New Journal of Physics
The open-access journal for physics

Regularizing portfolio optimization

Susanne Still[1,3] **and Imre Kondor**[2,3]

[1] Information and Computer Sciences, University of Hawaii at Mānoa, Honolulu, Hawaii, USA

[2] Collegium Budapest—Institute for Advanced Study and Department of Physics of Complex Systems, Eötvös University, Budapest, Hungary

E-mail: sstill@hawaii.edu and kondor@colbud.hu

New Journal of Physics **12** (2010) 075034 (15pp)
Received 20 October 2009
Published 30 July 2010
Online at http://www.njp.org/
doi:10.1088/1367-2630/12/7/075034

Abstract. The optimization of large portfolios displays an inherent instability due to estimation error. This poses a fundamental problem, because solutions that are not stable under sample fluctuations may look optimal for a given sample, but are, in effect, very far from optimal with respect to the average risk. In this paper, we approach the problem from the point of view of statistical learning theory. The occurrence of the instability is intimately related to over-fitting, which can be avoided using known regularization methods. We show how regularized portfolio optimization with the expected shortfall as a risk measure is related to support vector regression. The budget constraint dictates a modification. We present the resulting optimization problem and discuss the solution. The L2 norm of the weight vector is used as a regularizer, which corresponds to a diversification 'pressure'. This means that diversification, besides counteracting downward fluctuations in some assets by upward fluctuations in others, is also crucial because it improves the stability of the solution. The approach we provide here allows for the simultaneous treatment of optimization and diversification in one framework that enables the investor to trade off between the two, depending on the size of the available dataset.

[3] Author to whom any correspondence should be addressed.

1367-2630/10/075034+15$30.00

Contents

1. Introduction

Markowitz's portfolio selection theory [1, 2] is one of the pillars of theoretical finance. It has greatly influenced theory and practice in investment, capital allocation, index tracking and a number of other fields. Its two major ingredients are (i) seeking a trade-off between risk and reward and (ii) exploiting the cancelation between fluctuations of (anti-)correlated assets. In the original formulation of the theory, the underlying process was assumed to be multivariate normal. Accordingly, reward was measured in terms of the expected return, and risk in terms of the variance of the portfolio.

The fundamental problem of this scheme (shared by all the other variants that have been introduced since) is that the characteristics of the underlying process generating the distribution of asset prices are not known in practice, and therefore averages are replaced by sums over the available sample. This procedure is well justified as long as the sample size, T (i.e. the length of the available time series for each item), is sufficiently large compared to the size of the portfolio, N (i.e. the number of items). In that limit, sample averages asymptotically converge to the true average due to the central limit theorem.

Unfortunately, the nature of portfolio selection is not compatible with this limit. Institutional portfolios are large, with Ns in the range of hundreds or thousands, while considerations of transaction costs and non-stationarity limit the number of available data points to a couple of hundred at most. Therefore, portfolio selection works in a region where N and T are, at best, of the same order of magnitude. This, however, is not the realm of classical statistical methods. Portfolio optimization is closer to a situation which, borrowing a term from statistical physics, might be termed the 'thermodynamic limit', where N and T tend to infinity such that their ratio remains fixed.

It is evident that portfolio theory struggles with the same fundamental difficulty that underlies every complex modeling and optimization task: the high number of dimensions and the insufficient amount of information available about the system. This difficulty has existed in portfolio selection from the early days, and a plethora of methods have been proposed to cope with it, e.g. single and multi-factor models [3], Bayesian estimators [4]–[17] or, more recently, tools borrowed from random matrix theory [18]–[23]. In the thermodynamic regime, estimation errors are large, sample-to-sample fluctuations are huge, and results obtained from one sample do not generalize well and can be quite misleading concerning the true process.

The same problem has received considerable attention in the area of machine learning. We discuss how the observed instabilities in portfolio optimization (elaborated in section 2) can be understood and remedied by looking at portfolio theory from the point of view of machine learning.

Portfolio optimization is a special case of regression, and therefore can be understood as a machine learning problem (see section 3). In machine learning, as well as in portfolio optimization, one wishes to minimize the *actual risk*, which is the risk (or error) evaluated by taking the ensemble average. This quantity, however, cannot be computed from the data, only the *empirical risk* can. The difference between the two is not necessarily small in the thermodynamic limit, so that a small empirical risk does not automatically guarantee small actual risk [24].

Statistical learning theory [24]–[26] finds upper bounds on the generalization error that hold with a certain accuracy. These error bounds quantify the expected generalization performance of a model, and they decrease with decreasing *capacity* of the function class that is being fitted to the data. Lowering the capacity therefore lowers the error bound and thereby improves generalization. The resulting procedure is also referred to as regularization and essentially prevents over-fitting (see section 4).

In the thermodynamic limit, portfolio optimization needs to be regularized. We show in section 5 how the above-mentioned concepts, which find their practical application in support vector machines [27, 28], can be used for portfolio optimization. Support vector machines constitute an extremely powerful class of learning algorithms, which have met with considerable success. We show that regularized portfolio optimization (RPO), using the expected shortfall as a risk measure, is almost identical to support vector regression (SVR), apart from the budget constraint. We provide the modified optimization problem, which can be solved by linear programming.

In section 6, we discuss the financial meaning of the regularizer: minimizing the L2 norm of the weight vector corresponds to a diversification pressure. We also discuss alternative constraints that could serve as regularizers in the context of portfolio optimization.

Taking this machine learning angle allows one to organize a variety of ideas in the existing literature on portfolio optimization filtering methods into one systematic and well-developed framework. There are basically two choices to be made: (i) which risk measure to use and (ii) which regularizer. These choices result in different methods, because different optimization problems are being solved.

