Hoeffding and Bernstein Races for Selecting Policies in Evolutionary Direct Policy Search

Verena Heidrich-Meisner, Christian Igel

Institut für Neuroinformatik
Ruhr-Universität Bochum, Germany
http://www.neuroinformatik.rub.de

Outline

Variable Metric Evolution Strategies for Episodic Reinforcement Learning

Racing Algorithms for Uncertainty Handling

Experiments

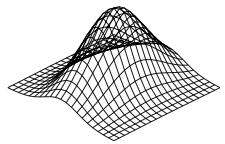
Conclusions & Future Work

Outline

Evolutionary Strategies

- core of random search: search distribution over candidate solutions (i.e., policy parameters)
- evolutionary algorithms: search distribution parameterized by μ candidate solutions in parent population and parameters of variation operators
- we consider real-valued evolution strategies (ESs)
- in each iteration, λ new solutions $\boldsymbol{x}_k \in \mathbb{R}^n, 1 \leq k \leq \lambda$ are generated

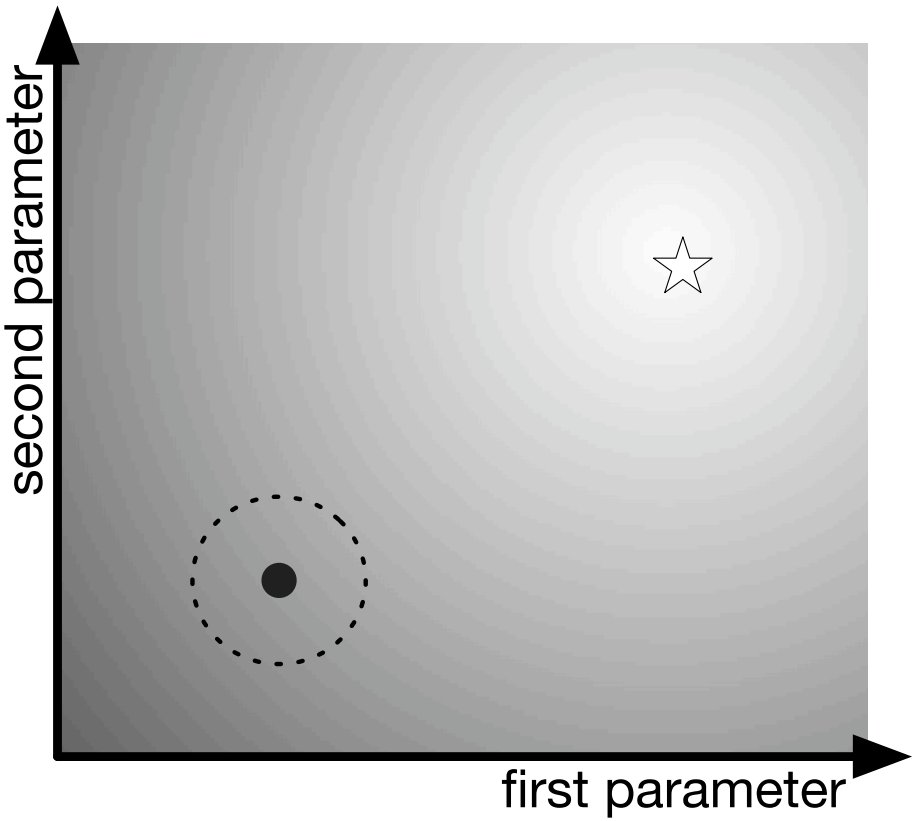
$$\boldsymbol{x}_k \sim \mathsf{N}\left(\sum_{i=1}^{\mu} w_i \boldsymbol{x}_{i\text{th-best-parent}}, \sigma^2 \boldsymbol{C}\right) =$$



  with $\boldsymbol{C} \in \mathbb{R}^{n \times n}$, $\sigma \in \mathbb{R}^+$ and convex weights w_i
- non-elitist (μ, λ)-selection: the best μ of the λ offspring survive

Covariance Matrix Adaptation ES (CMA-ES)

- matrix C is updated to make recent beneficial steps more likely
- global step size σ is updated independently on faster timescale
- highly efficient, allows for small population sizes
- invariant under translation, rotation, flipping of search space, and under order preserving transformations of objective function

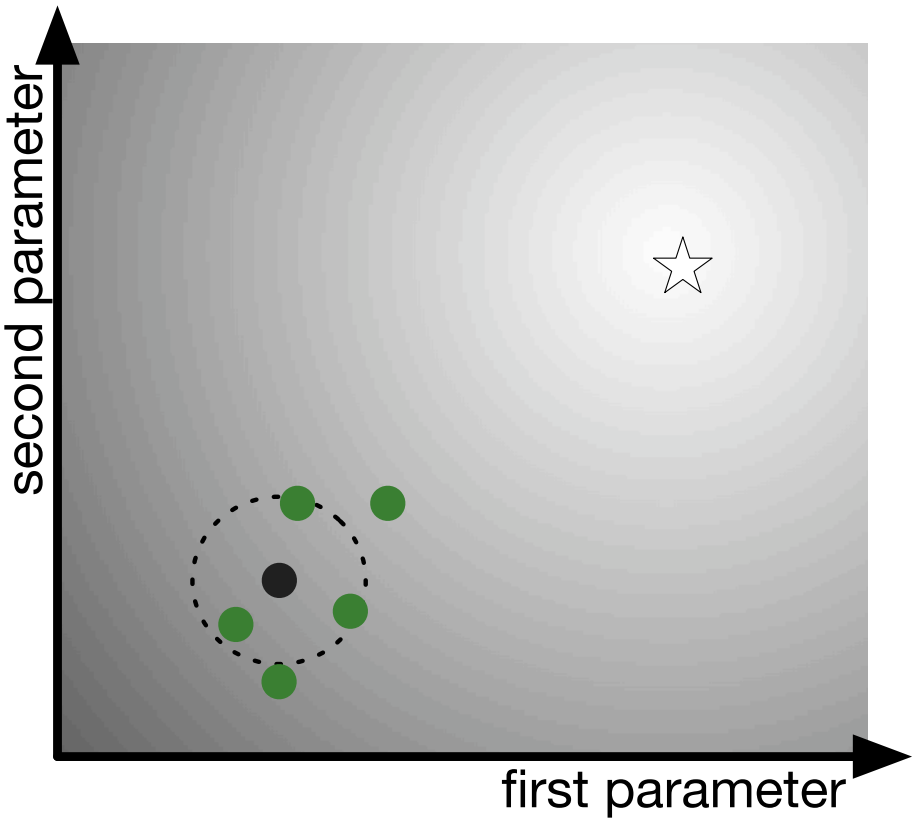


Hansen et al.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), Evolutionary Computation 11(1), 2003

