

# Regularizing Portfolio Optimization.

**Susanne Still**

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

E-mail: [sstill@hawaii.edu](mailto:sstill@hawaii.edu)

**Imre Kondor**

Collegium Budapest–Institute for Advanced Study, Budapest, Hungary

E-mail: [kondor@colbud.hu](mailto:kondor@colbud.hu)

## **Abstract.**

The optimization of large portfolios displays an inherent instability 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. We approach this problem from the point of view of statistical learning theory. The occurrence of this instability is intimately related to over-fitting which can be avoided using known regularization methods. We show how regularized portfolio optimization using 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. In the finance context this corresponds to a diversification "pressure", which 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 data set.

## 1. Introduction

Markowitz' portfolio selection theory [1, 2] is one of the pillars of theoretical finance. It has greatly influenced the thinking 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 cancellation 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, 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  $N$ 's in the range of hundreds or thousands, while considerations of transaction costs and non-stationarity limit the number of data points  $T$  available for computing covariances to a couple of hundreds 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 rather closer to a situation which, by 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 is underlying basically every complex modeling and optimization task: the high number of dimensions and the insufficient amount of information available about the system. This difficulty has been around 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, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15], or, more recently, tools borrowed from random matrix theory [16, 17, 18, 19, 20, 21]. In the thermodynamic regime, estimation errors are large, sample to sample fluctuations are huge, 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 and in modern statistics. We discuss how the observed instabilities in portfolio optimization (elaborated in Section 2) can be understood (Section 3) and remedied (Section 4) by looking at portfolio theory from the point of view of machine learning.

Our logic is this: We note in Section 3 that since portfolio optimization is a special case of regression, it can be viewed as a machine learning problem. 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, can not be computed from the data, only the *empirical risk* can. The difference between the two is not necessarily small in the thermodynamic limit. Statistical learning theory [22, 23, 24] 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. This procedure is often 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 [25, 26], can be used for portfolio optimization. Support vector machines constitute an extremely powerful class of learning machines which have met with considerable success. We show that regularized portfolio optimization, using the expected shortfall as a risk measure is almost identical to support vector regression, apart from the budget constraint. We provide the modified optimization problem and discuss its solution.

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 can serve as regularizers in the context of portfolio optimization.

Taking this machine learning angle allows one to organize a variety of ideas, insights and methods 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. Regularized portfolio optimization with the variance risk measure has been implemented, under different names, for example, by Bayesian estimators and shrinkage methods [4, 5, 6, 27, 7, 8, 9, 10, 11, 12, 13, 14, 15]. When the L2 norm is used as a regularizer, it is closely related to Bayesian ridge regression (with the difference of the additional budget constraint), which uses a Gaussian prior. Other norms can be used in place of the L2 norm [15]. In statistics, using the L1 norm has been popularized as the "LASSO" (least absolute shrinkage and selection operator) [28], and methods that use any  $L_p$  norm are also known as the "bridge" [29].

## 2. Preliminaries – Instability of classical portfolio optimization.

Portfolio optimization in large institutions operates in what we called 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 [30, 31] as to make the trade-off between risk and

return largely illusory. Therefore, following a number of authors [7, 32, 33, 34, 35], we focus on the minimum variance 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, 38, 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, 42, 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 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:  $x_i^k$ ,  $i = 1, \dots, N$ ,  $k = 1, \dots, T$ , and given the risk functional  $F(\mathbf{w} \cdot \mathbf{x})$ , find a weighted sum (the portfolio),  $\mathbf{w} \cdot \mathbf{x}$ ,<sup>‡</sup> such that it minimizes the *actual* risk

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

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 underly the generation of the data. In practice, one then minimizes the *empirical* risk, replacing ensemble averages by sample averages:

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

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 [22, 23, 24]. (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 rough idea is this: a function class has larger capacity if it can potentially fit more different types of data sets. Higher capacity comes, however, at the cost of potentially over fitting the data. The capacity measures how powerful a function class is, and thereby also how easy it is to learn a model of that class. Capacity can be measured, for example, by the Vapnik-Chervonenkis (VC-) dimension [22], 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, 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, 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,  $h = n_{\text{max}}$ , for which this can be done. Two points can trivially be separated by a line. Three points that are not arranged collinear can still be separate for any of the 8 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 can not 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 are much larger than the capacity of the learning machine,  $h/T \ll 1$ , a small empirical risk guarantees small actual risk

<sup>‡</sup> Notation: bold face symbols are understood to denote vectors.

