Benchmarking the Novel CMA-ES Restart Strategy Using the Search History on the BBOB Noiseless Testbed

Takahiro Yamaguchi   Youhei Akimoto

Shinshu University, Japan

Introduction: CMA-ES with Restart Strategy

Covariance Matrix Adaptation Evolution Strategy (CMA-ES)

1. Generate candidate solutions $(x_i^{(t)})_{i=1,2,\ldots,\lambda}$ from $\mathcal{N}(m^{(t)}, (\sigma^{(t)})^2 \mathbf{C}^{(t)})$
2. Evaluate $f(x_i^{(t)})$ and sort them, $f(x_{1:\lambda}) < \cdots < f(x_{\lambda:\lambda})$.
3. Update the distribution parameters $\theta^{(t)} = (m^{(t)}, (\sigma^{(t)})^2 \mathbf{C}^{(t)})$ using the ranking of candidate solutions.

Restart strategies: almost necessities for multimodal black-box functions.

- increasing the population size: helpful for multimodal functions with well global structure
- decreasing the initial step-size: helpful for multimodal functions with weak global structure

Introduction: CMA-ES with Restart Strategy

Existing (successful) restart strategies:

IPOP: Doubles the population size every restart

effective on well-structured multimodal functions

BIPOP: IPOP regime + LS regime (start with a smaller step-size)

effective on well-structured multimodal functions (IPOP regime)

effective on weak-structured multimodal functions (LS regime)

Our Proposal: Utilizing the Search History

- to early stop overlapping restarts (new termination criterion)
- to shrink the initial step-size to prevent overlapping restarts

Search History

record the distribution parameters

History of Normalized Parameters

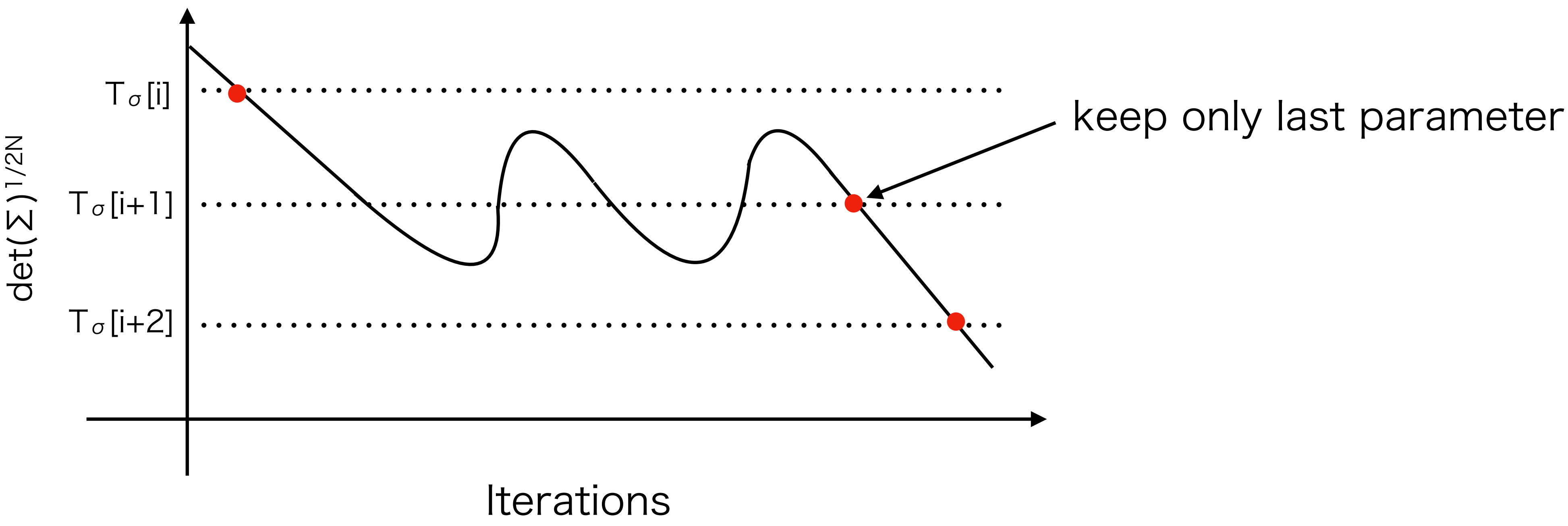
History of Normalized Distribution Parameters (m, Σ)

m: the mean vector

Σ: the normalized covariance matrix $\sigma^2 C / \alpha$ (α: normalization factor)

When to Record the Parameters?

- predefined target of $\det(\Sigma)^{1/2N}$: $T_\sigma = \det(\Sigma^{(0)})^{1/2N} \times [1, 10^{-1}, ...,]$
- every time $\det(\Sigma)^{1/2N}$ crosses $T_\sigma[i]$ from above



History of Normalized Parameters

After J restarts

#Restart	$T_\sigma[0]$	$T_\sigma[1]$	...	$T_\sigma[k]$	$T_\sigma[k+1]$	...	$T_\sigma[n_\sigma-1]$
1	(m, Σ)	(m, Σ)		(m, Σ)	(m, Σ)		(m, Σ)
2	(m, Σ)	(m, Σ)		(m, Σ)	-		-
⋮							
J	(m, Σ)	(m, Σ)		(m, Σ)	(m, Σ)		(m, Σ)

- at most J entries for each target $T_\sigma[k]$
- some entries are missing due to early termination

Termination Criterion Using Search History

detect and stop overlapping restarts

Termination Criterion: Basic Idea

After J restarts

#Restart	$T_\sigma[0]$	$T_\sigma[1]$	...	$T_\sigma[k]$	$T_\sigma[k+1]$	...	$T_\sigma[n_\sigma-1]$
1	(m, Σ)	(m, Σ)		(m, Σ)	(m, Σ)		(m, Σ)
2	(m, Σ)	(m, Σ)		(m, Σ)	-		-
⋮							
J	(m, Σ)	(m, Σ)		(m, Σ)	(m, Σ)		(m, Σ)
J+1st Restart	(m, Σ)	(m, Σ)	terminate!				

