Foundations of Big Data and Machine Learning in Finance, Statistics, and Beyond. Lecture Series

Semyon Malamud
SFI, EPFL, and CEPR

1 Learning Outcomes

- Modeling with High-Dimensional Predictors: Understand modern methods for large-scale data and factor models in finance.
- Overfitting and Complexity: Recognize the role of over-parameterized models in predictive performance, portfolio choice, and decision making.
- Neural Networks: Interpret deep vs. shallow learning architectures and their application to financial data. Deep Learning and Feature Learning.
- Factor Models: Construct and analyze high-dimensional factor models for the cross-section of returns.
- Equilibrium: Assess how complexity corrections change asset pricing fundamentals and how complexity leads to tractable equilibria with nonlinear parameter learning.

2 Lecture Breakdown

Lecture 1: Overfitting, Double Descent, Model Complexity, and Inductive Biases

We discuss the evolution of model design in machine learning over the last decades, the discovery of double descent, and scaling laws. We then demonstrate that similar results hold in the realm of finance: Bigger (more complex) models perform better out-of-sample in terms of their Sharpe Ratios. We then discuss the key regularization property of over-parameterized (more parameters than observations) models and their inductive biases.
Key References:

- Belkin, Mikhail, Daniel Hsu, Siyuan Ma, and Soumik Mandal. ”Reconciling modern machine-learning practice and the classical bias–variance trade-off.” Proceedings of the National Academy of Sciences 116, no. 32 (2019): 15849-15854.
- Nakkiran, Preetum, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. ”Deep double descent: Where bigger models and more data hurt.” Journal of Statistical Mechanics: Theory and Experiment 2021, no. 12 (2021): 124003.
- Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. ”Scaling laws for neural language models.” arXiv preprint arXiv:2001.08361 (2020).
- Malamud, Semyon, Kelly, Bryan T., & Zhou, Kanying (2024). “The Virtue of Complexity in Return Prediction.” *Journal of Finance*, 79(1), 459-503.

Lecture 2: Regularization, Model Selection, Sparsity, Non-Linearities, and Random Features

This lecture covers techniques to control model complexity and avoid overfitting in high-dimensional settings through forms of penalization (shrinkage). We discuss how linear algebra magic can be used to find optimal shrinkage. While imposing forms of sparsity can help (Freyberger, Neuhierl, and Weber, 2020), this can be problematic in high dimensions (Xiu and Shen, 2025).
We then introduce the simplest form of non-linearities- random features- and show how this method can be extremely powerful but may also fail trastically in high dimensions due to the curse of dimensionality and the *inability of random feature methods to perform feature learning.*
Key References:

- Kelly, Bryan T., Semyon Malamud, Mohammad Pourmohammadi, and Fabio Trojani. Universal portfolio shrinkage. No. w32004. National Bureau of Economic Research, 2024.
- Gu, Shihao, Kelly, Bryan T., & Xiu, Dacheng. (2020). “Empirical Asset Pricing via Machine Learning.” *Review of Financial Studies*, 33(5), 2223–2273.
- Xiu, Dacheng, and Zhouyu Shen. (2025). ”Can Machines Learn Weak Signals?”, working paper.
- Malamud, Semyon, Kelly, Bryan, & Zhou, Kanying (2024). “The Virtue of Complexity in Return Prediction.” *Journal of Finance*, 79(1), 459-503.

Lecture 3: Implicit Regularization, The Virtue of Complexity, The Magic of High Dimensions. Basics of Random Matrix Theory

In this lecture, we highlight empirical findings that greater model complexity can be beneficial for predicting returns. Contrary to conventional wisdom, increasing model parameters (even beyond the number of observations) can raise out-of-sample performance — the "virtue of complexity." We examine the theoretical justification for this and review evidence that high-complexity ML models substantially outperform simpler models in forecasting tasks. **Key References:**

- Malamud, Semyon, Kelly, Bryan, & Zhou, Kanying (2024). "The Virtue of Complexity in Return Prediction." *Journal of Finance*, 79(1), 459-503.
- Lettau, Martin, & Pelger, Markus. (2020). Factors that fit the time series and cross-section of stock returns. The Review of Financial Studies, 33(5), 2274-2325.
- Onatski, Alexei. "Testing hypotheses about the number of factors in large factor models." Econometrica 77.5 (2009): 1447-1479.
- Onatski, Alexei, and Chen Wang. "Alternative asymptotics for cointegration tests in large VARs." Econometrica 86.4 (2018): 1465-1478.
- Onatski, Alexei, and Chen Wang. "Spurious factor analysis." Econometrica 89.2 (2021): 591-614.

Lecture 4: Kernel Methods, Shallow Learning, Curse of Dimensionality

We introduce and discuss kernel methods and their key role in understanding over-parametrization and generalization properties of Machine Learning models. We discuss the surprising link between kernel methods and shallow neural networks and introduce the surprising "Plato's cave" result, where each machine learning model in high dimensions, instead of recovering the ground truth, can only recover its "shadow." This naturally leads us to talk about the alignment between a model and the data and how to characterize it. **Key References:**

- El Karoui, Noureddine. "The spectrum of kernel random matrices." (2010): Annals of Statistics 38(1): 1-50
- Misiakiewicz, Theodor. "Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression." arXiv preprint arXiv:2204.10425 (2022).

- Mei, Song, Theodor Misiakiewicz, and Andrea Montanari. "Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration." Applied and Computational Harmonic Analysis 59 (2022): 3-84.