While we focus here on the popular expected shortfall risk measure (in section 5), the variance has a long history as an important risk measure in finance. Several existing filtering methods that use the variance risk measure essentially implement regularization, without necessarily stating so explicitly. The only work we found that explicitly mentions regularization in the context of portfolio optimization [7] has not received much notice in the ensuing, closely related, literature. It is easy to show that when the L2 norm is used as a regularizer, then regularizing portfolio optimization results in a method that is closely related (with the difference of the additional budget constraint) to Bayesian ridge regression, which uses a Gaussian prior on the weights. The work on covariance shrinkage, such as [8]–[11], falls into the same category. Other priors can be used [17], which can be expected to lead to different results (for an insightful comparison see, e.g., [29]). Using the L1 norm has been popularized in statistics as the least absolute shrinkage and selection operator (LASSO) [29], and methods that use any Lp norm are also known as the 'bridge' [30].

2. Preliminaries—instability of classical portfolio optimization

Portfolio optimization in large institutions operates in what we call the thermodynamic limit, where both the number of assets and the number of data points are large, with their ratio a certain, typically not very small, number. The estimation problem for the mean is so serious [31, 32] as to make the trade-off between risk and return largely illusory. Therefore, following a number of authors [8, 9], [33]–[35], we focus on the minimum risk portfolio and drop the usual constraint on the expected return. This is also in line with previous work (see [36] and references therein), and makes the treatment simpler without compromising the main conclusions. An extension of the results to the more general case is straightforward.

Nevertheless, even if we forget about the expected return constraint, the problem still remains that covariances have to be estimated from finite samples. It is an elementary fact from linear algebra that the rank of the empirical $N \times N$ covariance matrix is the smaller of N and T. Therefore, if $T < N$, the covariance matrix is singular and the portfolio selection task becomes meaningless. The point $T = N$ thus separates two regions: for $T > N$, the portfolio problem has a solution, whereas for $T < N$, it does not.

Even if T is larger than N, but not *much* larger, the solution to the minimum variance problem is unstable under sample fluctuations, which means that it is not possible to find the optimal portfolio in this way. This instability of the estimated covariances, and hence of the optimal solutions, has been generally known in the community; however, the full depth of the problem has only been recognized recently, when it was pointed out that the average estimation error diverges at the critical point $N = T$ [37]–[39].

In order to characterize the estimation error, Kondor and co-workers used the ratio q_0^2 between (i) the risk, evaluated at the optimal solution obtained by portfolio optimization using finite data, and (ii) the true minimal risk. This quantity is a measure of generalization performance, with perfect performance when $q_0^2 = 1$, and increasingly bad performance as q_0^2 increases. As found numerically in [38] and demonstrated analytically by random matrix theory techniques in [40], the quantity q_0 is proportional to $(1 - N/T)^{-1/2}$ and diverges when T goes to N from above.

The identification of the point $N = T$ as a phase transition [36, 41] allowed for the establishment of a link between portfolio optimization and the theory of phase transitions, which helped to organize a number of seemingly disparate phenomena into a single coherent picture with a rich conceptual content. For example, it has been shown that the divergence is not a special feature of the variance, but persists under all the other alternative risk measures that have been investigated so far: historical expected shortfall, maximal loss, mean absolute deviation, parametric VaR, expected shortfall and semivariance [36], [41]–[43]. The critical value of the N/T ratio, at which the divergence occurs, depends on the particular risk measure and on any parameter that the risk measure may depend on (such as the confidence level in expected shortfall). However, as a manifestation of universality, the power law governing the divergence of the estimation error is independent of the risk measure [36, 41, 42], the covariance structure of the market [39] and the statistical nature of the underlying process [44]. Ultimately, this line of thought led to the discovery of the instability of coherent risk measures [45].

3. Statistical reasons for the observed instability in portfolio optimization

As mentioned above, for the simplicity and clarity of the treatment, we do not impose a constraint on the expected return, and only look for the global minimum risk portfolio. This task can be formalized as follows. Given a fixed budget, customarily taken to be unity, given T past measurements of the returns of N assets, x_i^k, $i = 1, \ldots, N$, $k = 1, \ldots, T$, and given the risk functional $F(\mathbf{w} \cdot \mathbf{x})$, find a weighted sum (the portfolio), $\mathbf{w} \cdot \mathbf{x}$,[4] such that it minimizes the *actual* risk

$$R(\mathbf{w}) = \langle F(\mathbf{w} \cdot \mathbf{x}) \rangle_{p(\mathbf{x})}, \tag{1}$$

under the constraint that $\sum_i w_i = 1$. The central problem is that one does not know the distribution $p(\mathbf{x})$, which is assumed to underlie the generation of the data. In practice, one then minimizes the *empirical* risk, replacing ensemble averages by sample averages:

$$R_{\text{emp}}(\mathbf{w}) = \frac{1}{T} \sum_{k=1}^{T} F(\mathbf{w} \cdot \mathbf{x}^{(k)}). \tag{2}$$

Now, let us interpret the weight vector as a linear model. The model class given by the linear functions has a *capacity* h, which is a concept that has been introduced by Vapnik and Chervonenkis in order to measure how powerful a learning machine is [24]–[26]. (In the statistical learning literature, a learning machine is thought of as having a function class at its disposal, together with an induction principle and an algorithmic procedure for the implementation thereof [46].) The capacity measures how powerful a function class is, and thereby also how easy it is to learn a model of that class. The rough idea is this: a learning machine has larger capacity if it can potentially fit more different types of datasets. Higher capacity comes, however, at the cost of potentially over-fitting the data. Capacity can be measured, for example, by the Vapnik–Chervonenkis (VC) dimension [24], which is a combinatoric measure that counts how many data points can be separated in all possible ways by any function of a given class.

To make the idea tangible for linear models, we focus on two dimensions ($N = 2$). For each number of points, n, one can choose the geometrical arrangement of the points in the plane freely. Once it is chosen, points are labeled by one of two labels, let us say ‘red’ and ‘blue’. Can a line separate the red points from the blue points for *any* of the 2^n different ways in which the points could be colored? The VC dimension is the largest number of points for which this can be done. Two points can trivially be separated by a line. Three points that are not arranged collinearly can still be separate for any of the eight possible labelings. However, for four points this is no longer the case, since there is no geometrical arrangement for which one could not find a labeling that cannot be separated by a line. The VC dimension is 3, and in general, for linear models in N dimensions, it is $N + 1$ [46, 47].

