Comparison of Ordinal and Metric Gaussian Process Regression as Surrogate Models for CMA Evolution Strategy

Zbyněk Pitra$^{1,2,3}$, Lukáš Bajer$^{1,4}$, Jakub Repický$^{1,4}$, Martin Holeňa$^1$

$^1$Institute of Computer Science, Czech Academy of Sciences
$^2$Faculty of Nuclear Sciences and Physical Engineering
$^3$National Institute of Mental Health
$^4$Faculty of Mathematics and Physics, Charles University

Prague, Czech Republic

GECCO 2017
Contents

1 DTS-CMA-ES

2 Surrogate models
   • Metric Gaussian Processes
   • Ordinal Gaussian Processes

3 Experimental results

Z Pitra, L Bajer, J Repický, M Holeňa
Comparison Ordinal vs. Metric GP for CMA-ES
DTS-CMA-ES

Initialize: standard CMA-ES initialization with population doubled

\textbf{while} not terminate

1. CMA-ES sampling of population $x_i \sim \mathcal{N}(m, \sigma^2 C)$, for $i = 1, \ldots, \lambda$
DTS-CMA-ES

**Initialize**: standard CMA-ES initialization with population doubled

**while** not terminate

1. CMA-ES sampling of population $x_i \sim \mathcal{N}(\mathbf{m}, \sigma^2 C)$, for $i = 1, \ldots, \lambda$
2. train the first model $f_{M1}$ on the so-far original-evaluated points
DTS-CMA-ES

**Initialize:** standard CMA-ES initialization with population doubled

while not terminate

1. CMA-ES sampling of population $\mathbf{x}_i \sim \mathcal{N}(\mathbf{m}, \sigma^2 \mathbf{C})$, for $i = 1, \ldots, \lambda$
2. train the first model $f_{M_1}$ on the so-far original-evaluated points
3. get mean $\hat{\mu}_i$ and variance $\hat{s}_i^2$ of all $\mathbf{x}_i$ with the model $f_{M_1}$
4. select the most promising $\lceil \alpha \lambda \rceil$ points accord. to the model $f_{M_1}$
5. evaluate the chosen points with the original fitness $f$
6. re-train the second model $f_{M_2}$ with these new points
7. predict the fitness for the non-original-evaluated points with $f_{M_2}$
8. CMA-ES update of $\mathbf{m}$, $\sigma$, $\mathbf{C}$
DTS-CMA-ES

**Initialize**: standard CMA-ES initialization with population doubled

**while** not terminate

1. CMA-ES sampling of population $\mathbf{x}_i \sim \mathcal{N}(\mathbf{m}, \sigma^2 \mathbf{C})$, for $i = 1, \ldots, \lambda$
2. train the first model $f_{M_1}$ on the so-far original-evaluated points
3. get mean $\hat{\mu}_i$ and variance $\hat{s}_i^2$ of all $\mathbf{x}_i$ with the model $f_{M_1}$
4. select the most promising $\lceil \alpha \lambda \rceil$ points accord. to the model $f_{M_1}$

![Criterion ranking according to 1st model](image)
DTS-CMA-ES

**Initialize**: standard CMA-ES initialization with population doubled

**while** not terminate

1. CMA-ES sampling of population \( x_i \sim \mathcal{N}(m, \sigma^2 C) \), for \( i = 1, \ldots, \lambda \)
2. train the first model \( f_{M1} \) on the so-far original-evaluated points
3. get mean \( \hat{\mu}_i \) and variance \( \hat{s}_i^2 \) of all \( x_i \) with the model \( f_{M1} \)
4. select the most promising \( \lceil \alpha \lambda \rceil \) points accord. to the model \( f_{M1} \)
5. evaluate the chosen points with the original fitness \( f \)

Z Pitra, L Bajer, J Repický, M Holeňa

Comparison Ordinal vs. Metric GP for CMA-ES	3
DTS-CMA-ES

**Initialize**: standard CMA-ES initialization with population doubled

while not terminate

1. CMA-ES sampling of population $x_i \sim \mathcal{N}(m, \sigma^2 C)$, for $i = 1, \ldots, \lambda$
2. train the first model $f_{M1}$ on the so-far original-evaluated points
3. get mean $\hat{\mu}_i$ and variance $\hat{s}_i^2$ of all $x_i$ with the model $f_{M1}$
4. select the most promising $\lceil \alpha \lambda \rceil$ points accord. to the model $f_{M1}$
5. evaluate the chosen points with the original fitness $f$
6. re-train the second model $f_{M2}$ with these new points
**DTS-CMA-ES**