Suttorp, Hansen, Igel: Efficient Covariance Matrix Update for Variable Metric Evolution Strategies, Machine Learning 75, 2009

Covariance Matrix Adaptation ES (CMA-ES)

- matrix C is updated to make recent beneficial steps more likely
- global step size σ is updated independently on faster timescale
- highly efficient, allows for small population sizes
- invariant under translation, rotation, flipping of search space, and under order preserving transformations of objective function

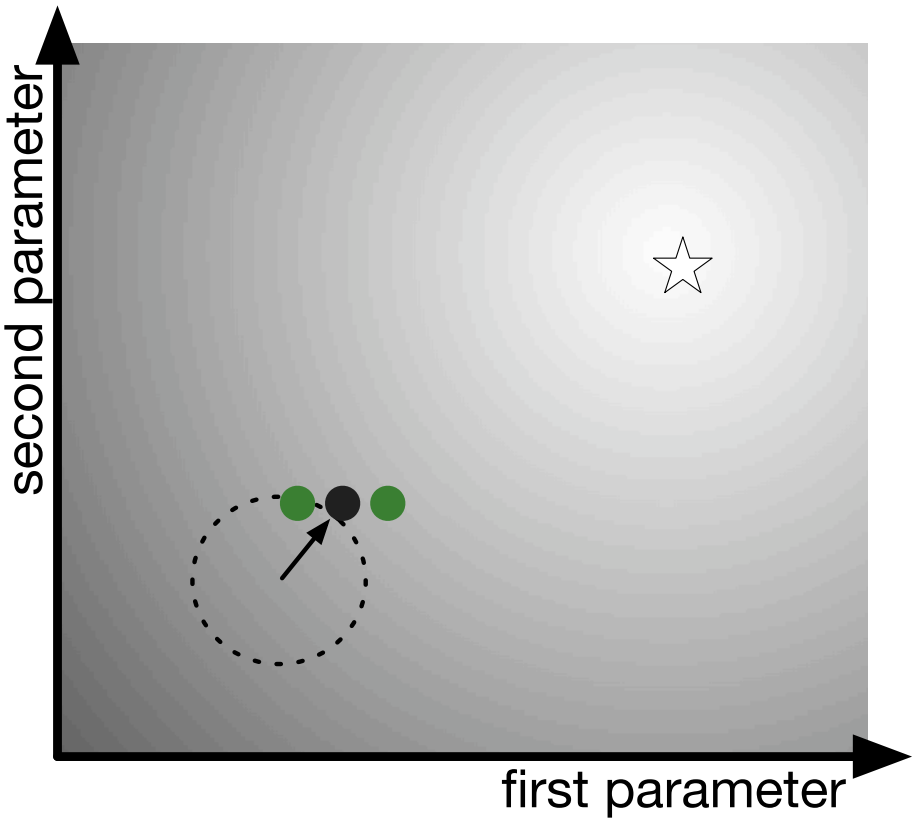


Hansen et al.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), Evolutionary Computation 11(1), 2003

Suttorp, Hansen, Igel: Efficient Covariance Matrix Update for Variable Metric Evolution Strategies, Machine Learning 75, 2009

Covariance Matrix Adaptation ES (CMA-ES)

- matrix C is updated to make recent beneficial steps more likely
- global step size σ is updated independently on faster timescale
- highly efficient, allows for small population sizes
- invariant under translation, rotation, flipping of search space, and under order preserving transformations of objective function

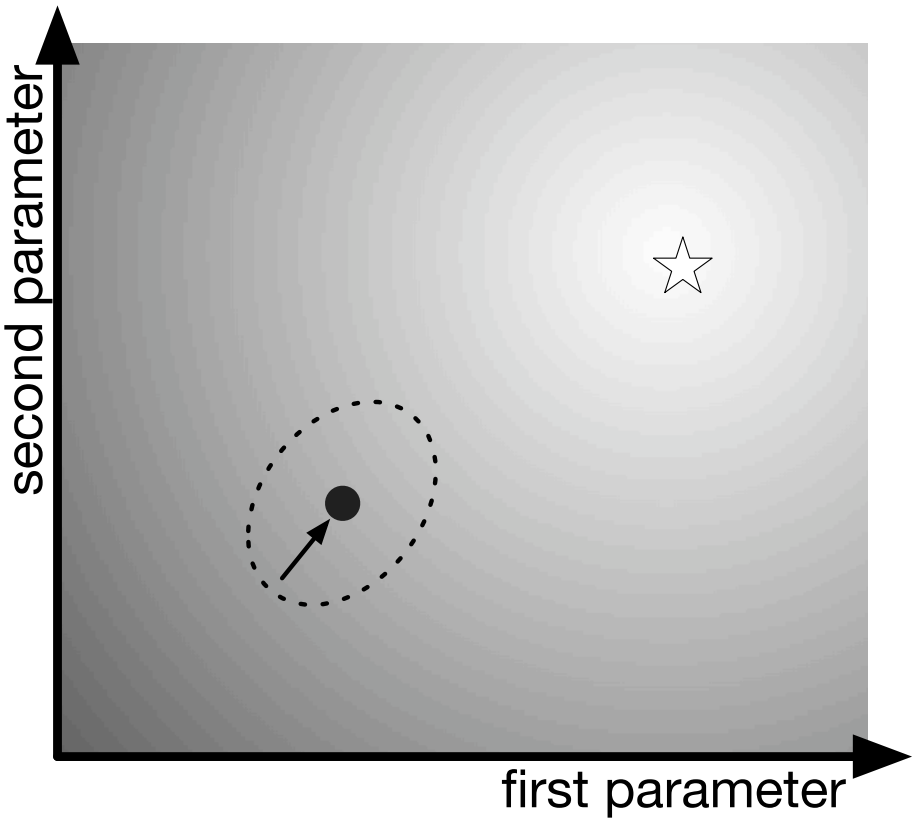


Hansen et al.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), Evolutionary Computation 11(1), 2003

Suttorp, Hansen, Igel: Efficient Covariance Matrix Update for Variable Metric Evolution Strategies, Machine Learning 75, 2009

Covariance Matrix Adaptation ES (CMA-ES)

- matrix C is updated to make recent beneficial steps more likely
- global step size σ is updated independently on faster timescale
- highly efficient, allows for small population sizes
- invariant under translation, rotation, flipping of search space, and under order preserving transformations of objective function

