IMPERIAL

IMPERIAL COLLEGE LONDON

DEPARTMENT OF MATHEMATICS

Random features based asset pricing models

Author: Matias Data

A thesis submitted for the degree of

MSc in Mathematics and Finance, 2023-2024

Declaration

The work contained in this thesis is my own work unless otherwise stated.

Acknowledgements

First and foremost, I extend my deepest gratitude to my supervisor, Lukas Gonon, for his invaluable guidance and support during the development of this thesis.
I am also immensely grateful to the friends I made over the past year: Lorenzo, Meghal, and Edward (a.k.a. the Toucans). Your help and company have been greatly appreciated.
Lastly, I would like to express my heartfelt thanks to my girlfriend and my parents for their unwavering support and encouragement.

Abstract

This thesis investigates the use of random features, a machine learning technique, to produce asset pricing models in high-dimensional settings. Traditional models like CAPM and Fama-French struggle with the "factor zoo", a vast number of potential explanatory variables that challenge conventional statistical methods. Random features simplify model training while maintaining strong theoretical foundations, offering a new way to estimate stochastic discount factors (SDFs) effectively. This research develops a simple approach to SDF estimation, providing theoretical guarantees and empirical evidence of its effectiveness on historical data. The findings suggest that random features methods can deliver performance on par with more sophisticated and computationally expensive approaches, offering promising new asset pricing models.

Contents

List of Figures

List of Tables

Introduction

Asset pricing theory tries to answer the fundamental question of finding a price for an uncertain cashflow. Empirical asset pricing seeks to understand how securities (e.g. stocks, bonds, etc.) are valued and what factors drive their expected returns. The theory is based on the fundamental notion of no arbitrage, which implies that expected returns vary due to different exposures (i.e., betas) to the stochastic discount factor (SDF).

The holy grail in asset pricing has been to estimate a stochastic discount factor that can account for the expected returns of all assets. Practically, this involves overcoming several problems. The SDF could inherently rely on all available information (past and present), implying that it is a function of a potentially vast number of variables. The number of potential explanatory variables is indeed large and is known as the "factor zoo" (Cochrane 2011). Moreover, the excess return of individual stocks has a low signal-to-noise ratio, i.e. the mean is much smaller than its variance, thus making the estimation process much harder.

Traditional models, such as the Capital Asset Pricing Model (CAPM) and multi-factor models like Fama-French three and five factor models (Fama and French 1993, Fama and French 2015), have long dominated the field, relying on a relatively small number of observable factors to explain the cross-section of returns. However, these models often struggle to fully capture the complexities of financial markets, particularly as the number of potential explanatory variables, like stock characteristics, grows. Forty years of research have produced a "factor zoo" of hundreds of characteristic-based factors, also known as anomalies, which is any strategy that generates a significant positive alpha relative to the Fama-French three factor model (Novy-Marx and Velikov 2015). Cochrane (2011) calls this problem "The Multidimensional Challenge".

Conventional statistical methods, such as ordinary least squares (OLS), are often inadequate in this high-dimensional context, where the number of observations is typically much smaller than the number of explanatory variables. In practice, it is common to have only a few hundred observations for monthly stock characteristics and returns, while considering

more than ten thousand potential factors. This imbalance poses significant difficulties for traditional modelling approaches, as these methods tend to overfit noise, resulting in an excellent in-sample fit but poor out-of-sample forecasts.

Amid the recent surge in popularity of machine learning, driven by advancements in computing power and data availability, a new body of literature has appeared that applies these methods to tackle this high dimensional problem, including shrinkage or regularization methods (e.g. ridge, lasso and bayesian regression, Kozak et al. 2020), PCA (Lettau and Pelger 2020), random forests (Bryzgalova et al. 2019), kernel methods (Kozak 2020), random features (Didisheim et al. 2023), deep learning (Gu et al. 2020,Kelly et al. 2024), or even generative adversarial networks (GANs, Chen et al. 2024). This integration of machine learning approaches represents a significant shift in the field of asset pricing, enabling researchers to enhance the predictive power of asset pricing models.

In this thesis, we formulate the problem using the machine learning technique of random features. Random features, also known as random feature neural networks, were introduced by Rahimi and Recht (2007). They can be understood as feedforward neural networks with a single hidden layer in which the hidden weights are fixed randomly (i.e., they are not trained) and only the output weights are trained. This drastically simplifies training, as it typically results in a convex problem that often has closed-form solutions. In contrast, feedforward neural networks (FFNs) must be trained using variants of stochastic gradient descent (SGD), which introduces optimization errors that are challenging to analyze mathematically.

Random features are also closely related to kernel methods; for each activation function and weight distribution, one can associate a positive definite kernel. This connection is one of the main reasons the method has exhibited both theoretical and empirical robustness. As demonstrated by Rahimi and Recht (2008a) and Rahimi and Recht (2008b), and reviewed in Chapter 2, this method has theoretical guarantees that show it can approximate functions within a dense subspace of its associated reproducing kernel Hilbert space (RKHS) arbitrarily well. Moreover, recent results from Gonon et al. (2023) and Gonon (2023) establish universal approximation results for random features under smoothness hypothesis of the target function.

Our approach is grounded in the work of Kozak (2020) and Didisheim et al. (2023) which are closely related and are reviewed in Chapter 1. Kozak (2020) extends the ridge regression approach from Kozak et al. (2020) by introducing kernel methods to handle interactions and non-linear features derived from basic characteristics. Didisheim et al. (2023) has a close

formulation of the problem and introduce random features and its associated kernel. Our approach is similar to that of Didisheim et al. (2023), but it is considerably simpler. We do not re-rank random features as they do, and we evaluate different activation functions and random weights. Moreover, the analysis of their algorithm is complex and is based on random matrix theory. In comparison, our theory is based on Rademacher complexity bounds and the approximation results from Rahimi and Recht (2008a), Rahimi and Recht (2008b), and Gonon et al. (2023). The main results of this thesis include approximation, generalization, and learning bounds for our random features based SDFs, as detailed in Chapter 3. These results are derived under mild and relatively general assumptions, and are supported empirically by our findings in Chapter 4.

This thesis is structured as follows: Chapter 1 begins with a review of key literature on asset pricing and stochastic discount factors, then introduces characteristic-based SDFs and discusses penalised estimators obtained from different approaches; in Chapter 2 a review of statistical learning theory is done, focusing on Rademacher complexity bounds and introducing the main results on random features; Chapter 3 contains approximation and generalization bounds for random features based SDFs; finally in Chapter 4, the empirical analysis applies these methods to historical data, evaluating model performance.

Chapter 1

Characteristic-based SDFs

1.1 Stochastic discount factors

Consider a market with assets (e.g. stocks) indexed by $i = 1,\dots,N_t$. Trade occurs at discrete times $t = 1,\dots,T$. Let $R^e_{t+1,i}$ denote the excess return of asset i at time $t+1$, i.e. $R^e_{t+1,i} = R_{t+1,i} - R^f_{t+1}$ where R^f_{t+1} denotes the risk-free return. Denote by $(\Omega,\Sigma,\mathbb{P},(\mathscr{F}_t)_t)$ a filtered probability space where these random variables live and $\mathscr{F}_t$ represents all information available up to time t.

Definition 1.1.1. A stochastic discount factor (SDF) $(M_t)_t$ is an adapted process which satisfies that

$$\mathbb{E}_t[M_{t+1}R^e_{t+1,i}] = 0, \tag{1.1.1}$$

for all tradable assets $i = 1,\dots,N_t$.

A tradable SDF is an SDF which is a tradable payoff. Since condition 1.1.1 does not determines the mean of M_{t+1}, consider an SDF of the form

$$M_{t+1} = 1 - b_t^\top (R^e_{t+1} - \mathbb{E}_t[R^e_{t+1}]),$$

where R^e_{t+1} denotes the $N_t \times 1$ vector of excess returns and $b_t \in m\mathscr{F}_t$ (i.e., these are known at time t). Then, it follows that

$$\begin{aligned} 0 = \mathbb{E}_t[M_{t+1}R^e_{t+1}] = \mathbb{E}_t[R^e_{t+1}(1 - b_t^\top (R^e_{t+1} - \mathbb{E}_t[R^e_{t+1}])^\top] \\ = \mathbb{E}_t[R^e_{t+1}] - \mathbb{E}_t[R^e_{t+1}(R^e_{t+1} - \mathbb{E}_t[R^e_{t+1}])^\top]b_t = \mathbb{E}_t[R^e_{t+1}] - \mathbb{V}_t[R^e_{t+1}]b_t, \end{aligned}$$

thus $b_t = \mathbb{V}_t[R^e_{t+1}]^{-1}\mathbb{E}_t[R^e_{t+1}]$. Hence we obtain that the SDF coefficients are also the weights of a mean-variance efficient portfolio.

Suppose at each time we observe d asset characteristic-based instruments, that we capture in a matrix $Z_t \in m\mathscr{F}_t$ of dimensions $N_t \times d$. Given a feature map $\Phi : \mathbb{R}^d \to \mathbb{R}^L$ (where usually $L \gg d$), we can consider the features $X_t = \Phi(Z_t)$ which is a matrix of dimensions $N_t \times L$. Associated with these features we have L factor portfolios $F_{t+1} = X_t^\top R^e_{t+1} = \Phi(Z_t)^\top R^e_{t+1}$. More precisely,

$$F_{t+1,\ell} = \sum_{i=1}^{N_t} \Phi(Z_{t,i})_\ell R^e_{t+1,i},$$

for $\ell = 1, \dots, L$, i.e. in the factor F_ℓ we weight the i-th asset at time $t+1$ according to its ℓ-th feature $\Phi(Z_{t,i})_\ell$ at time t. These factors are by definition managed portfolios, i.e. they are tradable portfolios as their weights are known at time t (i.e., $\mathscr{F}_t$-measurable).

The definition of the specific characteristic-based instruments will be deferred until Chapter 4. For the moment, think of any characteristic that might have some explanatory power for excess returns of stocks, for example the size (i.e., market capitalization, defined as price times shares outstanding) or the dividend yield (i.e., dividend as a fraction of price). Then, stocks are ranked with respect to such characteristic to define the instruments. There are hundreds of such characteristics available (see Novy-Marx and Velikov 2015 and Jensen et al. 2023). Similarly, the specific choice of the feature map $\Phi : \mathbb{R}^d \to \mathbb{R}^L$ will be addressed later, but it will be based on random features. Consequently, the number of features L could be as large as our computational resources and memory capacity permit.

The SDF definition implies infinitely many unconditional moment conditions, i.e.

$$\mathbb{E}[M_{t+1} R^e_{t+1,i} Y_t] = 0,$$

for all bounded $\mathscr{F}_t$-measurable random variable Y_t and all assets $i = 1, \dots, N_t$. In particular, if we consider as test functions our features $\Phi(Z_t)$, then

$$\mathbb{E}[M_{t+1} F_{t+1,\ell}] = \sum_{i=1}^{N_t} \mathbb{E}[M_{t+1} R^e_{t+1,i} \Phi(Z_{t,i})_\ell] = 0,$$

thus $\mathbb{E}[M_{t+1} F_{t+1}] = 0$. If our feature space is sufficiently ample, we might be able to find an SDF that lives in the linear span of the characteristic-based factors.

Definition 1.1.2. (Cochrane 2009, Section 8.3) An *unconditional (or fixed weight) linear asset pricing model* asks for an SDF $M_{t+1} = 1 - b^\top (F_{t+1} - \mathbb{E}[F_{t+1}])$ for $b \in \mathbb{R}^L$ (fixed) such that $\mathbb{E}[M_{t+1} F_{t+1}] = 0$.

By far the most famous example of an asset pricing model is the Capital Asset Pricing Model (CAPM) of Sharpe (1964)), where there is just one factor, the "market", and where the

weights of each asset are proportional to their market capitalization. Other famous models include Fama-French three (Fama and French 1993) and five factor (Fama and French 2015) models. The three factor models adds two factors, SMB (Small Minus Big, based on the size) and HML (High Minus Low, based on value). The five factor model adds the factor RMW (Robust Minus Weak, based on profitability) and CMA (Conservative Minus Aggressive). See Kenneth French data library for details on how the factors are constructed [1].

Traditional factors were usually expressed in terms of expected-return beta representations rather than in the language of stochastic discount factors. See Cochrane (2009, Section 6.1) for equivalence results between stochastic discount factors, expected-return beta representations and mean variance efficient portfolios.

As an alternative to 1.1.2, some authors ask for the SDF to be written as $M_{t+1} = 1 - b^\top F_{t+1}$. This parametrisation is not exactly equivalent as the weights are not the same and moreover it does not have mean one, however one can work with either in practice. Thus, we might express the SDF as follows,

$$M_{t+1} = 1 - (\Phi(Z_t)b)^\top R^e_{t+1} = 1 - \sum_{i=1}^{N_t} \omega_{t,i} R^e_{t+1,i},$$

where $\omega_t \in \mathbb{R}^{N_t}$ is the vector of asset-specific weights given by $\omega_{t,i} = \sum_{i=1}^{N_t} \Phi(Z_{t,i})_\ell b_\ell$. Hence, our main assumption here is that the weights ω_t of the SDF are a linear function of the (derived) features. However, the features themselves might be highly non-linear functions of the basic instruments. It turns out one can postulate an SDF of the following form,

$$M_{t+1} = 1 - \sum_{i=1}^{N_t} \omega(Z_{t,i}) R^e_{t+1,i}, \tag{1.1.2}$$

for some function $\omega : \mathbb{R}^d \to \mathbb{R}$, and we will see in Chapter 2 that random features can approximate a target function arbitrarily well under some relatively mild hypothesis on ω. The main assumption in equation 1.1.2 is that the characteristic-based instruments determine the weights of an asset on the SDF. Since we can add as many instruments and transformations (e.g., lags, interactions) this is not very restrictive.

However, it's important to note that the expected return of a stock cannot directly depend on firm characteristics. Instead, expected returns are correlated with these characteristics, which is why we focus on managed portfolios associated with them. If stock returns were directly dependent on characteristics like size, one could theoretically buy small com-

[1] `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_5_factors_2x3.html`

panies with high expected returns, consolidate them into a large holding company, and then distribute low returns to shareholders while pocketing the difference as the holding manager. In reality, this strategy fails because the large holding company would still behave like a portfolio of small firms, and its returns will still be like those of a portfolio of small firms, as discussed by Cochrane (2009, Section 5.1). Our model rules out such scenarios by explaining expected returns through their correlation with the stochastic discount factor (SDF), which itself depends on how a stock correlates with characteristic-based factors, and how much each factor impacts the SDF.

By definition 1.1.2 and a similar computation as before, it follows that the SDF coefficients are given by $b = \mathbb{V}[F_{t+1}]^{-1}\mathbb{E}[F_{t+1}]$. Given a sample of T observations, a naïve estimator is thus given by

$$\overline{\mu} = \frac{1}{T}\sum_{t=1}^{T} F_t,$$

$$\overline{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} (F_t - \overline{\mu})(F_t - \overline{\mu})^\top,$$

$$\hat{b} = \overline{\Sigma}^{-1}\overline{\mu}.$$

This estimator will perform very poorly in practice unless the number of factors L is very small relative to the sample size T. However, if we reduce the number of factors as in traditional models, then it becomes less plausible that an SDF is approximated by one living in this small linear subspace. To deal with the case of a large number of factors L several approaches have been considered in the literature.

1.2 KNS shrinkage estimator

In Kozak et al. (2020), the problem is stated in the framework of Bayesian statistics. The authors introduce a family of priors over the mean returns of factors as follows,

$$\mu \sim \mathcal{N}\left(0, \frac{\kappa^2}{\tau}\Sigma^\eta\right),$$

where Σ is the covariance matrix of the factors (assume it is known), $\tau = \mathrm{tr}[\Sigma]$ and κ is a constant that controls the scale of μ. In practice Σ is unknown and has to be estimated in sample, so it is replaced by $\overline{\Sigma}$ or more generally some estimator $\hat{\Sigma}$. They economically motivate the choice of $\eta = 2$, but other values (e.g. $\eta \geq 2$) remain plausible. Assuming a multivariate normal likelihood $F_t \sim \mathcal{N}(\mu, \Sigma)$, they consider a Bayesian regression

$$\overline{\mu} = \Sigma b + \varepsilon$$

where $\mu = \Sigma b$, thus $\varepsilon \sim \mathcal{N}\left(0, \frac{1}{T}\Sigma\right)$. Recall the following formula for Bayesian regression.

Proposition 1.2.1. Consider a linear regression model:

$$y = Xg + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, \Sigma)$ with known Σ and assume a prior $g \sim \mathcal{N}(0, \Sigma_g)$. Then the posterior distribution of g given the data y is normal with mean given by

$$g_p = \left(X^\top \Sigma^{-1} X + \Sigma_g^{-1}\right)^{-1} X^\top \Sigma^{-1} y \tag{1.2.1}$$

and posterior variance $\Sigma_p = (X^\top \Sigma^{-1} X + \Sigma_g^{-1})^{-1}$.

If $\eta = 2$, then the prior distribution over is $b = \Sigma^{-1}\mu \sim \mathcal{N}\left(0, \frac{\kappa^2}{\tau} I\right)$, thus the posterior distribution has mean

$$\hat{b} = (\Sigma + \gamma I)^{-1} \overline{\mu}, \tag{1.2.2}$$

and variance $(\Sigma + \gamma I)^{-1}$, where $\gamma = \frac{\tau}{\kappa^2 T}$. We can interpret this estimator as the solution of a penalised regression minimizing the *Hansen-Jagannathan distance* (Hansen and Jagannathan 1997) with an L^2 penalty

$$\hat{b} = \arg\min_b \left((\bar{\mu} - \Sigma b)^\top \Sigma^{-1} (\bar{\mu} - \Sigma b) + \gamma \|b\|^2\right). \tag{1.2.3}$$

We can understand better this estimator if we first perform a change of basis into that of principal component (PC) factors, by applying the spectral theorem to the covariance matrix. That is, we decompose $\Sigma = Q \Lambda Q^\top$ with Λ diagonal and Q an orthogonal matrix. Define the PC factors by $P_{t+1} = Q^\top F_{t+1}$. Then,

$$\hat{b}_{P_j} = \frac{\overline{\mu}_{P_j}}{\lambda_j + \gamma} = \left(\frac{\lambda_j}{\lambda_j + \gamma}\right)\left(\frac{\overline{\mu}_{P_j}}{\lambda_j}\right). \tag{1.2.4}$$

We observe that the solution is a shrinkage of the naïve (OLS) solution $\hat{b}^{OLS}_{P_j} = \frac{\overline{\mu}_{P_j}}{\lambda_j}$, where $\overline{\mu}_P = Q^\top \overline{\mu}$ are the mean of the PC factors. The shrinkage factor $\frac{\lambda_j}{\lambda_j + \gamma}$ is much stronger for small eigenvalues.

Notice that if we take $\eta = 3$, then $b = \Sigma^{-1}\mu \sim \mathcal{N}\left(0, \frac{\kappa^2}{\tau}\Sigma\right)$ whence

$$\hat{b} = \left(T\Sigma + \frac{\tau}{\kappa^2}\Sigma^{-1}\right)^{-1} T\overline{\mu} = (\Sigma^2 + \gamma I)^{-1} \Sigma \overline{\mu},$$

which is the solution of the ridge regression problem

$$\hat{b} = \arg\min_b \left(\|\bar{\mu} - \Sigma b\|^2 + \gamma \|b\|^2 \right).$$

They also consider adding an L^1 penalty term to the estimator of equation 1.2.3, resulting in the following optimisation problem

$$\hat{b} = \arg\min_b \frac{1}{2} \left[(\bar{\mu} - \Sigma b)^\top \Sigma^{-1} (\bar{\mu} - \Sigma b) + \gamma_2 \|b\|_2^2 \right] + \gamma_1 \|b\|_1. \tag{1.2.5}$$

This estimator will look for an SDF which is sparse in the characteristics (or more generally features) space, as only a few features will have non-zero coefficients. In general the authors find that this method performs poorly suggesting that there's no easy way to approximate the SDF with a small number of characteristics. If instead we first rotate the factors into the PC factors, then the problem will look for a sparse solution in PC factors (as opposed to equation 1.2.3, this problem has no rotational invariance). The problem is thus

$$\hat{b}_P = \arg\min_{b_P} \quad \frac{1}{2} \left[(\bar{\mu} - \Lambda b_P)^\top \Lambda^{-1} (\bar{\mu} - \Lambda b_P) + \gamma_2 \|b_P\|_2^2 \right] + \gamma_1 \|b_P\|_1.$$

This problem has a closed form solution given by

$$\hat{b}_{P_j} = \frac{\operatorname{sgn}(\hat{b}^{OLS}_{P_j}) \left(\hat{b}^{OLS}_{P_j} - \frac{\gamma_1}{\lambda_j} \right)_+}{\left(1 + \frac{\gamma_2}{\lambda_j} \right)}, \quad \hat{b}^{OLS}_{P_j} = \frac{\overline{\mu}_{P_j}}{\lambda_j}. \tag{1.2.6}$$

In either problems, the SDF or the MVE portfolio is characterised by the PC factors and their singular values, with appropiate scaling given by equations 1.2.4 and 1.2.6 respectively.

1.3 Maximum Sharpe ratio regression

Now we consider the approach introduced in Didisheim et al. (2023) and Kelly et al. (2024). As before, we have a matrix of features $X_t = \Phi(Z_t)$ which defines factor portfolios $F_{t+1} = X_t^\top R^e_{t+1}$. The authors parametrise the SDF as $M_{t+1} = 1 - b^\top F_{t+1}$, thus knowing population moments the optimal coefficients are given by $b = \mathbb{E}[F_{t+1} F_{t+1}^\top]^{-1} \mathbb{E}[F_{t+1}]$. Denote by F the $L \times T$ matrix of of factor returns. First notice that we can find these coefficients with an OLS regression of the constant vector $\mathbb{1} \in \mathbb{R}^T$ against the factors

$$\mathbb{1} = F^\top b + \varepsilon.$$

Hence the solution is given by

$$\hat{b} = \left(FF^\top\right)^{-1} F\mathbb{1} = \left(\frac{1}{T}FF^\top\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} F_t.$$

This naïve estimator will result in overfitting in sample and will perform poorly out of sample. When $L > T$ the system will generically have infinite solutions that perfectly fit the data in sample. Consider instead now a ridge regression estimator for this problem, i.e.

$$\hat{b} = \arg\min_b \frac{1}{T}\left\|\mathbb{1} - F^\top b\right\|^2 + \gamma\|b\|^2. \tag{1.3.1}$$

Thus, its solution is given by

$$\hat{b} = \left(FF^\top + (\gamma T)I\right)^{-1} F\mathbb{1} = \left(\frac{1}{T}FF^\top + \gamma I\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} F_t. \tag{1.3.2}$$

The estimator in equation 1.3.2 looks quite similar to that of equation 1.2.2. The main difference is that 1.2.2 is based on the covariance matrix of factors and 1.3.2 is instead based on the second moment matrix. This stems from the fact that they estimate different parametrisations of the SDF.

1.4 Kernel methods and random features

When the number of features L becomes too large (e.g. $L \sim 10^5$ or larger) then the formulas of equations 1.2.2 and 1.3.2 become prohibitively expensive. Computing the inverse of the regularised $L \times L$ matrix $B = \frac{1}{T}FF^\top$ (or $\hat{\Sigma}$) is $O(L^3)$, even storing this matrix in memory is $O(L^2)$ which can be infeasible in practice. Consider instead the matrix $\tilde{B} = \frac{1}{T}F^\top F$ which is $T \times T$. It turns out that we can compute $(B + \gamma I)^{-1}$ in terms of the eigenvalue decomposition of $\tilde{B}$.

Proposition 1.4.1. (Didisheim et al. 2023, Lemma 1) Let $X \in \mathbb{R}^{n\times m}$, let $X = UDV^\top$ be its compact SVD decomposition. Then

$$(XX^\top + \gamma I)^{-1} = U(D^2 + \gamma I)^{-1}U^\top + \frac{1}{\gamma}(I - UU^\top)$$

Proof. Let r be the rank of X, and $X^\top X = \sum_{i=1}^{r} \lambda_i q_i q_i^\top$ be the spectral decomposition of $X^\top X$ (q_i are orthonormal). Then $V = [q_1, \ldots, q_r] \in \mathbb{R}^{m\times r}$, $D = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ and $u_i = \frac{1}{\sqrt{\lambda_i}} X q_i$ (thus $U = [u_1, \ldots, u_r] \in \mathbb{R}^{n\times r}$) give the compact SVD of X.

