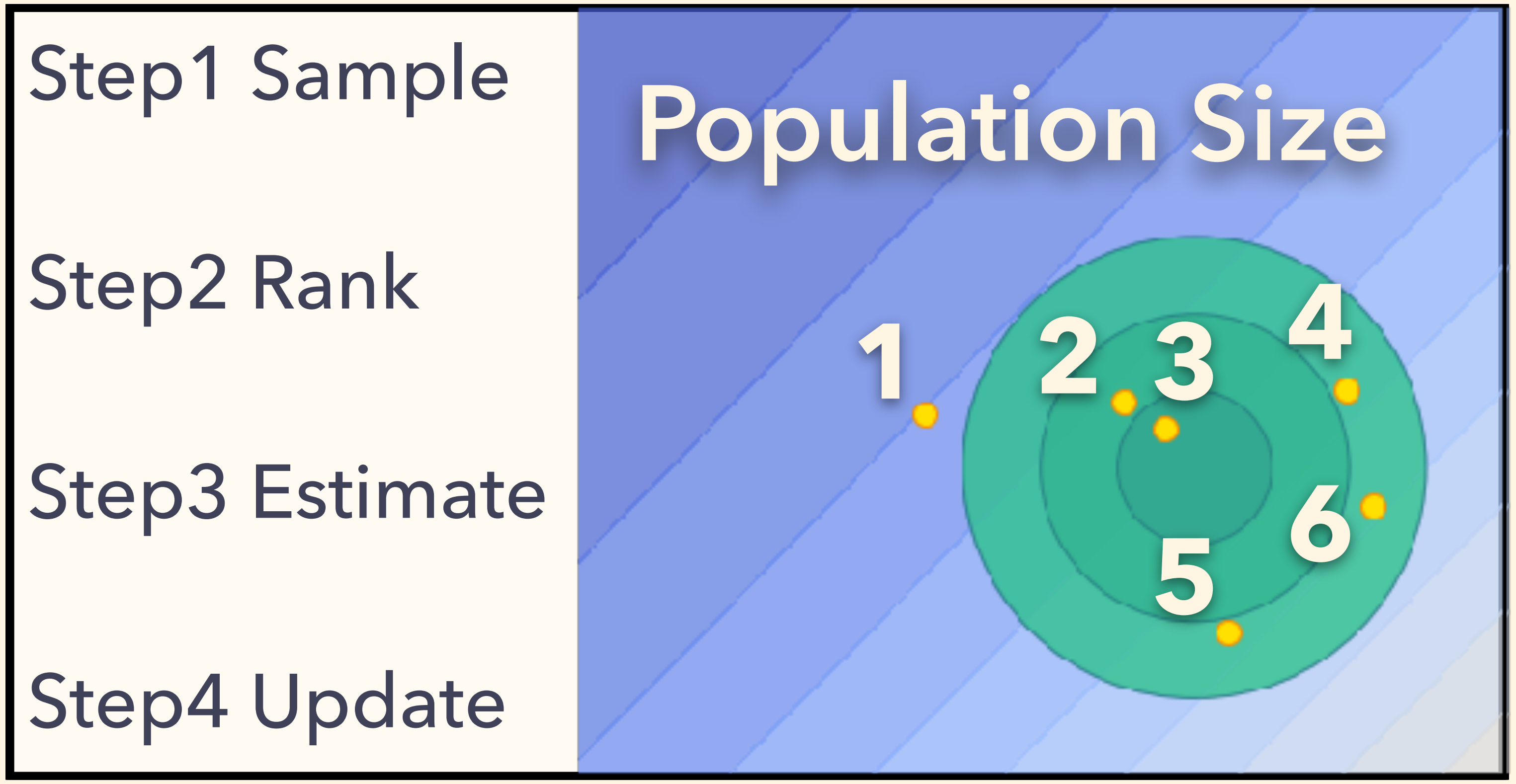
Benchmarking the PSA-CMA-ES on the BBOB Noiseless Testbed

Kouhei Nishida,   Youhei Akimoto
Shinshu University,   University of Tsukuba

CMA-ES

- It maintains a multivariate normal distribution $\mathcal{N}(m, \Sigma)$



$$\Sigma = \sigma^2 C$$

m: mean vector
σ : step-size
C: covariance matrix

- All of its hyper-parameters have their default values
  i.e. the learning rate, the population size
- The population size needs tuning
  if the objective function is a noisy or multimodal function
  [Hansen 2004]

CMA-ES: Population Size Tuning

Approach to Avoid Tuning by Users

- To utilize a multi-run strategy with different population sizes
- To adapt the population size

BIPOP-CMA-ES

First run: CMA-ES with the default population size

→ unimodal functions

Additional runs:

- CMA-ES with an increased population size
  → well-structured multimodal or noisy functions

- CMA-ES with a relatively small step-size and population size
  → weakly-structured multimodal functions

CMA-ES: Population Size Tuning

Approach to Avoid Tuning by Users

- To utilize a multi-run strategy with different population sizes
- To adapt the population size

PSA-CMA-ES [Nishida2018, Thursday 19, ENUM4]

- Based on tendency of the parameter update

Key Observation
On multimodal functions and noisy functions, the parameter update has less tendency than on noiseless unimodal functions.