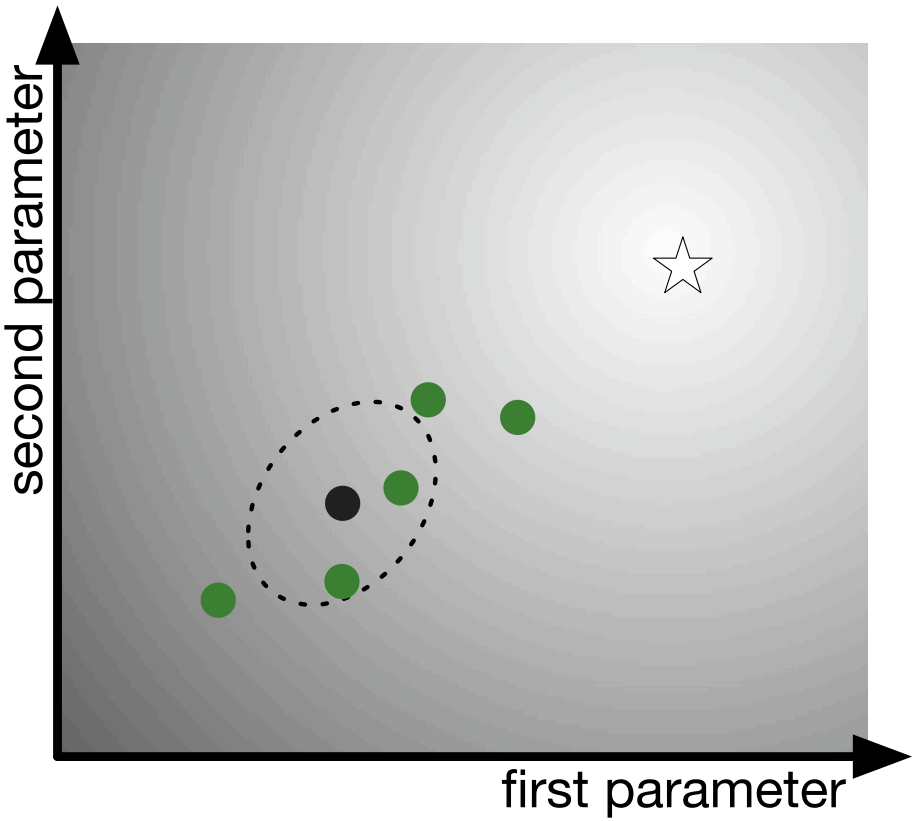


Hansen et al.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), Evolutionary Computation 11(1), 2003

Suttorp, Hansen, Igel: Efficient Covariance Matrix Update for Variable Metric Evolution Strategies, Machine Learning 75, 2009

Covariance Matrix Adaptation ES (CMA-ES)

- matrix C is updated to make recent beneficial steps more likely
- global step size σ is updated independently on faster timescale
- highly efficient, allows for small population sizes
- invariant under translation, rotation, flipping of search space, and under order preserving transformations of objective function

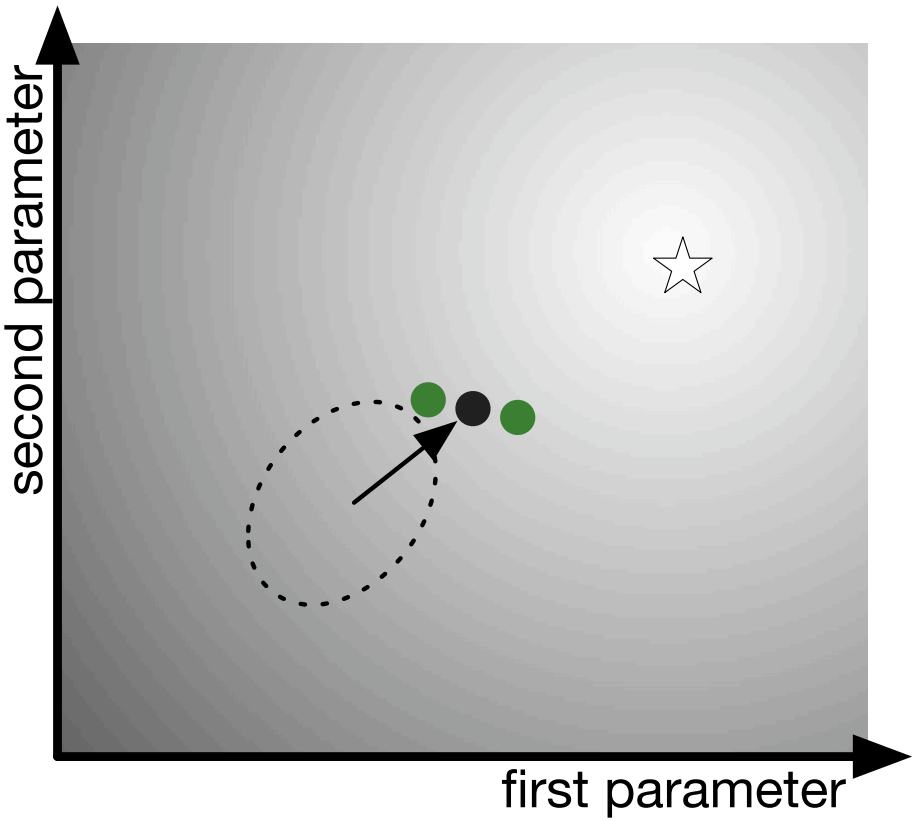


Hansen et al.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), Evolutionary Computation 11(1), 2003

Suttorp, Hansen, Igel: Efficient Covariance Matrix Update for Variable Metric Evolution Strategies, Machine Learning 75, 2009

Covariance Matrix Adaptation ES (CMA-ES)

- matrix C is updated to make recent beneficial steps more likely
- global step size σ is updated independently on faster timescale
- highly efficient, allows for small population sizes
- invariant under translation, rotation, flipping of search space, and under order preserving transformations of objective function

