Towards Large Scale Continuous EDA: A Random Matrix Theory Perspective

A. Kabán, J. Bootkrajang, R.J. Durrant
School of Computer Science
The University of Birmingham
Birmingham B15 2TT, UK
http://www.cs.bham.ac.uk/~axk

Outline

- Introduction - EDA & the challenges of high dimensional search spaces
- A new approach from Random Matrix Theory
  - New search operators for EDA
  - Algorithm
  - Analysis
- Experiments
- Summary & future work
Introduction (1)

• Large scale continuous global optimisation is key to many areas in engineering, science & industry; Scaling up Evolutionary Algorithms (EA) is a major challenge of the field.
Introduction (1)

• Large scale continuous global optimisation is key to many areas in engineering, science & industry; Scaling up Evolutionary Algorithms (EA) is a major challenge of the field.

• Estimation of Distribution Algorithms (EDA) - a major branch of EA with unique advantages in principle - build a model of fittest individuals & sample the new generation from it.
Introduction (1)

- Large scale continuous global optimisation is key to many areas in engineering, science & industry; Scaling up Evolutionary Algorithms (EA) is a major challenge of the field.

- Estimation of Distribution Algorithms (EDA) - a major branch of EA with unique advantages in principle - build a model of fittest individuals & sample the new generation from it.

- In high dimensions, model building is extremely difficult - EDAs decline quickly beyond 50-100 dimensions.
• Large scale continuous global optimisation is key to many areas in engineering, science & industry; Scaling up Evolutionary Algorithms (EA) is a major challenge of the field.

• Estimation of Distribution Algorithms (EDA) - a major branch of EA with unique advantages in principle - build a model of fittest individuals & sample the new generation from it.

• In high dimensions, model building is extremely difficult - EDAs decline quickly beyond 50-100 dimensions.

• Current practice of EDA employs independence models or limited dependence in exchange for feasibility, e.g:
- Univariante mixed Gaussian & Lévy search distribution (Wang & Li, CEC’08) - cannot tackle non-separable problems (Mühlenbein et al, ’99; Echegoyen et al, CEC’11)
- Univariate mixed Gaussian & Lévy search distribution (Wang & Li, CEC’08) - cannot tackle non-separable problems (Mühlenbein et al, ’99; Echegoyen et al, CEC’11)
- sep-CMA-ES (Ros & Hansen, PPSN’08) & related CMA methods - applicable to rotated separable problems, which is just a subset of generic nonseparable problems
- Univariate mixed Gaussian & Lévy search distribution (Wang & Li, CEC’08) - cannot tackle non-separable problems (Mühlenbein et al, ’99; Echegoyen et al, CEC’11)
- sep-CMA-ES (Ros & Hansen, PPSN’08) & related CMA methods - applicable to rotated separable problems, which is just a subset of generic nonseparable problems
- Bayesian factorisations (Bosman, GECCO’09)
- Univariante mixed Gaussian & Lévy search distribution (Wang & Li, CEC’08) - cannot tackle non-separable problems (Mühlenbein et al, ’99; Echegoyen et al, CEC’11)
- sep-CMA-ES (Ros & Hansen, PPSN’08) & related CMA methods - applicable to rotated separable problems, which is just a subset of generic nonseparable problems
- Bayesian factorisations (Bosman, GECCO’09)
- Pre-defined block-diagonal covariance, thresholding (Dong et al, IEEE TEC 2013)
- Univariate mixed Gaussian & Lévy search distribution (Wang & Li, CEC’08) - cannot tackle non-separable problems (Mühlenbein et al, ’99; Echegoyen et al, CEC’11)
- sep-CMA-ES (Ros & Hansen, PPSN’08) & related CMA methods - applicable to rotated separable problems, which is just a subset of generic nonseparable problems
- Bayesian factorisations (Bosman, GECCO’09)
- Pre-defined block-diagonal covariance, thresholding (Dong et al, IEEE TEC 2013)

- In large scale optimisation, the state of the art best performers are EC with cooperative coevolution & multi-level coevolution (Yang, Tang & Yao, 2008), and hybrid methods with local searches (e.g. Molina, Lozano, & Herrera, 2010).
Introduction (2)

Our aim is to pin down some of the fundamental roots of the problem, & make a start on developing a new and generic framework to yield effective EDA-type global optimisers for large scale continuous problems.
**Introduction (2)**

Our **aim** is to pin down some of the fundamental roots of the problem, & make a start on developing a new and generic framework to yield effective EDA-type global optimisers for large scale continuous problems.