In the regime in which the number of data points is much larger than the capacity of the learning machine, $h/T \ll 1$, a small empirical risk guarantees a small actual risk [24]. For linear functions through the origin that are otherwise unconstrained, the VC dimension grows with N. In the thermodynamic regime, where N/T is not very small, minimizing the empirical risk does not necessarily guarantee a small actual risk [24]. Therefore, it is not guaranteed to produce a solution that generalizes well to other data drawn from the same underlying distribution.

[4] Notation: bold face symbols are understood to denote vectors.

In solving the optimizing problem that minimizes the *empirical* risk, equation (2) in the regime in which N/T is not very small, portfolio optimization *over-fits* the observed data. It thereby finds a solution that essentially pays attention to the seeming correlations in the data which come from estimation noise due to finite sample effects, rather than from real structure. The solution is thus different for different realizations of the data, and does not necessarily come close to the actual optimal portfolio.

4. Overcoming the instability

The generalization error can be bounded from above (with a certain probability) by the empirical error, plus a confidence term that increases monotonically with some measure of the capacity, and depends on the probability with which the bound holds [48]. Several different bounds have been established, connected with different measures of capacity, see, e.g., [47].

Poor generalization and over-fitting can be improved upon by decreasing the capacity of the model [25, 26], which helps to lower the generalization error. Support vector machines are a powerful class of algorithms that implement this idea.

We suggest that if one wants to find a solution to the portfolio optimization problem in the thermodynamic regime, then one should not only minimize the empirical risk alone, but also constrain the capacity of the portfolio optimizer (the linear model).

The capacity of a linear model is minimized when the length of the weight vector is minimized [25, 26]. Vapnik's concept of *structural risk minimization* [48] results in the support vector algorithm [27, 28], which finds the model with the smallest capacity that is consistent with the data, that is, the model with the smallest $\|w\|^2$. This leads to a convex constrained optimization problem [27, 28], which can be solved using linear programming.

5. Regular portfolio optimization (RPO) with the expected shortfall risk measure

While the original Markowitz's formulation [1] measures risk by the variance, many other risk measures have been proposed since. Today, the most widely used risk measure, both in practice and in regulation, is Value at Risk (VaR) [49, 50]. VaR has, however, been criticized for its lack of convexity, see, e.g., [51]–[53], and an axiomatic approach, leading to the introduction of the class of coherent risk measures, was put forward [51]. Expected shortfall, essentially a conditional average measuring the average loss above a high threshold, has been demonstrated to belong to this class [54]–[56].

Expected shortfall has been steadily gaining in popularity in recent years. The regularization that we propose here is intended to cure its weak point, the sensitivity to sample fluctuations, at least for reasonable values of the ratio N/T.

Choose the risk functional $F(z) = z\theta(z-\alpha_\beta)$, where α_β is a threshold, such that a given fraction β of the (empirical) loss distribution over z lies above α_β. One now wishes to minimize the average over the remaining tail distribution, containing the fraction $\nu := 1-\beta$, and defines the expected shortfall as

$$ES = \min_{\epsilon}\left[\epsilon + \frac{1}{\nu T}\sum_{k=1}^{T}\frac{1}{2}\left(-\epsilon - \mathbf{w}\cdot\mathbf{x}^{(k)} + |-\epsilon - \mathbf{w}\cdot\mathbf{x}^{(k)}|\right)\right]. \qquad (3)$$

The term in the sum implements the θ function, while ν in the denominator ensures normalization of the tail distribution. It has been pointed out [57] that this optimization problem maps onto solving the linear program:

$$\min_{\mathbf{w},\xi,\epsilon}\left[\frac{1}{T}\sum_{k=1}^{T}\xi_k+\nu\epsilon\right] \tag{4}$$

$$\text{s.t.}\quad \mathbf{w}\cdot\mathbf{x}^{(k)}+\epsilon+\xi_k\geqslant 0;\qquad \xi_k\geqslant 0; \tag{5}$$

$$\sum_i w_i=1. \tag{6}$$

We propose implementing regularization by including the minimization of $\|\mathbf{w}\|^2$. This can be done using a Lagrange multiplier, C, to control the trade-off—as we relax the constraint on the length of the weight vector, we can, of course, make the empirical error go to zero and retrieve the solution to the minimal expected shortfall problem. The new optimization problem reads

$$\min_{\mathbf{w},\xi,\epsilon}\left[\frac{1}{2}\|\mathbf{w}\|^2+C\left(\frac{1}{T}\sum_{k=1}^{T}\xi_k+\nu\epsilon\right)\right] \tag{7}$$

$$\text{s.t.}\quad -\mathbf{w}\cdot\mathbf{x}^{(k)}\leqslant\epsilon+\xi_k; \tag{8}$$

$$\xi_k\geqslant 0;\quad \epsilon\geqslant 0; \tag{9}$$

$$\sum_i w_i=1. \tag{10}$$

The problem is mathematically almost identical to an SVR algorithm called ν-SVR. There are two differences: (i) the budget constraint is added and (ii) the loss function is asymmetric. Expected shortfall is an asymmetric version of the ϵ-intensive loss, used in SVR, defined as the maximum of $\{0; |f(\mathbf{x})-y|-\epsilon\}$, where $f(\mathbf{x})$ is the interpolant and y is the measured value (response). In that sense, ϵ measures an allowable error below which deviations are discarded[5].

The use of asymmetric risk measures in finance is motivated by the consideration that investors are not afraid of upside fluctuations. However, to make the relationship to SVR as clear as possible, we will first solve the more general symmetrized problem, before restricting our treatment to the completely asymmetric case, corresponding to expected shortfall. In addition, one may argue that focusing exclusively on large negative fluctuations might not be advisable even from a financial point of view, especially when one does not have sufficiently large samples. In a relatively small sample, it may happen that a particular item, or a certain combination of items, dominates the rest, i.e. produces a larger return than any other item in the portfolio at each time point, even though no such dominance exists on longer time scales. The probability of such an apparent arbitrage increases with the ratio N/T, and when it occurs it may encourage an investor acting on a lopsided risk measure to take up very large long positions in the dominating item(s), which may turn out to be detrimental in the long run. This is the essence of the argument that has led to the discovery of the instability of coherent and downside risk measures [43, 45].