[22]. 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 [22]. Therefore it is not guaranteed to produce a solution that generalizes well to other data drawn from the same underlying distribution.

In solving the optimizing problem that minimizes the *empirical* risk, Eq. (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 [48] that is monotonically increasing with some measure of the capacity, and depends on the probability with which the bound holds. 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 [23, 24], which helps to lower the generalization error. Support vector machines are a powerful class of algorithms that implement this idea, which is also known in statistics as regularization.

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

How can portfolio optimization be regularized? Portfolio optimization is essentially a regression problem, and therefore we can apply statistical learning theory, in particular the work on support vector regression.

Note first that the capacity of a linear model class for which the length of the weight vector is restricted to  $\|w\|^2 \leq A$  has an upper bound which is smaller than the capacity of unconstrained linear models [23, 24]. The capacity is minimized when the length of the weight vector is minimized [23, 24]. Vapnik's concept of *structural risk minimization* [48] results in the support vector algorithm [25, 26] which finds the model with the smallest capacity that is consistent with the data, that is the model with smallest  $\|w\|^2$ . This leads to a convex constrained optimization problem [25, 26] which can be solved using linear programming.

## 5. Regularized portfolio optimization with the expected shortfall risk measure.

While the original Markowitz' 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, 52, 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, 55, 56].

Expected shortfall has been steadily gaining popularity in recent years. The regularization 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  $\alpha_\beta$  is a threshold, such that a given fraction  $\beta$  of the (empirical) loss-distribution over  $z$  lies above  $\alpha_\beta$ . One now wishes to minimize the average over the remaining tail distribution, containing the fraction  $\nu := 1 - \beta$ , and defines the expected shortfall as

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

The term in the sum implements the  $\theta$ -function, while  $\nu$  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:

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

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

$$\sum_i w_i = 1 \quad (6)$$

We propose to implement 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:

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

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

$$\xi_k \geq 0; \quad \epsilon \geq 0; \quad (9)$$

$$\sum_i w_i = 1. \quad (10)$$

The problem is mathematically almost identical to a support vector regression (SVR) algorithm called  $\nu$ -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  $\epsilon$ -intensive loss, used in support vector regression, defined as the maximum of  $\{0; |f(\mathbf{x}) - y| - \epsilon\}$ , where  $f(\mathbf{x})$  is the interpolant, and  $y$  the measured value (response). In that sense  $\epsilon$  measures an allowable error below which deviations are discarded.§

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 support vector regression 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 on 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].

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

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

and additional constraints have to be added to Eqs. (8) to (10):

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

This problem corresponds to  $\nu$ -SVR, a well understood regression method [60], with the only difference that the budget constraint, Eq. (10) is added here. In the finance context the associated loss might be called *symmetric tail average*. Solving the regularized expected shortfall minimization problem, Eqs. (7)–(10) is a special case of solving the regularized STA minimization problem, Eq. (11) with the constraints Eqs. (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, Eqs. (7)–(10).

§ The mathematical similarity between minimum expected shortfall without regularization and the  $E\nu$ -SVM algorithm [58] was pointed out, but incorrectly, in [59]. There is an important difference between the two optimization problems. In  $E\nu$ -SVM, the length of the weight vector,  $\|\mathbf{w}\|$ , is constrained, which implements capacity control. In the pure expected shortfall minimization, Eq. (4), this is not done. Instead, the total budget  $\sum_i w_i$  is fixed. This difference is not correctly identified in the proof of the central theorem (Theorem 1) in [59].

