On the Capacity of Evolution Strategies to Statistically Learn the Landscape

Ofer M. Shir
School of Computer Science,
Tel-Hai College, and
The Galilee Research Institute
- Migal, Upper Galilee, Israel
ofersh@telhai.ac.il

Jonathan Roslund
Laboratoire Kastler Brossel,
Université Pierre et Marie
Curie,
Paris, France
jroslund@lkb.upmc.fr

Amir Yehudayoff
Department of Mathematics
Technion - Israel Institute of
Technology,
Haifa, Israel
amir.yehudayoff@gmail.com

ABSTRACT

We investigate the covariance matrix when constructed by Evolution Strategies (ESs) operating with the selection operator alone. We model continuous generation of candidate solutions about quadratic basins of attraction, with deterministic selection of the decision vectors that minimize the objective function values. Our goal is to rigorously show that accumulation of winning individuals carries the potential to reveal valuable information about the search landscape. We first show that the statistically-constructed covariance matrix over such winning decision vectors shares the same eigenvectors with the Hessian matrix about the optimum. We then provide an analytic approximation of this covariance matrix for a non-elitist multi-child $(1, \lambda)$-strategy, which holds for a large population size λ.

Keywords

Theory of evolution strategies; statistical learning; covariance matrix adaptation; landscape Hessian; limit distributions of order statistics; extreme value distributions

1. STATISTICAL LANDSCAPE LEARNING

We outline the *research question* that we target:

> What is the relation between the statistically-learned covariance matrix and the landscape Hessian [2] if a single winner is selected in each iteration assuming generated samples that follow an isotropic Gaussian (no adaptation)?

Let $J : \mathbb{R}^n \to \mathbb{R}$ denote the objective function subject to minimization. We assume that J is minimized at the location $\vec{x}^*$, which is assumed for simplicity to be the origin. The objective function may be *Taylor-expanded* about the optimum. We model the n-dimensional basin of attraction about $\vec{x}^*$ by means of a quadratic approximation. We assume that this expansion is precise $J\left(\vec{x}-\vec{x}^*\right) = J\left(\vec{x}\right) = \vec{x}^T \cdot \mathcal{H} \cdot \vec{x}$, with $\mathcal{H}$ being the landscape Hessian about the optimum.

The canonical non-elitist single-parent ES search process operates in the following manner: The ES generates λ search-points $\vec{x}_1, \ldots, \vec{x}_\lambda$ in each iteration, based upon Gaussian sampling with respect to the given search-point. We are especially concerned with the canonical operation, which adds a normally distributed *mutation* $\vec{z} \sim \mathcal{N}\left(\vec{0}, \mathbf{I}\right)$. Following the evaluation of those λ search points with respect to $J\left(\vec{x}\right)$, the best (minimal) individual is selected and recorded as $\vec{y} = \arg\min\{J(\vec{x}_i)\}$. Finally, let ω denote the *winning* objective function value, $\omega = J(\vec{y}) = \min\{J_1,\ J_2,\ \ldots,\ J_\lambda\}$, where $J_i = J(\vec{x}_i)$. **We distinguish between the optimization phase, which aims to arrive at the optimum and is not discussed here, to the statistical learning of the basin − which lies in the focus of this study.**

The length of a mutation vector, $\sqrt{\vec{z}^T \vec{z}}$, obeys the so-called χ-distribution with n degrees of freedom. We assume a *quadratic* basin of attraction, where the Hessian matrix is positive definite with the following eigendecomposition form,

$$\mathcal{H} = \mathcal{U}\mathcal{D}\mathcal{U}^{-1} \qquad \mathcal{D} = \operatorname{diag}\left[\Delta_1, \ldots, \Delta_n\right], \tag{1}$$

with $\{\Delta_i\}_{i=1}^n$ being the eigenvalues. The random variable $\psi = J(\vec{z})$ obeys a generalized χ^2-distribution, whose CDF is known to follow an *approximation*,

$$F_{\chi^2}\left(\psi\right) = \frac{\Upsilon^\eta}{\Gamma\left(\eta\right)} \int_0^\psi t^{\eta-1} \exp\left(-\Upsilon t\right)\ \mathrm{d}t, \tag{2}$$

with Υ and η accounting for matching the first two moments of $\vec{z}^T \mathcal{H} \vec{z}$: $\Upsilon = \frac{1}{2}\frac{\sum_{i=1}^n \Delta_i}{\sum_{i=1}^n \Delta_i^2}, \quad \eta = \frac{1}{2}\frac{\left(\sum_{i=1}^n \Delta_i\right)^2}{\sum_{i=1}^n \Delta_i^2}$.

