# Towards Large Scale Continuous EDA: A Random Matrix Theory Perspective

Ata Kabán
School of Computer Science
University of Birmingham
School of Computer Science
Edgbaston, UK, B15 2TT
A.Kaban@cs.bham.ac.uk

Jakramate Bootkrajang
School of Computer Science
University of Birmingham
School of Computer Science
Edgbaston, UK, B15 2TT
JXB008@cs.bham.ac.uk

Robert J. Durrant
School of Computer Science
University of Birmingham
School of Computer Science
Edgbaston, UK, B15 2TT
R.J.Durrant@cs.bham.ac.uk

## ABSTRACT

Estimation of distribution algorithms (EDA) are a major branch of evolutionary algorithms (EA) with some unique advantages in principle. They are able to take advantage of correlation structure to drive the search more efficiently, and they are able to provide insights about the structure of the search space. However, model building in high dimensions is extremely challenging and as a result existing EDAs lose their strengths in large scale problems.

Large scale continuous global optimisation is key to many real-world problems of modern days. Scaling up EAs to large scale problems has become one of the biggest challenges of the field.

This paper pins down some fundamental roots of the problem and makes a start at developing a new and generic framework to yield effective EDA-type algorithms for large scale continuous global optimisation problems. Our concept is to introduce an ensemble of *random projections* of the set of fittest search points to low dimensions as a basis for developing a new and generic divide-and-conquer methodology. This is rooted in the theory of random projections developed in theoretical computer science, and will exploit recent advances of non-asymptotic random matrix theory.

## Categories and Subject Descriptors

G.1.6 [**Optimization**]: unconstrained optimization; I.2.6 [**Learning**]: Parameter Learning

## General Terms

Algorithms, Theory

## Keywords

large scale optimisation, estimation of distribution algorithms, random projections, random matrix theory

## 1. INTRODUCTION

Estimation of distribution algorithms (EDAs) are population-based stochastic black-box optimisations methods that have been recognised as a major paradigm of Evolutionary Computation (EC) [21]. Contrary to the majority of traditional EC that takes no advantage of any correlation structure of the fittest sample, there is no crossover or mutation in EDAs – instead, EDAs guide the search for the global optimum by estimating the distribution of the fittest sample and drawing new candidates from this distribution.

However, as the search space dimensionality increases EDA type methods decline very quickly. Indeed, attempts to use the full power of continuous EDA are scarce when the search space exceeds 50-100 dimensions. Current practice of EDA most often resorts to independence models or models with some pre-defined limited dependency structure [29, 6, 11, 24] in exchange for feasibility even in moderate scale problems. Some authors employ heavy tail search distributions, for example [29] propose a univariate EDA (UMDAc) with Gaussian and Lévy search distribution for large scale EDA. While this improves the exploration ability to some extent, a univariate model unfortunately means that non-separable problems cannot be tackled adequately – fact both proved theoretically [20, 21] and shown experimentally [10]. A method that goes beyond this is sep-CMA-ES [24]. It imposes a diagonal constraint on the covariance in a different way than UMDAc – therefore by construction it is designed for rotated separable problems. Note, however, that rotated separable problems are not equivalent to nonseparable problems in general.

Large scale continuous optimisation problems are one of the most important concerns in EC research in the recent years because they appear in many real-world problems such as computational vision, data mining, bio-computing, atmospheric sciences, and robotics. Many optimisation methods suffer from the curse of dimensionality and deteriorate quickly when dimension $d > 100$. The state of the art best performers are EC methods that use cooperative co-evolution [30], multi-level co-evolution [31], and hybrid methods that include local searches [22].

## 2. THE CHALLENGES

In order to appreciate the issues involved with estimating a high dimensional distribution (e.g. a Gaussian) we must build on their intrinsic properties, which defeat the intuition rooted in low dimensional experiences. There are several


*GECCO'13,* July 6–10, 2013, Amsterdam, The Netherlands.

fundamental reasons for the failure of EDA in large scale problems, including the following:

(1) Estimation problems: The sample size required to produce a reliable estimate of the distribution of high fitness individuals grows exponentially with the dimension of the search space [15]. If sample size is insufficient, the eigenvalues of the covariance are misestimated [27] and bogus correlations appear in the maximum likelihood covariance estimates.

(2) Geometric problems: When dimension $d$ is large, the contrast between pairwise distances may vanish [4, 12]. This is exemplified by the data piling problems in high dimensional space observed in [14].

(3) Computational problems: The computation cost of sampling from a full $d$-dimensional Gaussian distribution is $\mathcal{O}(d^3)$ [9] and this becomes prohibitive when $d$ is very large. In addition, the problem characteristics may often change with the dimensionality [25].

## 2.1 Reachability of the global optimum

Let $x^* \in \mathbb{R}^d$ denote the global optimum and let $B(x^*, \epsilon)$ be the $d$-dimensional ball with radius $\epsilon$ around it. Consider the multivariate Gaussian search distribution. By definition,

$$\Pr_{x\sim\mathcal{N}(\mu,\Sigma)}[\|x - x^*\| \leqslant \epsilon] = \int_{x\in B(x^*,\epsilon)} \mathcal{N}(x|\mu, \Sigma)dx \qquad (1)$$

is the probability that a draw from the search distribution parametrised by $\mu$ and $\Sigma$ falls in the $\epsilon$-neighbourhood of the global optimum.

In current practice, the parameters $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ are maximum likelihood estimates (MLE) from $\tilde{N}$ selected search points of the population. Hence $\Sigma$ is a matrix valued random variable, i.e. a random matrix. It can be analysed with the existing tools of Random Matrix Theory (RMT) that were also useful in analyses of covariance estimation [27, 23].

