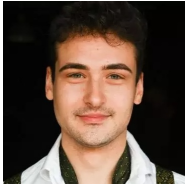
Mind the U Curve: Why We Stop Gradient Descent Early

Shirley Xiaoqi Liu

Joint work with:



Alex Buna

Patrick Rebeschini

Meet Our Group Seminar Series
Department of Statistics, University of Oxford

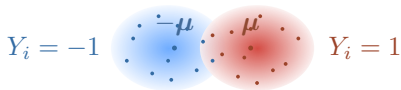
27 May 2026

A canonical classification task

A canonical classification task

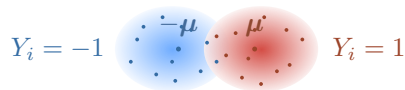
- **IID samples** $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$**.** Binary Gaussian mixtures

A canonical classification task



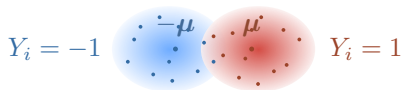
- **IID samples** $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$**.** Binary Gaussian mixtures

A canonical classification task



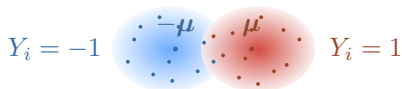
- **IID samples** $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$**.** Binary Gaussian mixtures $\boldsymbol{X}_i = Y_i \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i \in \mathbb{R}^d$ with $Y_i \sim \text{uniform}\{\pm 1\}$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$

A canonical classification task



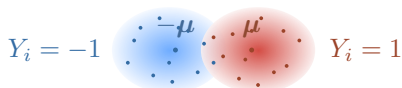
- **IID samples** $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$**.** Binary Gaussian mixtures $\boldsymbol{X}_i = Y_i \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i \in \mathbb{R}^d$ with $Y_i \sim \text{uniform}\{\pm 1\}$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
- **Regime.** $d \geq n$

A canonical classification task



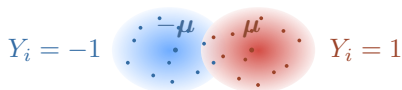
- **IID samples** $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$**.** Binary Gaussian mixtures $\boldsymbol{X}_i = Y_i \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i \in \mathbb{R}^d$ with $Y_i \sim$ uniform$\{\pm 1\}$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
- **Regime.** $d \geq n \Rightarrow$ samples are linearly-separable

A canonical classification task



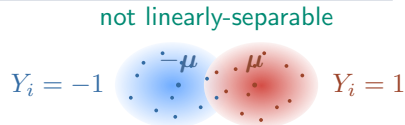
- **IID samples** $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$**.** Binary Gaussian mixtures $\boldsymbol{X}_i = Y_i \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i \in \mathbb{R}^d$ with $Y_i \sim \text{uniform}\{\pm 1\}$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
- **Regime.** $d \geq n \Rightarrow$ samples are linearly-separable $\Rightarrow$ gradient descent (GD) on logistic loss produces iterates $(\boldsymbol{w}_t)_{t \in \mathbb{N}}$ with $\|\boldsymbol{w}_t\| \to \infty$ and direction $\frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}$ converging to max-margin classifier $\boldsymbol{x} \mapsto \text{sign}(\boldsymbol{x}^\top \boldsymbol{w}_\infty)$ Soudry et al. [2018], Shamir [2021]

A canonical classification task



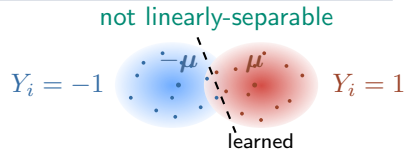
- **IID samples** $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$**.** Binary Gaussian mixtures $\boldsymbol{X}_i = Y_i \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i \in \mathbb{R}^d$ with $Y_i \sim \text{uniform}\{\pm 1\}$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
- **Regime.** $d \geq n \Rightarrow$ samples are linearly-separable $\Rightarrow$ gradient descent (GD) on logistic loss produces iterates $(\boldsymbol{w}_t)_{t \in \mathbb{N}}$ with $\|\boldsymbol{w}_t\| \to \infty$ and direction $\frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}$ converging to max-margin classifier $\boldsymbol{x} \mapsto \text{sign}(\boldsymbol{x}^\top \boldsymbol{w}_\infty)$ Soudry et al. [2018], Shamir [2021]

A canonical classification task



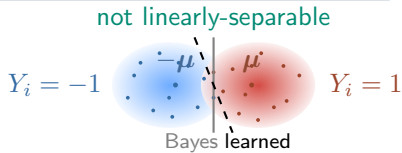
- **IID samples** $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$**.** Binary Gaussian mixtures $\boldsymbol{X}_i = Y_i \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i \in \mathbb{R}^d$ with $Y_i \sim \text{uniform}\{\pm 1\}$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
- **Regime.** $d \geq n \Rightarrow$ samples are linearly-separable $\Rightarrow$ gradient descent (GD) on logistic loss produces iterates $(\boldsymbol{w}_t)_{t \in \mathbb{N}}$ with $\|\boldsymbol{w}_t\| \to \infty$ and direction $\frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}$ converging to max-margin classifier $\boldsymbol{x} \mapsto \text{sign}(\boldsymbol{x}^\top \boldsymbol{w}_\infty)$ Soudry et al. [2018], Shamir [2021]

A canonical classification task



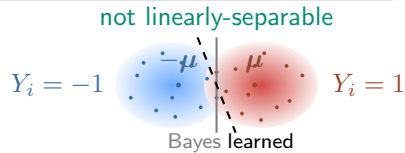
- **IID samples** $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$. Binary Gaussian mixtures $\boldsymbol{X}_i = Y_i \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i \in \mathbb{R}^d$ with $Y_i \sim \text{uniform}\{\pm 1\}$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
- **Regime.** $d \geq n \Rightarrow$ samples are linearly-separable $\Rightarrow$ gradient descent (GD) on logistic loss produces iterates $(\boldsymbol{w}_t)_{t \in \mathbb{N}}$ with $\|\boldsymbol{w}_t\| \to \infty$ and direction $\frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}$ converging to max-margin classifier $\boldsymbol{x} \mapsto \text{sign}(\boldsymbol{x}^\top \boldsymbol{w}_\infty)$ Soudry et al. [2018], Shamir [2021]

A canonical classification task



- **IID samples** $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$**.** Binary Gaussian mixtures $\boldsymbol{X}_i = Y_i \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i \in \mathbb{R}^d$ with $Y_i \sim \text{uniform}\{\pm 1\}$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
- **Regime.** $d \geq n \Rightarrow$ samples are linearly-separable $\Rightarrow$ gradient descent (GD) on logistic loss produces iterates $(\boldsymbol{w}_t)_{t \in \mathbb{N}}$ with $\|\boldsymbol{w}_t\| \to \infty$ and direction $\frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}$ converging to max-margin classifier $\boldsymbol{x} \mapsto \text{sign}(\boldsymbol{x}^\top \boldsymbol{w}_\infty)$ Soudry et al. [2018], Shamir [2021]