[5] The similarity between minimum expected shortfall and the Eν-SVM algorithm [58] was used in [59] to provide a motivation for Eν-SVM. They considered a different problem, where the budget constraint was disregarded, because their focus was not on portfolio optimization, but rather on improving the Eν-SVM algorithm.

According to the above, let us consider the general case where positive deviations are also penalized. The objective function, equation (7), then becomes

$$\min_{\mathbf{w},\xi,\epsilon}\left[\frac{1}{2}\|\mathbf{w}\|^2 + C\left(\frac{1}{T}\sum_{k=1}^{T}\left(\xi_k+\xi_k^*\right)+\nu\epsilon\right)\right], \tag{11}$$

and additional constraints have to be added to equations (8)–(10):

$$\mathbf{w}\cdot\mathbf{x}^{(k)} \leqslant \epsilon + \xi_k^*; \quad \xi_k^* \geqslant 0. \tag{12}$$

This problem corresponds to ν-SVR, a well-understood regression method [60], with the only difference being that the budget constraint, equation (10), is added here. In the finance context, the associated loss might be called *symmetric tail average* (STA). Solving the regularized expected shortfall minimization problem, equations (7)–(10), is a special case of solving the regularized STA minimization problem, equation (11), with the constraints equations (8)–(10) and (12). Therefore, we solve the more general problem first (section 5.1), before providing, in section 5.2, the solution to the regularized expected shortfall, equations (7)–(10).

5.1. Regularized symmetric tail average (STA) minimization

The solution to the regularized STA problem, equation (11) with the constraints equations (8)–(10) and (12), is found in analogy to SVR, following [60], by writing down the Lagrangian, using Lagrange multipliers, $\{\alpha,\alpha^*,\gamma,\lambda,\eta,\eta^*\}$, for the constraints. The solution is then a saddle point, i.e. minimum over primal and maximum over dual variables. The Lagrangian is different from the one that arises in ν-SVR in that it is modified by the budget constraint:

$$\begin{aligned} L[\mathbf{w},\xi,\xi^*,\epsilon,\alpha,\alpha^*,\gamma,\lambda,\eta,\eta^*] &= \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{T}\sum_{k=1}^{T}(\xi_k+\xi_k^*) + C\nu\epsilon - \lambda\epsilon + \gamma\left(\sum_i w_i - 1\right) \\ &\quad + \sum_{k=1}^{T}\alpha_k^*(\mathbf{w}\cdot\mathbf{x}^{(k)} - \epsilon - \xi_k^*) - \sum_{k=1}^{T}\alpha_k(\mathbf{w}\cdot\mathbf{x}^{(k)} + \epsilon + \xi_k) - \sum_{k=1}^{T}(\eta_k\xi_k + \eta_k^*\xi_k^*), \end{aligned} \tag{13}$$

$$\begin{aligned} &= F[\mathbf{w}] + \epsilon\left(C\nu - \lambda - \sum_{k=1}^{T}(\alpha_k+\alpha_k^*)\right) - \gamma \\ &\quad + \sum_{k=1}^{T}\left[\xi_k\left(\frac{C}{T}-\alpha_k-\eta_k\right) + \xi_k^*\left(\frac{C}{T}-\alpha_k^*-\eta_k^*\right)\right], \end{aligned} \tag{14}$$

with

$$F[\mathbf{w}] = \mathbf{w}\cdot\left(\frac{1}{2}\mathbf{w} - \left(\sum_{k=1}^{T}(\alpha_k-\alpha_k^*)\mathbf{x}^{(k)} - \gamma\mathbf{1}\right)\right), \tag{15}$$

where $\mathbf{1}$ denotes the unit vector of length N. Setting the derivative of the Lagrangian w.r.t. $\mathbf{w}$ to zero gives

$$\mathbf{w}_{\text{opt}} = \sum_{k=1}^{T}(\alpha_k-\alpha_k^*)\mathbf{x}^{(k)} - \gamma\mathbf{1}. \tag{16}$$

This solution for the optimal portfolio is sparse in the sense that, due to the Karush–Kuhn–Tucker conditions (see, e.g., [61]), only those points contribute to the optimal portfolio weights, for which the inequality constraints in (8), and the corresponding constraints in equation (12), are met exactly. The solution of $\mathbf{w}_{\text{opt}}$ contains only those points, and effectively ignores the rest. This sparsity contributes to the stability of the solution. RPO operates, in contrast to general regression, with a fixed budget. As a consequence, the Lagrange multiplier γ now appears in the optimal solution, equation (16). Compared to the optimal solution in SVR, $\mathbf{w}_{\text{SV}}$, the solution vector under the budget constraint, $\mathbf{w}_{\text{RPO}}$, is shifted by γ:

$$\mathbf{w}_{\text{RPO}} = \mathbf{w}_{\text{SV}} - \gamma \mathbf{1}. \tag{17}$$

Let us now consider the dual problem. The dual is, in general, a function of the dual variables, which are here $\{\alpha, \alpha^*, \gamma, \lambda, \eta, \eta^*\}$, although we will see in the following that some of these variables drop out. The dual is defined as $D := \min_{\mathbf{w},\xi,\xi^*,\epsilon} L[\mathbf{w}, \xi, \xi^*, \epsilon, \alpha, \alpha^*, \gamma, \lambda, \eta, \eta^*]$, and the dual problem is then to maximize D over the dual variables. We can replace the minimization over $\mathbf{w}$ by evaluating the Lagrangian at $\mathbf{w}_{\text{opt}}$. For that, we have to evaluate

$$F[\mathbf{w}_{\text{opt}}] = -\frac{1}{2}\|\mathbf{w}_{\text{opt}}\|^2 \tag{18}$$

$$= \left[-\frac{1}{2}\left(\sum_{k=1}^{T} (\alpha_k - \alpha_k^*)\mathbf{x}^{(k)} - \gamma \mathbf{1} \right)^2 \right]. \tag{19}$$

For the other terms in the Lagrangian, we have to consider different cases.