Further,by the mean value theorem for multivariate definite integrals ([3],pp. 401), there exists a point $\tilde{x} \in B(x^*, \epsilon)$ s.t. eq.(1) equals:

$$= \text{Volume}(B(x^*, \epsilon))\mathcal{N}(\tilde{x}|\mu, \Sigma) \qquad (2)$$

$$= \text{Volume}(B(x^*, \epsilon)) \prod_{i=1}^{d} \mathcal{N}(U_i(\tilde{x} - \mu)|0, \lambda_i) \qquad (3)$$

where $U_i$ denotes the $i$-th eigenvector of $\Sigma$ and $\lambda_i$ is its associated eigenvalue. Among the things to notice from this, by computing the partial derivatives of (3) w.r.t. the eigenvalues $\lambda_i$ one can find that the optimal value of the $i-th$ eigenvalue of $\Sigma$ is the square length of the projection of $\tilde{x}-\mu$ onto the corresponding eigendirection, i.e. $\lambda_i^{opt} = \|U_i(\tilde{x} - \mu)\|^2$. When $\|U_i(\tilde{x} - \mu)\|^2 > \lambda_i$ then the probability (1) of drawing a point in the $\epsilon$-neighbourhood of $x^*$ can be increased by increasing $\lambda_i$. When $\|U_i(\tilde{x} - \mu)\|^2 < \lambda_i$ then (1) can be increased by decreasing $\lambda_i$. Hence the eigenvalues of $\Sigma$ play the role of learning rates in Gaussian EDA.

Now, it is known from RMT that in small sample conditions the smallest eigenvalue is severely underestimated while the largest eigenvalue is overestimated. An example is shown in Figure 1. The extent of this misestimation is well understood in RMT, and based on this new methods have been developed (including most recent results [13], see also [19]) that are able to remedy the problem effectively even when $\Sigma$ is singular, using an ensemble of random projections of the covariance estimate.

The recent RMT-based methods to covariance estimation in small sample conditions [19, 13] also have several advantages over other statistical methods such as sparsity constraints and various other regularisation approaches in the given context: First, they do not impose unjustified constraints; secondly, they were found to outperform the optimal Ledoix-Wolf estimator in terms of approximating the true covariance [19] and in data classification [13]; thirdly, they lend themselves to parallel implementation that fits well with the algorithmic structure of population based search.

# 3. APPROACH

The main goal of this paper is to develop a radically new approach to large scale stochastic optimisation in application to EDA. Building on recent results in other areas, our concept is to introduce an ensemble of random non-adaptive dimensionality reducing projections of the fittest high dimensional search points as a basis for developing a new and generic divide-and-conquer methodology rooted in the theory of Random Projections and exploiting recent advances of non-asymptotic Random Matrix Theory and related fields.

At a high level, the rationale is as follows:

1. Random matrices that satisfy the Johnson-Lindenstrauss Lemma (JLL) [7] are approximate isometries. Hence, with appropriate choice of the target dimension, important structure such as Euclidean distances and dot products are approximately preserved in the reduced space. This makes it possible to capture correlations between the $d$-dimensional search variables in the $k << d$-dimensional space.

2. In the reduced space the distribution becomes 'more Gaussian', in a sense made precise in [8]. Also, both parameter estimation and sampling become feasible and computationally affordable, so there is no need to overly restrict the parametric form of the search distribution.

3. There is a natural smoothing effect that emerges when appropriately combining the ensemble of estimates from several random subspaces [18, 19, 13]. This will ensure that the exploration ability of the search distribution can be maintained even with small population sizes.

Random projections have been used in approximation theory since the 1970s [17]. In computer science, information theory, signal processing and more recently in machine learning, random matrices provide a mechanism for dimensionality reduction while preserving the essential information in the data [26]. Compared with other methods in that context, they lead to (1) faster algorithms that are (2) simpler to analyse, (3) lend themselves to parallel implementation, and (4) exhibit robustness (see [18] for a recent review). We aim to port and exploit these characteristics to high dimensional optimisation.

## 3.1 New search operators for EDA

Let $R \in \mathbb{R}^{k\times d}$ a random matrix with entries drawn i.i.d. from a Gaussian $\mathcal{N}(0, \sigma^2)$. When $d$ is large, the rows of this matrix are almost orthogonal. So if we choose $\sigma^2 = 1/d$ then $R$ well approximates a projection matrix from $\mathbb{R}^d$ to $\mathbb{R}^k$ where $k$ may be chosen much lower than $d$.

Further let $x_0 \in \mathbb{R}^d$ a point. Denote by $\mathcal{S}_{x_0}^R$ the subspace defined by $R$ that passes through $x_0$. We define new search operators as follows:

`Project`: takes an $R \in \mathbb{R}^{k\times d}$, an $x_0 \in \mathbb{R}^d$, and a sample $\mathcal{P}^{fit} = (x_i \in \mathbb{R}^d)_{i=1:\tilde{N}}$, and orthogonally projects $\mathcal{P}^{fit}$

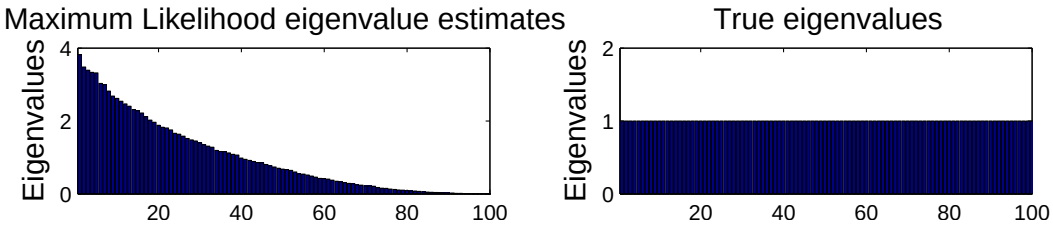


**Figure 1: Eigenvalue misestimation from $\tilde{N} = 100$ points in $d = 100$ dimensions.**

onto the subspace defined by $R$ that passes through $x_0$, i.e. returns $\mathcal{P}_R = (R^T R(x_i - x_0) + x_0)_{i=1:\tilde{N}}$.

`sEstimate`: takes a sample $\mathcal{P}_R$ that lives in a subspace and computes the ML parameter estimates $\hat{\theta}_R$ of its distribution $\mathcal{D}_R$ w.r.t. the restriction of the Lebesgue measure to the $k$-dimensional affine subspace defined by $R$, e.g. $\hat{\theta}_R = (\hat{\mu}_R, \hat{\Sigma}_R)$.

`sSample`: takes parameter estimates $\hat{\theta}_R$ obtained by `sEstimate` and returns a sample of $N$ $k$-dimensional points drawn i.i.d. from $\mathcal{D}_R$ with parameters $\hat{\theta}_R$.

`Combine`: takes populations from several $k$-dimensional subspaces $\mathcal{S}^{R_i}_{x_0}, i = 1, ..., M$ and returns a population that lives in the full search space $\mathbb{R}^d$.

Using these operators, the high level outline of our meta-algorithm is as follows.