#Restart	$T_\sigma[0]$	$T_\sigma[1]$	...	$T_\sigma[k]$	$T_\sigma[k+1]$	...	$T_\sigma[n_\sigma-1]$
1	close	far					
2	far	close					
⋮							
J	far	close					

- check if the current distribution is sufficiently close to the history
- terminate if they are regarded as close to the history $n_{\mathrm{KL}}^{\mathrm{check}}$ times in a row

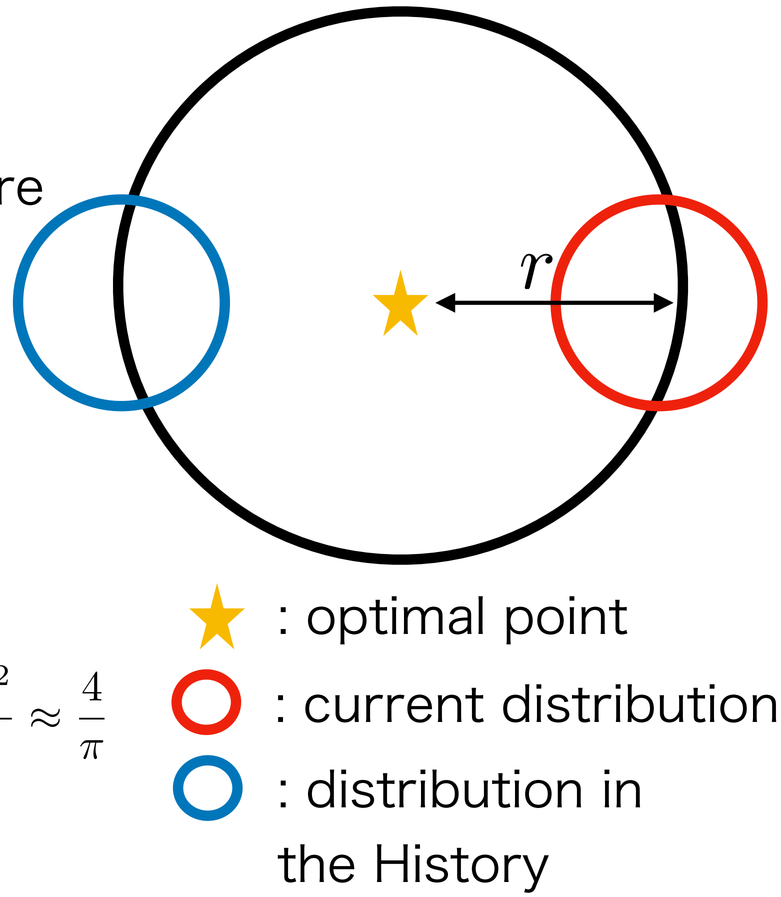
Termination Criterion: Similarity Check by KL-divergence

KL-divergence

$$D_{\mathrm{KL}}(\mathcal{N}_0 \parallel \mathcal{N}_1) = \frac{1}{2}\{(m_1 - m_0)^{\mathrm{T}}\Sigma_1^{-1}(m_1 - m_0) + \mathrm{Tr}(\Sigma_0^{-1}\Sigma_1) - N + \ln\det(\Sigma_0^{-1}\Sigma_1)\}$$

Threshold for KL-divergence

- We want to detect if two distributions are optimizing the same Sphere



KL-divergence on Sphere

- Optimal Step-Size Case

$$D_{\mathrm{KL}}(\mathcal{N}_0 \parallel \mathcal{N}_1) = \frac{1}{2}(m_1 - m_0)^{\mathrm{T}}\Sigma_1^{-1}(m_1 - m_0)$$
$$= \frac{N^2\alpha^2\|m_1 - m_0\|^2}{2\beta^2\mu_w^2 f(m_1)} = \frac{2N^2\alpha^2 f(m_1)}{\beta^2\mu_w^2 f(m_1)} = \frac{2}{\beta^2}\frac{N^2\alpha^2}{\mu_w^2} \approx \frac{4}{\pi}$$