Our **idea** is to introduce an ensemble of *random projections* of the set of fittest search points to low dimensions as a basis for a new divide-and-conquer methodology.
**Introduction (2)**

**Our aim** is to pin down some of the fundamental roots of the problem, & make a start on developing a new and generic framework to yield effective EDA-type global optimisers for large scale continuous problems.

**Our idea** is to introduce an ensemble of *random projections* of the set of fittest search points to low dimensions as a basis for a new divide-and-conquer methodology.

**Our methodology** is rooted in the theory of random projections developed in theoretical computer science, and will exploit recent advances of non-asymptotic random matrix theory.
**Introduction (2)**

Our **aim** is to pin down some of the fundamental roots of the problem, & make a start on developing a new and generic framework to yield effective EDA-type global optimisers for large scale continuous problems.

Our **idea** is to introduce an ensemble of *random projections* of the set of fittest search points to low dimensions as a basis for a new divide-and-conquer methodology.

Our **methodology** is rooted in the theory of random projections developed in theoretical computer science, and will exploit recent advances of non-asymptotic random matrix theory.

Our **results** are comparable (and surpass in some cases) with state-of-the-art best performers in large scale optimisation.
The Challenges

Why EDA fails in high dimensions?
The Challenges

Why EDA fails in high dimensions?

Warning: High dimensional spaces defeat our intuition rooted in low dimensional experiences.
The Challenges

Why EDA fails in high dimensions?

(1) **Estimation problems**: If sample size is insufficient, the eigenvalues of the covariance are misestimated (Vershynin ’10) and bogus correlations appear in the maximum likelihood covariance estimates.
The Challenges

Why EDA fails in high dimensions?

(1) **Estimation problems**: If sample size is insufficient, the eigenvalues of the covariance are misestimated (Vershynin ’10) 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 (Beyer et al, ’99; Durrant & Kaban, ’09; Hall et al, ’05).
The Challenges

Why EDA fails in high dimensions?

(1) **Estimation problems**: If sample size is insufficient, the eigenvalues of the covariance are misestimated (Vershynin ’10) 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 (Beyer et al, '99; Durrant & Kaban, ’09; Hall et al, ’05).

(3) **Computational problems**: The computation cost of sampling from a full $d$-dimensional Gaussian distribution is $O(d^3)$ - becomes prohibitive when $d$ is very large.
Probability of reaching the 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. Considering the multivariate Gaussian search distribution, the probability that a draw from the search distribution parametrised by \( \mu \) and \( \Sigma \) falls in the \( \epsilon \)-neighbourhood of the global optimum is:

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

(1)

where in current practice, \( \mu \in \mathbb{R}^d \) and \( \Sigma \in \mathbb{R}^{d \times d} \) are ML estimates from \( \tilde{N} \) selected points. (\( \Sigma \) is a matrix valued random variable)
Eigenvalues act as learning rates (1)

By the mean value theorem for multivariate definite integrals \( \exists \tilde{x} \in B(x^*, \epsilon) \) s.t.

\[
\Pr_{x \sim \mathcal{N}(\mu, \Sigma)}[\|x - x^*\| \leq \epsilon] = \text{Volume}(B(x^*, \epsilon))\mathcal{N}(\tilde{x}| \mu, \Sigma) \tag{2}
\]

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

where \( U_i \) are eigenvectors of \( \Sigma \) and \( \lambda_i \) its eigenvalues.

Note: The partial derivatives of (3) w.r.t. \( \lambda_i \) tell us how this probability depends on \( \lambda_i \):

- if \( \lambda_i < \|U_i(\tilde{x} - \mu)\|^2 \) then it is an increasing function of \( \lambda_i \)
- if \( \lambda_i > \|U_i(\tilde{x} - \mu)\|^2 \) then it is a decreasing function of \( \lambda_i \)
Eigenvalues act as learning rates (2)

The problem: In small sample conditions the smallest eigenvalue is \textit{systematically} and \textit{severely} underestimated while the largest eigenvalue is \textit{systematically} overestimated! – this is well known in Random Matric Theory (RMT).