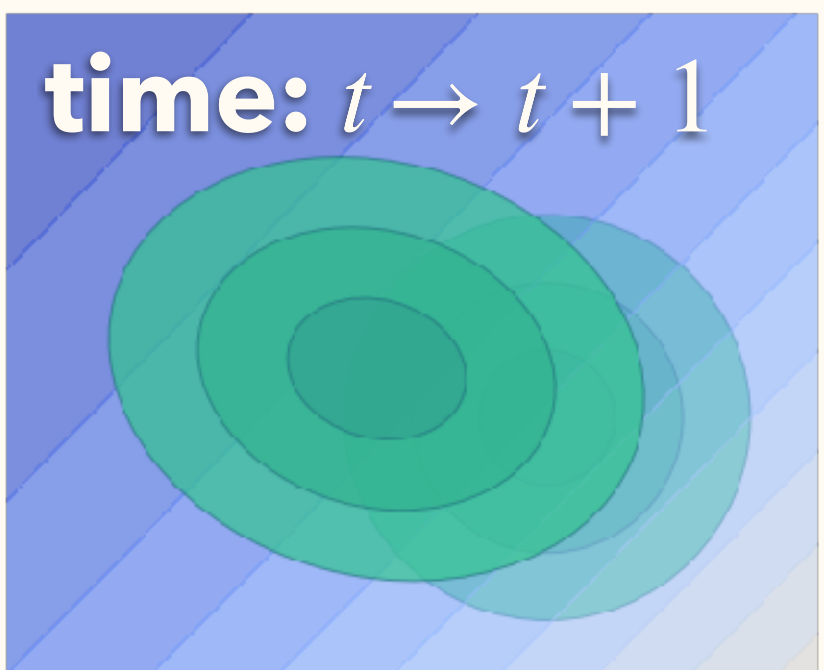
Population Size Adaptation

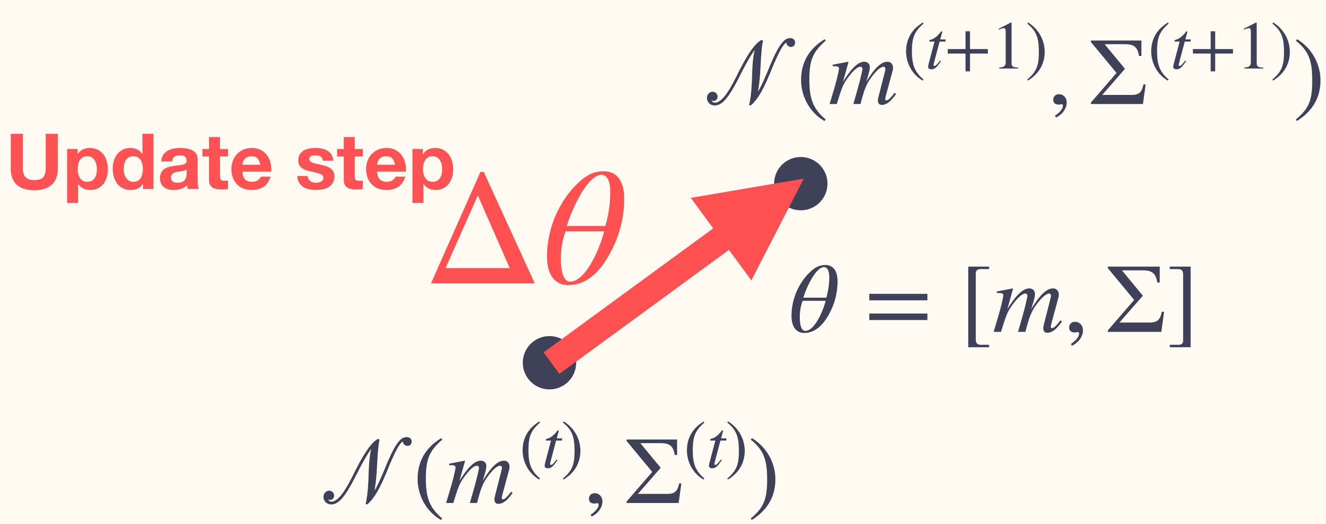
- Based on tendency of the parameter update

Key Observation

On multimodal functions and noisy functions, the parameter update has less tendency than on noiseless unimodal functions.

In the parameter space of the sampling distribution…





Population Size Adaptation

- Based on tendency of the parameter update

Key Observation
On multimodal functions and noisy functions, the parameter update has less tendency than on noiseless unimodal functions.

In the parameter space of the sampling distribution…

On
- noiseless unimodal function

Population Size Adaptation

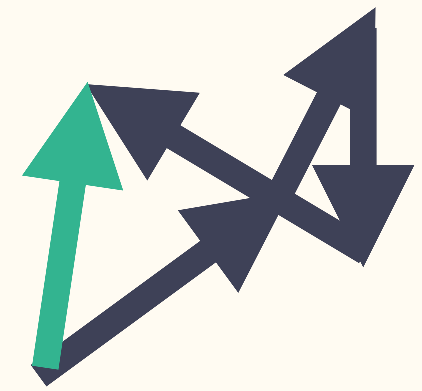
- Based on tendency of the parameter update

Key Observation

On multimodal functions and noisy functions, the parameter update has less tendency than on noiseless unimodal functions.

In the parameter space of the sampling distribution…

On
- multimodal functions
- noisy functions



PSA: Evolution Path

- It accumulates steps in the parameter space

$$p_\theta^{(t+1)} \leftarrow (1-\beta)\, p_\theta^{(t)} + \sqrt{\beta(2-\beta)} \frac{\mathcal{I}_{\theta^{(t)}}^{\frac{1}{2}} \Delta\theta^{(t+1)}}{\sqrt{\mathbb{E}[\|\mathcal{I}_{\theta^{(t)}}^{\frac{1}{2}} \Delta\theta^{(t+1)}\|^2]}}$$

β : cumulation factor

$\mathcal{I}_\theta$: Fisher information matrix under θ

$\mathbb{E}[\,\cdot\,]$: expectation under a random function $f(x) = \epsilon$

normalization factor

→ To absorb the effect of…

- Parameterization of the sampling distribution
- Change of the population size

under a random function

$$\|p_\theta\|^2 \approx 1$$

when λ is too large

$$\|p_\theta\|^2 \gg 1$$

λ: population size

PSA: Population Size Update

$$\lambda^{(t+1)} \leftarrow \lambda^{(t)} \exp\left(\beta\left(\gamma^{(t+1)} - \frac{\|p_\theta^{(t+1)}\|^2}{\alpha}\right)\right)$$

α : threshold

$\gamma^{(t)}$: normalization factor $\approx 1 \ (t \gg 1)$

$$\gamma^{(t+1)} \leftarrow (1-\beta)^2\gamma^{(t)} + \beta(2-\beta)$$

$\|p_\theta\|^2 < \alpha \Rightarrow$ The population size increases

$\|p_\theta\|^2 > \alpha \Rightarrow$ The population size decreases

→ the population size is adapted so that the parameter update has sufficient tendency

PSA: Step-size Correction

- Based on the quality gain analysis [Akimoto 2017]
  ⟶ The optimal step-size depends on the population size
- A practical step-size adaptation in the CMA-ES usually well follows the optimal value [Krause 2017]
- It implies that the step-size is increased when the population size increases, and vice versa.
- The step-size adaptation is corrupted by the population size adaptation.