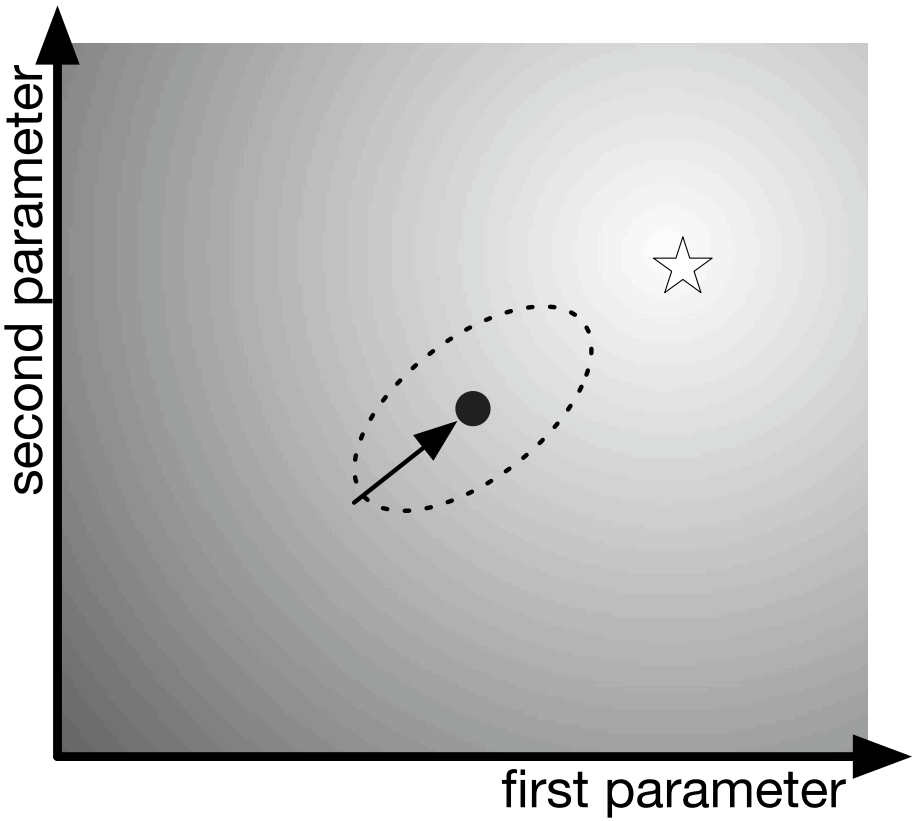


Hansen et al.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), Evolutionary Computation 11(1), 2003

Suttorp, Hansen, Igel: Efficient Covariance Matrix Update for Variable Metric Evolution Strategies, Machine Learning 75, 2009

Covariance Matrix Adaptation ES (CMA-ES)

- matrix C is updated to make recent beneficial steps more likely
- global step size σ is updated independently on faster timescale
- highly efficient, allows for small population sizes
- invariant under translation, rotation, flipping of search space, and under order preserving transformations of objective function

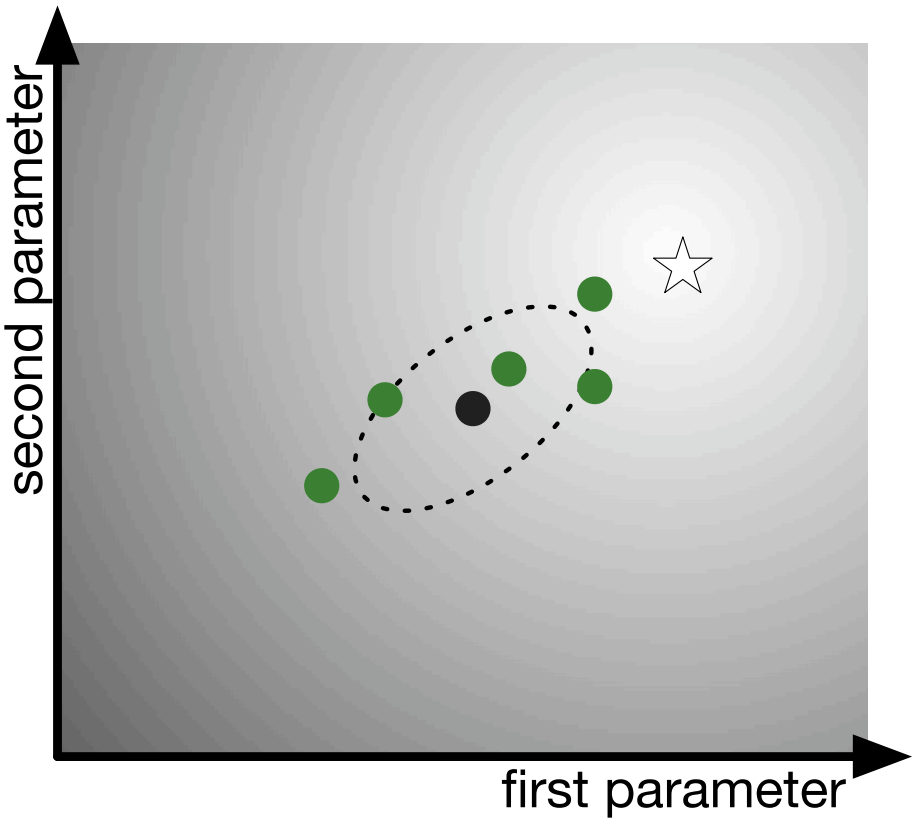


Hansen et al.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), Evolutionary Computation 11(1), 2003

Suttorp, Hansen, Igel: Efficient Covariance Matrix Update for Variable Metric Evolution Strategies, Machine Learning 75, 2009

Covariance Matrix Adaptation ES (CMA-ES)

- matrix C is updated to make recent beneficial steps more likely
- global step size σ is updated independently on faster timescale
- highly efficient, allows for small population sizes
- invariant under translation, rotation, flipping of search space, and under order preserving transformations of objective function

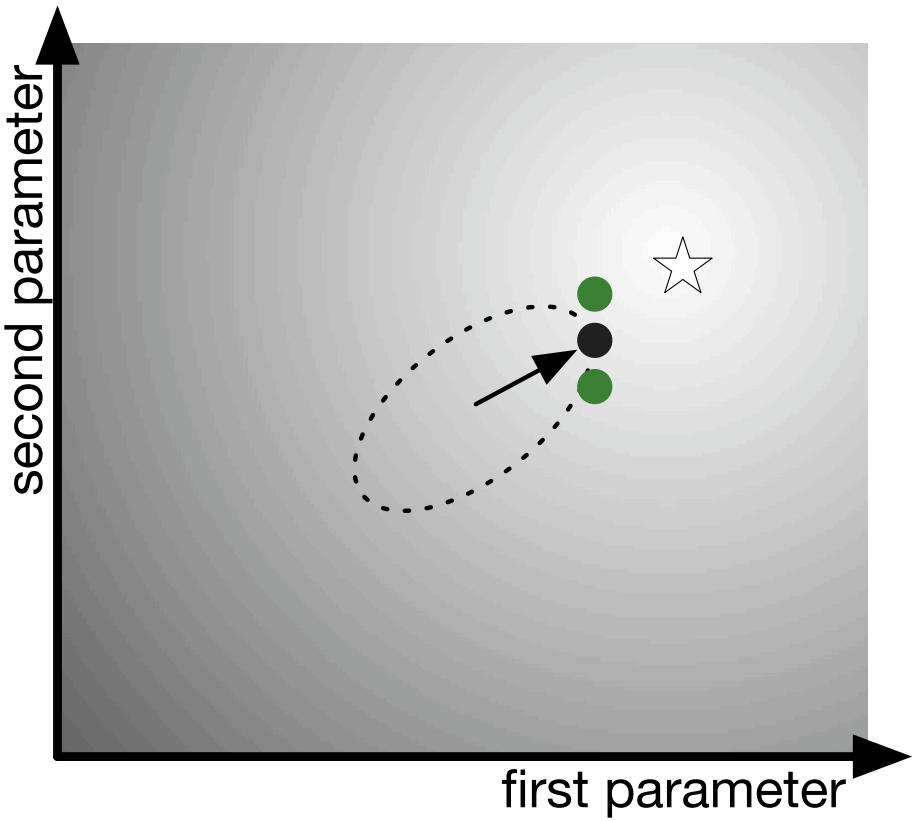


Hansen et al.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), Evolutionary Computation 11(1), 2003

