<Emerging Topic of Financial Machine Learning Research>

Virtue of Complexity

2025.09.24

Financial Risk Engineering Lab
양재원

목차

01 Preliminaries: Mathematical Foundations of Statistical Learning

- Machine Learning의 최종적인 목표는 generalization error를 최소화하는 것
  - Given Sample $(x_1, y_1), \ldots, (x_N, y_N) \sim \mathcal{D}$,
  - Find Function $h \in \mathcal{H}$ for new sample $(x, y) \sim \mathcal{D}$,
  - Minimizes $R(h) = \mathbb{E}[\mathcal{L}(h(x), y)]$ for some loss function $\mathcal{L}$

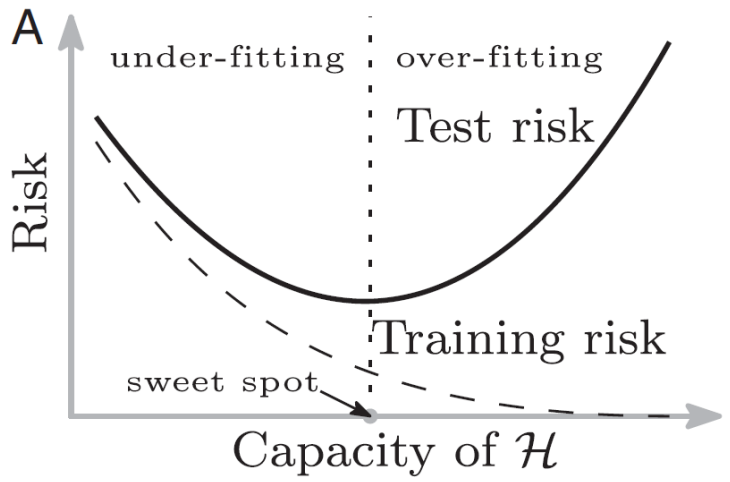
- Training error $\hat{R}(h) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(h(x_i), y_i))$로, ML 모델은 training error를 최소화하는 방향으로 학습됨

- 통계학습 이론에서 $R(h)$의 상한은 $\hat{R}(h)$와 $\mathcal{H}$의 capacity*의 합으로 표현됨→ bias-variance trade-off 발생

[1]

$$R(h) \leq \underbrace{\hat{R}(h)}_{\text{bias}} + \underbrace{O\left(\sqrt{\frac{VC(\mathcal{H}))}{N}}\right)}_{\text{variance}}$$

[2]



*$\mathcal{H}$의 capacity: 함수공간이 random data를 fitting하는 능력, Rademacher complexity나 VC dimension으로 측정됨

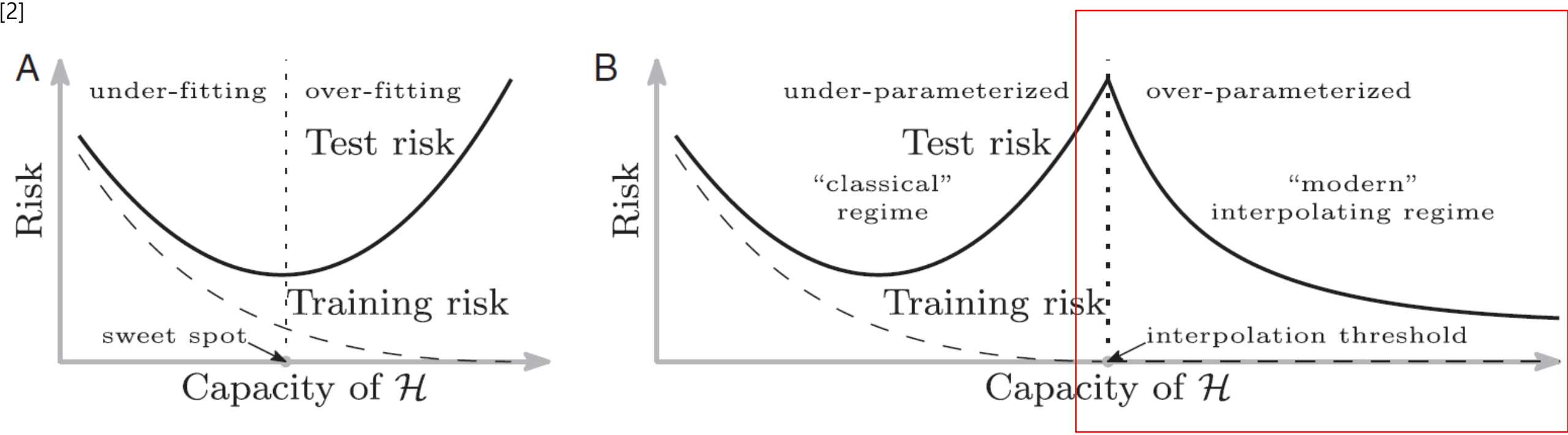
01 Preliminaries: Mathematical Foundations of Statistical Learning

- Deep Learning의 성공으로 통계학습 이론과 실제 현상의 괴리가 발생함
  - Deep Learning 모델은 전통적인 ML 모델들에 비해 $\mathcal{H}$의 capacity가 높음
  - 그럼에도 불구하고 training error와 generalization error 모두 전통적인 ML 모델에 비해 낮음

- 전통적인 통계학습 이론의 수정이 불가피해짐

Double Descent!

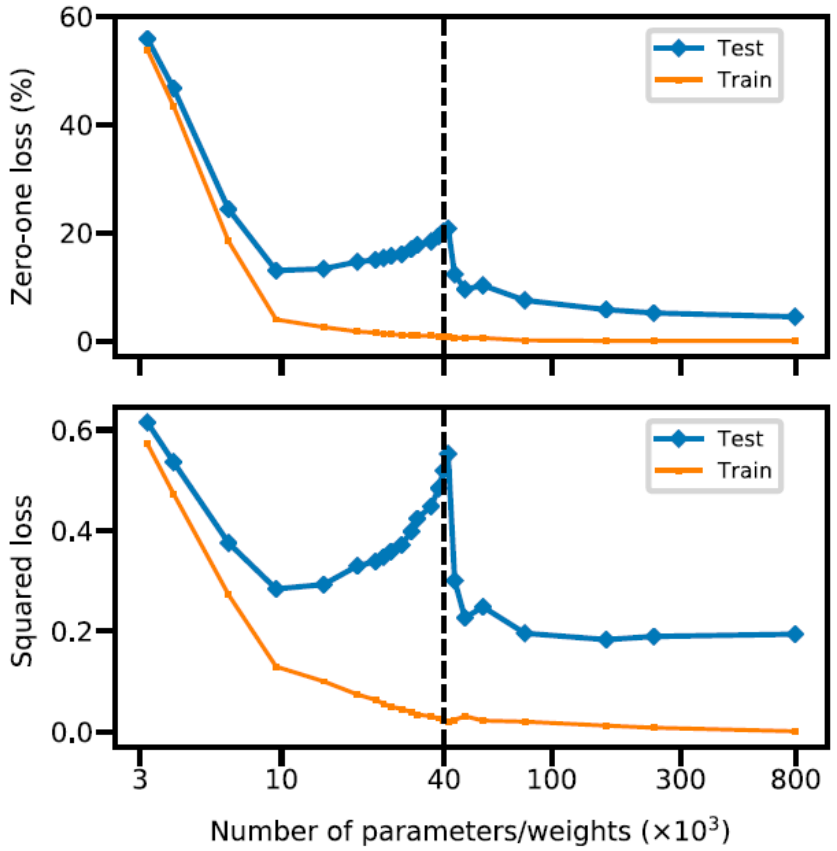
[2]