After updating the population size...

$$\sigma^{(t+1)} \leftarrow \sigma^{(t+1)} \cdot \frac{\sigma^*(\lambda^{(t+1)})}{\sigma^*(\lambda^{(t)})}$$

$$\sigma^*(\lambda) = \frac{c(\lambda) \cdot n \cdot \mu_w(\lambda)}{n - 1 + c(\lambda)^2 \cdot \mu_w(\lambda)}$$

$$c(\lambda) = -\sum_{i=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}]$$

PSA-CMA-ES

1. An iteration of CMA-ES

   A step in the parameter space

$$\Delta\theta = [\Delta m, \Delta\Sigma]$$

$$\Delta m = m^{(t+1)} - m^{(t)}$$

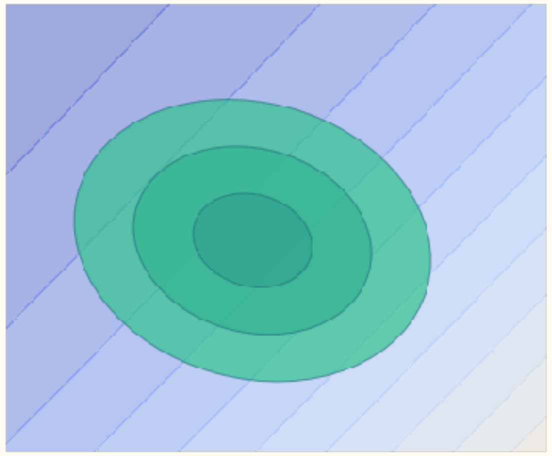
$$\Delta\Sigma = (\sigma^{(t+1)})^2 C^{(t+1)} - (\sigma^{(t)})^2 C^{(t)}$$

Step1 Sample

Step2 Rank

Step3 Estimate

Step4 Update



$$\mathcal{N}(m^{(t)}, (\sigma^{(t)})^2 C^{(t)})$$

$$\Downarrow$$

$$\mathcal{N}(m^{(t+1)}, (\sigma^{(t+1)})^2 C^{(t+1)})$$

2. Update the evolution path

   and the population size

$$p_\theta^{(t+1)} \leftarrow (1-\beta)\, p_\theta^{(t)} + \sqrt{\beta(2-\beta)} \frac{\mathcal{I}^{\frac{1}{2}}_{\theta^{(t)}} \Delta\theta^{(t+1)}}{\sqrt{\mathbb{E}[\| \mathcal{I}^{\frac{1}{2}}_{\theta^{(t)}} \Delta\theta^{(t+1)} \|^2]}}$$

$$\lambda^{(t+1)} \leftarrow \lambda^{(t)} \exp\left(\beta\left(\gamma^{(t+1)} - \frac{\|p_\theta^{(t+1)}\|^2}{\alpha}\right)\right)$$

3. Correct the step-size

$$\sigma^{(t+1)} \leftarrow \frac{\sigma^*(\lambda^{(t+1)})}{\sigma^*(\lambda^{(t)})} \sigma^{(t+1)}$$

Restart Strategy for PSA-CMA-ES

First run: CMA-ES with the default population size $(\sigma^{(0)} = 2)$

→ unimodal functions

Second run: PSA-CMA-ES $(\sigma^{(0)} = 2)$

→ well-structured multimodal

Max population size

$$\lambda_{\max} = 2^9 \cdot \lambda_{\mathrm{def}}$$

Additional runs:

PSA-CMA-ES with a relatively small step-size

$$\sigma^{(0)} = 2 \cdot 10^{-2 \cdot \mathcal{U}[0,1]}$$

→ weakly-structured multimodal functions

Simple Restart

All runs: PSA-CMA-ES $(\sigma^{(0)} = 2, \lambda_{\max} = \infty)$

Simulation

Common Setting

- Initialization: $m^{(0)} \sim \mathcal{U}[4,4)^D \quad (D\text{: problem dimension})$
- Termination:
  - The target function value is reached
  - The number of evaluation is over $10^6 \cdot D$
  - One of the termination conditions [Hansen 2009] is satisfied

