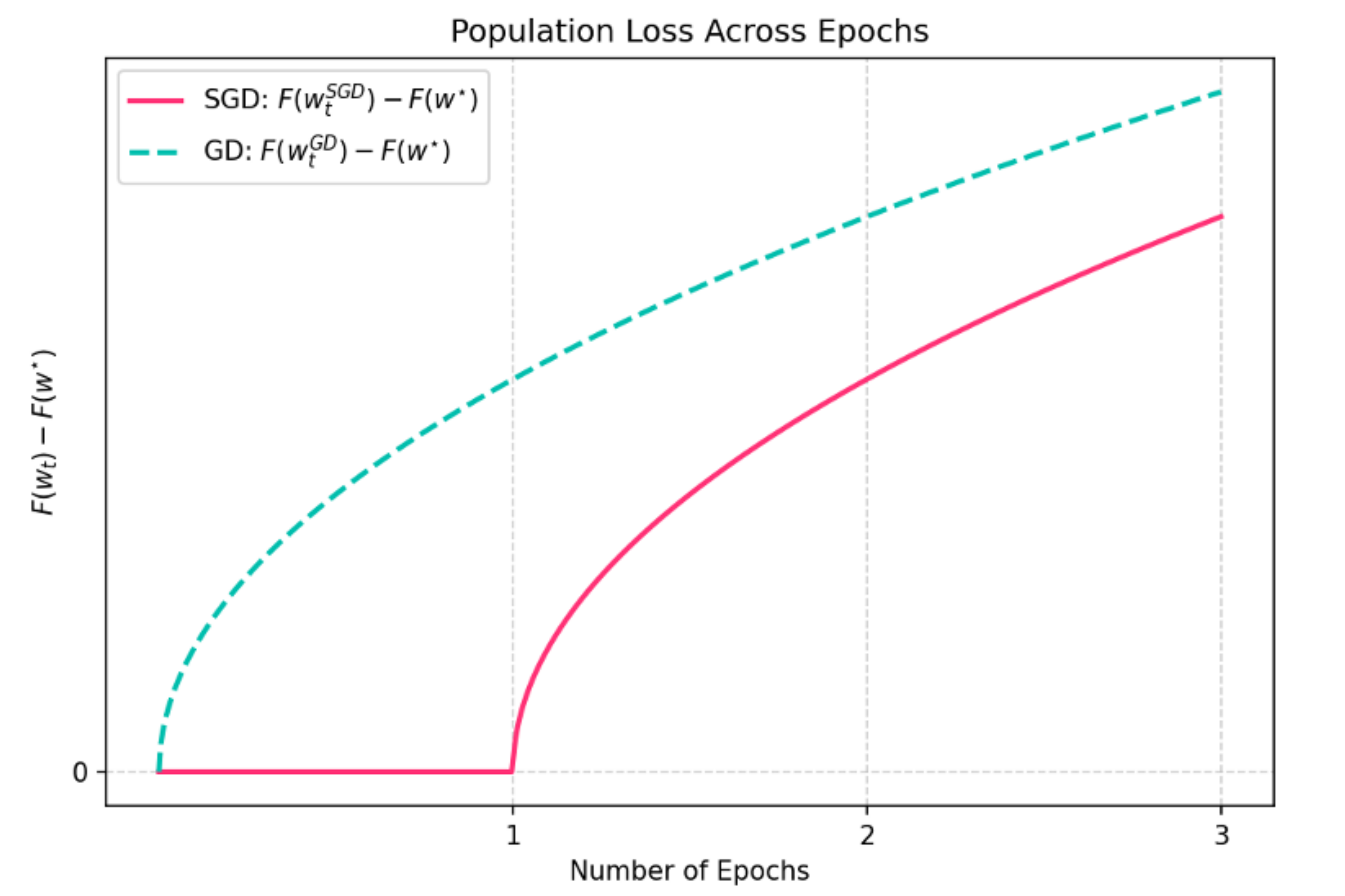


Multi-pass SGD: proof ideas

- GD observes the entire sample each step → overfitting may occur
  [Amir, K, Livni ’21, Schliserman, Sherman, K ’24, Livni ’24]
- SGD doesn't observe the entire training set in first pass → no overfitting

