Fundamentals of ESs' Statistical Learning

Ofer M. Shir (Tel-Hai College & Migal Institute)

Joint work with Amir Yehudayoff, Mathematics/Technion

ofersh@telhai.ac.il

Dagstuhl Seminar 17191
Theory of Randomized Optimization Heuristics
May 11 2017, Schloss Dagstuhl, DE

Outline

1. Introduction
   - Model
2. The Analytical Covariance Matrix
   - A Single Winner: $(1, \lambda)$-Selection
   - (μ, λ)-Truncation Selection
3. The Inverse Relation
   - Proving $\lim_{\lambda\to\infty} \alpha\mathcal{C}\mathcal{H} = \mathbf{I}$
4. Discussion
5. Backup Slides

ESs' statistical landscape learning

the classical hypothesis $\mathcal{C} \rightarrow \mathcal{H}^{-1}$

- Open question since the early development of ESs; widely discussed [Rudolph1992]

- Sheer amount of empirical evidence for this relation, in addition to extensive branding "C=inv(H)" made this hypothesis a practical *postulate* throughout the years

the classical hypothesis $\mathcal{C} \to \mathcal{H}^{-1}$

- Open question since the early development of ESs; widely discussed [Rudolph1992]

- Sheer amount of empirical evidence for this relation, in addition to extensive branding "C=inv(H)" made this hypothesis a practical *postulate* throughout the years

- Recent proofs published, yet limited to Derandomization (or Natural Gradient); they exercise IGO [Akimoto2012, Beyer2014]

- We seek the fundamentals of this learning capability and consider a theoretical model – which is not likely to reflect an everyday's heuristic – e.g., $\mathcal{C}$'s eigenvalues are $\Omega\left(1/\lambda^2\right)$ and λ tends to infinity

so why bother?

- IGO is an elephant gun – can we target an equivalent result in more fundamental ways? (David's stone+sling to knock Goliath down)

so why bother?

- IGO is an elephant gun – can we target an equivalent result in more fundamental ways? (David's stone+sling to knock Goliath down)

- "Going back to basics" using first principles of *probability theory* and *calculus* on a basic ES model

so why bother?

- IGO is an elephant gun – can we target an equivalent result in more fundamental ways? (David's stone+sling to knock Goliath down)
- "Going back to basics" using first principles of *probability theory* and *calculus* on a basic ES model
- Mathematically beautiful, but may also serve as a tool elsewhere in the future
- This work concerns the absolutely continuous case, but any theory guy should find it interesting :-)

statistical sampling by $(1, \lambda)$-selection

$t \leftarrow 0$

$\mathcal{S} \leftarrow \emptyset$

repeat

  for $k \leftarrow 1$ **to** λ **do**

    $\vec{x}_k^{(t+1)} \leftarrow \vec{x}_0 + \vec{z}_k, \quad \vec{z}_k \sim \mathcal{N}(\vec{0}, \mathbf{I})$

    $J_k^{(t+1)} \leftarrow \texttt{evaluate}\left(\vec{x}_k^{(t+1)}\right)$

  end

  $m_{t+1} \leftarrow \arg\min\left(\left\{J_\imath^{(t+1)}\right\}_{\imath=1}^{\lambda}\right)$

  $\mathcal{S} \leftarrow \mathcal{S} \cup \left\{\vec{x}_{m_{t+1}}^{(t+1)}\right\}$

  $t \leftarrow t + 1$

until $t \geq N_{\texttt{iter}}$

output: $\mathcal{C}^{\texttt{stat}} = \texttt{statCovariance}(\mathcal{S})$

first results [FOGA'17]

quadratic approximation; optimum's vicinity

$$J\left(\vec{x}\right) = J\left(\vec{x}-\vec{x}^{*}\right) = \vec{x}^{T}\cdot\mathcal{H}\cdot\vec{x} \tag{1}$$

first results [FOGA'17]

quadratic approximation; optimum's vicinity

$$J(\vec{x}) = J(\vec{x} - \vec{x}^{*}) = \vec{x}^{T} \cdot \mathcal{H} \cdot \vec{x} \tag{1}$$

main results

i. Rigorous formulation of the covariance matrix $\mathcal{C}$ over ESs' selected individuals ("winners")
ii. The covariance matrix and the Hessian commute and are simultaneously diagonalizable for any λ for (μ, λ)-selection
iii. For every invertible $\mathcal{H}$ and $\lambda \in \mathbb{N}$, there exists a constant $\alpha = \alpha(\mathcal{H}, \lambda) > 0$ such that