A canonical classification task



- **IID samples** $S := (\boldsymbol{X}_i, Y_i)_{i=1}^n$. Binary Gaussian mixtures $\boldsymbol{X}_i = Y_i \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i \in \mathbb{R}^d$ with $Y_i \sim \text{uniform}\{\pm 1\}$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
- **Regime.** $d \geq n \Rightarrow$ samples are linearly-separable $\Rightarrow$ gradient descent (GD) on logistic loss produces iterates $(\boldsymbol{w}_t)_{t \in \mathbb{N}}$ with $\|\boldsymbol{w}_t\| \to \infty$ and direction $\frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}$ converging to max-margin classifier $\boldsymbol{x} \mapsto \text{sign}(\boldsymbol{x}^\top \boldsymbol{w}_\infty)$ Soudry et al. [2018], Shamir [2021]
- **Goal.** Understand how the excess zero-one risk

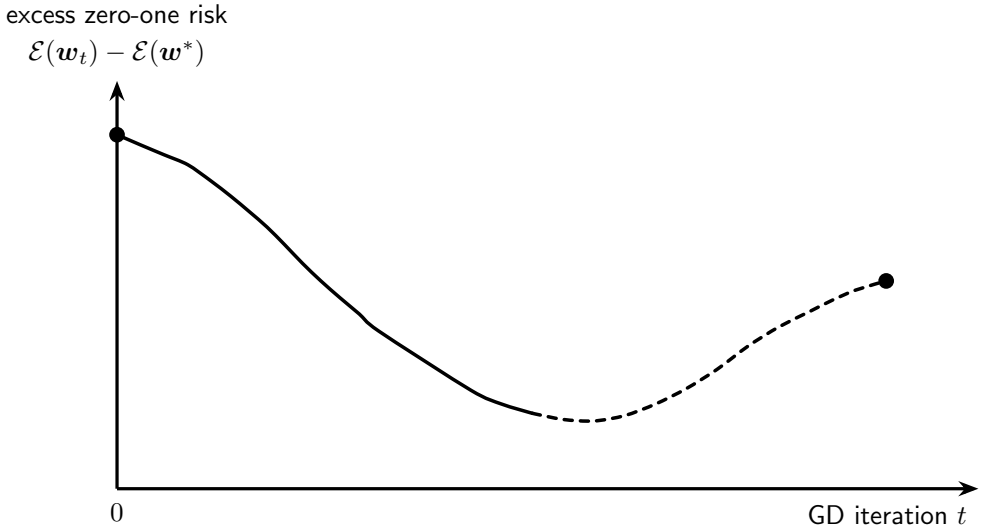
$$\mathcal{E}(\boldsymbol{w}_t) - \mathcal{E}(\boldsymbol{w}^*)$$

evolves along the GD trajectory, where $\mathcal{E}(\boldsymbol{w}) := \mathbb{P}\big(\text{sign}(\boldsymbol{X}_i^\top \boldsymbol{w}) \neq Y_i\big)$ and $\boldsymbol{w}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is Bayes classifier