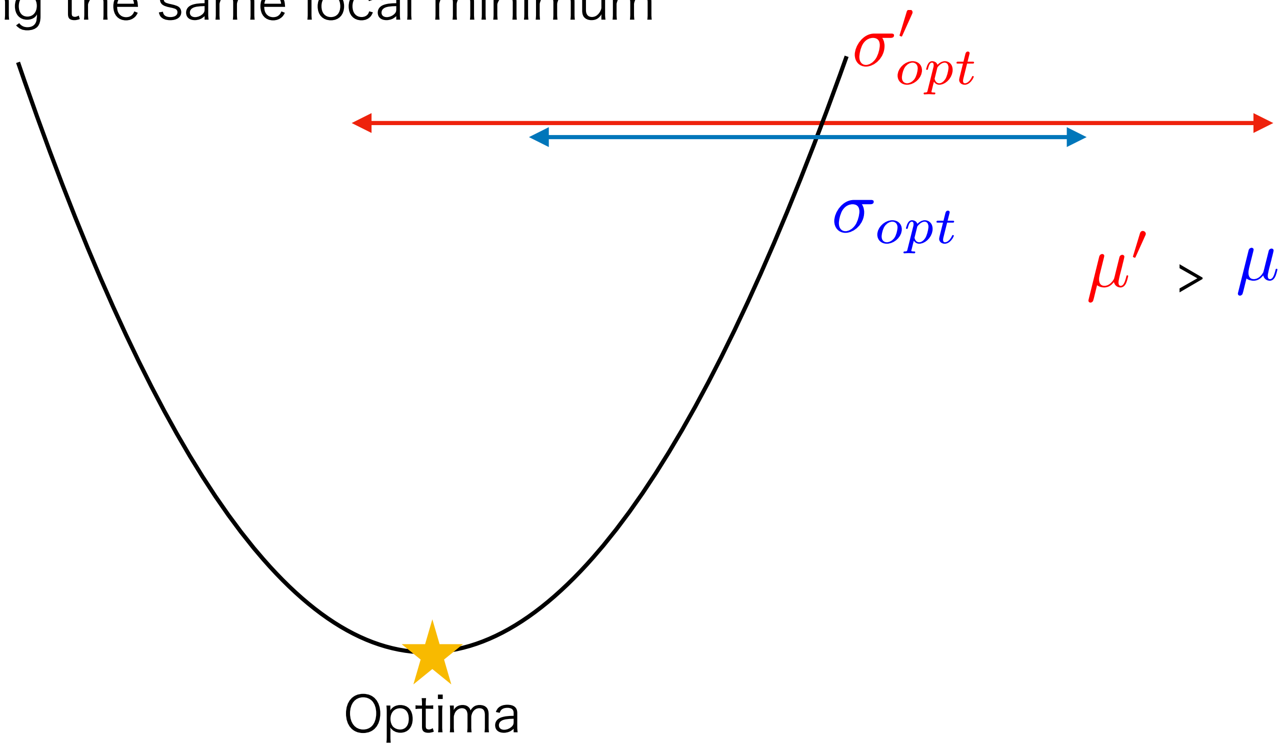
Based on this derivation, we set $\delta_{\mathrm{KL}}^{\mathrm{thre}} = 2$.

- regarded as close if KL(N_0 || N_1) ≦ $\delta_{\mathrm{KL}}^{\mathrm{thre}} = 2$

Normalization Factor α

Comparing two distributions with different population size?

- optimal step-size depends on the population size
- distributions won't be close
  even if they are searching the same local minimum



Normalized Parameter

$$\Sigma = \frac{\sigma^2}{\alpha^2}\mathbf{C} \quad \text{where} \quad \alpha = \frac{\mu_w}{N - 1 + \mu_w}$$

- reflect $\sigma^* \propto \mu_w = 1/\sum_{I=1}^{\lambda} w_i^2$ if $\mu_w \leqq N$
- reflect σ^* tends to constant if $\mu_w \geqq N$

Initial Step-Size Selection Using Search History

shrink the initial step-size
to prevent the overlapping search

Initial Normalized Step-Size Selection

Initial Step-Size Selection in BIPOP

- first run: $\sigma^{(0)}$
- IPOP regime: $\sigma^{(0)}$
- LS regime: $\sigma^{(0)} \times r$, r: random in (0.01, 1)

When to shrink the initial (normalized) step-size?

- overlapping restarts observed n_σ^{dec} times in a row
    - the current initial step-size is regarded as too large to escape from already searched big valley

How to shrink the initial (normalized) step-size?

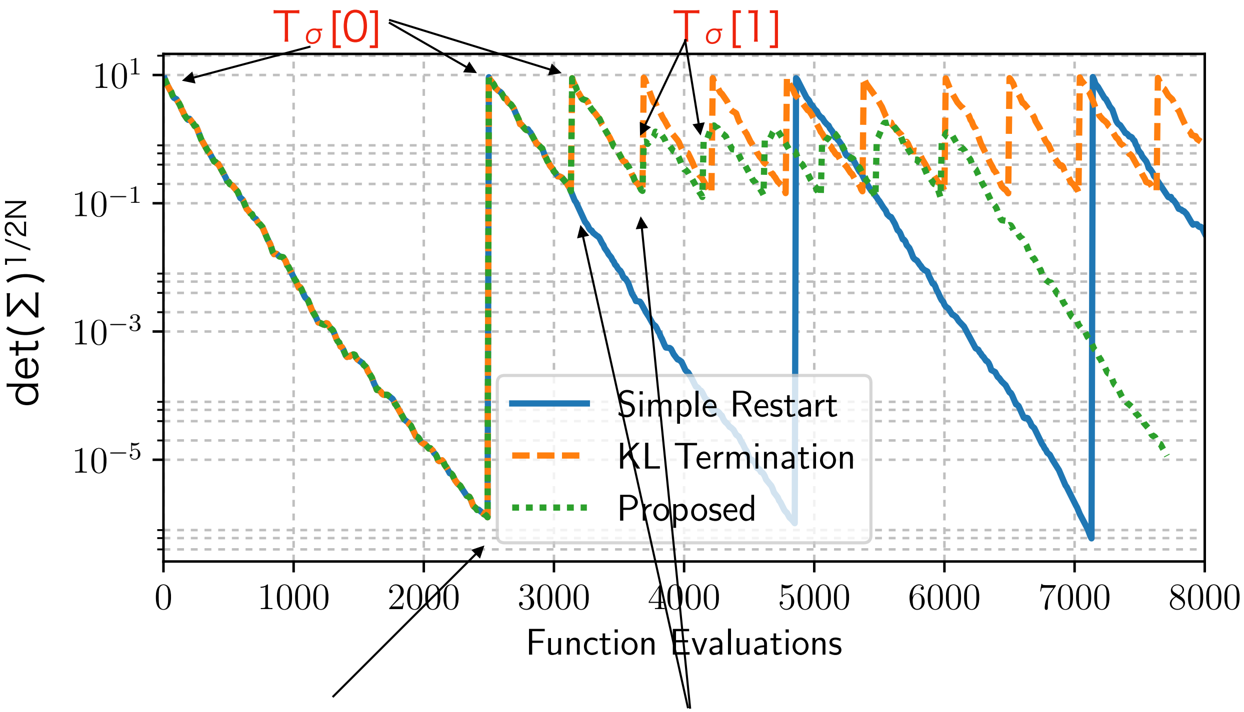
- first run: $\sigma^{(0)} = \alpha \times T_\sigma[0]$
- J^{th} run: $\sigma^{(0)} = \alpha \times T_\sigma[i]$
    - if overlapping restarts observed n_σ^{dec} times in a row: $\sigma^{(0)} = \alpha \times T_\sigma[i+1]$
    - if not: $\sigma^{(0)} = \alpha \times T_\sigma[i]$