Now consider the full SVD decomposition $X = \tilde{U}\tilde{D}\tilde{V}^\top$ of sizes $m \times m$, $m \times n$ and $n \times n$ respectively. $\tilde{U} = [U, U^*]$ and $\tilde{V} = [V, V^*]$ where U^* and V^* columns are orthonormal basis of $\ker(X^\top) = \mathrm{Im}(X)^\perp$ and $\ker(X)$ respectively, and $\tilde{D}$ is equal to D in its first r diagonal elements and elsewhere it is zero. Thus,

$$(XX^\top + \gamma I)^{-1} = \tilde{U}(\tilde{D}\tilde{D}^\top + \gamma I)\tilde{U}^\top$$
$$= U(D^2 + \gamma I)^{-1}U^\top + \frac{1}{\gamma}U^*(U^*)^\top = U(D^2 + \gamma I)^{-1}U^\top + \frac{1}{\gamma}(I - UU^\top),$$

where we used first that $(\tilde{D}\tilde{D}^\top + \gamma I)^{-1}$ is a diagonal matrix with values $1/(d_i^2 + \gamma)$ for $i = 1, \ldots, r$ and $1/\gamma$ for $r+1, \ldots, n$, and in the last equality we observe that $U^*(U^*)^\top = I - UU^\top$ as it is the orthogonal projection onto the null space of $X^\top$. □

As a consequence of 1.4.1 we find an alternative formula for the solution of ridge regression that is computationally more efficient for fat matrices (i.e., when the number of features exceeds the number of observations). Given that this is the most common scenario in our setting, we will use this alternative approach in practice.

Corollary 1.4.2. (Kelly et al. 2024, Lemma 1) Let $X \in \mathbb{R}^{n \times m}$. Given $y \in \mathbb{R}^n$ then the ridge regression solution of

$$\hat{\beta} = \arg\min_\beta \|y - X\beta\|^2 + \gamma\|\beta\|^2,$$

is given by

$$\hat{\beta} = (X^\top X + \gamma I)^{-1}X^\top y = X^\top(XX^\top + \gamma I)^{-1}y.$$

Proof. Let $X = UDV^\top$ the compact SVD decomposition of X. Then, it is well known that

$$\hat{\beta} = (X^\top X + \gamma I)^{-1}X^\top y = V(D^2 + \gamma I)^{-1}DU^\top y$$

Now using proposition 1.4.1 we get that

$$X^\top(XX^\top + \gamma I)^{-1}y = X^\top\left(U(D^2 + \gamma I)^{-1}U^\top + \frac{1}{\gamma}(I - UU^\top)\right)y$$
$$= VD(D^2 + \gamma I)U^\top y,$$

where we used that $X^\top(I - UU^\top) = 0$. This proves the equality of the two formulas for the ridge regression solution. □

Thus, when L is larger than T we can compute the eigenvalue decomposition of $\tilde{B}$ in order to compute $\hat{b}$ (take $X = \frac{1}{\sqrt{T}}F$ in proposition 1.4.1). Now we take a closer look at how

to compute the matrix $\tilde{B}$. First observe that,

$$F^\top F = \begin{bmatrix} F_1^\top F_1 & \dots & F_1^\top F_T \\ \vdots & \ddots & \vdots \\ F_T^\top F_1 & \dots & F_T^\top F_T \end{bmatrix} = \begin{bmatrix} (R_1^e)^\top \Phi(Z_0)\Phi(Z_0)^\top R_1^e & \dots & (R_1^e)^\top \Phi(Z_0)\Phi(Z_{T-1})^\top R_T^e \\ \vdots & \ddots & \vdots \\ (R_T^e)^\top \Phi(Z_{T-1})\Phi(Z_0)^\top R_1^e & \dots & (R_T^e)^\top \Phi(Z_{T-1})\Phi(Z_{T-1})^\top R_T^e \end{bmatrix}. \tag{1.4.1}$$

Thus, to compute $\tilde{B}$ we need to compute the factors $F_t = \Phi(Z_t)^\top R_{t+1}$, these cost $O(L \times N_t)$ operations each assuming each feature evaluation is computed in $O(1)$ operations. Hence to compute F we need $O(L \times N \times T)$ operations where N is the average number of assets. Then it costs $O(T \times L \times T)$ to compute $\tilde{B}$ and $O(T^3)$ to perform its eigenvalue decomposition. As these scale linearly in L, it is feasible to compute for large L, provided we have enough space to save the factor matrix in memory.

There is an alternative way of computing $F^\top F$ as in equation 1.4.1 in terms of a kernel which in some cases is independent of the number of features. For $z, z' \in \mathbb{R}^d$ define $k(z,z') = \langle \Phi(z), \Phi(z') \rangle$ where $\langle , \rangle$ denotes the inner product of $\mathbb{R}^L$. This is a positive definite kernel, see Mohri et al. (2018, Chapter 6) or Hardt and Recht (2022, Chapter 4) for an introduction to kernel methods in machine learning. Common examples of p.d. kernels are the polynomial kernel of degree q defined by

$$k(z,z) = (c + \langle z, z' \rangle)^q,$$

or the Gaussian kernel

$$k(z,z) = \exp(-c\|z - z'\|^2), \tag{1.4.2}$$

where $c > 0$ is a constant, considered in Kozak (2020). Suppose we can compute the kernel $k(z,z')$ in constant time $O(1)$ for $z, z' \in \mathbb{R}^d$. For example, in the polynomial kernel of degree q there are $L = \binom{d+q}{q}$ features corresponding to all monomials of degree less or equal than q, however $k(z,z')$ can be computed in $O(d)$ operations which for our purposes it is the same as constant time. Then we can compute the matrices $k(Z_t, Z_s)$ for $t, s = 0, \dots, T-1$, each with cost $O(N_t \times N_s)$. This is known as the *kernel trick*, which computes everything implicitely in terms of dot product of the transformed features $\Phi(z)$ without ever explicitly computing the transformation $\Phi(z)$ itself. The kernel trick allows algorithms to operate in very high-dimensional (even infinite-dimensional) spaces without the computational cost of explicitly transforming the data. Thus, the cost to compute

$$\tilde{B} = \frac{1}{T} F^\top F = \frac{1}{T} \begin{bmatrix} (R_1^e)^\top k(Z_0, Z_0) R_1^e & \dots & (R_1^e)^\top k(Z_0, Z_{T-1}) R_T^e \\ \vdots & \ddots & \vdots \\ (R_T^e)^\top k(Z_{T-1}, Z_0) R_1^e & \dots & (R_T^e)^\top k(Z_{T-1}, Z_{T-1}) R_T^e \end{bmatrix},$$

is $O(N^2T^2)$ which is independent of L.

We consider now random features introduced by Rahimi and Recht (2007). Specifically, consider a univariate activation function $\phi : \mathbb{R} \to \mathbb{R}$ (e.g. $\phi(x) = \mathrm{ReLU}(x) = \max(x, 0)$) and some distribution p over weights $w \in \Omega \subset \mathbb{R}^d$ (e.g. multivariate gaussian $N(0, I_d)$). Then given an i.i.d. sample $S = w_1, \ldots, w_L \sim p$ define the random features by

$$\phi_\ell(z) = \frac{1}{\sqrt{L}} \phi(w_\ell^\top z),$$

for $\ell = 1, \ldots, L$. Random features are also called random feature neural networks as they can be interpreted as feedforward neural networks with one hidden layer where the inner weights are left untrained. More specifically, one can associate a function

$$f(z; b, W) = b^\top \Phi(Wz) = \sum_{i=1}^{L} b_\ell \phi_\ell(z),$$

for each vector $b \in \mathbb{R}^L$ of outer weights to be trained, where $W \in \mathbb{R}^{L \times d}$ is the matrix of random weights stacked row-wise, see Figure 1.1.

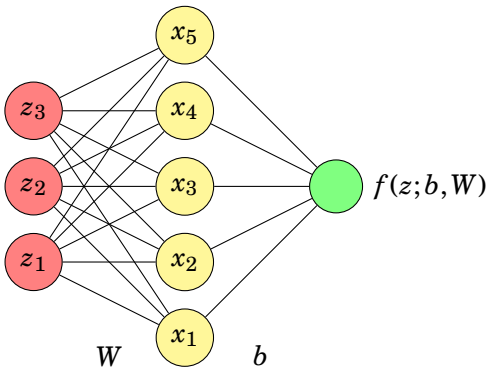


Figure 1.1: A random feature neural network represented as a feedforward neural network (FFN) with one hidden layer. The input layer is of dimension $d = 3$, while the number of random features (that is, the dimension of the hidden layer) is $L = 5$. The weight matrix $W = [w_1, \ldots, w_L]^\top \in \in \mathbb{R}^{L \times d}$ is not trained.

Then, associated to these random features there is a (random) kernel k_S that converges to the following (deterministic) limit

$$k_S(z, z') = \langle \Phi(z), \Phi(z') \rangle = \frac{1}{L} \sum_{\ell=1}^{L} \phi(W_\ell^\top z) \phi(W_\ell^\top z') \underset{L \to \infty}{\longrightarrow} \mathbb{E}[\phi(w^\top z) \phi(w^\top z')],$$

where $w \sim p$ by the SLLN. Surprisingly, this expectation has a closed formula in some cases. Rahimi and Recht (2007) show that the gaussian kernel defined in equation 1.4.2 can be approximated by cosine and sine activations (or complex exponentials) with gaussian weights. In the case of ReLU activations with gaussian weights there is a formula due to a

theorem of Cho and Saul (2009),

$$\mathbb{E}[\phi(w^\top z)\phi(w^\top z')] = \sin(\theta) + (\pi - \theta)\cos(\theta), \quad \theta = \cos^{-1}\left(\frac{z^\top z'}{\|z\|\|z'\|}\right),$$

i.e. the so called arc-cosine kernel of degree one. Thus, when we consider the limit of infinite random features, we obtain a kernel which can be computed in constant time. Hence, we can compute the matrix $\tilde{B}$ by equation 1.4 and thus we can compute the estimator $\hat{b}$ in this case.

To further elaborate, both regression approaches from Kozak (2020) (equation 1.2.3) and Didisheim et al. (2023) (equation 1.3.2) can be implemented in conjuntion with random features methods. In fact, Didisheim et al. (2023) employs random features with cosine and sine activations of different scales. It's important to note that Didisheim et al. (2023) re-ranks their random features (Didisheim et al. 2023, Equation 61), whereas Kozak et al. (2020) and Kozak (2020) do not re-rank but instead center (Kozak 2020, Subsection 2.2.2) or center and normalize (Kozak et al. 2020, Section 3.4) their derived features. While centering is straightforward, re-ranking disrupts the non-linearities in random features (e.g., causing all monotone activation functions to yield identical features) and alters the stronger exposures produced by certain model configurations. In our approach, we adhere to the regression model in equation 1.3.2 without applying any re-ranking or normalization. This modeling decision allows for a more straightforward theoretical analysis and the derivation of approximation bounds.

Chapter 2

Statistical learning theory

2.1 Rademacher complexity bounds

In this section we introduce the main results of Statistical Learning needed to understand why random features usually work well in practice. We start recalling McDiarmid's inequality, which is the main concentration inequality that is used to produce approximation and generalization bounds.

Theorem 2.1.1. (McDiarmid's inequality, Mohri et al. 2018, Theorem D.8) Let $X_1,\dots,X_n \in \mathcal{X}^n$ be i.i.d. random variables, and assume there are constants $c_1,\dots,c_n$ such that the function $f:\mathcal{X}^n \to \mathbb{R}$ satisfies the bounded differences property:

$$|f(x_1,\dots,x_i,\dots,x_n) - f(x_1,\dots,\tilde{x}_i,\dots,x_n)| \leq c_i,$$

for all $i = 1,\dots,n$ and any points, $x_1,\dots,x_n,\tilde{x}_i \in \mathcal{X}$. Then for all $\varepsilon > 0$ the following inequalities hold:

$$\mathbb{P}[f(X_1,\dots,X_n) - \mathbb{E}[f(X_1,\dots,X_n)] \geq \varepsilon] \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right),$$

$$\mathbb{P}[f(X_1,\dots,X_n) - \mathbb{E}[f(X_1,\dots,X_n)] \leq -\varepsilon] \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right),$$

and as a consequence,

$$\mathbb{P}[|f(X_1,\dots,X_n) - \mathbb{E}[f(X_1,\dots,X_n)]| \geq \varepsilon] \leq 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Remark 2.1.2. McDiarmid's inequality has a beatiful proof using martingales, specifically, the Doob's martingale of the sequence, see Mohri et al. (2018, Appendix D). However we

decided not to include it here as this technique is not used elsewhere.

Now we introduce the notion of Rademacher complexity which measures how well some function in a class $\mathcal{F}$ can be correlated with a random noise sequence. The typical example to have in mind is a family of loss functions associated to a parametric (or not) class of functions

$$\mathcal{F} = \{f_\theta : (x,y) \to L(h_\theta(x), y) \,|\, \theta \in \Theta\},$$

where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is some specified loss function (e.g. $L(y,y') = (y-y')^2$, etc.).

Definition 2.1.3. (Bartlett and Mendelson 2002, Definition 2) Let p be a probability distribution over a set $\mathcal{X}$ and suppose $S = X_1,\ldots,X_n$ is an i.i.d. sample drawn from p. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. Then define the empirical Rademacher complexity as the following random variable

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right| \,\middle|\, X_1,\ldots,X_n\right],$$

where $\sigma_1,\ldots,\sigma_n$ are i.i.d. uniform on $\{-1,1\}$ (also known as Rademacher random variables) independent from $X_1,\ldots,X_n$. Then, define the Rademacher complexity as $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}\left[\hat{\mathfrak{R}}_S(\mathcal{F})\right]$, thus by the tower property,

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right|\right].$$

Remark 2.1.4. Some authors such as Mohri et al. (2018) use an alternative definition of the (empirical) Rademacher complexity, dropping the absolute value and the 2 inside the expectation. These definitions are not equivalent but it can be shown that they differ at most by a constant of 2.

Now, we bound the supremum difference between the sample mean of a function on a class $\mathcal{F}$ and its expectation. The argument is classic and it is based on the idea of symmetrisation and is similar to the one given in the proof of Bartlett and Mendelson (2002, Theorem 8).

Proposition 2.1.5. Let p be a probability distribution over a set $\mathcal{X}$ and suppose $X_1,\ldots,X_n$ are i.i.d. samples drawn from p. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. Then,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right|\right] \le \mathfrak{R}_n(\mathcal{F}).$$

Proof. First we introduce $X'_1,\dots,X'_n$ a ghost i.i.d. sample from p independent from the sample $X_1,\dots,X_n$. Then,

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}[f(X)]\right|\right] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(X'_i)\right]\right|\right] \\
&\overset{(1)}{=} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(X_i)-\frac{1}{n}\sum_{i=1}^n f(X'_i)\,\middle|\,X_1,\dots,X_n\right]\right|\right] \\
&\overset{(2)}{\leq} \mathbb{E}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\frac{1}{n}\sum_{i=1}^n f(X'_i)\right|\,\middle|\,X_1,\dots,X_n\right]\right] \\
&\overset{(3)}{=} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\frac{1}{n}\sum_{i=1}^n f(X'_i)\right|\right],
\end{aligned}$$

where (1) follows since $X'_1,\dots,X'_n$ are independent of $X_1,\dots,X_n$, (2) follows from the monotonicity of conditional expectation, and (3) by the tower property. Now, since $Z_i = f(X_i) - f(X'_i)$ is a symmetric random variable (around 0), then $\sigma_i Z_i = \sigma_i(f(X_i)-f(X'_i))$ has the same distribution as Z_i for $\sigma_1,\dots,\sigma_n$ i.i.d. Rademacher random variables independent form $X_1,\dots,X_n$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\frac{1}{n}\sum_{i=1}^n f(X'_i)\right|\right] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i)-f(X'_i))\right|\right] \\
\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right|+\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X'_i)\right|\right] &\leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right|\right],
\end{aligned}$$

thus we get our result. □

One can also bound the difference between the empirical Rademacher complexity and the Rademacher complexity of a family of functions taking values in the interval $[0,1]$. The following is essentially a restatement of Bartlett and Mendelson (2002, Theorem 11).

Theorem 2.1.6. Let $X_1,\dots,X_n$ be an i.i.d. sample of random variables with values in $\mathcal{X}$. Let $\mathcal{F}$ be a class of functions $f:\mathcal{X}\to[0,1]$. Then,

$$\begin{aligned}
\mathbb{P}\left[\hat{\mathfrak{R}}_S(\mathcal{F})-\mathfrak{R}_n(\mathcal{F})\geq\varepsilon\right] &\leq \exp\left(\frac{-n\varepsilon^2}{2}\right), \\
\mathbb{P}\left[\hat{\mathfrak{R}}_S(\mathcal{F})-\mathfrak{R}_n(\mathcal{F})\leq-\varepsilon\right] &\leq \exp\left(\frac{-n\varepsilon^2}{2}\right).
\end{aligned}$$

Proof. Define $g(x_1,\dots,x_n) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^n \sigma_i f(x_i)\right|\right]$ where the expectation is over the i.i.d. Rademacher random variables $\sigma_1,\dots,\sigma_n$. Then g satisfies the bounded differences property

as follows,

$$\begin{aligned}
&|g(x_1,\dots,x_i,\dots,x_n) - g(x_1,\dots,\tilde{x}_i,\dots,x_n)| \\
&= \left| \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{2}{n}\sum_{i=1}^n \sigma_i f(x_i) \right| \right] - \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{2}{n}\sum_{i=1}^n \sigma_i f(x_i) + \frac{2}{n}\sigma_i(f(\tilde{x}_i) - f(x_i)) \right| \right] \right| \\
&\overset{(1)}{\leq} \left| \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left\{ \left| \frac{2}{n}\sum_{i=1}^n \sigma_i f(x_i) \right| - \left| \frac{2}{n}\sum_{i=1}^n \sigma_i f(x_i) + \frac{2}{n}\sigma_i(f(\tilde{x}_i) - f(x_i)) \right| \right\} \right] \right| \\
&\overset{(2)}{\leq} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{2}{n} |\sigma_i(f(\tilde{x}_i) - f(x_i))| \right] \leq \sup_{f\in\mathcal{F}} \frac{2}{n} |f(\tilde{x}_i) - f(x_i)| \leq \frac{2}{n},
\end{aligned}$$

where (1) follows since the difference of the suprema is less than the supremum of the difference and (2) is the reverse triangle inequality (i.e., $|\|x\| - \|y\|| \leq \|x - y\|$, holds in any normed vector space). Then, since $\mathbb{E}[g(X_1,\dots,X_n)] = \mathbb{E}[\hat{\mathfrak{R}}_S(\mathcal{F})] = \mathfrak{R}_n(\mathcal{F})$, the result follows by McDiarmid's inequality 2.1.1. □

We are now able to state and prove a "generalization bound", which bounds uniformly over a class $\mathcal{F}$ the empirical mean of a function and its expectation. It is essentially the same as Mohri et al. (2018, Theorem 3.3) with the caveat of observation 2.1.4.

Theorem 2.1.7. Let $X_1,\dots,X_n$ be an i.i.d. sample of random variables with values in $\mathcal{X}$. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to [0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over $X_1,\dots,X_n$, the following inequalities holds for all $f \in \mathcal{F}$:

$$\left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \leq \mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}},$$

and

$$\left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \leq \hat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

Proof. Define $g(x_1,\dots,x_n) = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}[f(X)] \right|$. We now show that g satisfies the bounded differences property as follows,

$$\begin{aligned}
&|g(x_1,\dots,x_i,\dots,x_n) - g(x_1,\dots,\tilde{x}_i,\dots,x_n)| \\
&= \left| \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}[f(X)] \right| - \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}[f(X)] + \frac{1}{n}(f(\tilde{x}_i) - f(x_i)) \right| \right| \\
&\leq \left| \sup_{f\in\mathcal{F}} \left\{ \left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}[f(X)] \right| - \left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}[f(X)] + \frac{1}{n}(f(\tilde{x}_i) - f(x_i)) \right| \right\} \right| \\
&\leq \sup_{f\in\mathcal{F}} \frac{1}{n} |f(\tilde{x}_i) - f(x_i)| \leq \frac{1}{n}.
\end{aligned}$$

Moreover, by Proposition 2.1.5 we have that $\mathbb{E}[g(X_1,\dots,X_n)] \le \mathfrak{R}_n(\mathcal{F})$. Thus, by McDiarmid's inequality 2.1.1, we have that

$$\mathbb{P}[g(X_1,\dots,X_n) \ge \mathfrak{R}_n(\mathcal{F}) + \varepsilon] \le \mathbb{P}[g(X_1,\dots,X_n) - \mathbb{E}[g(X_1,\dots,X_n)] \ge \varepsilon] \le \exp(-2n\varepsilon^2) =: \delta.$$

Rearranging we get that $\varepsilon = \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}$ and we get the first inequality. To get the second inequality we use a union bound with Theorem 2.1.6 as follows,

$$\begin{aligned}
&\mathbb{P}\left[g(X_1,\dots,X_n) \ge \hat{\mathfrak{R}}_S(\mathcal{F}) + 3\varepsilon\right] \\
&\le \mathbb{P}[g(X_1,\dots,X_n) \ge \mathfrak{R}_n(\mathcal{F}) + \varepsilon] + \mathbb{P}\left[\mathfrak{R}_n(\mathcal{F}) \ge \hat{\mathfrak{R}}_S(\mathcal{F}) + 2\varepsilon\right] \\
&= 2\exp(-2n\varepsilon^2) =: \delta.
\end{aligned}$$

Rearranging we get that $\varepsilon = \sqrt{\frac{\log(\frac{2}{\delta})}{2n}}$ and we get the second inequality. □

The next results shows how one can bound the Rademacher complexity of linear functions with coefficients (or weights) bounded in the euclidean norm. This follows from Bartlett and Mendelson (2002, Lemma 22) for the case of the euclidean inner product kernel $k(x,y) = x^\top y$.

Proposition 2.1.8. Let p be a probability distribution over a set $\mathcal{X}$ and suppose $X_1,\dots,X_n$ are i.i.d. samples drawn from p. Assume $\mathbb{E}[\|X\|_2^2] \le C^2$. Let $\mathcal{F} = \{x \to w^\top x \mid \|w\|_2 \le B\}$. Then,

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{\|w\|_2 \le B} \left|\frac{2}{n}\sum_{i=1}^n \sigma_i w^\top X_i\right|\right] \le \frac{2BC}{\sqrt{n}},$$

for $\sigma_1,\dots,\sigma_n$ are independent Rademacher random variables.

Proof. We first bound by Cauchy-Schwartz inequality,

$$\mathbb{E}\left[\sup_{\|w\|_2 \le B} \left|\frac{2}{n}\sum_{i=1}^n \sigma_i w^\top X_i\right|\right] \le \mathbb{E}\left[\sup_{\|w\|_2 \le B} \|w\|_2 \left\|\frac{2}{n}\sum_{i=1}^n \sigma_i X_i\right\|_2\right] \le 2B\,\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i X_i\right\|_2\right].$$

Then, we compute

$$\begin{aligned}
2B\,\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i X_i\right\|_2\right] &\overset{(1)}{\le} 2B\sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \sigma_i X_i\right\|_2^2\right]} \\
&\overset{(2)}{\le} 2B\sqrt{\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\|X_i\|_2^2\right]} \overset{(3)}{\le} \frac{2BC}{\sqrt{n}},
\end{aligned}$$

(1) Jensen's inequality for $\sqrt{\cdot}$, (2) because this is the variance of a sum of i.i.d. random variables and $\mathbb{E}[\|\sigma_i X_i\|_2^2] = \mathbb{E}[\sigma_i^2]\mathbb{E}[\|X_i\|_2^2] = \mathbb{E}[\|X_i\|_2^2]$ and (3) because $\mathbb{E}[\|X\|_2^2] \le C^2$. □

The following result, sometimes referred to as "Talagrand's Lemma", is particularly useful to bound the (empirical) Rademacher complexity of families of functions which are obtained via compositions with contractions (or more generally, Lipschitz functions).