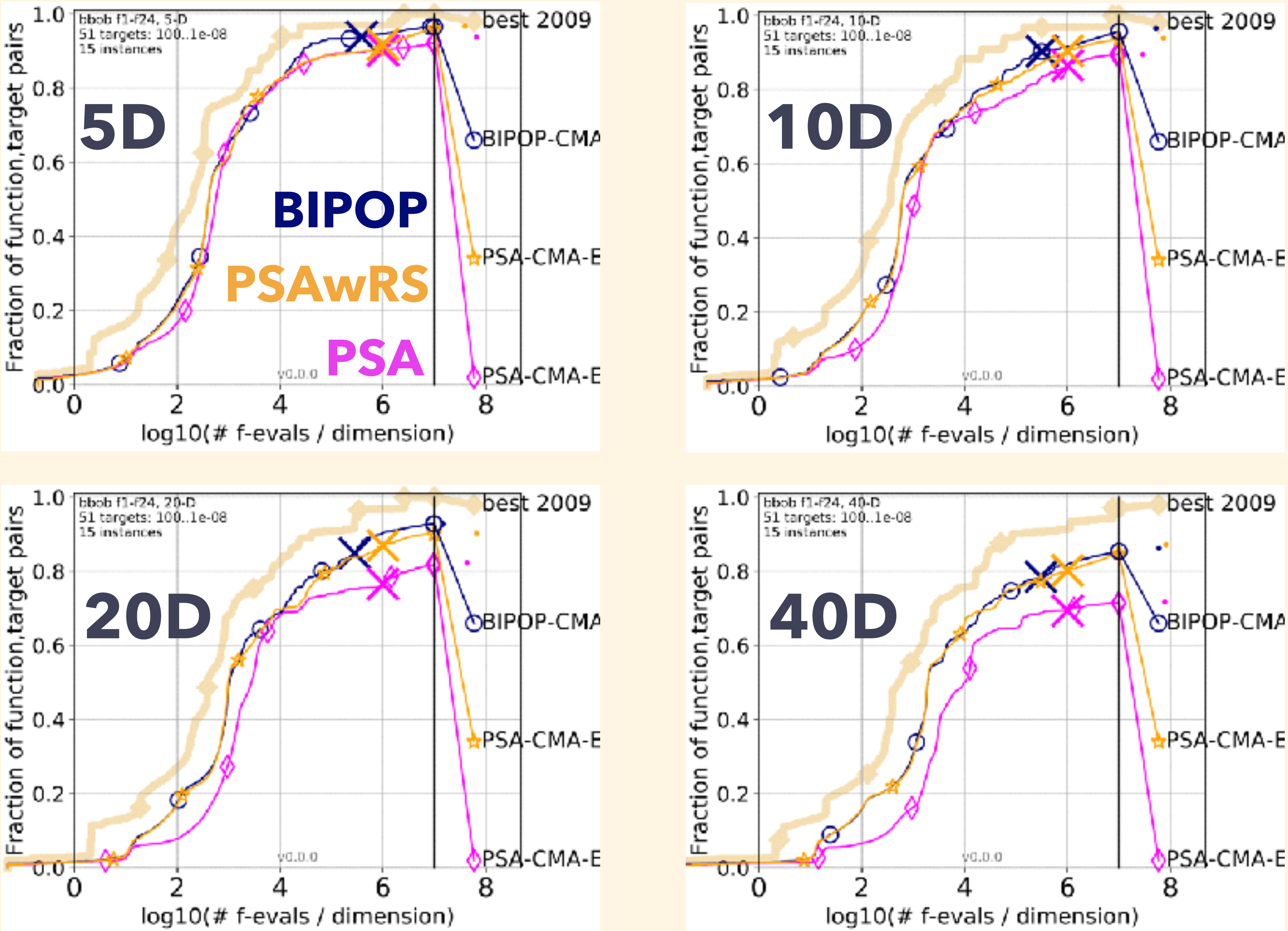
Algorithm Variants

PSA: PSA-CMA-ES with the simple restart

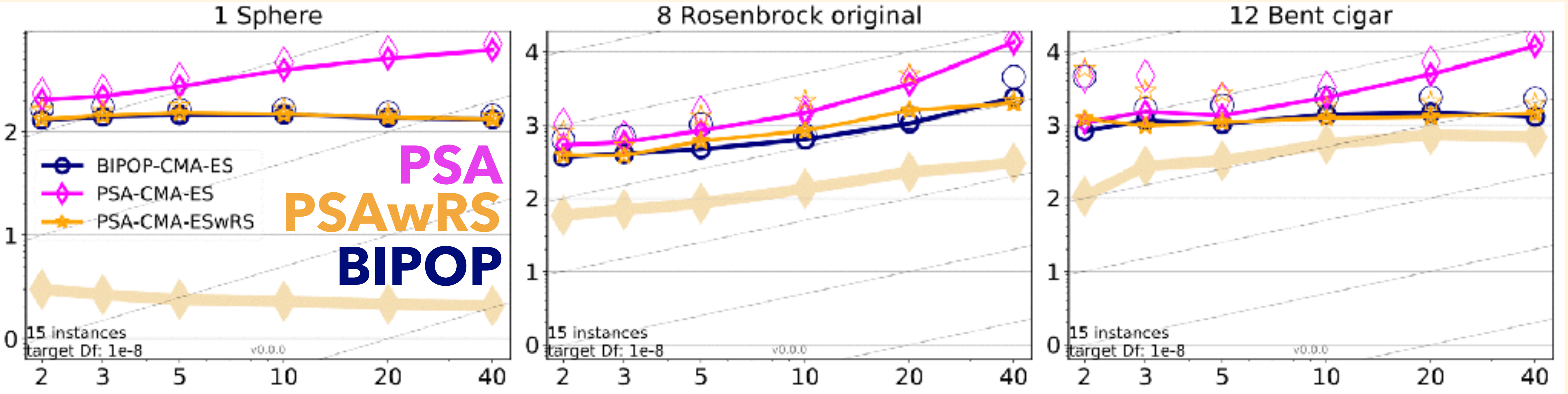
PSAwRS: PSA-CMA-ES with the proposed restart strategy

BIPOP: BIPOP-CMA-ES [Hansen 2009]