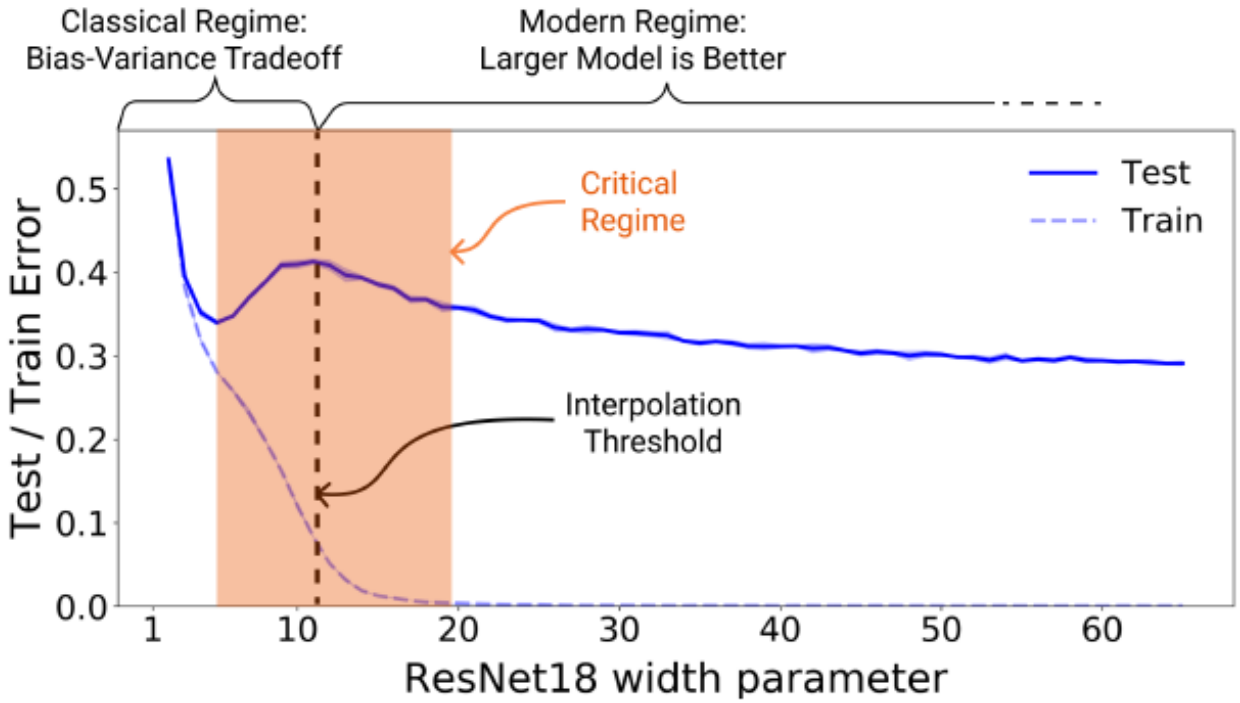
02 Double Descent: Virtue of Complexity in Statistical Learning

- **Double Descent**란 interpolating regime*에서 모델 복잡도가 증가할수록 generalization error가 감소하는 현상으로 다양한 문헌 [2, 3]에서 이 현상이 보고됨

[2] Belkin et al., 2019



[3] Nakkiran et al., 2019



*interpolating regime : Training Error가 0이거나 0에 매우 근접한 상태

02 Double Descent: Virtue of Complexity in Statistical Learning

- **Problem**: 기존 통계학습 이론으로 설명 불가능한 현상 발생
  - Double Descent 현상은 다양한 ML/DL 모델에서 실제로 관측되는 현상
  - 고전 통계학습 이론(bias-variance trade-off)으로는 이 현상을 설명할 수 없음
  - 실제로 관찰되는 중요한 현상에 대한 이론적 공백이 발생

- **Solution**: Double Descent 현상에 대한 이론적 토대 구축
  - 수학적 분석이 용이한 선형 모델에서 Double Descent 현상을 설명하려는 시도가 우선적으로 이루어짐
  - 선형 모델 분석은 Neural Network에서의 Double Descent 이해에 중요한 발판이 됨
  - (참고) 선형 모델과 Neural Network 간 연관성을 밝힌 주요 연구:
    - [4, 5] Random Fourier Feature Regression은 Kernel Regression을 효과적으로 모사
    - [6] Neural Network의 학습 과정이 Kernel Regression과 거의 일치함

→Double Descent 현상을 선형 모델에서 이론적으로 설명하려는 시도가 이어짐 [7, 8]

※ 후술하는 연구 [9,10]에서 선형 모델의 복잡도(파라미터 수)를 조절을 위해 Random Fourier Feature를 이용

$$S_{i,t} = \left[\sin(\gamma \omega_i' G_t), \cos(\gamma \omega_i' G_t)\right]', \quad \omega_i \sim i.i.d.N(0, I),$$

G_t is original input data,
(# of ω_i drawn) controls model complexity

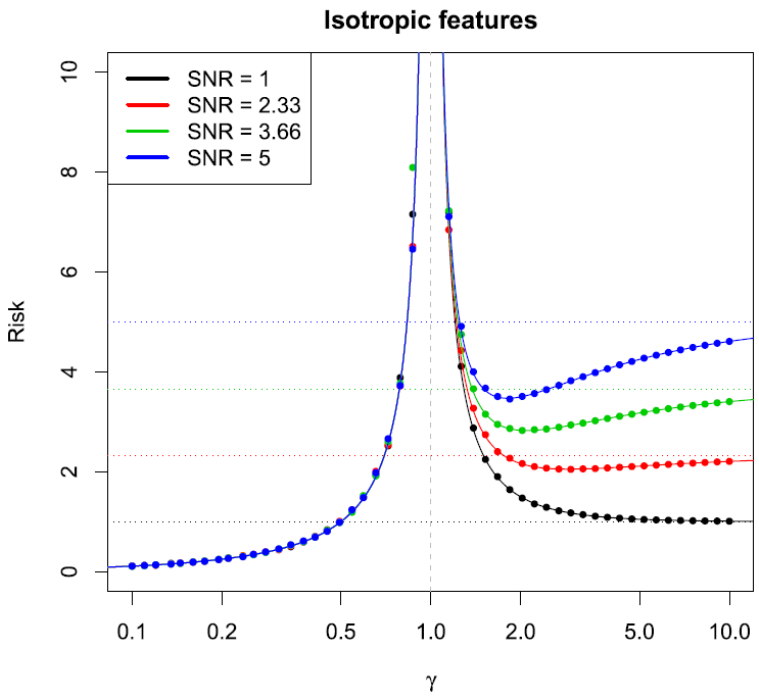
02 Double Descent: Virtue of Complexity in Statistical Learning

- Hastie et al. [9]는 이전 연구 [7,8]를 확장, well-specified/misspecified linear model에서 ridgeless solution* 의 일반화 성능을 model complexity의 함수로 표현→ Double Descent 현상에 대한 이론적 설명

well-specified linear model

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n$$

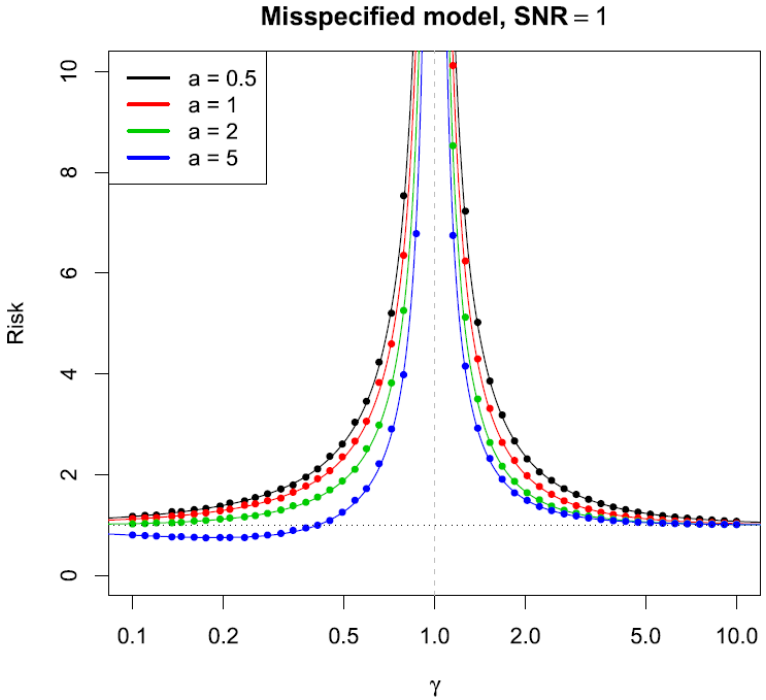
$$n, p \to \infty \quad p/n \to \gamma$$