Theorem 2.1.9. (Ledoux and Talagrand 1991, Theorem 4.12) Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ convex and increasing. Let $\varphi_i : \mathbb{R} \to \mathbb{R}$ be contractions such that $\varphi_i(0) = 0$. Then for any bounded subset $T \subset \mathbb{R}^n$ we have that

$$\mathbb{E}\left[F\left(\frac{1}{2}\sup_{t\in T}\left|\sum_{i=1}^{n}\sigma_i\varphi_i(t_i)\right|\right)\right] \leq \mathbb{E}\left[F\left(\sup_{t\in T}\left|\sum_{i=1}^{n}\sigma_i t_i\right|\right)\right]$$

where $t = (t_1, \ldots, t_n) \in T$ and $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables.

An immediate corollary is the following result.

Corollary 2.1.10. (Bartlett and Mendelson 2002, Theorem 12.4) Let $\phi : \mathbb{R} \to \mathbb{R}$ be K-Lipschitz such that $\phi(0) = 0$, then $\mathfrak{R}_n(\phi \circ \mathcal{F}) \leq 2K\mathfrak{R}_n(\mathcal{F})$.

2.2 Complexity bounds for non-i.i.d. processes

The main results of the previous section can be generalized to the non-i.i.d. case, such as Mohri and Rostamizadeh (2008) for the case of stationary β-mixing processes. This will be used later as the i.i.d. assumption does not hold in practice for most financial time series. First recall the notion of a (strictly) stationary sequence.

Definition 2.2.1. (Stationarity) A sequence of random variables $\mathbf{X} = (X_t)_{t\in\mathbb{Z}}$ is said to be stationary if for any t and non-negative integers m and k the vectors $(X_t, \ldots, X_{t+m})$ and $(X_{t+k}, \ldots, X_{t+m+k})$ have the same distribution.

Notice that in particular, for a stationary sequence $\mathbf{X}$ the distribution of X_t is the same for all times $t \in \mathbb{Z}$. Next, we recall the concept of β-mixing sequences introduced in Yu (1994). Informally, a stationary sequence is β-mixing if the dependence between samples decreases with time. In other words, if two samples are far away from each other then they are roughly independent.

Definition 2.2.2. (β-mixing, Yu 1994, Definition 2.2) Let $\mathbf{X} = (X_t)_{t\in\mathbb{Z}}$ be a stationary sequence, denote by σ_l the sigma algebra generated by the random variables X_t for $t \leq l$, i.e. $\sigma_l = \sigma(\{X_t \mid t \leq l\})$ and similarly $\sigma'_l = \sigma(\{X_t \mid t \geq l\})$. Then, for k any positive integer, define the k-th β-mixing coefficient of $\mathbf{X}$ as follows

$$\beta(k) = \sup_l \mathbb{E}\left[\sup_{A\in\sigma'_{l+k}} \left|\mathbb{P}[A \mid \sigma_l] - \mathbb{P}[A]\right|\right].$$

We say that $\mathbf{X}$ is β-mixing if $\beta(k) \to 0$ as $k \to \infty$.

Example 2.2.3. If $\mathbf{X} = (X_t)_{t\in\mathbb{Z}}$ is a strictly stationary countable-state Markov chain, then $\mathbf{X}$ is irreducible and aperiodic if and only if it is β-mixing (see Bradley 2005, Theorem 3.2).

The definition of Rademacher complexity can be extended in an almost verbatim manner to stationary sequences. Notice that since the distribution of a sample from a stationary sequence only depends on the sample size (i.e., it is independent of t) one can take any sample of size m from it to define the Rademacher complexity.

Definition 2.2.4. (Mohri and Rostamizadeh 2008, Definition 3) Given a stationary sequence $\mathbf{X} = (X_t)_{t\in\mathbb{Z}}$ taking values over a set $\mathcal{X}$ and let $S = (X_t,\dots,X_{t+m-1})$ be a sample of size m this process. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. Define the function

$$R(x_1,\dots,x_m) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right|\right],$$

for $x = (x_1,\dots,x_m) \in \mathcal{X}^m$, where $\sigma_1,\dots,\sigma_m$ are i.i.d. Rademacher random variables. Then define the empirical Rademacher complexity as the random variable

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = R(S) = R(X_t,\dots,X_{t+m-1}).$$

Finally, define the Rademacher complexity as $\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}\left[\hat{\mathfrak{R}}_S(\mathcal{F})\right]$.

While the empirical Rademacher complexity can be estimated from a sample of a stationary sequence, the Rademacher complexity can be even harder to compute than in the case of i.i.d. sequences. Nonetheless, the key feature of stationary β-mixing sequences is that one can for most purposes replace the original sequence by a sequence of independent blocks and the difference in expectations can be controlled by the β-mixing coefficients.

This method is known as the independent blocks technique. The method consists of splitting the sample S in two samples S_0 and S_1, each consisting of μ blocks of a consecutive points, thus $m = 2\mu a$. These are defined as follows,

$$\begin{aligned} S_0 &= (X_1^{(0)},\dots,X_\mu^{(0)}), \qquad \text{where } X_i^{(0)} = (X_{(2i-2)a+1},\dots,X_{(2i-2)a+a}),\\ S_1 &= (X_1^{(1)},\dots,X_\mu^{(1)}), \qquad \text{where } X_i^{(1)} = (X_{(2i-1)a+1},\dots,X_{(2i-1)a+a}), \end{aligned}$$

for $i = 1,\dots,\mu$, see Table 2.1. Then, we can replace S_0 by a sequence $\tilde{S}_0 = (\tilde{X}_1^{(0)},\dots,\tilde{X}_\mu^{(0)})$ where the blocks $\left(\tilde{X}_k^{(0)}\right)_{k=1}^{\mu}$ are mutually independent, but each block $\tilde{X}_k^{(0)}$ has the same distribution as $X_k^{(0)}$. For a sufficiently fast β-mixing distribution, and sufficiently large

blocks sizes $a > 0$, the expectation of a function defined on these blocks is only slightly change if we replace S_0 with $\tilde{S}_0$.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
$X_1^{(0)}$			$X_1^{(1)}$			$X_2^{(0)}$			$X_2^{(1)}$		

Table 2.1: A visual representation of the independent block technique. A sample S of size $m = 12$ is split into $\mu = 2$ blocks of size $a = 3$.

With this technique Mohri and Rostamizadeh (2008) are able to extend the generalization bound of Theorem 2.1.7 to the case of a stationary β-mixing sequence. The cost we pay is that we have to split the sequence into independent blocks as above, δ must be larger than $2(\mu - 1)\beta(a)$, thus we need $\beta(a)$ to be small so as to have a high probability bound, and the size of the independent sample for the bound is μ instead of m.

Theorem 2.2.5. (Mohri and Rostamizadeh 2008, Theorems 1 and 2) Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $[0,1]$. Then for any sample of size m drawn from a stationary β-mixing sequence, and for any $\mu, a > 0$ such that $m = 2\mu a$, and $\delta > 2(\mu - 1)\beta(a)$, with probability at least $1 - \delta$ the following inequality hold for all $f \in \mathcal{F}$,

$$\mathbb{E}[f(X_t)] - \frac{1}{m}\sum_{i=1}^{m} f(X_{t+i-1}) \leq \mathfrak{R}_\mu^{\tilde{D}}(\mathcal{F}) + \sqrt{\frac{\log\left(\frac{2}{\delta'}\right)}{2\mu}},$$

where $\delta' = \delta - 2(\mu - 1)\beta(a)$ and $\mathfrak{R}_\mu^{\tilde{D}}(\mathcal{F})$ denotes the Rademacher complexity of a sample of size μ of i.i.d. distributed random variables as X_t. Moreover, with probability at least $1 - \delta$,

$$\mathbb{E}[f(X_t)] - \frac{1}{m}\sum_{i=1}^{m} f(X_{t+i-1}) \leq \hat{\mathfrak{R}}_{S_\mu}(\mathcal{F}) + 3\sqrt{\frac{\log\left(\frac{4}{\delta'}\right)}{2\mu}},$$

where $\delta' = \delta - 4(\mu - 1)\beta(a)$ and S_μ is a subsample of μ points separated by a gap of $2a - 1$ points (say, extract the first point of each block of S_0).

Remark 2.2.6. Alternatively, less restrictive weak dependence assumptions, like those introduced in Gonon et al. (2020, Assumptions 1 and 2), could be considered. Moreover, a generalization of Rademacher complexity is defined for reservoir systems with a causal Bernoulli shift structure in Gonon et al. (2020). This approach is particularly relevant because many standard time-series are non-mixing, and even for those that are, mixing is not a testable property, and estimating mixing parameters is challenging in practice.

2.3 Random features

Let $\{\phi(\cdot,w) \mid w \in \Omega\}$ be a family of functions over a compact set $\mathcal{X} \subset \mathbb{R}^d$, parametrized by a set Ω. We are interested in approximating functions which have an integral representation

$$f(x) = \int_\Omega \alpha(w)\phi(x,w)\,dw,$$

for some $\alpha : \Omega \to \mathbb{R}$. Given a probability distribution p over the parameter space Ω, we can define a norm for such functions as follows

$$\|f\|_p = \sup_{w\in\Omega} \left|\frac{\alpha(w)}{p(w)}\right|. \tag{2.3.1}$$

Denote by $\mathcal{F}_p$ the vector space of functions $f(x) = \int_\Omega \alpha(w)\phi(x,w)\,dw$ such that $\|f\|_p < \infty$ (i.e., with finite p norm), it is easy to check that $\mathcal{F}_p$ is thus a normed space. We will approximate f by functions of the following form

$$\hat{f}(x) = \frac{1}{L}\sum_{\ell=1}^{L} b_\ell \phi(x,w_\ell),$$

where $w_1,\dots,w_L$ are i.i.d. sampled from p and we can only choose the coefficients $b = (b_1,\dots,b_L)$.

The following theorem gives an approximation result in the $\mathscr{L}^2$ sense.

Theorem 2.3.1. (Rahimi and Recht 2008a, Lemma 1) Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact and μ be a probability measure on $\mathcal{X}$. Let $\{\phi(\cdot,w) \mid w \in \Omega\}$ be a family of functions and p a probability distribution over Ω. Let $f \in \mathcal{F}_p$ as defined in equation 2.3.1. Suppose $\sup_{x\in\mathcal{X},w\in\Omega} |\phi(x,w)| \le 1$. Then for any $\delta > 0$, with probability at least $1-\delta$ over $w_1,\dots,w_L$ drawn i.i.d. from p, there exists $b_1,\dots,b_L$ such that $\hat{f}(x) = \frac{1}{L}\sum_{\ell=1}^{L} b_\ell \phi(x,w_\ell)$ satisfies

$$\|\hat{f} - f\|_{\mathscr{L}^2(\mathcal{X},\mu)} \le \frac{\|f\|_p}{\sqrt{L}}\left(1 + \sqrt{2\log\left(\frac{1}{\delta}\right)}\right),$$

where $\|f\|_p = \sup_{w\in\Omega}\left|\frac{\alpha(w)}{p(w)}\right|$. Moreover, $b = (b_1,\dots,b_L)$ is such that $\|b\|_\infty \le \|f\|_p$.

Proof. Let $f(x) = \int_\Omega \alpha(w)\phi(x,w)\,dw$. Then, given $w_1,\dots,w_L$ drawn i.i.d. from p, define $f_\ell(x) = b_\ell\phi(x,w_\ell)$ where $b_\ell = \frac{\alpha(w_\ell)}{p(w_\ell)}$ are importance sampling chosen weights, and define

$\hat{f} = \frac{1}{L}\sum_{i=1}^{L} f_\ell$. Then,

$$\mathbb{E}_p[f_\ell(x)] = \mathbb{E}_p\left[\frac{\alpha(w)}{p(w)}\phi(x,w)\right] = \int_\Omega \frac{\alpha(w)}{p(w)}\phi(x,w)p(w)\,dw = \int_\Omega \alpha(w)\phi(x,w)\,dw = f(x),$$

so in particular $\mathbb{E}_p[\hat{f}(x)] = f(x)$ for all $x \in \mathcal{X}$. Also, since $\sup_{x\in\mathcal{X}, w\in\Omega} |\phi(x,w)| \leq 1$ we have that

$$\|f_\ell\|_{\mathscr{L}^2(\mathcal{X},\mu)} = \sqrt{\int_{\mathcal{X}} (b_\ell \phi(x,w_\ell))^2\, d\mu(x)} \leq \|f\|_p.$$

Consider the function

$$g(w_1,\ldots,w_L) = \|\hat{f} - f\|_{\mathscr{L}^2(\mathcal{X},\mu)} = \left\| \frac{1}{L}\sum_{\ell=1}^{L} f_\ell - f \right\|_{\mathscr{L}^2(\mathcal{X},\mu)},$$

where the functions f_ℓ are defined as above. We show that g satisfies the bounded differences property. Let $\tilde{f}_i(x) = \frac{\alpha(\tilde{w}_i)}{p(\tilde{w}_i)}\phi(x,\tilde{w}_i)$ for $\tilde{w}_i \in \Omega$. If $\tilde{f}$ denotes the average of the functions f_ℓ where we replace f_i by $\tilde{f}_i$ then,

$$\begin{aligned} |g(w_1,\ldots,w_L) - g(w_1,\ldots,\tilde{w}_i,\ldots,w_L)| &= |\|\hat{f} - f\|_2 - \|\tilde{f} - f\|_2| \\ &\leq \|\hat{f} - \tilde{f}\|_2 = \frac{1}{L}\|f_i - \tilde{f}_i\|_2 \leq \frac{2\|f\|_p}{L}, \end{aligned}$$

where we used that f_i and $\tilde{f}_i$ are in the ball of radius $\|f\|_p$ in $\mathscr{L}^2(\mathcal{X},\mu)$. Now we bound the mean of g as follows,

$$\begin{aligned} \mathbb{E}[g(w_1,\ldots,w_L)] = \mathbb{E}\left[\left\|\frac{1}{L}\sum_{\ell=1}^{L} f_\ell - f\right\|_2\right] &\overset{(1)}{\leq} \sqrt{\mathbb{E}\left[\left\|\frac{1}{L}\sum_{\ell=1}^{L}(f_\ell - f)\right\|_2^2\right]} \\ &\overset{(2)}{=} \sqrt{\frac{1}{L^2}\sum_{\ell=1}^{L}\mathbb{E}[\|f_\ell - f\|_2^2]} \overset{(3)}{\leq} \frac{\|f\|_p}{\sqrt{L}}, \end{aligned}$$

where (1) is Jensen's inequality for the square root, (2) because this is the variance of a sum of i.i.d. random variables, and to show (3) observe that

$$\mathbb{E}[\|f_\ell - f\|_2^2] \leq \mathbb{E}[\|f_\ell\|_2^2] \leq \|f\|_p^2,$$

using Fubini and the fact that $f(x) = \mathbb{E}[f_\ell(x)]$ and the mean minimizes the squared distance to a constant.

Thus, by McDiarmid's inequality 2.1.1,

$$\mathbb{P}\left(\|\hat{f}-f\|_2 \geq \frac{\|f\|_p}{\sqrt{L}}+\varepsilon\right) \leq \mathbb{P}\left(g(w_1,\ldots,w_L) \geq \mathbb{E}[g(w_1,\ldots,w_L)]+\varepsilon\right) \leq \exp\left(\frac{-\varepsilon^2 L}{2\|f\|_p^2}\right) =: \delta.$$

Rearranging terms we get that $\varepsilon = \frac{\|f\|_p}{\sqrt{L}}\sqrt{2\log\left(\frac{1}{\delta}\right)}$. Thus with probability $1-\delta$,

$$\|\hat{f}-f\|_2 \leq \frac{\|f\|_p}{\sqrt{L}}\left(1+\sqrt{2\log\left(\frac{1}{\delta}\right)}\right).$$

□

Given the feature map $\phi : \mathcal{X}\times\Omega \to \mathbb{R}$ and a probability measure p over Ω, we can define an associated p.d. kernel as follows,

$$k(x,y) = \mathbb{E}_p[\phi(x,w)\phi(y,w)] = \int_\Omega \phi(x,w)\phi(y,w)\,p(w)\,dw. \tag{2.3.2}$$

Associated to this kernel there is a Hilbert space $\mathcal{H}$ called its reproducing kernel Hilbert space (RKHS). Formally, this Hilbert space is constructed as the closure of the linear span of the functions $k_x : y \to k(x,y)$ for $x \in \mathcal{X}$. It turns out that for this particular kernel, $\mathcal{H}$ can be given a more explicit description.

Proposition 2.3.2. (Rahimi and Recht 2008b, Proposition 4.1) Let $\phi : \mathcal{X}\times\Omega \to \mathbb{R}$ be a feature map such that $\sup_{x\in\mathcal{X},w\in\Omega} |\phi(x,w)| \leq 1$. Then, the RKHS $\mathcal{H}$ of the p.d. kernel k defined by equation 2.3.2 is given by the set of functions $f(x) = \int_\Omega \alpha(x)\phi(x,w)\,dw$ such that $\int_\Omega \frac{\alpha(w)^2}{p(w)}dw < \infty$, with the inner product

$$\langle f,g\rangle_{\mathcal{H}} = \int_\Omega \frac{\alpha(w)\beta(w)}{p(w)}dw,$$

for $g(x) = \int_\Omega \beta(x)\phi(x,w)\,dw$.

Proof. To show that $\mathcal{H}$ is the RKHS it is enough to show that

1. For all $x \in \mathcal{X}$, the function $k_x = k(x,-) : \mathcal{X} \to \mathbb{R}$ is in $\mathcal{H}$.

2. For all $x \in \mathcal{X}$ and $f \in \mathcal{H}$ the reproducing property holds:

$$f(x) = \langle f, k_x\rangle_{\mathcal{H}}.$$

Given $x \in \mathcal{X}$, then

$$k_x(y) = k(x,y) = \int_\Omega \phi(x,w)\phi(y,w)\,p(w)\,dw = \int_\Omega \alpha(w)\phi(y,w)\,dw,$$

where $\alpha(w) = \phi(x,w)p(w)$. Moreover,

$$\int_\Omega \frac{\alpha(w)^2}{p(w)} dw = \int_\Omega (\phi(x,w))^2 p(w)\, dw \leq 1,$$

as $|\phi(x,w)| \leq 1$ for all (x,w). Thus $k_x \in \mathcal{H}$ and (1) follows.

Now given $f(x) = \int \alpha(w)\phi(x,w)\, dw$, then

$$\langle f, k_x \rangle_{\mathcal{H}} = \int_\Omega \frac{\alpha(w)\phi(x,w)p(w)}{p(w)}\, dw = f(x),$$

so the reproducing property (2) holds. □

Moreover, one can show that the space $\mathcal{F}_p$ is dense in the RKHS.

Proposition 2.3.3. (Rahimi and Recht 2008b, Theorem 4.2) Let $\phi : \mathcal{X} \times \Omega \to \mathbb{R}$ be a feature map such that $\sup_{x \in \mathcal{X}, w \in \Omega} |\phi(x,w)| \leq 1$. Let $\mathcal{F}_p$ be the normed space of functions defined in equation 2.3.1. Then, $\mathcal{F}_p$ is a dense subspace of the RKHS $\mathcal{H}$ of the kernel k defined in equation 2.3.2.

Proof. First notice that if $f \in \mathcal{F}_p$ then

$$\int_\Omega \frac{\alpha(w)^2}{p(w)} dw = \int_\Omega \left(\frac{\alpha(w)}{p(w)}\right)^2 p(w) dw \leq \|f\|_p^2,$$

thus $f \in \mathcal{H}$. To show that $\mathcal{F}_p$ is dense it suffices to show that k_x are in $\mathcal{F}_p$, as their linear span is dense in $\mathcal{H}$. Now, $k_x(y) = \int_\Omega \alpha(w)\phi(y,w)\, dw$ for $\alpha(w) = \phi(x,w)p(w)$, and thus

$$\|k_x\|_p = \sup_{w \in \Omega} \left| \frac{\alpha(w)}{p(w)} \right| = \sup_{w \in \Omega} |\phi(x,w)| \leq 1.$$

Hence $k_x \in \mathcal{F}_p$ and the result follows. □

Remark 2.3.4. Theorem 2.3.1 can be relaxed to the hypothesis that $\|\phi(-,w)\|_{\mathscr{L}^2(\mathcal{X},\mu)} \leq 1$ for all $w \in \Omega$. Also, Propositions 2.3.2 and 2.3.3 hold under the weaker hypothesis that $\phi(x,-) \in \mathscr{L}^2(\Omega, p)$ for all $x \in X$, although the proof for the latter result is different as k_x is no longer necessarily in $\mathcal{F}_p$.

The next result shows that under slightly stronger hypothesis than Theorem 2.3.1, one can approximate a function in $\mathcal{F}_p$ uniformly, i.e. with respect to the ∞-norm of uniform convergence.

Theorem 2.3.5. (Rahimi and Recht 2008b, Theorem 3.2) Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact set. Let $\phi(x,w) = \phi(w^\top x)$ for $\phi : \mathbb{R} \to \mathbb{R}$ K-Lipschitz, $\phi(0) = 0$ and $|\phi| \leq 1$. Suppose p has finite second

moment. Fix $f \in \mathcal{F}_p$ as defined in equation 2.3.1. Then for any $\delta > 0$, with probability at least $1-\delta$ over $w_1,\dots,w_L$ drawn i.i.d. from p, there exists $b_1,\dots,b_L$ such that $\hat{f}(x) = \frac{1}{L}\sum_{\ell=1}^{L} b_\ell \phi(x,w_\ell)$ satisfies

$$\|f-\hat{f}\|_{\mathcal{X},\infty} \le \frac{\|f\|_p}{\sqrt{L}}\left(\sqrt{2\log\left(\frac{1}{\delta}\right)} + 4KBC\right),$$

where $B = \sup_{x\in X}\|x\|_2$ and $\sqrt{\mathbb{E}[\|w\|_2^2]} \le C$. Moreover, $b = (b_1,\dots,b_L)$ is such that $\|b\|_\infty \le \|f\|_p$.