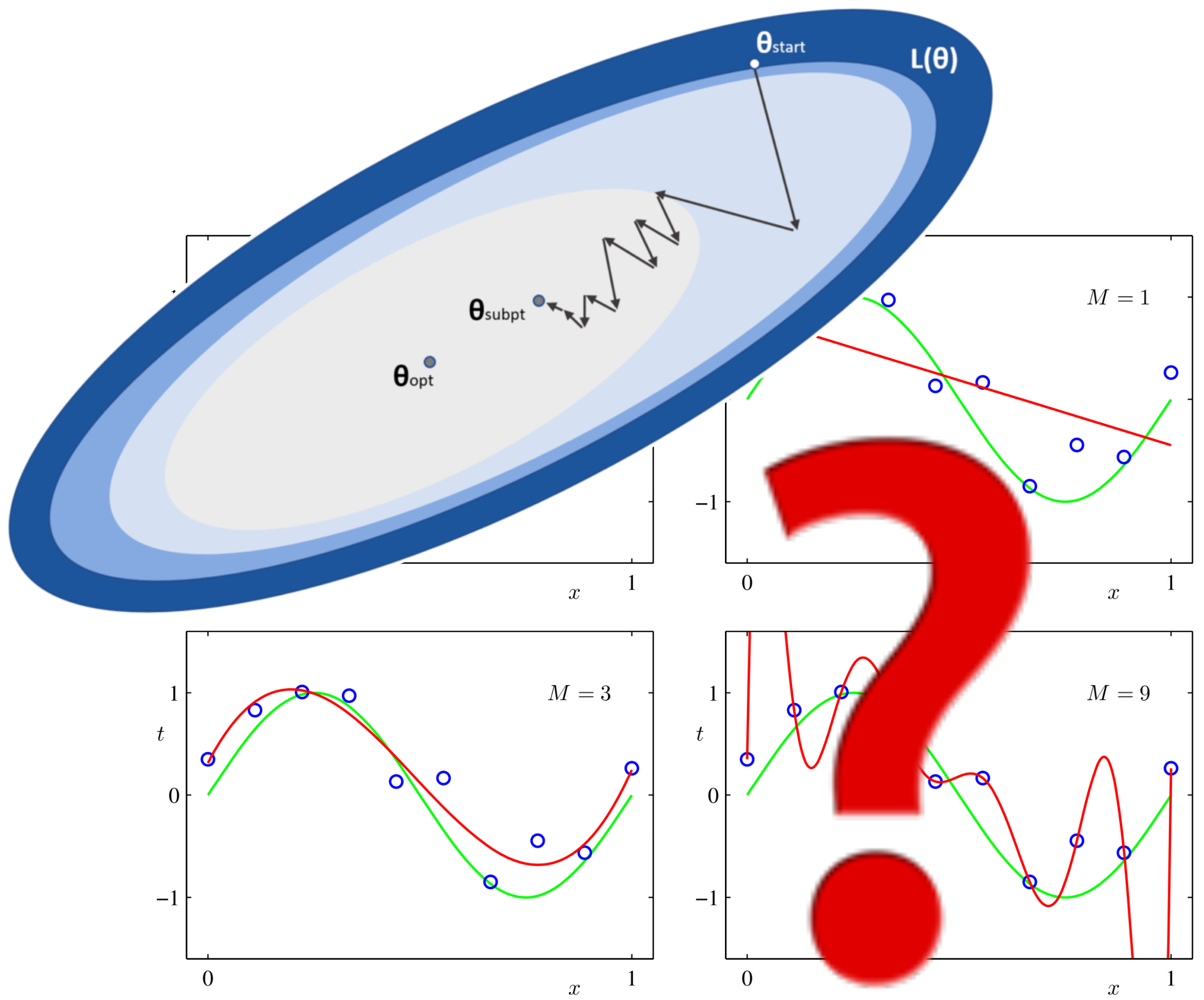
Multi-pass SGD: proof ideas

- GD observes the entire sample each step → overfitting may occur [Amir, K, Livni '21, Schliserman, Sherman, K '24, Livni '24]
- SGD doesn't observe the entire training set in first pass → no overfitting