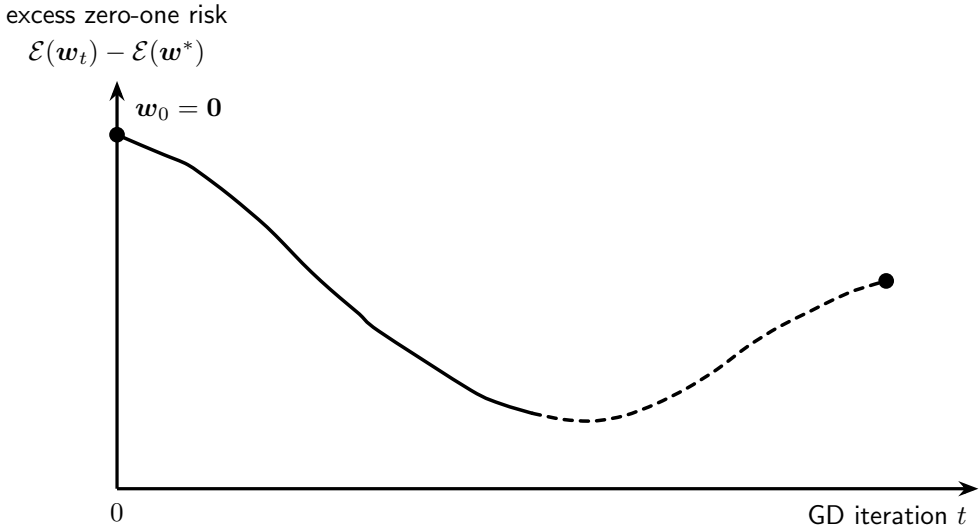
The schematic U curve

The schematic U curve

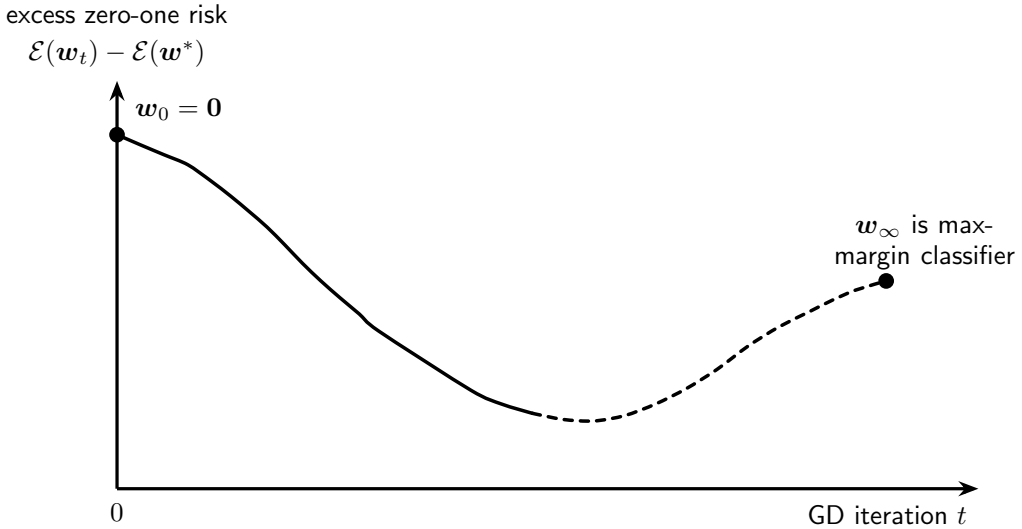
The schematic U curve



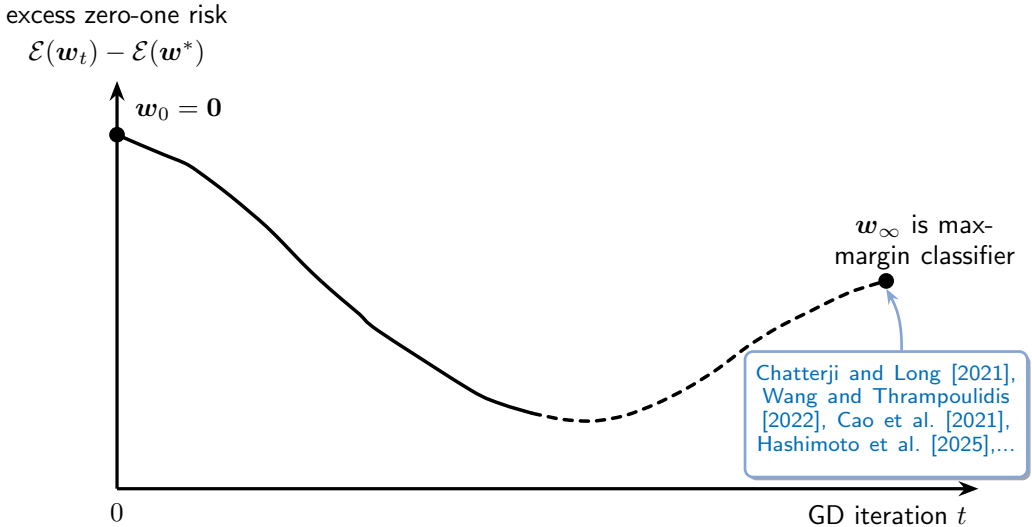
The schematic U curve



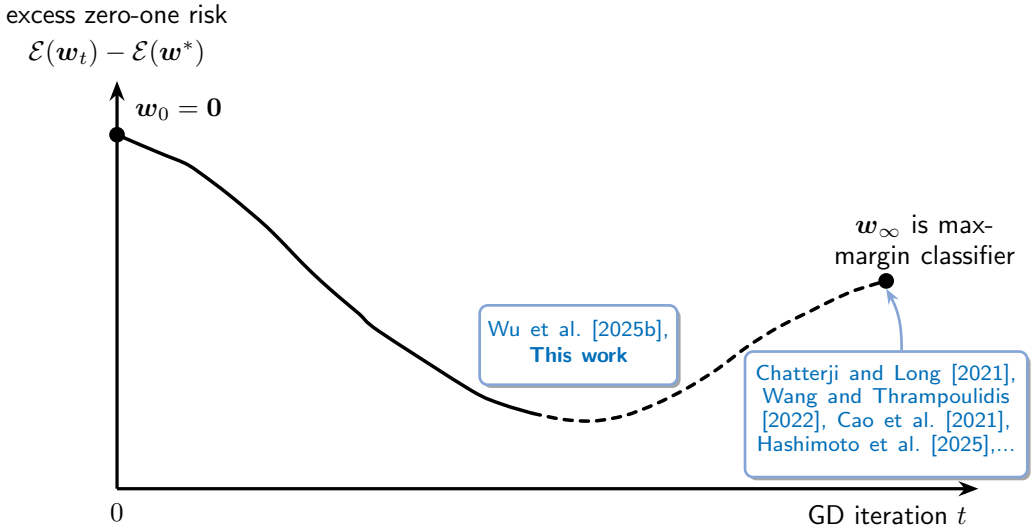
The schematic U curve



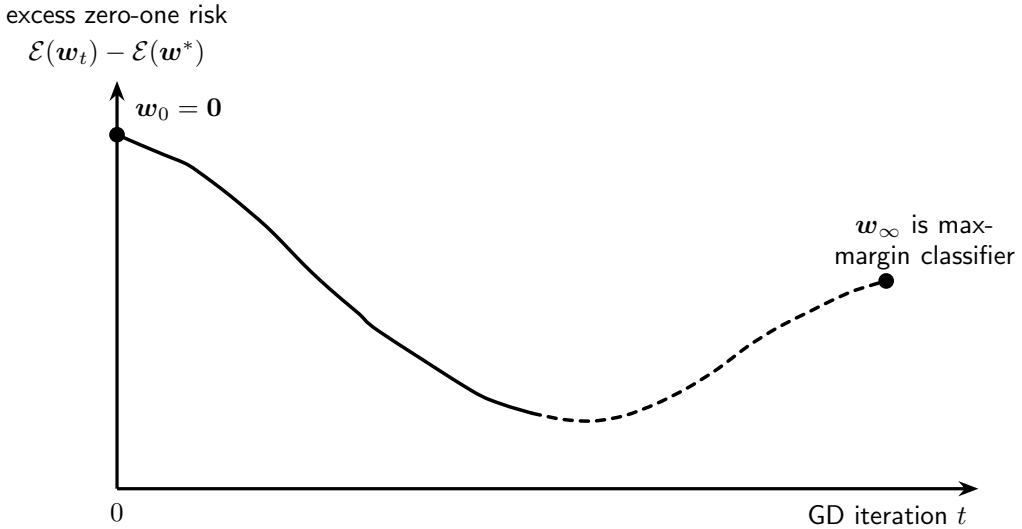
The schematic U curve



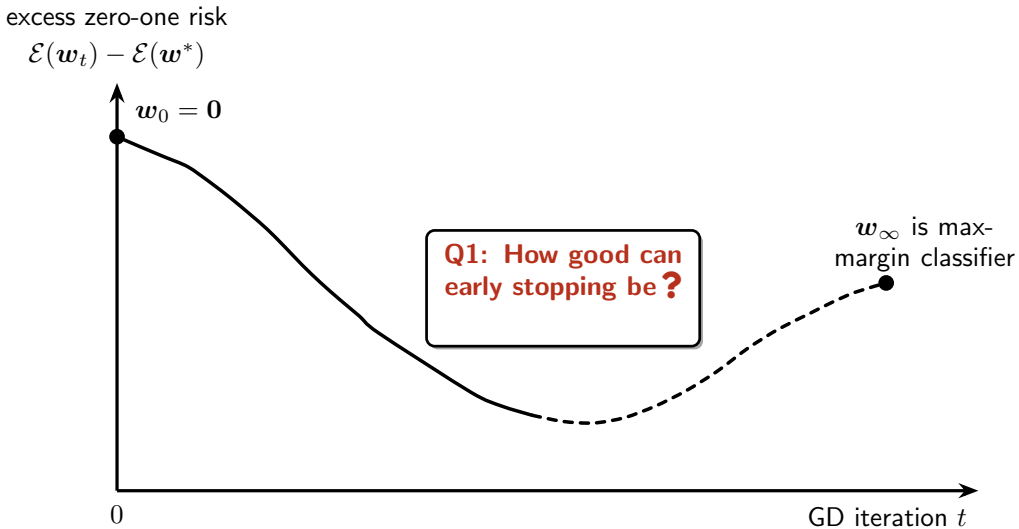