Proof. Let $f(x) = \int_\Omega \alpha(w)\phi(x,w)dw$. Consider the function

$$g(w_1,\dots,w_L) = \|\hat{f}-f\|_\infty = \sup_{x\in\mathcal{X}}\left|\frac{1}{L}\sum_{\ell=1}^{L} f_\ell(x) - f(x)\right|,$$

where the functions f_ℓ are given by importance sampling $f_\ell(x) = \frac{\alpha(w_\ell)}{p(w_\ell)}\phi(x,w_\ell)$ as before. Then we observe that g satisfies the bounded differences property, since

$$\begin{aligned}
&|g(w_1,\dots,w_L) - g(w_1,\dots,\tilde{w}_i,\dots,w_L)| = |\|\hat{f}-f\|_\infty - \|\tilde{f}-f\|_\infty| \\
&\le \|\hat{f}-\tilde{f}\|_\infty = \frac{1}{L}\sup_{x\in\mathcal{X}}\left|\frac{\alpha(w_i)}{p(w_i)}\phi(x,w_i) - \frac{\alpha(\tilde{w}_i)}{p(\tilde{w}_i)}\phi(x,\tilde{w}_i)\right| \le \frac{2\|f\|_p}{L}.
\end{aligned}$$

Now, we bound its expectation using the Rademacher complexity bound of Proposition 2.1.5,

$$\mathbb{E}[g(w_1,\dots,w_L)] = \mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{L}\sum_{\ell=1}^{L} f_\ell(x) - f(x)\right|\right] \le 2\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{L}\sum_{\ell=1}^{L}\sigma_\ell f_\ell(x)\right|\right],$$

where $\sigma_1,\dots,\sigma_L$ are Rademacher random variables. Specifically, our class of functions are $\left\{w \to \frac{\alpha(w)}{p(w)}\phi(x,w) \,|\, x\in\mathcal{X}\right\}$ from Ω to $\mathbb{R}$, and the random variables are $w_1,\dots,w_L$ sampled i.i.d. from p and recall that $f(x) = \mathbb{E}[f_\ell(x)]$ so Proposition 2.1.5 applies. We will now apply Theorem 2.1.9 to bound the RHS. Fix $w_1,\dots,w_L$, and let $\varphi_\ell(t_\ell) = \frac{1}{\|f\|_p K}\frac{\alpha(w_\ell)}{p(w_\ell)}\phi(t_\ell)$, for $t\in\mathbb{R}^L$ are contractions such that $\varphi_\ell(0) = 0$. Let $T = \left\{(w_1^\top x,\dots,w_L^\top x) \,|\, x\in\mathcal{X}\right\}$ which is bounded since $\mathcal{X}$ is compact. Then, by Theorem 2.1.9 we get that

$$\mathbb{E}_\sigma\left[\frac{1}{2}\sup_{x\in\mathcal{X}}\left|\sum_{\ell=1}^{L}\sigma_\ell\varphi_\ell(w_\ell^\top x)\right|\right] \le \mathbb{E}_\sigma\left[\sup_{x\in\mathcal{X}}\left|\sum_{\ell=1}^{L}\sigma_\ell w_\ell^\top x\right|\right].$$

Multiply by $\frac{2\|f\|_p K}{L}$ and take expectation over $w_1,\dots,w_L$ to get

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{L}\sum_{\ell=1}^{L}\sigma_\ell f_\ell(x)\right|\right] \le 2\|f\|_p K\,\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{L}\sum_{\ell=1}^{L}\sigma_\ell w_\ell^\top x\right|\right].$$

Then we compute

$$\mathbb{E}[g(w_1,\dots,w_L)] \le 4K\|f\|_p \,\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{L}\sum_{\ell=1}^{L}\sigma_\ell w_\ell^\top x\right|\right] \overset{(1)}{\le} \|f\|_p \frac{4KBC}{\sqrt{L}},$$

where (1) follows from Proposition 2.1.8 where $B = \sup_{x\in\mathcal{X}}\|x\|_2$ and $\sqrt{\mathbb{E}[\|w\|_2^2]} \le C$.

Thus, by McDiarmid's inequality 2.1.1,

$$\mathbb{P}\left(\|\hat{f} - f\|_\infty \ge \|f\|_p \frac{4KBC}{\sqrt{L}} + \varepsilon\right) \le \mathbb{P}(g(w_1,\dots,w_L) \ge \mathbb{E}[g(w_1,\dots,w_L)] + \varepsilon) \le \exp\left(\frac{-\varepsilon^2 L}{2\|f\|_p^2}\right) =: \delta$$

Rearranging terms we get that $\varepsilon = \frac{\|f\|_p}{\sqrt{L}}\sqrt{2\log\left(\frac{1}{\delta}\right)}$. Thus with probability $1-\delta$,

$$\|\hat{f} - f\|_\infty \le \frac{\|f\|_p}{\sqrt{L}}\left(\sqrt{2\log\left(\frac{1}{\delta}\right)} + 4KBC\right).$$

□

However, there is different recent approach to random features, developed in Gonon et al. (2023) and Gonon (2023), which gives universal approximation and generalization results based on smoothness assumptions on the target function.

Denote by $\mathcal{W}^{k,2}(\mathbb{R}^d)$ for $k \in \mathbb{N}$ the Sobolev space of functions $u : \mathbb{R}^d \to \mathbb{R}$ whose partial derivatives $D^\alpha u$ for any multi-index $(\alpha_1,\dots,\alpha_d)$ such that $\alpha_1 + \cdots + \alpha_d \le k$, satisfies $D^\alpha u \in \mathcal{L}^2(\mathbb{R}^d)$. Recall that for $u \in \mathcal{L}^1(\mathbb{R}^d)$, $\mathcal{F}(u)(\xi) = \int_{\mathbb{R}^d} e^{-i\langle\xi,x\rangle}u(x)dx$ for $\xi \in \mathbb{R}^d$ denotes its Fourier transform. The space $\mathcal{W}^{k,2}(\mathbb{R}^d)$ can be alternatively described as the functions $u \in \mathcal{L}^2(\mathbb{R}^d)$ such that the norm

$$\|u\|_k = \left(\int_{\mathbb{R}^d} |\mathcal{F}(u)(\xi)|^2(1+\|\xi\|^2)^k d\xi\right)^{1/2}, \tag{2.3.3}$$

is finite. The following theorem gives an approximation result in the $\mathcal{L}^2$ sense, compare with Theorem 2.3.1. Informally, the theorem says that if the target function f is sufficiently smooth (i.e., k large and $\|f\|_k$ bounded), we sample from a bounded random variable X, and we choose a ReLU neural network with one hidden layer whose weights are uniform on a ball, and sufficiently large biases (not trained as well), then there are coefficients such that the target function is close on average to our random feature neural network.

Theorem 2.3.6. (Gonon et al. 2023, Corollary 2) Let $k \ge \frac{d}{2} + 1 + \varepsilon$ for some $\varepsilon > 0$ and $f \in \mathcal{W}^{k,2}(\mathbb{R}^d) \cap \mathcal{L}^1(\mathbb{R}^d)$. Assume that X is an $\mathbb{R}^d$-valued random variable with $\|X\|_2 \le M$ $\mathbb{P}$-almost surely. Let $R = L^{1/(2k-2\varepsilon+1)}$, suppose $W = (w_1,\dots,w_L)$ are i.i.d. random variables uniformly distributed over $B_R \subset \mathbb{R}^d$, suppose $\xi = (\xi_1,\dots,\xi_L)$ are i.i.d. uniformly distributed over the interval $[-\max(MR,1), \max(MR,1)]$, assume that W, ξ are independent and let

$\phi : \mathbb{R} \to \mathbb{R}$ given by $\phi(x) = \max(x, 0)$. Then, there exists a random $\sigma(W, \xi)$-measurable vector $b = (b_1, \ldots, b_L)$ and a constant $C > 0$ (depending on d and M but independent of f, L) such that

$$\mathbb{E}[\|f(X) - \hat{f}_{W,\xi}(X)\|^2]^{1/2} \leq C\|f\|_k L^{-1/\alpha},$$

with $\alpha = 2 + \frac{d+1}{k-d/2-\varepsilon}$, where $\hat{f}_{W,\xi}(x) = \frac{1}{L}\sum_{\ell=1}^{L} b_\ell \phi(w_\ell^\top x + \xi_\ell)$ and the expectation is over X, W, and ξ. Moreover, $\|b\|_\infty \leq 2\|f\|_{\mathscr{L}^1(\mathbb{R}^d)}$ and C is explicitly given by Gonon et al. (2023, Equation 43).

Remark 2.3.7. Gonon et al. (2023, Corollary 2) does not explicitely states that the vector $b = (b_1, \ldots, b_L)$ is $\sigma(W, \xi)$-measurable or the bound $\|b\|_\infty \leq 2\|f\|_{\mathscr{L}^1(\mathbb{R}^d)}$, however these two properties follow from the bound on Gonon et al. (2023, Equation 21-22). Explicitely, first b_ℓ is given by importance sampling, i.e. the Radon-Nikodym derivative $b_\ell = \frac{d\alpha}{d\pi}(w_\ell, \xi_\ell)$, thus is $\sigma(w_\ell, \xi_\ell)$-measurable. Second, Gonon et al. (2023, Equation 21) gives a bound

$$\left|\frac{d\alpha}{d\pi}(w, \xi)\right| \leq \left(\mathbb{1}_{(-M\|w\|,0)}(u) + 2\sqrt{2}\mathbb{1}_{[-1,1]}(u)\right)\frac{1}{\pi_{\mathbb{R}}(u)} g(w).$$

In this case $\pi_{\mathbb{R}}$ is the uniform density over the interval $[-\max(MR, 1), \max(MR, 1)]$, thus it is bounded by 1/2. The numerator can be bounded by $1 + 2\sqrt{2}$ times $\sup_{w \in B_R} |g(w)|$ where g is the Fourier transform of f. Now,

$$\sup_{w \in B_R} |g(w)| = \sup_{w \in \mathbb{R}^d} \left|\int_{\mathbb{R}^d} e^{-i\langle w, x\rangle} f(x)\, dx\right| \leq \int_{\mathbb{R}^d} |f(x)|\, dx = \|f\|_{\mathscr{L}^1(\mathbb{R}^d)},$$

thus the bound follows.

As a direct consequence of Theorem 2.3.6 and Markov's inequality we obtain the following bound.

Corollary 2.3.8. Under the same assumptions as in Theorem 2.3.6, with probability at least $1 - \delta$ over $W = (w_1, \ldots, w_L)$ and $\xi = (\xi_1, \ldots, \xi_L)$, the random feature neural network satisfies the inequality

$$\|f - \hat{f}_{W,\xi}\|_{\mathscr{L}^2(\mathbb{R}^d, \mu_X)} = \left(\int_{\mathbb{R}^d} (f(x) - \hat{f}_{W,\xi}(x))^2 \mu_X(dx)\right)^{1/2} \leq \frac{C\|f\|_k}{L^{1/\alpha}\sqrt{\delta}}.$$

Proof. Simply apply Markov's inequality to the random variable $\|f - \hat{f}_{W,\xi}\|^2_{\mathscr{L}^2(\mathbb{R}^d, \mu_X)}$ and use

Theorem 2.3.6 to bound it's expectation as follows

$$\mathbb{P}\left[\|f-\hat{f}_{W,\xi}\|_{\mathscr{L}^2(\mathbb{R}^d,\mu_X)}>\varepsilon\right]=\mathbb{P}\left[\int_{\mathbb{R}^d}(f(x)-\hat{f}_{W,\xi}(x))^2\,\mu_X(dx)>\varepsilon^2\right]\leq\left(\frac{C\|f\|_k}{L^{1/\alpha}\varepsilon}\right)^2=:\delta.$$

Rearranging we get that $\varepsilon=\frac{C\|f\|_k}{L^{1/\alpha}\sqrt{\delta}}$, and thus with probability at least $1-\delta$,

$$\|f-\hat{f}_{W,\xi}\|_{\mathscr{L}^2(\mathbb{R}^d,\mu_X)}\leq\frac{C\|f\|_k}{L^{1/\alpha}\sqrt{\delta}}.$$

□

Chapter 3

Random features SDF

3.1 Approximation bounds

Suppose we sample random features $\phi(z, w_\ell)$ for $w_1, \dots, w_L$ drawn i.i.d. from the distribution p. Define the associated random factors as

$$F_{t+1,\ell} = \frac{1}{LN_t} \sum_{i=1}^{N_t} \phi(Z_{t,i}, w_\ell) R^e_{t+1,i}, \tag{3.1.1}$$

for $\ell = 1, \dots, L$. Call F_{t+1} the $L \times 1$ vector of factors. Notice that solving for $M_{t+1} = 1 - b^\top F_{t+1}$ such that $\mathbb{E}[M_{t+1} F_{t+1}] = 0$ is equivalent to solve the minimization

$$\hat{b} = \arg\min_b \mathbb{E}\left[\|1 - b^\top F_{t+1}\|^2\right]. \tag{3.1.2}$$

We now argue why minimising this quantity might make sense in order to find an "approximate SDF". Suppose

$$M^*_{t+1} = 1 - F^*_{t+1} = 1 - (\omega^*_t)^\top R^e_{t+1},$$

is the true (tradable) SDF, i.e. $\omega^*_t \in m\mathscr{F}_t$ and $\mathbb{E}_t[M^*_{t+1} R^e_{t+1,i}] = 0$ for all i and t. Let $M_{t+1} = 1 - b^\top F_{t+1}$ for some coefficients $b \in \mathbb{R}^L$. Then,

$$\begin{aligned} \mathbb{E}\left[\|M_{t+1}\|^2\right] &= \mathbb{E}\left[\left\|M_{t+1} - M^*_{t+1} + M^*_{t+1}\right\|^2\right] \\ &= \mathbb{E}\left[\left\|M^*_{t+1}\right\|^2\right] + \mathbb{E}\left[\left\|M_{t+1} - M^*_{t+1}\right\|^2\right] + 2\mathbb{E}\left[M^*_{t+1}(M_{t+1} - M^*_{t+1})\right] \\ &= \mathbb{E}\left[\left\|M^*_{t+1}\right\|^2\right] + \mathbb{E}\left[\left\|M_{t+1} - M^*_{t+1}\right\|^2\right], \end{aligned} \tag{3.1.3}$$

where in the last step we observe that $\mathbb{E}\left[M^*_{t+1}(M_{t+1} - M^*_{t+1})\right] = 0$ as

$$M_{t+1} - M^*_{t+1} = (\omega_t - \omega^*_t)^\top R^e_{t+1},$$

where $\omega_{t,i} = \frac{1}{LN_t}\sum_{\ell=1}^{L} b_\ell \phi(Z_{t,i}, w_\ell) \in m\mathcal{F}_t$ are the asset-specific portfolio weights of the approximate SDF, thus is an excess return and hence is orthogonal to the true SDF. It follows from this computation that minimising equation 3.1.2 is equivalent to minimising the distance to the true SDF M^*_{t+1}.

Now we state the assumptions we need in order to get effective learning error bounds. First we specify the properties of the feature map ϕ and of the random weights $w \sim p$. These determine the kernel k defined in equation 2.3.2, its associated RKHS $\mathcal{H}$ and the normed space $\mathcal{F}_p$ defined in equation 2.3.1.

Assumption 3.1.1. There is a K-Lipschitz function $\phi : \mathbb{R} \to \mathbb{R}$ such that $\phi(0) = 0$ and $|\phi| \leq 1$. Let the associated feature map be $\phi(z, w) = \phi(w^\top z)$ for $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ compact and $w \in \Omega \subseteq \mathbb{R}^d$. Let p be a distribution over Ω. We assume p to have finite second moments, i.e. if $w \sim p$ then $\sqrt{\mathbb{E}[\|w\|_2^2]} \leq C$.

Now, in order to have an approximate SDF given by our random factors we assume that the true SDF weights are a function of the characteristics Z_t, and moreover this function lives in the space $\mathcal{F}_p$.

Assumption 3.1.2. We assume there exists a function $f^* \in \mathcal{F}_p$ (i.e., there is $\alpha : \Omega \to \mathbb{R}$ such that $f^*(z) = \int_\Omega \alpha(w)\phi(z, w)dw$ and such that $\|f\|_p < \infty$), such that

$$M^*_{t+1} = 1 - F^*_{t+1} = 1 - \frac{1}{N_t} f^*(Z_t)^\top R^e_{t+1},$$

is the true SDF, i.e. $\mathbb{E}_t[M^*_{t+1} R^e_{t+1,i}] = 0$ for all i and t.

Notice that $f^*(Z_t) \in \mathbb{R}^{N_t}$ denotes the vector of applying f^* pointwise to each asset, i.e. $f^*(Z_t) = (f^*(Z_{t,1}), \ldots, f^*(Z_{t,N_t}))$. A larger (resp. lower) than one value of $f^*(Z_{t,i})$ means we overweight (resp. underweight) this asset with respect to an equally-weighted portfolio.

Lastly, we assume that returns are bounded so as to be able to apply the generalization bounds of Theorems 2.1.7 and 2.2.5. This is consistent with our winsorizing of returns as explained in section 4.1.

Assumption 3.1.3. Assume that excess returns are bounded $R^e_{t+1,i} \in [-1, 1]$ for all i, t.

Remark 3.1.4. This assumption can be relaxed to a bound like $R^e_{t+1,i} \in [-1, R_{\max}]$ for some constant $R_{\max}$, only at the cost of keeping track of such constant. Moreover, the boundedness of returns is not essential for our approximation results, for this it is enough to have returns with bounded second moments. It will, however, be essential for our generalization bounds, otherwise the theory becomes more involved.

We will then make use of the approximation theorems of Rahimi and Recht (Theorem 2.3.1 or Theorem 2.3.5) to bound the distance from our approximate SDF to the true SDF. How restrictive is the assumption that $f^* \in \mathcal{F}_p$ will depend on the kernel chosen, but recall that $\mathcal{F}_p$ is dense in the RKHS $\mathcal{H}$ by Proposition 2.3.3, and for a universal kernel $\mathcal{H}$ might be dense in the space of continuous functions, thus it might not be very restrictive after all.

Proposition 3.1.5. Suppose that Assumptions 3.1.1, 3.1.2 and 3.1.3 hold. Suppose that the weights $w_1,\dots,w_L$ are sampled i.i.d. from p. Then for any $\delta > 0$, with probability at least $1-\delta$ over $w_1,\dots,w_L$, there exists $b_1,\dots,b_L$ such that $\hat{f}(z) = \frac{1}{L}\sum_{\ell=1}^{L} b_\ell \phi(z,w_\ell)$ and the approximate SDF $M_{t+1} = 1 - \frac{1}{N_t}\hat{f}(Z_t)^\top R^e_{t+1}$ satisfies

$$\mathbb{E}\left[\|M_{t+1} - M^*_{t+1}\|^2\right] = \mathbb{E}\left[\|M_{t+1}\|^2\right] - \mathbb{E}\left[\|M^*_{t+1}\|^2\right] \leq \frac{\|f^*\|_p^2}{L}\left(\sqrt{2\log\left(\frac{1}{\delta}\right)} + 4KRC\right)^2,$$

where $R = \sup_{z\in\mathcal{Z}} \|z\|_2$, $\sqrt{\mathbb{E}[\|w\|_2^2]} \leq C$ and K is the Lipschitz constant of ϕ. Moreover $\|b\|_\infty \leq \|f^*\|_p$.

Proof. By our previous computation in equation 3.1.3,

$$\begin{aligned}\mathbb{E}\left[\|M_{t+1}\|^2\right] &= \mathbb{E}\left[\left\|M^*_{t+1}\right\|^2\right] + \mathbb{E}\left[\left\|M_{t+1} - M^*_{t+1}\right\|^2\right] \\ &= \mathbb{E}\left[\|M^*_{t+1}\|^2\right] + \frac{1}{N_t^2}\mathbb{E}\left[\left\|(f^*(Z_t) - \hat{f}(Z_t))^\top R^e_{t+1}\right\|^2\right],\end{aligned}$$

Now we bound this second term which is the distance between our approximate SDF and the true SDF. Using Hölder's inequality we get the bound

$$\frac{1}{N_t^2}\mathbb{E}\left[\left\|(f^*(Z_t) - \hat{f}(Z_t))^\top R^e_{t+1}\right\|^2\right] \leq \frac{1}{N_t^2}\mathbb{E}\left[\left\|f^*(Z_t) - \hat{f}(Z_t)\right\|_\infty^2\right]\mathbb{E}\left[\|R^e_{t+1}\|_1^2\right].$$

Then, we observe that

$$\mathbb{E}\left[\left\|f^*(Z_t) - \hat{f}(Z_t)\right\|_\infty^2\right] = \mathbb{E}\left[\max_{i=1,\dots,N_t}\left(f^*(Z_{t,i}) - \hat{f}(Z_{t,i})\right)^2\right] \leq \left\|f^* - \hat{f}\right\|^2_{\mathcal{Z},\infty},$$

where $\mathcal{Z} \subset \mathbb{R}^d$ is compact. Thus, we can apply Theorem 2.3.5 to bound this quantity. Explicitely, with probability at least $1-\delta$ over $w_1,\dots,w_L$ there exists $b_1,\dots,b_L$ such that the following inequality holds

$$\|f^* - \hat{f}\|_{\mathcal{Z},\infty} \leq \frac{\|f^*\|_p}{\sqrt{L}}\left(\sqrt{2\log\left(\frac{1}{\delta}\right)} + 4KRC\right), \tag{3.1.4}$$

where $R = \sup_{z \in \mathcal{Z}} \|z\|_2$ and $\sqrt{\mathbb{E}[\|w\|_2^2]} \leq C$ and $\phi : \mathbb{R} \to \mathbb{R}$ is K-Lipschitz. Moreover, b is such that $\|b\|_\infty \leq \|f^*\|_p$.

Since $R^e_{t+1,i} \in [-1,1]$, then

$$\mathbb{E}\left[\|R^e_{t+1}\|_1^2\right] \leq N_t^2 \mathbb{E}\left[\|R^e_{t+1}\|_\infty^2\right] \leq N_t^2. \tag{3.1.5}$$

Thus, combining the bounds from equations 3.1.4 and 3.1.5 we get that with probability $1-\delta$ over $w_1,\ldots,w_L$ drawn i.i.d. from p, there exists $b_1,\ldots,b_L$ such that

$$\mathbb{E}\left[\|M_{t+1}\|^2\right] \leq \mathbb{E}\left[\|M^*_{t+1}\|^2\right] + \frac{\|f^*\|_p^2}{L}\left(\sqrt{2\log\left(\frac{1}{\delta}\right)} + 4KRC\right)^2 = \mathbb{E}\left[\|M^*_{t+1}\|^2\right] + O\left(\frac{\|f^*\|_p^2}{L}\right).$$

□

Alternatively, we can deduce a similar approximation result from Theorem 2.3.1. Moreover, Assumption 3.1.1 can be substantially relaxed.

Proposition 3.1.6. Assume the feature map $\phi : \mathcal{Z} \times \Omega \to \mathbb{R}$ satisfies $|\phi(z,w)| \leq 1$ for all $(z,w) \in \mathcal{Z} \times \Omega$. Suppose that Assumptions 3.1.2 and 3.1.3 hold. Suppose $w_1,\ldots,w_L$ are sampled i.i.d. from p. Then for any $\delta > 0$, with probability at least $1-\delta$ over $w_1,\ldots,w_L$, there exists $b_1,\ldots,b_L$ such that $\hat{f}(z) = \frac{1}{L}\sum_{\ell=1}^{L} b_\ell \phi(z,w_\ell)$ and the approximate SDF $M_{t+1} = 1 - \frac{1}{N_t}\hat{f}(Z_t)^\top R^e_{t+1}$ satisfies

$$\mathbb{E}\left[\|M_{t+1} - M^*_{t+1}\|^2\right] = \mathbb{E}\left[\|M_{t+1}\|^2\right] - \mathbb{E}\left[\|M^*_{t+1}\|^2\right] \leq \frac{\|f^*\|_p^2}{L}\left(1 + \sqrt{2\log\left(\frac{N_t}{\delta}\right)}\right)^2.$$

Moreover $\|b\|_\infty \leq \|f^*\|_p$.

Proof. Instead of Hölder's inequality we can apply Cauchy-Schwarz to bound the distance $\mathbb{E}\left[\|M_{t+1} - M^*_{t+1}\|^2\right]$, thus we obtain the following inequality

$$\frac{1}{N_t^2}\mathbb{E}\left[\left\|(f^*(Z_t) - \hat{f}(Z_t))^\top R^e_{t+1}\right\|^2\right] \leq \frac{1}{N_t^2}\mathbb{E}\left[\left\|f^*(Z_t) - \hat{f}(Z_t)\right\|_2^2\right]\mathbb{E}\left[\|R^e_{t+1}\|_2^2\right].$$

First observe that

$$\mathbb{E}\left[\|R^e_{t+1}\|_2^2\right] \leq N_t \mathbb{E}\left[\|R^e_{t+1}\|_\infty^2\right] \leq N_t. \tag{3.1.6}$$

Moreover,

$$\mathbb{E}\left[\left\|f^*(Z_t) - \hat{f}(Z_t)\right\|_2^2\right] = \sum_{i=1}^{N_t} \mathbb{E}\left[\left(f^*(Z_{t,i}) - \hat{f}(Z_{t,i})\right)^2\right] = \sum_{i=1}^{N_t} \|f^* - \hat{f}\|^2_{\mathscr{L}^2(\mathcal{Z},\mu_{t,i})},$$

where $\mu_{t,i}$ is the distribution of $Z_{t,i}$. By the bound on Theorem 2.3.1, and using a union bound we get that

$$\mathbb{P}\left[\bigcup_{i=1}^{N_t}\left\{\|f^*-\hat{f}\|_{\mathscr{L}^2(\mathcal{Z},\mu_{t,i})}\geq\frac{\|f^*\|_p}{\sqrt{L}}+\varepsilon\right\}\right]\leq N_t\exp\left(\frac{-\varepsilon^2 L}{2\|f^*\|_p^2}\right)=:\delta.$$

Rearranging terms we get that $\varepsilon=\frac{\|f^*\|_p}{\sqrt{L}}\sqrt{2\log\left(\frac{N_t}{\delta}\right)}$. Thus with probability at least $1-\delta$ over $w_1,\ldots,w_L$ we have that

$$\mathbb{E}\left[\left\|f^*(Z_t)-\hat{f}(Z_t)\right\|_2^2\right]=\sum_{i=1}^{N_t}\|f^*-\hat{f}\|^2_{\mathscr{L}^2(\mathcal{Z},\mu_{t,i})}\leq\frac{N_t\|f^*\|_p^2}{L}\left(1+\sqrt{2\log\left(\frac{N_t}{\delta}\right)}\right)^2,\tag{3.1.7}$$

for all $i=1,\ldots,N_t$. Thus combining the bounds from equation 3.1.6 and 3.1.7 we get that

$$\frac{1}{N_t^2}\mathbb{E}\left[\left\|(f^*(Z_t)-\hat{f}(Z_t))^\top R^e_{t+1}\right\|^2\right]\leq\frac{\|f^*\|_p^2}{L}\left(1+\sqrt{2\log\left(\frac{N_t}{\delta}\right)}\right)^2,$$

with probability at least $1-\delta$ over $w_1,\ldots,w_L$, which proves the result. □

As an immediate consequence of Propositions 3.1.5 and 3.1.6 we can take minimum of both constants and obtain the following result.