1. Initialise population $\mathcal{P}$ by generating $N$ individuals uniformly randomly.
2. Let $\mathcal{P}^{fit}$ be the fittest $\tilde{N} < N$ individuals from $\mathcal{P}$.
3. For $i = 1, ..., M$ ($M \geqslant 1$) randomly oriented (affine) $k < d$-dimensional subspaces $\mathcal{S}^{x_0}_{R_i}$
   (a) Project $\mathcal{P}^{fit}$ onto $\mathcal{S}^{x_0}_{R_i}$
   (b) Produce $N$ new individuals on the subspace $\mathcal{S}^{x_0}_{R_i}$ using the sequence `sEstimate`; `sSample`.
4. Create the new population $\mathcal{P}$ using `Combine`.
5. If stopping criteria is met then Stop; else Goto 2.

We will instantiate this by taking the translation vector $x_0$ of the consecutive set of subspaces (in consecutive generations) to be the mean of $\mathcal{P}^{fit}$ in the previous generation. Further, in this work we instantiate the `Combine` operator as a scaled[1] average of the individuals produced on the individual subspaces (which may be done even without appeal to fitness evaluation within subspaces).

## 3.2 Algorithm

Denote by $\mathcal{P}^{fit} = [x_1, ..., x_{\tilde{N}}]$ the set of $\tilde{N}$ selected fit individuals, and let $N$ be the population size.
The following is an instantiation of the module for creating the new generation (steps 3-4 of the above).

[1]Orthogonal projection from $\mathbb{R}^d$ to $\mathbb{R}^k$ shortens the lengths of vectors by a factor of $\sqrt{k/d}$ and averaging $M$ i.i.d. points reduces their std by a factor of $\sqrt{M}$, hence a scaling factor of $\sqrt{(dM)/k}$ is needed to recover the original scale. This is the case when the entries of $R$ were drawn with variance $\sigma^2 = 1/d$ – equivalently, the scaling required otherwise is $\sqrt{M/(k\sigma^2)}$.

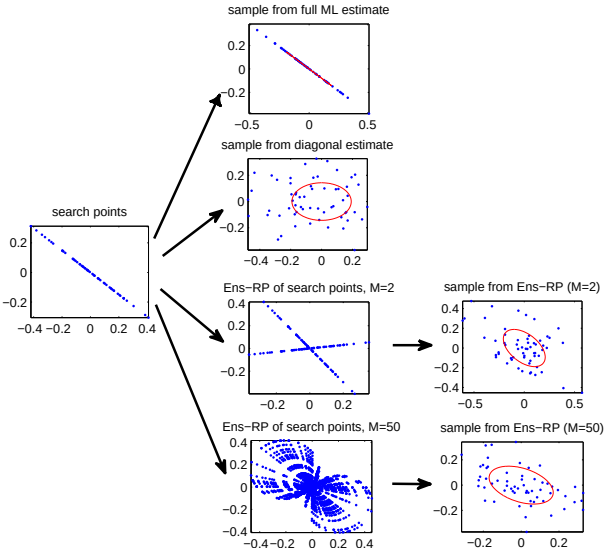


**Figure 2: Illustration of the use of random projections for sampling the new population.**

1. *Inputs*: $\mathcal{P}^{fit}, M, k$ // where $M \geqslant \lceil d/k \rceil$
2. Estimate $\mu := mean(\mathcal{P}^{fit})$
3. Generate $M$ independent random projection matrices $R_i, i = 1, ..., M$.
4. For i=1,...,M
   (a) Project the centred points into $k$-dimensions: $\mathbf{Y}^{R_i} := [R_i(x_n - \mu); n = 1, ..., \tilde{N}]$.
   (b) Estimate the $k \times k$ sample covariance $\Sigma^{R_i}$.
   (c) Sample $N$ new points $y^{R_i}_1, ..., y^{R_i}_N \sim_{i.i.d.} \mathcal{N}(0, \Sigma^{R_i})$.
5. Let the new population $\mathcal{P} := \sqrt{\frac{dM}{k}}[\frac{1}{M}\sum_{i=1}^{M} R_i^T y^{R_i}_1, ..., \frac{1}{M}\sum_{i=1}^{M} R_i^T y^{R_i}_N] + \mu$.
6. *Output*: $\mathcal{P}$

The working of this method is illustrated in Figure 2 – of course with the caveat that high dimensional geometry is hard to capture on a 2D figure – and should be read as follows. In large scale problems the number of fit points $\tilde{N}$ is always smaller than the dimension of the search space $d$, hence the fit individuals live in the $\tilde{N}$-dimensional subspace

of the search space determined by their span. The leftmost subplot illustrates a situation where some $\tilde{N}$ points live in a subspace (here 1D) of the overall space (here 2D). Hence, the maximum likelihood (ML) covariance estimate of the fit points is singular. Sampling points from a distribution with a singular covariance means that the next generation is confined in the same subspace. The middle upper subplot illustrates this. Now, if we impose a diagonality constraint, the diagonal covariance estimate is no longer singular in general. So the next generation is allowed in the full search space, although the directions in which it can spread is quite limited. This is seen in the middle second subplot. The remaining figures show what happens when we use a RP-ensemble. The first one shows a case where the number of random subspaces is the smallest that still spans the full search space, while the second one shows a case where a large number of random subspaces are used. In both cases, the fit points are projected onto each of the random subspaces, and a new generation is sampled within each subspace. The new individuals from these multiple worlds are then averaged, to give the ultimate new population shown on the rightmost plots. We see that the ML covariance estimate of this new population has a tendency to respect the orientation of the density of the parent population while it eliminates degeneracy.

## 3.3 Analysis of the algorithm to create the new generation

To understand the effect of the Algorithm in Sec. 3.2, we analyse it in the full search space by assembling a new search distribution in the original search space.

Fix the set of selected fit individuals $\mathcal{P}^{fit}$, and denote by $\Sigma$ the maximum likelihood estimate of their sample covariance. This covariance estimate is never computed explicitly throughout the algorithm, but it is useful for the theoretical analysis of this section.

