# On the Statistical Learning Ability of ESs

Ofer M. Shir (Tel-Hai & Migal)
Amir Yehudayoff (Technion-IIT)

ofersh@telhai.ac.il amir.yehudayoff@gmail.com

Foundations of Genetic Algorithms (FOGA-XIV)
January 12-15, 2017, Copenhagen, Denmark

# Outline

# ESs' statistical landscape learning

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

- Open question since the development of ESs
- Sheer amount of empirical evidence for this relation + 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]

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

- Open question since the development of ESs
- Sheer amount of empirical evidence for this relation + 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]

- Current study: "going back to basics" using first principles of probability theory on a classical ES model

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

- Open question since the development of ESs
- Sheer amount of empirical evidence for this relation + 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]

- Current study: "going back to basics" using first principles of probability theory on a classical ES model
- This work concerns the absolutely continuous case, but should still interest the discrete guys in the audience ...

# model

## quadratic approximation; optimum's vicinity

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

# model

**quadratic approximation; optimum's vicinity**

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

**sampling**
$\lambda$ search-points are generated in each iteration using isotropic mutations, $\vec{z} \sim \mathcal{N}(\vec{0}, \mathbf{I})$;
i.e., $\vec{x}_1, \ldots, \vec{x}_\lambda$ are independent and each is $\mathcal{N}(\vec{0}, \mathbf{I})$

# model

## quadratic approximation; optimum's vicinity

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

## sampling

$\lambda$ search-points are generated in each iteration using isotropic mutations, $\vec{z}\sim\mathcal{N}(\vec{0},\mathbf{I})$;
i.e., $\vec{x}_1,\ldots,\vec{x}_\lambda$ are independent and each is $\mathcal{N}(\vec{0},\mathbf{I})$

## truncation selection ("winners")

$$\vec{y}=\arg\min\left\{J(\vec{x}_1),\ J(\vec{x}_2),\ \ldots,\ J(\vec{x}_\lambda)\right\} \tag{2}$$

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

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

$t \leftarrow 0$

$\mathcal{S} \leftarrow \emptyset$

**repeat**

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

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

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

  **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** $t \geq N_{\mathtt{iter}}$

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

# probability functions

**isotropic case: $\mathcal{H} = \mathbf{I}$**

$\psi = J(\vec{z})$ is a random variable obeying the $\chi^2$-distribution:

$$F_{\chi^2}(\psi) = \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^{\psi} t^{\frac{n}{2}-1} \exp\left(-\frac{t}{2}\right) \, \mathrm{d}t \tag{4}$$

$$f_{\chi^2}(\psi) = \frac{1}{2^{n/2}\Gamma(n/2)} \psi^{n/2-1} \exp\left(-\frac{\psi}{2}\right) \tag{5}$$

# probability functions

**isotropic case:** $\mathcal{H} = \mathbf{I}$

$\psi = J(\vec{z})$ is a random variable obeying the $\chi^2$-distribution:

$$F_{\chi^2}(\psi) = \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^{\psi} t^{\frac{n}{2}-1} \exp\left(-\frac{t}{2}\right) \,\mathrm{d}t \tag{4}$$

$$f_{\chi^2}(\psi) = \frac{1}{2^{n/2}\Gamma(n/2)} \psi^{n/2-1} \exp\left(-\frac{\psi}{2}\right) \tag{5}$$

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

$$\begin{aligned} F_{\mathcal{H}\chi^2}(\psi) = \textstyle\int_0^{\infty} \frac{2}{\pi} \frac{\sin\frac{t\psi}{2}}{t} \cos\left(-t\psi + \frac{1}{2}\sum_{j=1}^{n} \tan^{-1} 2\Delta_j t\right) & \\ \times \prod_{j=1}^{n} \left(1 + \Delta_j^2 t^2\right)^{-\frac{1}{4}} \,\mathrm{d}t, & \end{aligned} \tag{6}$$

# approximation for the general case

$$F_{\tau\chi^2}\left(\psi\right) = \frac{\Upsilon^{\eta}}{\Gamma\left(\eta\right)} \int_0^{\psi} t^{\eta-1} \exp\left(-\Upsilon t\right)\, \mathrm{d}t \tag{7}$$

$$f_{\tau\chi^2}\left(\psi\right) = \frac{\Upsilon^{\eta}}{\Gamma\left(\eta\right)} \psi^{\eta-1} \exp\left(-\Upsilon \psi\right) \tag{8}$$

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

$$\Upsilon = \frac{1}{2} \frac{\sum_{i=1}^{n} \Delta_i}{\sum_{i=1}^{n} \Delta_i^2}, \quad \eta = \frac{1}{2} \frac{\left(\sum_{i=1}^{n} \Delta_i\right)^2}{\sum_{i=1}^{n} \Delta_i^2} \tag{9}$$