Corollary 3.1.7. Suppose that Assumptions 3.1.1, 3.1.2 and 3.1.3 hold. Suppose $w_1,\ldots,w_L$ are sampled i.i.d. from p. Then for any $\delta>0$, with probability at least $1-\delta$ over $w_1,\ldots,w_L$, there exists $b_1,\ldots,b_L$ such that $\hat{f}(z)=\frac{1}{L}\sum_{\ell=1}^{L}b_\ell\phi(z,w_\ell)$ and the approximate SDF $M_{t+1}=1-\frac{1}{N_t}\hat{f}(Z_t)^\top R^e_{t+1}$ satisfies

$$\mathbb{E}\left[\|M_{t+1}-M^*_{t+1}\|^2\right]=\mathbb{E}\left[\|M_{t+1}\|^2\right]-\mathbb{E}\left[\|M^*_{t+1}\|^2\right]\leq\frac{\|f^*\|_p^2 H}{L}$$

where H is the constant

$$H=\min\left\{\left(\sqrt{2\log\left(\frac{1}{\delta}\right)}+4KRC\right),\left(1+\sqrt{2\log\left(\frac{N_t}{\delta}\right)}\right)\right\}^2,$$

and $R=\sup_{z\in\mathcal{Z}}\|z\|_2$, $\sqrt{\mathbb{E}[\|w\|_2^2]}\leq C$ and K is the Lipschitz constant of ϕ. Moreover $\|b\|_\infty\leq\|f^*\|_p$.

Furthermore, as another option, we can deduce an approximation bound from the results of Gonon et al. (2023), see Theorem 2.3.6 and Corollary 2.3.8, which are based on smoothness and integrability conditions on the target function.

Proposition 3.1.8. Let $k \geq \frac{d}{2} + 1 + \varepsilon$ for some $\varepsilon > 0$ and $f^* \in \mathscr{W}^{k,2}(\mathbb{R}^d) \cap \mathscr{L}^1(\mathbb{R}^d)$. Assume that $Z_{t,i}$ is an $\mathbb{R}^d$-valued random variable with $\|Z_{t,i}\|_2 \leq K$ $\mathbb{P}$-almost surely for all assets $i = 1, \ldots, N_t$ for some constant $K > 0$. Assume that

$$M^*_{t+1} = 1 - F^*_{t+1} = 1 - \frac{1}{N_t} f^*(Z_t)^\top R^e_{t+1},$$

is the true SDF. Suppose $\mathbb{E}[\|R^e_{t+1}\|^2] \leq D$. Let $R = L^{1/(2k-2\varepsilon+1)}$, let $W = (w_1, \ldots, w_L)$ be i.i.d. random variables uniformly distributed over $B_R \subset \mathbb{R}^d$ and $\xi = (\xi_1, \ldots, \xi_L)$ be i.i.d. uniformly distributed over $[-\max(KR, 1), \max(KR, 1)]$. Assume that W, ξ are independent and let $\phi : \mathbb{R} \to \mathbb{R}$ given by $\phi(x) = \max(x, 0)$. Then for any $\delta > 0$, with probability at least $1 - \delta$ over W, ξ, there exists a vector $b = (b_1, \ldots, b_L)$ such that $\hat{f}_{W,\xi}(x) = \frac{1}{L} \sum_{\ell=1}^{L} b_\ell \phi(w_\ell^\top x + \xi_\ell)$ and the approximate SDF $M_{t+1} = 1 - \frac{1}{N_t} \hat{f}(Z_t)^\top R^e_{t+1}$ satisfies

$$\mathbb{E}\left[\|M_{t+1} - M^*_{t+1}\|^2\right] = \mathbb{E}\left[\|M_{t+1}\|^2\right] - \mathbb{E}\left[\|M^*_{t+1}\|^2\right] \leq \frac{C^2 \|f^*\|_k^2 D}{L^{2/\alpha}\delta},$$

where $\alpha = 2 + \frac{d+1}{k - d/2 - \varepsilon}$, C is a constant given by Gonon et al. (2023, Equation 43) and $\|f^*\|_k$ denotes the norm of the Sobolev space $\mathscr{W}^{k,2}(\mathbb{R}^d)$ defined in equation 2.3.3. Moreover, $\|b\|_\infty \leq 2\|f\|_{\mathscr{L}^1(\mathbb{R}^d)}$.

Proof. The proof follows the same strategy of that of Proposition 3.1.6, use Cauchy-Schwarz inequality and a union bound combined with Corollary 2.3.8. By a union bound on Corollary 2.3.8,

$$\mathbb{P}\left[\bigcup_{i=1}^{N_t} \left\{\|f^* - \hat{f}_{W,\xi}\|_{\mathscr{L}^2(\mathbb{R}^d, \mu_{Z_{t,i}})} \geq \varepsilon\right\}\right] \leq N_t \left(\frac{C\|f^*\|_k}{L^{1/\alpha}\varepsilon}\right)^2 =: \delta.$$

Rearranging we get that $\varepsilon = \frac{C\|f^*\|_k \sqrt{N_t}}{L^{1/\alpha}\sqrt{\delta}}$. Thus, with probability at least $1 - \delta$ over W, ξ we have that

$$\mathbb{E}\left[\left\|f^*(Z_t) - \hat{f}_{W,\xi}(Z_t)\right\|_2^2\right] = \sum_{i=1}^{N_t} \|f^* - \hat{f}_{W,\xi}\|^2_{\mathscr{L}^2(\mathbb{R}^d, \mu_{Z_{t,i}})} \leq \frac{C^2 \|f^*\|_k^2 N_t^2}{L^{2/\alpha}\delta}.$$

The bound on $\mathbb{E}\left[\|M_{t+1} - M^*_{t+1}\|^2\right]$ then follows by Cauchy-Schwarz. □

3.2 Generalization bounds

Now let us consider the estimation problem. We assume that the factors $(F_t)_t$ defined in equation 3.1.1 are i.i.d. to simplify our analysis first, and later consider less restrictive

assumptions. We want to bound the difference

$$\mathbb{E}\left[\left\|1-b^\top F_{t+1}\right\|^2\right]-\frac{1}{T}\sum_{t=1}^{T}(1-b^\top F_{t+1})^2,$$

for $\|b\|_2 \le B$. This follows since the ridge penalised regression of equation 1.3.1 is equivalent to minimizing

$$\frac{1}{T}\sum_{t=1}^{T}(1-b^\top F_{t+1})^2=\frac{1}{T}\left\|\mathbb{1}-F^\top b\right\|^2,$$

restricted to $\|b\|_2 \le B$ for some $B=B(\gamma)$. Explicitely, the equivalence is as follows. Given B, the solution to the bounded regression problem coincides with the unconstrained solution of minimal norm if such $\hat{b}$ satisfies $\|b\|_2 \le B$, otherwise it is given by

$$\hat{b}=\left(\frac{1}{T}FF^\top+\gamma I\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}F_t,$$

where γ is a non-negative $\sigma(w_1,\dots,w_L)$-measurable random variable such that $\|\hat{b}\|_2 = B$. This means that once the random weights have been chosen, γ is simply a constant.

To simplify our analysis let us introduce the following notations,

$$\mathscr{L}(b)=\mathbb{E}\left[\left\|1-b^\top F_{t+1}\right\|^2\right],\quad \hat{\mathscr{L}}(b)=\frac{1}{T}\sum_{t=1}^{T}(1-b^\top F_{t+1})^2 \tag{3.2.1}$$

for the expected loss and the empirical loss respectively. We observe that to get a good approximation bound in Propositions 3.1.5 and 3.1.6, we can restrict b to be in the ball $\|b\|_\infty \le \|f^*\|_p$. Thus, if we choose $B=\sqrt{L}\|f^*\|_p$ or larger, then there is an approximate SDF close to the true SDF M^*_{t+1}, as

$$\|b\|_2 \le \sqrt{L}\|b\|_\infty \le \sqrt{L}\|f^*\|_p = B. \tag{3.2.2}$$

Alternatively, under the assumptions of Proposition 3.1.8, we have that an approximate SDF might be found with highp probability for $\|b\|_2 \le B = 2\sqrt{L}\|f^*\|_{\mathscr{L}^1(\mathbb{R}^d)}$.

Assuming that the factors $F_2,\dots,F_{T+1}$ are i.i.d. over time, the generalization error can be bounded as in a bounded linear regression and the Rademacher complexity can be bounded for using Proposition 2.1.8.

Proposition 3.2.1. Suppose that Assumption 3.1.3 holds. Assume the factors $F_2,\dots,F_{T+1}$ are independent and identically distributed over time. Then with probability at least $1-\delta$

over $F_2, \ldots, F_{T+1}$ we have that

$$\sup_{\|b\| \leq B} \left| \mathscr{L}(b) - \hat{\mathscr{L}}(b) \right| \leq M^2 \left(\frac{8}{\sqrt{T}} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2T}} \right) = O\left(\frac{1}{\sqrt{T}} \right),$$

where $M = 1 + \frac{B}{\sqrt{L}}$.

Proof. First we bound the norm of the factors by Hölder's inequality,

$$|F_{t+1,\ell}| = \frac{1}{LN_t} \left| \sum_{i=1}^{N_t} \phi(Z_{t,i}, w_\ell) R^e_{t+1,i} \right| \leq \frac{1}{LN_t} \left\| \phi(Z_t, w_\ell) \right\|_1 \left\| R^e_{t+1} \right\|_\infty \leq \frac{1}{L}.$$

Then, we have a bound

$$|1 - b^\top F_{t+1}| \leq 1 + \|b\|_2 \|F_{t+1}\|_2 \leq 1 + B\sqrt{L} \frac{1}{L} \leq 1 + \frac{B}{\sqrt{L}} = M.$$

Now, we observe that the function $y \to (1-y)^2$ is $2M$-Lipschitz for y such that $|1-y| \leq M$. Thus, by Theorem 2.1.9 and Proposition 2.1.8, we can bound the Rademacher complexity as follows

$$\mathbb{E}\left[\sup_{\|b\| \leq B} \left| \frac{2}{T} \sum_{t=1}^{T} \sigma_{t+1} (1 - b^\top F_{t+1})^2 \right| \right] \leq 4M \, \mathbb{E}\left[\sup_{\|b\| \leq B} \left| \frac{2}{T} \sum_{t=1}^{T} \sigma_{t+1} b^\top F_{t+1} \right| \right] \leq \frac{8MB \sqrt{\mathbb{E}\left[\|F_{t+1}\|^2 \right]}}{\sqrt{T}}.$$

Now, $\sqrt{\mathbb{E}\left[\|F_{t+1}\|^2 \right]} \leq \frac{1}{\sqrt{L}}$, thus, putting all together we obtain the following bound for the Rademacher complexity,

$$\mathbb{E}\left[\sup_{\|b\| \leq B} \left| \frac{2}{T} \sum_{t=1}^{T} \sigma_{t+1} (1 - b^\top F_{t+1})^2 \right| \right] \leq \frac{8MB}{\sqrt{TL}} \leq \frac{8M^2}{\sqrt{T}} = O\left(\frac{1}{\sqrt{T}} \right).$$

Thus, we get a generalization bound by means of Theorem 2.1.7. That is, with probability at least $1 - \delta$ over $F_2, \ldots, F_{T+1}$ we have that

$$\sup_{\|b\| \leq B} \left| \mathscr{L}(b) - \hat{\mathscr{L}}(b) \right| \leq M^2 \left(\frac{8}{\sqrt{T}} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2T}} \right) = O\left(\frac{1}{\sqrt{T}} \right).$$

□

We can relax the i.i.d. assumption on the factors $F_2, \ldots, F_{T+1}$ to that of a stationary β-mixing sequence, and instead apply Theorem 2.2.5 to get a generalization bound.

Proposition 3.2.2. Suppose that Assumption 3.1.3 holds. Assume the factors $F_2, \ldots, F_{T+1}$ are a sample of a stationary β-mixing sequence. Let $T = 2\mu a$ for $\mu, a > 0$. Then with proba-

bility at least $1-\delta$ over $F_2,\dots,F_{T+1}$ we have that

$$\sup_{\|b\|\le B}\left|\mathscr{L}(b)-\hat{\mathscr{L}}(b)\right| \le M^2\left(\frac{8}{\sqrt{\mu}}+\sqrt{\frac{\log\left(\frac{4}{\delta'}\right)}{2\mu}}\right)=O\left(\frac{1}{\sqrt{\mu}}\right),$$

where $M=1+\frac{B}{\sqrt{L}}$ and $\delta'=\delta-4(\mu-1)\beta(a)$. Moreover, with probability at least $1-\delta$,

$$\sup_{\|b\|\le B}\left|\mathscr{L}(b)-\hat{\mathscr{L}}(b)\right| \le \hat{\mathfrak{R}}_{S_\mu}(\mathcal{F})+3M^2\sqrt{\frac{\log\left(\frac{8}{\delta'}\right)}{2\mu}},$$

where $\delta'=\delta-8(\mu-1)\beta(a)$ and S_μ is a subsample of μ points separated by a gap of $2a-1$ points.

Proof. We apply the generalization bound of Theorem 2.2.5. The Rademacher complexity $\mathfrak{R}_\mu^{\tilde{D}}(\mathcal{F})$ can be bounded as in Proposition 3.2.1, only that now our i.i.d. sample is of size μ, thus for $\mathcal{F}=\{x\to(1-b^\top x)^2 \mid \|b\|\le B\}$ we have that $\mathfrak{R}_\mu^{\tilde{D}}(\mathcal{F})\le\frac{8M^2}{\sqrt{\mu}}$. Then by Theorem 2.2.5,

$$\mathbb{P}\left[\sup_{\|b\|\le B}\frac{1}{M^2}\left(\mathscr{L}(b)-\hat{\mathscr{L}}(b)\right)\ge\frac{8}{\sqrt{\mu}}+\varepsilon\right]\le\exp\left(-2\mu\varepsilon^2\right)+2(\mu-1)\beta(a)$$

A similar bound is obtained for the class $-\mathcal{F}$ (i.e., minus the squared loss). Setting the right-hand side to be $\frac{\delta}{2}$, rearring we get that $\varepsilon=\sqrt{\frac{\log\left(\frac{4}{\delta'}\right)}{2\mu}}$ where $\delta'=\delta-4(\mu-1)\beta(a)$, using a union bound we have that with probability at least $1-\delta$, for all $\|b\|\le B$

$$\left|\mathscr{L}(b)-\hat{\mathscr{L}}(b)\right| \le M^2\left(\frac{8}{\sqrt{\mu}}+\sqrt{\frac{\log\left(\frac{4}{\delta'}\right)}{2\mu}}\right),$$

and thus we get the first inequality. The second inequality follows in a similar manner. □

3.3 Learning error bounds

Let us introduce some extra notation to that of equation 3.2.1, let us denote by $\mathscr{L}^*=\mathbb{E}\left[\left\|1-F_{t+1}^*\right\|^2\right]$ the global minimum of the expected loss, i.e., the expected squared norm of the true SDF. Denote $\hat{b}$ the empirical loss minimiser, in other words $\hat{b}=\underset{\|b\|\le B}{\arg\min}\,\hat{\mathscr{L}}(b)$, and denote by $b^*=\underset{\|b\|\le B}{\arg\min}\,\mathscr{L}(b)$ the expected loss minimiser within our model class. Then,

$$\begin{aligned}\mathscr{L}(\hat{b})-\mathscr{L}^* &= \left(\mathscr{L}(\hat{b})-\hat{\mathscr{L}}(\hat{b})\right)+\left(\hat{\mathscr{L}}(\hat{b})-\hat{\mathscr{L}}(b^*)\right)+\left(\hat{\mathscr{L}}(b^*)-\mathscr{L}(b^*)\right)+\left(\mathscr{L}(b^*)-\mathscr{L}^*\right)\\ &\le 2\sup_{\|b\|\le B}\left|\mathscr{L}(b)-\hat{\mathscr{L}}(b)\right|+\mathscr{L}(b^*)-\mathscr{L}^*.\end{aligned}\tag{3.3.1}$$

Thus, depending on the assumptions we make, by applying Propositions 3.2.1 or 3.2.2 we can bound the first term, and with Propositions 3.1.5 or 3.1.6 we can bound the last term, with probability $1-\delta$ over the factors $F_2,\dots,F_{T+1}$ and the weights $w_1,\dots,w_L$. Thus, we have the following result in the case of i.i.d. factors.

Proposition 3.3.1. Suppose that Assumptions 3.1.1, 3.1.2 and 3.1.3 hold. Suppose that $w_1,\dots,w_L$ are sampled i.i.d. from p. Assume the factors $F_2,\dots,F_{T+1}$ are independent and identically distributed over time. Let $B = \lambda\sqrt{L}\|f^*\|_p$ for some constant $\lambda \geq 1$. Then, with probability at least $1-\delta$ over the factors $F_2,\dots,F_{T+1}$ and the weights $w_1,\dots,w_L$ we have that

$$\mathbb{E}\left[\|\hat{M}_{t+1} - M^*_{t+1}\|^2\right] = \mathbb{E}\left[\|\hat{M}_{t+1}\|^2\right] - \mathbb{E}\left[\|M^*_{t+1}\|^2\right] \leq M^2\left(\frac{C_1}{\sqrt{T}} + \frac{C_2}{L}\right),$$

where $\hat{M}_{t+1} = 1-\hat{b}^\top F_{t+1}$ is the approximate SDF which solves $\min_{\|b\|\leq B} \frac{1}{T}\sum_{t=1}^{T}(1-b^\top F_{t+1})^2$, M^*_{t+1} is the true SDF, $M = 1+\lambda\|f^*\|_p$ and C_1, C_2 are the constants

$$C_1 = \left(16 + \sqrt{2\log\left(\frac{2}{\delta}\right)}\right), \qquad C_2 = \left(\sqrt{2\log\left(\frac{2}{\delta}\right)} + 4KRC\right)^2,$$

where $R = \sup_{z\in\mathcal{Z}}\|z\|_2$, $\sqrt{\mathbb{E}[\|w\|_2^2]} \leq C$ and K is the Lipschitz constant of ϕ.

Proof. We bound $\mathscr{L}(\hat{b}) - \mathscr{L}^*$ by means of equation 3.3.1 and apply a union bound on Propositions 3.2.1 and 3.1.5. Specifically, with probability at least $1-\frac{\delta}{2}$ over $F_2,\dots,F_{T+1}$ we have that

$$2\sup_{\|b\|\leq B}\left|\mathscr{L}(b) - \hat{\mathscr{L}}(b)\right| \leq \frac{M^2}{\sqrt{T}}\left(16 + \sqrt{2\log\left(\frac{2}{\delta}\right)}\right),$$

because of Proposition 3.2.1. Since $M = 1+\lambda\|f^*\|_p \geq \|f^*\|_p$, the bound of equation 3.2.2, and since b^* denotes the expected loss minimiser, with probability at least $1-\frac{\delta}{2}$ over $w_1,\dots,w_L$ we have that

$$\mathscr{L}(b^*) - \mathscr{L}^* \leq \frac{M^2}{L}\left(\sqrt{2\log\left(\frac{2}{\delta}\right)} + 4KRC\right)^2,$$

by Proposition 3.1.5. Thus, with probability at least $1-\delta$ we have the bound

$$\mathscr{L}(\hat{b}) - \mathscr{L}^* \leq M^2\left(\frac{C_1}{\sqrt{T}} + \frac{C_2}{L}\right).$$

□

By the same means, using Propositions 3.1.5 and 3.2.2 we obtain one of our main results.

Theorem 3.3.2. Suppose that Assumptions 3.1.1, 3.1.2 and 3.1.3 hold. Suppose $w_1,\dots,w_L$ are sampled i.i.d. from p. Assume the factors $F_2,\dots,F_{T+1}$ are a sample of a stationary

β-mixing sequence. Let $T = 2\mu a$ for $\mu, a > 0$. Let $B = \lambda\sqrt{L}\|f^*\|_p$ for some constant $\lambda \geq 1$. Then, with probability at least $1-\delta$ over the factors $F_2, \ldots, F_{T+1}$ and the weights $w_1, \ldots, w_L$ we have that

$$\mathbb{E}\left[\|\hat{M}_{t+1} - M^*_{t+1}\|^2\right] = \mathbb{E}\left[\|\hat{M}_{t+1}\|^2\right] - \mathbb{E}\left[\|M^*_{t+1}\|^2\right] \leq M^2\left(\frac{C_1}{\sqrt{\mu}} + \frac{C_2}{L}\right),$$

where $\hat{M}_{t+1} = 1 - \hat{b}^\top F_{t+1}$ is the approximate SDF which solves $\min_{\|b\| \leq B} \frac{1}{T}\sum_{t=1}^{T}(1 - b^\top F_{t+1})^2$, M^*_{t+1} is the true SDF, $M = 1 + \lambda\|f\|_p$ and C_1, C_2 are the constants

$$C_1 = \left(16 + \sqrt{2\log\left(\frac{8}{\delta'}\right)}\right), \qquad C_2 = \left(\sqrt{2\log\left(\frac{2}{\delta}\right)} + 4KRC\right)^2,$$

where $\delta' = \delta - 4(\mu - 1)\beta(a)$, $R = \sup_{z \in \mathcal{Z}} \|z\|_2$, $\sqrt{\mathbb{E}[\|w\|_2^2]} \leq C$ and K is the Lipschitz constant of ϕ.

By the same means, applying the approximation bound of Proposition 3.1.8 combined with Propositions 3.2.1 or 3.2.2 we obtain learning bounds under smoothness and integrability assumptions.

Proposition 3.3.3. Let $k \geq \frac{d}{2} + 1 + \varepsilon$ for some $\varepsilon > 0$ and $f^* \in \mathcal{W}^{k,2}(\mathbb{R}^d) \cap \mathcal{L}^1(\mathbb{R}^d)$. Assume that $Z_{t,i}$ is an $\mathbb{R}^d$-valued random variable with $\|Z_{t,i}\|_2 \leq K$ $\mathbb{P}$-almost surely for all assets $i = 1, \ldots, N_t$ for some constant $K > 0$. Assume that

$$M^*_{t+1} = 1 - F^*_{t+1} = 1 - \frac{1}{N_t} f^*(Z_t)^\top R^e_{t+1},$$

is the true SDF. Suppose that Assumption 3.1.3 holds. Let $R = L^{1/(2k-2\varepsilon+1)}$, let $W = (w_1, \ldots, w_L)$ be i.i.d. random variables uniformly distributed over $B_R \subset \mathbb{R}^d$ and $\xi = (\xi_1, \ldots, \xi_L)$ be i.i.d. uniformly distributed over $[-\max(KR, 1), \max(KR, 1)]$. Assume that W, ξ are independent and let $\phi : \mathbb{R} \to \mathbb{R}$ given by $\phi(x) = \max(x, 0)$. Assume the factors $F_2, \ldots, F_{T+1}$ are independent and identically distributed over time. Let $B = \lambda\sqrt{L}\|f^*\|_{\mathcal{L}^1(\mathbb{R}^d)}$ for some constant $\lambda \geq 2$. Then, with probability at least $1 - \delta$ over the factors $F_2, \ldots, F_{T+1}$ and the weights $w_1, \ldots, w_L$ we have that

$$\mathbb{E}\left[\|\hat{M}_{t+1} - M^*_{t+1}\|^2\right] = \mathbb{E}\left[\|\hat{M}_{t+1}\|^2\right] - \mathbb{E}\left[\|M^*_{t+1}\|^2\right] \leq \frac{M^2 C_1}{\sqrt{T}} + \frac{C_2}{L^{2/\alpha}},$$

where:

- The approximate SDF is defined as

$$\hat{M}_{t+1} = 1 - \frac{1}{N_t}\hat{f}(Z_t)^\top R^e_{t+1} = 1 - \hat{b}^\top F_{t+1},$$

for $\hat{f}_{W,\xi}(x) = \frac{1}{L}\sum_{\ell=1}^{L} \hat{b}_\ell \phi(w_\ell^\top x + \xi_\ell)$.