Now, it is not too difficult to show that, by construction, the new population, $\mathcal{P}$, obtained by the Algorithm in Sec. 3.2, is distributed i.i.d. as $\mathcal{N}(\mu, \frac{d}{k}[\frac{1}{M}\sum_{i=1}^{M} R_i^T R_i \Sigma R_i^T R_i])$. However, while $\Sigma$ is singular due to $\tilde{N}$ being much smaller than $d$, the matrix $\frac{d}{k}[\frac{1}{M}\sum_{i=1}^{M} R_i^T R_i \Sigma R_i^T R_i]$ is positive definite a.s. for $M \geqslant \lceil d/k \rceil$. Furthermore, we can analyse this matrix as a sum of positive semi-definite random matrices that concentrates around its expectation, $d/k\mathrm{E}[R^T R \Sigma R^T R]$, which is also the limit of the sum when $M \to \infty$.

### 3.3.1 *Infinitely many random projections*

Recall that the random projections $R_i$ are drawn i.i.d. Therefore, by the law of large numbers, the ensemble may be thought as a finite approximation of the following expectation:

$$\frac{1}{M}\sum_{i=1}^{M} R_i^T R_i \Sigma R_i^T R_i \underset{M\to\infty}{\longrightarrow} \mathrm{E}_R[R^T R \Sigma R^T R] \tag{4}$$

and we can understand the effect of the RP-ensemble by computing this expectation.

Denote $\rho = \mathrm{rank}(\Sigma)$. By the rotation-invariance of the random Gaussian matrix $R$, and denoting by $\Sigma = U\Lambda U^T$ the SVD decomposition of $\Sigma$. it is easy to rewrite $\mathrm{E}[R^T R \Sigma R^T R] = U\mathrm{E}[R^T R \Lambda R^T R]U^T$ – hence we see that the overall effect is an operation on the eigenvalues of the traditional EDA's $\Sigma$, and it is enough to analyse $\mathrm{E}[R^T R \Lambda R^T R]$. This is:

$$\mathrm{E}_R[R^T R \Lambda R^T R] = \sum_{i=1}^{\rho} \lambda_i \begin{bmatrix} \mathrm{E}[(r_1^T r_i)^2] & \dots & \mathrm{E}[(r_1^T r_i)(r_i^T r_d)] \\ \dots & & \dots \\ \mathrm{E}[(r_d^T r_i)(r_i^T r_1)] & \dots & \mathrm{E}[(r_d^T r_i)^2] \end{bmatrix} \tag{5}$$

The diagonal elements have the form $\mathrm{E}[(r_j^T r_i)^2]$. There are two cases:

Case $j = i$:

$$\begin{aligned} \mathrm{E}[(r_i^T r_i)^2] &= \mathrm{E}[(\sum_{j=1}^{k} r_{ji}^2)^2] = \sum_{j=1}^{k}\sum_{j'=1}^{k} \mathrm{E}[r_{ji}^2 r_{j'i}^2] \qquad (6) \\ &= \sum_{j=1}^{k}\sum_{j'=1,j'\neq j}^{k} \mathrm{E}[r_{ji}^2]\mathrm{E}[r_{j'i}^2] + \sum_{j=1}^{k}\mathrm{E}[r_{ji}^4] \\ &= \sigma^4 + 3\sigma^4 = \sigma^4(k^2+2k) \qquad (7) \end{aligned}$$

Case $j \neq i$:

$$\begin{aligned} \mathrm{E}[(r_i^T r_j)^2] &= \mathrm{E}[(\sum_{\ell=1}^{k} r_{\ell i} r_{\ell j})^2] = \sum_{\ell=1}^{k}\sum_{\ell'=1}^{k} \mathrm{E}[r_{\ell i} r_{\ell j} r_{\ell' i} r_{\ell' j}] \\ &= \sum_{\ell=1}^{k}\sum_{\ell'=1,\ell'\neq\ell}^{k} \mathrm{E}[r_{\ell i}]\mathrm{E}[r_{\ell j}]\mathrm{E}[r_{\ell' i}]\mathrm{E}[r_{\ell' j}] + \dots \\ &+ \sum_{\ell=1}^{k} \mathrm{E}[r_{\ell i}^2 r_{\ell j}^2] = \sigma^4. \qquad (8) \end{aligned}$$

Finally, the off-diagonal elements have the form $\mathrm{E}[(r_j^T r_i)(r_i^T r_\ell)]$ with $j \neq \ell$, and these evaluate to zero:

$$\begin{aligned} \mathrm{E}[(r_j^T r_i)(r_i^T r_\ell)] &= \mathrm{E}[(\sum_{m=1}^{k} r_{mi} r_{mj})(\sum_{m'=1}^{k} r_{m'i} r_{m'\ell})] \qquad (9) \\ &= \sum_{m=1}^{k}\sum_{m'=1}^{k} \mathrm{E}[r_{mi}]\mathrm{E}[r_{mj}]\mathrm{E}[r_{m'i}]\mathrm{E}[r_{m'\ell}] = 0 \end{aligned}$$

by the independence of the entries of $R$.

Hence,

$$\mathrm{E}[R^T R \Lambda R^T R] = \sum_{i=1}^{\rho} \lambda_i D_i \tag{10}$$

where $D_i$ is a diagonal matrix having its $(i,i)$-th element equal to $\sigma^4(k^2+2k)$ and all other diagonal elements equal to $\sigma^4 k$. After some algebra, this may be further rewritten as:

$$\mathrm{E}[R^T R \Lambda R^T R] = \sigma^4 k\,(\mathrm{Trace}(\Lambda) I_d + (k+1)\Lambda) \tag{11}$$

where $I_d$ is the $d$-dimensional identity matrix.

To sum up, for the choice $\sigma^2 = 1/d$ we get that:

$$\mathrm{E}[R^T R \Lambda R^T R] = \frac{k}{d}\left(\frac{\mathrm{Trace}(\Lambda)}{d} I_d + \frac{k+1}{d}\Lambda\right) \tag{12}$$

so by implication, we obtained a regularised version of the sample covariance estimate:

$$\mathrm{E}[R^T R \Sigma R^T R] = \frac{k}{d}\left(\frac{\mathrm{Trace}(\Sigma)}{d} I_d + \frac{k+1}{d}\Sigma\right) \tag{13}$$

In consequence, in the limit of $M \to \infty$ our new population $\mathcal{P}$ returned by the Algorithm in Sec. 3.2 will be distributed i.i.d. as $\mathcal{N}\left(\mu, \frac{\mathrm{Trace}(\Sigma)}{d} I_d + \frac{k+1}{d}\Sigma\right)$. Of course, when $M$ is finite the covariance obtained will concentrate around its

expectation hence it will be close to the estimate computed above. This can be quantified precisely using matrix-valued tail bounds [23, 28, 2]. However, the finite $M$ implementation can be run in parallel on $M$ separate cores, and so the net effect is to get samples from the regularised covariance without ever computing the maximum likelihood estimate, and without the need to explicitly sample from a $d \times d$ covariance (which would be an $\mathcal{O}(d^3)$ opreation).