### 5.1. Regularized Symmetric Tail Average Minimization

The solution to the regularized symmetric tail average problem, Eq. (11) with the constraints Eqs. (8)–(10) and (12), is found in analogy to support vector regression, following [60], by writing down the Lagrangean, 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 Lagrangean is different from the one that arises in  $\nu$ -SVR in that it is modified by the budget constraint:

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

$$= F[\mathbf{w}] + \epsilon \left( C\nu - \lambda - \sum_{k=1}^T (\alpha_k + \alpha_k^*) \right) - \gamma \quad (14)$$

$$+ \sum_{k=1}^T \left[ \xi_k \left( \frac{C}{T} - \alpha_k - \eta_k \right) + \xi_k^* \left( \frac{C}{T} - \alpha_k^* - \eta_k^* \right) \right]$$

with

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

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

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

This solution for the optimal portfolio is sparse in the sense that, due to the Karush-Kuhn-Tucker conditions, only those points contribute to the optimal portfolio weights, for which the inequality constraints in (8), and the corresponding constraints in Eq. (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. Regularized portfolio optimization (RPO) operates, in contrast to general regression, with a fixed budget. As a consequence, the Lagrange multiplier  $\gamma$  now appears in the optimal solution, Eq. (16). Compared to the optimal solution in support vector (SV) regression,  $\mathbf{w}_{\text{SV}}$ , the solution vector under the budget constraint,  $\mathbf{w}_{\text{RPO}}$ , is shifted by  $\gamma$ :

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

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  $D := \min_{\mathbf{w}, \xi, \xi^*, \epsilon, \alpha, \alpha^*, \gamma, \lambda, \eta, \eta^*} L[\mathbf{w}, \xi, \xi^*, \epsilon, \alpha, \alpha^*, \gamma, \lambda, \eta, \eta^*]$ , 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

$$\begin{aligned} F[\mathbf{w}_{\text{opt}}] &= -\frac{1}{2} \mathbf{w}_{\text{opt}}^2 \\ &= -\frac{1}{2} \sum_{k=1}^T \sum_{l=1}^T (\alpha_k - \alpha_k^*)(\alpha_l - \alpha_l^*) \mathbf{x}^{(k)} \mathbf{x}^{(l)} - \gamma \sum_{k=1}^T (\alpha_k - \alpha_k^*) \mathbf{x}^{(k)} \mathbf{1} - \frac{N}{2} \gamma^2. \end{aligned} \quad (18)$$

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

- (i) If  $\left(C\nu - \lambda - \sum_{k=1}^T (\alpha_k + \alpha_k^*)\right) < 0$ , then  $L$  can be minimized by letting  $\epsilon \rightarrow \infty$ , which means that  $D = -\infty$ .
- (ii) If  $\left(C\nu - \lambda - \sum_{k=1}^T (\alpha_k + \alpha_k^*)\right) \geq 0$ : The term  $\epsilon \left(C\nu - \lambda - \sum_{k=1}^T (\alpha_k + \alpha_k^*)\right)$  vanishes. Reason: 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 asterix).

- (i) If  $\left(\frac{C}{T} - \alpha_k^{(*)} - \eta_k^{(*)}\right) < 0$ , then  $L$  can be minimized by letting  $\xi_k^{(*)} \rightarrow \infty$ , which means that  $D = -\infty$ .
- (ii) If  $\left(\frac{C}{T} - \alpha_k^{(*)} - \eta_k^{(*)}\right) \geq 0$ , then  $\xi_k \left(\frac{C}{T} - \alpha_k^{(*)} - \eta_k^{(*)}\right) = 0$ . Reason: If the inequality holds strictly then  $L$  can be minimized by  $\xi_k^{(*)} = 0$ . If equality holds then it is trivially true.