# approximation for the general case

$$F_{\tau\chi^2}(\psi) = \frac{\Upsilon^\eta}{\Gamma(\eta)} \int_0^\psi t^{\eta-1} \exp(-\Upsilon t) \ \mathrm{d}t \qquad (7)$$

$$f_{\tau\chi^2}(\psi) = \frac{\Upsilon^\eta}{\Gamma(\eta)} \psi^{\eta-1} \exp(-\Upsilon \psi) \qquad (8)$$

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

$$\Upsilon = \frac{1}{2} \frac{\sum_{i=1}^n \Delta_i}{\sum_{i=1}^n \Delta_i^2}, \quad \eta = \frac{1}{2} \frac{\left(\sum_{i=1}^n \Delta_i\right)^2}{\sum_{i=1}^n \Delta_i^2} \qquad (9)$$

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

# the covariance matrix

# the analytical form

The origin is set at the parent search-point, which is located at the optimum:

$$\boxed{\mathcal{C}_{ij} = \int x_i x_j \texttt{PDF}_{\vec{y}}\left(\vec{x}\right) \mathrm{d}\vec{x}} \tag{10}$$

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

# the analytical form

The origin is set at the parent search-point, which is located at the optimum:

$$\boxed{\mathcal{C}_{ij} = \int x_i x_j \texttt{PDF}_{\vec{y}}\left(\vec{x}\right) \mathrm{d}\vec{x}} \tag{10}$$

$\texttt{PDF}_{\vec{y}}\left(\vec{x}\right)$ **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.

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

**Proposition 0**

$$\mathtt{PDF}_{\vec{y}}\left(\vec{x}\right) = \mathtt{PDF}_{\omega}\left(J\left(\vec{x}\right)\right) \cdot \frac{\mathtt{PDF}_{\vec{z}}\left(\vec{x}\right)}{\mathtt{PDF}_{\psi}\left(J\left(\vec{x}\right)\right)} \tag{11}$$

- $\mathtt{PDF}_{\omega}$ : density of the *winning* value $\omega$
- $\mathtt{PDF}_{\vec{z}}$ : density for generating an individual by *mutation*
- $\mathtt{PDF}_{\psi}$ : density of the objective function values (Eqs. 5 or 8)

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

**Proposition 0**

$$\mathtt{PDF}_{\vec{y}}\left(\vec{x}\right) = \mathtt{PDF}_{\omega}\left(J\left(\vec{x}\right)\right) \cdot \frac{\mathtt{PDF}_{\vec{z}}\left(\vec{x}\right)}{\mathtt{PDF}_{\psi}\left(J\left(\vec{x}\right)\right)} \tag{11}$$

- $\mathtt{PDF}_{\omega}$ : density of the *winning* value $\omega$
- $\mathtt{PDF}_{\vec{z}}$ : density for generating an individual by *mutation*
- $\mathtt{PDF}_{\psi}$ : density of the objective function values (Eqs. 5 or 8)

**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. $\omega$ is set to the minimum $J_\ell$, and $\vec{y}$ is set to $\vec{x}_\ell$

# simultaneous diagonalization: $(1, \lambda)$-selection

**Proposition 1**

The covariance matrix and the Hessian commute and are simultaneously diagonalizable, when the objective function follows the quadratic approximation.

# simultaneous diagonalization: $(1, \lambda)$-selection

**Proposition 1**
The covariance matrix and the Hessian commute and are simultaneously diagonalizable, when the objective function follows the quadratic approximation.

**sketch**:
i. the covariance reads:

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

ii. apply change of variables

$$\mathcal{U}^{-1}\mathcal{H}\mathcal{U} \equiv \mathrm{diag}\left[\Delta_1, \Delta_2, \ldots, \Delta_n\right], \quad \vec{\vartheta} = \mathcal{U}^{-1}\vec{x}, \quad \mathrm{d}\vec{\vartheta} = \mathrm{d}\vec{x}$$

iii. target $\mathcal{T}_{ij} = \left(\mathcal{U}^{-1}\mathcal{C}\mathcal{U}\right)_{ij}$ and show that it vanishes for any $i \neq j$ due to symmetry considerations.

# $(\mu, \lambda)$-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 $\mu$ values from this list.
- $\vec{y}_{1:\lambda}, \ldots, \vec{y}_{\mu:\lambda}$ are their corresponding vectors.

