Replica approach to mean-variance portfolio optimization

Replica approach to mean-variance portfolio optimization

Istvan Varga-Haszonits[1]*, Fabio Caccioli[2,3] and Imre Kondor[1,4]
1- Department of Physics of Complex Systems, Eötvös University, Budapest, Hungary
2- University College London, Department of Computer Science, London, WC1E 6BT, UK
3- Systemic Risk Centre, London School of Economics and Political Sciences, London, UK
4-Parmenides Foundation, Pullach, Germany

October 25, 2016

Abstract

We consider the problem of mean-variance portfolio optimization for a generic covariance matrix subject to the budget constraint and the constraint for the expected return, with the application of the replica method borrowed from the statistical physics of disordered systems. We find that the replica symmetry of the solution does not need to be assumed, but emerges as the unique solution of the optimization problem. We also check the stability of this solution and find that the eigenvalues of the Hessian are positive for $r = N/T < 1$, where N is the dimension of the portfolio and T the length of the time series used to estimate the covariance matrix. At the critical point $r = 1$ a phase transition is taking place. The out of sample estimation error blows up at this point as $1/(1-r)$, independently of the covariance matrix or the expected return, displaying the universality not only of the critical exponent, but also the critical point. As a conspicuous illustration of the dangers of in-sample estimates, the optimal in-sample variance is found to vanish at the critical point inversely proportional to the divergent estimation error.

1 Introduction

Starting with Markowitz's seminal paper [1], the problem of portfolio optimization has become the subject of a vast literature. The several hundred papers written on this problem apply widely different approaches; the list [2–46] is just a small selection from this literature.

The present paper belongs to a narrow subfield where the methods of statistical physics, in particular the replica method borrowed from the theory of disordered systems, are applied. In this approach one assumes that the returns on the securities in

*Present employer: MSCI, Budapest

the portfolio obey an idealized probability distribution, in the simplest case a Gaussian, and calculates the quantities of interest like the weights of the optimal portfolio, the minimal risk, sensitivity to changes in the returns, estimation error, etc., exactly. Of course, real-life returns are neither normal, nor stationary, therefore one can only hope that the results produced by the theory will give an idea of the behaviour of the various characteristic quantities, which can then be compared to simulation results and empirical data. Well-suited to the case of huge institutional portfolios, the statistical physics-inspired approach targets large portfolios composed of N different kind of securities, and assumes that the size T of the statistical samples is comparable to N, so that the ratio of the two, $r = N/T$, is a fixed value. In this high-dimensional limit sample fluctuations and the concomitant estimation error can be very large, even divergent. The problem of estimation error is therefore in the focus of the statistical physics-inspired approach which, due to the highly idealized nature of the probability distribution, may be expected to provide a lower bound for this error.

The first step along this path was taken by Ciliberti et al [47] who studied the large fluctuations and the resulting phase transition in the case of the special risk measure Expected Shortfall (ES), soon to be followed by a similar work [48] on the mean absolute deviation. In subsequent papers ES was optimized under various regularizers to suppress the instability of estimation, and to take into account the future market impact of an eventual liquidation of the portfolio [49], [50]. The works [47], [48], [49], [50] were mostly concerned with the instability and the accompanying phase transition, accordingly they considered i.i.d. returns (i.e. a diagonal covariance matrix with identical elements), so that to show up the phenomenon in the simplest possible setting. The papers [51], [52] changed focus, and still assuming i.i.d. returns, they mapped out the whole parameter space of ES, demonstrating the very serious estimation error problem even far away from the instability region. Finally, [53] considered also an ℓ_2 regularized version of ES, which allowed an insight into how variance-bias tradeoff is playing out in the case of this particular risk measure.

The prominence of the risk measure ES in all these papers was mainly motivated by its anticipated regulatory significance; and indeed, after a long consultation period the Basel Committee finally instituted ES as the global regulatory market risk measure in January 2016 [54]. Because of the focus on the regulatory context, the application of the replica method to the most obvious and simplest risk measure, the variance, has been left in the background. Nevertheless, it was not completely neglected: we performed the optimization of variance via replicas quite some years ago, but the results were left unpublished.[1] As we recently decided to carry out a systematic study of the sensitivity of different risk measures to estimation error (as it were, an analytical counterpart of the numerical study [36]) and also their responses to various regularization schemes, it is high time to publish the results for the variance now.[2]

[1] The results for variance formed part of the PhD thesis "The instability of risk measures" (in Hungarian) by one of us (I.V-H.), written under the supervision of I.K. and successfully defended at Eötvös University, Budapest, in 2009. Independently, F.C. also performed the replica treatment of the variance optimization at the Santa Fe Institute in 2011.

[2] During the preparation of this manuscript we came to learn about T. Shinzato's preprint posted on arXive in May 2016 [55], which is concerned with the replica optimization of the variance at the global minimum in the special case of i.i.d. normal returns and with a constraint promoting the concentration of

While the papers quoted in the previous paragraph all confined their treatment to the simplified case of independent random variables (diagonal covariance matrix) and studied the global minimal portfolio, here we present the replica treatment of the full Markowitz problem, that is we consider a generic full rank covariance matrix, and take into account not only the budget constraint but also the constraint on the expected return. Thereby we calculate the estimation error all along the efficient frontier.

The optimization of variance leads us to two very interesting results. In contrast to the usual situation, where at a certain point in the derivation one has to choose the replica symmetric minimum from among several seemingly possible ones and hope this choice leads to the correct solution, in the special case of the variance the symmetric solution emerges as the unique minimum. Since the variance is convex, it necessarily has a unique minimum, therefore the unique replica symmetric minimum found by the method must coincide with the correct result. The other remarkable feature of the solution is that the estimation error turns out to be the same along the whole efficient boundary, independently of the covariance matrix and the expected return.[3] In principle, one might have expected that a covariance matrix with several more or less strongly correlated returns (but still full rank) would display a reduced effective dimension, hence a shift in the critical point. The independence of the estimation error from the covariances and expected returns is the manifestation of a surprising degree of universality, analogous to the independence of the critical point of the minimax risk measure from the underlying probability distribution (see [36]), but also to the universality discovered in a wide class of high dimensional random geometric phase transitions by Donoho and Tanner [57].

When looking for a minimum, one should, in general, check the behaviour of the second derivatives. Although there could be no doubt about the extremum being a stable minimum in the case of the variance, we nevertheless calculate the Hessian (the matrix of second derivatives) at the extremum, and show that all its eigenvalues are positive. This is meant to be an example for a check one should always perform, but often neglects.

Our calculations are limited to the case where both the number N of assets in the portfolio and the sample size T (the length of the time series for the returns) are large, with their ratio $r = N/T$ a fixed number, less than 1. ($r = 1$ is the critical point of the problem, beyond which the covariance matrix ceases to be of full rank, and the optimization problem becomes meaningless.) We note that in the case of the variance, finite N and finite T results also exist (e.g. [30], [39], [43]) and in the appropriate limit they coincide with our results.

We also note that the Markowitz problem with the expected return constraint relaxed is equivalent to the problem of linear regression, see e.g. [29], [58], thus the replica approach can naturally cover regression type problems as well.

The plan of the paper is as follows. In Sec. 2 we formulate the statistical mechanics of portfolio optimization assuming that the covariance matrix and returns are known.

investment, equivalent to an ℓ_2 regularizer. Despite the similar subject and method, there is little overlap between his work and ours.

[3] Various special cases of this invariance have long been known to us from earlier numerical [27] and analytical work [56], but the result to be presented below is the most general proof of this surprising invariance.

Sec. 3 is the replica treatment of the problem when the returns are drawn from a multivariate normal distribution, with a generic covariance matrix, a budget constraint and a constraint on the expected return. Finally, Sec. 4 is a summary with a discussion of the nature of the phase transition and the explanation of its universality. As the main purpose of the paper is to exhibit the technique of replicas in the context of variance optimization, we include some of the details of the calculation in the main text rather than exiling them to Appendices.

2 The Markowitz problem as a statistical physics model

Portfolio optimization is a tradeoff between risk and return: it seeks to maximize the return of a portfolio at a given level of risk, or minimize the risk at a given level of return. If, following Markowitz [1], we assume that the returns are drawn from a multivariate Gaussian distribution and risk is measured in terms of the variance of the portfolio, then we are led to the following quadratic optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^N} \sum_{i=1}^{N}\sum_{j=1}^{N} \sigma_{ij} w_i w_j, \tag{1}$$

$$\sum_{i=1}^{N} w_i \mu_i = \mu, \tag{2}$$

$$\sum_{i=1}^{N} w_i = 1, \tag{3}$$

where N is the number of securities in the portfolio, μ_i is the expected value of the return on security i, μ is the expected return on the portfolio, σ_{ij} is the covariance matrix of returns, and w_i is the weight of security i in the portfolio. The solution of this problem is the set of weights that minimize the variance given a fixed return μ and a fixed budget expressed by the sum of weights being fixed at 1. Note that in this simple setup the weights are not constrained to be positive, so unlimited short selling is allowed.