(i) If $(C\nu - \lambda - \sum_{k=1}^{T}(\alpha_k + \alpha_k^*)) < 0$, then L can be minimized by letting $\epsilon \to \infty$, which means that $D = -\infty$.

(ii) If $(C\nu - \lambda - \sum_{k=1}^{T}(\alpha_k + \alpha_k^*)) \geqslant 0$, the term $\epsilon(C\nu - \lambda - \sum_{k=1}^{T}(\alpha_k + \alpha_k^*))$ vanishes. This is because if equality holds, this is trivially true, and if the inequality holds strictly, then L can be minimized by setting $\epsilon = 0$.

Similarly, for the other constraints (the notation $(*)$ means that this is true for variables with and without the asterisk) we reason as follows.

(i) If $(\frac{C}{T} - \alpha_k^{(*)} - \eta_k^{(*)}) < 0$, then L can be minimized by letting $\xi_k^{(*)} \to \infty$, which means that $D = -\infty$.

(ii) If $(\frac{C}{T} - \alpha_k^{(*)} - \eta_k^{(*)}) \geqslant 0$, then $\xi_k\, (\frac{C}{T} - \alpha_k^{(*)} - \eta_k^{(*)}) = 0$. This is because if the inequality holds strictly, then L can be minimized by $\xi_k^{(*)} = 0$. If the equality holds, then it is trivially true.

Altogether we have that either $D = -\infty$, or

$$D(\alpha, \alpha^*, \gamma) = \min_{\xi,\xi^*,\epsilon} F[\mathbf{w}_{\text{opt}}(\alpha, \alpha^*, \gamma)] = -\frac{1}{2}\|\mathbf{w}_{\text{opt}}\|^2 - \gamma \tag{20}$$

and

$$\sum_{k=1}^{T} (\alpha_k^* + \alpha_k) \leqslant C\nu - \lambda \tag{21}$$

and

$$\alpha_k^{(*)} + \eta_k^{(*)} \leqslant \frac{C}{T}. \tag{22}$$

Note that the variables $\xi_k^{(*)}$, $\eta_k^{(*)}$, ϵ, λ do not appear in $F[\mathbf{w}_{\mathrm{opt}}(\alpha, \alpha^*, \gamma)]$. The dual problem is therefore given by

$$\max_{\alpha,\alpha^*,\gamma} \left[-\frac{1}{2} \left(\sum_{k=1}^{T} (\alpha_k - \alpha_k^*) \mathbf{x}^{(k)} - \gamma \mathbf{1} \right)^2 - \gamma \right] \tag{23}$$

$$\text{s.t.} \quad \{\alpha_k, \alpha_k^*\} \in \left[0, \frac{C}{T} \right], \tag{24}$$

$$\sum_{k=1}^{T} (\alpha_k^* + \alpha_k) \leqslant C\nu. \tag{25}$$

The budget constraint implies that

$$\gamma = \frac{1}{N} \left(\sum_{k=1}^{T} (\alpha_k - \alpha_k^*) \sum_{i=1}^{N} x_i^{(k)} - 1 \right). \tag{26}$$

The optimal projection (= optimal portfolio) is given by

$$\mathbf{w}_{\mathrm{opt}} \cdot \mathbf{x} = \sum_{k=1}^{T} (\alpha_k - \alpha_k^*) \mathbf{x}^{(k)} \cdot \mathbf{x} - \frac{1}{N} \left(\sum_{k=1}^{T} (\alpha_k - \alpha_k^*) \sum_{i=1}^{N} x_i^{(k)} - 1 \right) \mathbf{1} \cdot \mathbf{x}. \tag{27}$$

For $N \to \infty$, the second term vanishes and the solution is the same as the solution in SVR. Note that the kernel-trick (see, e.g., [47]), which is used in support vector machines to find nonlinear models, hinges on the fact that only dot products of input vectors appear in the support vector expansion of the solution. As a consequence of the budget constraint, one can no longer use the kernel-trick (compare equation (27)). As long as we disregard derivatives, this is not a problem for portfolio optimization. Keep in mind, however, that the budget constraint introduces this otherwise undesirable property.

Support vector algorithms typically solve the dual form of the problem (for a recent survey, see [62]), which is in our case given by

$$\max_{\alpha,\alpha^*} -\frac{1}{2} \left[\sum_{k=1}^{T} \sum_{l=1}^{T} (\alpha_k - \alpha_k^*)(\alpha_l - \alpha_l^*) \left(\mathbf{x}^{(k)} \mathbf{x}^{(l)} - \frac{1}{N} \sum_{i=1}^{N} x_i^{(k)} \sum_{i=1}^{N} x_i^{(l)} \right) \right.$$
$$\left. + \sum_{k=1}^{T} (\alpha_k - \alpha_k^*) \frac{1}{N} \sum_{i=1}^{N} x_i^{(k)} \right] \tag{28}$$

The solution can be found numerically by linear programming, for which software packages are available [63]. This modified SVM-type problem can be solved by appropriately modifying existing methods. Solvers such as the ones discussed in [62, 64] can be used, but have to be adapted to this specific problem.

The regularized STA minimization problem (equation (11) with the constraints equations (8)–(10) and (12)) is, as we have shown here, directly related to SVR, which uses the ϵ-insensitive loss function. The ϵ-insensitive loss is stable to local changes for data points that fall outside the range specified by ϵ. This point is elaborated in section 3 in [60], and relates this method to robust estimation of the mean. It can also be extended to robust estimation of quantiles [60] by scaling of the slack variables ξ_k by μ and ξ_k^* by $1 - \mu$.