**Initialize:** standard CMA-ES initialization with population doubled

**while** not terminate

1. CMA-ES sampling of population $\mathbf{x}_i \sim \mathcal{N}(\mathbf{m}, \sigma^2 \mathbf{C})$, for $i = 1, \ldots, \lambda$

2. **train the first model** $f_{\mathcal{M}_1}$ on the so-far original-evaluated points

3. get mean $\hat{\mu}_i$ and variance $\hat{s}_i^2$ of all $\mathbf{x}_i$ with the **model** $f_{\mathcal{M}_1}$

4. select the most promising $\lceil \alpha \lambda \rceil$ points accord. to the **model** $f_{\mathcal{M}_1}$

5. evaluate the chosen points with the original fitness $f$

6. **re-train the second model** $f_{\mathcal{M}_2}$ with these new points

7. predict the fitness for the non-original-evaluated points with $f_{\mathcal{M}_2}$
DTS-CMA-ES

Initialize: standard CMA-ES initialization with population doubled

while not terminate

1. CMA-ES sampling of population $x_i \sim \mathcal{N}(m, \sigma^2 C)$, for $i = 1, \ldots, \lambda$
2. train the first model $f_{M1}$ on the so-far original-evaluated points
3. get mean $\hat{\mu}_i$ and variance $\hat{s}_i^2$ of all $x_i$ with the model $f_{M1}$
4. select the most promising $\lceil\alpha \lambda\rceil$ points accord. to the model $f_{M1}$
5. evaluate the chosen points with the original fitness $f$
6. re-train the second model $f_{M2}$ with these new points
7. predict the fitness for the non-original-evaluated points with $f_{M2}$
8. CMA-ES update of $m$, $\sigma$, $C$
Gaussian Process

GP is a stochastic process, where any finite collection of random variables has a joint Gaussian distribution

\[ f_{GP}(\mathbf{x}) \sim \text{GP}(\mu(\mathbf{x}), k(\mathbf{x}_1, \mathbf{x}_2)) \]

Defined by the mean function \( \mu(\mathbf{x}) \) (usually constant) and covariance function \( k(\mathbf{x}_1, \mathbf{x}_2) \) and their (hyper)parameters.
Gaussian Process

GP is a stochastic process, where any finite collection of random variables has a joint Gaussian distribution

\[ f_{\text{GP}}(\mathbf{x}) \sim \text{GP}(\mu(\mathbf{x}), k(\mathbf{x}_1, \mathbf{x}_2)) \]

Defined by the **mean function** \( \mu(\mathbf{x}) \) (usually constant) and **covariance function** \( k(\mathbf{x}_1, \mathbf{x}_2) \) and their (hyper)parameters

GP can express **uncertainty** of the prediction in a new point \( \mathbf{x} \): it gives a **probability distribution** of the output value
Given a set of $N$ training points $X_N = (x_1 \ldots x_N), \ x_i \in \mathbb{R}^d$, and corresponding measured values $y_N = (y_1, \ldots, y_N)^\top$ of a function $f$ being approximated

$$y_i = f(x_i), \quad i = 1, \ldots, N$$
given a set of \( N \) training points \( \mathbf{X}_N = (\mathbf{x}_1 \ldots \mathbf{x}_N), \mathbf{x}_i \in \mathbb{R}^d \), and corresponding measured values \( \mathbf{y}_N = (y_1, \ldots, y_N)^\top \) of a function \( f \) being approximated

\[
y_i = f(\mathbf{x}_i), \quad i = 1, \ldots, N
\]

GP considers vector of these function values as a sample from \( N \)-variate Gaussian distribution

\[
\mathbf{y}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_N)
\]
When considering a new point \((x^*, y^*)\), the prob. density of its \(f\)-values is 1D Gaussian

\[
p(y^* \mid X_N, x^*, y_N) \sim \mathcal{N}(\hat{\mu}_{N+1}, \hat{s}^2_{N+1})
\]
Gaussian Process prediction

When considering a new point \((x^*, y^*)\), the prob. density of its \(f\)-values is 1D Gaussian

\[
p(y^* \mid X_N, x^*, y_N) \sim \mathcal{N}(\hat{\mu}_{N+1}, \hat{s}^2_{N+1})
\]

with the mean and variance given by

\[
\begin{align*}
\hat{\mu}_{N+1} &= k^\top C_N^{-1} y_N, \\
\hat{s}^2_{N+1} &= \kappa - k^\top C_N^{-1} k
\end{align*}
\]

where

- \(C_N\) is GP covariance matrix – matrix of covariance function’s values \(k(x_i, x_j)\) for each pair \(x_i, x_j\)
- \(k\) is vector of covariance function’s values \(k(x^*, x_i)\) between the new point \(x^*\) and \(x_i \in X_N\)
- \(\kappa\) is the variance of the new point itself \(k(x^*, x^*)\)
Ordinal GP = Gaussian process $f_{GP}(x) \sim \text{GP}(\mu(x), k(x_1, x_2))$