### 3.3.2 Finitely many random projections

Here we bound the deviation of the assembled covariance with finite $M$ from its expectation computed above. This is summarised in the following result.

THEOREM 1. *Let $\Sigma$ be a positive semi-definite matrix of size $d \times d$ and rank $\rho$, and $R_i, i = 1, ..., M$ independent random projection matrices, each having entries drawn iid from $\mathcal{N}(0, 1/d)$, and denote by $\|\cdot\| = \lambda_{\max}(\cdot)$ the spectral norm of its argument.*

$$Pr\left\{\|\frac{1}{M}\sum_{i=1}^{M} R_i^T R_i \Sigma R_i^T R_i - E[R^T R \Sigma R^T R]\| \geqslant \epsilon \|E[R^T R \Sigma R^T R]\|\right\}$$
$$\leqslant d \exp\left\{-\epsilon^2 M^{\frac{1}{3}} \frac{\|E[R^T R \Sigma R^T R]\|}{4\tilde{K}}\right\} + 2M \exp\left\{-\frac{M^{\frac{1}{3}}}{2}\right\}$$

*where* $\tilde{K} = \|\Sigma\| \left(\frac{1}{M^{1/6}}(1 + \sqrt{\frac{k}{d}}) + \frac{1}{\sqrt{d}}\right)^2 \left(\frac{1}{M^{1/6}}(\sqrt{\frac{\rho}{d}} + \sqrt{\frac{k}{d}}) + \frac{1}{\sqrt{d}}\right)^2$ *is bounded w.r.t. $M$.*

PROOF. We will use the following result from random matrix theory about sums of independent random matrices:

THEOREM 2. *[28, 2] Let $X_i$ $d$-dimensional independent random positive-semi-definite matrices satisfying $\|X_i\| \leqslant 1$ a.s. Let $S_M = \sum_{i=1}^{M} X_i$, and $\Omega = \sum_{i=1}^{M} \|E[X_i]\|$. Then $\forall \epsilon \in (0, 1)$ we have:*

$$Pr(\|S_M - E[S_M]\| \geq \epsilon\Omega) \leqslant d \exp(-\epsilon^2\Omega/4) \qquad (14)$$

Observe, $\|R_i^T R_i \Sigma R_i^T R_i\|$ is not bounded a.s., so we cannot apply this result directly. However, this condition can be satisfied by exploiting concentration, as the following.

First, we note that this random variable has the same distribution as $\|R_i^T R_i \Lambda R_i^T R_i\|$ where $\Lambda$ is the diagonal matrix of eigenvalues of $\Sigma$. Here we used the rotation invariance of the Gaussian. Now, denote by $\underline{\Lambda}$ the $\rho \times \rho$ sub-matrix of $\Lambda$ that contains the non-zero diagonals, and by $\underline{R_i}$ the $k \times \rho$ sub-matrix of $R_i$ that are not wiped out by the zeros of $\Lambda$. Then we can write $\|R_i^T R_i \Lambda R_i^T R_i\| = \|R_i^T \underline{R_i} \Lambda \underline{R_i^T} R_i\|$, and we can bound this with high probability (w.r.t. the random draws of $R_i$):

$$\|R_i^T \underline{R_i} \Lambda \underline{R_i^T} R_i\| \leqslant \|\Sigma\| \cdot \|R_i^T R_i\| \cdot \|\underline{R_i^T R_i}\| \leqslant ...$$
$$\|\Sigma\| \cdot (1 + \sqrt{k/d} + \frac{\eta}{\sqrt{d}})^2 (\sqrt{\rho/d} + \sqrt{k/d} + \frac{\eta}{\sqrt{d}})^2 =: K(\eta)$$

with probability $1 - 2\exp(\eta^2/2)$, for any $\eta > 0$. Here we used twice the bound on the largest singular value of a Gaussian matrix with i.i.d. entries [27] (Corollary 5.35), applied to $R_i$ and $\underline{R_i}$ that have entries drawn from $\mathcal{N}(0, 1/d)$.

Now, let $X_i(\eta) := R_i^T R_i \Sigma R_i^T R_i / K(\eta)$. Then we have:

$$\|X_i(\eta)\| \leqslant 1 \text{ w.p. } 1 - 2\exp(-\eta^2/2) \qquad (15)$$

Hence, by union bound, $\|X_i(\eta)\| \leqslant 1$ w.p. $1 - 2M\exp(-\eta^2/2)$ uniformly for all $i = 1, ..., M$. This holds for any choice of $\eta > 0$, and we will eventually choose $\eta$ to kill the factor of $M$ as well as to (approximately) tighten the bound on the deviation bound we will derive.

We now apply the Theorem 2 conditionally on the event that $\|X_i(\eta)\| \leqslant 1, \forall i = 1, ..., M$, and use the bound on the probability that this condition fails. It is easy to see that $\Omega(\eta) = \frac{M}{K(\eta)} \|\mathrm{E}[R^T R \Sigma R^T R]\|$, and $\mathrm{E}[S_M(\eta)] = M \cdot \mathrm{E}[X_i(\eta)] = \frac{M}{K(\eta)} \mathrm{E}[R^T R \Sigma R^T R]$, and so we get:

$$\Pr\left\{\|\tfrac{1}{K(\eta)} \textstyle\sum_{i=1}^{M} R_i^T R_i \Sigma R_i^T R_i - \tfrac{M}{K(\eta)} \mathrm{E}[R^T R \Sigma R^T R]\| \right.$$
$$\left. \geqslant \epsilon \tfrac{M}{K(\eta)} \|\mathrm{E}[R^T R \Sigma R^T R]\|\right\} =$$
$$\Pr\left\{\|\tfrac{1}{M} \textstyle\sum_{i=1}^{M} R_i^T R_i \Sigma R_i^T R_i - \mathrm{E}[R^T R \Sigma R^T R]\| \geqslant \epsilon \|\mathrm{E}[R^T R \Sigma R^T R]\|\right\}$$
$$\leqslant d \exp\left\{-\epsilon^2 \tfrac{M}{4K(\eta)} \|\mathrm{E}[R^T R \Sigma R^T R]\|\right\} + 2M \exp\left\{-\tfrac{\eta^2}{2}\right\}$$

Finally, we choose $\eta = M^{1/6}$ and denote $\tilde{K} := K(M^{1/6})$, which yields the statement of Theorem 1. □

This analysis shows that we can use a finite number of random subspaces since we have control over the spectral distance between the resulting finite average of the $d$-dimensional rank-$k$ covariances and the infinite limit of this sum. Hence, we may expect a similar behaviour from a finite ensemble, which is pleasing. The practical implication, as we already mentioned earlier, is that a parallel implementation can be realised where the estimation and sampling within each subspace is run on a separate core.