Lecture 5: Deep vs. Shallow Learning; Neural Tangent Kernel; Feature Learning

We discuss a striking connection between kernel methods and deep learning. We then discuss how to train neural networks away from kernel regimes and how feature learning emerges in neural nets. We then discuss the implications of feature learning for asset pricing and how transformers perform feature learning in Large Language Models and for Predicting Returns.
Key References:

- Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." Advances in neural information processing systems 31 (2018).
- Radhakrishnan, Adityanarayanan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. "Mechanism for feature learning in neural networks and backpropagation-free machine learning models." Science 383, no. 6690 (2024): 1461-1467.
- Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. (2018). "Human decisions and machine predictions." *The Quarterly Journal of Economics*, 133(1), 237-293.
- Mullainathan, S., & Spiess, J. (2017). "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives*, 31(2), 87–106.
- Kelly, Bryan T., Boris Kuznetsov, Semyon Malamud, and Teng Andrea Xu. Artificial Intelligence Asset Pricing Models. No. w33351. National Bureau of Economic Research, 2025.

Lecture 6: High-Dimensional Factor Models, Portfolio Tangent Kernel, and the Complexity Wedge

This lecture applies big data concepts to the cross-section of asset returns via factor models. We discuss the tension between traditional unconditional factor models (with a small number of static factors) and conditional approaches that incorporate more information (time-varying or state-dependent factors). We discuss the limitations of dimensionality reduction techniques in high-dimensional settings (Lettau and Pelger, 2020) and the total collapse of APT-style arguments when the data is sufficiently complex (Didisheim et al., 2024).

Finally, we introduce the Portfolio Tangent Kernel (an analog of the Neural Tangent Kernel for portfolio optimization problems) and show how it can be used to derive deep insights about almost any machine learning model. We then focus on models with cross-predictability (Kelly et al., 2023) and the role of the attention mechanism for learning across stocks (Kelly et al., 2025). **Key References:**

- Feng, Guanhao, Stefano Giglio, and Dacheng Xi. (2020). Taming the Factor Zoo: A Test of New Factors. *Journal of Finance,* 75(3), 1327-1370.
- Bryzgalova, Svetlana, Victor DeMiguel, Sicong Li, and Markus Pelger. ”Asset-pricing factors with economic targets.” Available at SSRN 4344837 (2023).
- Chernov, Mikhail, Bryan T Kelly, Semyon Malamud, and Johannes Schwab, “A Test of the Efficiency of a Given Portfolio in High Dimensions,” Swiss Finance Institute Research Paper, 2025, (25-26)
- Lettau, M., & Pelger, M. (2020). Factors that fit the time series and cross-section of stock returns. The Review of Financial Studies, 33(5), 2274-2325.
- Onatski, Alexei. ”Testing hypotheses about the number of factors in large factor models.” Econometrica 77.5 (2009): 1447-1479.
- Onatski, Alexei, and Chen Wang. ”Alternative asymptotics for cointegration tests in large VARs.” Econometrica 86.4 (2018): 1465-1478.
- Onatski, Alexei, and Chen Wang. ”Spurious factor analysis.” Econometrica 89.2 (2021): 591-614.
- Didisheim, Antoine, Shikun Barry Ke, Bryan T. Kelly, and Semyon Malamud. APT or “AIPT”? the surprising dominance of large factor models. No. w33012. National Bureau of Economic Research, 2024.
- Kelly, Bryan, Boris Kuznetsov, Semyon Malamud, and Teng Andrea Xu. ”Large (and deep) factor models.” arXiv preprint arXiv:2402.06635 (2024).
- Didisheim, Antoine, Shikun Barry Ke, Bryan T. Kelly, and Semyon Malamud. APT or “AIPT”? the surprising dominance of large factor models. No. w33012. National Bureau of Economic Research, 2024.
- Kelly, Bryan T., Boris Kuznetsov, Semyon Malamud, and Teng Andrea Xu. Artificial Intelligence Asset Pricing Models. No. w33351. National Bureau of Economic Research, 2025.
- Kelly, Bryan, Semyon Malamud, and Lasse Heje Pedersen. ”Principal portfolios.” The Journal of Finance 78, no. 1 (2023): 347-387.

Lecture 7: Bayesian Learning, Gaussian Processes and Equilibrium Models in High Dimensions

In this closing lecture, we show how the techniques developed in the previous lectures can be used to study equilibria where agents use complex models. We show how Bayesian learning becomes tractable and exhibits striking hidden order in high dimensions. We then show how embedding Bayesian agents in standard equilibrium models allows us to study parameter learning in environments previously considered intractable. We also show how one can use these methods in non-parametric learning and establish surprising connections with Gaussian Processes. We discuss the surprising phenomena occurring in models where agents (like real-world humans) try to interpolate based on past observations.

- Farmer, Leland E, Emi Nakamura, and Jon Steinsson, “Learning about the long run,” Journal of Political Economy, 2024, 132 (10), 3334–3377.
- Moll, Benjamin, “The Trouble with Rational Expectations in Heterogeneous Agent Mod- els: A Challenge for Macroeconomics,” London School of Economics, mimeo, available at https://benjaminmoll.com, 2024.
- Molavi, Pooya, Alireza Tahbaz-Salehi, and Andrea Vedolin. ”Model complexity, expectations, and asset prices.” Review of Economic Studies 91, no. 4 (2024): 2462-2507.

Prerequisites

Basic probability and linear algebra. Some Python skills would also be useful, as we will be working with Jupyter Notebooks.

Contact Information

Professor: Semyon Malamud

Email: semyon.malamud@epfl.ch