We summarize our notation: The random vector $\vec{z}$ is a normal Gaussian mutation and $\psi = J(\vec{z})$. The random vectors $\vec{x}_1, \ldots, \vec{x}_\lambda$ are λ independent copies of $\vec{z}$, and $J_i = J(\vec{x}_i)$. The winner is $\vec{y}$, and $\omega = J(\vec{y})$. The matrix $\mathcal{H}$ is the Hessian about the optimum $\vec{x}^*$, and $\mathcal{C}$ is the covariance matrix of $\vec{y}$.

2. THE COVARIANCE MATRIX

Having the origin set at the parent search-point, which is located at the optimum, the covariance elements are thus defined as:

$$\boxed{\mathcal{C}_{ij} = \int x_i x_j \mathtt{PDF}_{\vec{y}}\left(\vec{x}\right) \mathrm{d}\vec{x}}, \tag{3}$$

where $\mathtt{PDF}_{\vec{y}}\left(\vec{x}\right)$ is an n-dimensional density function characterizing the *winning* decision variables about the optimum. In the *decision-space perspective*, the density function of a

GECCO '16 July 20-24, 2016, Denver, CO, USA

ACM ISBN 978-1-4503-4323-7/16/07.

DOI: http://dx.doi.org/10.1145/2908961.2909057

winning vector of decision variables $\vec{y}$ is related to the density of the *winning* function value ω as follows:

$$\boxed{\mathtt{PDF}_{\vec{y}}\left(\vec{x}\right) = \mathtt{PDF}_{\omega}\left(J\left(\vec{x}\right)\right) \cdot \frac{\mathtt{PDF}_{\vec{z}}\left(\vec{x}\right)}{\mathtt{PDF}_{\psi}\left(J\left(\vec{x}\right)\right)}}, \qquad (4)$$

with $\mathtt{PDF}_{\vec{z}}$ denoting the density function for generating an individual, and $\mathtt{PDF}_{\psi}$ denoting the density function of the objective function values (derived from Eq. (2)). A brief justification follows. The density functions satisfy the conditional relation: $\mathtt{PDF}_{\vec{y}}\left(\vec{x}\right) = \mathtt{PDF}_{\omega}\left(J\left(\vec{x}\right)\right) \cdot \mathtt{PDF}_{\vec{y}|\omega}\left(\vec{x} \mid J\left(\vec{x}\right)\right)$. Now consider the distribution of $[\vec{y};\omega]$ on $\mathbb{R}^{n+1}$. The density of $\vec{y}$ conditioned on the value of $J\left(\vec{y}\right)$ is that of a normal Gaussian subject to this conditioning, since we may sample $[\vec{y};\omega]$ by the following construction: First sample $\{J_1,\ldots,J_\lambda\}$ according to $\mathtt{PDF}_{\psi}$ independently. Then sample $\{\vec{x}_1,\ldots,\vec{x}_\lambda\}$ conditioned on the values of $J_1,\ldots,J_\lambda$ independently. Finally, J may be set to $J_\ell = \omega$ that is minimal and $\vec{y}$ set to the respective $\vec{x}_\ell$. Overall, the winning vector $\vec{y}$ conditioned on the winning value ω is generated in the same manner as a normally-distributed $\vec{z}$ conditioned on ψ.