# $(\mu, \lambda)$-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 $\mu$ 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 $\ell^{th}$-degree winners ($\ell > k$):

$$\boxed{\begin{aligned}\mathtt{PDF}_{\vec{y}_{k:\lambda},\vec{y}_{\ell:\lambda}}\left(\vec{x}_k, \vec{x}_\ell\right) = \mathtt{PDF}_{\omega_{k:\lambda},\omega_{\ell:\lambda}}\left(J\left(\vec{x}_k\right), J\left(\vec{x}_\ell\right)\right) \times \\ \times \left(\frac{\mathtt{PDF}_{\vec{z}}\left(\vec{x}_k\right)}{\mathtt{PDF}_{\psi}\left(J\left(\vec{x}_k\right)\right)}\right) \cdot \left(\frac{\mathtt{PDF}_{\vec{z}}\left(\vec{x}_\ell\right)}{\mathtt{PDF}_{\psi}\left(J\left(\vec{x}_\ell\right)\right)}\right)\end{aligned}} . \tag{12}$$

# simultaneous diagonalization: $(\mu, \lambda)$-selection

**Proposition 2**

The rank-$\mu$ covariance matrix and the Hessian commute and are simultaneously diagonalizable, when the objective function follows the quadratic approximation.

# simultaneous diagonalization: $(\mu, \lambda)$-selection

**Proposition 2**
The rank-$\mu$ covariance matrix and the Hessian commute and are simultaneously diagonalizable, when the objective function follows the quadratic approximation.

**sketch**:
i. the covariance reads (up to a factor):

$$\mathcal{C}_{ij} \propto \sum_{k<\ell\leq\mu} \int x_{k,i} x_{\ell,j} \mathtt{PDF}_{\vec{y}_{k:\lambda},\vec{y}_{\ell:\lambda}} \left(\vec{x}_k, \vec{x}_\ell\right) \mathrm{d}\vec{x}_k \mathrm{d}\vec{x}_\ell$$

ii. repeat proof steps of Proposition 1 and apply the same symmetry argumentation the inverse relation

# winning values' density & proposition 3

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

# winning values' density & proposition 3

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

$$\mathtt{CDF}_{\omega}(v) = 1 - \left(1 - \mathtt{CDF}_{\psi}(v)\right)^{\lambda} \tag{13}$$

$$\mathtt{PDF}_{\omega}(v) = \lambda \cdot \left(1 - \mathtt{CDF}_{\psi}(v)\right)^{\lambda-1} \cdot \mathtt{PDF}_{\psi}(v) \tag{14}$$

# winning values' density & proposition 3

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

$$\mathtt{CDF}_{\omega}(v) = 1 - \left(1 - \mathtt{CDF}_{\psi}(v)\right)^{\lambda} \tag{13}$$

$$\mathtt{PDF}_{\omega}(v) = \lambda \cdot \left(1 - \mathtt{CDF}_{\psi}(v)\right)^{\lambda - 1} \cdot \mathtt{PDF}_{\psi}(v) \tag{14}$$

**Proposition 3**:
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 3

Proposition 1 tells us that we may assume that both $\mathcal{H}$ and $\mathcal{C}$ are diagonalizable in the same base.

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

# proof sketch for proposition 3

i. assume $\mathcal{H}$ is diagonal and so off-diagonal of $\mathcal{C}\mathcal{H}$ vanish

ii. $\mathcal{C}_{ii} = \mathbb{E}\left[y_i^2\right] = \int x_i^2 \lambda (1 - \mathtt{CDF}_\psi(J(\vec{x})))^{\lambda-1} f(\|\vec{x}\|)$

iii. apply change of variables into $r_i = \sqrt{\Delta_i} \cdot x_i$ s.t. $\Delta_i \mathcal{C}_{ii} = c_{\mathcal{H}} \int r_i^2 \lambda (1 - \mathtt{CDF}_\psi(\|\vec{r}\|^2))^{\lambda-1} \exp\left(-\hat{J}(\vec{r})\right) \mathrm{d}\vec{r}$

iv. show that $\alpha\Delta_i\mathcal{C}_{ii} \geq 1 - \epsilon_1$ and $\alpha\Delta_i\mathcal{C}_{ii} \leq 1 + \epsilon_2$ ($\epsilon_1$ and $\epsilon_2$ tend to zero as $\lambda$ tends to infinity)

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

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

limit distributions of order statistics

# targeting $\texttt{PDF}_\omega\left(J\left(\vec{x}\right)\right)$

Using the explicit forms of $\texttt{CDF}_\psi$ and $\texttt{PDF}_\psi$, the desired density function $\texttt{PDF}_\omega\left(J\left(\vec{x}\right)\right)$ is obtained, however not in a closed form.