The schematic U curve



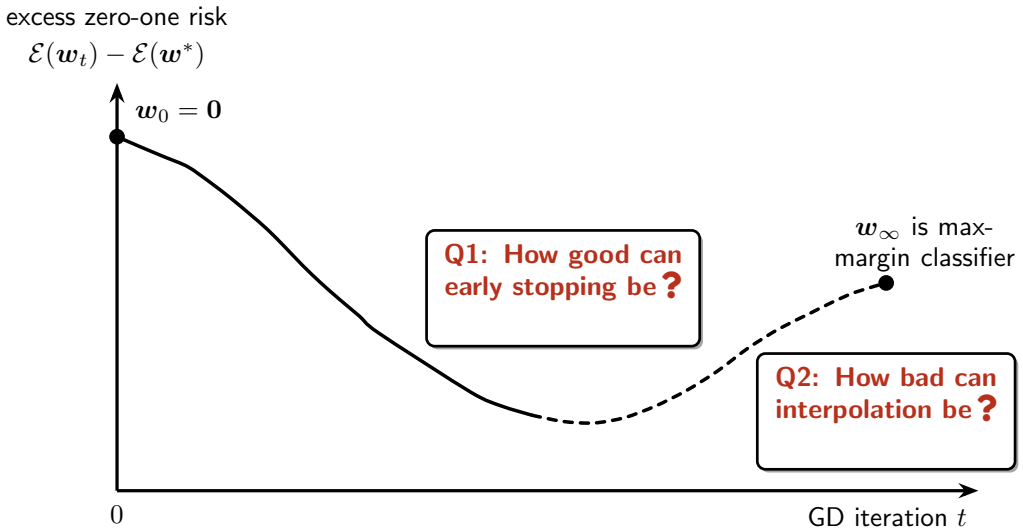
The schematic U curve



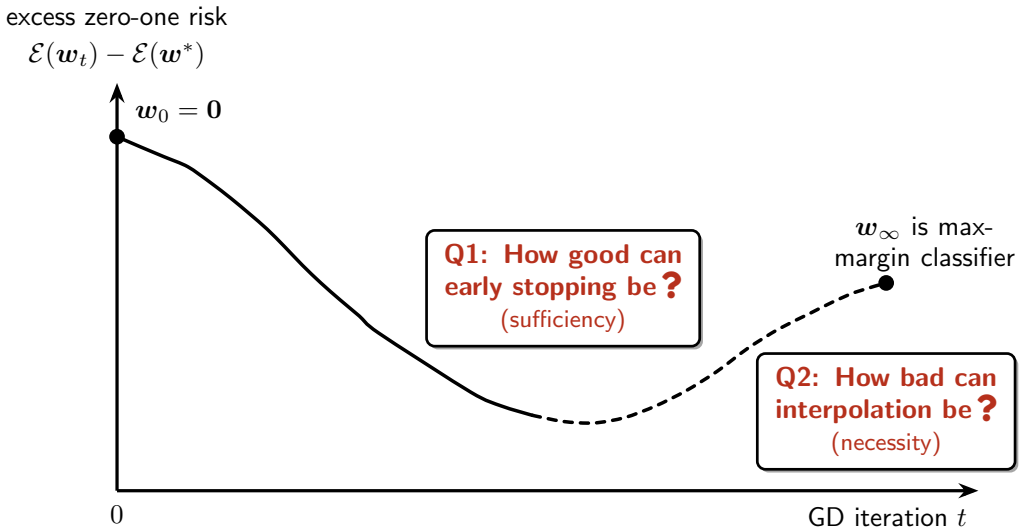
The schematic U curve



The schematic U curve

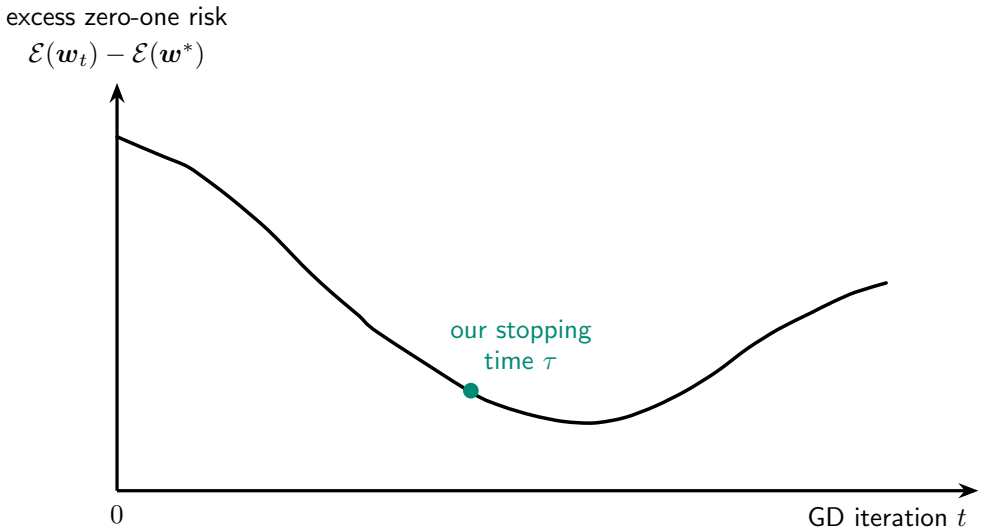


The schematic U curve

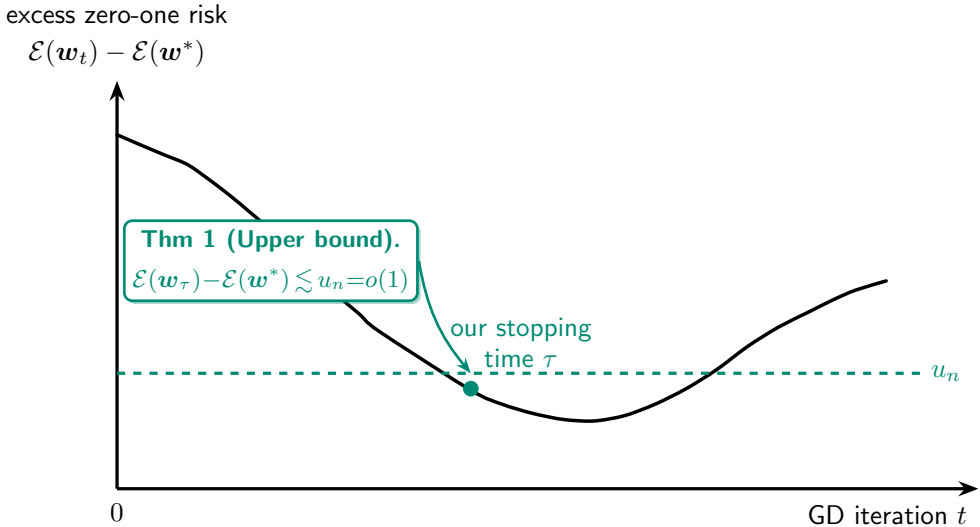


Q1: How good can early stopping be?

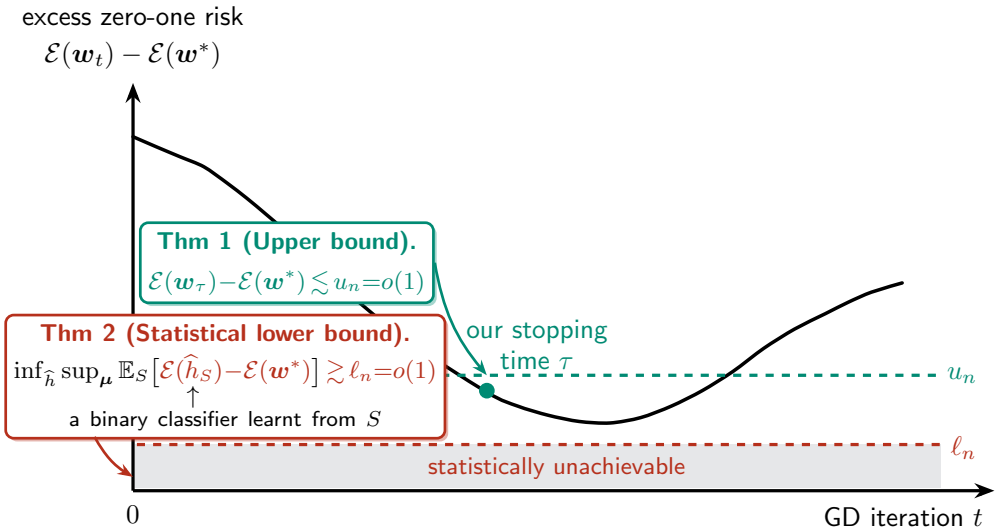
Q1: How good can early stopping be?