This scaling translates directly to the portfolio optimization problem, which is an extreme case: downside risk measures penalize only loss, not gain. The asymmetry in the loss function corresponds to $\mu = 1$.

5.2. Regularized expected shortfall

By this final change, we arrive at the RPO problem, equations (7)–(10), which we originally set out to solve. This is now easily solved in analogy to the previous paragraphs: the slack variables ξ_k^* disappear, together with the respective Lagrange multipliers that enforce constraints, including α_k^*. The optimal solution is now

$$\mathbf{w}_{\text{opt}} = \sum_{k=1}^{T} \alpha_k \mathbf{x}^{(k)} - \gamma \mathbf{1}, \tag{29}$$

with

$$\gamma = \frac{1}{N}\left(\sum_{k=1}^{T} \alpha_k \sum_{i=1}^{N} x_i^{(k)} - 1 \right). \tag{30}$$

The dual problem is given by

$$\max_{\alpha_k} -\frac{1}{2}\left[\sum_{k=1}^{T}\sum_{l=1}^{T} \alpha_k \alpha_l \left(\mathbf{x}^{(k)}\mathbf{x}^{(l)} - \frac{1}{N}\sum_{i=1}^{N} x_i^{(k)} \sum_{i=1}^{N} x_i^{(l)} \right) \right]$$

$$+ \sum_{k=1}^{T} \alpha k \frac{1}{N} \sum_{i=1}^{N} x_i^{(k)}, \tag{31}$$

which, like its symmetric counterpart, equation (28), can be solved by adjusting existing algorithms.

The formalism provides a free parameter, C, to set the balance between the original risk function and the regularizer. Its choice may depend on a number of factors, such as the investor's time horizon, the nature of the underlying data and, crucially, the ratio N/T. Intuitively, there must be a maximum allowable value $C_{\max}(N/T)$ for C, such that when one puts more emphasis on the data, $C > C_{\max}(N/T)$, then over-fitting will occur with high probability. It would be desirable to know an analytic expression for (a bound on) $C_{\max}(N/T)$. In practice, cross-validation methods are often employed in machine learning to set the value of C. Those methods are not free of problems (see, e.g., the treatment in [65]), and the optimal choice of this parameter remains an open problem.

6. Regularization corresponds to portfolio diversification

Above, we have controlled the capacity of the linear model by minimizing the L2 norm of the portfolio weight vector. In the finance context, minimizing

$$\|\mathbf{w}\|^2 = \sum_i w_i^2 \simeq \frac{1}{N_{\text{eff}}} \tag{32}$$

corresponds roughly to maximizing the effective number of assets, N_{eff}, i.e. to exerting a *pressure* towards portfolio diversification [66]. We conclude that diversification of the portfolio is crucial, because it serves to counteract the observed instability by acting as a regularizer.

Other constraints that penalize the length of the weight vector could alternatively be considered as a regularizer, in particular any Lp norm. The budget constraint *alone*, however, does not suffice as a regularizer, since it does not constrain the length of the weight vector. Adding a ban on short selling, $w_i \geqslant 0$, to the budget constraint, $\sum_i w_i = 1$, limits the allowable solutions to a finite volume in the space of weights and is equivalent to requiring that $\sum_i |w_i| \leqslant 1$.[6] It thereby imposes a limit on the L1 norm, that is, on the sum of the absolute amplitudes of long and short positions.

One may argue that it may be a good idea to use the L1 norm instead of the L2 norm, because that may make the solution sparser. However, the L1 norm has a tendency to make some of the weights vanish. Indeed, it has been shown that in the orthonormal design case (using the variance as the risk measure), an L1 regularizer will set some of the weights to zero, while an L2 regularizer will scale all the weights [29]. The spontaneous reduction of portfolio size has also been demonstrated in numerical simulations [67]: as one goes deeper and deeper into the regime where T is significantly smaller than N, under a ban on short selling, more and more of the weights will become zero. The same 'freezing out' of the weights has been observed in portfolio optimization [68] as an empirical fact.

It is important to stress that the vanishing of some of the weights does not reflect any structural property of the objective function; it is just a random effect: as clearly demonstrated by simulations [67], for a different sample a different set of weights vanishes. The angle of the weight vector fluctuates wildly from sample to sample. (The behavior of the solutions is similar for other limit systems as well.) This means that the solutions will be determined by the limit system and the random sample rather than by the structure of the market. So the underlying instability is merely 'masked', in that the solutions do not run away to infinity, but they are still unstable under sample fluctuations when T is too small. As it is certainly not in the interest of the investor to obtain a portfolio solution that sets weights to zero on the basis of unreliable information from small samples, the above observations speak strongly in favor of using the L2 norm over the L1 norm.

7. Conclusion

We have made the observation that the optimization of large portfolios minimizes the empirical risk in a regime where the dataset size is similar to the size of the portfolio. In that regime, a small empirical risk does not necessarily guarantee a small actual risk [24]. In this sense, naive portfolio optimization over-fits the data. Regularization can overcome this problem by reducing the capacity of the considered model class.

RPO has choices to make, not only about the risk function but also about the regularizer. Here, we have focussed on the increasingly popular expected shortfall risk measure. Using the L2 norm as a regularizer leads to a convex optimization problem that can be solved by linear programming. We have shown that RPO is then a variant of SVR. The differences are the asymmetry, due to the tolerance to large positive deviations, and the budget constraint, which is not present in regression.

Our treatment provides a novel insight into why diversification is so important. The L2 regularizer implements a pressure towards portfolio diversification. Therefore, from a statistical

[6] This point has been made independently in [17]. To see that the constraints $\sum_i w_i = 1 \wedge w_i \geqslant 0, \forall i$ are equivalent to $\sum_i |w_i| \leqslant 1 \wedge \sum_i w_i = 1$, firstly, note that $\sum_i w_i = 1 \wedge w_i \geqslant 0, \forall i \Rightarrow \sum_i |w_i| = 1$. Secondly, note that $\sum_i |w_i| \leqslant 1 \wedge \sum_i w_i = 1 \Rightarrow \sum_i (|w_i| - w_i) \leqslant 0 \Leftrightarrow \sum_{i\in\mathcal{I}} |w_i| \leqslant 0$, where $\mathcal{I} := \{i : w_i < 0\}$, and this can only be true if $\mathcal{I} = \emptyset$.

point of view, diversification is important as it is one way to control the capacity of the portfolio optimizer and thereby to find a solution that is more stable, and hence meaningful.