Demonstration on Double-Sphere

$$f(x) = \min(a^2\|x_o\|^2,\ \|x_l\|^2 + 1.0)$$

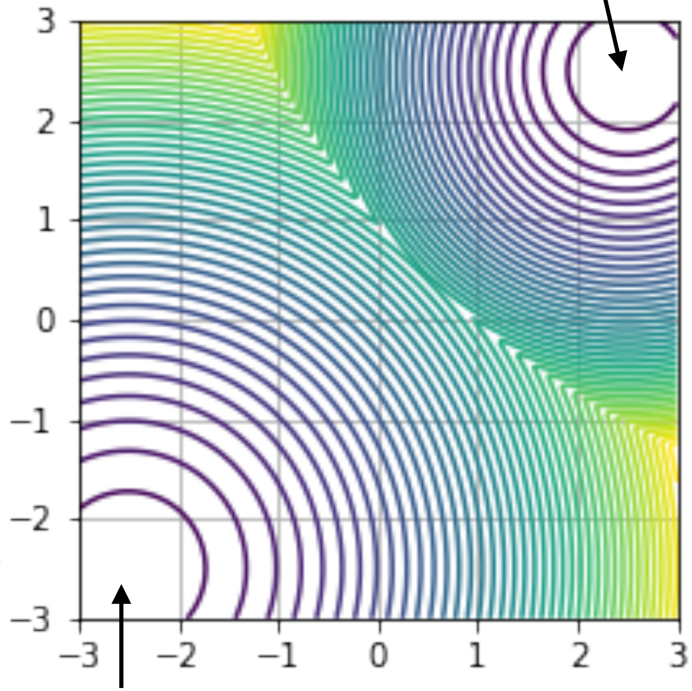
$a = 1.5,\ x_o = x - [2.5, \ldots, 2.5]$ and $x_l = x + [2.5, \ldots, 2.5]$

smaller basin of attraction for global optimum



terminated due to convergence criterion

terminated due to proposed criterion



lager basin of attraction for local optimum

BBOB Results

Termination Criteria

Termination Condition (Convergence Criteria)

tolf: $\text{median}(fiqr\_hist) < 10^{-11}$

tolfrel: $\text{median}(fiqr\_hist) < 10^{-12} * \text{abs}(\text{median}(fmin\_hist))$

- the objective function value differences are too small to sort them without being affected by numerical errors.

tolx: $\text{median}(xiqr\_hist) < 10^{-11}$

tolxrel: $\text{median}(xiqr\_hist) < 10^{-12} * \text{abs}(\text{median}(xmin\_hist))$

- the coordinate value differences are too small to update parameters without being affected by numerical errors.

Restart Scheme

	KL-Restart	KL-IPOP	KL-BIPOP	IPOP	BIPOP
Pop. Size	fixed λ_{def}^{2}	IPOP	BIPOP	IPOP	BIPOP
Init. σ	proposed	proposed	proposed	fixed	BIPOP
termination	convergence +proposed	convergence +proposed	convergence +proposed	convergence	convergence

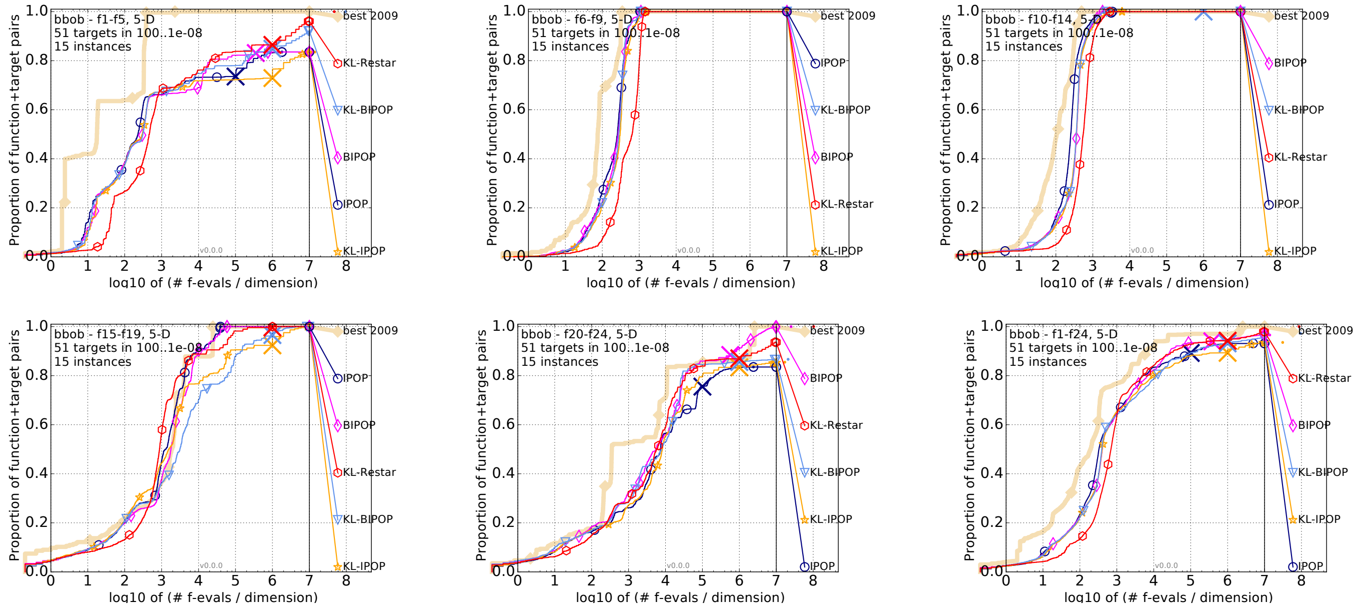
Maximum pop. size is $2^8 \times \lambda_{def}$ for IPOP regime