Illustration: Sample 100 points from a Gaussian with Identity covariance in 100 dimensions. Compute the eigenvalues of the sample covariance.
A New Approach - The high level rationale (1)

Exploit results on Random Projections from theoretical computer science.

Work in random subspaces of the search space.

Combine search points from several random subspaces.
What is Random Projection?

- Take a wide, flat matrix $R \in M_{k \times d}$ by picking the entries from the univariate Gaussian $\mathcal{N}(0, 1/d)$.

- Orthonormalise the rows of $R$ — in fact this is not needed, since high dimensional random vectors with i.i.d. entries are, with high probability, orthogonal and have similar lengths.

- Linearly transform each point $v \in \mathbb{R}^d$, by pre-multiplying it with $R$: $v \mapsto Rv \in \mathbb{R}^k$. 
Why Random Projection?

- Provided \( k \in \mathcal{O}(\log N) \), it preserves distances w.h.p. [Johnson-Lindenstrauss lemma]
Why Random Projection?

• Provided $k \in \mathcal{O}(\log N)$, it preserves distances w.h.p. [Johnson-Lindenstrauss lemma]

• Tractable to analysis.
Why Random Projection?

- Provided $k \in \mathcal{O}(\log N)$, it preserves distances w.h.p. [Johnson-Lindenstrauss lemma]

- Tractable to analysis.

- Combination of RP-ed covariances ‘repairs’ eigenvalue mis-estimation. [Marzetta, RMT]
Why Random Projection?

- Provided $k \in \mathcal{O}(\log N)$, it preserves distances w.h.p. [Johnson-Lindenstrauss lemma]

- Tractable to analysis.

- Combination of RP-ed covariances ‘repairs’ eigenvalue mis-estimation. [Marzetta, RMT]

- Computationally cheap & easy to implement.
Why Random Projection?

- Provided $k \in O(\log N)$, it preserves distances w.h.p. [Johnson-Lindenstrauss lemma]

- Tractable to analysis.

- Combination of RP-ed covariances ‘repairs’ eigenvalue mis-estimation. [Marzetta, RMT]

- Computationally cheap & easy to implement.

- Lends itself to parallel implementation.
Previous successes of RP in machine learning

- Dimensionality reduction [Bingham & Mannila ’01]

- Classification. e.g. [Fradkin et al ’03], [Goel et al ’05], [Schclar et al ’09]; [Arriaga & Vempala ’99], [Calderbank et al ’09], [Dur- rant & Kaban ’10], [Durrant & Kaban ICML ’13]

- Regression. e.g. [Maillard & Munos ’09], [Boutsidis ’09]; Re- inforcement Learning [Ghavamzadeh et al.’10]

- Clustering and Density estimation. e.g. [Indyk ’98], [Das- gupta ’99], [Kalai ’12], [Fern et al ’03]

- Low-rank matrix approximations [Halko & Tropp’11]
The Johnson-Lindenstrauss Lemma

The JLL is the following rather surprising fact:

**Theorem** [Johnson and Lindenstrauss, 1984] Let $\epsilon \in (0, 1)$. Let $N, k \in \mathbb{N}$ such that $k \geq C\epsilon^{-2}\log N$, for a large enough absolute constant $C$. Let $V \subseteq \mathbb{R}^d$ be a set of $N$ points. Then there exists a linear mapping $R: \mathbb{R}^d \rightarrow \mathbb{R}^k$, such that for all $u, v \in V$:

$$(1 - \epsilon)\|u - v\|_{\ell_2}^2 \leq \|Ru - Rv\|_{\ell_2^k}^2 \leq (1 + \epsilon)\|u - v\|_{\ell_2}^2$$

- With high probability random projection satisfies JLL [Dasgupta & Gupta '02] (proof by Chernoff bounding).