- trained on ordinal values $0, 1, \ldots, r$ instead of original $f$-values (including the following transformation)
- linearly mapped via set of additional parameters $\alpha_0, \alpha, b_1, \ldots, b_{r-1}$ onto the space of ordinal values $0, 1, \ldots, r$ as

$$f_{\text{ORD}}(x) = \alpha_0 - \alpha f_{GP}(x)$$

where $-\infty = b_0 < b_1 < \cdots < b_{r-1} < b_r = \infty$. 

Z Pitra, L Bajer, J Repický, M Holeňa

Comparison Ordinal vs. Metric GP for CMA-ES
Ordinal Gaussian Processes

Training

\( (x_i, y_i)^N_{i=1} \leftarrow \mathcal{A} \)

\( \mathcal{A} \) – original data archive

\{load data from archive\}
Ordinal Gaussian Processes

Training

1. \( (x_i, y_i)^N_{i=1} \leftarrow \mathcal{A} \)
2. \( \{ y_i^{\text{ord}} \}^N_{i=1} \leftarrow \text{cluster}(\{ y_i \}^N_{i=1}, r) \)

\( \mathcal{A} \) – original data archive
\( r \) – number of cluster levels

\( y_1^{\text{ord}} \)
\( y_2^{\text{ord}} \)
\( y_3^{\text{ord}} \)

{load data from archive}
Ordinal Gaussian Processes

Training

1. $\{(x_i, y_i)\}_{i=1}^{N} \leftarrow A$
2. $\{y_{i}^{\text{ord}}\}_{i=1}^{N} \leftarrow \text{cluster}(\{y_i\}_{i=1}^{N}, r)$
3. $(\alpha, \{\beta_j\}_{j=1}^{r-1}, \theta)^* \leftarrow \arg\max_{\alpha, \{\beta_j\}_{j=1}^{r-1}, \theta} \log \hat{L}(\{y_{i}^{\text{ord}}\}_{i=1}^{N} | \{x_i\}_{i=1}^{N}, \alpha, \{\beta_j\}_{j=1}^{r-1}, \theta)$

$A$ – original data archive
$r$ – number of cluster levels
$\alpha, \alpha_0$ – linear mapping parameters
$\beta_i = \alpha_0 + b_i$
$\theta$ – latent GP hyperparameters

$\hat{L}$ – log-likelihood

model trained through likelihood maximization

ordinal GP model

$b_3 = \infty$
$b_2$
$b_1$
$b_0 = -\infty$
$I_3$
$I_2$
$I_1$
Ordinal Gaussian Processes

Prediction

\[ \{x_i\}_{i=1}^{\lambda} \text{ – population to predict} \]
Ordinal Gaussian Processes

Prediction

\[ p_{i,k} \leftarrow P(f(x_i) \in I_k | x_i, \alpha, \{\beta_j\}_{j=1}^{r-1}, \theta) \]

\[ \forall k = 1, \ldots, r, \forall i = 1, \ldots, \lambda \]

\{x_i\}_{i=1}^\lambda \text{ – population to predict} \\
\( r \) \text{ – number of cluster levels} \\
\( \alpha, \alpha_0 \) \text{ – linear mapping parameters} \\
\( \beta_i = \alpha_0 + b_i \) \\
\( \theta \) \text{ – latent GP hyperparameters}
Ordinal Gaussian Processes

Prediction

1. \( p_{i,k} \leftarrow P(f(x_i) \in I_k | x_i, \alpha, \{\beta_j\}_{j=1}^{r-1}, \theta) \)
2. \( q_i \leftarrow \sum_{k=1}^{r} p_{i,k} \)

\( \{x_i\}_{i=1}^{\lambda} \) – population to predict  
\( r \) – number of cluster levels  
\( \alpha, \alpha_0 \) – linear mapping parameters  
\( \beta_i = \alpha_0 + b_i \)  
\( \theta \) – latent GP hyperparameters

Mapping a new population to intervals using probability  
Weighted prediction
Ordinal Gaussian Processes

Prediction

1. \( p_{i,k} \leftarrow P(f(\mathbf{x}_i) \in I_k | \mathbf{x}_i, \alpha, \{\beta_j\}_{j=1}^{r-1}, \theta) \) \quad \forall k = 1, \ldots, r, \forall i = 1, \ldots, \lambda