For each (re-)start of the algorithm, we initialize the mean vector $m \sim \mathcal{U}[-4, 4]^D$ and the covariance matrix $C = 2^2 I$. The maximum #f-call set to $10^6 D$.

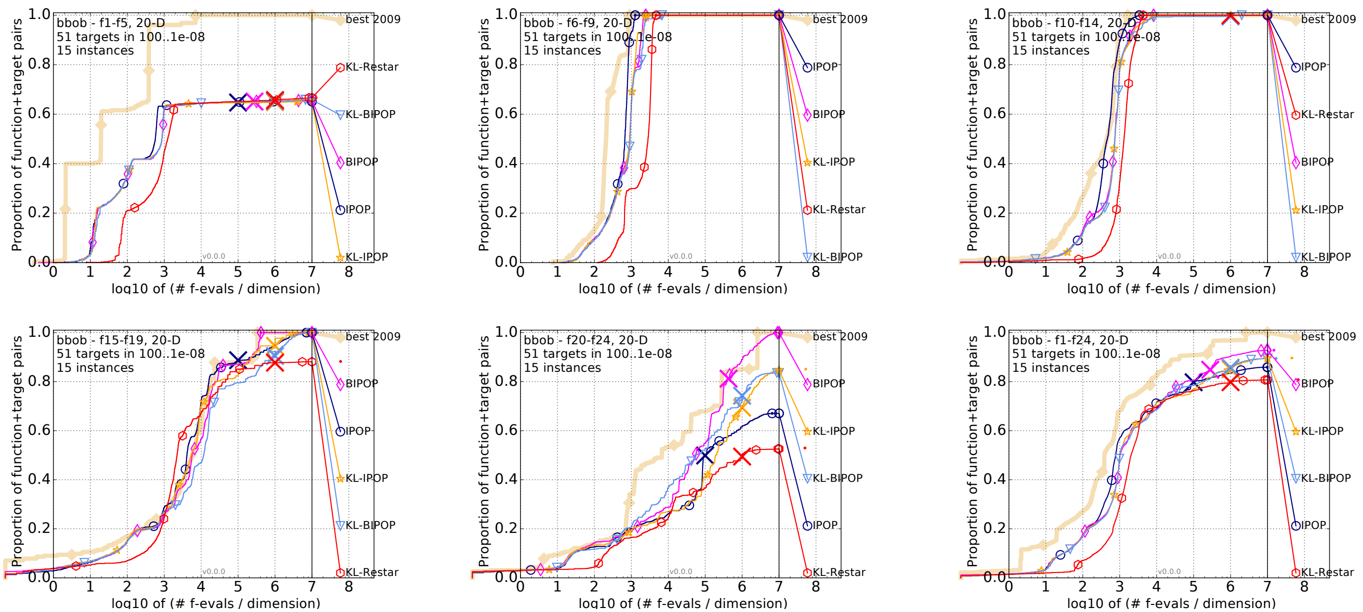
Data for IPOP and BIPOP are downloaded from the web page