- The bound on $k$ is essentially tight: $\forall N, \exists V$ s.t. $k \in \Omega(\epsilon^{-2}\log N/\log \epsilon^{-1})$ is required [Alon '03].
A New Approach - The high level rationale (2)

1. Random matrices that satisfy the Johnson-Lindenstrauss Lemma preserve pairwise Euclidean distances and dot products. We can use such random matrices to compress the search space, and still capture some of the correlation structure.
A New Approach - The high level rationale (2)

1. Random matrices that satisfy the Johnson-Lindenstrauss Lemma preserve pairwise Euclidean distances and dot products. We can use such random matrices to compress the search space, and still capture some of the correlation structure.

2. In the reduced space both parameter estimation and sampling become feasible and computationally affordable - no need to overly restrict the form of covariance in the search distribution.
A New Approach - The high level rationale (2)

1. Random matrices that satisfy the Johnson-Lindenstrauss Lemma preserve pairwise Euclidean distances and dot products. We can use such random matrices to compress the search space, and still capture some of the correlation structure.

2. In the reduced space both parameter estimation and sampling become feasible and computationally affordable - no need to overly restrict the form of covariance in the search distribution.

3. Combining estimates from several random subspaces covers the original search space while bypassing the problems of eigenvalue misestimation (Marzetta et al, IEEE T Info Theory ’11; Durrant & Kaban, submitted). Exploration becomes possible even with small population sizes.
New search operators for EDA

Project: takes an $R \in \mathbb{R}^{k \times d}$, an $x_0 \in \mathbb{R}^d$, and a sample $\mathcal{P}^{fit} = (x_i \in \mathbb{R}^d)_{i=1:\tilde{N}}$. Projects $\mathcal{P}^{fit}$ 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 $\hat{\theta}_R = (\hat{\mu}_R, \hat{\Sigma}_R)$ of its distribution $\mathcal{D}_R$ w.r.t. the restriction of the Lebesgue measure to the $k$-dimensional affine subspace defined by $R$.

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 $S^{R_i}_{x_0}, i = 1, ..., M$ and returns a population that lives in the full search space $\mathbb{R}^d$. 
Meta-Algorithm

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 $i = 1, \ldots, M$ ($M \geq 1$) randomly oriented (affine) $k < d$-dimensional subspaces through $x_0$, $S_{R_i}^{x_0}$

   (a) Project $\mathcal{P}^{fit}$ onto $S_{R_i}^{x_0}$

   (b) Produce $N$ new individuals on the subspace $S_{R_i}^{x_0}$ 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.
In line with the spirit of Gaussian EDA, we take $x_0$ to be the average of $P^{fit}$ from the previous generation.

We take Combine to be simply an average, scaled by $\sqrt{M/(k\sigma^2)}$ (to recover the original scale).
Analysis of Algorithm of creating the new generation (1)

- Infinitely many random projections

\[
\frac{1}{M} \sum_{i=1}^{M} R_i^T R_i \Sigma R_i^T R_i \xrightarrow{M \to \infty} \mathbb{E}_R[R^T R \Sigma R^T R]
\]

\[
= \frac{k}{d} \left( \frac{\text{Trace}(\Sigma)}{d} I_d + \frac{k + 1}{d} \Sigma \right)
\]

(details in the paper)
Analysis of Algorithm to create new generation (2)

- Finitely many random projections

**Theorem** 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 $\| \cdot \| = \lambda_{\text{max}}(\cdot)$ the spectral norm of its argument.

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

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

Proof is in the paper.
Experiments

CEC’10 benchmark functions, all $d = 1000$ dimensional.

Created from

1. Separable functions

2. Partially-separable functions, in which a small number of variables are dependent while all the remaining ones are independent

3. Partially-separable functions that consist of multiple independent sub-components, each of which is $m$-non-separable ($m=50$) – this category includes two subtypes: $d/(2m)$-group $m$-nonseparable, and $d/m$-group $m$-nonseparable functions

4. Fully nonseparable functions

20 functions in total, of which 12 are multi-modal - we test on these.
The multi-modal functions are created from shifted & group-rotated (!) versions of some base functions, scaled to d=1000:

Rastrigin function  Ackley function  Rosenbrock function

**Definition.** A nonseparable function $f(x)$ is called m-nonseparable function if at most $m$ of its parameters $x_i$ are not independent. A nonseparable function $f(x)$ is called fully-nonseparable function if any two of its parameters $x_i$ are not independent.
Results vs. the CEC’10 winner & state-of-the-art