In closing, we should mention that, although we used the truncation method here in this section, a more direct route might exists if the a.s. boundedness condition could be relaxed in the original Theorem 2.

# 4. EXPERIMENTS

## 4.1 Benchmark test functions

To test the potential of our idea and the ability of our algorithm to find a near-optimal solution in large-scale settings, we tested it on the suite of benchmark functions from the CEC'2010 competition on Large-Scale Global Optimisation [25]. This test suite consists of twenty 1000-dimensional functions of four different types:

1. Separable functions
2. Partially-separable functions, in which a small number of variables are dependent while all the remaining ones are independent (m=50)
3. Partially-separable functions that consist of multiple independent sub-components, each of which is $m$-nonseparable (m=50) – this category includes two subtypes: $d/(2m)$-group $m$-nonseparable, and $d/m$-group $m$-nonseparable functions
4. Fully nonseparable functions

See [25] for more details on these functions.

Of these, 12 functions are multimodal and 8 are unimodal. We will focus on the multimodal functions, as this is the category where we expect our methodology to be most beneficial – especially nonseparable multimodal functions.

We use a simple averaging combination of RP-EDAs as in the algorithm described in Section 3.2. We take the random

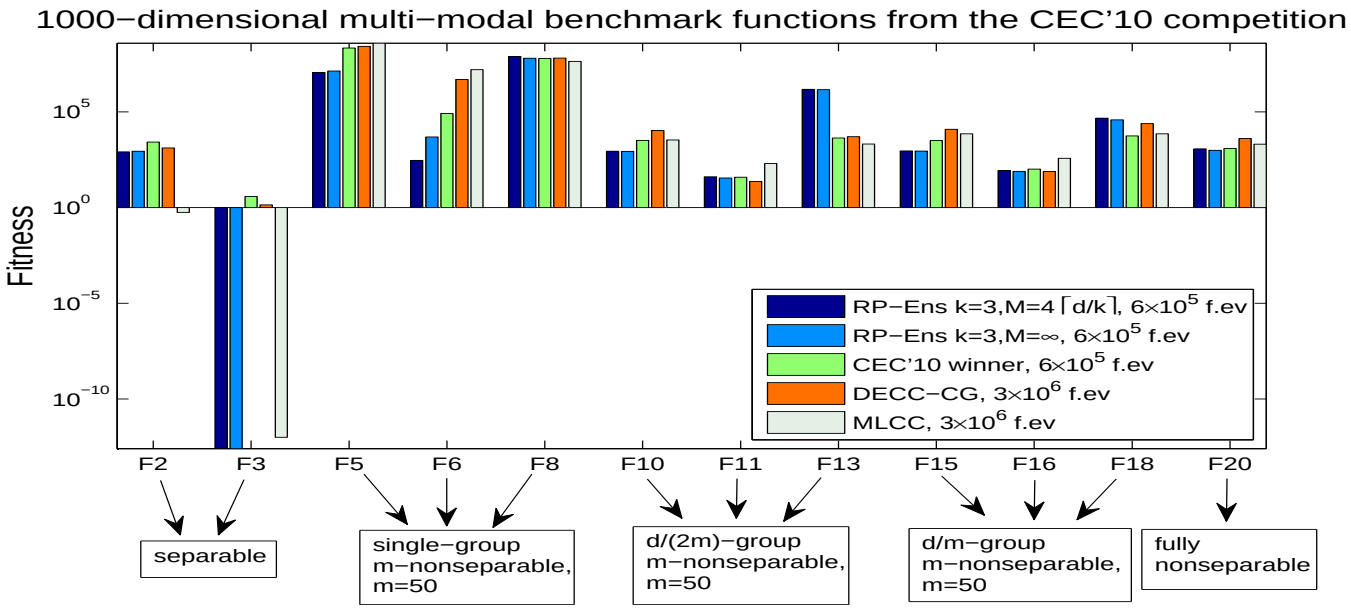


**Figure 3: Comparison of a variant of our new RP-Ensemble EDA algorithm with the CEC'10 large scale optimisation competition winner [22] on 12 multi-modal functions, after $6 \times 10^5$ function evaluations. Results of other state of the art co-evolutionary based MLCC and DECC-CG are also shown for reference; these are quoted from [22] and use $3 \times 10^6$ function evaluations.**

subspace dimension to be $k = 3$ and the number of subspaces is set to $M = 4 \cdot \lceil d/k \rceil = 1334$. The population size was set to $N = 300$ and $\tilde{N} = 75$, and we used truncation selection with elitism. We have set the number of function evaluations to $6 \times 10^5$.

Figure 3 summarises our results obtained on these 12 multimodal functions in terms of the average of the best fitness from 10 independent runs. Further, we include results from 10 independent runs of the limiting version $k = 3, M = \infty$, which we implemented using the analytic expression computed in Sec. 3.3.1, eq. (13) (with sampling done in the full $d$-dimensional space). We compare these results with the CEC'2010 winner algorithm [22] – a fairly sophisticated memetic algorithm based on local search chains – and two other state of the art co-evolutionary methods referenced on the competition's page, namely DECC-CG [30] and MLCC [31]. The results of the latter three methods are quoted from [22], these represent the average fitness over 25 independent runs. For DECC-CG and MLCC we only had access to results produced with $3 \times 10^6$ function evaluations, that is considerably more than our budget had been set to. Nevertheless, we see that our results still compare well to these too.

Thus, we see that our simple RP-Ensemble based EDA algorithm is highly competitive with the best state of the art methods for large scale optimisation – and even slightly outperforms the CEC'2010 competition winner on 8 out of these 12 multimodal benchmark functions. Furthermore, the version with largish but finite $M$ is nearly indistinguishable from that with infinite $M$.