- The vector $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_L)$ solves the constrained least squares problem

$$\hat{b} = \operatorname*{arg\,min}_{\|b\| \le B} \hat{\mathscr{L}}(b) = \operatorname*{arg\,min}_{\|b\| \le B} \frac{1}{T}\sum_{t=1}^{T}(1 - b^\top F_{t+1})^2.$$

- The constants α, M, C_1 and C_2 are defined as follows

$$\alpha = 2 + \frac{d+1}{k - d/2 - \varepsilon}, \quad , M = 1 + \lambda \|f^*\|_{\mathscr{L}^1(\mathbb{R}^d)},$$
$$C_1 = \left(16 + \sqrt{2\log\left(\frac{2}{\delta}\right)}\right), \quad C_2 = \frac{2C^2\|f^*\|_k^2 N_t}{\delta},$$

where C is explicitly given by Gonon et al. (2023, Equation 43).

Proof. Follows directly from Propositions 3.1.8 and 3.2.1 by a union bound as in Proposition 3.3.1, and observing that $\mathbb{E}[\|R^e_{t+1}\|_2^2] \le N_t$ under Assumption 3.1.3. □

Theorem 3.3.4. Let $k \ge \frac{d}{2} + 1 + \varepsilon$ for some $\varepsilon > 0$ and $f^* \in \mathscr{W}^{k,2}(\mathbb{R}^d) \cap \mathscr{L}^1(\mathbb{R}^d)$. Assume that $Z_{t,i}$ is an $\mathbb{R}^d$-valued random variable with $\|Z_{t,i}\|_2 \le K$ $\mathbb{P}$-almost surely for all assets $i = 1, \ldots, N_t$ for some constant $K > 0$. Assume that

$$M^*_{t+1} = 1 - F^*_{t+1} = 1 - \frac{1}{N_t} f^*(Z_t)^\top R^e_{t+1},$$

is the true SDF. Suppose that Assumption 3.1.3 holds. Let $R = L^{1/(2k-2\varepsilon+1)}$, let $W = (w_1, \ldots, w_L)$ be i.i.d. random variables uniformly distributed over $B_R \subset \mathbb{R}^d$ and $\xi = (\xi_1, \ldots, \xi_L)$ be i.i.d. uniformly distributed over $[-\max(KR,1), \max(KR,1)]$. Assume that W, ξ are independent and let $\phi : \mathbb{R} \to \mathbb{R}$ given by $\phi(x) = \max(x, 0)$. Assume the factors $F_2, \ldots, F_{T+1}$ are a sample of a stationary β-mixing sequence. Let $T = 2\mu a$ for $\mu, a > 0$. Let $B = \lambda\sqrt{L}\|f^*\|_{\mathscr{L}^1(\mathbb{R}^d)}$ for some constant $\lambda \ge 2$. Then, with probability at least $1 - \delta$ over the factors $F_2, \ldots, F_{T+1}$ and the weights $w_1, \ldots, w_L$ we have that

$$\mathbb{E}\left[\|\hat{M}_{t+1} - M^*_{t+1}\|^2\right] = \mathbb{E}\left[\|\hat{M}_{t+1}\|^2\right] - \mathbb{E}\left[\|M^*_{t+1}\|^2\right] \le \frac{M^2 C_1}{\sqrt{\mu}} + \frac{C_2}{L^{2/\alpha}},$$

where:

- The approximate SDF is defined as

$$\hat{M}_{t+1} = 1 - \frac{1}{N_t} \hat{f}(Z_t)^\top R^e_{t+1} = 1 - \hat{b}^\top F_{t+1},$$

  for $\hat{f}_{W,\xi}(x) = \frac{1}{L} \sum_{\ell=1}^{L} \hat{b}_\ell \phi(w_\ell^\top x + \xi_\ell)$.

- The vector $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_L)$ solves the constrained least squares problem

$$\hat{b} = \underset{\|b\| \le B}{\arg\min}\, \hat{\mathscr{L}}(b) = \underset{\|b\| \le B}{\arg\min}\, \frac{1}{T} \sum_{t=1}^{T} (1 - b^\top F_{t+1})^2.$$

- The constants α, M, C_1 and C_2 are defined as follows

$$\alpha = 2 + \frac{d+1}{k - d/2 - \varepsilon}, \quad , M = 1 + \lambda \|f^*\|_{\mathscr{L}^1(\mathbb{R}^d)}, \quad \delta' = \delta - 4(\mu - 1)\beta(a),$$
$$C_1 = \left(16 + \sqrt{2 \log\left(\frac{8}{\delta'}\right)}\right), \quad C_2 = \frac{2C^2 \|f^*\|_k^2 N_t}{\delta},$$

  and C is explicitly given by Gonon et al. (2023, Equation 43).

Proof. Follows directly from Propositions 3.1.8 and 3.2.2 by a union bound as in Proposition 3.3.1. □

Chapter 4

Empirics

4.1 Data

As starting point, we consider the set of U.S. firms available in the Center for Research in Securities Prices (CRSP) dataset. This time series has the prices, adjusted returns, of U.S. stocks since 1925. The stock characteristics we use to construct managed portfolios are from the Financial Ratios Firm Level dataset by Wharton Research Data Services (WRDS). This dataset has 69 financial ratios for each firm from 1970 onwards classified in seven categories: valuation, profitability, capitalisation, financial soundness, solvency, liquidity, efficiency and others. For example, the dataset includes valuation ratios such as Book/-Market (bm), Price/Book (ptm), profitability ratios such as Return on Equity (roe), Return on Assets (roa), etc. See Appendix A for the complete list of financial ratios. We supplement these 69 ratios with 12 past monthly returns in months $t-1$ through $t-12$, thus we have a set of $d = 81$ basic asset characteristics. We denote $c_{i,t}^{j}$ the characteristic j (where $j = 1,\ldots,d$) of asset i (where $i = 1,\ldots,N_t$) at time t (where $t = 1,\ldots,T$).

We restrict our sample to large cap stocks, as trading small cap stocks involves high bid-ask spreads and low liquidity, thus a high Sharpe ratio achieved with small cap stocks might not be exploitable. In practice, we restrict ourselves to the universe of stocks with a market capitalisation above 0.01% of the total market capitalisation, this leaves us with about 700 to 900 stocks on each month. This large capitalization group is defined as in Kozak et al. (2020) or Chen et al. (2024). It is in between the mega and large capitalization groups considered in Didisheim et al. (2023).

The CRSP dataset provide us with delisting returns for stocks that are delisted due to mergers, exchange, performance, etc. If a delisting return is missing and is due to performance, we set the delisting return to -100%. This is the most conservative estimate,

however, it is only applied once in the large cap universe. Alternatively, one can use a -30% proxy as suggested by Shumway (1997) based on average perfomance delisting returns. Other delisting returns are typically small (e.g. mergers have an average positive 2% delisting return) and when there is a missing delisting return not due to performance we proxy it to 0%.

Some care must be taken to compound the standard returns and delisting returns of a stock on its last month. The monthly delisting return typically includes both the return from the start of the month until the delisting date, as well as the delisting return itself. If the delisting occurs before the last trading day of the month, these returns are combined in the delisting return, and the standard return is missing. However, if the delisting happens on the last trading day of the month, both the standard return and the delisting return are reported separately. In this case, these two returns must be compounded to obtain the total return for the delisted stock. See Beaver et al. (2007) for more details.

We use the one month treasury bill rate as a proxy for the risk free rate R^f_{t+1}, and substract it from returns to obtain the excess return of each asset.

Lastly, we winsorize excess returns of assets so that they do not exceed 100% on each month, to reduce the impact of these outliers on the final estimates, as done in Jensen et al. (2023). Such returns happen although they are exceptionally rare for large cap stocks, only 0.035% of our sample or about 1 return every 3000 exceeds this bound. The test data remains unaltered when evaluating the model, i.e. we apply this winsorizing only to the training data.

4.2 Instruments from characteristics

Now we explain how the instrument matrix $Z_t \in \mathbb{R}^{N_t \times d}$ is defined from observed characteristics. At each time t, we observe d characteristics $c^j_{i,t}$ for $j = 1,\dots,d$ and each stock $i = 1,\dots,N_t$. There are usually missing values, so for each characteristic j we have $N_t(j)$ non-missing values that we rank from 1 to $N_t(j)$ and normalise it as follows

$$rc^j_{i,t} = \frac{\text{rank}(c^j_{i,t})}{N_t(j)+1}.$$

Notice that $\bar{rc}_t^j = \frac{1}{2}$ is the mean of the normalised rankings. Then, we center these ranks and normalise to have unit 1-norm:

$$Z_{i,t}^j = \frac{rc_{i,t}^j - \frac{1}{2}}{\sum_{i=1}^{N} \left| rc_{i,t}^j - \frac{1}{2} \right|}.$$

We also define $Z_{i,t}^j = 0$ for all missing values, so that stocks with a missing value have no exposure to the respective factor, and we get $Z_t \in \mathbb{R}^{N_t \times d}$. The resulting instruments have zero mean (i.e., $\sum_{i=1}^{N} Z_{i,t}^j = 0$) and have fixed leverage as the absolute exposure is one (i.e., $\sum_{i=1}^{N} |Z_{i,t}^j| = 1$). Moreover, the instruments are bounded as $Z_{i,t}^j \in \left[-\frac{1}{2}, \frac{1}{2}\right]$, thus $Z_{i,t} \in \left[-\frac{1}{2}, \frac{1}{2}\right]^d$ for all assets. This definition is the same as in Kozak (2020, Section 4.2) and is also used in Kozak (2020) and Didisheim et al. (2023).

It is worth noting that Chen et al. (2024) employs a different methodology by including only stocks with complete firm characteristics available for a given month. This method has the advantage of excluding stocks with small market capitalizations and avoids potential artificial time-series fluctuations in the instrument-based factors that can arise from data imputation. However, while the instrument factors have zero exposure to missing values, this is no longer the case for the (transformed) features.

4.3 Model configurations

With our theory, we obtain different approximate SDFs depending on the weights distribution $w \sim p$ and the activation function $\phi : \mathbb{R} \to \mathbb{R}$. We consider the following weight distributions:

a) $w \sim N(0, I_d)$, i.e. standard d-dimensional gaussian.

b) $w \sim U(B^d)$, i.e. uniform over the unit ball $B^d = \{w \in \mathbb{R}^d \,|\, w^\top w \leq 1\}$.

c) $w \sim U(S^{d-1})$, i.e. uniform over the sphere $S^{d-1} = \{w \in \mathbb{R}^d \,|\, w^\top w = 1\}$.

d) $w \sim t_\nu(0, I_d)$, i.e. a multivariate standard t-student distribution with ν degrees of freedom (with $\nu > 2$ so it has finite second moments).

With respect to activation functions, we consider the following list:

i) $\phi(x) = \text{ReLU}(x) = \max(x, 0)$, i.e. the rectified linear unit (ReLU) function.

ii) $\phi(x) = \sigma(x) = \frac{1}{1+e^{-x}}$, i.e. the sigmoid or logistic function.

iii) $\phi(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, i.e. the hyperbolic tangent function.

iv) $\phi(x) = (\cos(x), \sin(x))$, i.e. cosine and sine functions.

We observe that the ReLU function does not satisfy Assumption 3.1.1 as it is unbounded. However, if we consider weights as in *b*) or *c*), then for $z \in \mathcal{Z} = [-0.5, 0.5]^d$ we have the bound $|w^\top z| \leq \frac{\sqrt{d}}{2}$. Thus, effectively it behaves as if $|\phi| \leq 1$ up to a constant of normalisation. On the other hand, the sigmoid and hyperbolic tangent functions are bounded by 1 and thus satisfy Assumption 3.1.1 with any distribution of the list.

Our baseline configuration is given by a ReLU activation function with weights uniformly distributed over the unit ball (i.e., *i*) and *b*)). We will also explore alternative configurations for comparison. We will also evaluate the performance of configurations that do not satisfy all our theoretical hypothesis, such as the ReLU activation with gaussian or t-student weights, or the cosine-sine activation from Rahimi and Recht (2007) which is implicitely a $\mathbb{C}$-valued activation (i.e., the complex exponential). The cosine-sine activation with gaussian weights (*iv*) and *a*)) is a nice example as it corresponds in the limit to the gaussian kernel used in Kozak (2020). Finally, we assess the performance of a ReLU activation with weights $W = (w_1, \ldots, w_L)$ uniformly distributed on the unit ball and biases $\xi = (\xi_1, \ldots, \xi_L)$ uniform over $[-\sqrt{d}, \sqrt{d}]$. This configuration is closely aligned with that of Proposition 3.1.8, since if $z \in \mathcal{Z} = [-0.5, 0.5]^d$, then $\|z\|_2 \leq \sqrt{d}\|z\|_\infty = \frac{\sqrt{d}}{2}$. Additionally, $R = L^{1/(2k-2\varepsilon+1)} \leq 2$ for the values considered, since by assumption $k \geq \frac{d}{2} + 1 + \varepsilon$ which implies that $2k - 2\varepsilon + 1 \geq d + 3 = 84$, thus R is close to one.

We consider the number of random features $L_k = 2^k d$ for $k = -2, -1, 0, 1, \ldots, 7$ to evaluate the performance of the algorithm as the number of factors increase. Thus, the maximum number of random features considered is $L = 128d = 10{,}368$. When it comes to the regularisation paramenter γ, we consider a grid of 30 evenly spaced numbers on a log scale from 10^{-10} to 10^3. The implementation of the algorithms is available in a github repository [1].

4.4 Empirical results

4.4.1 Simple split

We first consider a simple split of the data, training our SDF with returns from February 1970 to January 2000, i.e. a fixed window of $T = 360$ months, and we evaluate its performance out of sample from February 2000 to December 2023. We recall that the SDF coefficients are those of a mean-variance efficient (MVE) portfolio (see Section 1.1, Cochrane

[1] `https://github.com/matiasdata/RandomFeatures`

2009, Section 6.1 or Chen et al. 2024, I. A.). Figure 4.1 displays the annualized Sharpe ratio achieved by the estimated MVE portfolios as a function of the regularisation parameter γ. The maximum Sharpe ratio achieved is 1.02, and it is comparable to those obtained by Chen et al. (2024, Table IV) with a GAN based SDF using 46 anomaly characteristics and 124 macroeconomic time series (restricted to large cap stocks, i.e. $\geq 0.01\%$ of total market cap) and Kozak (2020, Figure 8) using a radial kernel approach as described in section 1.4 on 50 anomaly characteristics.

Moreover, we look at the time series of returns obtained by the MVE portfolio for the γ that achieves the maximum Sharpe ratio in Figure 4.2. Specifically, Figure 4.2 displays the Year-over-Year return and cumulative return that this portfolio achieves out of sample. The optimal weights $\hat{\omega}_t = \Phi(Z_t)b$ are typically large and correspond to highly leveraged portfolios, thus we normalise these to have unit leverage ($\|\hat{\omega}_t\|_1 = 1$), i.e. the sum of long and short positions is equal to one dollar. This has no impact on the Sharpe ratio and controls the leverage which is a key measure when risk-managing such positions. Alternatively, one can normalized the weights so that the portfolio has the same standard deviation as the aggregate market (or a benchmark index), as done by Kozak et al. (2020), although this can be done exactly ex-post or only approximately in real-time.

Lastly, in table 4.1 we report the Jensen's alpha of our MVE portfolio for the CAPM and Fama-French 5-factor models, both alphas are above 10% and are statistically significant. The portfolio weights are normalized so that it has the same volatility as the market in that period.

Sharpe ratio	CAPM α	Z-score	Fama-French α	Z-score
1.02	13.38% (2.99%)	4.47	11.46% (3.06%)	3.74

Table 4.1: Out-of-sample (OOS) annualized Sharpe ratio, annualized CAPM α, Fama-French 5 factors α (standard errors in parenthesis) and respective Z-scores for the optimal γ portfolio in a simple split training. Portfolio returns are normalized to have the same volatility as the market. We use $L = 128d = 10{,}368$ random features. Random weights are chosen uniformly over the unit ball (i.e., $w \sim U(B^d)$), and a ReLU activation function is used (i.e., $\phi(x) = \text{ReLU}(x)$). The data is split once, optimal parameters are determined in-sample (IS), and held constant out-of-sample (OOS). In-sample period: February 1970 to January 2000. Out-of-sample period: February 2000 to December 2023.

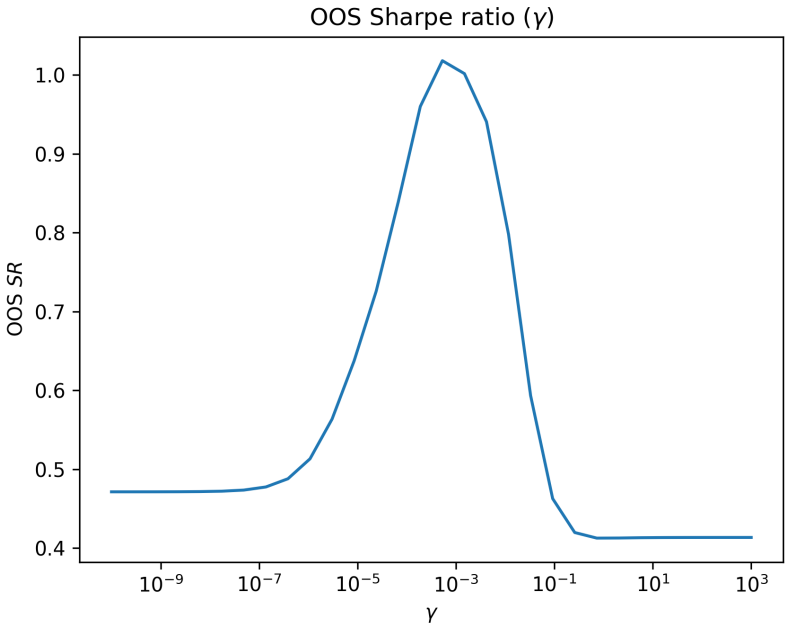
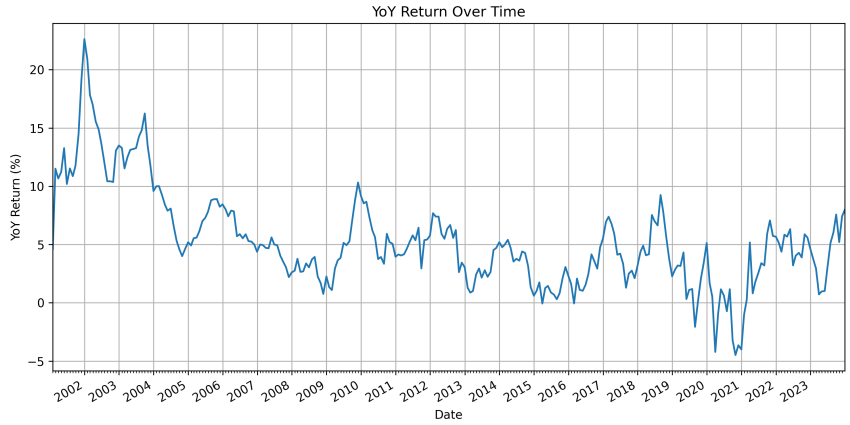
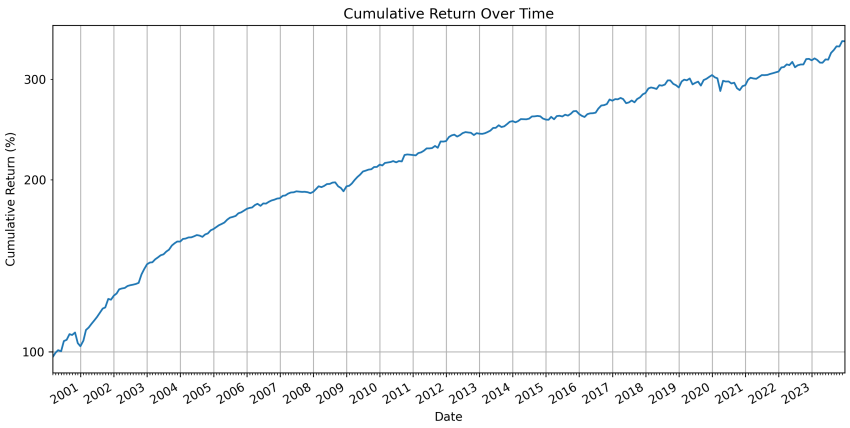


Figure 4.1: Out-of-sample (OOS) annualized Sharpe ratios of the estimated MVE portfolios as a function of the regularisation parameter γ, for $L = 128d = 10{,}368$ random features. Random weights are chosen uniform over the unit ball (i.e., $w \sim U(B^d)$) and we use a ReLU activation function (i.e., $\phi(x) = \text{ReLU}(x)$). The data is splitted once, optimal parameters are found IS and are held constant OOS. In-sample period: February 1970 to January 2000. Out-of-sample period: February 2000 to December 2023.



(a)



(b)

Figure 4.2: (a) Out-of-sample (OOS) Year-over-Year (YoY) return and (b) Out-of-sample (OOS) cumulative return for the portfolio with optimal γ, using $L = 128d = 10{,}368$ random features. The portfolio leverage is fixed at 1, i.e., $\|\hat{\omega}_t\|_1 = 1$. Random weights are chosen uniformly over the unit ball (i.e., $w \sim U(B^d)$), and a ReLU activation function is used (i.e., $\phi(x) = \mathrm{ReLU}(x)$). The data is split once, optimal parameters are determined in-sample (IS), and held constant out-of-sample (OOS). In-sample period: February 1970 to January 2000. Out-of-sample period: February 2000 to December 2023.

4.4.2 Rolling window

In contrast to the simple split approach, where we trained the SDF once on a fixed in-sample (IS) period and evaluated it on an out-of-sample (OOS) period, the rolling window approach continuously re-estimates the model. Specifically, we use a rolling window of 360 months, where the SDF is trained on the most recent 360 months of data and then evaluated on the next month. This process is repeated by shifting the window forward by one month, allowing the model to update its parameters as new data becomes available.

Figure 4.3 presents the out-of-sample annualized Sharpe ratios for the estimated mean-variance efficient (MVE) portfolios as a function of the regularisation parameter γ, similar to the analysis in the simple split approach. The maximum Sharpe ratio achieved under the rolling window method is 1.28, which is considerably higher than the 1.02 obtained in the simple split, indicating a better performance when we re-estimate the model every month.

Additionally, in Figure 4.4, we observe the Year-over-Year (YoY) and cumulative returns for the portfolio that achieves the maximum Sharpe ratio under the rolling window approach. This figure shows that the cumulative return is higher achieving a 275% return versus 249% for the simple split. The improvement in the SR is mainly achieved with a reduction of volatility of the MVE portfolio.

Finally, in Table 4.2, we report the Jensen's alpha for both the CAPM and Fama-French 5-factor models under the rolling window approach, along with their respective Z-scores. Similar to the simple split analysis, these alphas are statistically significant and above 10%, but with the rolling window, we observe an even higher CAPM alpha of 18.79% and a corresponding higher Z-score. This suggests that re-estimating the model every month might capture more of the time-varying risks and opportunities in the market compared to a fixed in-sample training period.