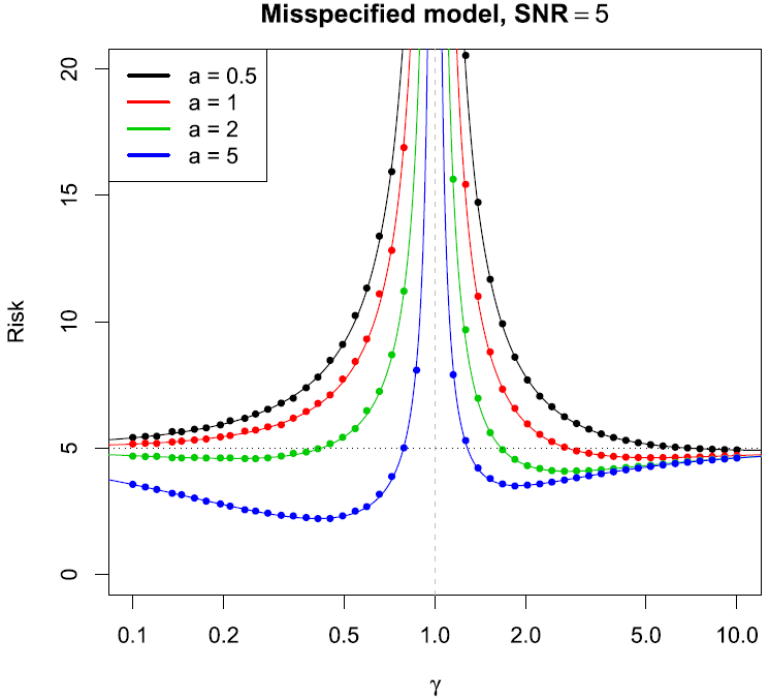


misspecified linear model

$$y_i = x_i^T \beta + w_i^T \theta + \epsilon_i, \quad i = 1, \ldots, n$$

$$n, p \to \infty \quad p/n \to \gamma$$

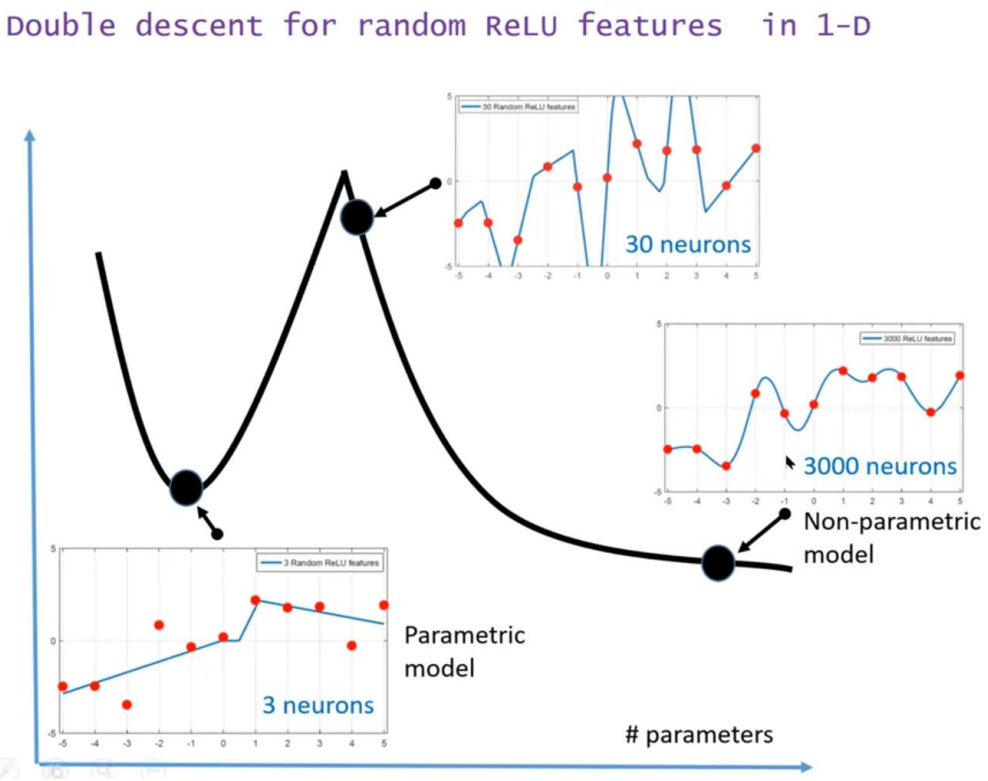




*ridgeless solution: 모든 training Sample을 fitting하면서 가장 L2-norm이 작은 solution

02 Double Descent: Virtue of Complexity in Statistical Learning

- Neural Network에서 Double Descent 현상을 직관적으로 설명은 가능하나, 이론적 해석은 여전히 불충분함



PROPOSITION 1. *Initialize* $\beta^{(0)} = 0$, *and consider running gradient descent on the least squares loss, yielding iterates*

$$\beta^{(k)} = \beta^{(k-1)} + tX^T\left(y - X\beta^{(k-1)}\right), \quad k = 1, 2, 3, \ldots,$$

where we take $0 < t \leq 1/\lambda_{\max}(X^T X)$ *(and* $\lambda_{\max}(X^T X)$ *is the largest eigenvalue of* $X^T X$*). Then* $\lim_{k\to\infty} \beta^{(k)} = \hat{\beta}$, *the min-norm least squares solution in* (6).

$$\hat{\beta} = \arg\min\{\|b\|_2 : b \text{ minimizes } \|y - Xb\|_2^2\}. \tag{6}$$

over-parameterized 모델에 대한 gradient descent 방식의 학습은
β를 min-norm least square solution(ridgeless solution)으로 수렴시킴

03 Virtue of Complexity: Application to Financial Machine Learning

- 금융시계열의 특성과 전통적인 모델링 관습
  - 금융시계열의 특성: Data Volume versus Signal-to-Noise Ratio(SNR) Trade-off
    - Data Poor Environment→ ML 모델은 data poor 환경에서는 잘 작동하지 않음
    - Weak Signal-to-Noise Ratio→ ML 기반으로 모델링하는 경우 noise까지 overfitting할 우려가 있음
  - 전통적인 금융시계열 모델링 관습(Occam's razor)
    - 복잡한 모델보다는 해석가능한 단순한 모델 선호(Parsimony principle)
    - Risk management 관점에서도 복잡한 모델은 variance가 크기때문에 비선호됨

- 금융시계열 모델링에서 복잡한(over-parameterized) 모델의 부상
  - 수익률 예측, 포트폴리오 최적화 등에서 Neural Network 기반 모델의 주목할 만한 성과가 최근 다수 보고됨
  - 복잡한 모델이 단순한 모델에 비해 예측 성능 뿐만 아니라 Economic Value 측면에서도 더 나은 성과를 보임
  - 자연스럽게 연구자들은 "금융시계열 모델링에서도 복잡한 모델이 더 좋을까?"라는 질문에 도달함

03 Virtue of Complexity: Application to Financial Machine Learning

- **Research Question**: 금융시계열 예측에서 복잡한 모델이 이론·실증적으로 더 나은가?
- Kelly et al. [10]는 위의 Research Question을 이론·실증적으로 검증하고자 함
  - **Goal**: 단일자산 수익률 예측 모델(ridge regression)의 예측 성능 및 투자 성과를 model complexity로 표현
  - well-specified model: true DGP*의 모든 signal을 포함하여 모델링한 경우(model complexity $c = P/T$)
  - misspecified model: $P_1 (< P)$개의 signal만 사용하여 모델링한 경우(model complexity= cq, where $q = P_1/P$)

true DGP model	$R_{t+1} = S_t'\beta + \varepsilon_{t+1}$ $S_t \in \mathbb{R}^P$ (signal) $\beta \in \mathbb{R}^P$ (estimator)

well-specified model	$R_{t+1} = S_t'\beta + \varepsilon_{t+1}$ $\hat{\beta}(z) = (zI + \frac{1}{T}\sum_{t=1}^{T} S_t S_t')^{-1} \frac{1}{T}\sum_{t=1}^{T} S_t R_{t+1}$
misspecified model	$R_{t+1} = (S_t^{(1)})'\beta + \varepsilon_{t+1}$, $S_t^{(1)} \in \mathbb{R}^{P_1}, P_1 < P$ $\hat{\beta}(z) = (zI + \frac{1}{T}\sum_{t=1}^{T} S_t^{(1)}(S_t^{(1)})')^{-1} \frac{1}{T}\sum_{t=1}^{T} S_t^{(1)} R_{t+1}$