In order to gain more insights about the behavour of our RP-Ensemble based EDA algorithm it is useful inspect the evolution of the best fitness accross generations. In Figure 4 we plotted these comparatively with UMDAc. We included the Sphere function as a representative of easy functions, as well as one representative from the three most non-separable categories from the CEC'2010 benchmarks. We compare UMDAc, and four variations of our algorithm: $k = 3, M = 1000$, $k = 3, M = \lceil d/k \rceil$, $k = 15, M = \lceil d/k \rceil$ – that is, the latter two versions use the smallest number of random subspaces that still span the full search space a.s. – and $k = 3, M = \infty$. As before, we use population sizes of $N = 300, \tilde{N} = 75$. A study of how best to set these parameters remains for future research.

From Figure 4 we can see a clear tendency of our RP-Ensemble-EDA algorithms to escape early convergence that is typical of UMDAc, and reach at better fitness values. The premature convergence of UMDAc is not surprising, and was indeed noted previously in the literature – the imposition of a diagonal covariance limits the exploration ability of UMDAc. In turn, our RP-Ensemble-EDA algorithms display a better ability to explore the search space for all variations of $k$ and $M$ settings tested here. It is particularly pleasing that the versions with finite number of random subspaces are also performing well and are indeed not far from the performance we observe for infinite number of random subspaces.

## 4.2 Application to mislabelled gene array classification

Optimisation underpins many areas of science and engineering. Efficient specialised algorithms only exist for certain types of problems. In data modelling and data mining, for example, data sets become increasingly complex, and new tasks need new models to be optimised using the data. Large scale black-box optimisation can be enabling in such domains.

Here we consider the problem of classification with labelling errors for gene microarrays. Previous studies re-

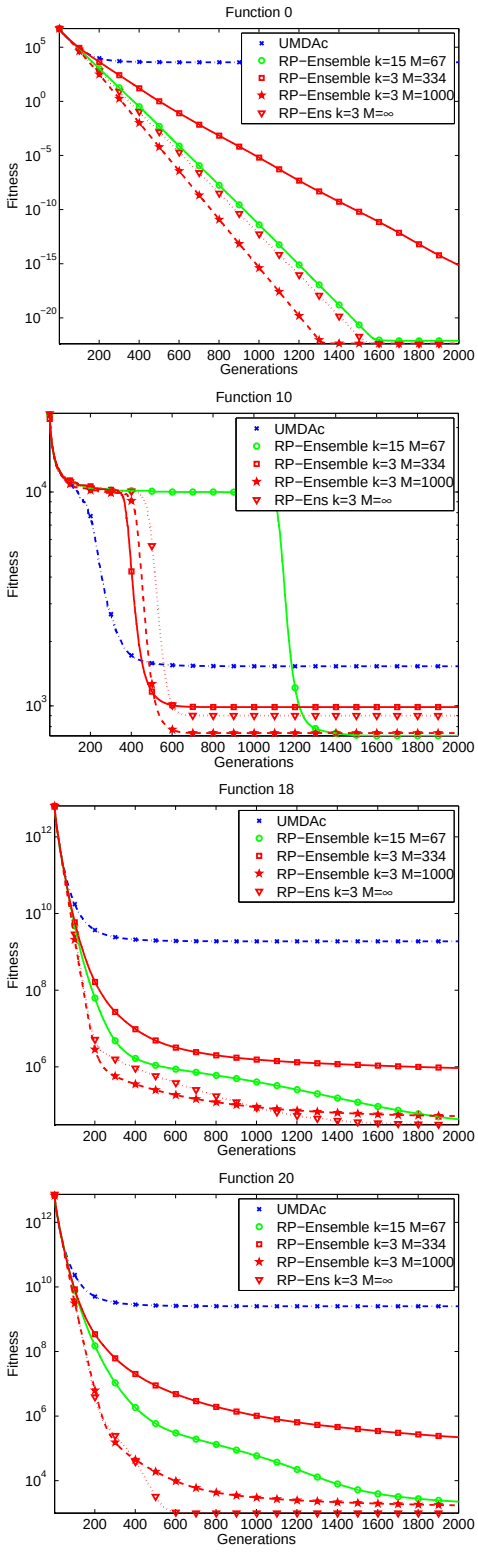


**Figure 4: Behaviour of our new RP-Ensemble EDA methods versus UMDAc on 1000-dimensional test functions. We see the RP-Ensemble based EDAs with finite number of random subspaces are close to the performance observed for infinite number of random subspaces.**

| RP-Ens k=3 | UMDAc | Method from [5] |
|---|---|---|
| 14.63±2.68 | 20.02 ±2.62 | 18.75±1.1 |

**Table 1: Classification of mislabelled colon cancer gene arrays. The average and standard error is given from 10 repeated runs on a random split into 50 training points and 12 test points.**

ported that labelling errors are not uncommon in microarray data sets [5]. In such cases the training set may become misleading, and the ability of classifiers to make reliable inferences from the data is compromised. Yet, very few methods are currently available in the bioinformatics literature to deal with this problem. A model based approach was recently developed to counter the effects of labelling errors [5]. Although the approach builds on a logistic regression with Lasso penalty – which is a convex objective – the inclusion of a model of label noise transforms this into a non-convex multimodal objective.

We apply our RP-Ens-EDA algorithm to colon cancer classification in the presence of label noise. The task aims to distinguish between normal tissue and tumour. The data is 2000-dimensional and contains 62 points (40 tumour, 22 normal). Our search space, i.e. the parameter space is, thus, 2000-dimensional. According to [1] there is biological evidence that the samples T2, T30, T33, T36, T37, N8, N12, N34, N36 may be mislabelled.

We define the fitness function to be the likelihood of the label-robust classifier of [5] (section 2.1). This function is both non-convex, and non-smooth at the origin.

We split the data into 80% training points and 20% test points randomly, we train our classifier on the training points by optimising the model likelihood using RP-Ens-EDA (or a competing optimiser) with population size $N = 100, \tilde{N} = 10$, using $10^5$ function evaluations (re-estimating the label flipping probabilities after each 2000 generations). Table 1 presents the mean and standard errors of misclassification rates from 10 independent trials. Although these are preliminary results, and a larger number of trials will need to be run in order to determine if the differences are statistically significant, we nevertheless can see that RP-Ens is a promising new method with better performance on average. The last column gives for reference the result using a specialised optimiser we developed in [5], tested here under the same conditions as our RP-ENS based EDA and UMDAc.