Suttorp, Hansen, Igel: Efficient Covariance Matrix Update for Variable Metric Evolution Strategies, Machine Learning 75, 2009

Covariance Matrix Adaptation ES (CMA-ES)

- matrix C is updated to make recent beneficial steps more likely
- global step size σ is updated independently on faster timescale
- highly efficient, allows for small population sizes
- invariant under translation, rotation, flipping of search space, and under order preserving transformations of objective function

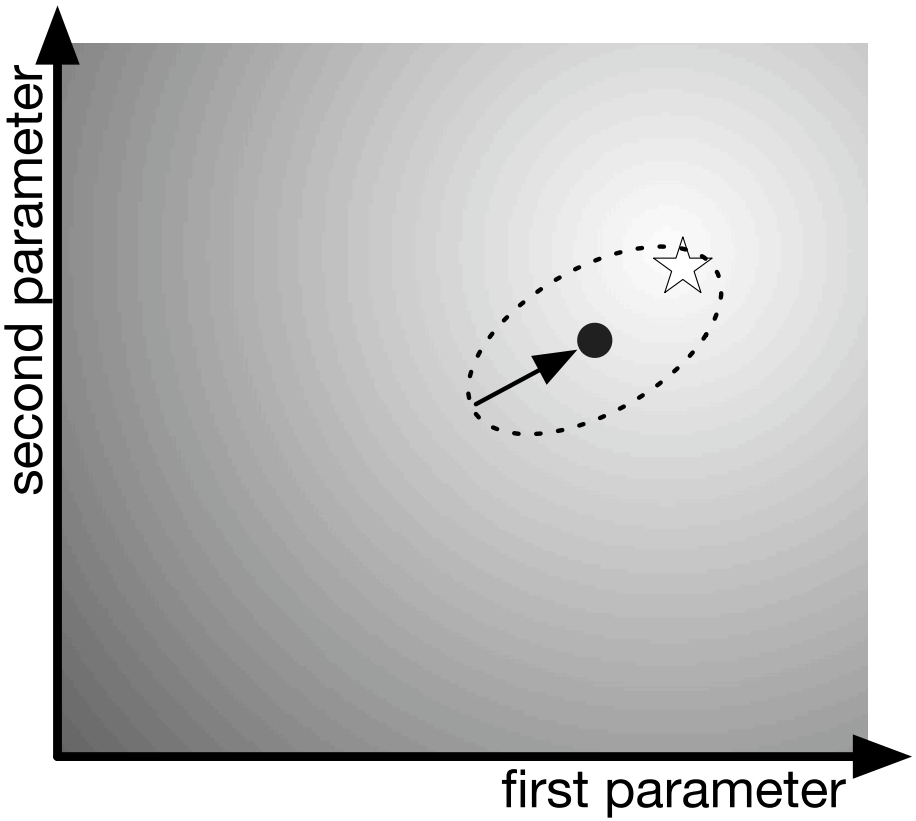


Hansen et al.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), Evolutionary Computation 11(1), 2003

Suttorp, Hansen, Igel: Efficient Covariance Matrix Update for Variable Metric Evolution Strategies, Machine Learning 75, 2009

Covariance Matrix Adaptation ES (CMA-ES)

- matrix C is updated to make recent beneficial steps more likely
- global step size σ is updated independently on faster timescale
- highly efficient, allows for small population sizes
- invariant under translation, rotation, flipping of search space, and under order preserving transformations of objective function



Hansen et al.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES), Evolutionary Computation 11(1), 2003

Suttorp, Hansen, Igel: Efficient Covariance Matrix Update for Variable Metric Evolution Strategies, Machine Learning 75, 2009

The CMA-ES for RL...

1. works as a direct policy search method resembling policy gradient methods,

Heidrich-Meisner, Igel: Variable-metric Evolution Strategies for Direct Policy Search. MSRL 2009

Heidrich-Meisner, Igel: Neuroevolution Strategies for Episodic Reinforcement Learning. Journal of Algorithms, 2009, doi:10.1016/j.jalgor.2009.04.002

The CMA-ES for RL...

1. works as a direct policy search method resembling policy gradient methods,
2. is comparatively robust w.r.t. tuning of meta-parameters,

Heidrich-Meisner, Igel: Variable-metric Evolution Strategies for Direct Policy Search. MSRL 2009

Heidrich-Meisner, Igel: Neuroevolution Strategies for Episodic Reinforcement Learning. Journal of Algorithms, 2009, doi:10.1016/j.jalgor.2009.04.002

The CMA-ES for RL...

1. works as a direct policy search method resembling policy gradient methods,
2. is comparatively robust w.r.t. tuning of meta-parameters,
3. is based on ranking policies, which is less susceptible to uncertainty & noise compared to estimating a value function or a gradient of a performance measure w.r.t. policy parameters,

Heidrich-Meisner, Igel: Variable-metric Evolution Strategies for Direct Policy Search. MSRL 2009

Heidrich-Meisner, Igel: Neuroevolution Strategies for Episodic Reinforcement Learning. Journal of Algorithms, 2009, doi:10.1016/j.jalgor.2009.04.002

The CMA-ES for RL...

1. works as a direct policy search method resembling policy gradient methods,
2. is comparatively robust w.r.t. tuning of meta-parameters,
3. is based on ranking policies, which is less susceptible to uncertainty & noise compared to estimating a value function or a gradient of a performance measure w.r.t. policy parameters,
4. is a variable metric algorithm learning an appropriate coordinate system for searching better policies,

Heidrich-Meisner, Igel: Variable-metric Evolution Strategies for Direct Policy Search. MSRL 2009

Heidrich-Meisner, Igel: Neuroevolution Strategies for Episodic Reinforcement Learning. Journal of Algorithms, 2009, doi:10.1016/j.jalgor.2009.04.002

The CMA-ES for RL...