By a similar argument, the term  $\gamma$  in Eq. (14) disappears in the Dual. Altogether we have that either  $D = -\infty$ , or

$$D(\alpha, \alpha^*, \gamma) = \min_{\xi, \xi^*, \epsilon} F[\mathbf{w}_{\text{opt}}(\alpha, \alpha^*, \gamma)] = F[\mathbf{w}_{\text{opt}}(\alpha, \alpha^*, \gamma)] \quad (19)$$

$$\text{AND } \sum_{k=1}^T (\alpha_k^* + \alpha_k) \leq C\nu - \lambda \quad (20)$$

$$\text{AND } \alpha_k^{(*)} + \eta_k^{(*)} \leq \frac{C}{T}. \quad (21)$$

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

$$\max_{\alpha, \alpha^*, \gamma} \left[ -\frac{1}{2} \left( \sum_{k=1}^T (\alpha_k - \alpha_k^*) \mathbf{x}^{(k)} - \gamma \mathbf{1} \right)^2 \right]. \quad (22)$$

$$\text{s.t. } \{\alpha_k, \alpha_k^*\} \in \left[0, \frac{C}{T}\right] \quad (23)$$

$$\sum_{k=1}^T (\alpha_k^* + \alpha_k) \leq C\nu. \quad (24)$$

We can analytically maximize over  $\gamma$  and obtain for the optimal value

$$\gamma = \frac{1}{N} \left( \sum_{k=1}^T (\alpha_k - \alpha_k^*) \sum_{i=1}^N x_i^{(k)} - 1 \right) \quad (25)$$

The optimal projection (= optimal portfolio) is given by

$$\mathbf{w} \cdot \mathbf{x} = \sum_{k=1}^T (\alpha_k - \alpha_k^*) \mathbf{x}^{(k)} \cdot \mathbf{x} - \frac{1}{N} \left( \sum_{k=1}^T (\alpha_k - \alpha_k^*) \sum_{i=1}^N x_i^{(k)} - 1 \right) \mathbf{1} \cdot \mathbf{x}. \quad (26)$$

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 Eq. (26)). 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 [61]), which is in our case given by

$$\begin{aligned} \max_{\alpha, \alpha^*, \gamma} & -\frac{1}{2} \left[ \sum_{k=1}^T \sum_{l=1}^T (\alpha_k - \alpha_k^*) (\alpha_l - \alpha_l^*) \left( \mathbf{x}^{(k)} \mathbf{x}^{(l)} - \frac{1}{N} \sum_{i=1}^N x_i^{(k)} \sum_{i=1}^N x_i^{(l)} \right) \right. \\ & \left. + \frac{2}{N} \sum_{k=1}^T (\alpha_k - \alpha_k^*) \sum_{i=1}^N x_i^{(k)} \right] \\ & \{\alpha_k, \alpha_k^*\} \in \left[ 0, \frac{C}{T} \right]; \\ & \sum_{k=1}^T (\alpha_k^* + \alpha_k) \leq C\nu. \end{aligned} \quad (27)$$

For  $N \rightarrow \infty$  the problem becomes *identical* to  $\nu$ -SVR, which can be solved by linear programming, for which software packages are available [62]. For finite  $N$ , it can still be solved with existing methods, because it is quadratic in the  $\alpha_k$ 's. Solvers such as the ones discussed in [63, 61] can be used, but have to be adapted to this specific problem.

The regularized symmetric tail average minimization problem (Eq. (11) with the constraints Eqs. (8)–(10) and (12)) is, as we have shown here, directly related to support vector regression which uses the  $\epsilon$ -insensitive loss function. The  $\epsilon$ -insensitive loss is stable to local changes for data points that fall outside the range specified by  $\epsilon$ . 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  $\xi_k$  by  $\mu$  and  $\xi_k^*$  by  $1 - \mu$ , respectively.

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 regularized portfolio optimization problem, Eqs. (7)–(10), which we originally set out to solve. This is now easily solved in analogy to the previous paragraphs: the slack variables  $\xi_k^*$  disappear, together with the respective Lagrange multipliers which enforce constraints, including  $\alpha_k^*$ . The optimal solution is now

$$\mathbf{w}_{\text{opt}} = \sum_{k=1}^T \alpha_k \mathbf{x}^{(k)} - \gamma \mathbf{1}, \quad (28)$$

with

$$\gamma = \frac{1}{N} \left( \sum_{k=1}^T \alpha_k \sum_{i=1}^N x_i^{(k)} - 1 \right). \quad (29)$$