Sharpe ratio	CAPM α	Z-score	Fama-French α	Z-score
1.28	18.79% (3.20%)	5.87	13.71% (3.21%)	4.27

Table 4.2: Out-of-sample (OOS) annualized Sharpe ratio, annualized CAPM α, Fama-French 5 factors α (standard errors in parenthesis) and respective Z-scores for the optimal γ portfolio in a rolling window training. Portfolio returns are normalized to have the same volatility as the market. We use $L = 128d = 10{,}368$ random features. Random weights are chosen uniformly over the unit ball (i.e., $w \sim U(B^d)$), and a ReLU activation function is used (i.e., $\phi(x) = \text{ReLU}(x)$). Training is performed on a rolling window of $T = 360$ months, and evaluated on the next month, from February 2000 to December 2023.

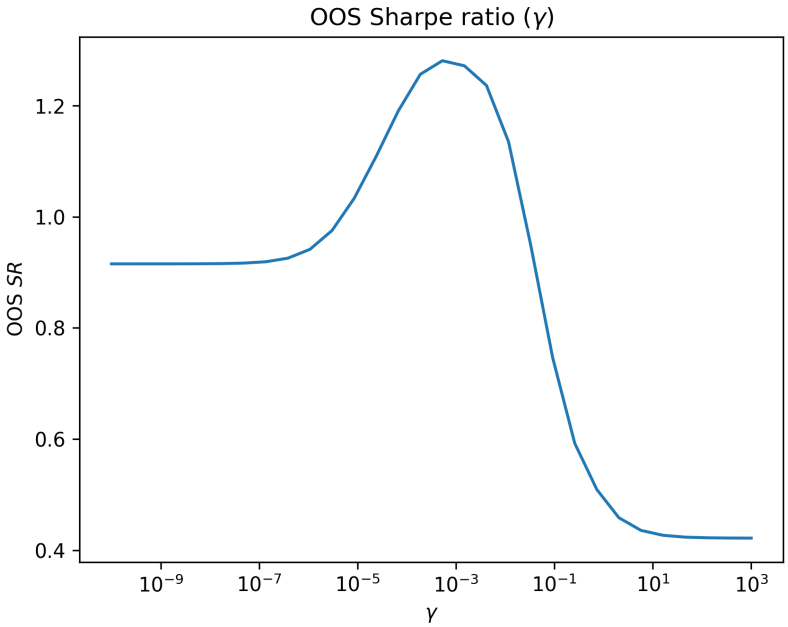
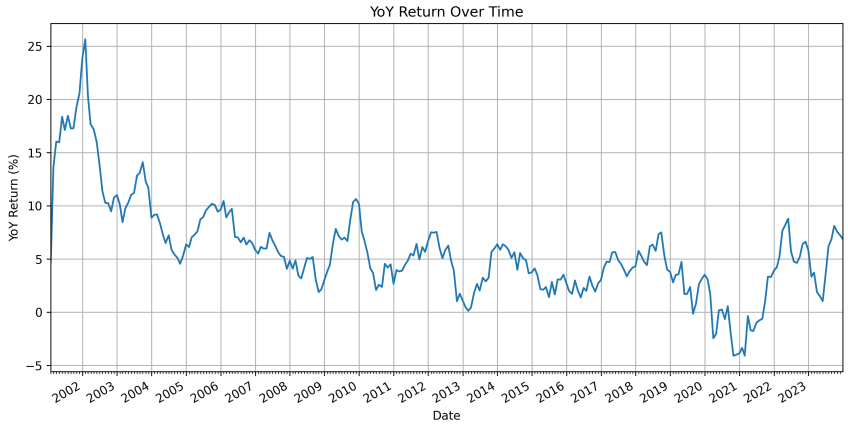
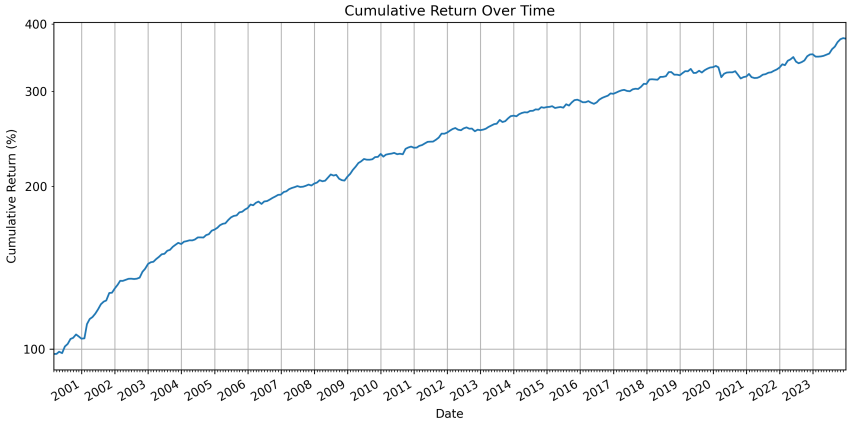


Figure 4.3: Out-of-sample (OOS) annualized Sharpe ratios as a function of the regularisation parameter γ, for $L = 128d = 10{,}368$ random features. Random weights are chosen uniform over the unit ball (i.e., $w \sim U(B^d)$) and we use a ReLU activation function (i.e., $\phi(x) = \text{ReLU}(x)$). Training is performed on a rolling window of $T = 360$ months, and evaluated on the next month, from February 2000 to December 2023.



(a)



(b)

Figure 4.4: (a) Out-of-sample (OOS) Year-over-Year (YoY) return and (b) Out-of-sample (OOS) cumulative return for the portfolio with optimal γ, using $L = 128d = 10{,}368$ random features. The portfolio leverage is fixed at 1, i.e., $\|\hat{w}_t\|_1 = 1$. Random weights are chosen uniformly over the unit ball (i.e., $w \sim U(B^d)$), and a ReLU activation function is used (i.e., $\phi(x) = \mathrm{ReLU}(x)$). Training is performed on a rolling window of $T = 360$ months, and evaluated on the next month, from February 2000 to December 2023.

4.4.3 Different number of factors

In this section, we examine how the Sharpe ratio of the MVE portfolio varies as a function of both the regularization parameter γ and the ratio L/d of the number of random features L to the number of basic characteristics d.

Figure 4.5 illustrates this relationship, showing that the SR improves almost monotonically as L/d increases. This suggests that incorporating more random features improves the model's approximation ability to capture the underlying data structure, leading to better out-of-sample performance, consistent with our theory of Chapter 3 and the "virtue of complexity" observed in Didisheim et al. (2023). We observe that the maximum Sharpe ratio of 1.33 is achieved with 648 random features, is only slightly higher than the 1.28 obtained by the most complex model.

Furthermore, we evaluate if the convergence rate of the out-of-sample empirical loss aligns with that of our Theorems 3.3.2 and 3.3.4. We consider the out-of-sample squared norm of our SDF:

$$E_{OOS}[\|\hat{M}\|^2] = \frac{1}{T_{OOS}} \sum_{t=T}^{T+T_{OOS}} \left(1 - \frac{1}{N_t} \hat{f}(Z_t)^\top R^e_{t+1}\right)^2.$$

From Theorem 3.3.2 (or Theorem 3.3.4 for k large enough) we expect that $\sqrt{E_{OOS}[\|\hat{M}\|^2]}$ should decrease to the constant $\sqrt{\mathbb{E}[\|M^*_t\|^2]}$ with the number of factors (L) at a rate of $O(1/\sqrt{L})$ and a rate of $O(1/T^{1/4})$ for the training sample size (i.e., T, recall that μ is proportional to T).

To evaluate this convergence, we increase linearly the (rolling window) training sample size T from 12 months to 360 months (30 years), and we allow the number of random features to increase according to $L = c\sqrt{T}$, where c is set to $1000/\sqrt{360}$ so that 360 months corresponds to 1000 features. Under this configuration, we expect $E_{OOS}[\|\hat{M}\|^2]$ to decrease at a $O(1/\sqrt{L})$.

Figure 4.6 plots $\sqrt{E_{OOS}[\|\hat{M}\|^2]}$ against the number of features L, along with a curve of the form $y(L) = c_1 + \frac{c_2}{\sqrt{L}}$ where c_1 and c_2 are fitted to the data by a least squares criterion. The results indicate that the empirical loss decreases at the predicted $O(1/\sqrt{L})$ rate of our results, albeit with some noise.

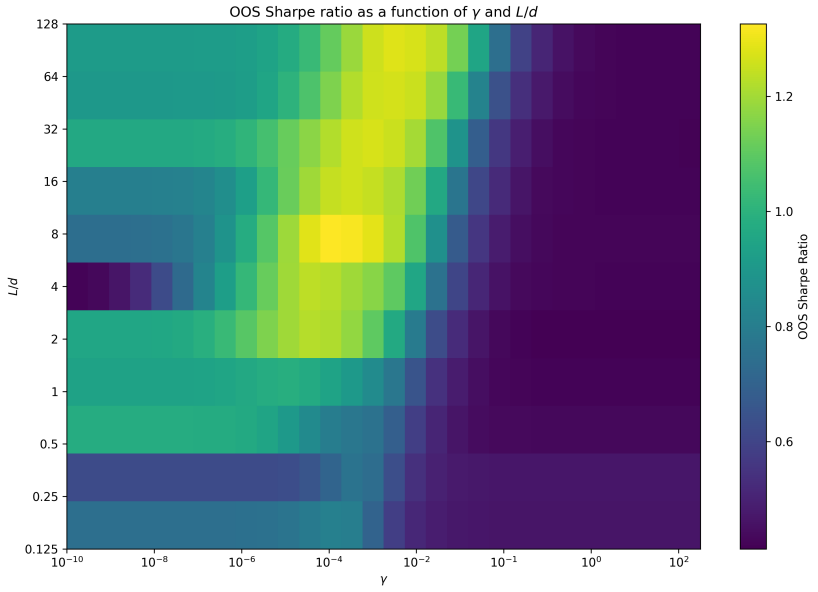


Figure 4.5: Out-of-sample (OOS) annualized Sharpe ratio as a function of γ and L/d. The model has $L = 128d = 10{,}368$ random features, random weights are chosen uniform over the unit ball (i.e., $w \sim U(B^d)$) and we use a ReLU activation function (i.e., $\phi(x) = \text{ReLU}(x)$). Training is performed on a rolling window of $T = 360$ months, and evaluated on the next month, from February 2000 to December 2023.

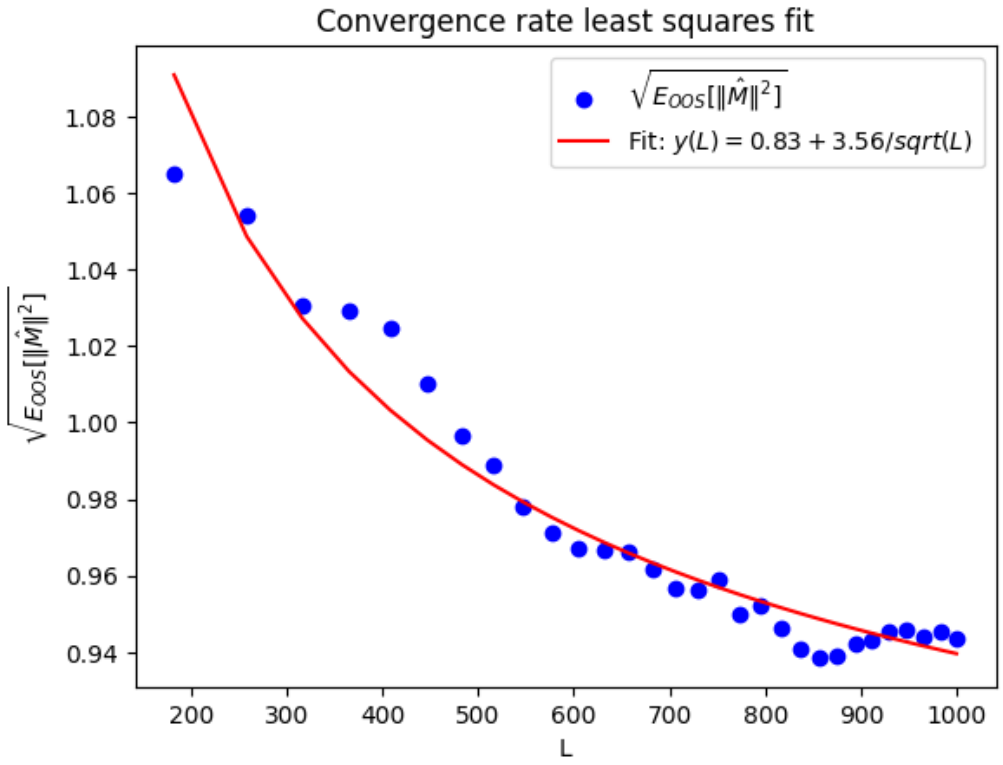


Figure 4.6: Out-of-sample (OOS) loss, i.e. $\sqrt{E_{OOS}[\|\hat{M}\|^2]}$, as a function of the number of random features. Random weights are chosen uniform over the unit ball (i.e., $w \sim U(B^d)$) and we use a ReLU activation function (i.e., $\phi(x) = \text{ReLU}(x)$). The number of random features goes from 182 to 1000 and increases as $L = c\sqrt{T}$ where T is the rolling window size. γ is fixed at the optimal value of Section 4.4.2 ($\gamma \approx 0.0005$). Training is performed on a rolling window of T months for $T = 12, 24, \ldots, 360$. Evaluation is done in the period from February 2000 to December 2023.

4.4.4 Cross validation

The results obtained from the simple split or rolling window approaches, although informative, present an idealized scenario as the optimal regularisation parameter γ is known only after evaluating the entire out-of-sample period, thus it's Sharpe ratio is not actually achievable. In other words, in a real-world setting, this knowledge isn't available at the time the portfolio needs to be executed as γ has to be fixed in advance. To address this, we apply 5-fold cross-validation within our in-sample period to select the optimal γ, making the portfolio tradable, i.e. the process is implementable on-line. In the cross-validation approach, the model is trained on a rolling window of 360 months, and every 12 months, the regularisation parameter γ is re-selected using 5-fold cross-validation on the in-sample data.

Figure 4.7 shows the evolution of the selected regularisation parameter γ over time, illustrating how cross-validation dynamically adjusts the regularisation parameter. Even though the out-of-sample SR achieved using cross-validation (1.275) is slightly lower than the one obtained in the rolling window method (1.281), the difference is marginal, indicating that the efficiency loss from not knowing the optimal γ ahead of time is small. We observe that the chosen γ is stable over time as it only changes a few times, and are close to the optimal fixed value of $\gamma \approx 0.0005$ found in subsection 4.4.2.

Further supporting this, Figures 4.8 (a) and (b) present the out-of-sample Year-over-Year (YoY) and cumulative returns for the portfolio where γ is chosen through cross-validation. The cumulative return is slightly higher at 291%, compared to the 275% achieved with the rolling window approach.

Furthermore, Table 4.3 reports Jensen's alpha for both the CAPM and Fama-French 5-factor models under the cross-validation method, along with their respective Z-scores. The CAPM alpha is 18.06%, which, although slightly lower than the 18.79% observed in the rolling window approach, remains highly significant. This demonstrates that cross-validation, gives a tradable approximate MVE portfolio without much efficiency loss.

Finally, Figure 4.9 plots the out-of-sample rolling three-year and five-year Sharpe ratio of the MVE portfolio and the market portfolio. We observe that the performance of the SDF is extraordinarily high in the 2000s', where it achieves a SR between 2 and 3. On the 2010s' performance drops but its is still about 1.5 on average and is still outperforming the market by a considerable amount. However, since 2020 the SR drops (particularly during the COVID crisis) and has been close to 0.5 on average, only to recover by the end of 2023.

This might be due to the extraordinary circumstances of this crisis, where these tail

risks where not priced appropriately by this SDF. An alternative explanation is that it might be due to the "publication decay", as documented by McLean and Pontiff 2016 and Jensen et al. 2023). This phenomenon occurs where after a new factor that predicts the cross-section of stock returns is published, the returns of a portfolio based on it are significantly reduced, by about 50% on average. Given that machine learning methods exploiting similar characteristics were published around that time, it is plausible that this has negatively impacted the performance of our SDF. This serves as a word of caution against relying on these methods for machine-learning-driven investments.

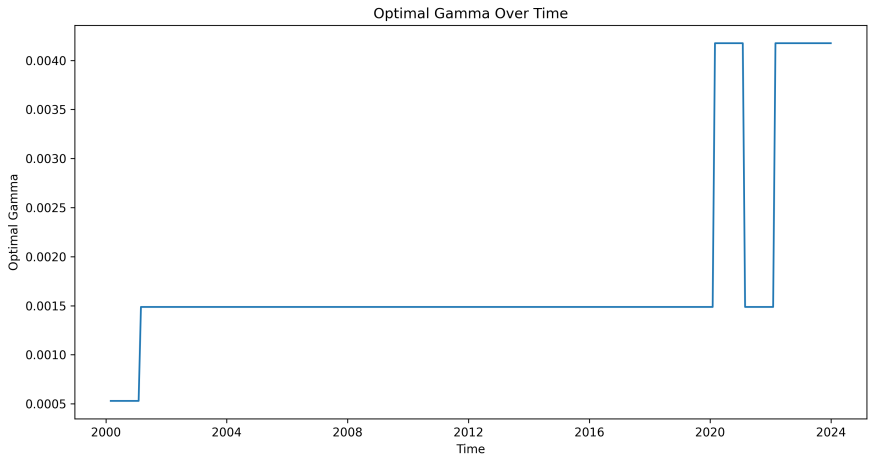
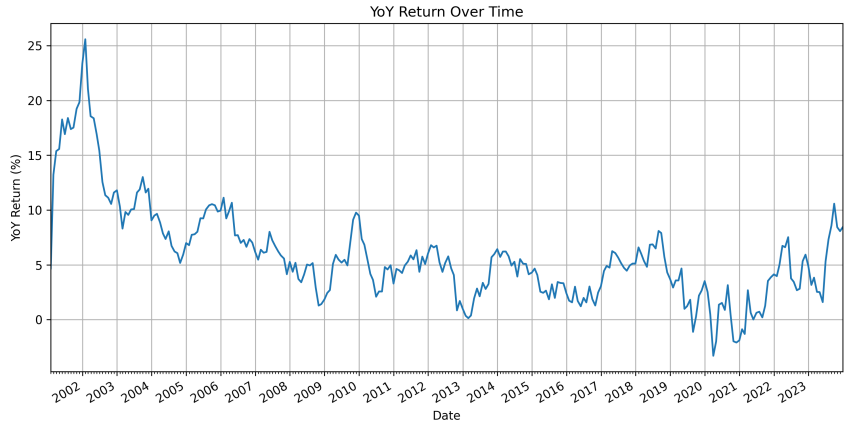


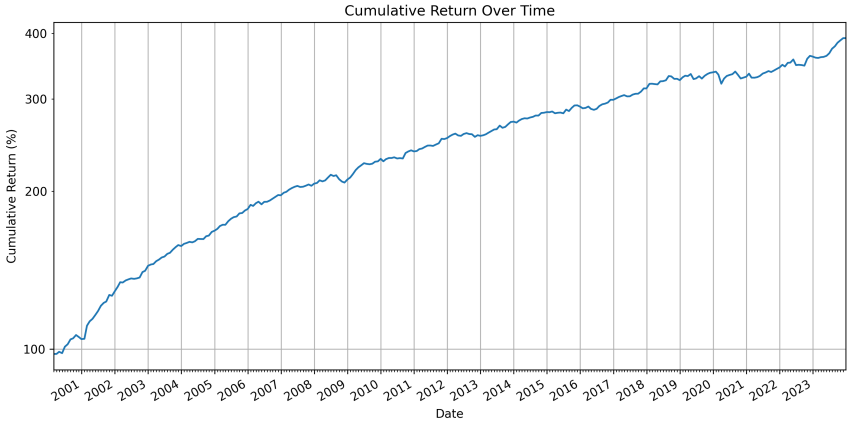
Figure 4.7: Regularisation parameter γ as a function of time selected with cross-validation every 12 months. The model has $L = 128d = 10{,}368$ random features, random weights are chosen uniform over the unit ball (i.e., $w \sim U(B^d)$) and we use a ReLU activation function (i.e., $\phi(x) = \text{ReLU}(x)$). Training is performed every month on a rolling window of $T = 360$ months, and every 12 months γ is selected with 5-fold cross validation.

Sharpe ratio	CAPM α	Z-score	Fama-French α	Z-score
1.28	18.06% (3.11%)	5.81	11.76% (3.01%)	3.90

Table 4.3: Out-of-sample (OOS) annualized Sharpe ratio, annualized CAPM α, Fama-French 5 factors α (standard errors in parenthesis) and respective Z-scores for the optimal γ portfolio chosen with cross validation. Portfolio returns are normalized to have the same volatility as the market. We use $L = 128d = 10{,}368$ random features. Random weights are chosen uniformly over the unit ball (i.e., $w \sim U(B^d)$), and a ReLU activation function is used (i.e., $\phi(x) = \text{ReLU}(x)$). Training is performed every month on a rolling window of $T = 360$ months, and every 12 months γ is selected with 5-fold cross validation.

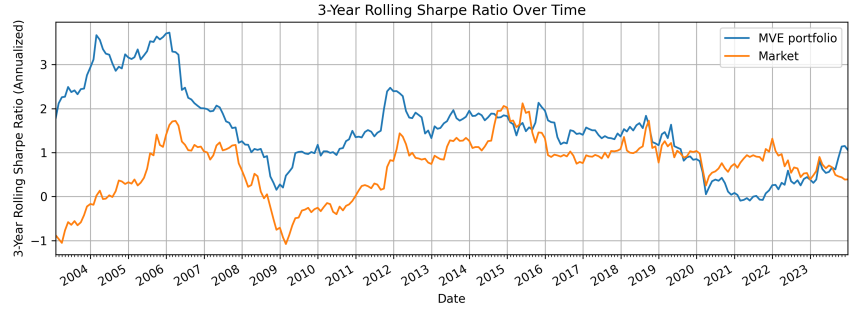


(a)

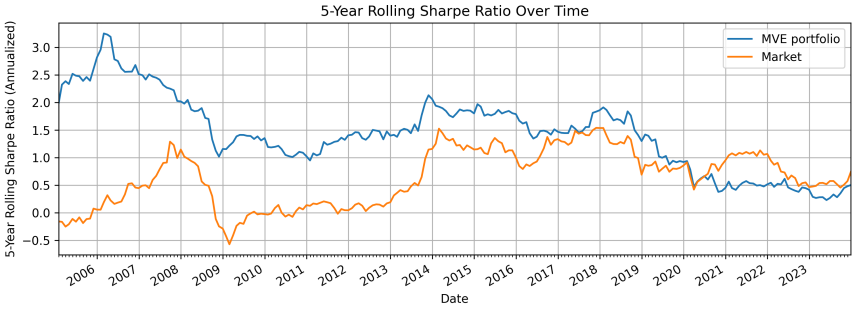


(b)

Figure 4.8: (a) Out-of-sample (OOS) Year-over-Year (YoY) return and (b) Out-of-sample (OOS) cumulative return for the portfolio with γ chosen through cross-validation, using $L = 128d = 10{,}368$ random features. The portfolio leverage is fixed at 1, i.e., $\|\hat{\omega}_t\|_1 = 1$. Random weights are chosen uniformly over the unit ball (i.e., $w \sim U(B^d)$), and a ReLU activation function is used (i.e., $\phi(x) = \text{ReLU}(x)$). Training is performed every month on a rolling window of $T = 360$ months, and every 12 months γ is selected with 5-fold cross validation.



(a)



(b)

Figure 4.9: OOS (a) three-year and (b) five-year rolling Sharpe ratio for the portfolio with γ chosen through cross-validation, using $L = 128d = 10{,}368$ random features. The portfolio leverage is fixed at 1, i.e., $\|\hat{\omega}_t\|_1 = 1$. Random weights are chosen uniformly over the unit ball (i.e., $w \sim U(B^d)$), and a ReLU activation function is used (i.e., $\phi(x) = \text{ReLU}(x)$). Training is performed every month on a rolling window of $T = 360$ months, and every 12 months γ is selected with 5-fold cross validation.