$$\lim_{\lambda \to \infty} \alpha \mathcal{C} \mathcal{H} = \mathbf{I}.$$

current model

general quadratic approximation

$$J(\vec{x}) = (\vec{x} - \vec{x}^*)^T \cdot \mathcal{H} \cdot (\vec{x} - \vec{x}^*) \tag{2}$$

current model

general quadratic approximation

$$J\left(\vec{x}\right) = \left(\vec{x} - \vec{x}^{*}\right)^{T} \cdot \mathcal{H} \cdot \left(\vec{x} - \vec{x}^{*}\right) \tag{2}$$

truncation selection ("winners")

$$\vec{y} = \arg\min\left\{J(\vec{x}_1),\ J(\vec{x}_2),\ \ldots,\ J(\vec{x}_\lambda)\right\} \tag{3}$$

$$\omega = J(\vec{y}) = \min\left\{J(\vec{x}_1),\ J(\vec{x}_2),\ \ldots,\ J(\vec{x}_\lambda)\right\} \tag{4}$$

the covariance matrix

the analytical form

The expectation vector of the winner is defined by its i^{th} element:

$$\mathcal{E}_i = \int x_i \texttt{PDF}_{\vec{y}}\left(\vec{x}\right) \mathrm{d}\vec{x} \ , \tag{5}$$

$$\boxed{\mathcal{C}_{ij} = \int (x_i - \mathcal{E}_i)(x_j - \mathcal{E}_j) \texttt{PDF}_{\vec{y}}\left(\vec{x}\right) \mathrm{d}\vec{x}} \ . \tag{6}$$

$\texttt{PDF}_{\vec{y}}\left(\vec{x}\right)$ is an n-dimensional density function characterizing the *winning* decision variables about the optimum.

the analytical form

The expectation vector of the winner is defined by its i^{th} element:

$$\mathcal{E}_i = \int x_i \texttt{PDF}_{\vec{y}}\left(\vec{x}\right) \mathrm{d}\vec{x} \ , \tag{5}$$

$$\boxed{\mathcal{C}_{ij} = \int (x_i - \mathcal{E}_i)(x_j - \mathcal{E}_j) \texttt{PDF}_{\vec{y}}\left(\vec{x}\right) \mathrm{d}\vec{x}} \ . \tag{6}$$

$\texttt{PDF}_{\vec{y}}\left(\vec{x}\right)$ is an n-dimensional density function characterizing the *winning* decision variables about the optimum.

One of the primary goals is to fully understand this expression and utilize it.

winners' density in $(1, \lambda)$-selection

Proposition 0

$$\mathtt{PDF}_{\vec{y}}\left(\vec{x}\right) = \mathtt{PDF}_{\omega}\left(J\left(\vec{x}\right)\right) \cdot \frac{\mathtt{PDF}_{\vec{z}}\left(\vec{x}\right)}{\mathtt{PDF}_{\psi}\left(J\left(\vec{x}\right)\right)} \tag{7}$$

- $\mathtt{PDF}_{\omega}$: density of the *winning* value ω
- $\mathtt{PDF}_{\vec{z}}$: density for generating an individual by *mutation*
- $\mathtt{PDF}_{\psi}$: density of the objective function values (Eqs. 13 or 16)

winners' density in $(1,\lambda)$-selection

Proposition 0

$$\mathtt{PDF}_{\vec{y}}\left(\vec{x}\right) = \mathtt{PDF}_{\omega}\left(J\left(\vec{x}\right)\right) \cdot \frac{\mathtt{PDF}_{\vec{z}}\left(\vec{x}\right)}{\mathtt{PDF}_{\psi}\left(J\left(\vec{x}\right)\right)} \tag{7}$$

- $\mathtt{PDF}_{\omega}$: density of the *winning* value ω
- $\mathtt{PDF}_{\vec{z}}$: density for generating an individual by *mutation*
- $\mathtt{PDF}_{\psi}$: density of the objective function values (Eqs. 13 or 16)

sketch: consider the distribution of $[\vec{y};\omega]$ on $\mathbb{R}^{n+1}$
i. sample $\{J_1,\ldots,J_\lambda\}$ according to $\mathtt{PDF}_{\psi}$ independently
ii. sample $\{\vec{x}_1,\ldots,\vec{x}_\lambda\}$ conditioned on $J_1,\ldots,J_\lambda$ independently
iii. ω is set to the minimum J_ℓ, and $\vec{y}$ is set to $\vec{x}_\ell$

(μ, λ)-selection

- $J_{1:\lambda} \leq J_{2:\lambda} \leq \ldots \leq J_{\lambda:\lambda}$ are the order statistics obtained by sorting the objective function values.
- $\omega_{1:\lambda}, \ldots, \omega_{\mu:\lambda}$ are the first μ values from this list.
- $\vec{y}_{1:\lambda}, \ldots, \vec{y}_{\mu:\lambda}$ are their corresponding vectors.