We will refer to the above optimization problem as the Markowitz problem. If the covariance matrix and the expected returns are given, the Markowitz problem can be easily solved by the method of Lagrange multiplyers. As the covariance matrix is positive definite, the objective function is strictly convex, so the problem has a unique solution. This solution was first derived by Merton [3].

Introducing the notations $A = \sum_{i,j} \sigma_{ij}^{-1}$, $B = \sum_{i,j} \sigma_{ij}^{-1} \mu_j$ and $C = \sum_{i,j} \sigma_{ij}^{-1} \mu_i \mu_j$

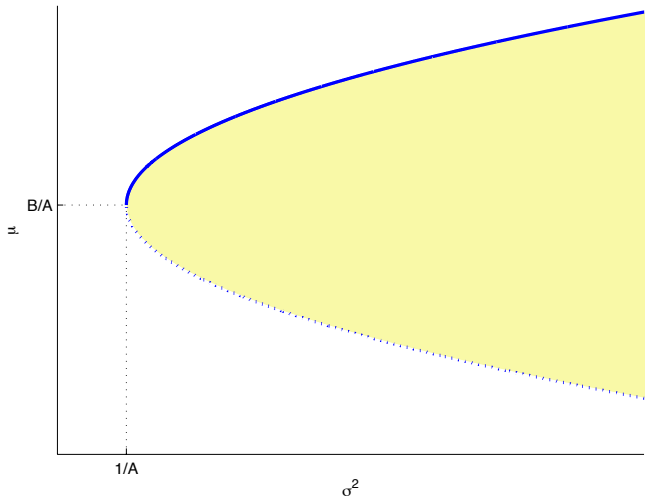

Figure 1: The set of attainable portfolios (yellow area), marginal portfolios (continuous and dotted blue line) and efficient portfolios (continuous blue line) on the risk-return plane.

the solution is:

$$w_i^*(\mu) = \sum_{j=1}^{N} \sigma_{ij}^{-1} \left[\lambda^*(\mu) + \eta^*(\mu)\mu_j\right], \tag{4}$$

$$\lambda^*(\mu) = \frac{C - B\mu}{AC - B^2}, \tag{5}$$

$$\eta^*(\mu) = \frac{A\mu - B}{AC - B^2}. \tag{6}$$

The variance of the optimal portfolio for a given expected return μ is:

$$\sigma^{*2}(\mu) = \sum_{i=1}^{N}\sum_{j=1}^{N} \sigma_{ij} w_i^*(\mu) w_j^*(\mu) = \frac{A}{AC - B^2}\left(\mu - \frac{B}{A}\right)^2 + \frac{1}{A}, \tag{7}$$

where $AC - B^2 > 0$.

The financial meaning of the result is illustrated in Figure 1. The set of those portfolios whose expected return μ and risk σ satisfy the inequality $\sigma^2 \geq \sigma^{*2}(\mu)$ are called attainable portfolios, because these are the ones that can be composed out of the N securities under the budget constraint (3). The solutions of the constrained optimization problem (1) are called marginal portfolios, because for these $\sigma^2 = \sigma^{*2}(\mu)$ holds, so they constitute the boundary of the set of attainable portfolios.

As shown by the figure, if $\sigma^2 > 1/A$ there are two marginal portfolios for every value of the standard deviation σ. Of these, the one with the larger expected return

will be efficient, since for a given risk this will have the largest return among the attainable portfolios. The efficient portfolios will therefore be those for which the conditions $\mu > B/A$ and $\sigma^2 = \sigma^{*2}(\mu)$ are fulfilled simultaneously. These are shown by the continuous blue line in the figure.

In order to provide a bridge to the formalism presented in the next section, we now rephrase the standard Markowitz problem of optimal portfolio selection. Following Kirkpatrick et al. [59], we convert the problem of optimizing the variance into a problem in statistical physics. For this purpose, we regard the variance (with a factor $\frac{1}{2}$ included for convenience) as the Hamiltonian of a fictitious physical system:

$$\mathcal{H} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_{ij}w_i w_j + \sum_{i=1}^{N} h_i w_i, \tag{8}$$

where w_i are the portfolio weights and σ_{ij} are the elements of the covariance matrix, which, in this Section, we continue to regard as given. The h_i's are external fields conjugate to the variables w_i. In the case of $h_1 = h_2 = ... = h_N = 0$ the Hamiltonian is half of the risk (measured in terms of the variance) of the portfolio $\mathbf{w}$. The optimization of the variance corresponds to finding the ground state of the above Hamiltonian.

Assume the system is subject to a Boltzmann distribution at inverse temperature β. Imposing the budget constraint $\sum_i w_i = 1$ and the constraint on the expected return $\sum_i \mu_i w_i = \mu$ the partition function can be written as:

$$Z = \int_{-\infty}^{\infty}\prod_i dw_i \exp\left(-\frac{1}{2}\beta\sum_{i,j}\sigma_{ij}w_i w_j - \beta\sum_i h_i w_i\right)\delta\left(\sum_i w_i - 1\right)\delta\left(\sum_i \mu_i w_i - \mu\right), \tag{9}$$

where the index i runs from 1 to N. By the Fourier representation of the Dirac-δ we obtain the following Gaussian integral:

$$Z = \int_{-\infty}^{\infty}d\eta\int_{-\infty}^{\infty}d\lambda\int_{-\infty}^{\infty}\prod_i dw_i \exp\left(-\frac{1}{2}\beta\sum_{i,j}\sigma_{ij}w_i w_j - \beta\sum_i h_i w_i\right)\times$$
$$\times\exp\left[-i\lambda\left(\sum_i w_i - 1\right) - i\eta\left(\sum_i \mu_i w_i - \mu\right)\right]. \tag{10}$$

Having calculated the integral, the free energy, defined by the relation $F = -\beta^{-1}\ln Z$,

can be written as:

$$F = -\frac{N}{2\beta}\ln\left(\frac{2\pi}{\beta}\right) + \frac{1}{2\beta}\mathrm{Tr}\ln(\sigma) - \frac{1}{2\beta}\ln\left(AC-B^2\right) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_{ij}^{-1}h_i h_j + \\ + \frac{1}{AC-B^2}\left[A\left(\sum_{i,j}\sigma_{ij}^{-1}\mu_i h_j + \mu\right)^2 + C\left(\sum_{i,j}\sigma_{ij}^{-1}h_j + 1\right)^2 - \right. \\ \left. -2B\left(\sum_{i,j}\sigma_{ij}^{-1}\mu_i h_j + \mu\right)\left(\sum_{i,j}\sigma_{ij}^{-1}h_j + 1\right)\right], \quad (11)$$

where we used the notations $A = \sum_{i,j}\sigma_{ij}^{-1}$, $B = \sum_{i,j}\sigma_{ij}^{-1}\mu_j$ and $C = \sum_{i,j}\sigma_{ij}^{-1}\mu_i\mu_j$. The weights of the optimal portfolio are the ensemble averages of the variables w_i at zero temperature and zero external field. They can be obtained by taking the derivative of the free energy with respect to the conjugate fields h_i:

$$\left.\frac{\partial F}{\partial h_i}\right|_{h_1=h_2=\ldots=h_N=0} = \langle w_i\rangle, \quad (12)$$

which gives

$$\langle w_i\rangle = \sum_j \sigma_{ij}^{-1}\left(\frac{A\mu - B}{AC-B^2}\mu_j + \frac{C-B\mu}{AC-B^2}\right). \quad (13)$$

It is remarkable that $\langle w_i\rangle$ does not depend on β, thus the thermal average produces the optimal weights not only at zero, but also at finite temperature. (A side remark: this is like the thermal average of the atoms' positions in a harmonic solid. Thermal expansion needs anharmonic terms.) As we shall see, however, the curve of the efficient portfolios can be recovered only in the low temperature limit. For this, let us write the free energy in zero field:

$$F = -\frac{N}{2\beta}\ln\left(\frac{2\pi}{\beta}\right) + \frac{1}{2\beta}\mathrm{Tr}\ln(\sigma) - \frac{1}{2\beta}\ln\left(AC-B^2\right) + \frac{1}{2}\frac{1}{AC-B^2}\left(A\mu^2 - 2B\mu + C\right). \quad (14)$$

At zero temperature the free energy is the minimum of the Hamiltonian, that is the optimal value of the risk given the expected value of the return μ :

$$\sigma^*(\mu) = 2\lim_{\beta\to\infty} F = \frac{A\mu^2 - 2B\mu + C}{AC-B^2}, \quad (15)$$

which, after some rearrangement, is seen to agree with the formula for the efficient frontier derived above.