## 5. OUTLOOK AND FUTURE WORK

We presented a new framework for designing and developing EDA-type methods for large scale optimisation. Our approach is to employ multiple random projections of the fit individuals, and carry out the estimation and sampling operations in low dimensional spaces, where these are both efficient and reliable – as opposed to working in the original high dimensional space. We derived some theoretical analysis that show the effect of our divide-and-conquer methodology when re-assembled and understood in the full high-dimensional search space. Finally, we presented empirical results using a very simple instantiation of our proposed framework, which demonstrated its effectiveness. On a battery of 12 multimodal test functions from the large scale CEC'10 competition we obtained results that are competitive to the best state of the art. We also presented a real-

world application to the problem of high dimensional gene array classification in the presence of labelling errors. We believe these results may give a new perspective to research on EDA-type model building optimisation algorithms, and future work is aimed at better understanding and exploiting its potential.

## 6. REFERENCES


[1] U. Alon, et. al. Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. In Proc. of the National Academy of Sciences of the United States of America, 96(12), 1999, pp. 6745-6750.

[2] R. Ahlswede, A. Winter, Strong converse for identication via quantum channels, IEEE Trans. Information Theory 48, 2002, pp. 568-579.

[3] T.M. Apostol. Mathematical Analysis. T. M. Apostol Mathematical Analysis, Addison-Wesley, 1957.

[4] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is nearest neighbor meaningful? in: Proc. Int. Conf. Database Theory, 1999, pp. 217-235.

[5] J. Bootkrajang and A. Kabán. Classification of Mislabelled Microarrays using Robust Sparse Logistic Regression. Bioinformatics. (accepted)

[6] P. Bosman. On empirical memory design, faster selection of bayesian factorizations and parameter-free Gaussian EDAs. Proc. of GECCO-2009. ACM, 2009, pp. 389-396.

[7] S. Dasgupta. Learning Mixtures of Gaussians. In Annual Symposium on Foundations of Computer Science (FOCS), volume 40, pages 634-644, 1999.

[8] P. Diaconis and D. Freedman. Asymptotics of graphical projection pursuit. The Annals of Statistics, 12(3):793-815, 1984.

[9] W. Dong and X. Yao. Unified eigen analysis on multivariate Gaussian based estimation of distribution algorithms. Information Sciences, 178(15):3000-3023, August 2008.

[10] C. Echegoyen, Q. Zhang, A. Mendiburu, R. Santana, J.A. Lozano. On the limits of effectiveness in estimation of distribution algorithms. Proc. IEEE CEC 2011: 1573-1580.

[11] W. Dong, T. Chen, P. Tino, X. Yao. Scaling Up Estimation of Distribution Algorithms For Continuous Optimization. CoRR abs/1111.2221, 2011.

[12] R.J. Durrant, and A. Kabán. When is 'nearest neighbour' meaningful: A converse theorem and implications Journal of Complexity, Volume 25, Issue 4, Pages 385-397.

[13] R.J. Durrant and A. Kabán. Random Projections as Regularizers: Learning a Linear Discriminant Ensemble from Fewer Observations than Dimensions. Technical Report CSR-12-01, U. Birmingham, 2012.

[14] Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. J. Roy. Statist. Soc. Ser. B 67, 427-444.

[15] T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. (2nd Edition), Springer-Verlag, 2008.

[16] P. Larrañaga and J. A. Lozano. Estimation of distribution algorithms: A new tool for evolutionary computation. Kluwer, 2002.

[17] Lorentz, G.G., von Golitschek, M., Makovoz, Yu.: Constructive Approximation: Advanced Problems, vol. 304. Springer, Berlin, 1996.

[18] M.W. Mahoney. Randomized Algorithms for Matrices and Data. Foundations and Trends in Machine Learning: Vol. 3: No 2, 2011, pp 123-224.

[19] T.L. Marzetta, G.H. Tucci, and S.H. Simon. A Random Matrix Theoretic Approach to Handling Singular Covariance Estimates. IEEE Trans. Information Theory, 57(9), 2011.

[20] H. Mühlenbein, T. Mahnig. Convergence Theory and Application of the Factorized Distribution Algorithm, Journal of Computing and Information Technology, 7, pp. 19-32, 1999.

[21] P. Larrañga and J. A. Lozano. Estimation of distribution algorithms: A new tool for evolutionary computation. Kluwer Academic Publishers, Boston, 2002.

[22] D. Molina, M. Lozano, and F. Herrera. MA-SW-Chains: Memetic Algorithm Based on Local Search Chains for Large Scale Continuous Global Optimization. IEEE WCCI'10, 2010.

[23] N. Srivastava, R. Vershynin, Covariance estimation for distributions with 2+epsilon moments, Annals of Probability, to appear.

[24] R. Ros, N. Hansen. A simple modification in cma-es achieving linear time and space complexity. Proc. PPSN, 2008, pp. 296-305.

[25] K. Tang, Xiaodong Li, P. N. Suganthan, Z. Yang and T. Weise. Benchmark Functions for the CEC'2010 Special Session and Competition on Large Scale Global Optimization. Technical Report, Nature Inspired Computation and Applications Laboratory, USTC, China, http://nical.ustc.edu.cn/cec10ss.php, 2009.

[26] S. Vempala. The random projection method. DIMACS Series of Discrete Mathematics and Theoretical Computer Science., vol. 65. AMS, 2004.

[27] R. Vershynin, Introduction to the non-asymptotic analysis of random matrices. Chapter 5 of the book Compressed Sensing, Theory and Applications, ed. Y. Eldar and G. Kutyniok. Cambridge University Press, 2012. pp. 210–268. [arXiv:1011.3027, Aug 2010]

[28] R. Vershynin. A note on sums of independent random matrices after Ahlswede-Winter. http://www-personal.umich.edu/řomanv/teaching/reading-group/ahlswede-winter.pdf

[29] Y. Wang and B. Li. A restart univariate estimation of distribution algorithm: sampling under mixed Gaussian and Lt'evy probability distribution. In Proceedings of the 2008 IEEE CEC, 2008, pp. 3917-3924.

[30] Z. Yang, K. Tang, and X. Yao. Large scale evolutionary optimization using cooperative coevolution. Information Sciences 178, 2008, pp. 2985-2999.

[31] Z. Yang, K. Tang, and X. Yao. Multilevel cooperative coevolution for large scale optimization. IEEE CEC 2008, pp. 1663-1670.