In summary, the method that we have outlined in this paper allows for the unified treatment of optimization and diversification in one principled formalism. It shows how known methods from modern statistics can be used to improve the practice of portfolio optimization.

Acknowledgments

We thank Leon Bottou for helpful discussions and comments on the manuscript. This work has been supported by the 'Cooperative Center for Communication Networks Data Analysis', a NAP project sponsored by the National Office of Research and Technology under grant no. KCKHA005. SS thanks the Collegium Budapest for hosting her during this collaboration, and the community at the Collegium for providing a creative and inspiring atmosphere.

References

[1] Markowitz H 1952 Portfolio selection *J. Finance* **7** 77–91

[2] Markowitz H 1959 *Portfolio Selection: Efficient Diversification of Investments* (New York: Wiley)

[3] Elton E J and Gruber M J 1995 *Modern Portfolio Theory and Investment Analysis* (New York: Wiley)

[4] Jobson J D and Korkie B 1979 Improved estimation for Markowitz portfolios using James–Stein type estimators *Proc. Am. Stat. Assoc. (Bus. Econ. Stat.)* **1** 279–84

[5] Jorion P 1986 Bayes–Stein estimation for portfolio analysis *J. Financ. Quant. Anal.* **21** 279–92

[6] Frost P A and Savarino J E 1986 An empirical Bayes approach to efficient portfolio selection *J. Financ. Quant. Anal.* **21** 293–305

[7] Macrae R and Watkins C 1999 Safe portfolio optimization *Proc. IX Int. Symp. of Applied Stochastic Models and Data Analysis: Quantitative Methods in Business and Industry Society, ASMDA-99 (14–17 June 1999, Lisbon, Portugal)* ed H Bacelar-Nicolau, F C Nicolau and J Janssen (Portugal: INE, Statistics National Institute) p. 435

[8] Jagannathan R and Ma T 2003 Risk reduction in large portfolios: why imposing the wrong constraints helps *J. Finance* **58** 1651–84

[9] Ledoit O and Wolf M 2003 Improved estimation of the covariance matrix of stock returns with an application to portfolio selection *J. Empir. Finance* **10** 603–21

[10] Ledoit O and Wolf M 2004 A well-conditioned estimator for large-dimensional covariance matrices *J. Multivariate Anal.* **88** 365–411

[11] Ledoit O and Wolf M 2004 Honey, I shrunk the sample covariance matrix *J. Portfolio Manage.* **31** 110

[12] DeMiguel V, Garlappi L and Uppal R 2009 Optimal versus naive diversification: how inefficient is the 1/N portfolio strategy? *Rev. Financ. Stud.* **22.5** 1915–1953

[13] Garlappi L, Uppal R and Wang T 2007 Portfolio selection with parameter and model uncertainty: a multi-prior approach *Rev. Financ. Stud.* **20** 41–81

[14] Golosnoy V and Okhrin Y 2007 Multivariate shrinkage for optimal portfolio weights *Eur. J. Finance* **13** 441–58

[15] Kan R and Zhou G 2007 Optimal portfolio choice with parameter uncertainty. *J. Financ. Quant. Anal.* **42** 621–56

[16] Frahm G and Memmel Ch 2009 Dominating estimators for the global minimum variance portfolio, 2009. Deutsche Bundesbank, Discussion Paper, Series 2: Banking and Financial Studies http://ideas.repec.org/p/zbw/bubdp2/200901.html

[17] DeMiguel V, Garlappi L, Nogales F J and Uppal R 2009 A generalized approach to portfolio optimization: improving performance by constraining portfolio norms *Manage. Sci.* **55** 798–812

[18] Laloux L, Cizeau P, Bouchaud J-Ph and Potters M 1999 Noise dressing of financial correlation matrices *Phys. Rev. Lett.* **83** 1467–70

[19] Plerou V, Gopikrishnan P, Rosenow B, Amaral L A N and Stanley H E 1999 Universal and non-universal properties of cross-correlations in financial time series *Phys. Rev. Lett.* **83** 1471

[20] Laloux L, Cizeau P, Bouchaud J-P and Potters M 2000 Random matrix theory and financial correlations *Int. J. Theor. Appl. Finance* **3** 391

[21] Plerou V, Gopikrishnan P, Rosenow B, Amaral L A N, Guhr T and Stanley H E 2002 Random matrix approach to cross-correlations in financial data *Phys. Rev.* E **65** 066126

[22] Burda Z, Goerlich A and Jarosz A 2004 Signal and noise in correlation matrix *Physica* A **343** 295

[23] Potters M and Bouchaud J-Ph 2005 Financial applications of random matrix theory: old laces and new pieces *Acta Phys. Pol.* B **36** 2767

[24] Vapnik V and Chervonenkis A 1971 On the uniform convergence of relative frequencies of events to their probabilities *Theory Probab. Appl.* **16** 264–80

[25] Vapnik V 1995 *The Nature of Statistical Learning Theory* (New York: Springer)

[26] Vapnik V 1998 *Statistical Learning Theory* (New York: Wiley)

[27] Boser B E, Guyon I M and Vapnik V N 1992 A training algorithm for optimal margin classifiers *Proc. 5th Annu. ACM Workshop on Computational Learning Theory* ed D Haussler (New York: ACM) pp 144–52

[28] Cortes C and Vapnik V 1995 Support vector networks *Mach. Learn.* **20** 273–97

[29] Tibshirani R 1996 Regression shrinkage and selection via the lasso *J. R. Stat. Soc.* B **58** 267–88

[30] Frank I and Friedman J 1993 A statistical view of some chemometrics regression tools *Technometrics* **35** 109–48