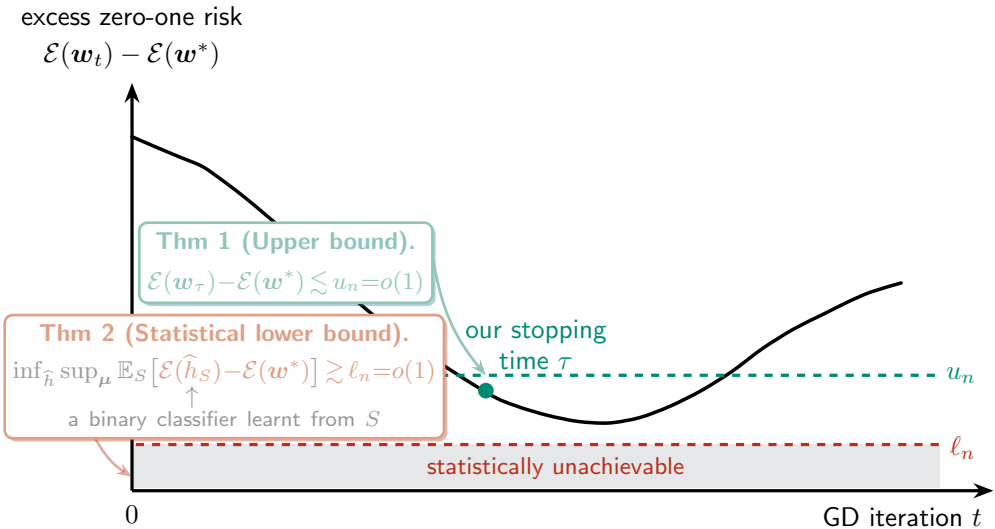
Q1: How good can early stopping be?



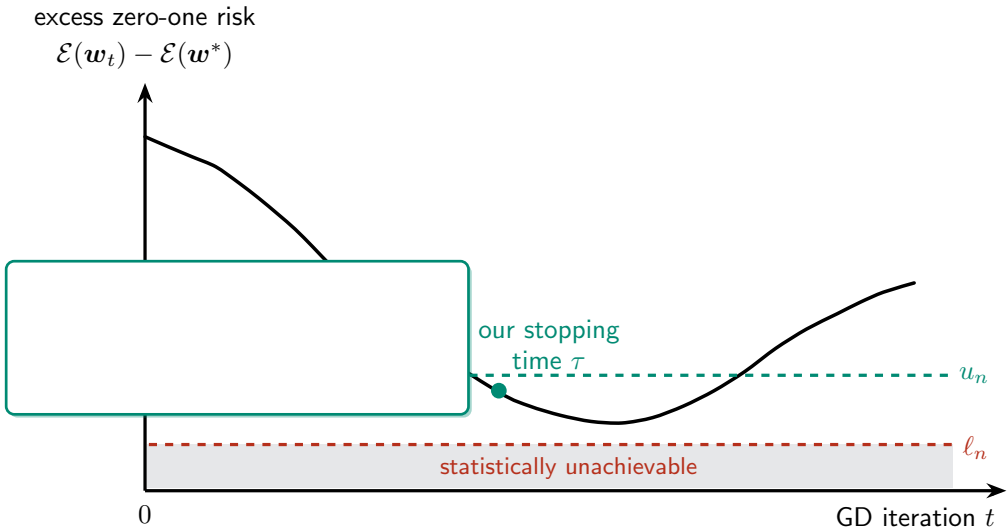
Q1: How good can early stopping be?



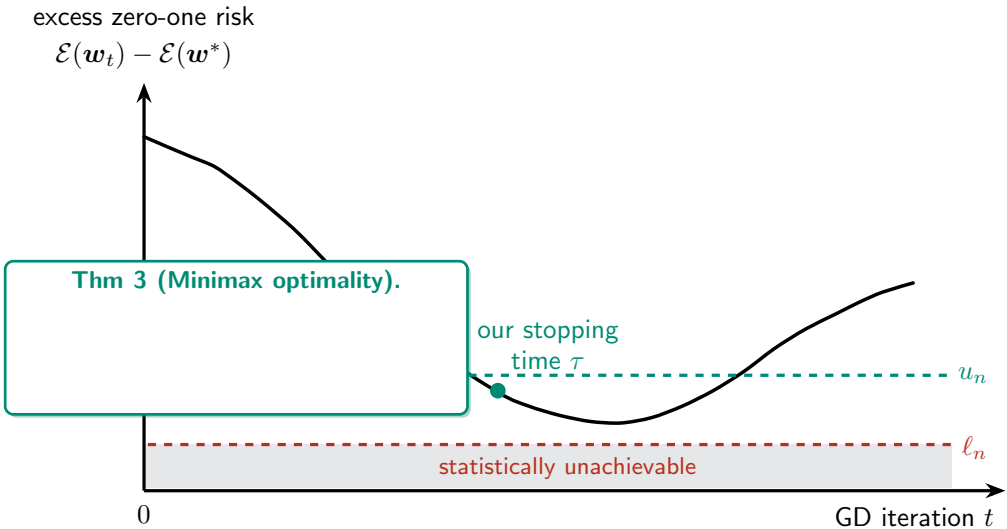
Q1: How good can early stopping be?



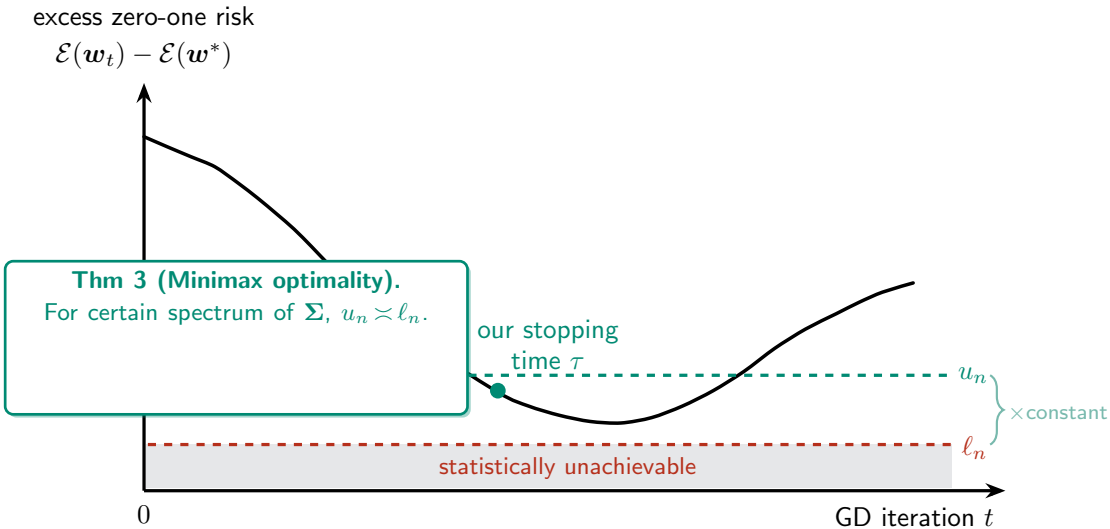
Q1: How good can early stopping be?