(μ, λ)-selection

- $J_{1:\lambda} \leq J_{2:\lambda} \leq \ldots \leq J_{\lambda:\lambda}$ are the order statistics obtained by sorting the objective function values.
- $\omega_{1:\lambda}, \ldots, \omega_{\mu:\lambda}$ are the first μ values from this list.
- $\vec{y}_{1:\lambda}, \ldots, \vec{y}_{\mu:\lambda}$ are their corresponding vectors.

To study the covariance in this case, we consider the pairwise density of the k^{th}-degree and ℓ^{th}-degree winners ($\ell > k$):

$$\boxed{\begin{aligned} \mathtt{PDF}_{\vec{y}_{k:\lambda},\vec{y}_{\ell:\lambda}}\left(\vec{x}_k, \vec{x}_\ell\right) = \mathtt{PDF}_{\omega_{k:\lambda},\omega_{\ell:\lambda}}\left(J\left(\vec{x}_k\right), J\left(\vec{x}_\ell\right)\right) \times \\ \times \left(\frac{\mathtt{PDF}_{\vec{z}}\left(\vec{x}_k\right)}{\mathtt{PDF}_{\psi}\left(J\left(\vec{x}_k\right)\right)}\right) \cdot \left(\frac{\mathtt{PDF}_{\vec{z}}\left(\vec{x}_\ell\right)}{\mathtt{PDF}_{\psi}\left(J\left(\vec{x}_\ell\right)\right)}\right) \end{aligned}} \cdot \tag{8}$$

the inverse relation

winning values' density & proposition 1

For simplicity, we consider $(1, \lambda)$-selection.

winning values' density & proposition 1

For simplicity, we consider $(1, \lambda)$-selection.

$$\mathtt{CDF}_{\omega}(v) = 1 - (1 - \mathtt{CDF}_{\psi}(v))^{\lambda} \tag{9}$$

$$\mathtt{PDF}_{\omega}(v) = \lambda \cdot (1 - \mathtt{CDF}_{\psi}(v))^{\lambda-1} \cdot \mathtt{PDF}_{\psi}(v) \tag{10}$$

winning values' density & proposition 1

For simplicity, we consider $(1, \lambda)$-selection.

$$\mathtt{CDF}_{\omega}(v) = 1 - \left(1 - \mathtt{CDF}_{\psi}(v)\right)^{\lambda} \tag{9}$$

$$\mathtt{PDF}_{\omega}(v) = \lambda \cdot \left(1 - \mathtt{CDF}_{\psi}(v)\right)^{\lambda-1} \cdot \mathtt{PDF}_{\psi}(v) \tag{10}$$

Proposition 1:
For every invertible $\mathcal{H}$ and $\lambda \in \mathbb{N}$, there exists a constant $\alpha = \alpha(\mathcal{H}, \lambda) > 0$ such that

$$\lim_{\lambda \to \infty} \alpha \mathcal{C} \mathcal{H} = \mathbf{I}.$$

intuition for proving proposition 1

We first target a diagonal $\mathcal{H}$

For a large λ, the winner $\vec{y}$ is close to the optimum, which in turn implies that $(\mathcal{C}\mathcal{H})_{ii}$ does not actually depend on i.

For the general case, we ought to show that both $\mathcal{H}$ and $\mathcal{C}$ are diagonalizable in the same base under the same limit conditions (not shown here).

proof sketch for proposition 1

i. firstly, assume $\mathcal{H}$ is diagonal and apply change of variables $r_i = \sqrt{\Delta_i} \cdot (x_i - x_i^*)$

ii. $\mathcal{E}_i - x_i^* = \frac{c_{\mathcal{H}}}{\sqrt{\Delta_i}} \int r_i \lambda (1 - \mathtt{CDF}_\psi(\|\vec{r}\|^2))^{\lambda-1} \exp\left(-\hat{J}(\vec{r})\right) \mathrm{d}\vec{r}$

iii. show that $|\mathcal{E}_i - x_i^*| \leq \epsilon_1 \sqrt{\mathcal{C}_{ii}}$

iv. bound the off-diagonal terms $\mathcal{C}_{ij} \leq \epsilon_2 \sqrt{\mathcal{C}_{ii}\mathcal{C}_{jj}}$
v. show that $\alpha\Delta_i\mathcal{C}_{ii} \geq 1 - \epsilon_3$ and $\alpha\Delta_i\mathcal{C}_{ii} \leq 1 + \epsilon_4$ (ϵ_3 and ϵ_4 tend to zero as λ tends to infinity)