Next, we seek an *approximation* for $\texttt{PDF}_\omega\left(J\left(\vec{x}\right)\right)$, in order to calculate $\mathcal{C}_{ij}$ when $\lambda$ tends to infinity.

$$\mathcal{L}_\lambda\left(v\right) = 1 - \left(1 - \texttt{CDF}_\psi\left(v\right)\right)^\lambda$$

# targeting $\texttt{PDF}_\omega\left(J\left(\vec{x}\right)\right)$

Using the explicit forms of $\texttt{CDF}_\psi$ and $\texttt{PDF}_\psi$, the desired density function $\texttt{PDF}_\omega\left(J\left(\vec{x}\right)\right)$ is obtained, however not in a closed form.

Next, we seek an *approximation* for $\texttt{PDF}_\omega\left(J\left(\vec{x}\right)\right)$, in order to calculate $\mathcal{C}_{ij}$ when $\lambda$ tends to infinity.

$$\mathcal{L}_\lambda\left(v\right) = 1 - \left(1 - \texttt{CDF}_\psi\left(v\right)\right)^\lambda$$

$$\lim_{\lambda \longrightarrow \infty} \mathcal{L}_\lambda\left(v\right) = \begin{cases} 0 & \text{if } \texttt{CDF}_\psi\left(v\right) = 0 \\ 1 & \text{if } \texttt{CDF}_\psi\left(v\right) > 0 \end{cases}$$

# targeting $\mathtt{PDF}_\omega\left(J\left(\vec{x}\right)\right)$

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

Next, we seek an *approximation* for $\mathtt{PDF}_\omega\left(J\left(\vec{x}\right)\right)$, in order to calculate $\mathcal{C}_{ij}$ when $\lambda$ tends to infinity.

$$\mathcal{L}_\lambda\left(v\right) = 1 - \left(1 - \mathtt{CDF}_\psi\left(v\right)\right)^\lambda$$

$$\lim_{\lambda \longrightarrow \infty} \mathcal{L}_\lambda\left(v\right) = \begin{cases} 0 & \text{if } \mathtt{CDF}_\psi\left(v\right) = 0 \\ 1 & \text{if } \mathtt{CDF}_\psi\left(v\right) > 0 \end{cases}$$

normalization will be needed to avoid degeneracy (the distributions tend to the origin).

# *von-Mises* family of distributions

**theorem** [Fisher-Tippett]
the generalized extreme value distributions (GEVD) are the only non-degenerate family of distributions satisfying this limit:

$$\mathcal{L}_\kappa\left(v;\kappa_1,\kappa_2,\kappa_3\right)=1-\exp\left\{-\left[1+\kappa_3\left(\frac{v-\kappa_1}{\kappa_2}\right)\right]^{1/\kappa_3}\right\} \tag{15}$$

# *von-Mises* family of distributions

**theorem** [Fisher-Tippett]
the generalized extreme value distributions (GEVD) are the only non-degenerate family of distributions satisfying this limit:

$$\mathcal{L}_{\kappa}\left(v;\kappa_1,\kappa_2,\kappa_3\right)=1-\exp\left\{-\left[1+\kappa_3\left(\frac{v-\kappa_1}{\kappa_2}\right)\right]^{1/\kappa_3}\right\} \qquad (15)$$

determination of shape parameter:

$$\kappa_3=\lim_{\varepsilon\longrightarrow 0}-\log_2\frac{\texttt{CDF}_{\psi}^{-1}\left(\varepsilon\right)-\texttt{CDF}_{\psi}^{-1}\left(2\varepsilon\right)}{\texttt{CDF}_{\psi}^{-1}\left(2\varepsilon\right)-\texttt{CDF}_{\psi}^{-1}\left(4\varepsilon\right)},$$

- If $\kappa_3>0$, $\texttt{CDF}_{\psi}$ belongs to the Weibull domain,
- if $\kappa_3=0$, $\texttt{CDF}_{\psi}$ belongs to the Gumbel domain, and
- if $\kappa_3<0$, $\texttt{CDF}_{\psi}$ belongs to the Frechét domain.

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

**Proposition 4**:
For the isotropic and transformed $\chi^2$ distributions, $F_{\chi^2}\left(\psi\right),\ F_{\tau\chi^2}\left(\psi\right)$, the limits exist and read $\kappa_3 = 2/n$.

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

**Proposition 4**:
For the isotropic and transformed $\chi^2$ distributions, $F_{\chi^2}\left(\psi\right),\ F_{\tau\chi^2}\left(\psi\right)$, the limits exist and read $\kappa_3 = 2/n$.