*DGP: Data Generating Process

※ 투자 성과는 마켓 타이밍 전략의 수익률($R_{t+1}^{\pi_t} = S_t'\hat{\beta}(z)R_{t+1}$)기반 측정

03 Virtue of Complexity: Application to Financial Machine Learning

- **Main Results** of Kelly et al. [10]
  - Random Matrix Theory를 기반으로 수익률 예측성능(R^2)을 model complexity에 대한 함수로 표현
  - 더 나아가 마켓 타이밍 전략의 투자성과(Sharpe Ratio, i.e. SR)를 model complexity에 대한 함수로 표현

well-specified model

In the limit as $T,\ P \to \infty$, *and* $P/T \to c$, *we have*

$$\mathcal{E}(z;c) = \lim E[\hat{\pi}_t R_{t+1} | \hat{\beta}(z)] = b_* \nu(z;c),$$

$$\mathcal{L}(z;c) = \lim E[\hat{\pi}_t^2 | \hat{\beta}(z)] = b_* \hat{\nu}(z;c) - c\nu'(z;c),$$

$$R^2(z;c) = \frac{2\mathcal{E}(z;c) - \mathcal{L}(z;c)}{1 + b_* \psi_{*,1}},$$

$$\mathcal{V}(z;c) := \lim E\left[(\hat{\pi}_t(z) R_{t+1})^2 | \hat{\beta}\right] = 2(\mathcal{E}(z;c))^2 + \left(1 + b_* \psi_{*,1}\right)\mathcal{L}(z;c)$$

$$SR(z;c) = \frac{\mathcal{E}(z;c)}{\sqrt{\mathcal{V}(z;c)}} = \frac{1}{\sqrt{2 + \left(1 + b_* \psi_{*,1}\right)\frac{\mathcal{L}(z;c)}{(\mathcal{E}(z;c))^2}}}$$

misspecified model

In the limit $T, P, P_1 \to \infty,\ P/T \to c$, *and* $P_1/P \to q \in (0, 1]$

$$\mathcal{E}(z;cq;q) := \lim E[\hat{\pi}_t(z) R_{t+1} | \hat{\beta}] = b_* q \left(\nu(z;cq;q) + \frac{(cq)^{-1} \xi_{2,1}(z;cq;q)}{1 + \xi(z;cq;q)} \right)$$

$$\mathcal{L}(z;cq;q) := \lim E[\hat{\pi}_t(z)^2 | \hat{\beta}] = q \left(b_* \hat{\nu}(z;cq;q) - c(1 + b_*[\psi_{*,1}(1) - q\psi_{*,1}(q)])\nu'(z;cq;q) \right) + \Delta(z;cq;q)$$

$$R^2(z;cq;q) = \frac{2\mathcal{E}(z;cq;q) - \mathcal{L}(z;cq;q)}{1 + b_* \psi_{*,1}(1)}$$

$$\mathcal{V}(z;cq;q) := \lim E\left[(\hat{\pi}_t(z) R_{t+1})^2\right] = 2(\mathcal{E}(z;cq;q))^2 + \left(1 + b_* \psi_{*,1}\right)\mathcal{L}(z;cq;q)$$

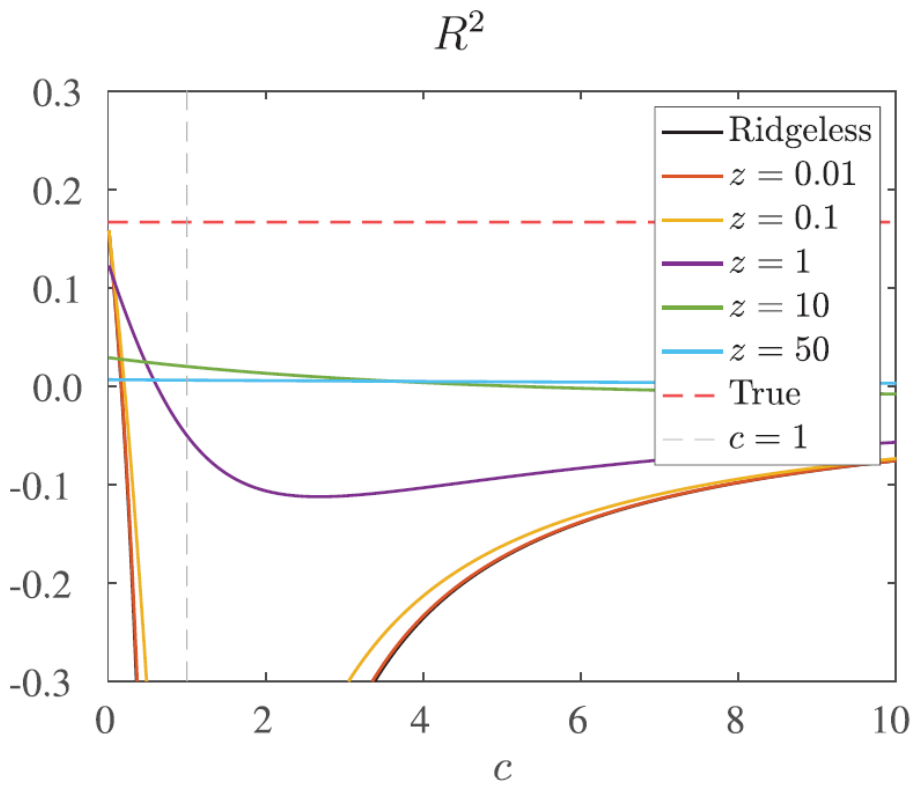
$$SR(z;cq;q) = \frac{\mathcal{E}(z;cq;q)}{\sqrt{\mathcal{V}(z;cq;q)}} = \frac{1}{\sqrt{2 + \left(1 + b_* \psi_{*,1}\right)\frac{\mathcal{L}(z;cq;q)}{(\mathcal{E}(z;cq;q))^2}}}$$

03 Virtue of Complexity: Application to Financial Machine Learning

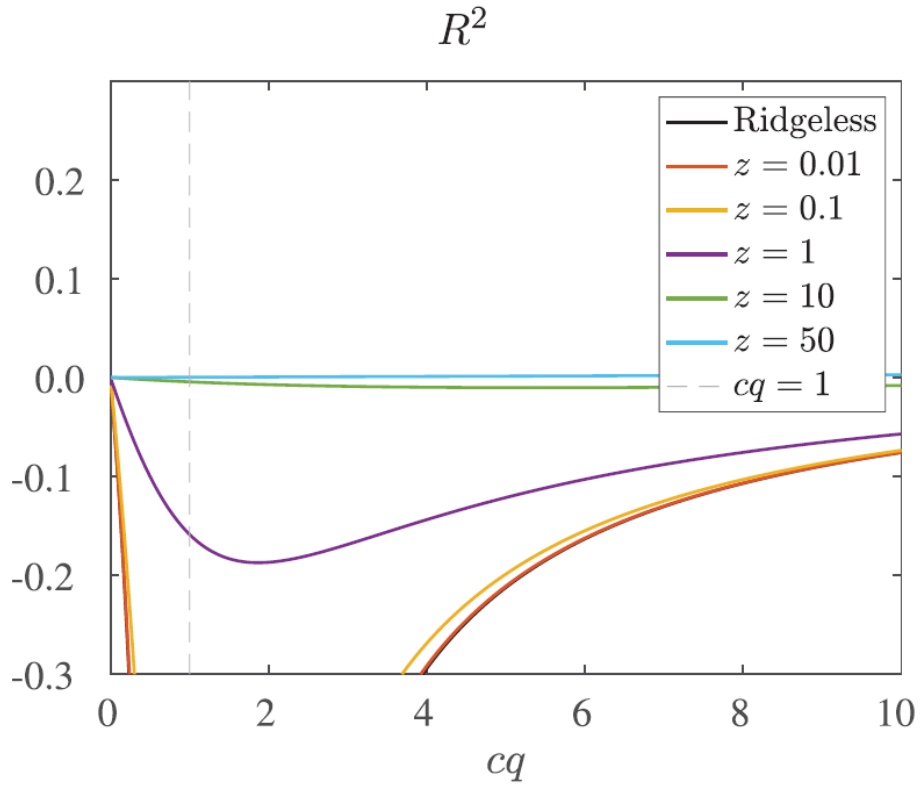
※ $b_* = 0.2, \psi$가 identity matrix인 경우

- **Theoretical Analysis** of Prediction Accuracy
  - under-parameterized model: model complexity가 증가할수록 R^2 감소하는 경향
  - over-parameterized model: model complexity가 증가할수록 R^2 증가하는 경향

