Istituto Dalle Molle di studi sull'intelligenza artificiale
IDSIA
USI/SUPSI

Parameter-based Value Functions

Francesco Faccio, Louis Kirsch and Jürgen Schmidhuber
{francesco, louis, juergen}@idsia.ch

@FaccioAI
@LouisKirschAI
@SchmidhuberAI

Problem and Motivation

- **Reinforcement Learning (RL):** find optimal policy $\pi*$
- **Policy optimization:** given a class of policies, find the policy parameters maximizing $J(\pi_\theta)$ (Sutton et al., 1999):

$$J(\pi_\theta) = \int_\mathcal{S} \mu_0(s) V^{\pi_\theta}(s)\, \mathrm{d}s = \int_\mathcal{S} \mu_0(s) \int_\mathcal{A} \pi_\theta(a|s) Q^{\pi_\theta}(s,a)\, \mathrm{d}a\, \mathrm{d}s$$

- **Problem:** value functions are defined for a single policy. During optimization, the information on previous policies is potentially lost

Off-policy RL

- Given data obtained from a behavioral policy π_b, find optimal policy π_{θ^*}
- The objective to maximize becomes:

$$J(\pi_\theta) = \int_\mathcal{S} d^{\pi_b}(s) V^{\pi_\theta}(s)\, \mathrm{d}s = \int_\mathcal{S} d^{\pi_b}(s) \int_\mathcal{A} \pi_\theta(a|s) Q^{\pi_\theta}(s,a)\, \mathrm{d}a\, \mathrm{d}s,$$

where $d^{\pi_b}(s)$ is the limiting distribution under π_b

- **Problem:** when computing $\nabla_\theta J(\pi_\theta)$, traditional off-policy policy gradients ignore $\nabla_\theta Q^{\pi_\theta}(s,a)$: the gradient of the action-value function
- When the policy is stochastic, the gradient is often **approximated** (Degris et al., 2012) by:

$$\nabla_\theta J(\pi_\theta) \approx \mathbb{E}_{s\sim d^{\pi_b}(s), a\sim\pi_b(.|s)} \left[\frac{\pi_\theta(a|s)}{\pi_b(a|s)} \left(Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s) \right) \right]$$

- When the policy is deterministic, the gradient is often **approximated** (Silver et al., 2014) by:

$$\nabla_\theta J_b(\pi_\theta) \approx \mathbb{E}_{s\sim d^{\pi_b}(s)} \left[\nabla_a Q^{\pi_\theta}(s,a)|_{a=\pi_\theta(s)} \nabla_\theta \pi_\theta(s) \right]$$

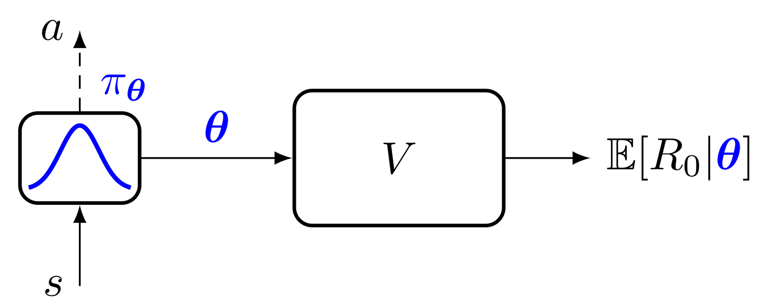
PVFs

- We augment traditional value functions by giving as input also the policy parameters
- **PSSVF**: Parameter based start-state-value function
$$V(\theta) := \mathbb{E}[R_0|\theta]$$
- **PSVF**: Parameter based state-value function
$$V(s,\theta) := \mathbb{E}[R_t|s_t = s, \theta]$$
- **PAVF**: Parameter based action-value function
$$Q(s,a,\theta) := \mathbb{E}[R_t|s_t = s, a_t = a, \theta]$$
- Parameter-based value functions (**PBVFs**) are **defined for any policy** and **can generalize** in the policy space
- The term $\nabla_\theta Q(s,a,\theta)$ can be **directly computed**
- PSSVF directly estimates the RL objective
- PSVF and PAVF are able to both **perform direct search in parameter space** AND use **Temporal Difference** for learning