The dual problem is given by

$$\begin{aligned} \max_{\alpha_k} & -\frac{1}{2} \left[ \sum_{k=1}^T \sum_{l=1}^T \alpha_k \alpha_l \left( \mathbf{x}^{(k)} \mathbf{x}^{(l)} - \frac{1}{N} \sum_{i=1}^N x_i^{(k)} \sum_{i=1}^N x_i^{(l)} \right) + \frac{2}{N} \sum_{k=1}^T \alpha_k \sum_{i=1}^N x_i^{(k)} \right] \\ \text{s.t. } & \alpha_k \in \left[ 0, \frac{C}{T} \right]; \quad \sum_{k=1}^T \alpha_k \leq C\nu. \end{aligned} \quad (30)$$

which, like its symmetric counterpart, Eq. (27), 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 investors time horizon, the nature of the underlying data, and, crucially, on 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, for example, the treatment in [64]), 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

$$\|\mathbf{w}\|^2 = \sum_i w_i^2 \simeq \frac{1}{N_{\text{eff}}} \quad (31)$$

corresponds roughly to maximizing the effective number of assets,  $N_{\text{eff}}$ , i.e. to exerting a *pressure* towards portfolio diversification [65]. 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 penalizes 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,  $w_i \geq 0$ , 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  $\sum_i |w_i| \leq 1$ .<sup>||</sup> 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, in principle, 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 [28]. The spontaneous reduction of portfolio size has also been demonstrated in numerical simulations [66]: 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 [67] 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 [66], 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 which 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 data set size is similar to the size of the portfolio. In that regime, a small empirical risk does not necessarily guarantee a small actual risk [22]. In this sense portfolio optimization overfits the data. Regularization can overcome this problem by reducing the capacity of the considered models.

Regularized portfolio optimization has choices to make, not only about the risk function, but also about the regularizer. One possible regularizer that leads to a convex optimization problem which can be solved with linear programming is the L2 norm of the weight vector. We have shown that with the L2 norm, regularized portfolio optimization

<sup>||</sup> This point has been made independently by [15].

with the expected shortfall as a risk measure is a variant of support vector regression. The differences are an 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: not only does it counteract downward fluctuations in one asset by upward fluctuations in another asset [2], but it also contributes to the stability of the solution. We have shown that the L2 regularizer implements a pressure towards portfolio diversification. Therefore, from a statistical 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 which is more stable, and hence meaningful.

In summary, the method 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.

## 8. Acknowledgements

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.

- [1] H. Markowitz. Portfolio selection. *Journal of Finance*, 7:77–91, 1952.
- [2] H. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. J. Wiley and Sons, New York, 1959.
- [3] E. J. Elton and M. J. Gruber. *Modern Portfolio Theory and Investment Analysis*. Wiley, New York, 1995.
- [4] J.D. Jobson and B. Korkie. Improved Estimation for Markowitz Portfolios Using James-Stein Type Estimators. *Proceedings of the American Statistical Association (Business and Economic Statistics)*, 1:279–284, 1979.
- [5] P. Jorion. Bayes-Stein Estimation for Portfolio Analysis. *Journal of Financial and Quantitative Analysis*, 21:279–292, 1986.
- [6] P.A. Frost and J.E. Savarino. An Empirical Bayes Approach to Efficient Portfolio Selection. *Journal of Financial and Quantitative Analysis*, 21:293–305, 1986.
- [7] O. Ledoit and M. Wolf. Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.
- [8] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, 88:365–411, 2004.
- [9] O. Ledoit and M. Wolf. Honey, I Shrunk the Sample Covariance Matrix. *J. Portfolio Management*, 31:110, 2004.
- [10] V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus Naive Diversification: How Efficient is the 1/N Portfolio Strategy? *Review of Financial Studies*, 2007.
- [11] L. Garlappi, R. Uppal, and T. Wang. Portfolio selection with parameter and model uncertainty: a multi-prior approach. *Review of Financial Studies*, 20:41–81, 2007.