1. works as a direct policy search method resembling policy gradient methods,
2. is comparatively robust w.r.t. tuning of meta-parameters,
3. is based on ranking policies, which is less susceptible to uncertainty & noise compared to estimating a value function or a gradient of a performance measure w.r.t. policy parameters,
4. is a variable metric algorithm learning an appropriate coordinate system for searching better policies,
5. is applicable if function approximators are non-differentiable, and

Heidrich-Meisner, Igel: Variable-metric Evolution Strategies for Direct Policy Search. MSRL 2009

Heidrich-Meisner, Igel: Neuroevolution Strategies for Episodic Reinforcement Learning. Journal of Algorithms, 2009, doi:10.1016/j.jalgor.2009.04.002

The CMA-ES for RL...

1. works as a direct policy search method resembling policy gradient methods,
2. is comparatively robust w.r.t. tuning of meta-parameters,
3. is based on ranking policies, which is less susceptible to uncertainty & noise compared to estimating a value function or a gradient of a performance measure w.r.t. policy parameters,
4. is a variable metric algorithm learning an appropriate coordinate system for searching better policies,
5. is applicable if function approximators are non-differentiable, and
6. extracts a search direction from the scalar reward signals.

Heidrich-Meisner, Igel: Variable-metric Evolution Strategies for Direct Policy Search. MSRL 2009

Heidrich-Meisner, Igel: Neuroevolution Strategies for Episodic Reinforcement Learning. Journal of Algorithms, 2009, doi:10.1016/j.jalgor.2009.04.002

Outline

Uncertainty handling

- Uncertainty and noise (e.g., caused by random rewards and transitions, random initialization, and noisy state observations) require evaluation of a policy based on several roll-outs.

Uncertainty handling

- Uncertainty and noise (e.g., caused by random rewards and transitions, random initialization, and noisy state observations) require evaluation of a policy based on several roll-outs.
- If the number of roll-outs is too small, the selection pressure is too weak to drive learning, if it is too high, learning becomes inefficient.

Uncertainty handling

- Uncertainty and noise (e.g., caused by random rewards and transitions, random initialization, and noisy state observations) require evaluation of a policy based on several roll-outs.
- If the number of roll-outs is too small, the selection pressure is too weak to drive learning, if it is too high, learning becomes inefficient.
- Thus, we ask:
  1. How many roll-outs should be performed in one selection procedure (i.e., per generation)?

Uncertainty handling

- Uncertainty and noise (e.g., caused by random rewards and transitions, random initialization, and noisy state observations) require evaluation of a policy based on several roll-outs.
- If the number of roll-outs is too small, the selection pressure is too weak to drive learning, if it is too high, learning becomes inefficient.
- Thus, we ask:
  1. How many roll-outs should be performed in one selection procedure (i.e., per generation)?
  2. How should these roll-outs be distributed among the candidate solutions?

How to distribute roll-outs?

Given

- a set of λ policies,
- a maximum number of roll-outs,
- upper and lower bounds on the quality (i.e., fitness).

We want to distribute the roll-outs such that we identify the best μ policies with a given confidence $1 - \delta$.

How to distribute roll-outs?

Given

- a set of λ policies,
- a maximum number of roll-outs,
- upper and lower bounds on the quality (i.e., fitness).

We want to distribute the roll-outs such that we identify the best μ policies with a given confidence $1 - \delta$.

Idea: extend *racing algorithms* to select a set of solutions!

Maron, Moore: The Racing Algorithm: Model Selection for Lazy Learners, Artificial Intelligence Review 11, 1997

Mnih, Szepesvári, Audibert: Empirical Bernstein Stopping, ICML 2008

How to distribute roll-outs?

Given

- a set of λ policies,
- a maximum number of roll-outs,
- upper and lower bounds on the quality (i.e., fitness).

We want to distribute the roll-outs such that we identify the best μ policies with a given confidence $1 - \delta$.

Idea: extend *racing algorithms* to select a set of solutions!

Maron, Moore: The Racing Algorithm: Model Selection for Lazy Learners, Artificial Intelligence Review 11, 1997

Mnih, Szepesvári, Audibert: Empirical Bernstein Stopping, ICML 2008

Ingredients:

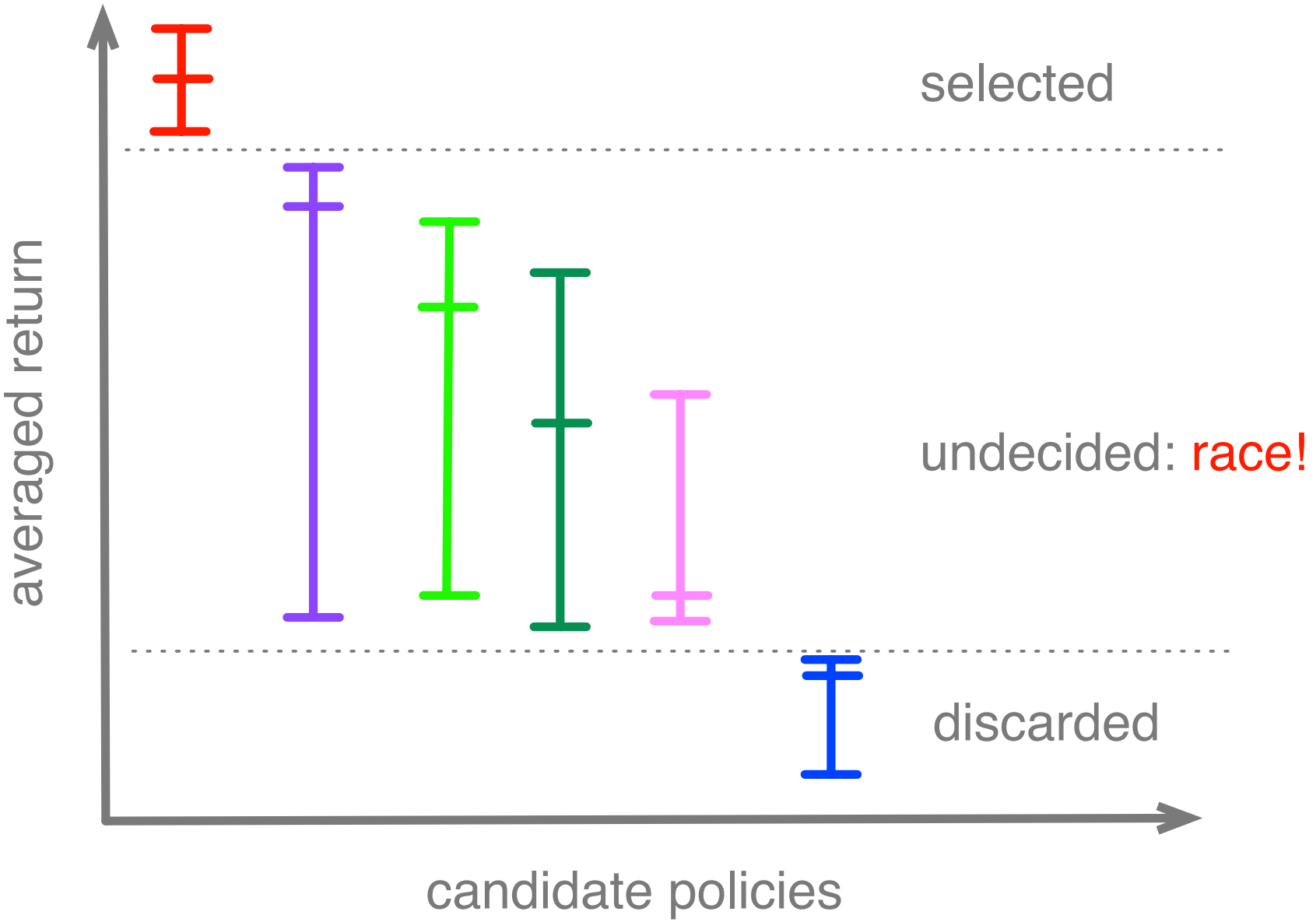
- Hoeffding and empirical Bernstein bounds
- union bound