Results on 5D and 20D

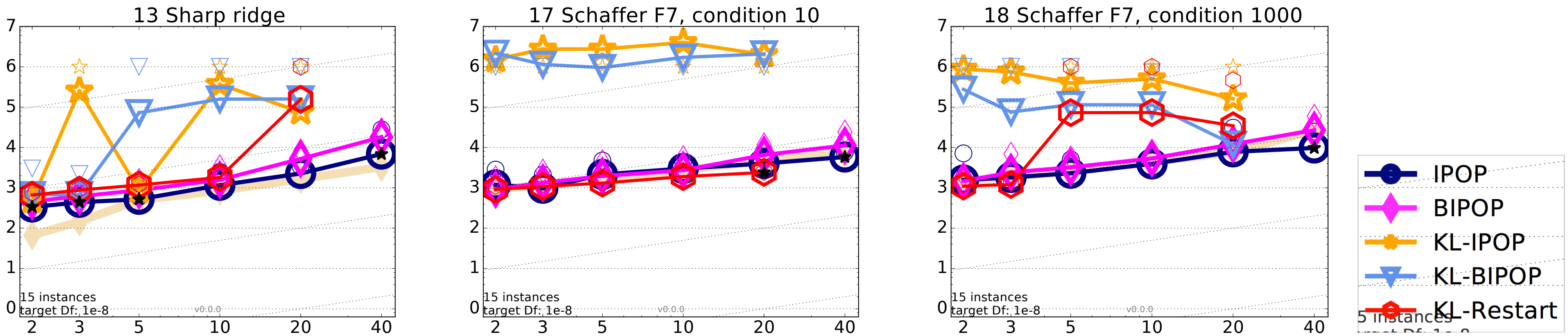
5D



20D

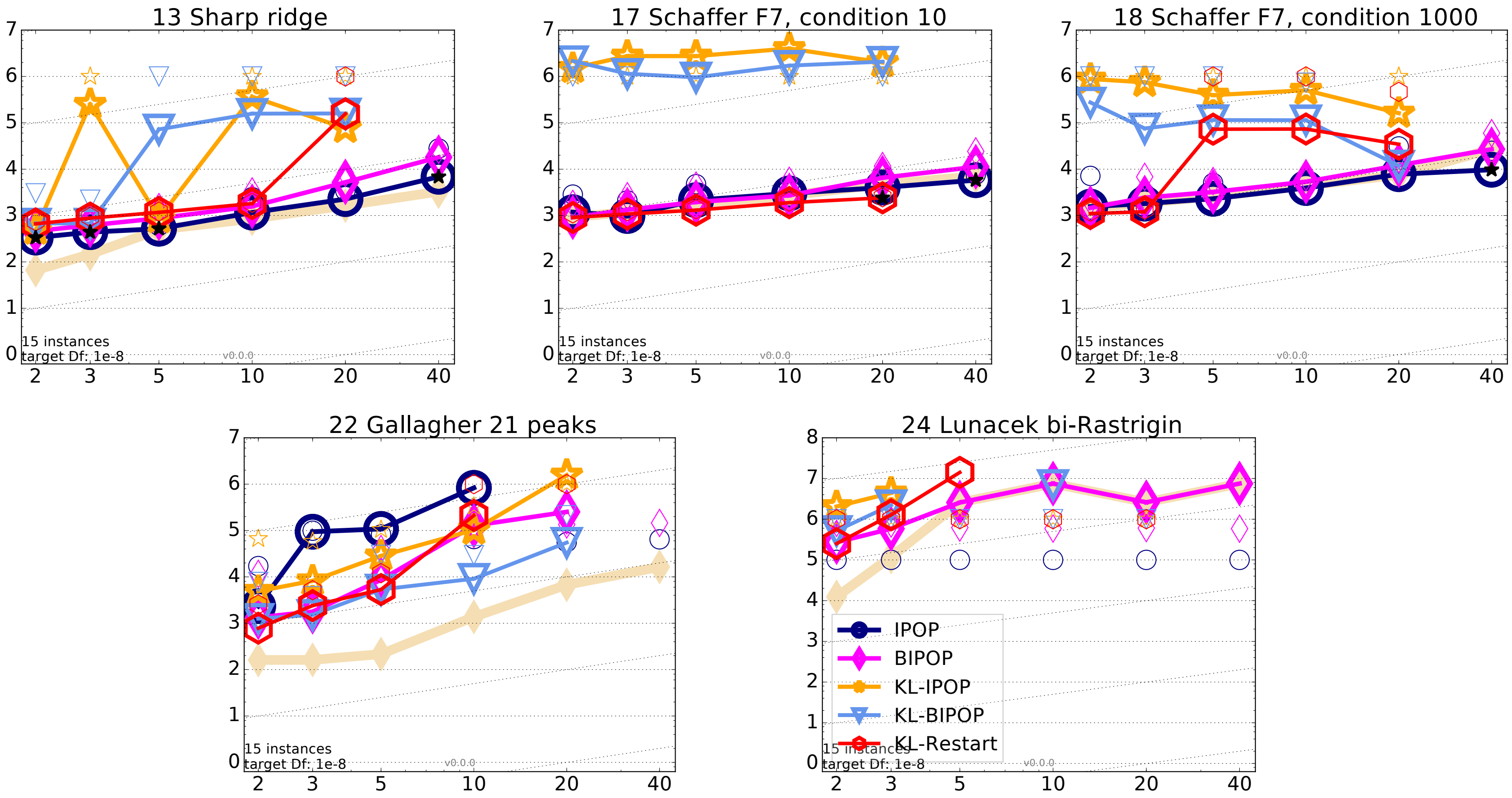


KL-Restart vs KL-IPOP vs KL-BIPOP



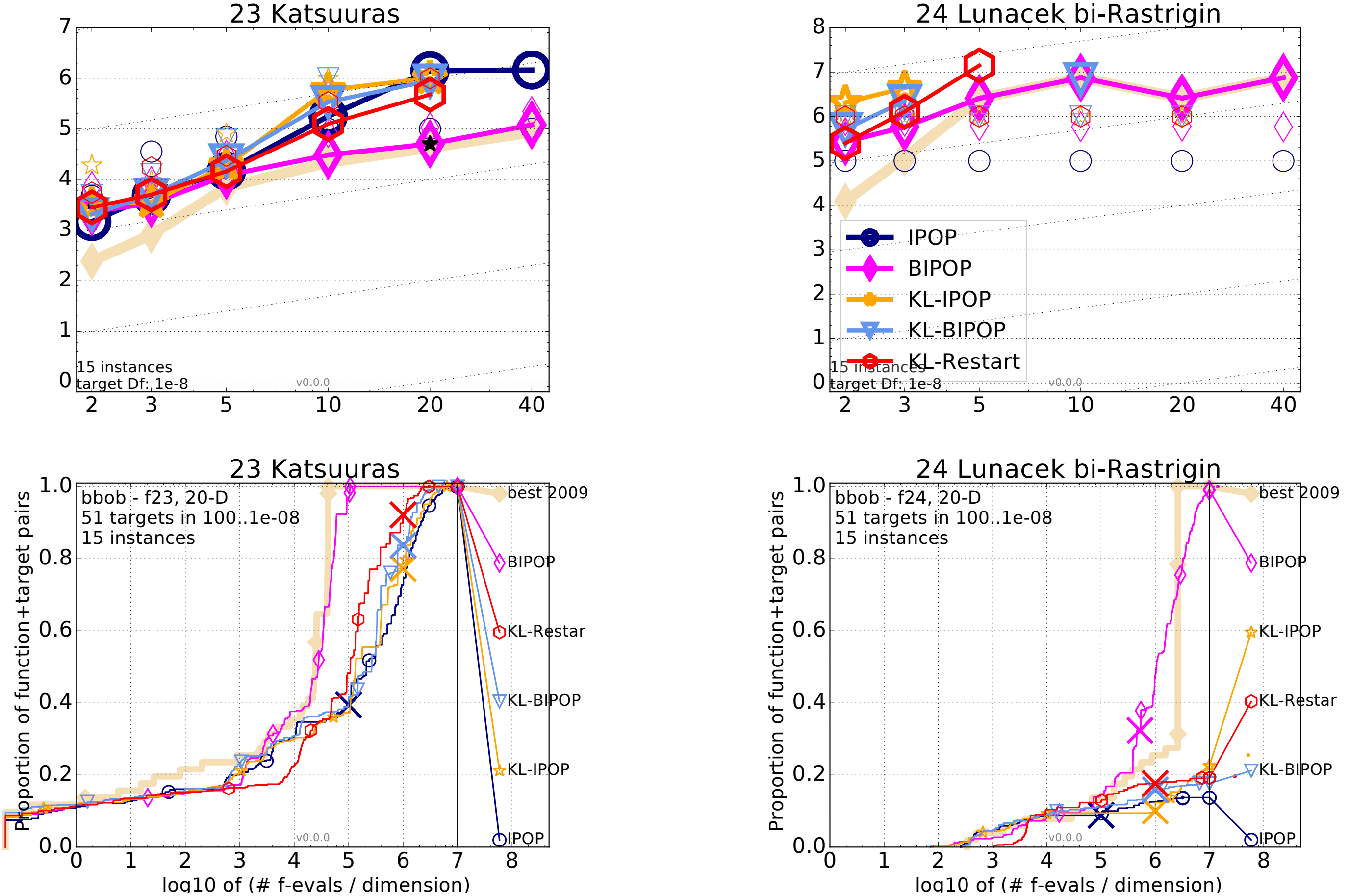
- KL-Restart: more FEs on unimodal functions due to the pop. size.
- f_{13}, f_{17} and f_{18}: KL-IPOP and KL-BIPOP suffered from early termination, while KL-Restart