well-specified model



misspecified model

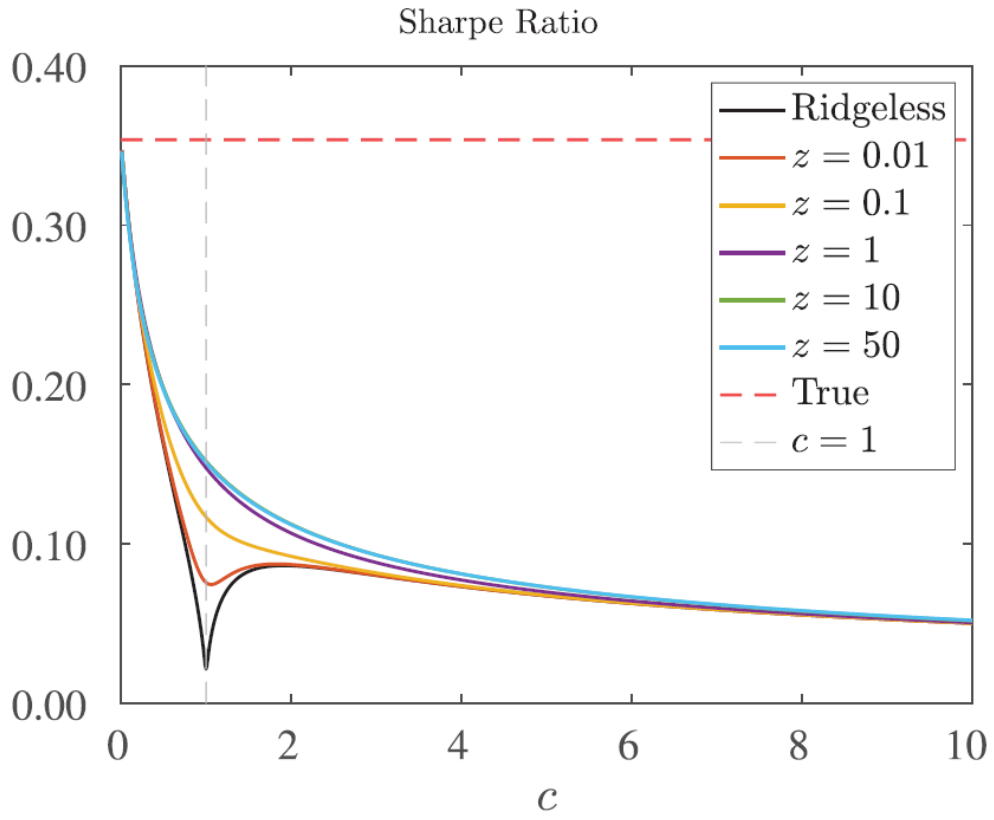


03 Virtue of Complexity: Application to Financial Machine Learning

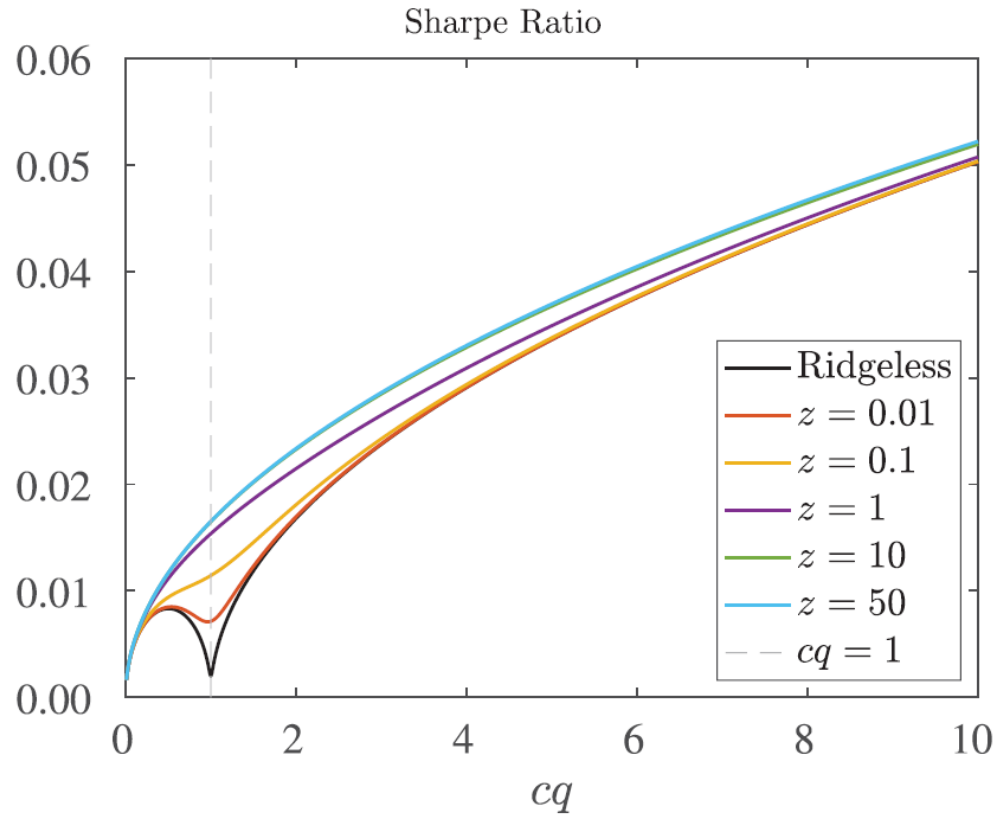
※ $b_* = 0.2, \psi$가 identity matrix인 경우
※ 투자 성과는 마켓 타이밍 전략의 수익률($R_{t+1}^{\pi_t} = S_t' \hat{\beta}(z) R_{t+1}$)기반 측정

- **Theoretical Analysis** of Investment Performance
  - well-specified model: model complexity가 증가할수록 Sharpe Ratio 감소하는 경향
  - misspecified model: model complexity가 증가할수록 Sharpe Ratio 증가하는 경향