**Corollary**:
Under the GEVD approximation for $\lambda \to \infty$, by normalizing the random variable to $\tilde{v} = \left(v - b^*_{\lambda}\right)/a^*_{\lambda}$ and using the tail-index result, $1/\kappa_3 = \frac{n}{2}$, a single winning event is described by:

$$\begin{aligned} \texttt{CDF}_{\omega}^{\text{GEVD}}\left(\tilde{v}\right) &= 1 - \exp\left(-\tilde{v}^{\,\frac{n}{2}}\right) \\ \boxed{\texttt{PDF}_{\omega}^{\text{GEVD}}\left(\tilde{v}\right) = \frac{n}{2}\tilde{v}^{\,\frac{n}{2}-1}\exp\left(-\tilde{v}^{\,\frac{n}{2}}\right)} \end{aligned} \tag{16}$$

# $\mathcal{C}_{ij}$ approximated for $(1, \lambda)$

$$\mathcal{C}_{ij} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_i x_j \frac{n}{2} \tilde{J}(\vec{x})^{\frac{n}{2}-1} \exp\left[-\tilde{J}(\vec{x})^{\frac{n}{2}}\right] \times$$
$$\times \frac{\frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}\vec{x}^T\vec{x}\right)}{\frac{\Upsilon^\eta}{\Gamma(\eta)} J(\vec{x})^{\eta-1} \exp\left(-\Upsilon J(\vec{x})\right)} \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n \tag{17}$$

# $\mathcal{C}_{ij}$ approximated for $(1,\lambda)$

$$\mathcal{C}_{ij} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_i x_j \frac{n}{2} \tilde{J}(\vec{x})^{\frac{n}{2}-1} \exp\left[-\tilde{J}(\vec{x})^{\frac{n}{2}}\right] \times \\ \times \frac{\frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}\vec{x}^T\vec{x}\right)}{\frac{\Upsilon^\eta}{\Gamma(\eta)} J(\vec{x})^{\eta-1} \exp\left(-\Upsilon J(\vec{x})\right)} \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n \tag{17}$$

$J$ is assumed here to satisfy $J(\vec{x}) = \vec{x}^T \cdot \mathcal{H} \cdot \vec{x}$ ; $a_\lambda^* = F_{\chi^2}^{-1}\left(\frac{1}{\lambda}\right)$:

# $\mathcal{C}_{ij}$ approximated for $(1, \lambda)$

$$\mathcal{C}_{ij} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_i x_j \frac{n}{2} \tilde{J}(\vec{x})^{\frac{n}{2}-1} \exp\left[-\tilde{J}(\vec{x})^{\frac{n}{2}}\right] \times \times \frac{\frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}\vec{x}^T\vec{x}\right)}{\frac{\Upsilon^\eta}{\Gamma(\eta)} J(\vec{x})^{\eta-1} \exp\left(-\Upsilon J(\vec{x})\right)} \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n \tag{17}$$

$J$ is assumed here to satisfy $J(\vec{x}) = \vec{x}^T \cdot \mathcal{H} \cdot \vec{x}$ ; $a^*_\lambda = F^{-1}_{\chi^2}\left(\frac{1}{\lambda}\right)$:

$$\boxed{\mathcal{C}_{ij} = \Phi_{\mathcal{C}} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_i x_j \left(\vec{x}^T\mathcal{H}\vec{x}\right)^{\frac{n}{2}-\eta} \times \times \exp\left[\Upsilon\vec{x}^T\mathcal{H}\vec{x} - \left(\frac{\vec{x}^T\mathcal{H}\vec{x}}{a^*_\lambda}\right)^{\frac{n}{2}} - \frac{1}{2}\vec{x}^T\vec{x}\right] \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n} \tag{18}$$

# integration

For a general positive-definite $\mathcal{H}$, the integral in Eq. 18 has an unknown closed form; it is easy to see that it commutes with $\mathcal{H}$.

# integration

For a general positive-definite $\mathcal{H}$, the integral in Eq. 18 has an unknown closed form; it is easy to see that it commutes with $\mathcal{H}$.

isotropic case $\mathcal{H} = h_0\mathbf{I}$:

$$\mathcal{C}^{(\mathcal{H}=h_0\mathbf{I})} = \frac{\Gamma(\frac{n}{2}) \cdot \Gamma\left(1+\frac{2}{n}\right) \cdot \phi\left(n\right) \cdot a_{\lambda}^{*}}{2\pi^{n/2}} \cdot \mathcal{H}^{-1} \tag{19}$$

with

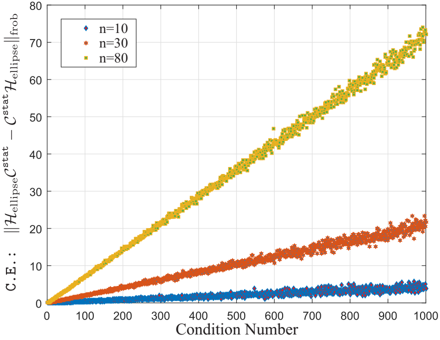
$$\phi\left(n\right) = \begin{cases} \frac{\pi^m}{m!} & n = 2m \\ \frac{2^{m+1}\pi^m}{1\cdot3\cdot5\cdots(2m+1)} & n = 2m+1 \end{cases} .$$