often finds the target function value at the first (re-)start, hence it works better than KL-IPOP and KL-BIPOP.

KL-IPOP vs IPOP



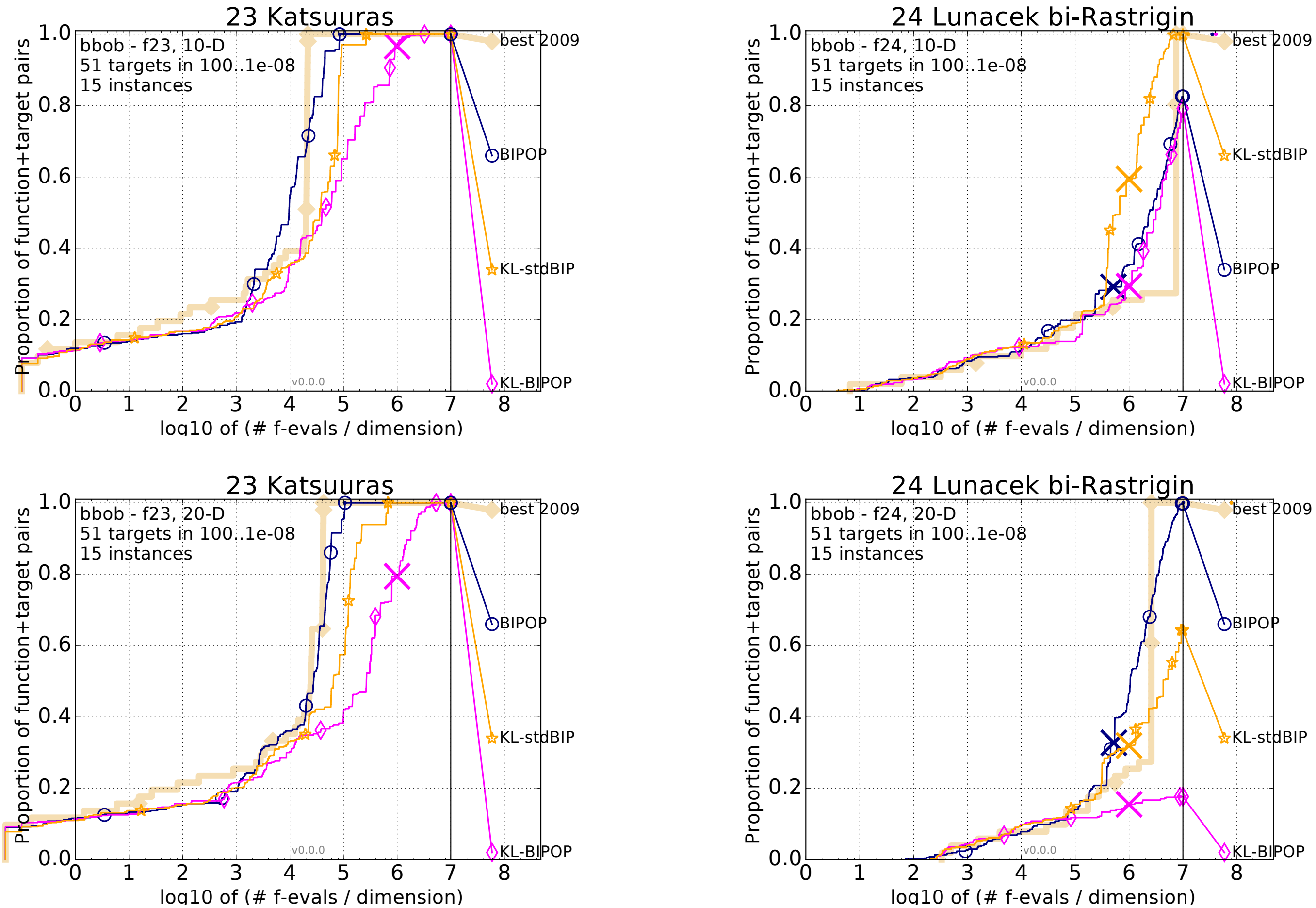
- KL-IPOP solved f_{22}, and f_{24} for N ≦ 5 with fewer number of FEs than IPOP
- IPOP is significantly better on f_{13}, f_{17} and f_{18} than KL-IPOP
  - due to too early stopping