- [12] V. Golosnoy and Y. Okhrin. Multivariate shrinkage for optimal portfolio weights. *The European Journal of Finance*, 13:441–458, 2007.
- [13] R. Kan and G. Zhou. Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 42:621–656, 2007.
- [14] G. Frahm and Ch. Memmel. Dominating estimators for the global minimum variance portfolio, 2009. Deutsche Bundesbank, Discussion Paper, Series 2: Banking and Financial Studies.
- [15] V. DeMiguel, L. Garlappi, F. J. Nogales, and R. Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55:798–812, 2009.
- [16] L. Laloux, P. Cizeau, J.-Ph. Bouchaud, and M. Potters. Noise Dressing of Financial Correlation Matrices. *Phys. Rev. Lett.*, 83:1467–1470, 1999.
- [17] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, and H.E. Stanley. Universal and Non-Universal Properties of Cross-Correlations in Financial Time Series. *Phys. Rev. Lett.*, 83:1471, 1999.
- [18] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters. Random Matrix Theory and Financial Correlations. *International Journal of Theoretical and Applied Finance*, 3:391, 2000.
- [19] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley. A Random Matrix Approach to Cross-Correlations in Financial Time-Series. *Phys. Rev. E*, 65:066136, 2000.
- [20] Z. Burda, A. Goerlich, and A. Jarosz. Signal and noise in correlation matrix. *Physica*, A343:295, 2004.
- [21] M. Potters and J.-Ph. Bouchaud. Financial applications of random matrix theory: Old laces and new pieces. *Acta Phys. Pol.*, B36:2767, 2005.
- [22] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.
- [23] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.
- [24] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.
- [25] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proc. 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.
- [26] C. Cortes and V. Vapnik. Support Vector Networks. *Machine Learning*, 20:273–297, 1995.
- [27] R. Macrae and C. Watkins. Safe portfolio optimization. In H. Bacelar-Nicolau, F. C. Nicolau, and J. Janssen, editors, *Proceedings of the IX International Symposium of Applied Stochastic Models and Data Analysis: Quantitative Methods in Business and Industry Society, ASMDA-99, 14-17 June 1999, Lisbon, Portugal*, page 435. INE, Statistics National Institute, Portugal, 1999.
- [28] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.*, 58(1):267–288, 1996.
- [29] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–148, 1993.
- [30] V.K. Chopra and W.T. Ziemba. The Effect of Errors in Means, Variances, and Covariances on Optimal Portfolio Choice. *Journal of Portfolio Management*, 19:611, 1993.
- [31] R.C. Merton. On Estimating the Expected Return on the Market: An Exploratory Investigation. *Journal of Financial Economics*, 8:323361, 1980.
- [32] R. Jagannathan and T. Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58:1651–1684, 2003.
- [33] Y. Okhrin and W. Schmieid. Distribution properties of portfolio weights. *Journal of Econometrics*, 134:235–256, 2006.
- [34] A. Kempf and C. Memmel. Estimating the Global Minimum Variance Portfolio. *Schmalenbach Business Review*, 58:332348, 2006.
- [35] G. Frahm. Linear Statistical Inference for Global and Local Minimum Variance Portfolios. *Statistical Papers*, 2008. DOI: 10.1007/s00362-008-0170-z.
- [36] I. Kondor, S. Pafka, and G. Nagy. Noise sensitivity of portfolio selection under various risk