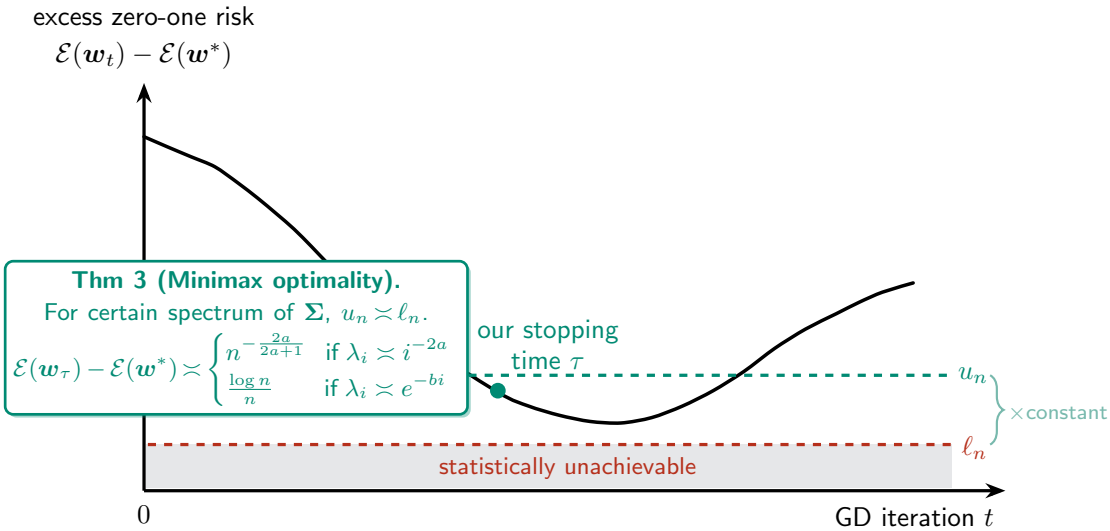
Q1: How good can early stopping be?



Q1: How good can early stopping be?



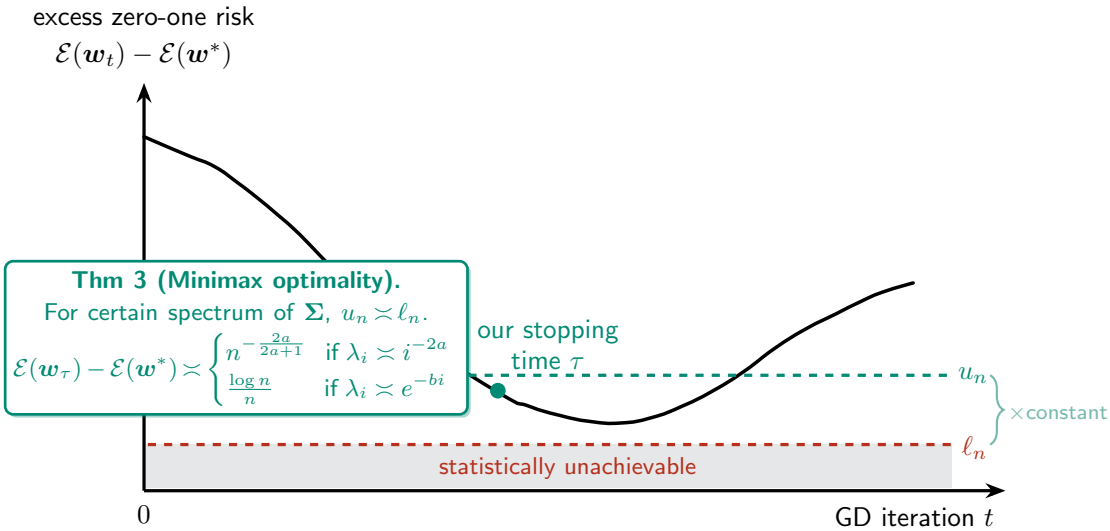
Q1: How good can early stopping be?



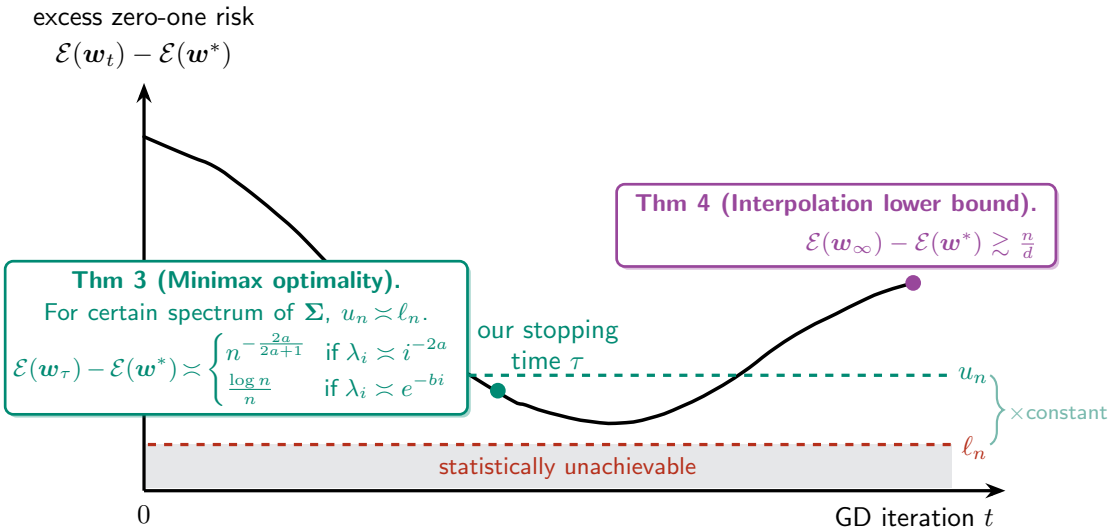
Q1: How good can early stopping be?

Q2: How bad can interpolation be?

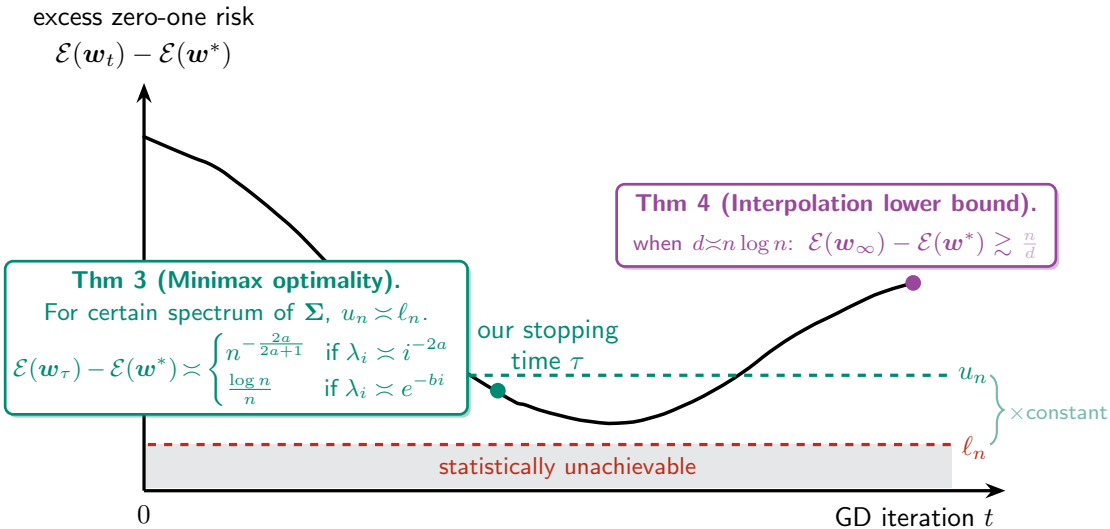
Q2: How bad can interpolation be?