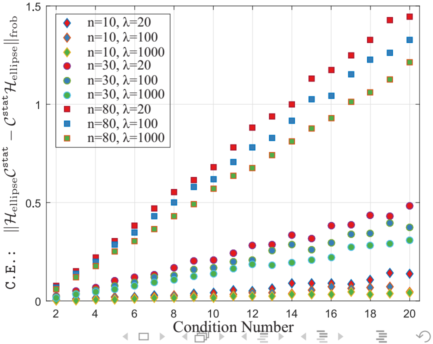
numerical validation

# eigendecomposition and commutator errors

$\{10, 30, 80\}$–dimensional separable ellipses $(\mathcal{H}_{\text{ellipse}})_{ii} = c^{\frac{i-1}{n-1}}$ with $N_{\texttt{iter}} = 10^5$.
[LEFT] $c = 2 \ldots 1000$ using $\lambda = 100$
[RIGHT] $c = 2 \ldots 20$ over $\lambda = \{20, 100, 1000\}$

Measure: $\texttt{C.E.:} \quad \|\mathcal{H}_{\text{ellipse}}\mathcal{C}^{\texttt{stat}} - \mathcal{C}^{\texttt{stat}}\mathcal{H}_{\text{ellipse}}\|_{\text{frob}}$

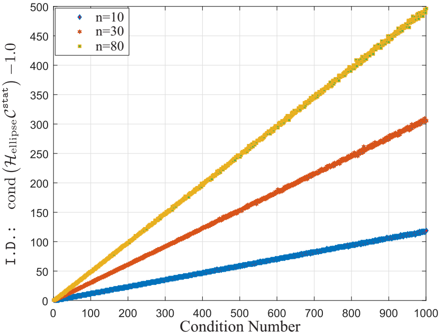


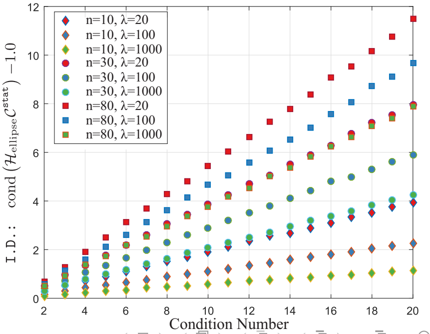



# the inverse relation under a large population

$\{10, 30, 80\}$–dimensional separable ellipses $(\mathcal{H}_{\text{ellipse}})_{ii} = c^{\frac{i-1}{n-1}}$ with $N_{\texttt{iter}} = 10^5$.
[LEFT] $c = 2 \ldots 1000$ using $\lambda = 100$
[RIGHT] $c = 2 \ldots 20$ over $\lambda = \{20, 100, 1000\}$

Measure: $\texttt{I.D.:}\quad \text{cond}\left(\mathcal{H}_{\text{ellipse}}\mathcal{C}^{\texttt{stat}}\right) - 1.0$



# statistical distributions assessment

We consider four quadratic basins of attraction:

(H-1) $n = 3$, $\mathcal{H}_1 = \left[\sqrt{2}/2\ 0.25\ 0.1;\ 0.25\ 1\ 0;\ 0.1\ 0\ \sqrt{2}\right]$

(H-2) $n = 10$, $\mathcal{H}_2 = \text{diag}\left[1.0, 1.5, \ldots, 5.5\right]$

(H-3) $n = 30$, $\mathcal{H}_3 = \text{diag}\left[\vec{I}^{10}, 2 \cdot \vec{I}^{10}, 3 \cdot \vec{I}^{10}\right]$

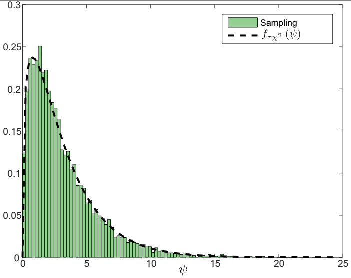
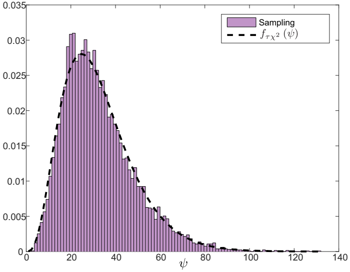
(H-4) $n = 100$, $\mathcal{H}_4 = 2.0 \cdot \mathbf{I}^{100\times 100}$