4.4.5 Results for alternative model configurations

In this section, we present the out-of-sample (OOS) performance metrics for various model configurations, focusing on different activation functions and random weight distributions. We compare the results across a simple split training, a rolling window training, and in a cross-validation approach. The results across various model configurations are summarized in Tables 4.4, 4.5, and 4.6.

These tables show that ReLU-based models exhibit stable performance regardless of the type of random weight distribution used. As seen in Table 4.6, the t-student distribution with $\nu = 4$ slightly outperforms other configurations, particularly in cross-validation, where it achieves the highest Sharpe ratio (SR).

The hyperbolic tangent activation function, as indicated in all three tables, consistently underperform compared to ReLU, although it clearly produces a different approximate SDF with considerably lower market exposure. On the other hand, the sigmoid and cosine-sine activation yield results that are decidedly close to the ReLU configurations. The results for ReLU with gaussian or t-student weights, although not covered by our theory, exhibit similar performance, indicating that our theoretical results might hold under less restrictive hypothesis.

Overall, the results demonstrate that the models are robust and deliver strong out-of-sample performance across different configurations, with ReLU remaining a reliable choice, regardless of the random weight distribution used.

Model	SR	CAPM β	CAPM α	Z-score	FF α	Z-score
ReLU $U(B^d)$	1.02	0.43	13.38%	4.47	11.46%	3.74
		(0.054)	(2.99%)		(3.06%)	
ReLU t-student	1.01	0.40	13.53%	4.45	11.69%	3.76
$\nu = 4$		(0.054)	(3.04%)		(3.11%)	
ReLU gaussian	1.01	0.48	12.91%	4.45	10.73%	3.62
		(0.052)	(2.90%)		(2.96%)	
Tanh gaussian	0.80	0.37	10.42%	3.38	10.18%	3.19
		(0.055)	(3.08%)		(3.19%)	
Sigmoid gaussian	0.99	0.43	12.81%	4.30	11.47%	3.74
		(0.053)	(2.98%)		(3.07%)	
Cosine-sine gaussian	0.98	0.43	12.85%	4.29	11.51%	3.73
		(0.054)	(2.99%)		(3.09%)	
ReLU $U(B^d)$	0.99	0.42	12.94%	4.31	11.60%	3.80
with biases		(0.054)	(3.00%)		(3.05%)	

Table 4.4: Out-of-sample (OOS) annualized Sharpe ratio (SR), CAPM β, annualized CAPM α, Fama-French 5 factors α (standard errors in parenthesis) and respective Z-scores for the optimal γ portfolio in a simple split training for alternative model configurations. Portfolio returns are normalized to have the same volatility as the market. We use $L = 128d = 10{,}368$ random features. Different activation functions and random weights are considered. The data is split once, optimal parameters are determined in-sample (IS), and held constant out-of-sample (OOS). In-sample period: February 1970 to January 2000. Out-of-sample period: February 2000 to December 2023.

Model	SR	CAPM β	CAPM α	Z-score	FF α	Z-score
ReLU $U(B^d)$	1.28	0.25	18.79%	5.87	13.71%	4.27
		(0.057)	(3.20%)		(3.21%)	
ReLU t-student	1.29	0.25	18.98%	5.93	13.69%	4.28
$\nu = 4$		(0.057)	(3.29%)		(3.19%)	
ReLU gaussian	1.28	0.24	18.79%	5.84	13.91%	4.31
		(0.058)	(3.21%)		(3.23%)	
Tanh gaussian	1.09	0.15	16.36%	5.00	13.00%	3.84
		(0.059)	(3.27%)		(3.38%)	
Sigmoid gaussian	1.27	0.27	18.51%	5.81	13.96%	4.31
		(0.057)	(3.19%)		(3.24%)	
Cosine-sine gaussian	1.26	0.26	18.43%	5.77	13.90%	4.28
		(0.057)	(3.20%)		(3.25%)	
ReLU $U(B^d)$	1.27	0.28	18.51%	5.82	13.90%	4.31
with biases		(0.057)	(3.18%)		(3.23%)	

Table 4.5: Out-of-sample (OOS) annualized Sharpe ratio, CAPM β, annualized CAPM α, Fama-French 5 factors α (standard errors in parenthesis) and respective Z-scores for the optimal γ portfolio in a cross validation training for alternative model configurations. Portfolio returns are normalized to have the same volatility as the market. We use $L = 128d = 10{,}368$ random features. Different activation functions and random weights are considered. Training is performed every month on a rolling window of $T = 360$ months, and every 12 months γ is selected with 5-fold cross validation.

Model	SR	CAPM β	CAPM α	Z-score	FF α	Z-score
ReLU $U(B^d)$	1.28	0.34 (0.056)	18.06% (3.11%)	5.81	11.76% (3.01%)	3.91
ReLU t-student $\nu = 4$	1.30	0.33 (0.056)	18.54% (3.13%)	5.92	12.13% (3.03%)	4.01
ReLU gaussian	1.24	0.38 (0.055)	17.19% (3.06%)	5.61	11.12% (2.98%)	3.73
Tanh gaussian	1.03	0.09 (0.059)	15.91% (3.29%)	4.83	12.06% (3.36%)	3.58
Sigmoid gaussian	1.27	0.35 (0.055)	17.91% (3.10%)	5.77	12.85% (3.10%)	4.14
Cosine-sine gaussian	1.26	0.34 (0.056)	17.80% (3.11%)	5.72	12.74% (3.11%)	4.09
ReLU $U(B^d)$ with biases	1.25	0.36 (0.055)	17.51% (3.09%)	5.67	12.33% (3.07%)	4.01

Table 4.6: Out-of-sample (OOS) annualized Sharpe ratio, CAPM β, annualized CAPM α, Fama-French 5 factors α (standard errors in parenthesis) and respective Z-scores for the optimal γ portfolio in a rolling window training for alternative model configurations. Portfolio returns are normalized to have the same volatility as the market. We use $L = 128d = 10{,}368$ random features. Different activation functions and random weights are considered. Training is performed on a rolling window of $T = 360$ months, and evaluated on the next month, from February 2000 to December 2023.

Conclusion

This thesis has explored the application of random features, also known as random feature neural networks, to asset pricing models within high-dimensional contexts. The empirical analysis demonstrated that random features-based methods could achieve comparable performance to more complex and computationally intensive techniques. Furthermore, the theoretical contributions of this thesis, including approximation, generalization, and learning bounds, provide a solid foundation for the practical use of random features in asset pricing.

However, open questions remain, such as extending these bounds under less restrictive assumptions. Results from Gonon et al. (2023), Gonon (2023) and Györfi et al. (2002) could prove to be really helpful in this direction. Additionally, it would be interesting to analyze the approach proposed by Kozak (2020) using the theory developed in this thesis. Another promising direction would be to investigate modifying the GAN model proposed by Chen et al. (2024) by incorporating random features.

In conclusion, the use of random features in asset pricing models represents a significant advancement, offering a practical and theoretically sound alternative to traditional methods.

Appendix A

Variable definitions

Below is the list of characteristics from the Financial Ratios Firm Level dataset by Wharton Research Data Services (WRDS). This dataset has 69 financial ratios for each firm classified in seven categories: valuation, profitability, capitalisation, financial soundness, solvency, liquidity, efficiency and others. Definitions are taken from the dataset manual[1].

1. **Dividend Payout Ratio** (dpr) - Valuation. Dividends as a fraction of income before extra items.
2. **Trailing P/E to Growth (PEG) ratio** (peg_trailing) - Valuation. Price-to-Earnings, excluding extraordinary items (diluted) to 3-year past EPS growth.
3. **Book/Market** (bm) - Valuation. Book value of equity as a fraction of market value of equity.
4. **Shiller's Cyclically Adjusted P/E Ratio** (capei) - Valuation. Multiple of market value of equity to 5-year moving average of net income.
5. **Dividend Yield** (divyield) - Valuation. Dividend rate as a fraction of price.
6. **Enterprise Value Multiple** (evm) - Valuation. Multiple of enterprise value to earnings before interest, taxes, depreciation, and amortisation (EBITDA).
7. **Price/Cash flow** (pcf) - Valuation. Multiple of market value of equity to net cash flow from operating activities.
8. **P/E (Diluted, Excl. EI)** (pe_exi) - Valuation. Price-to-Earnings, excluding extraordinary items (diluted).
9. **P/E (Diluted, Incl. EI)** (pe_inc) - Valuation. Price-to-Earnings, including extraordinary items (diluted).

[1] https://wrds-www.wharton.upenn.edu/documents/793/WRDS_Industry_Financial_Ratio_Manual.pdf

10. **Price/Operating Earnings (Basic, Excl. EI)** (pe_op_basic) - Valuation. Price-to-Operating-Earnings, excluding extraordinary items (diluted).
11. **Price/Operating Earnings (Diluted, Excl. EI)** (pe_op_dil) - Valuation. Price-to-Operating-Earnings, including extraordinary items (diluted).
12. **Price/Sales** (ps) - Valuation. Multiple of market value of equity to sales.
13. **Price/Book** (ptb) - Valuation. Multiple of market value of equity to book value of equity.
14. **Effective Tax Rate** (efftax) - Profitability. Income Tax as a fraction of Pretax Income.
15. **Gross Profit/Total Assets** (gprof) – Profitability. Gross profitability as a fraction of total assets.
16. **After-tax Return on Average Common Equity** (aftret_eq) – Profitability. Net income as a fraction of average of common equity based on most recent two periods.
17. **After-tax Return on Total Stockholders Equity** (aftret_equity) - Profitability. Net income as a fraction of average of total shareholders' equity based on most recent two periods.
18. **After-tax Return on Invested Capital** (aftret_invcapx) – Profitability. Net income plus interest expenses as a fraction of invested capital.
19. **Gross Profit Margin** (gpm) – Profitability. Gross profit as a fraction of sales.
20. **Net Profit Margin** (npm) – Profitability. Net income as a fraction of sales.
21. **Operating Profit Margin After Depreciation** (opmad) – Profitability. Operating income after depreciation as a fraction of sales.
22. **Operating Profit Margin Before Depreciation** (opmbd) – Profitability. Operating income before depreciation as a fraction of sales.
23. **Pre-tax Return on Total Earning Assets** (pretret_earnat) – Profitability. Operating income after depreciation as a fraction of average total earnings assets (TEA) based on most recent two periods, where TEA is defined as the sum of property plant and equipment and current assets.
24. **Pre-tax return on Net Operating Assets** (pretret_noa) – Profitability. Operating income after depreciation as a fraction of average net operating assets (NOA) based on most recent two periods, where NOA is defined as the sum of property plant and equipment and current assets minus current liabilities.
25. **Pre-tax Profit Margin** (ptpm) – Profitability. Pretax income as a fraction of sales.

26. **Return on Assets** (roa) – Profitability. Operating income before depreciation as a fraction of average total assets based on most recent two periods.

27. **Return on Capital Employed** (roce) – Profitability. Earnings before interest and taxes as a fraction of average capital employed based on most recent two periods, where capital employed is the sum of debt in long-term and current liabilities and common/ordinary equity.

28. **Return on Equity** (roe) – Profitability. Net income as a fraction of average book equity based on most recent two periods, where book equity is defined as the sum of total parent stockholders' equity and deferred taxes and investment tax credit.

29. **Capitalization Ratio** (capital_ratio) – Capitalization. Total long-term debt as a fraction of the sum of total long-term debt, common/ordinary equity and preferred stock.

30. **Common Equity/Invested Capital** (equity_invcap) – Capitalization. Common equity as a fraction of invested capital.

31. **Long-term Debt/Invested Capital** (debt_invcap) – Capitalization. Long-term debt as a fraction of invested capital.

32. **Total Debt/Invested Capital** (totdebt_invcap) – Capitalization. Total debt (long-term and current) as a fraction of invested capital.

33. **Inventory/Current Assets** (invt_act) – Financial soundness. Inventories as a fraction of current assets.

34. **Receivables/Current Assets** (rect_act) – Financial soundness. Accounts receivables as a fraction of current assets.

35. **Free Cash Flow/Operating Cash Flow** (fcf_ocf) – Financial soundness. Free cash flow as a fraction of operating cash flow, where free cash flow is defined as the difference between operating cash flow and capital expenditures.

36. **Operating CF/Current Liabilities** (ocf_lct) – Financial soundness. Operating cash flow as a fraction of current liabilities.

37. **Cash Flow/Total Debt** (cash_debt) – Financial soundness. Operating cash flow as a fraction of total debt.

38. **Cash Balance/Total Liabilities** (cash_lt) – Financial soundness. Cash balance as a fraction of total liabilities.

39. **Cash Flow Margin** (cfm) – Financial soundness. Income before extraordinary items and depreciation as a fraction of sales.

40. **Short-Term Debt/Total Debt** (short_debt) – Financial soundness. Short-term debt as a fraction of total debt.

41. **Profit Before Depreciation/Current Liabilities** (profit_lct) – Financial soundness. Operating income before D&A as a fraction of current liabilities.

42. **Current Liabilities/Total Liabilities** (curr_debt) – Financial soundness. Current liabilities as a fraction of total liabilities.

43. **Total Debt/EBITDA** (debt_ebitda) – Financial soundness. Gross debt as a fraction of EBITDA.

44. **Long-term Debt/Book Equity** (dltt_be) – Financial soundness. Long-term debt to book equity.

45. **Interest/Average Long-term Debt** (int_debt) – Financial soundness. Interest as a fraction of average long-term debt based on most recent two periods.

46. **Interest/Average Total Debt** (int_totdebt) – Financial soundness. Interest as a fraction of average total debt based on most recent two periods.

47. **Long-term Debt/Total Liabilities** (lt_debt) – Financial soundness. Long-term debt as a fraction of total liabilities.

48. **Total Liabilities/Total Tangible Assets** (lt_ppent) – Financial soundness. Total liabilities to total tangible assets.

49. **Total Debt/Equity** (de_ratio) – Solvency. Total liabilities to shareholders' equity (common and preferred).

50. **Total Debt/Total Assets** (debt_assets) – Solvency. Total debt as a fraction of total assets.

51. **Total Debt/Total Assets** (debt_at) - Solvency. Total liabilities as a fraction of total assets.

52. **Total Debt/Capital** (debt_capital) – Solvency. Total debt as a fraction of total capital, where total debt is defined as the sum of accounts payable and total debt in current and long-term liabilities, and total capital is defined as the sum of total debt and total equity (common and preferred).

53. **After-tax Interest Coverage** (intcov) – Solvency. Multiple of after-tax income to interest and related expenses.

54. **Interest Coverage Ratio** (intcov_ratio) – Solvency. Multiple of earnings before interest and taxes to interest and related expenses.

55. **Cash Conversion Cycle (Days)** (cash_conversion) – Liquidity. Inventories per daily COGS plus account receivables per daily sales minus account payables per daily COGS.

56. **Cash Ratio** (cash_ratio) – Liquidity. Cash and short-term investments as a fraction of current liabilities.

57. **Current Ratio** (curr_ratio) – Liquidity. Current assets as a fraction of current liabilities.

58. **Quick Ratio (Acid Test)** (quick_ratio) – Liquidity. Quick ratio: current assets net of inventories as a fraction of current liabilities.

59. **Asset Turnover** (at_turn) – Efficiency. Sales as a fraction of the average total assets based on the most recent two periods.

60. **Inventory Turnover** (inv_turn) – Efficiency. COGS as a fraction of the average Inventories based on the most recent two periods.

61. **Payables Turnover** (pay_turn) – Efficiency. COGS and change in inventories as a fraction of the average of accounts payable based on the most recent two periods.

62. **Receivables Turnover** (rect_turn) – Efficiency. Sales as a fraction of the average of accounts receivables based on the most recent two periods.

63. **Sales/Stockholders Equity** (sale_equity) – Efficiency. Sales per dollar of total stockholders' equity.

64. **Sales/Invested Capital** (sale_invcap) – Efficiency. Sales per dollar of invested capital.

65. **Sales/Working Capital** (sale_nwc) – Efficiency. Sales per dollar of working capital, defined as difference between current assets and current liabilities.

66. **Accruals/Average Assets** (accrual) - Other. Accruals as a fraction of average total assets based on most recent two periods.

67. **Research and Development/Sales** (rd_sale) – Other. R&D expenses as a fraction of sales.

68. **Advertising Expenses/Sales** (adv_sale) – Other. Advertising expenses as a fraction of sales.

69. **Labor Expenses/Sales** (staff_sale) – Other. Labor expenses as a fraction of sales.

References

Bartlett, Peter L and Shahar Mendelson (2002). "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *Journal of Machine Learning Research* 3.Nov, pp. 463–482. URL: https://www.jmlr.org/papers/volume3/bartlett02a/bartlett02a.pdf.

Beaver, William, Maureen McNichols, and Richard Price (2007). "Delisting returns and their effect on accounting-based market anomalies". In: *Journal of Accounting and Economics* 43.2, pp. 341–368. ISSN: 0165-4101. DOI: https://doi.org/10.1016/j.jacceco.2006.12.002. URL: https://www.sciencedirect.com/science/article/pii/S0165410106000930.

Bradley, Richard C. (2005). "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions". In: *Probability Surveys* 2.none, pp. 107–144. DOI: 10.1214/154957805100000104. URL: https://doi.org/10.1214/154957805100000104.

Bryzgalova, Svetlana, Markus Pelger, and Jason Zhu (2019). "Forest through the trees: Building cross-sections of stock returns". DOI: http://dx.doi.org/10.2139/ssrn.3493458. URL: https://ssrn.com/abstract=3493458.

Chen, Luyang, Markus Pelger, and Jason Zhu (2024). "Deep Learning in Asset Pricing". In: *Management Science* 70.2, pp. 714–750. DOI: 10.1287/mnsc.2023.4695. URL: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2023.4695.

Cho, Youngmin and Lawrence Saul (2009). "Kernel Methods for Deep Learning". In: *Advances in Neural Information Processing Systems*. Ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta. Vol. 22. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2009/file/5751ec3e9a4feab575962e78e006250d-Paper.pdf.

Cochrane, John H (2009). *Asset pricing: Revised edition*. Princeton university press.

— (2011). "Presidential Address: Discount Rates". In: *The Journal of Finance* 66.4, pp. 1047–1108. DOI: https://doi.org/10.1111/j.1540-6261.2011.01671.x.

URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2011.01671.x.

Didisheim, Antoine, Shikun Barry Ke, Bryan T Kelly, and Semyon Malamud (2023). "Complexity in factor pricing models". DOI: http://dx.doi.org/10.2139/ssrn.4388526. URL: https://ssrn.com/abstract=4388526.

Fama, Eugene F. and Kenneth R. French (1993). "Common risk factors in the returns on stocks and bonds". In: *Journal of Financial Economics* 33.1, pp. 3–56. ISSN: 0304-405X. DOI: https://doi.org/10.1016/0304-405X(93)90023-5. URL: https://www.sciencedirect.com/science/article/pii/0304405X93900235.

— (2015). "A five-factor asset pricing model". In: *Journal of Financial Economics* 116.1, pp. 1–22. ISSN: 0304-405X. DOI: https://doi.org/10.1016/j.jfineco.2014.10.010. URL: https://www.sciencedirect.com/science/article/pii/S0304405X14002323.

Gonon, Lukas (2023). "Random Feature Neural Networks Learn Black-Scholes Type PDEs Without Curse of Dimensionality". In: *Journal of Machine Learning Research* 24.189, pp. 1–51. URL: http://jmlr.org/papers/v24/21-0987.html.

Gonon, Lukas, Lyudmila Grigoryeva, and Juan-Pablo Ortega (2020). "Risk Bounds for Reservoir Computing". In: *Journal of Machine Learning Research* 21.240, pp. 1–61. URL: http://jmlr.org/papers/v21/19-902.html.

— (2023). "Approximation bounds for random neural networks and reservoir systems". In: *The Annals of Applied Probability* 33.1, pp. 28–69. DOI: 10.1214/22-AAP1806. URL: https://doi.org/10.1214/22-AAP1806.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (Feb. 2020). "Empirical Asset Pricing via Machine Learning". In: *The Review of Financial Studies* 33.5, pp. 2223–2273. ISSN: 0893-9454. DOI: 10.1093/rfs/hhaa009. eprint: https://academic.oup.com/rfs/article-pdf/33/5/2223/33209812/hhaa009.pdf. URL: https://doi.org/10.1093/rfs/hhaa009.

Györfi, László, Michael Kohler, Adam Krzyzak, Harro Walk, et al. (2002). *A distribution-free theory of nonparametric regression*. Vol. 1. Springer. DOI: https://doi.org/10.1007/b97848.

Hansen, Lars Peter and Ravi Jagannathan (1997). "Assessing Specification Errors in Stochastic Discount Factor Models". In: *The Journal of Finance* 52.2, pp. 557–590. DOI: https://doi.org/10.1111/j.1540-6261.1997.tb04813.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1997.tb04813.x.

URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1997.tb04813.x.

Hardt, Moritz and Benjamin Recht (2022). *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen (2023). "Is There a Replication Crisis in Finance?" In: *The Journal of Finance* 78.5, pp. 2465–2518. DOI: https://doi.org/10.1111/jofi.13249. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13249.

Kelly, Bryan, Boris Kuznetsov, Semyon Malamud, and Teng Andrea Xu (2024). "Large (and Deep) Factor Models". DOI: http://dx.doi.org/10.2139/ssrn.4679269. URL: https://ssrn.com/abstract=4679269.

Kozak, Serhiy (2020). "Kernel trick for the cross-section". DOI: http://dx.doi.org/10.2139/ssrn.3307895. URL: https://ssrn.com/abstract=3307895.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh (2020). "Shrinking the cross-section". In: *Journal of Financial Economics* 135.2, pp. 271–292. ISSN: 0304-405X. DOI: https://doi.org/10.1016/j.jfineco.2019.06.008. URL: https://www.sciencedirect.com/science/article/pii/S0304405X19301655.

Ledoux, Michel and Michel Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Vol. 23. Springer Science & Business Media. DOI: https://doi.org/10.1007/978-3-642-20212-4.

Lettau, Martin and Markus Pelger (2020). "Estimating latent asset-pricing factors". In: *Journal of Econometrics* 218.1, pp. 1–31. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2019.08.012. URL: https://www.sciencedirect.com/science/article/pii/S0304407620300051.

McLean, R. David and Jeffrey Pontiff (2016). "Does Academic Research Destroy Stock Return Predictability?" In: *The Journal of Finance* 71.1, pp. 5–32. DOI: https://doi.org/10.1111/jofi.12365. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12365. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12365.

Mohri, Mehryar and Afshin Rostamizadeh (2008). "Rademacher Complexity Bounds for Non-I.I.D. Processes". In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2008/file/7eacb532570ff6858afd2723755ff790-Paper.pdf.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2018). *Foundations of machine learning*. MIT press.

Novy-Marx, Robert and Mihail Velikov (Nov. 2015). "A Taxonomy of Anomalies and Their Trading Costs". In: *The Review of Financial Studies* 29.1, pp. 104–147. ISSN: 0893-9454. DOI: 10.1093/rfs/hhv063. eprint: https://academic.oup.com/rfs/article-pdf/29/1/104/24451084/hhv063.pdf. URL: https://doi.org/10.1093/rfs/hhv063.

Rahimi, Ali and Benjamin Recht (2007). "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

— (2008a). "Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning". In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2008/file/0efe32849d230d7f53049ddc4a4b0c60-Paper.pdf.

— (2008b). "Uniform approximation of functions with random bases". In: *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 555–561. DOI: 10.1109/ALLERTON.2008.4797607.

Sharpe, William F. (1964). "Capital Asset Prices: A Theory of Market Equilibrium under conditions of risk". In: *The Journal of Finance* 19.3, pp. 425–442. DOI: https://doi.org/10.1111/j.1540-6261.1964.tb02865.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1964.tb02865.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1964.tb02865.x.

Shumway, Tyler (1997). "The Delisting Bias in CRSP Data". In: *The Journal of Finance* 52.1, pp. 327–340. ISSN: 00221082, 15406261. URL: http://www.jstor.org/stable/2329566.

Yu, Bin (1994). "Rates of Convergence for Empirical Processes of Stationary Mixing Sequences". In: *The Annals of Probability* 22.1, pp. 94–116. ISSN: 00911798, 2168894X. URL: http://www.jstor.org/stable/2244496.