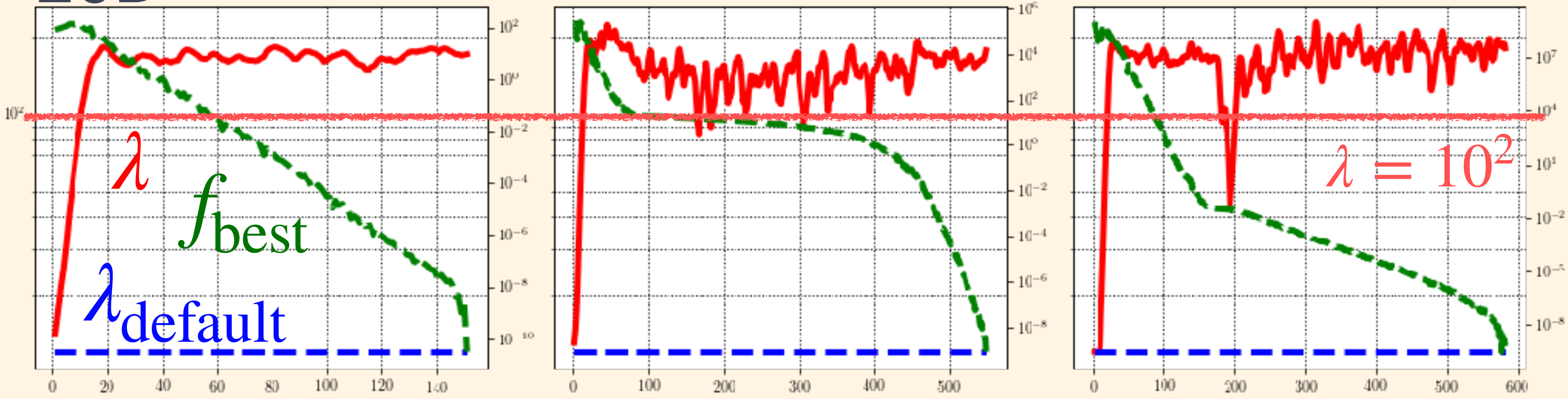
Overall Performance (f1-f24)