well-specified model



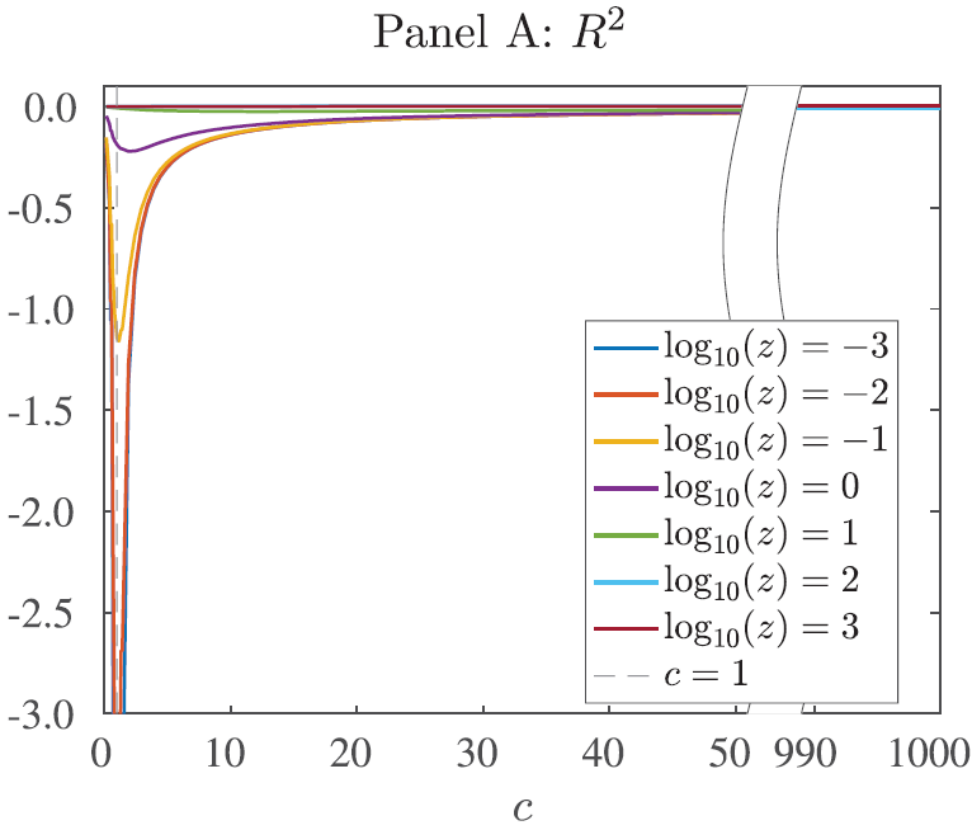
misspecified model

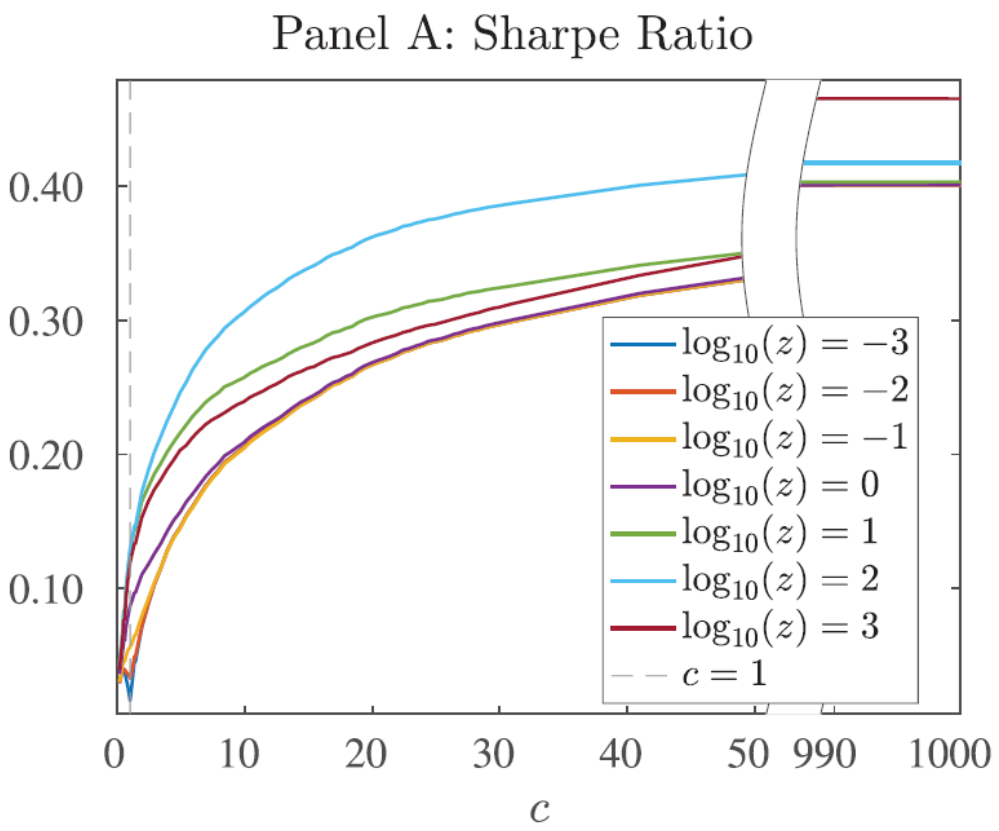


03 Virtue of Complexity: Application to Financial Machine Learning

Data Description: CRSP value-weighted index (1926–2020, monthly) with 15 predictor variables from Goyal and Welch (2008).
Random Fourier Feature was adopted to control model complexity.

- **Empirical Analysis** of the US Stock market
  - empirical R^2, SR plot은 misspecified model의 theoretical R^2, SR plot과 유사한 패턴을 보임
  - 즉, 실제 데이터에서도 model complexity를 늘릴수록 예측 성능 및 투자 성과가 좋아지는 현상이 관찰됨





03 Virtue of Complexity: Application to Financial Machine Learning

- 후속 연구 [11, 12]에서도 Financial Machine Learning에서 Virtue of Complexity 현상의 실증적 증거 발견됨

[11] Didisheim et al., 2024

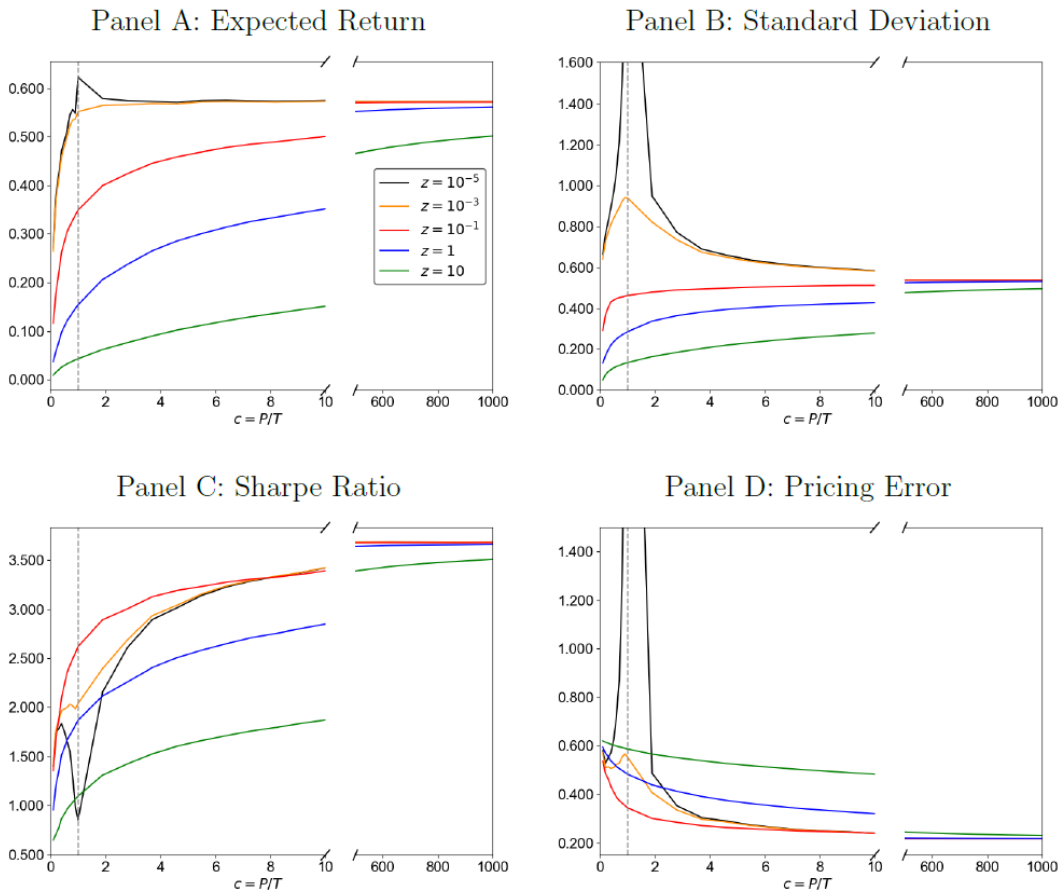


Figure 2: Out-of-sample Performance of Complex Factor Models

[12] Kelly et al., 2024

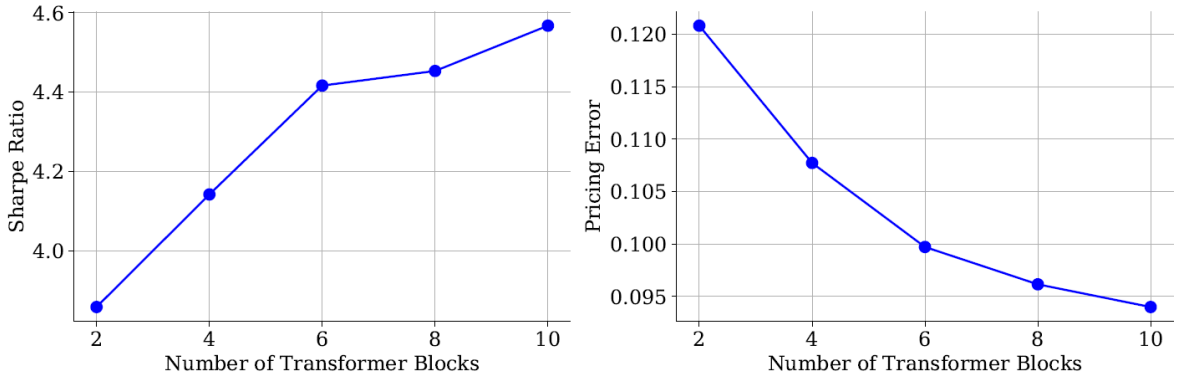


Figure 6: **Complexity and Transformer Model Performance**

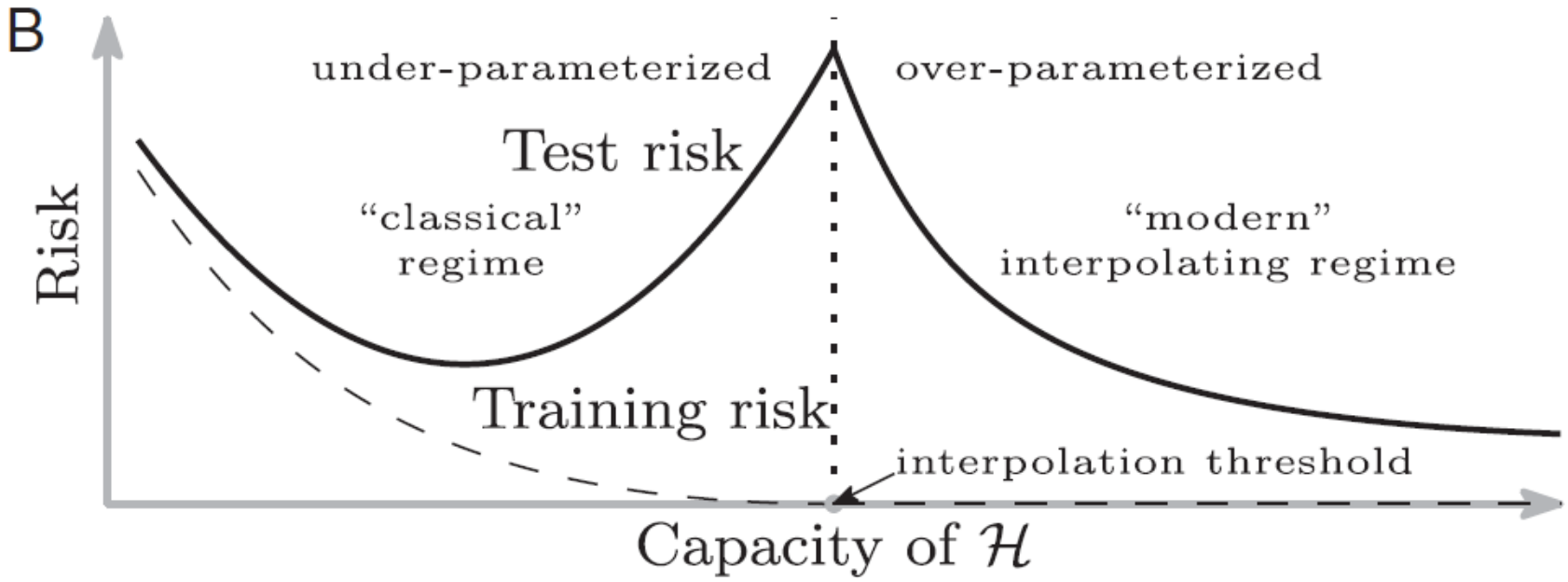
03 Virtue of Complexity: Application to Financial Machine Learning