Sketch of basic algorithm

While maximum number of evaluations is not reached:

1. (re)evaluate each policy which is not tagged *selected* or *discarded*
2. compute confidence intervals for fitness, keep tightest interval
3. *select* each policy better with probability $1 - \delta$ than $\lambda - \mu$ others
4. *discard* each policy worse with probability $1 - \delta$ than μ others

Selection race, (3,6)-selection



Empirical Bernstein bound

For example, in racing round t a confidence interval for the ith policy is given by

$$\pm\hat{\sigma}_{i,t}\sqrt{2\frac{\log(3n_{\mathsf{b}})-\log\delta}{t}}+3R\frac{\log(3n_{\mathsf{b}})-\log\delta}{t}$$

with estimated variance $\hat{\sigma}_{i,t}$, fitness range R, and estimated number of tests n_{b}.

Mnih, Szepesvári, Audibert: Empirical Bernstein Stopping, ICML 2008

Empirical Bernstein bound

For example, in racing round t a confidence interval for the ith policy is given by

$$\pm\hat{\sigma}_{i,t}\sqrt{2\frac{\log(3n_{\mathsf{b}})-\log\delta}{t}}+3R\frac{\log(3n_{\mathsf{b}})-\log\delta}{t}$$

with estimated variance $\hat{\sigma}_{i,t}$, fitness range R, and estimated number of tests n_{b}.

Mnih, Szepesvári, Audibert: Empirical Bernstein Stopping, ICML 2008

For each policy, the lowest upper and highest lower bound so far are stored.

Empirical Bernstein bound

For example, in racing round t a confidence interval for the ith policy is given by

$$\pm\hat{\sigma}_{i,t}\sqrt{2\frac{\log(3n_{\mathsf{b}})-\log\delta}{t}}+3R\frac{\log(3n_{\mathsf{b}})-\log\delta}{t}$$

with estimated variance $\hat{\sigma}_{i,t}$, fitness range R, and estimated number of tests n_{b}.

Mnih, Szepesvári, Audibert: Empirical Bernstein Stopping, ICML 2008

For each policy, the lowest upper and highest lower bound so far are stored.

Constant n_{b} depends on the a priori chosen maximum number of roll-outs and should be as small as possible!

Main result

Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda\}$ be a set of individuals with fitness values almost surely between a and b and $\delta \in]0, 1]$. If the selection procedure tags a set $\mathbb{S}$ of policies as *selected*, then with probability of at least $1 - \delta$ these elements belong to the best μ out of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda\}$ in terms of mean performance.

Main result

Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda\}$ be a set of individuals with fitness values almost surely between a and b and $\delta \in]0, 1]$. If the selection procedure tags a set $\mathbb{S}$ of policies as *selected*, then with probability of at least $1 - \delta$ these elements belong to the best μ out of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda\}$ in terms of mean performance.

What if the procedure fails to identify μ best individuals within n_b evaluation with the given confidence? But how to choose n_b, i.e., how many roll-outs per generation?

Main result

Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda\}$ be a set of individuals with fitness values almost surely between a and b and $\delta \in]0, 1]$. If the selection procedure tags a set $\mathbb{S}$ of policies as *selected*, then with probability of at least $1 - \delta$ these elements belong to the best μ out of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda\}$ in terms of mean performance.

What if the procedure fails to identify μ best individuals within n_b evaluation with the given confidence? But how to choose n_b, i.e., how many roll-outs per generation?

If $|\mathbb{S}| < \mu$ the overall number of evaluations per generation should be increased and may be decreased otherwise.