Unimodal Functions

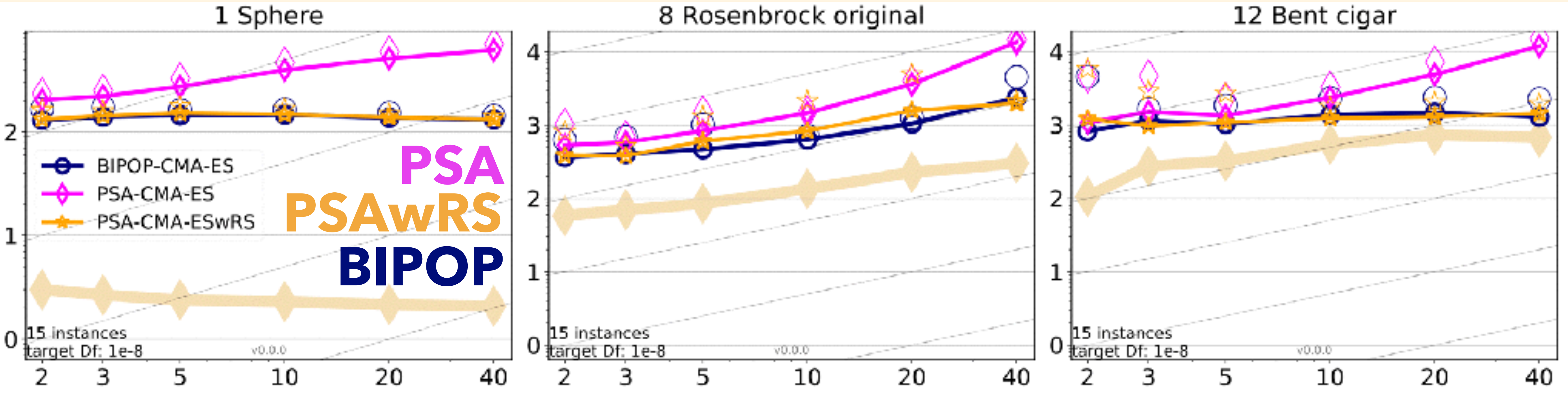


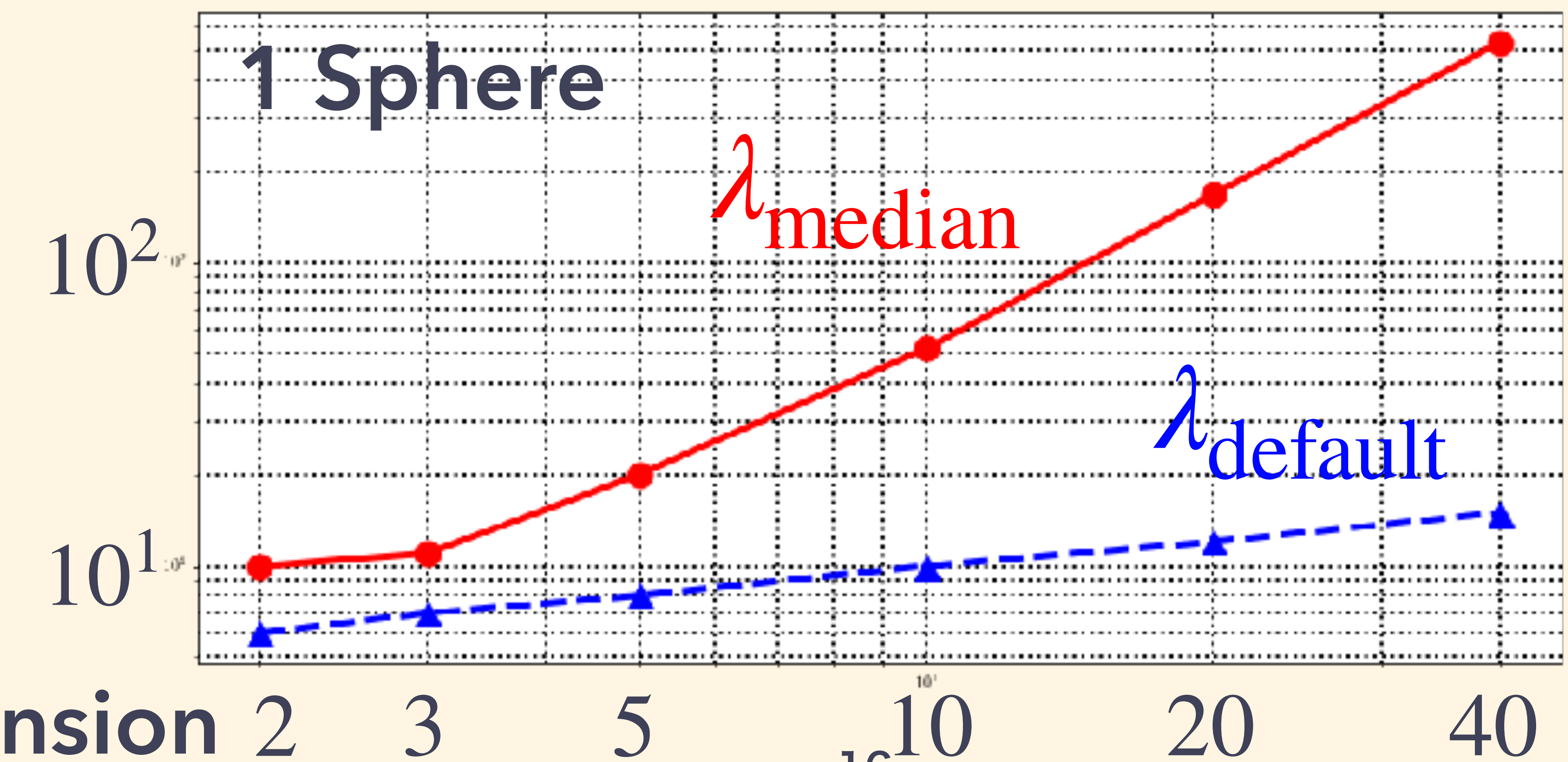
20D



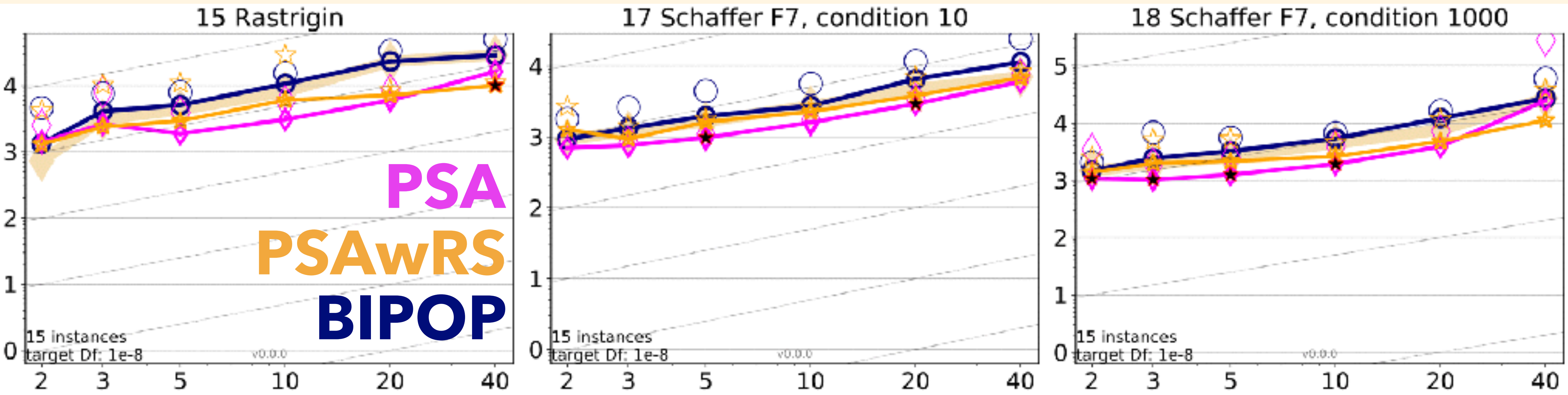
Number of Iteration

Unimodal Functions

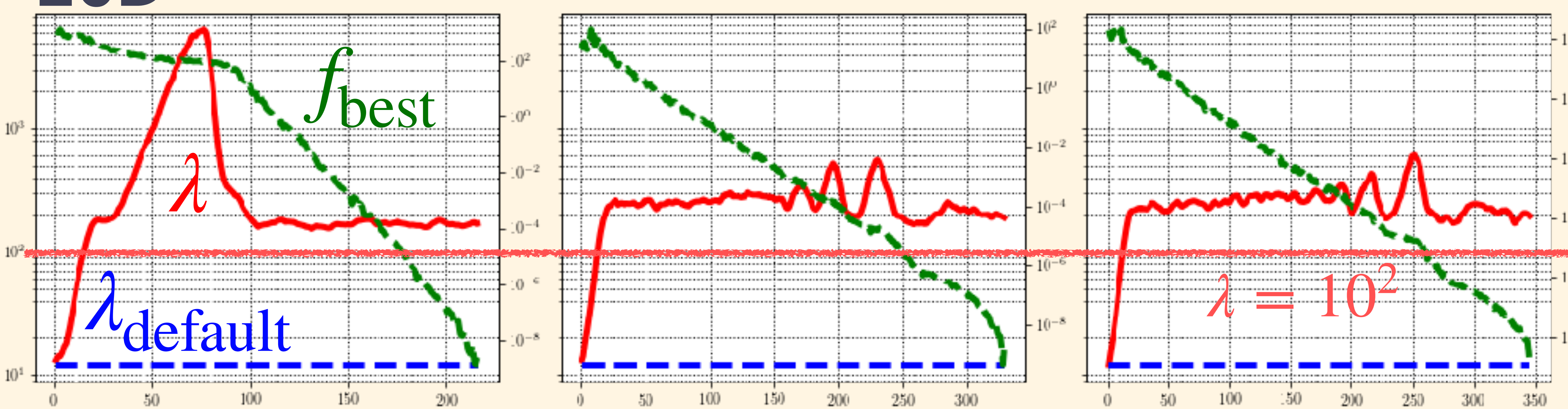




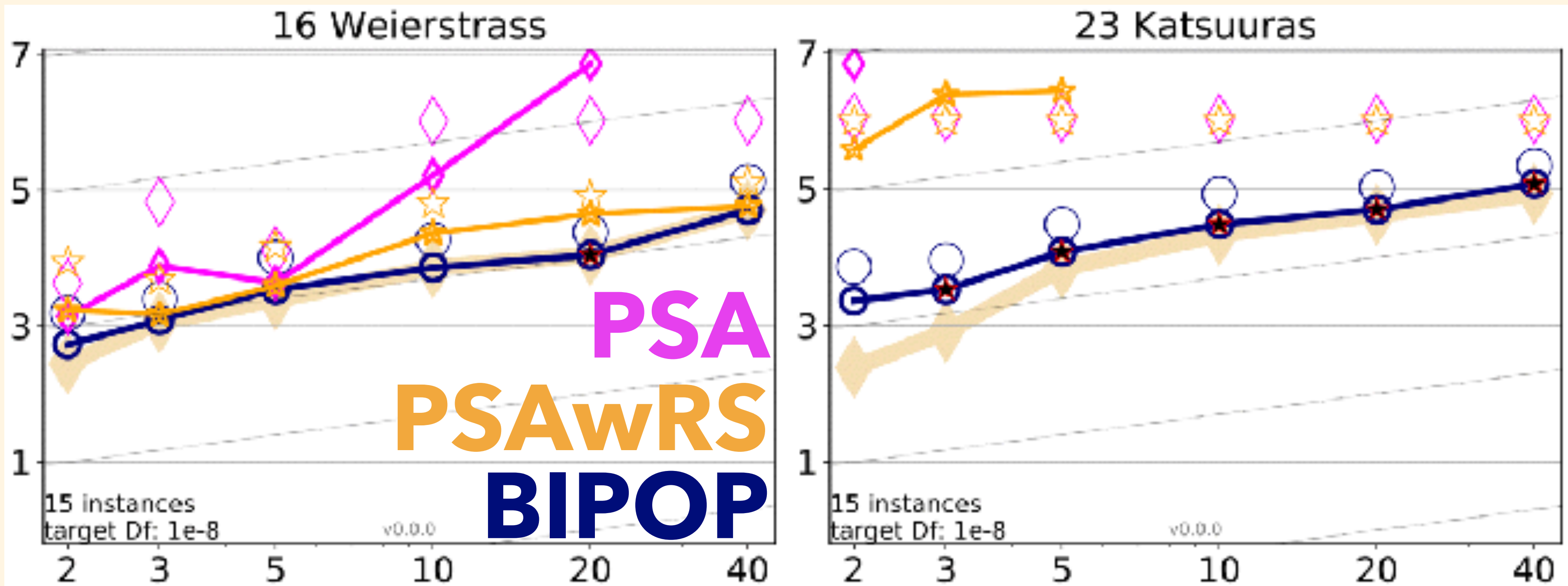
Well-structured Multimodal Functions