THEOREM 1. *The covariance matrix and the Hessian are* ***commuting matrices*** *when the objective function follows the quadratic approximation.*

PROOF. Given the density function in Eq. (4), the objective function is assumed to satisfy $J\left(\vec{x}\right) = \vec{x}^T \cdot \mathcal{H} \cdot \vec{x}$, and the covariance matrix reads:

$$\mathcal{C}_{ij} = \int x_i x_j \mathtt{PDF}_{\omega}\left(\vec{x}^T \cdot \mathcal{H} \cdot \vec{x}\right) \cdot \frac{\mathtt{PDF}_{\vec{z}}\left(\vec{x}\right)}{\mathtt{PDF}_{\psi}\left(\vec{x}^T \cdot \mathcal{H} \cdot \vec{x}\right)} \mathrm{d}\vec{x}. \qquad (5)$$

Consider the orthogonal matrix $\mathcal{U}$, which diagonalizes $\mathcal{H}$ into $\mathcal{D}$ and possesses a determinant of value 1 (as in Eq. (1)):

$$\vec{\vartheta} = \mathcal{U}^{-1}\vec{x}, \qquad \mathrm{d}\vec{\vartheta} = \mathrm{d}\vec{x}.$$

We target the integral $\mathcal{I}_{ij} = \left(\mathcal{U}^{-1}\mathcal{C}\mathcal{U}\right)_{ij}$ and apply a change of variables into $\vec{\vartheta}$ (after changing order of summations):

$$\mathcal{I}_{ij} = \frac{1}{\sqrt{(2\pi)^n}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \vartheta_i \vartheta_j \exp\left(-\frac{1}{2}\vec{\vartheta}^T\vec{\vartheta}\right) \times$$
$$\times \frac{\mathtt{PDF}_{\omega}\left(\vec{\vartheta}^T \cdot \mathcal{D} \cdot \vec{\vartheta}\right)}{\mathtt{PDF}_{\psi}\left(\vec{\vartheta}^T \cdot \mathcal{D} \cdot \vec{\vartheta}\right)} \mathrm{d}\vartheta_1 \mathrm{d}\vartheta_2 \cdots \mathrm{d}\vartheta_n. \qquad (6)$$

$\mathcal{I}_{ij}$ vanishes for any $i \neq j$ due to symmetry considerations: the overall integrand is an *odd* function, because all the terms are *even* functions, except for ϑ_j, ϑ_i when they differ. Therefore, the integration over the entire domain yields zero. Hence, $\mathcal{I}$ is the diagonalized form of $\mathcal{C}$, with $\mathcal{U}$ holding the eigenvectors. $\mathcal{C}$ is thus diagonalized by the same eigenvectors as $\mathcal{H}$, and therefore, by definition, they are *commuting matrices*, as claimed. □

3. ANALYTICAL APPROXIMATION

We provide an approximation for $\mathtt{PDF}_{\omega}\left(J\left(\vec{x}\right)\right)$ and consequently for $\mathtt{PDF}_{\vec{y}}\left(\vec{x}\right)$ to calculate the covariance matrix using Eq. (3). We consider here a non-elitist multi-child selection with λ offspring. The distribution function of the winning event amongst λ candidates and its derived density are:

$$\begin{array}{l} \mathtt{CDF}_{\omega}\left(\psi\right) = \mathtt{Pr}\left\{\omega \leq \psi\right\} = 1 - \left(1 - \mathtt{CDF}_{\psi}\left(\psi\right)\right)^{\lambda} \\ \mathtt{PDF}_{\omega}\left(\psi\right) = \lambda \cdot \left(1 - \mathtt{CDF}_{\psi}\left(\psi\right)\right)^{\lambda-1} \cdot \mathtt{PDF}_{\psi}\left(\psi\right). \end{array} \qquad (7)$$

Upon substituting the explicit forms into $\mathtt{CDF}_{\psi}$ and $\mathtt{PDF}_{\psi}$ using Eq. (2), the desired density function $\mathtt{PDF}_{\omega}\left(J\left(\vec{x}\right)\right)$ is obtained, however not in a closed form.