- (참고) Kelly et al. [10]에 대한 비판 vs. Kelly and Malamud [13]의 반박

연구	Kelly et al. [10]에 대한 비판	Kelly and Malamud [13]의 반박
Buncic [14]	➢ zero-intercept 가정을 없애고, 평가지표를 집계하는 방식을 바꾸면 단순한 모델이 더 높은 SR을 달성	➢ zero-intercept 가정은 잘 알려진 모멘텀 효과에 따라서 모델이 학습되는 것을 방지하기 위한 가정, 실제로 intercept를 포함하는 모델은 모멘텀 전략과 높은 상관관계를 보임 ➢ Buncic이 제안하는 집계방식에서 단순 모델은 결국 앙상블 모델이 되고, 이는 실질적으로는 복잡 모델로 볼 수 있음 ➢ 단순한 모델은 SR은 높을지 몰라도 다른 투자 성과 지표들(Information ratio 등)은 복잡한 모델 대비 현저히 낮음
Cartea et al. [15]	➢ 예측 변수에 측정 오차가 존재하면 모델의 복잡성이 증가할수록 이론적인 SR은 감소 ➢ 실제로 예측 변수에 일정 수준 이상의 노이즈를 추가해서 실험하면 모델의 복잡성이 증가할수록 SR이 감소	➢ Cartea et al.의 이론은 Kelly et al. [10]에서 전개한 이론의 특별한 경우 ➢ 노이즈를 충분히 추가하면 어떤 모델이든 무너지게 됨 ➢ 인공적으로 노이즈를 추가하여 만든 데이터를 활용한 실험 결과는 실제 데이터에서 얻어진 실험 결과와 상반됨
Nagel [16]	➢ 복잡한 모델의 성공은 예측 변수의 지속성에 기반한 변동성 모멘텀 전략의 결과로 진정한 학습이 된 것이 아님 ➢ 인공적으로 수익률 데이터에 반전을 주면 복잡한 모델은 여전히 모멘텀 전략을 생성하여 음의 수익률 기록 ➢ 간단한 모멘텀 전략으로 복잡한 모델을 모사할 수 있음	➢ 예측 변수의 지속성과 공분산 구조는 보존하지만 실제 수익률과 무관한 가짜 예측변수를 활용하면 예측력 사라짐 ➢ 예측 대상만 변경하면 예측 변수와 예측 대상 사이의 패턴을 모델이 포착하지 못하는 것은 당연함 ➢ Nagel이 제안하는 전략은 일종의 증류(distillation)

→"Virtue of Complexity"에 대한 이론·실증적 타당성을 두고 학계에서 활발한 논쟁이 현재도 이어지는 중

04 Summary and Discussion



	Pre-Deep Learning Era (under-parameterized model)	Post-Deep Learning Era (over-parameterized model)
Statistical Learning	Bias-Variance Trade-off	Double Descent
Financial Machine Learning	Occam's Razor(Parsimony Principle)	**Virtue of Complexity(?)**

04 Summary and Discussion

- Main Theoretical Results
  - **Double Descent**: over-parameterized misspecified linear model에서, signal이 충분히 혼합되고 sample size와 모델의 parameter 수가 모두 무한히 클 때, generalization error는 모델 복잡도(=number of parameters divided by sample size)가 증가함에 따라 감소
  - **Virtue of Complexity**: over-parameterized misspecified linear model을 활용한 수익률 예측에서 동일한 현상이 R^2에 대해서도 나타나며, 해당 모델 기반 투자 전략의 Sharpe Ratio는 모델 복잡도가 증가함에 따라 상승
  - **Important Notes on Theoretical Results**: 예측 target과 무관한 feature를 무작정 추가해도 성능이 개선됨을 의미하지 않으며, 데이터의 양이 많고 품질(SNR)이 좋을수록 ML·DL 방법론의 성능이 개선되는 것은 여전히 불변의 진리

- Further Research related to "Virtue of Complexity" Theory
  - Finite Sample Size에서의 이론적 검증

    실제 금융시계열의 표본 수는 제한적이므로 data poor 환경에서 복잡한 모델의 일반화 성능에 대한 이론적 검증 필요
  - Neural Network 모델에서의 이론적 검증

    Neural Network의 넓이(width)나 깊이(depth)에 대해 복잡성의 미덕이 일관되게 성립하는지에 대한 이론적 분석 필요
  - Cross sectional 모델에서의 이론적 검증

    다양한 자산을 대상 Cross sectional 분석에서도 모델의 복잡성이 성과에 미치는 영향에 대한 이론적 연구 필요

Reference

[1] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. in Adaptive Computation and Machine Learning series. Cambridge, MA, USA: MIT Press, 2018. Accessed: Sept. 12, 2025. [Online]. Available: https://mitpress.mit.edu/9780262039406/foundations-of-machine-learning/

[2] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15849–15854, Aug. 2019, doi: 10.1073/pnas.1903070116.

[3] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep Double Descent: Where Bigger Models and More Data Hurt," presented at the International Conference on Learning Representations, Sept. 2019. Accessed: Sept. 12, 2025. [Online]. Available: https://openreview.net/forum?id=B1g5sA4twr

[4] A. Rahimi and B. Recht, "Random Features for Large-Scale Kernel Machines," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2007. Accessed: Sept. 18, 2025. [Online]. Available: https://papers.nips.cc/paper_files/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html

[5] H. Avron, M. Kapralov, C. Musco, C. Musco, A. Velingker, and A. Zandieh, "Random Fourier Features for Kernel Ridge Regression: Approximation Bounds and Statistical Guarantees," in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, July 2017, pp. 253–262. Accessed: Sept. 18, 2025. [Online]. Available: https://proceedings.mlr.press/v70/avron17a.html

[6] A. Jacot, F. Gabriel, and C. Hongler, "Neural Tangent Kernel: Convergence and Generalization in Neural Networks," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2018. Accessed: Sept. 18, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html

[7] M. Belkin, D. Hsu, and J. Xu, "Two Models of Double Descent for Weak Features," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 4, pp. 1167–1180, Jan. 2020, doi: 10.1137/20M1336072.

[8] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 117, no. 48, pp. 30063–30070, Dec. 2020, doi: 10.1073/pnas.1907378117.

[9] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *Ann. Statist.*, vol. 50, no. 2, Apr. 2022, doi: 10.1214/21-AOS2133.

[10] B. Kelly, S. Malamud, and K. Zhou, "The Virtue of Complexity in Return Prediction," *The Journal of Finance*, vol. 79, no. 1, pp. 459–503, 2024, doi: 10.1111/jofi.13298.

[11] A. Didisheim, S. Ke, B. T. Kelly, and S. Malamud, "APT or 'AIPT'? The Surprising Dominance of Large Factor Models," Mar. 13, 2023, *Social Science Research Network, Rochester, NY*: 4388526. doi: 10.2139/ssrn.4388526.

[12] B. T. Kelly, B. Kuznetsov, S. Malamud, and T. A. Xu, "Artificial Intelligence Asset Pricing Models," Jan. 06, 2025, *Social Science Research Network, Rochester, NY*: 5089371. doi: 10.2139/ssrn.5089371.

[13] B. T. Kelly and S. Malamud, "Understanding The Virtue of Complexity," July 01, 2025, *Social Science Research Network, Rochester, NY*: 5346842. doi: 10.2139/ssrn.5346842.

[14] D. Buncic, "Simplified: A Closer Look at the Virtue of Complexity in Return Prediction," Apr. 30, 2025, *Social Science Research Network, Rochester, NY*: 5239006. doi: 10.2139/ssrn.5239006.