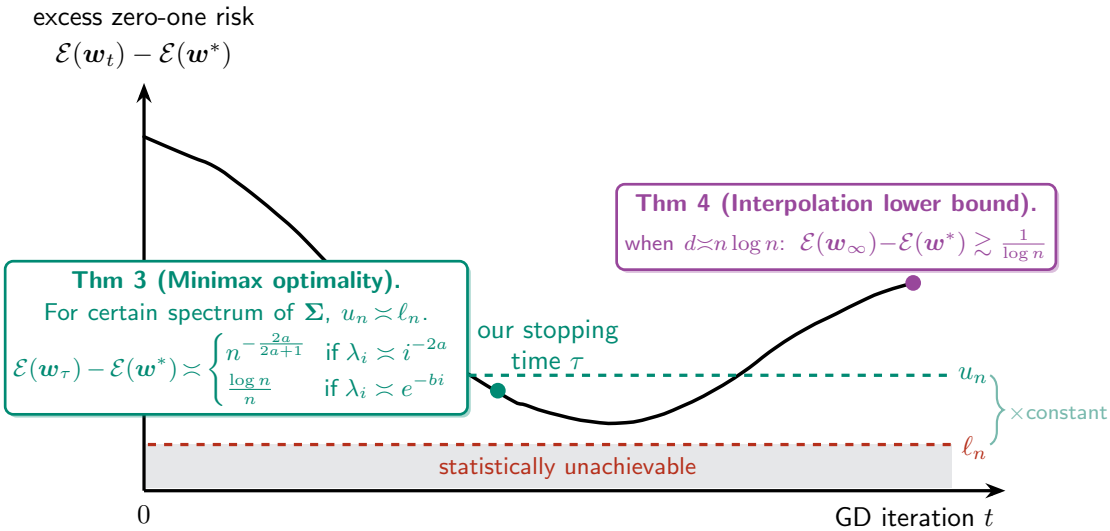
Q2: How bad can interpolation be?



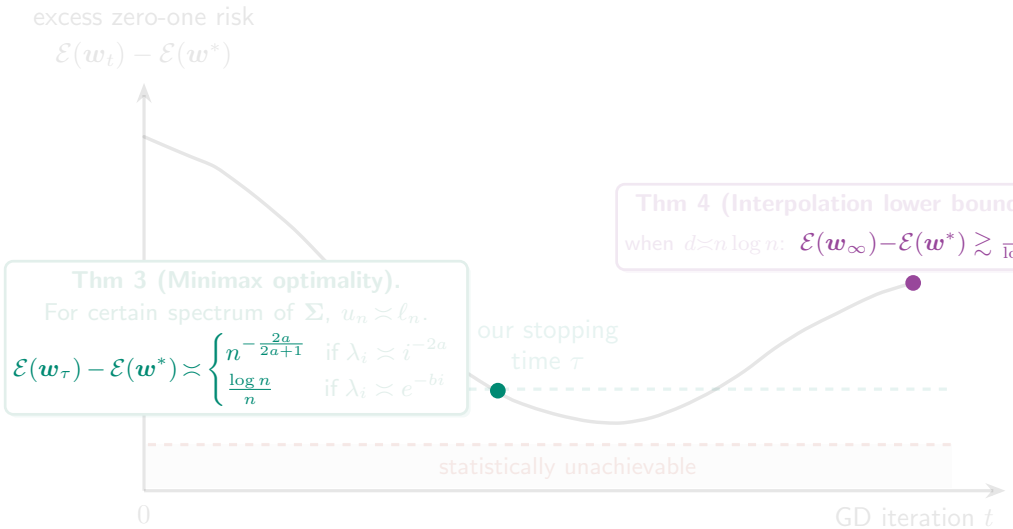
Q2: How bad can interpolation be?



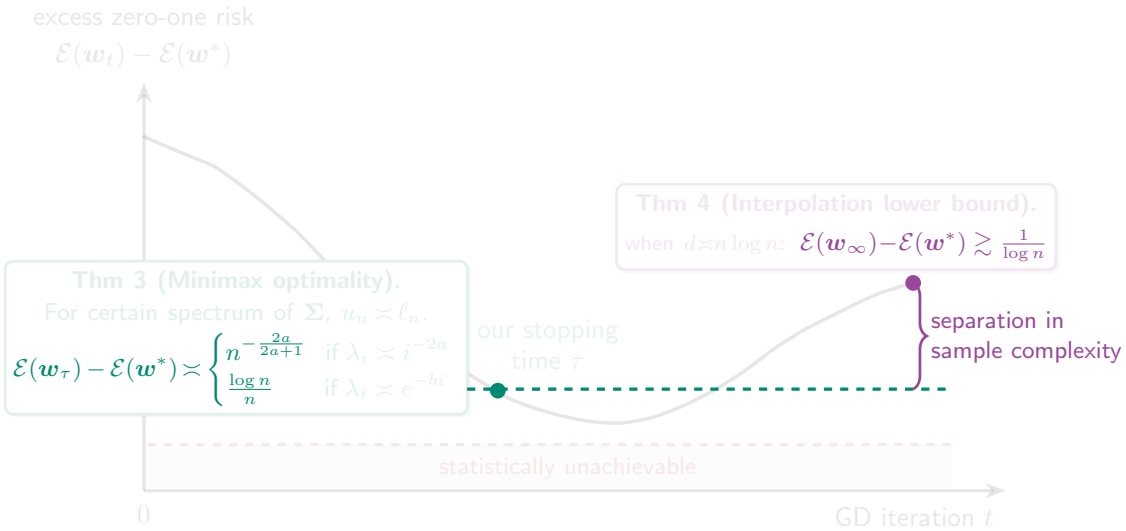
Q2: How bad can interpolation be?



Q2: How bad can interpolation be?

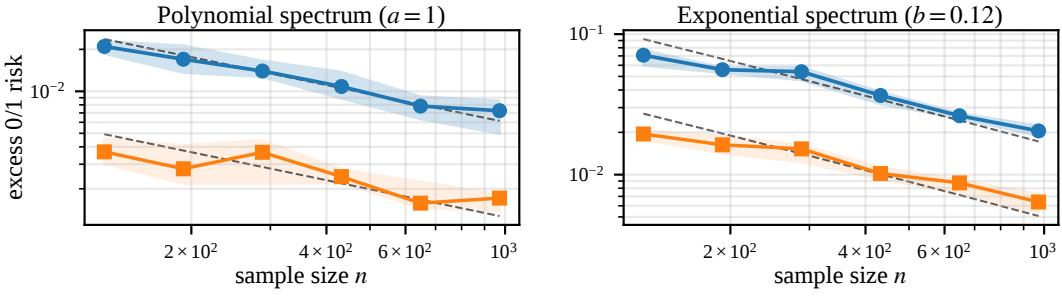


Q2: How bad can interpolation be?



Theory is validated by experiments

Theory is validated by experiments



$$\mathcal{E}(\boldsymbol{w}_\tau) - \mathcal{E}(\boldsymbol{w}^*) \asymp \begin{cases} n^{-\frac{2a}{2a+1}} & \text{if } \lambda_i \asymp i^{-2a} \\ \frac{\log n}{n} & \text{if } \lambda_i \asymp e^{-bi} \end{cases}$$

Summary

Summary: sufficiency & necessity of early stopping

Summary: sufficiency & necessity of early stopping

Sufficiency:

Summary: sufficiency & necessity of early stopping

Sufficiency:

- Early-stopped GD achieves minimax-optimal excess risk for certain data spectrum (e.g., $\lambda_i \asymp i^{-2a}$, $\lambda_i \asymp e^{-bi}$).