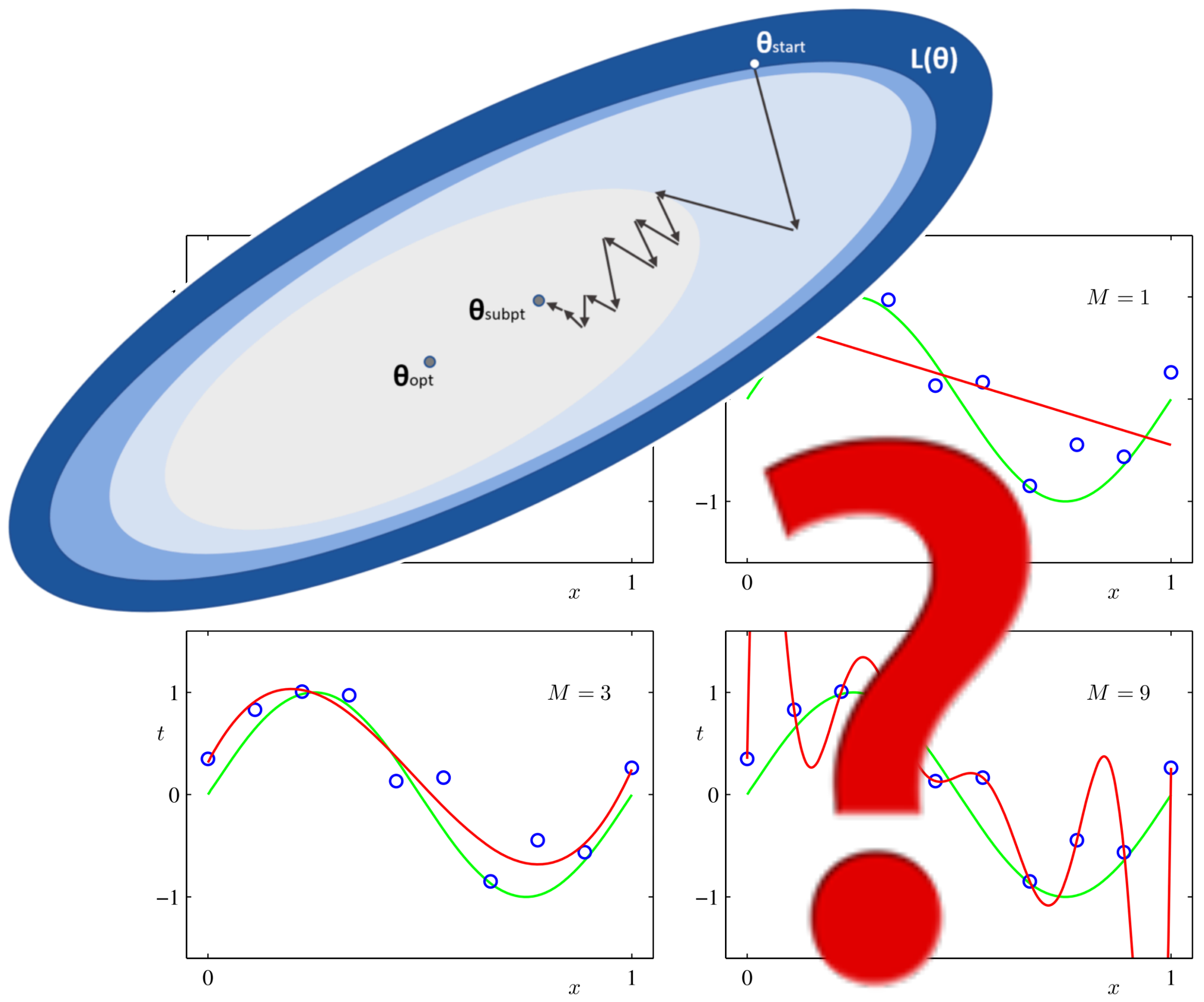
- **Key idea:** use first pass to "touch" and memorize entire sample
- Construct loss function s.t. SGD steps $\approx$ remain at init while memorizing
- Once sample has been memorized, overfitting starts

Takeaways



- Classical SGD doesn't "fit" in conventional statistical learning theory
- Perspective to contemporary discussion on generalization in modern ML
- Emerging picture: first pass is in "regret regime", later passes governed by algorithmic stability
- More results for (full batch) Gradient Descent, more general full-batch methods, Sharpness-aware algorithms (SAM), …

Takeaways



- Classical SGD doesn't "fit" in conventional statistical learning theory
- Perspective to contemporary discussion on generalization in modern ML
- Emerging picture: first pass is in "regret regime", later passes governed by algorithmic stability
- More results for (full batch) Gradient Descent, more general full-batch methods, Sharpness-aware algorithms (SAM), …

Thanks!

Funded by the European Union

European Research Council
Established by the European Commission