3 Sensitivity to noise and the replica method

In contrast to the assumption in the previous section, in real life the returns and covariances are not known, but have to be estimated from empirical samples. Assume,

therefore, that we observe the price changes of our securities in T ($T < \infty$) subsequent (nonoverlapping and equal) periods, and denote the relative price change of security i in period t by x_{it} ($i = 1, 2, ..., N$ s $t = 1, 2, ..., T$). Of course, the returns x_{it} will also be normally distributed variables, furthermore $\mathbb{E}[x_{it}] = \mu_i$ (for all t) and $\mathbb{E}[x_{it}x_{js}] - \mathbb{E}[x_{it}]\mathbb{E}[x_{js}] = \sigma_{ij}\delta_{ts}$, because the returns are assumed serially independent (their autocorrelation is zero). Accordingly, the unbiased estimate of the parameters of the distribution is given by the formulae

$$\hat{\mu}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}, \tag{16}$$

$$\hat{\sigma}_{ij} = \frac{1}{T-1}\sum_{t=1}^{T} \left(x_{it} - \hat{\mu}_i\right)\left(x_{jt} - \hat{\mu}_j\right) \tag{17}$$

Replacing the true distribution by the estimated one based on a finite sample will unavoidably introduce estimation error in the values of the optimal weights and also in the value of the variance itself. We will denote the optimal estimated weights by $\hat{w}_i^*$ and the estimated optimal variance by $(\hat{\sigma}^*)^2$ in the following. Pafka and Kondor [25] (see also the considerations in [52]) introduced the quantity

$$q_0 = \frac{\sum_{i,j}\sigma_{ij}\hat{w}_i^*\hat{w}_j^*}{\sum_{i,j}\sigma_{ij}w_i^*w_j^*}. \tag{18}$$

as a measure of out of sample estimation error. As in the numerator we have the true covariance matrix multiplied by the estimated weights, whereas in the denominator the same true covariance matrix is multiplied by the "true" weights (that minimize the variance), it is clear that $q_0 \geq 1$, where for finite samples the equality holds with zero probability. Thus the number $\sqrt{q_0} - 1$ determines the relative error in the risk in the estimated portfolio. Since the estimated covariance matrix and the corresponding optimal weights fluctuate from sample to sample, the quantity q_0 itself will also be a random variable. The first two moments of q_0 can give an idea about how sensitive the Markowitz model is to estimation error. One of the main results of the present calculation will be to derive the expectation value of q_0 for an arbitrary covariance matrix and verify its universality (its independence of the covariance matrix and the return), a possibility first raised on the basis of numerical experiments by [27]. As for the second moment, its behaviour can be inferred from the results in [30], [39], or [43]: in the thermodynamic limit the variance of q_0 vanishes, in the parlance of the theory of disordered systems, q_0 self-averages.

In addition to the estimation error, we may even worry about whether the optimization can be carried out at all, that is whether the estimated covariance matrix $\hat{\sigma}_{ij}$ preserves the positive definiteness of the true σ_{ij}, so that to remain invertible. According to elementary linear algebra, the condition for the positive definiteness of the estimated covariance matrix is $T \geq N$, because the rank of $\hat{\sigma}_{ij}$ is, with probability one, equal to $\min\{N, T\}$. This means we must have at least as many, or more, observations for each security as the dimension N of the portfolio, a very natural requirement. Accordingly, the ratio $r = N/T$ will play a crucial role in the following. For very small values of r, when we have plenty of data, we are in the realm of ordinary statistics, and

by force of the central limit theorem our estimates will converge to their true values. If r is not small enough, we will be working in the high dimensional regime, where the estimates may strongly deviate from the true values, and as we approach $r = 1$ from below we may expect the estimation error to blow up. Beyond this critical value of r the optimization of the variance cannot be performed. During the long history of portfolio theory, financial mathematics, statistics and computer science introduced a plethora of methods to deal with this difficulty; ultimately all these procedures boil down to a modification of the original problem (via dimensional reduction, regularization, etc.) so as to remove the instability at the price of permitting a, hopefully limited, bias. As the purpose of the present paper is to demonstrate the application of statistical physics methods to the optimization of the variance, we do not concern ourselves with these methods here, and keep to the original framework of mean-variance optimization.

A characteristic feature of the replica method that we are going to apply in the following is that at a certain point one has to calculate a saddle point integral where the dimension N is let go to infinity. Then, to keep the ratio r below its critical value 1, we have to consider very large samples. This means we will be interested in the "thermodynamic" limit where both N and T are large, but their ratio stays finite, smaller than 1. For this reason, we may safely neglect the 1 in the denominator in (17).

Let us now consider the optimization of the variance estimated from empirical samples. Introducing the notation $u_t = \sum_i w_i(x_{it} - \mu_i)$ the problem can be formulated as

$$\min_{\mathbf{u},\mathbf{w}} \frac{1}{T}\sum_{t=1}^{T}\left(u_t - \frac{1}{T}\sum_{s=1}^{T} u_s\right)^2, \tag{19}$$

$$\sum_{i=1}^{N} w_i(x_{it} - \mu_i) = u_t, \tag{20}$$

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} w_i x_{it} = N\mu, \tag{21}$$

$$\sum_{i=1}^{N} w_i = N. \tag{22}$$

Note the change of normalization in the budget constraint (22). The purpose of this modification is to ensure that the order of magnitude of the weights remain $O(1)$ in the thermodynamic limit. Let us write up the partition function corresponding to this problem, following the recipe in the previous section:

$$Z \propto \int \prod_t du_t \int \prod_t d\hat{u}_t e^{-\frac{\beta}{2T}\sum_t \left(u_t - \frac{1}{T}\sum_s u_s\right)^2} \int \prod_i dw_i e^{i\sum_t \hat{u}_t\left[u_t - \sum_i w_i(x_{it}-\mu_i)\right]}$$
$$\int d\eta e^{-i\eta\left(\frac{1}{T}\sum_t u_t + \sum_i w_i\mu_i - N\mu\right)} \int d\lambda e^{-i\lambda\left(\sum_i w_i - N\right)}, \tag{23}$$

where the integrals run from $-\infty$ to ∞. We have omitted here an unimportant constant factor in Z (hence the notation $\propto$ instead of $=$). We will do so also in the following.

We wish to average Z over the random sample $\{x_{it}\}$ whose joint distribution function is:

$$f\left(\{x_{it}\}\right) = (2\pi)^{-NT/2} \left(\det \sigma\right)^{-T/2} e^{-\frac{1}{2}\sum_{i,j,t} \sigma_{ij}^{-1}(x_{it}-\mu_i)(x_{jt}-\mu_j)}. \tag{24}$$

We assume that σ_{ij} is strictly positive definite. In order to facilitate the derivation, we express the variables x_{ij} through the standard normal variables z_{it}:

$$x_{it} = \sum_j D_{ij} z_{jt} + \mu_i, \tag{25}$$

where D_{ij} is the Cholesky decomposition of the true covariance matrix σ_{ij} of the returns, that is, by definition $\sigma_{ij} = \sum_k D_{ik} D_{jk}$. On the basis of the formula (24) we can see, by a simple change of variables, that the joint probability distribution of the random variables z_{it} is indeed

$$f\left(\{z_{it}\}\right) = (2\pi)^{-NT/2} e^{-\frac{1}{2}\sum_{i,t} z_{it}^2}. \tag{26}$$

Expressed through the standard normal variables z_{it} the partition function becomes

$$Z \propto \int \prod_t du_t \int \prod_t d\hat{u}_t e^{-\frac{\beta}{2T}\sum_t \left(u_t - \frac{1}{T}\sum_s u_s\right)^2} \int \prod_i dv_i e^{i\sum_t \hat{u}_t \left(u_t - \sum_i v_i z_{it}\right)} \int d\eta e^{-i\eta\left(\frac{1}{T}\sum_t u_t + \sum_i v_i \theta_i - N\mu\right)} \int d\lambda e^{-i\lambda\left(\sum_i v_i d_i - N\right)}, \tag{27}$$

where we have changed variables $v_i = \sum_j w_j D_{ji}$, and introduced the notations $d_i = \sum_j D_{ij}^{-1}$ and $\theta_i = \sum_j D_{ij}^{-1}\mu_j$.

3.1 The replica method

Our goal is to calculate the average of free energy density over the samples in the thermodynamic limit (N/T constant and $N \to \infty$):

$$\mathbb{E}[f] = -\lim_{N\to\infty} \frac{1}{N\beta} \mathbb{E}\left[\ln Z\right]. \tag{28}$$

The direct calculation of the expected value of the logarithm of a random variable is difficult, but by the relation

$$\ln Z = \lim_{n\to 0} \frac{Z^n - 1}{n} \tag{29}$$

we can reduce the problem to the calculation of $\mathbb{E}[Z^n]$. If n is a natural number then Z^n is the partition function of n identical, independent copies or replicas (hence the name of the method) of the original system, which by (27) can be written as:

$$Z^n \propto \int \prod_{a,t} du_t^a \int \prod_{a,t} d\hat{u}_t^a e^{-\frac{\beta}{2T}\sum_{a,t}\left(u_t^a - \frac{1}{T}\sum_s u_s^a\right)^2} \int \prod_{a,i} dv_i^a e^{i\sum_{a,t} \hat{u}_t^a \left(u_t^a - \sum_i v_i^a z_{it}\right)} \int \prod_a d\eta^a e^{-i\sum_a \eta^a \left(\frac{1}{T}\sum_t u_t^a + \sum_i v_i^a \theta_i - N\mu\right)} \int \prod_a d\lambda^a e^{-i\sum_a \lambda^a \left(\sum_i v_i^a d_i - N\right)}, \tag{30}$$

The replica indices a run from 1 to n. This expression can be easily averaged over the z_{it}. In the following, after performing the integrals, we shall bring the replica-partition function to the form $\mathbb{E}[Z^n] \propto \exp\left(-\beta N n g(\beta, N/T, N) + O(n^2)\right)$. Substituting this expression into the equation (29), and reinterpreting the number n as a real, we can carry out the limit $n \to 0$. As a result, we find that the function g in the exponent is, up to an unimportant additive constant, the average of the free energy density, that is $\mathbb{E}[f(\beta, r)] = const + \lim_{N\to\infty} g(\beta, r, N)$, where we used the notation $r = N/T$.