We numerically assess the following distributions:

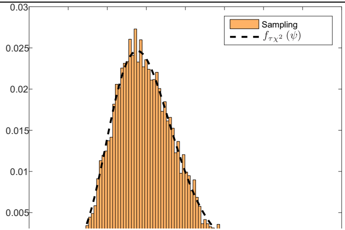
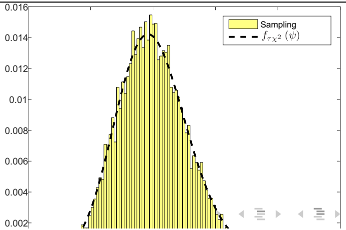
(i) density of $J(\vec{x})$ over a single iteration: $f_{\tau\chi^2}$

(ii) density of winning events: $\texttt{PDF}_{\omega}$ vs. $\texttt{PDF}_{\omega}^{\text{GEVD}}$

# (i) density of $J(\vec{x})$ over a single iteration: $f_{\tau\chi^2}$

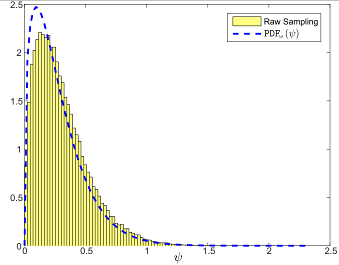
| $\mathcal{H}_1$ $(n=3)$ | $\mathcal{H}_2$ $(n=10)$ |
|---|---|
|  |  |

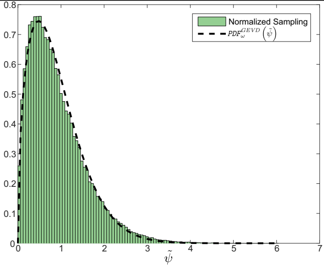


| $\mathcal{H}_3$ $(n=30)$ | $\mathcal{H}_4$ $(n=100)$ |
|---|---|
|  |  |

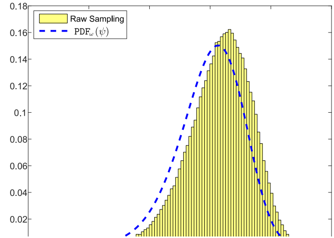
# (ii) density of winning events: $\mathtt{PDF}_{\omega}$ vs. $\mathtt{PDF}_{\omega}^{\mathrm{GEVD}}$

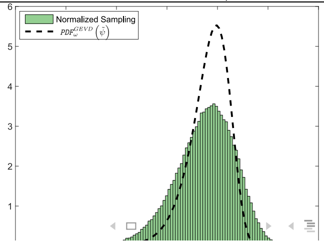
## $\mathcal{H}_1$ $(n=3)$ $\left(\lambda=20,\ N_{\mathtt{iter}}=10^5\right)$



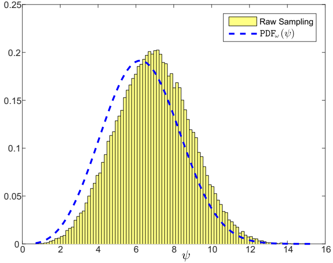

## $\mathcal{H}_3$ $(n=30)$ $\left(\lambda=1000,\ N_{\mathtt{iter}}=2\cdot 10^5\right)$

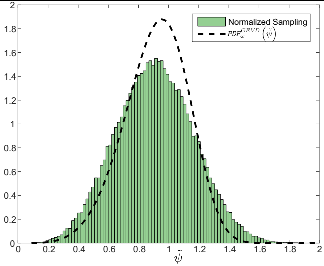


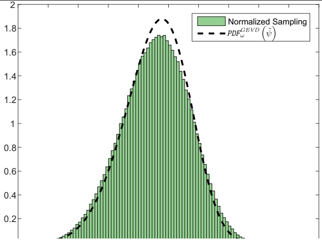
# density of winning events: $\mathcal{H}_2$ *feat.* various settings

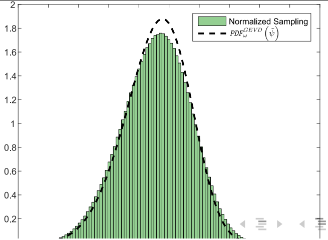
$\left(\lambda = 1000,\ N_{\texttt{iter}} = 10^5\right)$




$\left(\lambda = 10000,\ N_{\texttt{iter}} = 10^6\right)$ vs. $\left(\lambda = 20000,\ N_{\texttt{iter}} = 5 \cdot 10^6\right)$