vi. secondly, for a non-diagonal $\mathcal{H}$,

$$\lim_{\lambda\to\infty} \alpha\mathcal{C}\mathcal{H} - \mathbf{I} = \lim_{\lambda\to\infty} \mathcal{U}\left(\alpha\mathcal{T}\mathcal{D} - \mathbf{I}\right)\mathcal{U}^{-1} = 0$$

discussion

i. $\mathcal{C}$ and $\mathcal{H}$ commute (for any λ near the optimum, for $\lambda \to \infty$ elsewhere).
this learning capability stems only from two components:
(1) isotropic Gaussian mutations, and (2) rank-based selection.
* learning the landscape is an inherent property of classical ESs.
** it does not require Derandomization (adaptation) nor IGO (proofs)

discussion

i. $\mathcal{C}$ and $\mathcal{H}$ commute (for any λ near the optimum, for $\lambda \to \infty$ elsewhere).
this learning capability stems only from two components:
(1) isotropic Gaussian mutations, and (2) rank-based selection.
* learning the landscape is an inherent property of classical ESs.
** it does not require Derandomization (adaptation) nor IGO (proofs)

ii. $\lim_{\lambda\to\infty} \alpha\mathcal{C}\mathcal{H} = \mathbf{I}$; this approximation has two parts:
(1) guaranteeing that $\mathcal{C}^{\texttt{stat}}$ is pointwise ϵ-close to $\mathcal{C}$ with confidence $1-\delta$. the eigenvalues of $\mathcal{C}$ are at least $\Omega(1/\lambda^2)$; for $\mathcal{C}^{\texttt{stat}}$ to meaningfully approach $\mathcal{C}$ it requires $\epsilon \ll 1/\lambda^2$.
$\Longrightarrow$ number of samples required for this part is polynomial in $\lambda, 1/\epsilon, \ln(n)$ and $\ln(1/\delta)$.
(2) guaranteeing that $\mathcal{C}$ is pointwise ϵ-close to $\alpha\mathcal{H}^{-1}$, $\alpha\left(\lambda, \mathcal{H}\right) > 0$.
$\Longrightarrow$ upper bound on the number of samples required for this part depends on ϵ, λ and on the spectrum of $\mathcal{H}$.

limit distributions of order statistics

In order to calculate $\mathcal{C}_{ij}$ when λ tends to infinity, it is possible to approximate $\texttt{PDF}_{\omega}\left(J\left(\vec{x}\right)\right)$, by considering $\mathcal{L}_{\lambda}\left(v\right) = 1 - \left(1 - \texttt{CDF}_{\psi}\left(v\right)\right)^{\lambda}$ at $\lim_{\lambda \longrightarrow \infty} \mathcal{L}_{\lambda}\left(v\right)$:

limit distributions of order statistics

In order to calculate $\mathcal{C}_{ij}$ when λ tends to infinity, it is possible to approximate $\texttt{PDF}_{\omega}\left(J\left(\vec{x}\right)\right)$, by considering $\mathcal{L}_{\lambda}\left(v\right) = 1 - \left(1 - \texttt{CDF}_{\psi}\left(v\right)\right)^{\lambda}$ at $\lim_{\lambda \longrightarrow \infty} \mathcal{L}_{\lambda}\left(v\right)$:

theorem [Fisher-Tippett]

the generalized extreme value distributions (GEVD) are the only non-degenerate family of distributions satisfying this limit:

$$\mathcal{L}_{\kappa}\left(v;\kappa_1,\kappa_2,\kappa_3\right) = 1 - \exp\left\{-\left[1+\kappa_3\left(\frac{v-\kappa_1}{\kappa_2}\right)\right]^{1/\kappa_3}\right\} \tag{11}$$

$\Longrightarrow$ $\texttt{CDF}_{\psi}$ belongs to Weibull

limit distributions of order statistics

In order to calculate $\mathcal{C}_{ij}$ when λ tends to infinity, it is possible to approximate $\texttt{PDF}_{\omega}\left(J\left(\vec{x}\right)\right)$, by considering $\mathcal{L}_{\lambda}\left(v\right) = 1 - \left(1 - \texttt{CDF}_{\psi}\left(v\right)\right)^{\lambda}$ at $\lim_{\lambda \longrightarrow \infty} \mathcal{L}_{\lambda}\left(v\right)$:

theorem **[Fisher-Tippett]**

the generalized extreme value distributions (GEVD) are the only non-degenerate family of distributions satisfying this limit:

$$\mathcal{L}_{\kappa}\left(v; \kappa_1, \kappa_2, \kappa_3\right) = 1 - \exp\left\{-\left[1 + \kappa_3\left(\frac{v - \kappa_1}{\kappa_2}\right)\right]^{1/\kappa_3}\right\} \qquad (11)$$

$\Longrightarrow$ $\texttt{CDF}_{\psi}$ belongs to Weibull

This tool is hardly ever exercised amongst our scholars; Rudolph utilized it in his book [Rudolph1997].