2. \( q_i \leftarrow \sum_{k=1}^{r} p_{i,k} \)

3. \( \{\mathbf{x}_{i:\lambda}\}_{i=1}^{\lambda} \leftarrow \text{order } \{\mathbf{x}_i\}_{i=1}^{\lambda} \text{ according to } q_1:1 \leq q_2:1 \leq \cdots \leq q_{\lambda}:1 \)

\( \{\mathbf{x}_i\}_{i=1}^{\lambda} \) – population to predict
\( r \) – number of cluster levels
\( \alpha, \alpha_0 \) – linear mapping parameters
\( \beta_i = \alpha_0 + b_i \)
\( \theta \) – latent GP hyperparameters
Experimental settings

- Noiseless part of the BBOB
- 100 FE/D budget
- Algorithms
  - CMA-ES
  - DTS-CMA-ES
  - Ord-N-DTS – no clustering
  - Ord-Q-DTS – quantile-based clustering
  - Ord-H-DTS – agglomerative hierarchical clustering
Experimental settings

- Noiseless part of the BBOB
- 100 FE/D budget
- Algorithms
  - CMA-ES
  - DTS-CMA-ES
  - Ord-N-DTS – no clustering
  - Ord-Q-DTS – quantile-based clustering
  - Ord-H-DTS – agglomerative hierarchical clustering
- Ordinal settings
  - $\lambda$ ordinal levels
  - Matérn GP kernel
Experimental results on BBOB (2 D)

- bbob - f1-f24, 2-D
- 31 target RLs/dim: 0.5..50
- from refalgs/best2009-bbob.tar.gz
- 15 instances

Comparison Ordinal vs. Metric GP for CMA-ES
Experimental results on BBOB (5 D)

Proportion of function+target pairs

log10 of (# f-evals / dimension)

bbob - f1-f24, 5-D
31 target RLs/dim: 0.5..50
from refalgs/best2009-bbob.tar.gz
15 instances

Ord-N-DTS
Ord-H-DTS
Ord-Q-DTS
DTS-CMA-E
CMA-ES

best 2009

Comparison Ordinal vs. Metric GP for CMA-ES
Experimental results on BBOB (10 D)

- bobb - f1-f24, 10-D
- 31 target RLs/dim: 0.5..50
- from refalgs/best2009-bbob.tar.gz
- 15 instances

- Ord-H-DTS
- Ord-N-DTS
- Ord-Q-DTS
- DTS-CMA-E
- best 2009

- CMA-ES

Comparison Ordinal vs. Metric GP for CMA-ES
ECDF results on the whole BBOB (5 D)

**Separable**
- bobb - f1-f5, 5-D
- 31 target RLs/dim: 0.5..50
- from refalgs/best2009-bbbo.tar.gz
- 15 instances

**Moderate**
- bobb - f6-f9, 5-D
- 31 target RLs/dim: 0.5..50
- from refalgs/best2009-bbbo.tar.gz
- 15 instances

**Ill-conditional**
- bobb - f10-f14, 5-D
- 31 target RLs/dim: 0.5..50
- from refalgs/best2009-bbbo.tar.gz
- 15 instances

**Multi-modal**
- bobb - f15-f19, 5-D
- 31 target RLs/dim: 0.5..50
- from refalgs/best2009-bbbo.tar.gz
- 15 instances

**Weakly structured multi-modal**
- bobb - f20-f24, 5-D
- 31 target RLs/dim: 0.5..50
- from refalgs/best2009-bbbo.tar.gz
- 15 instances
Results on f6 and f22

2 3 5 10 20 40
0
1
2
3
6 Attractive sector

15 instances
target RL/dim: 10

Ord-N-DTS
Ord-H-DTS
Ord-Q-DTS
DTS-CMA-ES
CMA-ES

2 3 5 10 20 40
0
1
2
3
22 Gallagher 21 peaks

15 instances
target RL/dim: 10

Ord-N-DTS
Ord-H-DTS
Ord-Q-DTS
DTS-CMA-ES
CMA-ES

Z Pitra, L Bajer, J Repický, M Holeňa

Comparison Ordinal vs. Metric GP for CMA-ES
Conclusions

- Effect of different clustering methods not crucial
- Performance of the ordinal GP models is considerably lower than the standard GP models with few exceptions (e.g., attractive sector $f_6$)
- Further investigation:
  - Adaptive switch between metric and ordinal models
Thank you!

z.pitra@gmail.com  
bajeluk@gmail.com  
j.repicky@gmail.com  
martin@cs.cas.cz