The meaning of the expression "reinterpreting the number n as a real" is that we perform an analytical continuation from the set of natural numbers to the reals, an operation whose result is not necessarily unique. The analytic continuation is the weak link in the chain of manipulations making up the replica method; without a rigorous proof of the uniqueness of analytic continuation the method can only be regarded as heuristic. A guarantee of the uniqueness of the continuation can only come from imposing some additional constraints on the problem. We conjecture that this constraint is the convexity of the objective function, which guarantees that the optimization problem has a single solution, and it is hard to imagine how the analytic continuation could lead to a different one. As the variance is convex, the replica method should produce the correct results for its minimization. This, of course, does not constitute a proof, but rigorous proofs exist in similar, and even much more complicated, problems in the theory of disordered systems [60], [61], [62], which lends support to our conjecture. Furthermore, the results to be derived below agree with the rigorous results by [30], [39], [43] in the thermodynamic limit, and also with the numerical experiments by [36].

3.2 Averaging over the samples

Let us average the replica partition function (30) with the density function (26). For this, we have to calculate the following integral:

$$(2\pi)^{-NT/2} \int \prod_{i,t} dz_{it} e^{-\frac{1}{2}\sum_{i,t} z_{it}^2 - i\sum_{a,i,t} \hat{u}_t^a v_i^a z_{it}} = e^{-\frac{1}{2}\sum_{it}\left(\sum_a v_i^a \hat{u}_t^a\right)^2}. \tag{31}$$

Let us introduce the overlap matrix by:

$$Q^{ab} = \frac{1}{N}\sum_{i=1}^{N} v_i^a v_i^b, \tag{32}$$

and transform (31) as follows:

$$e^{-\frac{1}{2N}\sum_{it}\left(\sum_a v_i^a \hat{u}_t^a\right)^2} \propto \int \prod_{ab} dQ^{ab} \int \prod_{ab} d\tilde{Q}^{ab} \\ e^{\frac{i}{2}N\sum_{a,b}\tilde{Q}^{ab}\left(Q^{ab} - \frac{1}{N}\sum_i v_i^a v_i^b\right) - \frac{N}{2}\sum_{ab} Q^{ab}\sum_t \hat{u}_t^a \hat{u}_t^b}. \tag{33}$$

Plugging this back into (30), after some rearrangement we obtain

$$\mathbb{E}[Z^n] \propto \int \prod_{ab} dQ^{ab} \int \prod_{ab} d\hat{Q}^{ab} e^{\frac{1}{2}N\sum_{ab}\hat{Q}^{ab}Q^{ab}} \int \prod_{a,t} du_t^a e^{-\frac{\beta}{2T}\sum_{a,t}\left(u_t^a - \frac{1}{T}\sum_s u_s^a\right)^2}$$
$$\int \prod_{a,t} d\hat{u}_t^a e^{-\frac{N}{2}\sum_{a,b}Q^{ab}\sum_t \hat{u}_t^a\hat{u}_t^b + i\sum_{a,t}\hat{u}_t^a u_t^a} \int \prod_a d\eta^a \int \prod_a d\lambda^a e^{i\sum_a \eta^a\left(N\mu - \frac{1}{T}\sum_t u_t^a\right) + iN\sum_a \lambda^a}$$
$$\int \prod_{a,i} dv_i^a e^{-\frac{1}{2}\sum_{a,b}\hat{Q}^{ab}\sum_i v_i^a v_i^b - i\sum_{a,i}(\eta^a\theta_i + \lambda^a d_i)v_i^a}, \quad (34)$$

where we have applied the replacement $\hat{Q}^{ab} = i\tilde{Q}^{ab}$; accordingly the integration with respect to the variables $\hat{Q}^{ab}$ is along the imaginary axis from $-i\infty$ to $i\infty$. The Gaussian integrals over $\hat{u}_t^a$, v_i^a, η^a and λ^a can then be performed easily. The only thing to watch out for is that as $\hat{Q}^{ab}$ is imaginary, in some of the Gaussian integrals the parameter standing in the place of the standard deviation will also be imaginary. This, however, does not pose a problem, because the contour of integration over $\hat{Q}^{ab}$ can be deformed so as to make it run to the right of the imaginary axis and return to the imaginary axis only at $\pm i\infty$. Thus, for finite values of $\hat{Q}^{ab}$ $\mathrm{Re}\left(\hat{Q}^{ab}\right) > 0$ and all the Gaussian integrals will be meaningful. Finally, we end up with the result:

$$\mathbb{E}[Z^n] \propto \int \prod_{ab} dQ^{ab} \int \prod_{ab} d\hat{Q}^{ab} e^{-N\left[G(\mathbf{Q},\hat{\mathbf{Q}};\beta) + O(1/N)\right]}, \quad (35)$$

where

$$G(\mathbf{Q},\hat{\mathbf{Q}};\beta) = \lim_{N\to\infty}\left[-\frac{1}{2}\sum_{a,b}\hat{Q}^{ab}\left(Q^{ab} - N\sigma^{*2}(\mu)\right) + \right.$$
$$\left. + \frac{1}{2r}\mathrm{Tr}\ln\mathbf{Q} + \frac{1}{2}\mathrm{Tr}\ln\hat{\mathbf{Q}} - \frac{1}{N}\ln A(\mathbf{Q},\hat{\mathbf{Q}};\beta)\right], \quad (36)$$

$$A(\mathbf{Q},\hat{\mathbf{Q}};\beta) = \int \prod_{a,t} du_a^t e^{-\frac{\beta}{2T}\sum_{at}\left(u_t^a - \frac{1}{T}\sum_s u_s^a\right)^2 - \frac{1}{2N}\sum_{a,b}\left[\mathbf{Q}^{-1}\right]^{ab}\sum_t u_t^a u_t^b}$$
$$e^{-\frac{1}{2}\sum_{ab}\hat{Q}^{ab}\left[\frac{\alpha^*}{T^2}\sum_{t,s}u_t^a u_s^b - \frac{N\eta^*(\mu)}{T}\sum_t(u_t^a + u_t^b)\right]}. \quad (37)$$

and the limit $N \to \infty$ is taken such that $r = N/T = const$. The quantities $\eta^*(\mu)$ and $\sigma^*(\mu)$ are defined by (6), and (7), respectively, and, in terms of the notations in Sec. 2, $\alpha^* = A/(AC - B^2)$. As a reminder: $N\sigma^*(\mu)$ is nothing but the true risk of the portfolio (corresponding to $r \to 0$). (The factor N appears because of the modified budget constraint). In the following, we shall assume that $\sigma^{*2}(\mu)$, $\eta^*(\mu)$ and α^* are of the order of $O(1/N)$. (For example, for $\sigma_{ij} = \delta_{ij}$ and $\mu_i = const$ this is so automatically.) It can be shown that this is merely a technical assumption which does not influence the validity of the results.

As we want to determine $A(\mathbf{Q},\hat{\mathbf{Q}};\beta)$ only in the thermodynamic limit we can omit the terms less than $O(N)$ to obtain:

$$A(\mathbf{Q},\hat{\mathbf{Q}};\beta)=\int\prod_{a,t}du_a^t e^{-\frac{1}{2N}\sum_{a,b,t,s}\left[\left(\beta r+[\mathbf{Q}^{-1}]^{ab}\right)\delta_{ts}+\frac{r^2}{N}\left(\alpha^*\hat{Q}^{ab}-\beta\delta^{ab}\right)\right]u_t^a u_s^b+\eta^*(\mu)r\sum_{a,b,t}\hat{Q}^{ab}u_t^b}\propto$$

$$\propto e^{-\frac{N}{2r}\left[\mathrm{Tr}\ln\left(\beta r\mathbf{I}+\mathbf{Q}^{-1}\right)+O(1/N)\right]}, \quad (38)$$

where $\mathbf{I}$ is the $n\times n$ identity matrix. Plugging this result back into (36) we get:

$$G(\mathbf{Q},\hat{\mathbf{Q}};\beta)=\frac{1}{2}\left[-\sum_{a,b}\hat{Q}^{ab}\left(Q^{ab}-\nu(\mu)\right)+\mathrm{Tr}\ln\hat{\mathbf{Q}}+\frac{1}{r}\mathrm{Tr}\ln\left(\beta r\mathbf{Q}+\mathbf{I}\right)\right], \quad (39)$$

where $\nu(\mu)=\lim_{N\to\infty}N\sigma^{*2}(\mu)$.