[15] Á. Cartea, Q. Jin, and Y. Shi, "The Limited Virtue of Complexity in a Noisy World," Apr. 02, 2025, *Social Science Research Network, Rochester, NY*: 5202064. doi: 10.2139/ssrn.5202064.

[16] S. Nagel, "Seemingly Virtuous Complexity in Return Prediction," June 20, 2025, *Social Science Research Network, Rochester, NY*: 5335012. doi: 10.2139/ssrn.5335012.

Appendix

Belkin et al., 2020

$$\mathbb{E}[(y - \boldsymbol{x}^* \hat{\boldsymbol{\beta}})^2] = \begin{cases} (\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2) \cdot \left(1 + \frac{p}{n-p-1}\right) & \text{if } p \leq n-2; \\ +\infty & \text{if } n-1 \leq p \leq n+1; \\ \|\boldsymbol{\beta}_T\|^2 \cdot \left(1 - \frac{n}{p}\right) + (\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2) \cdot \left(1 + \frac{n}{p-n-1}\right) & \text{if } p \geq n+2. \end{cases}$$

Bartlett et al., 2020

$$R(\hat{\theta}) \leq c \left(\|\theta^*\|^2 \|\Sigma\| \max \left\{ \sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}} \right\} \right) + c \log(1/\delta) \sigma_y^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right)$$

Appendix

Hastie et al., 2022

PROPOSITION 1. *Initialize* $\beta^{(0)} = 0$, *and consider running gradient descent on the least squares loss, yielding iterates*

$$\beta^{(k)} = \beta^{(k-1)} + tX^T\left(y - X\beta^{(k-1)}\right), \quad k = 1, 2, 3, \ldots,$$

where we take $0 < t \leq 1/\lambda_{\max}(X^T X)$ (*and* $\lambda_{\max}(X^T X)$ *is the largest eigenvalue of* $X^T X$). *Then* $\lim_{k\to\infty} \beta^{(k)} = \hat{\beta}$, *the min-norm least squares solution in* (6).

well-specified models

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n$$

$$\mathbb{E}(x_i) = 0, \mathrm{Cov}(x_i) = \Sigma$$

$$\mathbb{E}(\epsilon_i) = 0, \mathrm{Var}(\epsilon_i) = \sigma^2$$

$$p/n \to \gamma \quad \|\beta\|_2^2 = r^2 \quad \mathrm{SNR} = \|\beta\|_2^2/\sigma^2$$

$$R_X(\hat{\beta}; \beta) \to \begin{cases} \sigma^2 \dfrac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2\left(1 - \dfrac{1}{\gamma}\right) + \sigma^2 \dfrac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases}$$

misspecified models

$$y_i = x_i^T \beta + w_i^T \theta + \epsilon_i, \quad i = 1, \ldots, n$$

$$\mathrm{Cov}((x_i, w_i)) = \Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xw} \\ \Sigma_{xw}^T & \Sigma_w \end{bmatrix}$$

$$p/n \to \gamma \quad r^2 = \|\beta\|_2^2 + \|\theta\|_2^2 \quad \kappa = \|\beta\|_2^2/r^2$$

$$R_X(\hat{\beta}; \beta, \theta) \to \begin{cases} r^2(1-\kappa) + \left(r^2(1-\kappa) + \sigma^2\right)\dfrac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2(1-\kappa) + r^2\kappa\left(1 - \dfrac{1}{\gamma}\right) + \left(r^2(1-\kappa) + \sigma^2\right)\dfrac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases}$$

Appendix

Kelly et al., 2024

ASSUMPTION 2: *There exist independent random vectors $X_t \in \mathbb{R}^P$ with four finite first moments, and a symmetric, P-dimensional positive semidefinite matrix Ψ such that*

$$S_t = \Psi^{1/2} X_t.$$

$$\Psi = \begin{pmatrix} \Psi_{1,1} & \Psi_{1,2} \\ \Psi_{1,2}' & \Psi_{2,2} \end{pmatrix} \quad \begin{matrix} \pi_t(\beta) = S_t'\beta \\ R_{t+1}^{\pi} = \pi_t R_{t+1} \end{matrix}$$

THEOREM 1 (Virtue of Complexity): *Suppose that signals are sufficiently mixed (so that $H(x;q)$ does not depend on q) and $\mathrm{tr}(\Psi_{1,2}\Psi_{2,1}) = o(P)$. Then, with the optimal amount of shrinkage z_*, the Sharpe ratio $SR(z_*(q;c);cq;q)$ and $R^2(z_*(q;c);cq;q)$ are strictly monotone increasing and concave in $q \in [0,1]$.*

$$S_{i,t} = \left[\sin(\gamma \omega_i' G_t), \cos(\gamma \omega_i' G_t)\right]', \quad \omega_i \sim i.i.d. N(0, I),$$

Appendix

Kelly and Malamud, 2025

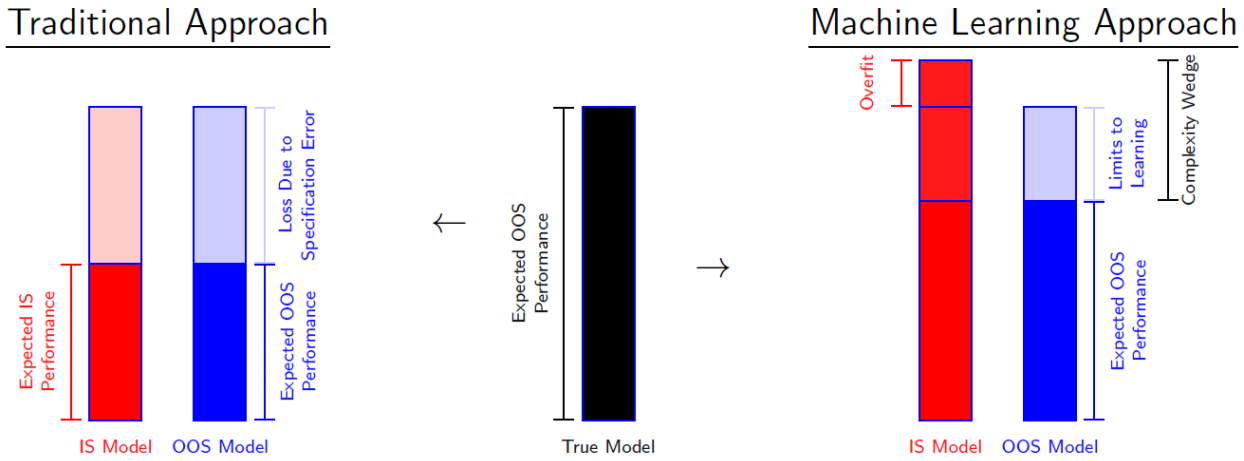


Figure 2: Limits to Learning, Overfit, and the Complexity Wedge

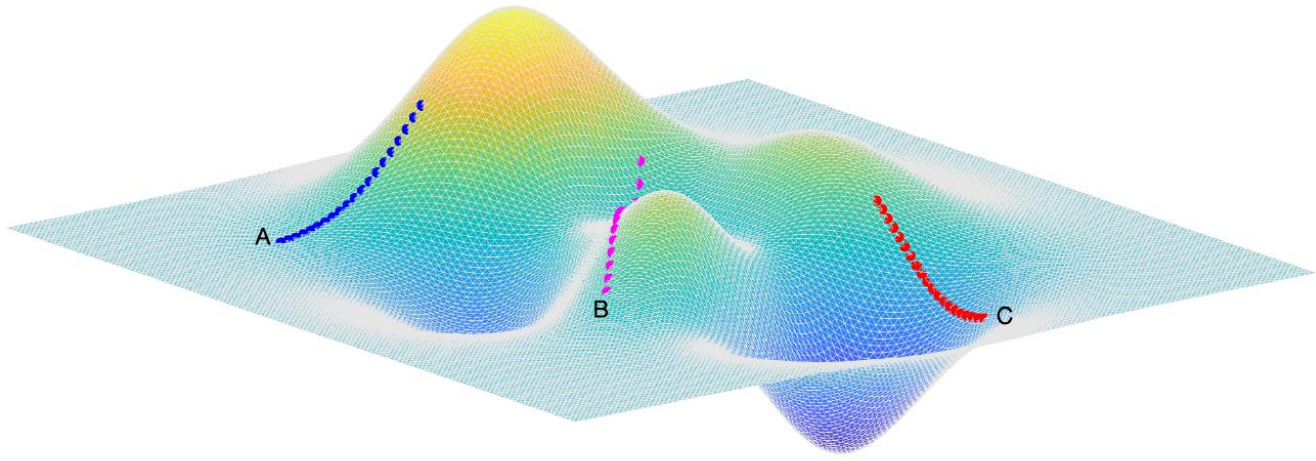


Figure 16: Limits To Learning and Ensembles