Summary: sufficiency & necessity of early stopping

Sufficiency:

- Early-stopped GD achieves minimax-optimal excess risk for certain data spectrum (e.g., $\lambda_i \asymp i^{-2a}$, $\lambda_i \asymp e^{-bi}$).
- Matching upper and lower bounds for early-stopped GD (Thm. 3)

Summary: sufficiency & necessity of early stopping

Sufficiency:

- Early-stopped GD achieves minimax-optimal excess risk for certain data spectrum (e.g., $\lambda_i \asymp i^{-2a}$, $\lambda_i \asymp e^{-bi}$).
- Matching upper and lower bounds for early-stopped GD (Thm. 3)

Necessity:

Summary: sufficiency & necessity of early stopping

Sufficiency:

- Early-stopped GD achieves minimax-optimal excess risk for certain data spectrum (e.g., $\lambda_i \asymp i^{-2a}$, $\lambda_i \asymp e^{-bi}$).
- Matching upper and lower bounds for early-stopped GD (Thm. 3)

Necessity:

- Interpolation requires exponentially worse sample complexity in mildly over-parameterized regimes (e.g., $d \asymp n \log n$).

Summary: sufficiency & necessity of early stopping

Sufficiency:

- Early-stopped GD achieves minimax-optimal excess risk for certain data spectrum (e.g., $\lambda_i \asymp i^{-2a}$, $\lambda_i \asymp e^{-bi}$).
- Matching upper and lower bounds for early-stopped GD (Thm. 3)

Necessity:

- Interpolation requires exponentially worse sample complexity in mildly over-parameterized regimes (e.g., $d \asymp n \log n$).
- Interpolation lower bound (Thm. 4)

Summary: sufficiency & necessity of early stopping

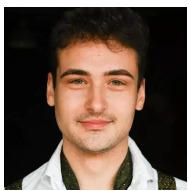
Sufficiency:

- Early-stopped GD achieves minimax-optimal excess risk for certain data spectrum (e.g., $\lambda_i \asymp i^{-2a}$, $\lambda_i \asymp e^{-bi}$).
- Matching upper and lower bounds for early-stopped GD (Thm. 3)

Necessity:

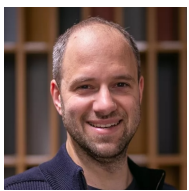
- Interpolation requires exponentially worse sample complexity in mildly over-parameterized regimes (e.g., $d \asymp n \log n$).
- Interpolation lower bound (Thm. 4)

Full paper: Alex Buna, Shirley Xiaoqi Liu, and Patrick Rebeschini. Minimax Optimal Early-Stopped Gradient Descent for Gaussian Mixture Classification, 2026. *(To appear)*



Alex

Shirley



Patrick

Sample covariates $\mathbf{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]^\top \in \mathbb{R}^{n \times d}$, step size $\gamma > 0$, iteration number t.

Least-squares regression

Objective:

$$\widehat{\mathcal{L}}(\boldsymbol{w}) = \frac{1}{2n} \|\mathbf{X}\boldsymbol{w} - \boldsymbol{Y}\|^2$$

$\Rightarrow$ gradient *linear* in $\boldsymbol{w}$
$\Rightarrow$ GD telescopes to **closed form**:

$$\boldsymbol{w}_t = \underbrace{\left(\boldsymbol{I} - \left(\boldsymbol{I} - \tfrac{\gamma}{n}\mathbf{X}^\top\mathbf{X}\right)^t\right)}_{\text{spectral filter, explicit in } t} (\mathbf{X}^\top\mathbf{X})^\dagger \mathbf{X}^\top \boldsymbol{Y}$$

As $t \to \infty$, $\boldsymbol{w}_t \to \boldsymbol{w}_\infty = (\mathbf{X}^\top\mathbf{X})^\dagger \mathbf{X}^\top \boldsymbol{Y}$
is *min-ℓ_2-norm interpolator*

Linear classification

Objective:

$$\widehat{\mathcal{L}}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-Y_i \boldsymbol{X}_i^\top \boldsymbol{w}}\right)$$

$\Rightarrow$ gradient *nonlinear* in $\boldsymbol{w}$
$\Rightarrow$ **no** closed form, only a recursion:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \frac{\gamma}{n} \sum_{i=1}^{n} \underbrace{\sigma(-Y_i \boldsymbol{X}_i^\top \boldsymbol{w}_t)}_{\text{nonlinear in } \boldsymbol{w}_t} Y_i \boldsymbol{X}_i$$

As $t \to \infty$, $\frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|} \to \frac{\boldsymbol{w}_\infty}{\|\boldsymbol{w}_\infty\|}$
$\boldsymbol{w}_\infty = \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \|\boldsymbol{w}\|$ s.t. $Y_i \boldsymbol{X}_i^\top \boldsymbol{w} \geq 1 \ \forall i$
is *max-margin classifier*

Table: Overview of the most closely related prior work and positioning of our work.

		Least-squares regression	Linear classification
Benign overfitting		Bartlett et al. [2020], Belkin et al. [2020], Hastie et al. [2022], Tsigler and Bartlett [2023], Muthukumar et al. [2020]...	Muthukumar et al. [2021], Hsu et al. [2021], Chatterji and Long [2021], Wang and Thrampoulidis [2022], Cao et al. [2021], Hashimoto et al. [2025], ...
Early-stopped GD	**Implicit regularisation**	Yao et al. [2007], Suggala et al. [2018], ...	Wu et al. [2025b]
	Statistical optimality	Bühlmann and Yu [2003], Raskutti et al. [2014], Dieuleveut and Bach [2016], Wu et al. [2025a], ...	**This work**

References I

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Peter Bühlmann and Bin Yu. Boosting with the l2 loss. *Journal of the American Statistical Association*, 98(462):324–339, 2003. doi: 10.1198/016214503000125.

Yuan Cao, Quanquan Gu, and Misha Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In *Advances in Neural Information Processing Systems*, 2021.

Niladri S. Chatterji and Philip M. Long. Finite-sample Analysis of Interpolating Linear Classifiers in the Overparameterized Regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363 – 1399, 2016. doi: 10.1214/15-AOS1391.

Ichiro Hashimoto, Stanislav Volgushev, and Piotr Zwiernik. Universality of benign overfitting in binary linear classification. *arXiv preprint arXiv:2501.10538*, 2025.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

References II

Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 91–99. PMLR, 2021.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020. doi: 10.1109/JSAIT.2020.2984716.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: does the loss function matter? *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Ohad Shamir. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22(85):1–20, 2021.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

References III

Arun Sai Suggala, Adarsh Prasad, and Pradeep Ravikumar. Connecting optimization and regularization paths. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 10631–10641, Red Hook, NY, USA, 2018. Curran Associates Inc.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

Ke Wang and Christos Thrampoulidis. Binary Classification of Gaussian Mixtures: Abundance of Support Vectors, Benign Overfitting, and Regularization. *SIAM Journal on Mathematics of Data Science*, 4(1): 260–284, 2022. doi: 10.1137/21M1415121. URL https://doi.org/10.1137/21M1415121.

Jingfeng Wu, Peter L Bartlett, Jason D Lee, Sham M Kakade, and Bin Yu. Risk comparisons in linear regression: Implicit regularization dominates explicit regularization. *arXiv preprint arXiv:2509.17251*, 2025a.

Jingfeng Wu, Peter L. Bartlett, Matus Telgarsky, and Bin Yu. Benefits of early stopping in gradient descent for overparameterized logistic regression. In *Proceedings of the 42nd International Conference on Machine Learning*. JMLR.org, 2025b.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On Early Stopping in Gradient Descent Learning. *Constructive Approximation*, 26(2):289–315, August 2007. ISSN 1432-0940. doi: 10.1007/s00365-006-0663-2.