Acknowledgements to Jonathan Roslund.

danke

probability functions

isotropic case: $\mathcal{H} = \mathbf{I}$
$\psi = J(\vec{z})$ is a random variable obeying the χ^2-distribution:

$$F_{\chi^2}(\psi) = \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^{\psi} t^{\frac{n}{2}-1} \exp\left(-\frac{t}{2}\right) \, \mathrm{d}t \tag{12}$$

$$f_{\chi^2}(\psi) = \frac{1}{2^{n/2}\Gamma(n/2)} \psi^{n/2-1} \exp\left(-\frac{\psi}{2}\right) \tag{13}$$

probability functions

isotropic case: $\mathcal{H} = \mathbf{I}$
$\psi = J(\vec{z})$ is a random variable obeying the χ^2-distribution:

$$F_{\chi^2}(\psi) = \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^{\psi} t^{\frac{n}{2}-1} \exp\left(-\frac{t}{2}\right) \, \mathrm{d}t \tag{12}$$

$$f_{\chi^2}(\psi) = \frac{1}{2^{n/2}\Gamma(n/2)} \psi^{n/2-1} \exp\left(-\frac{\psi}{2}\right) \tag{13}$$

general case: $\mathcal{H} = \mathcal{U}\mathcal{D}\mathcal{U}^{-1}, \qquad \mathcal{D} = \mathbf{diag}\left[\Delta_1, \ldots, \Delta_n\right]$

$$\begin{aligned} F_{\mathcal{H}\chi^2}(\psi) = \textstyle\int_0^{\infty} \frac{2}{\pi} \frac{\sin\frac{t\psi}{2}}{t} \cos\left(-t\psi + \frac{1}{2}\sum_{j=1}^{n} \tan^{-1} 2\Delta_j t\right) \\ \times \prod_{j=1}^{n} \left(1 + \Delta_j^2 t^2\right)^{-\frac{1}{4}} \, \mathrm{d}t, \end{aligned} \tag{14}$$

approximation for the general case

$$F_{\tau\chi^2}(\psi) = \frac{\Upsilon^\eta}{\Gamma(\eta)} \int_0^\psi t^{\eta-1} \exp(-\Upsilon t)\ \mathrm{d}t \tag{15}$$

$$f_{\tau\chi^2}(\psi) = \frac{\Upsilon^\eta}{\Gamma(\eta)} \psi^{\eta-1} \exp(-\Upsilon \psi) \tag{16}$$

Υ and η account for the first two moments of $\vec{z}^T \mathcal{H} \vec{z}$:

$$\Upsilon = \frac{1}{2} \frac{\sum_{i=1}^n \Delta_i}{\sum_{i=1}^n \Delta_i^2}, \quad \eta = \frac{1}{2} \frac{\left(\sum_{i=1}^n \Delta_i\right)^2}{\sum_{i=1}^n \Delta_i^2} \tag{17}$$

approximation for the general case

$$F_{\tau\chi^2}\left(\psi\right) = \frac{\Upsilon^\eta}{\Gamma\left(\eta\right)} \int_0^\psi t^{\eta-1} \exp\left(-\Upsilon t\right)\, \mathrm{d}t \tag{15}$$

$$f_{\tau\chi^2}\left(\psi\right) = \frac{\Upsilon^\eta}{\Gamma\left(\eta\right)} \psi^{\eta-1} \exp\left(-\Upsilon \psi\right) \tag{16}$$

Υ and η account for the first two moments of $\vec{z}^T \mathcal{H} \vec{z}$:

$$\Upsilon = \frac{1}{2} \frac{\sum_{i=1}^n \Delta_i}{\sum_{i=1}^n \Delta_i^2}, \quad \eta = \frac{1}{2} \frac{\left(\sum_{i=1}^n \Delta_i\right)^2}{\sum_{i=1}^n \Delta_i^2} \tag{17}$$

Accuracy depends on the eigenvalues' $\{\Delta_i\}$ *standard deviation*.