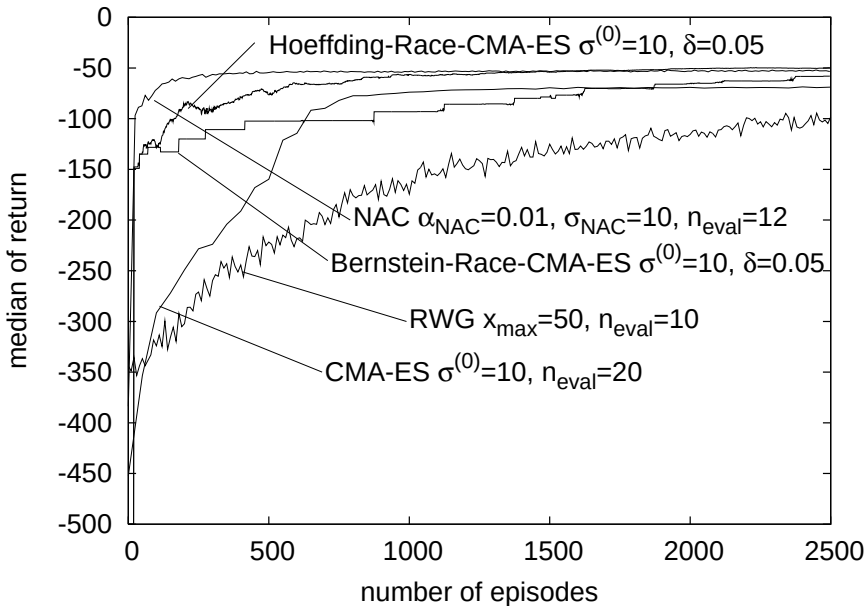
Outline

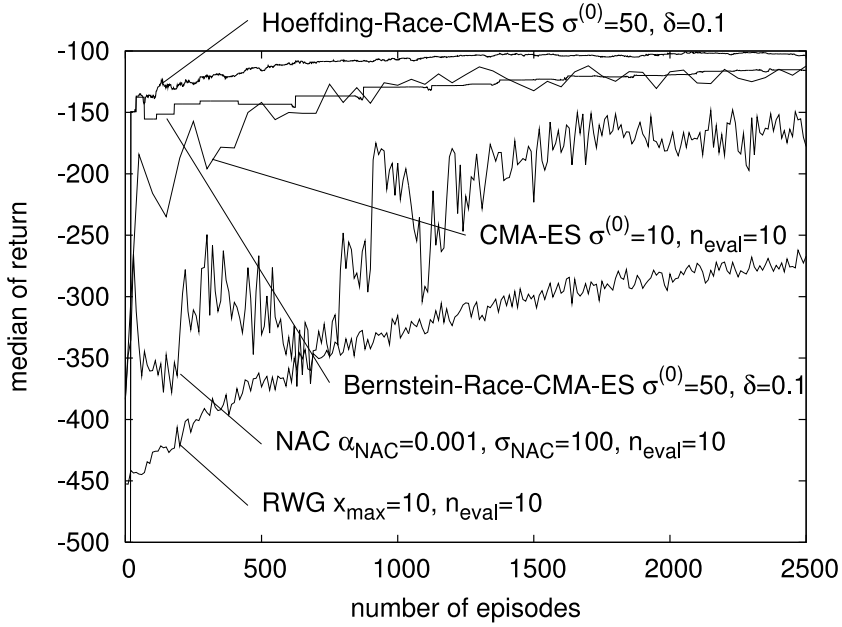
Preliminary experiments I

- first experiments on different instances of Mountain Car and Swimmer benchmark problems
- comparison with std. CMA-ES, Natural Actor-Critic (NAC), random weight guessing (RWG); best parameter configuration determined by grid-search for each method

Peters, Schaal: Natural Actor-Critic, Neurocomputing 71, 2008

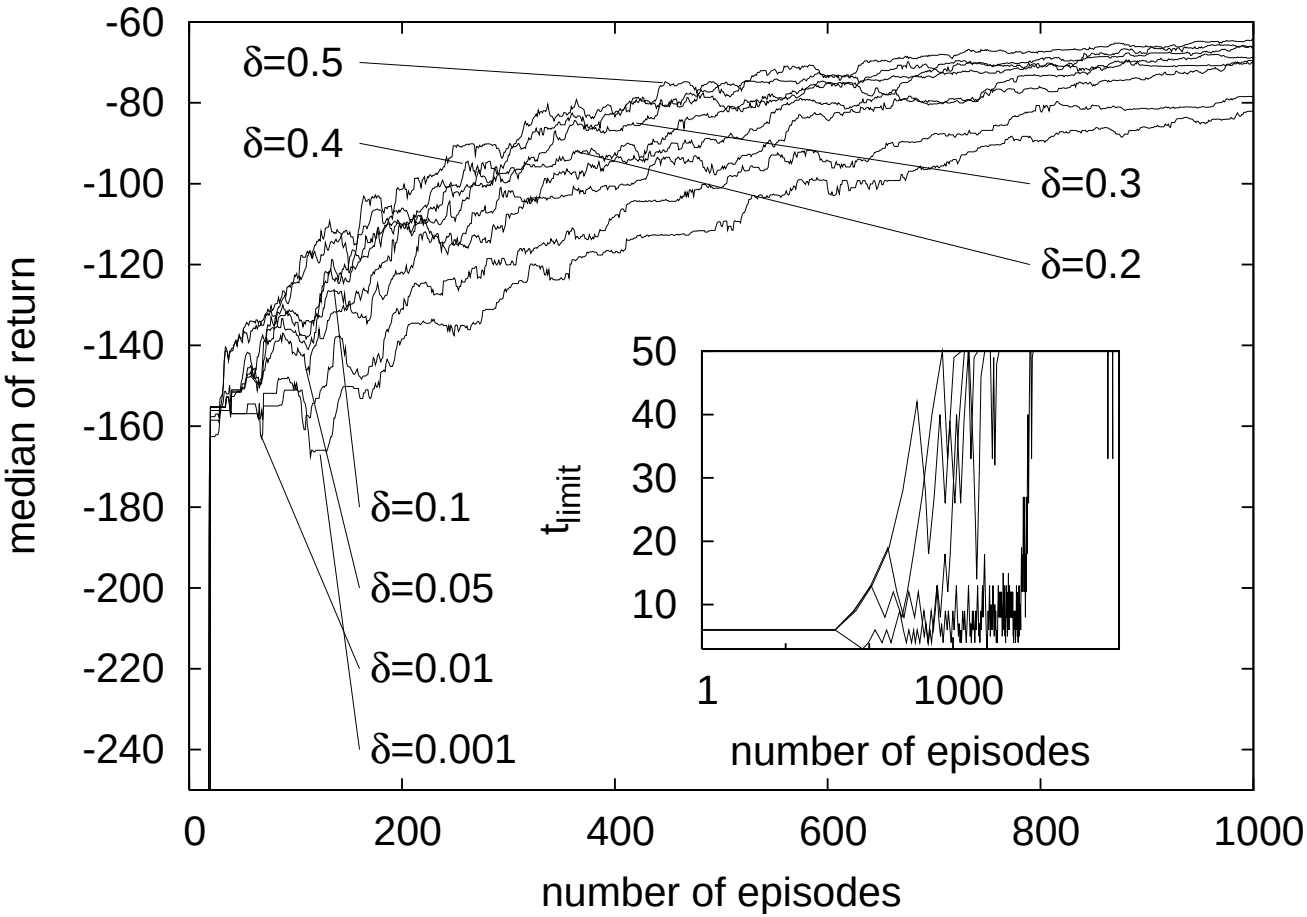
Mountain Car task without and with noisy observations





Preliminary experiments II

Mountain Car task with noise



Outline

Conclusions & Future Work

Future work will include

- more test problems,
- the combination of Hoeffding and empirical Bernstein bounds, and
- comparisons with the UH-CMA-ES for controlling the number of roll-outs.

  Hansen et al.: A Method for Handling Uncertainty in Evolutionary Optimization with an Application to Feedback Control of Combustion, IEEE TEC, 2009

Conclusions & Future Work

Future work will include

- more test problems,
- the combination of Hoeffding and empirical Bernstein bounds, and
- comparisons with the UH-CMA-ES for controlling the number of roll-outs.

  Hansen et al.: A Method for Handling Uncertainty in Evolutionary Optimization with an Application to Feedback Control of Combustion, IEEE TEC, 2009

We conclude:

1. CMA-ES is great for episodic RL.
2. Selection races provide a statistically sound, general answer to the question of how to distribute evaluations among individuals for selection under uncertainty.
3. Combining selection races and CMA-ES works even better than standard CMA-ES for episodic RL in the presence of uncertainty.