3.3 The "physical" meaning of the overlap matrix

Before continuing, let us take a closer look at how the matrix elements of $\mathbf{Q}$ can be interpreted. Let us consider two replicas with indices a and b. Let the vectors $\mathbf{v}^a$ and $\mathbf{v}^b$ be the configurations of the two systems. Then from $v_i^a=\sum_j w_j^a D_{ji}$ the overlap between replicas a and b in terms of the portfolio weights is

$$Q^{ab}=\frac{1}{N}\sum_{i=1}^{N}v_i^a v_i^b=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_{ij}w_i^a w_j^b. \quad (40)$$

Therefore, NQ^{ab} is nothing but the true covariance (calculated on the basis of complete information, $r\to 0$) of the portfolios $\mathbf{w}^a$ and $\mathbf{w}^b$. In particular, NQ^{aa} is the true variance (risk) of the portfolio $\mathbf{w}^a$.

As the portfolio weights have been normalized as $\sum_i w_i=N$, the standard deviation of the true optimum will be $N\sigma^*(\mu)$, instead of $\sigma^*(\mu)$. Then the estimation error q_0 of the portfolio $\mathbf{w}^a$ will be a simple function of Q^{aa}:

$$q_0(\mathbf{w}^a)=\frac{Q^{aa}}{N\sigma^{*2}(\mu)}. \quad (41)$$

The expected error of the estimated optimum is therefore equal to the above expression in the thermodynamic limit, at zero temperature. As the energy surface is strictly convex, we expect that at low temperature every replica tends to the same minimum, therefore Q^{aa} will be independent of the replica index. (In the next subsection this will be explicitly shown to be the case.) Accordingly,

$$\mathbb{E}\left[q_0\right]=\frac{1}{\nu(\mu)}\lim_{\beta\to\infty}\lim_{N\to\infty}Q^{aa} \quad (42)$$

independently of a. The equilibrium value of Q^{aa} will be calculated in the next subsection.

3.4 The saddle point

As we wish to calculate the integral (35) in the limit $N \to \infty$, we can use the saddle point method:

$$\int \prod_{ab} dQ^{ab} \int \prod_{ab} d\hat{Q}^{ab} e^{-N\left[G(\mathbf{Q},\hat{\mathbf{Q}};\beta)+O(1/N)\right]} \sim e^{-N \min_{\mathbf{Q},\hat{\mathbf{Q}}} G(\mathbf{Q},\hat{\mathbf{Q}};\beta)}, \tag{43}$$

which becomes exact in the thermodynamic limit. The saddle point conditions are:

$$2\frac{\partial G_\beta}{\partial Q^{ab}} = -\widehat{Q}^{ab} + \beta \left[(r\beta\mathbf{Q} + \mathbf{I})^{-1} \right]^{ba} = 0, \tag{44}$$

$$2\frac{\partial G_\beta}{\partial \widehat{Q}^{ab}} = -Q^{ab} + \nu(\mu) + \left[\widehat{\mathbf{Q}}^{-1} \right]^{ba} = 0. \tag{45}$$

The solution is not hard and leads us to the following result:

$$Q^{ab}_{sp} = \frac{1}{1-r}\left[\nu(\mu) + \beta^{-1}\delta^{ab}\right], \tag{46}$$

$$\hat{Q}^{ab}_{sp} = \beta(1-r)\left[\frac{\beta r \nu(\mu)}{\beta r \nu(\mu) + 1} + \delta^{ab}\right], \tag{47}$$

where the subscript *sp* signifies the saddle point. We can see therefore that at the saddle point the overlap matrix and its conjugate are invariant w.r.t. the permutation of replicas, the only thing that matters is whether we are considering the same replica or distinct ones. In other words, the saddle point is replica symmetric. As we mentioned in the previous section, this is a consequence of the fact that, independently of the sample, the Hamiltonian and the constraints determine a single, unique ground state, provided $r < 1$, that is the sample size is sufficiently large.

It can be seen furthermore that the difference between the diagonal and off-diagonal elements of $\mathbf{Q}_{sp}$ is proportional to β^{-1}, so it vanishes at zero temperature, which is explained by the fact that all the replicas settle into the same equilibrium state, so the difference between the self-overlap and the overlap of different replicas will become zero.

At this point, from (42) we can explicitly determine $\mathbb{E}[q_0]$:

$$\mathbb{E}[q_0] = \lim_{\beta\to\infty} \frac{1}{1-r}\left(1 + \frac{1}{\beta\nu(\mu)}\right) = \frac{1}{1-r}. \tag{48}$$

This result is in agreement with the results obtained in the special case of the global minimum and i.i.d. zero mean returns in [25] and [63]. The most surprising feature of (48) is that for large portfolios the average estimation error depends only on the ratio N/T, completely independently of the parameters σ_{ij} and μ_i of the distribution, and also of the expected return μ.

In order to obtain the free energy density, let us substitute (44) and (45) into (39). Writing $Q^{ab} = q + \Delta q \delta^{ab}$ we have $\mathrm{Tr}\ln\mathbf{Q} = n\left(q/\Delta q + \ln\Delta q\right) + O(n^2)$, and after some algebra we find

$$G(\mathbf{Q}_{sp}, \hat{\mathbf{Q}}_{sp}; \beta) = \frac{1}{2} n\beta\nu(\mu)(1-r) + O(n^2). \tag{49}$$

According to Subsection 3.1 we can then write the free energy density as follows:

$$\mathbb{E}[f(\beta, r)] = \frac{f_0}{\beta} + \lim_{n\to 0} \frac{1}{n\beta} G(\mathbf{Q}_{sp}, \hat{\mathbf{Q}}_{sp}; \beta) = \frac{f_0}{\beta} + \frac{1}{2}\nu(\mu)(1-r), \tag{50}$$

where the additive term f_0 comes from the multiplicative constants in $\mathbb{E}[Z^n]$. Reviewing the integrals we can see that f_0 does not depend on β, and because of the limits $n \to 0$ and $N \to \infty$ it does not depend on n, N and T either.

The zero temperature limit of the free energy is the minimum value of the Hamiltonian, so for large N the estimated risk at the optimum, averaged over the samples, is:

$$\mathbb{E}\left[\hat{\sigma}^{*2}(\mu)\right] = 2N \lim_{\beta\to\infty} \mathbb{E}[f(\beta, r)] = N^2 \sigma^{*2}(\mu)(1-r). \tag{51}$$

The estimated in-sample loss is thus seen to vanish inversely proportionally to the out of sample estimation error at the critical point for a generic covariance matrix and all along the efficient frontier. This phenomenon was first observed in numerical experiments and confirmed analytically for a diagonal covariance matrix at the global minimum of the variance in [25]. The same inverse proportionality was found in the case of the ES risk measure [51]. These findings demonstrate how grossly in-sample estimates can underestimate risk.

We conclude this subsection with and important remark. The fact that we could determine the extremum of (39), was a very lucky happenstance, due to the simplicity of the objective function. In more complicated cases the usual procedure is to assume replica symmetry, express the $G(\mathbf{Q}, \hat{\mathbf{Q}}; \beta)$ as a function of the diagonal and off-diagonal elements of $\mathbf{Q}$ and $\hat{\mathbf{Q}}$, and minimize over these variables. As the saddle point is, in general, not necessarily replica symmetric, one has to investigate the stability of the replica symmetric saddle point. In some cases (for example, in spin glasses) replica symmetry breaks down at a non-zero temperature, and in the region below this a symmetry breaking solution becomes stable [64]. Our present problem is much simpler, and the stability of the replica-symmetric saddle point is preserved all the way down to zero temperature. This is demonstrated in the next subsection.