1000–dimensional multi-modal benchmark functions from the CEC’10 competition

Results of RP-Ensemble-EDA using $N = 300, \tilde{N} = 75; 6 \times 10^5$ function evaluations.
More details - a $d/(2m)$ group m-nonseparable function ($m=50$)

$d/(2m)$-group shifted & m-rotated Rastrigin function.
More details - a d/m group m-nonseparable function (m=50)

Function 18

- UMDAc
- RP−Ensemble k=15 M=67
- RP−Ensemble k=3 M=334
- RP−Ensemble k=3 M=1000
- RP−Ens k=3 M=∞

d/m-group shifted & m-dimensional Rosenbrock function.
More details - a fully nonseparable function

Shifted d-dimensional Rosenbrock function.
Application to mislabelled gene array classification

Data sets become increasingly complex, new tasks need new models to be optimised.
Application to mislabelled gene array classification

Data sets become increasingly complex, new tasks need new models to be optimised.

We considered classification in the presence of labelling errors – literature reports this is a problem in e.g. biomedical domains. Training set becomes misleading.
Application to mislabelled gene array classification

Data sets become increasingly complex, new tasks need new models to be optimised.

We considered classification in the presence of labelling errors – literature reports this is a problem in e.g. biomedical domains. Training set becomes misleading.

To account for label errors we built a better model – logistic regression with LASSO penalty (convex) but including a model of label errors (no longer convex + multimodal + non-smooth at 0).
Application to mislabelled gene array classification

Data sets become increasingly complex, new tasks need new models to be optimised.

We considered classification in the presence of labelling errors – literature reports this is a problem in e.g. biomedical domains. Training set becomes misleading.

To account for label errors we built a better model – logistic regression with LASSO penalty (convex) but including a model of label errors (no longer convex + multimodal + non-smooth at 0).

Existing specialised algorithms no longer applicable.
Colon Cancer data: 62 points = 40 tumour & 22 normal, 2000 dimensions.

Split the data in 80% training & 20% testing. Training = optimisation of the model likelihood – a 2000-dimensional continuous optimisation problem.
Colon Cancer data: 62 points = 40 tumour & 22 normal, 2000 dimensions.

Split the data in 80% training & 20% testing. Training = optimisation of the model likelihood – a 2000-dimensional continuous optimisation problem.

We use our RP-Ens with $N = 100$, $\tilde{N} = 10$, $k = 3; 10^5$ function evaluations.
Colon Cancer data: 62 points = 40 tumour & 22 normal, 2000 dimensions.

Split the data in 80% training & 20% testing. Training = optimisation of the model likelihood – a 2000-dimensional continuous optimisation problem.

We use our RP-Ens with $N = 100, \tilde{N} = 10, k = 3; 10^5$ function evaluations.

Mean & std of misclassification rates over 10 repeated runs:

<table>
<thead>
<tr>
<th></th>
<th>RP-Ens $k=3$</th>
<th>UMDAc</th>
<th>Adaptation of LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>14.63±2.68</td>
<td>20.02 ±2.62</td>
<td>18.75±1.1</td>
</tr>
</tbody>
</table>
Conclusions & future work

We presented a new framework for designing & developing new EDA-type methods for large scale continuous optimisation.
Conclusions & future work

We presented a new framework for designing & developing new EDA-type methods for large scale continuous 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 reliably & efficiently.
Conclusions & future work

We presented a new framework for designing & developing new EDA-type methods for large scale continuous 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 reliably & efficiently.

We derived some theoretical analysis that shows the effect of our divide-and-conquer methodology when re-assembled and understood in the full search space.
Conclusions & future work

We presented a new framework for designing & developing new EDA-type methods for large scale continuous 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 reliably & efficiently.

We derived some theoretical analysis that shows the effect of our divide-and-conquer methodology when re-assembled and understood in the full search space.

Empirical results on a battery of 12 multimodal test functions from the large scale CEC’10 competition, and a real-world application have been competitive to the best state of the art.
New perspective to research on EDA-type model building optimisation algorithms, and future work is aimed at better understanding and exploiting its potential.
Selected references