- measures. *Journal of Banking and Finance*, 31:1545–1573, 2007.
- [37] S. Pafka and I. Kondor. Noisy covariance matrices and portfolio optimization. *Eur. Phys. J., B* 27:277–280, 2002.
- [38] S. Pafka and I. Kondor. Noisy covariance matrices and portfolio optimization ii. *Physica, A* 319:487–494, 2003.
- [39] S. Pafka and I. Kondor. Estimated correlation matrices and portfolio optimization. *Physica, A* 343:623–634, 2004.
- [40] Z. Burda, J. Jurkiewicz, and M. A. Nowak. Is econophysics a solid science? *Acta Physica Polonica, B* 34:87–132, 2003.
- [41] M. Mezard S. Ciliberti, I. Kondor. On the feasibility of portfolio optimization under expected shortfall. *Quantitative Finance*, 7:389–396, 2007.
- [42] M. Mezard S. Ciliberti. Risk minimization through portfolio replication. *Eur. Phys. J., B* 57:175–180, 2007.
- [43] I. Varga-Haszonits and I. Kondor. The instability of downside risk measures. *J. Stat. Mech.*, P12007, 2008.
- [44] I. Varga-Haszonits and I. Kondor. Noise Sensitivity of Portfolio Selection in Constant Conditional Correlation GARCH models. *Physica, A*385:307–318, 2007.
- [45] I. Kondor and I. Varga-Haszonits. Feasibility of portfolio optimization under coherent risk measures, 2008. submitted to *Quantitative Finance*.
- [46] B. Schölkopf. *Support vector learning*. GMD-Bericht ; 287. Oldenbourg, München, Germany, 1997. Dissertation: Berlin, Techn. Univ., Diss., 1997.
- [47] B. Schölkopf, C. J.C. Burges, and A. J. Smola. *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, 1999.
- [48] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, 1974. [In Russian. German translation available from Akademie-Verlag, Berlin, 1979.].
- [49] P. Jorion. *VaR: The New Benchmark for Managing Financial Risk*. McGraw-Hill, New York, 2000.
- [50] J.P. Morgan and Reuters. Riskmetrics. Technical Document available at <http://www.riskmetrics.com>.
- [51] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent Measures of Risk. *Mathematical Finance*, 9:203–228, 1999.
- [52] P. Embrechts. Extreme Value Theory: Potential and Limitations as an Integrated Risk Measurement Tool. *Derivatives Use, Trading and Regulation*, 6:449–456, 2000.
- [53] C. Acerbi, C. Nardio, and C. Sirtori. Expected Shortfall as a Tool for Financial Risk Management., 2001. unpublished.
- [54] C. Acerbi. Spectral Measures of Risk: a Coherent Representation of Subjective Risk Aversion. *Journal of Banking and Finance*, 26(7):1505–1518, 2002.
- [55] C. Acerbi and D. Tasche. On the Coherence of Expected Shortfall. *Journal of Banking and Finance*, 26(7):1487–1503, 2002.
- [56] C. Acerbi. Coherent representations of subjective risk-aversion. In G. Szegö, editor, *Risk Measures for the 21st Century*. John Wiley and Sons., 2004.
- [57] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.
- [58] F. Perez-Cruz, J. Weston, D.J.L. Herrmann, and B. Schölkopf. Extension of the nu-svm range for classification. In *Advances in Learning Theory: Methods, Models and Applications*, volume 190 of *NATO Science Series III: Computer and Systems Sciences*, pages 179–196. IOS Press, Amsterdam, 2003.
- [59] A. Takeda and M. Sugiyama.  $\nu$ -support vector machine as conditional value-at-risk minimization, 2008.
- [60] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 05 2000.

- [61] Léon Bottou and Chih-Jen Lin. Support vector machine solvers. In Léon Bottou, Olivier Chapelle, Dennis DeCoste, and Jason Weston, editors, *Large Scale Kernel Machines*, pages 301–320. MIT Press, Cambridge, MA., 2007.
- [62] Robert J. Vanderbei. LOQO users manual.x. Software available at <http://www.princeton.edu/~rvdb/loqo/LOQO.html>.
- [63] Antoine Bordes, Seyda Ertekin, Jason Weston, and Léon Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6:1579–1619, September 2005.
- [64] Y. Bengio and Y. Grandvalet. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. In S. Becker, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16 (NIPS'03)*. MIT Press, Cambridge, MA., 2004.
- [65] J.-Ph. Bouchaud and M. Potters. *Theory of Financial Risk - From Statistical Physics to Risk Management*. Cambridge University Press, Cambridge, UK, 2000.
- [66] N. Gulyas and I. Kondor. Portfolio instability and linear constraints, 2007. submitted to Physica A.
- [67] B. Scherer and R. D. Martin. *Introduction to Modern Portfolio Optimization With NUOPT and S-PLUS*. Springer, 2005.