3.5 The stability of the replica-symmetric saddle point

The saddle point is stable, if the functional $G(\mathbf{Q}, \hat{\mathbf{Q}}; \beta)$ takes its minimum at the point $\mathbf{Q}_{sp}$. Since the conjugate variable is complex and the contour of integration, originally running along the imaginary axis, can be deformed, it does not matter whether the extremum of G with respect to $\hat{\mathbf{Q}}$ is a minimum or maximum. Therefore, it is sufficient to check the stability of $G(\mathbf{Q}; \beta) = G(\mathbf{Q}, \hat{\mathbf{Q}}_{sp}; \beta)$ at $\mathbf{Q}_{sp}$. Expressing $\hat{\mathbf{Q}}$ from (45) and substituting into (39) we obtain

$$G(\mathbf{Q}; \beta) = \frac{1}{2}\left[-n - \mathrm{Tr}\ln\left(\mathbf{Q} - \nu(\mu)\mathbf{U}\right) + \frac{1}{r}\mathrm{Tr}\ln\left(\beta r\mathbf{Q} + \mathbf{I}\right)\right], \tag{52}$$

where $\mathbf{U}$ denotes the $n \times n$ matrix with all its elements equal to 1. Let us now calculate the Hess matrix of $G(\mathbf{Q};\beta)$ (absorbing a factor 2 for simplicity):

$$H^{ab,cd} = 2\frac{\partial^2}{\partial Q^{ab}\partial Q^{cd}} G(\mathbf{Q};\beta) = \left[(\mathbf{Q}-\nu(\mu)\mathbf{U})^{-1}\right]^{ac}\left[(\mathbf{Q}-\nu(\mu)\mathbf{U})^{-1}\right]^{bd} - \\ - r\beta^2\left[(\beta r\mathbf{Q}+\mathbf{I})^{-1}\right]^{ac}\left[(\beta r\mathbf{Q}+\mathbf{I})^{-1}\right]^{bd}. \quad (53)$$

Substituting $\mathbf{Q}_{sp}$ into the Hess matrix, we get $H_{sp}^{ab,cd} = R^{ac}R^{bd}$, where

$$R^{ab} = \rho + \Delta\rho\delta^{ab} = \beta(1-r)^{3/2}\left(-\frac{\beta r\nu(\mu)}{\beta r\nu(\mu)n+1} + \delta^{ab}\right). \quad (54)$$

The condition for the stability of the saddle point is that $H_{sp}^{ab,cd}$ be strictly positive definite. In order to check this, let us solve the eigenvalue problem

$$\sum_{c=1}^{n}\sum_{d=1}^{n} H_{sp}^{ab,cd} S_m^{cd} = \lambda_m S_m^{ab}, \quad (55)$$

where λ_m are the eigenvalues, and the number of the eigenvectors S_m^{ab} is n^2. We must find that all the eigenvalues are positive. Because of the symmetry of R^{ab} the matrix $H_{sp}^{ab,cd}$ is also symmetric, i.e. $H_{sp}^{ab,cd} = H_{sp}^{cd,ab}$, thus an orthonormed basis can be selected from the eigenvectors S_m^{ab}. In addition to this, the Hessian also displays an important further symmetry, namely $H_{sp}^{ab,cd} = H_{sp}^{ba,dc}$. An immediate consequence of this is that whenever S^{ab} is an eigenvector, its transposed S^{ba}, its symmetric part $(S^{ab}+S^{ba})/2$ and antisymmetric part $(S^{ab}-S^{ba})/2$ are also eigenvectors, and they belong to the same eigenvalue. The $n(n+1)/2$ dimensional space of symmetric matrices, and the $n(n-1)/2$ dimensional space of antisymmetric matrices are thus invariant subspaces of the Hessian. According to elementary considerations, the invariant subspaces of $H_{sp}^{ab,cd}$ are as shown in the Table below:

Subspace	Eigenvalue	Eigenvectors	Multiplicity
I	$\lambda = (\rho n + \Delta\rho)^2$	$S^{ab} = 1$ for all a and b	1
II	$\lambda = (\rho n + \Delta\rho)\Delta\rho$	$S^{ab} =$ $\begin{cases} 1-n & \text{if } a=b=k, \\ \frac{2-n}{2} & \text{if either } a=k, \text{ or } b=k, \\ 1 & \text{if } a\neq k \text{ and } b\neq k. \end{cases}$ $k = 1, 2, ..., n-1$	$n-1$
III	$\lambda = \Delta\rho^2$	Those symmetric matrices, which leave the vector $(1,1,1,...,1)$ invariant.	$\frac{1}{2}n(n-1)$
IV	$\lambda = (\rho n + \Delta\rho)\Delta\rho$	$S^{ab} = \delta^{ak} - \delta^{bk}$ $k = 1, 2, ..., n-1$	$n-1$
V	$\lambda = \Delta\rho^2$	Those matrices S^{ab} , for which $\sum_{b=1}^{n} S^{ab} = 0$ for arbitrary a.	$\frac{1}{2}(n-1)(n-2)$

Symmetric matrices belong to subspaces I, II and III, and by adding up the multiplicities we can see that these subspaces span the full $n(n+1)/2$ dimensional linear space of $n \times n$ symmetric matrices. Similarly, one can convince herself that the invariant subspaces IV and V span the full $n(n-1)/2$ dimensional space of the $n \times n$ antisymmetric matrices. The grand total of multiplicities is n^2, so we have found all the eigenvalues and eigenvectors.

Therefore, the Hessian $H_{sp}^{ab,cd}$ has three different eigenvalues:

$$\lambda_1 = (\rho n + \Delta\rho)^2 = \frac{\beta^2(1-r)^3}{\left(\beta r\nu(\mu)n + 1\right)^2}, \tag{56}$$

$$\lambda_2 = (\rho n + \Delta\rho)\Delta\rho = \frac{\beta^2(1-r)^3}{\beta r\nu(\mu)n + 1}, \tag{57}$$

$$\lambda_3 = \Delta\rho^2 = \beta^2(1-r)^3. \tag{58}$$

Of these, λ_1 has multiplicity 1, λ_2 is $2(n-1)$-fold degenerate, while λ_3 is $(n-1)^2$-fold degenerate. The eigenvalues are positive as long as $r < 1$. Therefore the replica symmetric saddle point is stable at any temperature, provided $r < 1$, that is $T > N$. At the critical point $r = 1$ all three eigenvalues vanish, the model becomes unstable against fluctuations in any directions. With this we have given the characterisation of the noise sensitivity of the Markowitz problem in the thermodynamic limit.

4 Summary and discussion

Several years after the first publications on the statistical physics approach to the optimization of the Expected Shortfall [47] and Mean Absolute Deviation [48] risk measures, this paper addressed the problem of optimizing the variance. In contrast to previous studies, which focused on the global minimum, in the present paper the simplicity of the objective function allowed us to extend the replica method to the full Markowitz problem, by considering a generic covariance matrix and, in addition to the budget constraint, also the constraint on the expected return of the portfolio. From the point of view of the method itself, the most interesting feature we encountered was that the replica symmetric solution emerged as the unique solution, without having to assume this symmetry in advance. It was also straightforward to check the stability of this solution, with a foregone conclusion. These features point in the direction of a possible rigorous foundation of the replica method in the case of convex risk measures.

The solution we found displayed a phase transition, a sharp change in the nature of the optimization problem. For the sake of the physicists among our readers, we point out that the statistical physics model corresponding to variance optimization is a version of the so called Gaussian model, familiar from introductory courses on phase transitions. It is an N-component ϕ^2 model in zero spatial dimensions, or, alternatively, a Ginzburg-Landau model, again with no spatial dependence and without the usual ϕ^4, or any other higher order, term. What makes the model somewhat less than trivial is the fact that the components of the covariance matrix are not given in advance, but have to be estimated from finite samples, and the resulting free energy has to be averaged over these samples. This last step corresponds to quenched averaging – hence the application of replicas.

There are two further features that distinguish variance optimization from the Gaussian model:

- the budget constraint, meaning the components of the weight vector sum to 1 (or any other arbitrarily chosen fixed number),
- the expected return of the portfolio is fixed.

In the absence of these constraints the optimization task would be completely trivial: the solution would be the null-vector, as indeed, in the Gaussian model the ground state configuration corresponds to the zero value of the order parameter components. If we were dealing with a spin model less naive than the Gaussian, we should impose some other type of constraint, such as stipulating that the components of the weight vector only take the values ± 1 (Ising spins), or are subject to a soft bimodal distribution (coarse grained cell spins), or subject to a global Euclidean norm constraint (spherical model). However, we are not treating a spin model here, in the simple setup which we are considering and which corresponds to the text-book version of the Markowitz problem, the portfolio weights can take any real value, positive or negative, as long as they satisfy the above two constraints.

Negative weights correspond to what are called "short positions" in finance. Unlimited short selling is seldom permitted, but without short positions it would be impossible to hedge risk, and also impossible to fully exploit correlations existing between the various assets in the investment universe.

In real life there are limitations on short positions (typically on long, i.e. positive weight positions, as well). These limits depend on the type of financial institution (hedge funds are permissive, while pension funds are forbidden to assume short positions), but also on the general financial and legal environment. The presence of such constraints elevates the difficulty of optimization to a qualitatively different level, where only numerical methods have been available so far. Our approach is analytical and the present paper is concerned with the simplest version of the Markowitz portfolio optimization problem, as laid out in eqs. (1), (2) and (3). However, we can see already at this stage that the technique of replicas will allow us to incorporate the most frequently occurring constraints in the analytic approach, thus the present paper should be regarded as a starting platform for such an extended program.