# validating the approximated integral

for the isotropic case, $\mathcal{C}^{\texttt{stat}}$ for the 100-dimensional case (H-4) was constructed using $\lambda = 5000$ and over $5 \cdot 10^5$ iterations to obtain a diagonal with an expected value

$$0.5617 \pm 0.0012$$

Eq. 19 obtained a value of

$$0.5680$$

# validating the integral for $\mathcal{H}_1$

| | |
|---|---|
| $\mathcal{C}^{Eq41} = \begin{pmatrix} 0.1618 & -0.0367 & -0.0107 \\ -0.0367 & 0.1179 & 0.0024 \\ -0.0107 & 0.0024 & 0.0804 \end{pmatrix}$ | $\mathcal{U}^{Eq41} = \begin{pmatrix} 0.1692 & -0.4680 & 0.8674 \\ 0.0981 & -0.8677 & -0.4873 \\ 0.9807 & 0.1675 & -0.1010 \end{pmatrix}$ |
| $\mathcal{C}^{\mathtt{stat}}_{\{N_{\mathtt{iter}}=10^5\}} = \begin{pmatrix} 0.1532 & -0.0350 & -0.0104 \\ -0.0350 & 0.1120 & 0.0026 \\ -0.0104 & 0.0026 & 0.0764 \end{pmatrix}$ error = 0.0115 | $\mathcal{U}^{\mathtt{stat}}_{\{N_{\mathtt{iter}}=10^5\}} = \begin{pmatrix} 0.1726 & -0.4704 & 0.8654 \\ 0.0945 & -0.8666 & -0.4899 \\ 0.9805 & 0.1664 & -0.1051 \end{pmatrix}$ error = 0.0077 |
| $\mathcal{C}^{\mathtt{stat}}_{\{N_{\mathtt{iter}}=5\cdot 10^5\}} = \begin{pmatrix} 0.1527 & -0.0344 & -0.0102 \\ -0.0344 & 0.1116 & 0.0023 \\ -0.0102 & 0.0023 & 0.0763 \end{pmatrix}$ error = 0.0123 | $\mathcal{U}^{\mathtt{stat}}_{\{N_{\mathtt{iter}}=5\cdot 10^5\}} = \begin{pmatrix} 0.1716 & -0.4681 & 0.8669 \\ 0.0984 & -0.8674 & -0.4878 \\ 0.9802 & 0.1690 & -0.1028 \end{pmatrix}$ error = 0.0034 |
| $\mathcal{C}^{\mathtt{stat}}_{\{N_{\mathtt{iter}}=5\cdot 10^6\}} = \begin{pmatrix} 0.1530 & -0.0346 & -0.0100 \\ -0.0346 & 0.1116 & 0.0023 \\ -0.0100 & 0.0023 & 0.0760 \end{pmatrix}$ error = 0.0121 | $\mathcal{U}^{\mathtt{stat}}_{\{N_{\mathtt{iter}}=5\cdot 10^6\}} = \begin{pmatrix} 0.1662 & -0.4691 & 0.8674 \\ 0.0942 & -0.8680 & -0.4875 \\ 0.9816 & 0.1627 & -0.1001 \end{pmatrix}$ error = 0.0071 |
| $\mathcal{H}_1\mathcal{C}^{Eq41} = \begin{pmatrix} 0.1042 & 0.0038 & 0.0011 \\ 0.0038 & 0.1087 & -0.0003 \\ 0.0011 & -0.0003 & 0.1126 \end{pmatrix}$ I.D. = 0.1061 | $\mathcal{U}^{\mathcal{H}_1} = \begin{pmatrix} 0.1692 & -0.4680 & 0.8674 \\ 0.0981 & -0.8677 & -0.4873 \\ 0.9807 & 0.1675 & -0.1010 \end{pmatrix}$ error = 0.0 |

# wrapping-up

# discussion

i. $\mathcal{C}$ and $\mathcal{H}$ commute (for any $\lambda$).
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 $\lambda$).
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 $\epsilon$-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 $\epsilon$-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 $\epsilon, \lambda$ and on the spectrum of $\mathcal{H}$.

# next steps

i. what mechanisms can increase the convergence rates?

# next steps

i. what mechanisms can increase the convergence rates?
ii. analogue phenomena near a general point:

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

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

similar behavior was indeed observed in simulations.

# next steps

i. what mechanisms can increase the convergence rates?
ii. analogue phenomena near a general point:

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

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

similar behavior was indeed observed in simulations.

* we possess a proof sketch for the general case.

Acknowledgements to Jonathan Roslund.

**tak**