[31] Chopra V K and Ziemba W T 1993 The effect of errors in means, variances and covariances on optimal portfolio choice *J. Portfolio Manage.* **19** 611

[32] Merton R C 1980 On estimating the expected return on the market: an exploratory investigation *J. Financ. Econ.* **8** 323361

[33] Okhrin Y and Schmied W 2006 Distribution properties of portfolio weights *J. Econometrics* **134** 235–56

[34] Kempf A and Memmel C 2006 Estimating the global minimum variance portfolio *Schmalenbach Bus. Rev.* **58** 332–48

[35] Frahm G 2008 Linear statistical inference for global and local minimum variance portfolios *Stat. Papers* doi: 10.1007/s00362-008-0170-z

[36] Kondor I, Pafka S and Nagy G 2007 Noise sensitivity of portfolio selection under various risk measures *J. Bank. Finance* **31** 1545–73

[37] Pafka S and Kondor I 2002 Noisy covariance matrices and portfolio optimization *Eur. Phys. J.* B **27** 277–80

[38] Pafka S and Kondor I 2003 Noisy covariance matrices and portfolio optimization ii *Physica* A **319** 487–94

[39] Pafka S and Kondor I 2004 Estimated correlation matrices and portfolio optimization *Physica* A **343** 623–34

[40] Burda Z, Jurkiewicz J and Nowak M A 2003 Is econophysics a solid science? *Acta Phys. Pol.* B **34** 87–132

[41] Mezard M, Ciliberti S and Kondor I 2007 On the feasibility of portfolio optimization under expected shortfall *Quant. Finance* **7** 389–96

[42] Mezard M and Ciliberti S 2007 Risk minimization through portfolio replication *Eur. Phys. J.* B **57** 175

[43] Varga-Haszonits I and Kondor I 2008 The instability of downside risk measures *J. Stat. Mech.* P12007

[44] Varga-Haszonits I and Kondor I 2007 Noise sensitivity of portfolio selection in constant conditional correlation GARCH models *Physica* A **385** 307–18

[45] Kondor I and Varga-Haszonits I 2008 Feasibility of portfolio optimization under coherent risk measures *Quant. Finance* submitted

[46] Schölkopf B 1997 *Support Vector Learning (GMD-Bericht* 287) (München, Germany: Oldenbourg) http://www.kernel-machines.org/papers/book_ref.ps.gz

[47] Schölkopf B, Burges C J C and Smola A J 1999 *Advances in Kernel Methods Support—Vector Learning* (Cambridge, MA: MIT Press)

[48] Vapnik V and Chervonenkis A 1979 *Theory of Pattern Recognition* (Moscow: Nauka) (In Russian. German translation available from Akademie-Verlag, Berlin: 1979)

[49] Jorion P 2000 *VaR: The New Benchmark for Managing Financial Risk* (New York: McGraw-Hill)

[50] Morgan J P and Reuters Riskmetrics Technical Document available at http://www.riskmetrics.com

[51] Artzner P, Delbaen F, Eber J-M and Heath D 1999 Coherent measures of risk *Math. Finance* **9** 203–28

[52] Embrechts P 2000 Extreme value theory: potential and limitations as an integrated risk measurement tool *Derivatives Use, Trading Regul.* **6** 449–56

[53] Acerbi C, Nordio C and Sirtori C 2001 Expected shortfall as a tool for financial risk management unpublished

[54] Acerbi C 2000 Spectral measures of risk: a coherent representation of subjective risk aversion *J. Bank. Finance* **26** 1505–18

[55] Acerbi C and Tasche D 2002 On the coherence of expected shortfall *J. Bank. Finance* **26** 1487–503

[56] Acerbi C 2004 Coherent representations of subjective risk-aversion *Risk Measures for the 21st Century* ed G Szegö (New York: Wiley)

[57] Rockafellar R T and Uryasev S 2000 Optimization of conditional value-at-risk *J. Risk* **2** 21–41

[58] Perez-Cruz F, Weston J, Herrmann D J L and Schölkopf B 2003 Extension of the nu-SVM range for classification *Advances in Learning Theory: Methods, Models and Applications (NATO Science Series III: Computer and Systems Sciences* vol 190) (Amsterdam: IOS Press) pp 179–96

[59] Takeda A and Sugiyama M 2008 ν-support vector machine as conditional value-at-risk minimization *Proc. 25th Int. Conf. on Machine learning (ICML)* vol 307 ed A McCallum and S Roweis (Omnipress) pp 1056–1063

[60] Schölkopf B, Smola A J, Williamson R C and Bartlett P L 2000 New support vector algorithms *Neural Comput.* **12** 1207–45

[61] Bertsekas D P 1995 *Nonlinear Programming* (Belmont, MA: Athena Scientific)

[62] Bottou L and Lin C-J 2007 Support vector machine solvers *Large Scale Kernel Machines* ed L Bottou, O Chapelle, D DeCoste and J Weston (Cambridge, MA: MIT Press) pp 301–20

[63] Vanderbei R J *LOQO User's Manual. X.* Software available at http://www.princeton.edu/rvdb/loqo/LOQO.html

[64] Bordes A, Ertekin S, Weston J and Bottou L 2005 Fast kernel classifiers with online and active learning *J. Mach. Learn. Res.* **6** 1579–619

[65] Bengio Y and Grandvalet Y 2004 No unbiased estimator of the variance of K-fold cross-validation *Advances in Neural Information Processing Systems 16 (NIPS'03)* ed S Becker, L Saul and B Schölkopf (Cambridge, MA: MIT Press)

[66] Bouchaud J-Ph and Potters M 2000 *Theory of Financial Risk From—Statistical Physics to Risk Management* (Cambridge: Cambridge University Press)

[67] Gulyas N and Kondor I 2007 Portfolio instability and linear constraints *Physica* A unpublished

[68] Scherer B and Martin R D 2005 *Introduction to Modern Portfolio Optimization With NUOPT and S-PLUS* (Berlin: Springer)