It should also be pointed out that the optimization of variance does not necessarily display a phase transition. If, for example, the probability distribution of returns is known and if it is sufficiently narrow so that the covariance matrix exists, then it is positive definite and no instability arises. This is the classical statistical setup where the number of data T is much larger than the dimension N of the portfolio, which in our language corresponds to the limit $r \to 0$. With large institutional portfolios with N in the hundreds or thousands, and with the length T of the time series limited by stationarity considerations to at most a few hundred, we are in the high-dimensional limit where the ratio r is not small and the covariance matrix has to be estimated from observed data by the maximum likelihood estimator in eq. (17). The estimated covariance matrix will be positive definite with probability 1, as long as $N \leq T$. However, when N becomes larger than T, the covariance matrix becomes positive semidefinite and $N - T$ zero modes appear (the number of unknowns, N, exceeds the number of equations, T). This is where we enter the other "phase" (in terms of the ratio N/T this corresponds to the critical point $r = 1$). This other phase cannot be called the ordered phase, because no order emerges, instead a continuum of solutions appears: any linear combination of the zero modes becomes a solution of the optimization problem and the portfolio variance becomes identically zero. (The vanishing of all the eigenvalues of the Hessian as r approaches 1 is an indication of this flat phase space landscape.) This strictly corresponds to what happens in the

Gaussian model, where the restoring force disappears at the critical point, and there being no higher order term in the Hamiltonian to constrain the fluctuations of the order parameter, they become infinitely large.

According to the above, the optimization becomes meaningless above the critical point. High-dimensional statistics has developed powerful methods to remedy this difficulty by various regularization procedures, (see e.g. [65]). Technically, these regularizers show up as portfolio constraints, and as such are beyond the scope of the present paper.

As must be clear from the foregoing discussion, the instability at $r = 1$ does not depend on the distribution of returns: it is a purely linear algebraic or geometric phenomenon, with the rank of the covariance matrix becoming less than its dimension.

One might then ask why should we call this simple instability a phase transition at all? Part of the answer is that the completely similar transition into a featureless phase in the Gaussian model has always been regarded as a phase transition – at least for pedagogical purposes. More seriously, this transition is one example of a host of algorithmic phase transitions (see e.g. [66]) and a large number of sharp transitions in random geometry that are also called phase transitions (see e.g. [57], or the more theoretical paper [67]).

Furthermore, there are a few features that this transition shares with bona fide phase transitions. One of these is the universality of the exponent of the scaling law for the estimation error $1/(1-r)$. The value of this exponent seems to be the same in a number of similar problems: it was obtained by numerical simulations in the case of portfolio optimization for the variance as risk measure in refs. [22, 25, 27, 36], for Expected Shortfall in [47, 49, 50, 52], for the minimax risk measure in [52], even for non-stationary underlying processes in a GARCH framework [37]. The invariance of this exponent for these different "Hamiltonians" is analogous to the independence of the usual critical exponents of microscopic details.

We have, however, found another, stronger sort of universality here: the independence of the critical point itself (the critical point does not depend on the parameters of the distribution of the returns, it depends only on the ratio N/T). This finding is in accord with the results of large scale numerical experiments on variance optimization with fat-tailed, Student-distributed returns [56]. Likewise, the behavior of Expected Shortfall in the region where it can be optimized is certainly different depending on the distribution of returns, but the phase boundary along which the estimation error blows up is again insensitive to the distribution of returns [52].

As further, independent support for the universality of the critical point in these sort of geometric problems, we recall the example of the minimax risk measure, the best combination of the worst losses. In addition to being a well defined portfolio optimization problem [11], the minimax has been shown [36] to be equivalent to the following problem in high dimensional random geometry: What is the probability that T random hyperplanes thrown into and N-dimensional space make a convex polytope? The answer was found by Schmidt and Mattheiss [68] for any N and T. Their result shows that this probability is large for $T > N$, small for $T < N$, and the transition becomes sharp for N and T large, with a critical ratio $N/T = 1/2$, independently of the distribution of the parameters of the random planes, provided this distribution is symmetric. There are a large number of random geometric transitions with universal

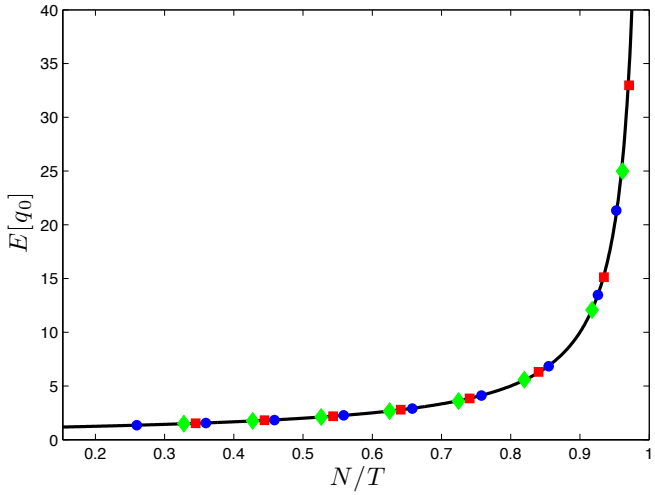

Figure 2: Estimation error as a function of $r = N/T$: comparison between simulations and theoretical result for different structures of the covariance matrix. The solid black line represents the analytical formula of the estimation error given by equation (48). Blue dots: numerical simulations for the global minimum (i.e. the constraint on the expected return is not enforced) of the variance of a system with $N = 100$ and a diagonal covariance matrix with all diagonal elements equal to $1/N$, in order to make the total variance of the portfolio of order $\mathcal{O}(1)$. Red squares: numerical simulations of a system with $N = 100$, $\mu_i = 0.04$ for all i, and a diagonal covariance matrix with all diagonal elements equal to $1/N$. Green diamonds: numerical simulations of a system with $N = 100$, $\mu_i = 0.04$ for all i, and a block-diagonal covariance matrix with four blocks. The diagonal elements of the covariance matrix are equal to $1/N$, the covariance between assets within the same block is set to $0.1/N$, and the covariance between assets of different blocks is zero. All results from simulations are averaged over 1000 realizations. The agreement between numerical and analytical results is very good.

critical points discovered by Donoho and Tanner [57].

Similarly, the optimization of the variance can be rephrased as follows [13]: we have an N-dimensional random ellipsoid and two intersecting random hyperplanes. Optimization corresponds to finding the point of tangency of the plane of intersection with the ellipsoid. As long as $N \leq T$, the ellipsoid is non-degenerate, and as a convex object it has a unique point of tangency with the plane. When $N > T$, one or more principal axes of the ellipsoid become infinite, accordingly the ellipsoid goes over into a cylindrical object with one or more principal directions along which its curvature is zero, therefore a tangent plane will have a continuum of points in common with the ellipsoid. Rephrased like this, the problem is again one in high-dimensional random geometry. This view of the problem makes it clear that the loss of uniqueness and the appearance of the continuum of solutions cannot depend on the probability distribution of the returns; it depends only on the rank of the covariance matrix, that is whether N is smaller or larger than T.

It is always true that the out of sample estimation error is larger than the in-sample

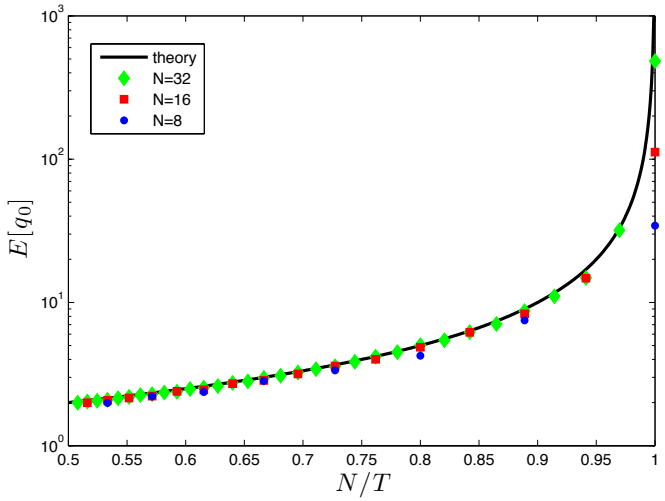

Figure 3: Estimation error as a function of $r = N/T$: comparison between simulations and theoretical result for different values of N. The solid black line represents the analytical formula of the estimation error given by equation (48). Blue dots: numerical simulations of a system with $N = 8$. Red squares: numerical simulations of a system with $N = 16$. Green diamonds: numerical simulations of a system with $N = 32$. All cases refer to the portfolio corresponding to the global minimum of the risk functional. In all cases the covariance matrix is diagonal with diagonal elements equal to $1/N$ (in order to make the total variance of the portfolio of order $\mathcal{O}(1)$) and results are averaged over 1000 realizations. Far from the critical point the agreement between numerical and analytical results is very good already for N quite small. Closer to the critical point the agreement improves quickly with N.

error. Our analytic result is a clear illustration of this: while the out of sample error diverges as $1/(1-r)$, the "optimal" value of the variance tends to zero as $(1-r)$ times its true value. (It is identically zero beyond the critical point where the optimal weight vector belongs to the null-space of the covariance matrix.)

Although we have run extensive numerical simulations on special cases of this problem (for various covariance matrices) in refs. [25, 27, 36], in order to give an idea about the speed of convergence to the large N limit, we provide some illustrations to conclude this discussion.