KL-BIPOP vs BIPOP



- difference between KL-BIPOP and BIPOP is similar to the difference between KL-IPOP and IPOP
- on f_{23} and f_{24}, BIPOP is superior to KL-BIPOP

KL-BIPOP vs BIPOP



KL-stdBIPOP: BIPOP + proposed termination mechanism

- KL-stdBIPOP performs better than KL-BIPOP on f_{23} and f_{24}
  - problem of KL-BIPOP on f_{23} and f_{24} is due to init. σ selection mechanism

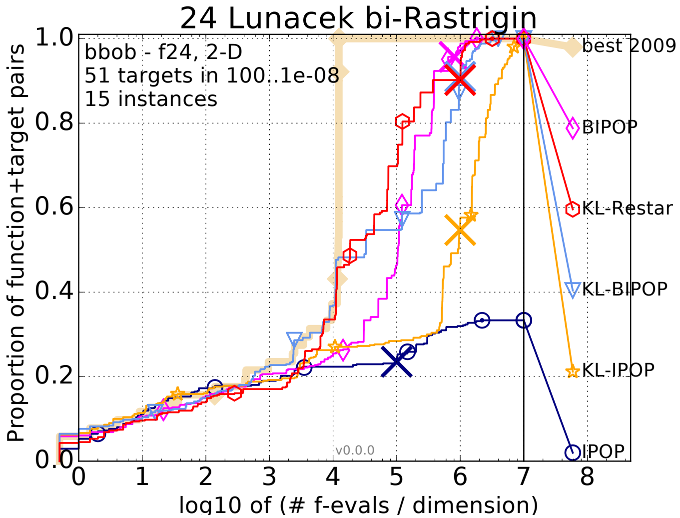
Conclusion

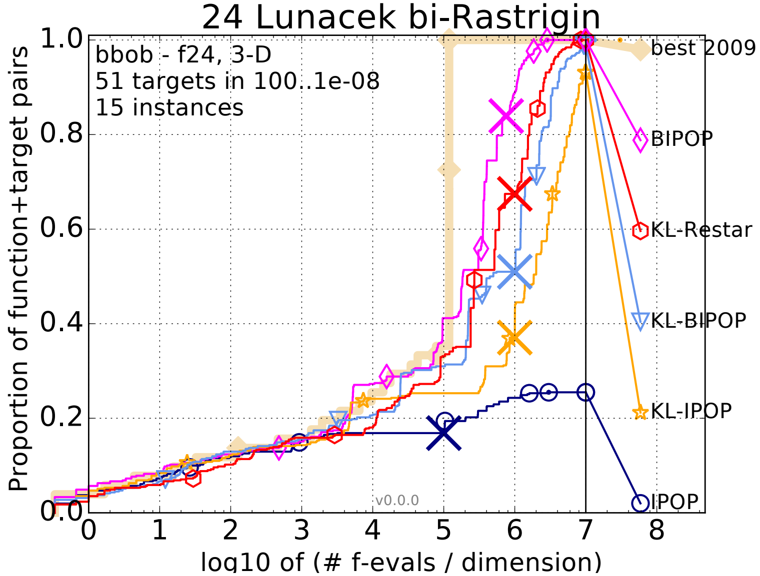
Advantage

- promising performance on f_{22} (21 peak): weak global structure with a relatively small number of local minima

Disadvantage

- too early stopping on f_{13}(sharp ridge), f_{17}, f_{18} (Schaffer)
  - termination criterion needs to be improved
- initial step-size control mechanism not properly working for f_{23} and f_{24}
  - initial step-size control mechanism needs to be improved


24 Lunacek bi-Rastrigin
bbob - f24, 2-D
51 targets in 100..1e-08
15 instances
Proportion of function+target pairs
log10 of (# f-evals / dimension)
best 2009
BIPOP
KL-Restar
KL-BIPOP
KL-IPOP
IPOP
v0.0.0


24 Lunacek bi-Rastrigin
bbob - f24, 3-D
51 targets in 100..1e-08
15 instances
Proportion of function+target pairs
log10 of (# f-evals / dimension)
best 2009
BIPOP
KL-Restar
KL-BIPOP
KL-IPOP
IPOP
v0.0.0