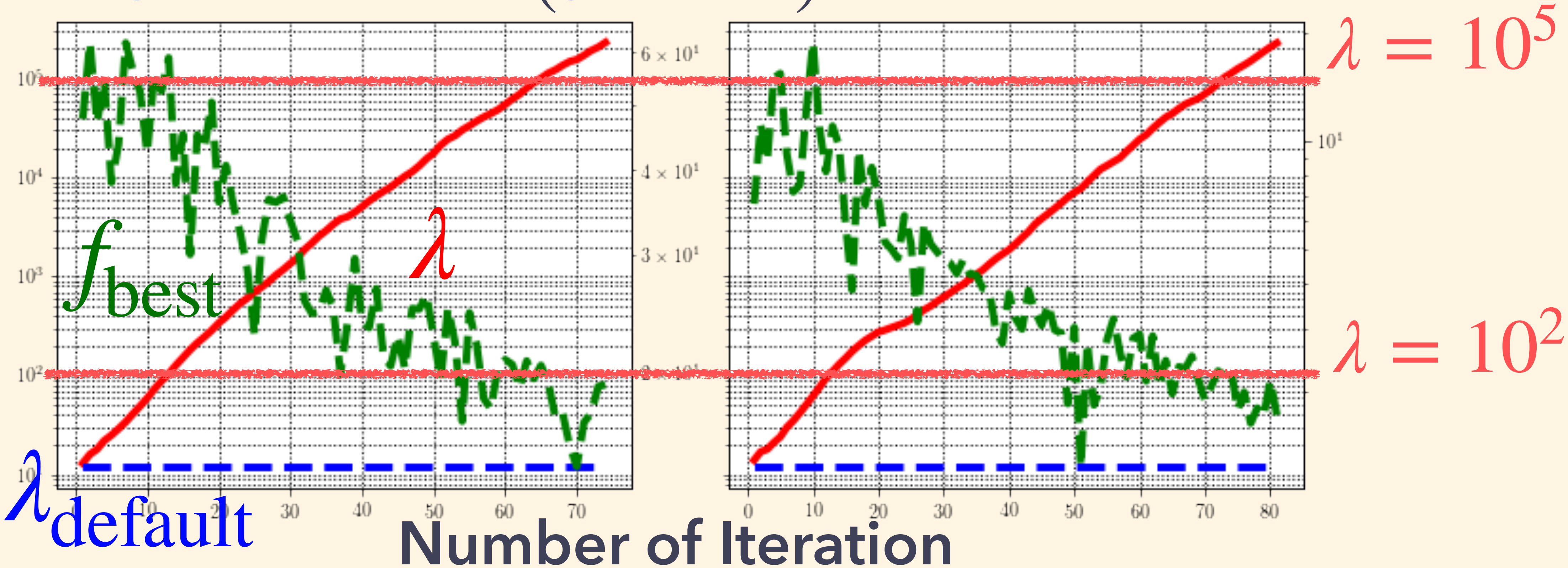
20D



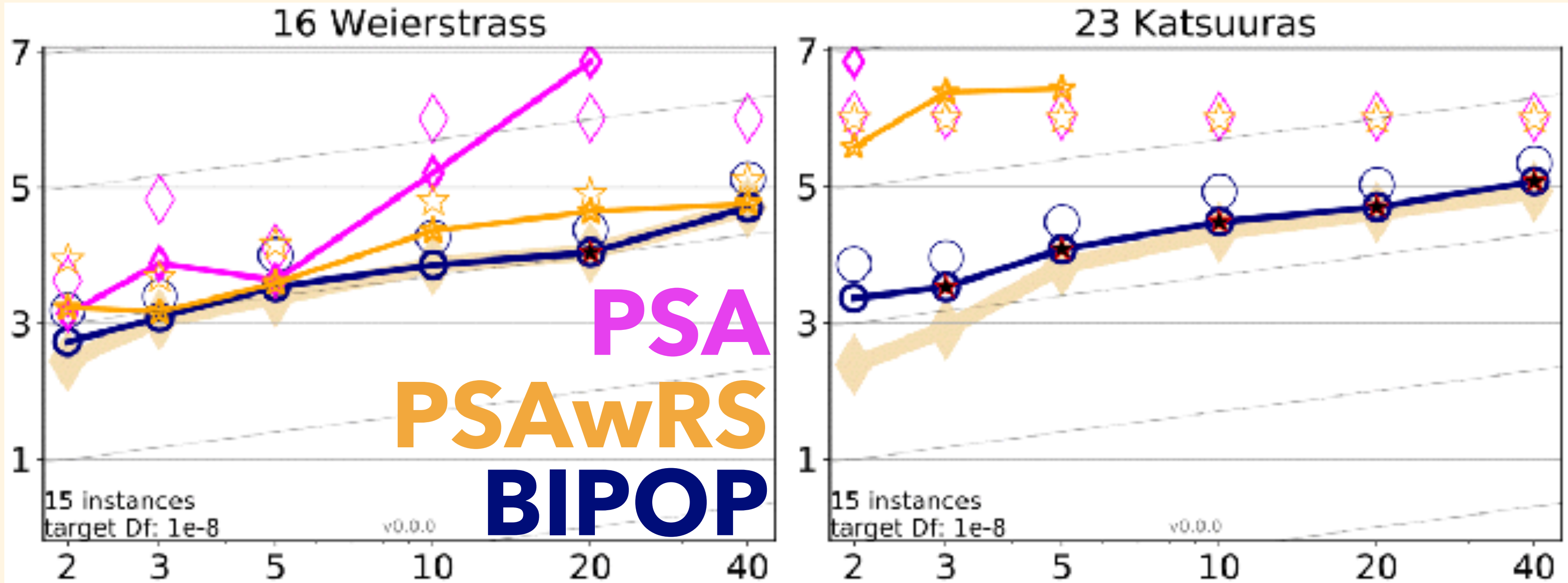
Repetitive Multimodal Functions



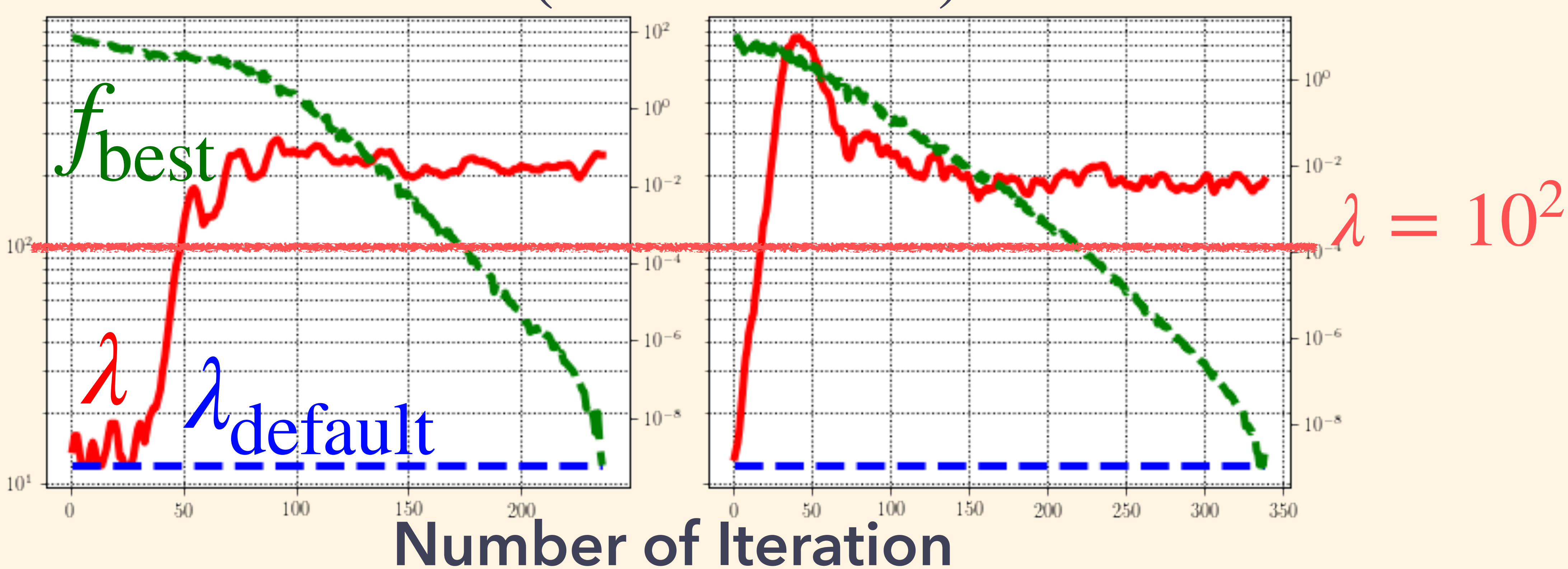
20D $(\sigma^{(0)} = 2)$



Repetitive Multimodal Functions



20D $(\sigma^{(0)} = 2/100)$



Summary

- PSA-CMA-ESwRS is comparable with BIPOP-CMA-ES.

On unimodal functions

- PSA-CMA-ES performs worse as dimension gets greater.

On well-structured multimodal functions

- PSA-CMA-ES works better than BIPOP-CMA-ES.

On repetitive multimodal functions

- An initial step-size is important to avoid inefficient increase of the population size.

Future Work

- To investigate the hyper-parameter setting