In figure 2 we show a comparison between the estimation error given by eq. (48) and that computed from numerical simulations performed for a portfolio of $N = 100$ assets for different structures of the covariance matrix. The figure makes it clear that the formula computed using the replica calculation describes very well the behavior of the estimation error as a function of the parameter r.

Finally, in figure 3 we show how the estimation error computed from numerical simulations converges to the analytical expression upon increasing N. From the plot it is clear that far from the transition the agreement is extremely good already for N as small as 8. Closer to the transition a larger number of assets is required, but the convergence appears to be quite fast.

With the replica approach to variance optimization hereby established, we can turn to finding a cure for the divergent estimation error by regularization. This is left for a subsequent publication.

Acknowledgement

I.K. thanks Risi Kondor, Bálint Komjáti and Christoph Memmel for helpful discussions. F.C. acknowledges support of the Economic and Social Research Council (ESRC) in funding the Systemic Risk Centre (ES/K002309/1).

References

[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] R. C. Merton. An analytic derivation of the efficient portfolio frontier. *Journal of Financial and Quantitative Analysis*, 7:1851–1872, 1972.

[4] J. P. Dickinson. The reliability of estimation procedures in portfolio analysis. *Journal of Financial and Quantitative Analysis*, 9:447 – 462, 1974.

[5] 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.

[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] P. Jorion. Bayes-stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis*, 21:279–292, 1986.

[8] H. Konno. Portfolio optimization using l1 risk function, 1988. Technical Report IHSS 88-9, Tokyo Institute of Technology.

[9] H. Konno and H. Yamazaki. Mean-absolute deviation portfolio optimization model and its application to tokyo stock market. *Management Science*, 37:519–531, 1991.

[10] E. J. Elton and M. J. Gruber. *Modern Portfolio Theory and Investment Analysis*. Wiley, New York, 1995.

[11] M. R. Young. A minimax portfolio selection rule with linear programming solution. *Management science*, 44(5):673–683, 1998.

[12] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999.

[13] A. Gábor and I. Kondor. Portfolios with nonlinear constraints and spin glasses. *Physica A: Statistical Mechanics and its Applications*, 274(1):222–228, 1999.

[14] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters. Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, 83:1467–1470, 1999.

[15] 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.*, 83:1471–74, 1999.

[16] 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*, 65:066126, 2000.

[17] P. Jorion. *VaR: The New Benchmark for Managing Financial Risk.* McGraw-Hill, New York, 2000.

[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] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

[20] C. Acerbi and D. Tasche. Expected shortfall: a natural coherent alternative to value at risk. *Economic Notes*, 31(2):379–388, 2002.

[21] C. Acerbi and D. Tasche. On the coherence of expected shortfall. *Journal of Banking and Finance*, 26(7):1487–1503, 2002.

[22] S. Pafka and I. Kondor. Noisy covariance matrices and portfolio optimization. *Eur. Phys. J.*, B 27:277–280, 2002.

[23] R. Jagannathan and T. Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58:1651–1684, 2003.

[24] 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.

[25] S. Pafka and I. Kondor. Noisy covariance matrices and portfolio optimization II. *Physica*, A 319:487–494, 2003.

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

[27] S. Pafka and I. Kondor. Estimated correlation matrices and portfolio optimization. *Physica*, A 343:623–634, 2004.

[28] B. Scherer and R. D. Martin. *Introduction to Modern Portflio Optimization With NUOPT and S-PLUS.* Springer, 2005.

[29] A. Kempf and C. Memmel. Estimating the global minimum variance portfolio. *Schmalenbach Business Review*, 58:332–348, 2006.

[30] Y. Okhrin and W. Schmid. Distributional properties of portfolio weights. *Journal of Econometrics*, 134:235 – 256, 2006.

[31] E. J. Elton, M. J. Gruber, S. J. Brown, and W. N. Goetzmann. *Modern portfolio theory and investment analysis.* Wiley: Hoboken, NJ, 2007.

[32] 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.

[33] V. Golosnoy and Y. Okhrin. Multivariate shrinkage for optimal portfolio weights. *The European Journal of Finance*, 13:441–458, 2007.

[34] R. Kan and G. Zhou. Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 42(3):621–656, 2007.

[35] P. Kolm, F. J. Fabozzi, D. A. Pachamanova, and S. M. Focardi. Robust portfolio optimization. *Journal of Portfolio Management*, 2007.

[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] I. Varga-Haszonits and I. Kondor. Noise sensitivity of portfolio selection in constant conditional correlation GARCH models. *Physica*, A385:307–318, 2007.

[38] G. Frahm. Linear Statistical Inference for Global and Local Minimum Variance Portfolios. *Statistical Papers*, 2008. DOI: 10.1007/s00362-008-0170-z.

[39] G. K. Basak, R. Jagannathan, and T. Ma. A jackknife estimator for tracking error variance of optimal portfolios constructed using estimated inputs. *Management Science*, 55(6):990–1002, 2009.

[40] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Science*, 106(30):12267–12272, 2009.

[41] 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(5):798–812, 2009.

[42] V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus naive diversification: how efficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22(22):1915–1953, 2009.

[43] G. Frahm and C. Memmel. Dominating estimators for minimum-variance portfolios. *Journal of Econometrics*, 159(2):289–302, 2010.

[44] I. Kondor and I. Varga-Haszonits. Instability of portfolio optimization under coherent risk measures. *Advances in Complex Systems*, 13(03):425–437, 2010.

[45] S. Still and I. Kondor. Regularizing portfolio optimization. *New Journal of Physics*, 12(7):075034, 2010.

[46] J. Bun, J-P. Bouchaud, and M. Potters. My beautiful laundrette: Cleaning correlation matrices for portfolio optimization. *available at https://www.researchgate.net/publication/302339055*, 2016.

[47] S. Ciliberti, I. Kondor, and M. Mézard. On the feasibility of portfolio optimization under expected shortfall. *Quantitative Finance*, 7:389–396, 2007.

[48] S. Ciliberti and M. Mézard. Risk minimization through portfolio replication. *Eur. Phys. J.*, B 57:175–180, 2007.

[49] F. Caccioli, S. Still, M. Marsili, and I. Kondor. Optimal liquidation strategies regularize portfolio selection. *The European Journal of Finance*, 19(6):554–571, 2013.

[50] Fabio Caccioli, Imre Kondor, Matteo Marsili, and Susanne Still. Liquidity risk and instabilities in portfolio optimization. *International Journal of Theoretical and Applied Finance*, 19(05):1650035, 2016.

[51] I. Kondor, F. Caccioli, G. Papp, and M. Marsili. Contour map of estimation error for expected shortfall. *Available at http://ssrn.com/abstract=2567876 and http://arxiv.org/abs/1502.0621*, 2015.

[52] F. Caccioli, I. Kondor, and G. Papp. Portfolio optimization under expected shortfall: contour maps of estimation error. *arXiv preprint arXiv:1510.04943*, 2015.

[53] F. Papp, G. Caccioli and I. Kondor. Variance-bias trade-off in portfolio optimization under expected shortfall with ℓ_2 regularization. *available at http:// arXiv:1602.08297v1 [q-fin.PM]*, 2016.

[54] Basel Committee on Banking Supervision. *Minimum capital requirements for market risk*. Bank for International Settlements, 2016.

[55] T. Shinzato. Minimal investment risk of portfolio optimization problem with budget and investment concentration constraints. *available at arXiv:1605.06845v1 [q-fin.PM]*, 2016.

[56] B. Komjati. The instability of portfolio selection (unpublished msc thesis). Technical report, Eötvös University, Budapest, 2008.

[57] D. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of The Royal Society A, Mathematical Physical and Engineering Sciences*, 367:4273–93, 2009.

[58] I. Kondor and I. Varga-Haszonits. Divergent estimation error in portfolio optimization and in linear regression. *Eur. Phys. J.*, B64:601–605, 2008.

[59] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220:671–680, 1983.

[60] F. Guerra and F. L. Toninelli. The thermodynamic limit in mean field spin glass model. *Commun. Math. Phys.*, 230:71–79, 2002.

[61] F. Guerra and F. L. Toninelli. The infinite volume limit in generalized mean field disordered models. *Markov Proc. Rel. Fields*, 9:195–207, 2003.

[62] M. Talagrand. *Spin glasses: a challenge for mathematicians: cavity and mean field models*, volume 46. Springer Science & Business Media, 2003.

[63] Z. Burda, J. Jurkiewicz, and M. A. Nowak. Is econophysics a solid science? *Acta Physics Polonica*, B 34:87 – 131, 2003.

[64] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin glass theory and beyond*. World Scientific Lecture Notes in Physics Vol. 9, World Scientific, Singapore, 1987.

[65] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[66] M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[67] D. Amelunxen, M. Lotz, M. B. McCoy, and Joel A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. Technical report, DTIC Document, 2013.

[68] B. K. Schmidt and T.H. Mattheiss. The probability that a random polytope is bounded. *Mathematics of Operations research*, 2(3):292–296, 1977.