We treat the derived winners' distribution for large sample sizes, i.e., when the population size tends to infinity [3], and adhere to the Fisher-Tippett theorem. We consider *minimal generalized extreme value distributions* ($\mathrm{GEVD}_{\mathrm{min}}$) [1], and are able to show that **under the GEVD approximation**, $\lambda \rightarrow \infty$, upon normalizing the random variable to $\tilde{\psi} = \left(\omega - b^*_\lambda\right)/a^*_\lambda$ and using the tail index $\frac{n}{2}$, **the CDF and PDF forms for the single winning event read**:

$$\begin{array}{l} \mathtt{CDF}_{\omega}^{\mathrm{GEVD}}\left(\tilde{\psi}\right) = 1 - \exp\left(-\tilde{\psi}^{\frac{n}{2}}\right) \\ \mathtt{PDF}_{\omega}^{\mathrm{GEVD}}\left(\tilde{\psi}\right) = \frac{n}{2}\tilde{\psi}^{\frac{n}{2}-1}\exp\left(-\tilde{\psi}^{\frac{n}{2}}\right) \end{array}. \qquad (8)$$

GEVD convergence is ensured by the constants $a^*_\lambda = F^{-1}_{\chi^2}\left(\frac{1}{\lambda}\right)$, $b^*_\lambda = 0$. J is assumed to satisfy $J\left(\vec{x}\right) = \vec{x}^T \cdot \mathcal{H} \cdot \vec{x}$, and must be normalized only for $\mathtt{PDF}_{\omega}$ by means of $\tilde{J}(\vec{x}) \equiv \left(J(\vec{x})\right)/a^*_\lambda$; then Eq. (3) may be rewritten using Eq. (4):

$$\boxed{\begin{array}{l} \mathcal{C}_{ij} = \displaystyle\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} N_{\mathcal{C}} x_i x_j \left(\vec{x}^T\mathcal{H}\vec{x}\right)^{\frac{n}{2}-\eta} \times \\ \times \exp\left[\Upsilon\vec{x}^T\mathcal{H}\vec{x} - \left(\frac{\vec{x}^T\mathcal{H}\vec{x}}{a^*_\lambda}\right)^{\frac{n}{2}} - \frac{1}{2}\vec{x}^T\vec{x}\right] \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n. \end{array}} \qquad (9)$$

with a normalizing constant $N_{\mathcal{C}} = \frac{n\Gamma(\eta)}{2\Upsilon^\eta\left(a^*_\lambda\right)^{\frac{n}{2}-1}\sqrt{(2\pi)^n}}$.

For the *isotropic case*, $\mathcal{H} = h_0\mathbf{I}$, the integration is straightforward ($\eta = \frac{n}{2}$, $\Upsilon = \frac{1}{2h_0}$) – the attained covariance is proportional to the inverse Hessian, multiplied by a factor:

$$\mathcal{C}^{(\mathcal{H}=h_0\mathbf{I})} = \frac{\Gamma(\frac{n}{2}) \cdot \Gamma\left(1 + \frac{2}{n}\right) \cdot c\left(n\right) \cdot a^*_\lambda}{2\pi^{n/2}} \cdot \mathcal{H}^{-1}, \qquad (10)$$

wherein

$$c\left(n\right) = \left\{ \begin{array}{ll} \frac{\pi^m}{m!} & n = 2m \\ \frac{2^{m+1}\pi^m}{1\cdot3\cdot5\cdots(2m+1)} & n = 2m+1 \end{array} \right. . \qquad (11)$$

For the *general case* of any positive-definite Hessian $\mathcal{H}$, the integral in Eq. (9) has an unknown closed form.

The reported results herein have been numerically validated to a satisfactory level for different landscape Hessians at various dimensions. The validation comprised the commuting property devised by Theorem 1, the accuracy of the *approximated* distributions described at Eqs. (2) and (8), and the approximated integral of Eq. (9), which was successfully corroborated for the isotropic case at a spectrum of dimensions and for the general case at $n = 3$.

Possible directions for future research include exploration of this model subject to non-isotropic Gaussian mutations, and its investigation for a multi-parent (μ,λ)-strategy.

4. REFERENCES

[1] E. Castillo, A. S. Hadi, N. Balakrishnan, and J. M. Sarabia. *Extreme Value and Related Models with Applications in Engineering and Science.* John Wiley and Sons, 2004.

[2] G. Rudolph. On Correlated Mutations in Evolution Strategies. In *Parallel Problem Solving from Nature - PPSN II*, pages 105–114, Amsterdam, 1992. Elsevier.

[3] A. Zhigljavsky and A. Žilinskas. *Stochastic Global Optimization.* Springer Optimization and Its Applications. Springer US, 2007.