PSSVF

- Stochastic or deterministic policies
- Find the policy π_θ maximizing $J(\pi_\theta)$:

$$J(\pi_\theta) = \mathbb{E}[R_0|\theta] = V(\theta)$$



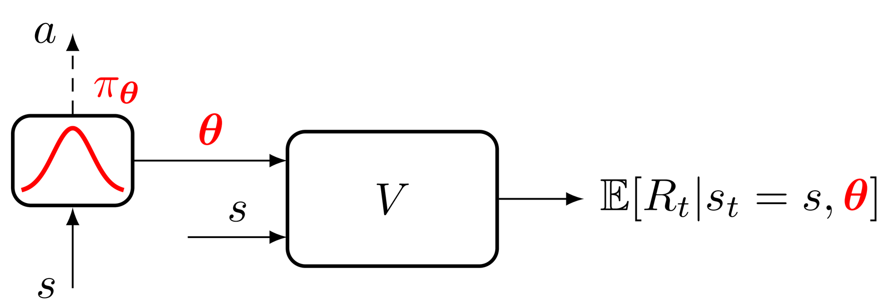
- Taking the gradient of $J(\pi_\theta)$ we obtain:

$$\nabla_\theta J(\pi_\theta) = \nabla_\theta V(\pi_\theta)$$

PSVF

- Stochastic or deterministic policies
- Find the policy π_θ maximizing $J(\pi_\theta)$:

$$J(\pi_\theta) = \int_\mathcal{S} d^{\pi_b}(s) V(s,\theta)\, \mathrm{d}s$$



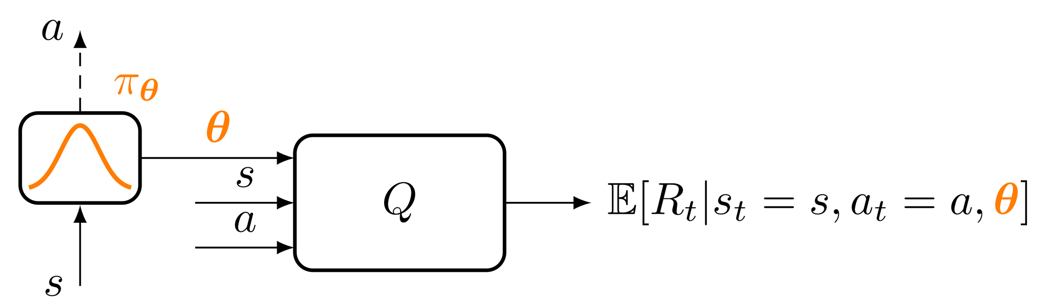
- Taking the gradient of $J(\pi_\theta)$ we obtain:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim d^{\pi_b}(s)}[\nabla_\theta V(s,\theta)]$$

Stochastic PAVF

- Stochastic policies
- Find the policy π_θ maximizing $J(\pi_\theta)$:

$$J(\pi_\theta) = \int_\mathcal{S} d^{\pi_b}(s) \int_\mathcal{A} \pi_\theta(a|s) Q(s,a,\theta)\, \mathrm{d}a\, \mathrm{d}s$$



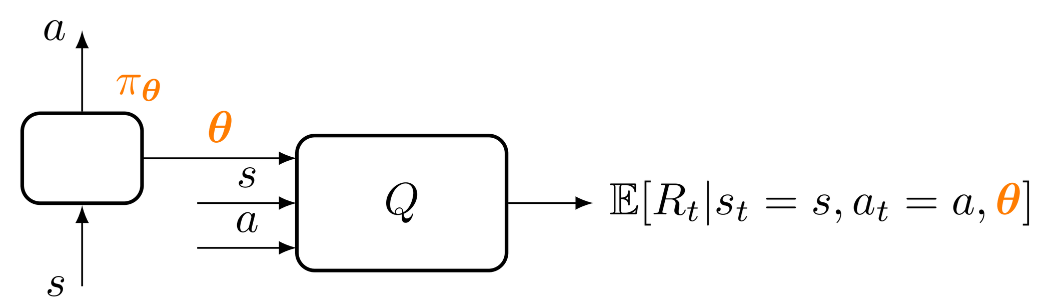
- Taking the gradient of $J(\pi_\theta)$ we obtain:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim d^{\pi_b}(s), a\sim\pi_b(.|s)} \left[\frac{\pi_\theta(a|s)}{\pi_b(a|s)} \left(Q(s,a,\theta) \nabla_\theta \log \pi_\theta(a|s) + \nabla_\theta Q(s,a,\theta) \right) \right]$$

Deterministic PAVF

- Deterministic policies
- Find the policy π_θ maximizing $J(\pi_\theta)$:

$$J(\pi_\theta) = \int_\mathcal{S} d^{\pi_b}(s) Q(s, \pi_\theta(s), \theta)\, \mathrm{d}s$$

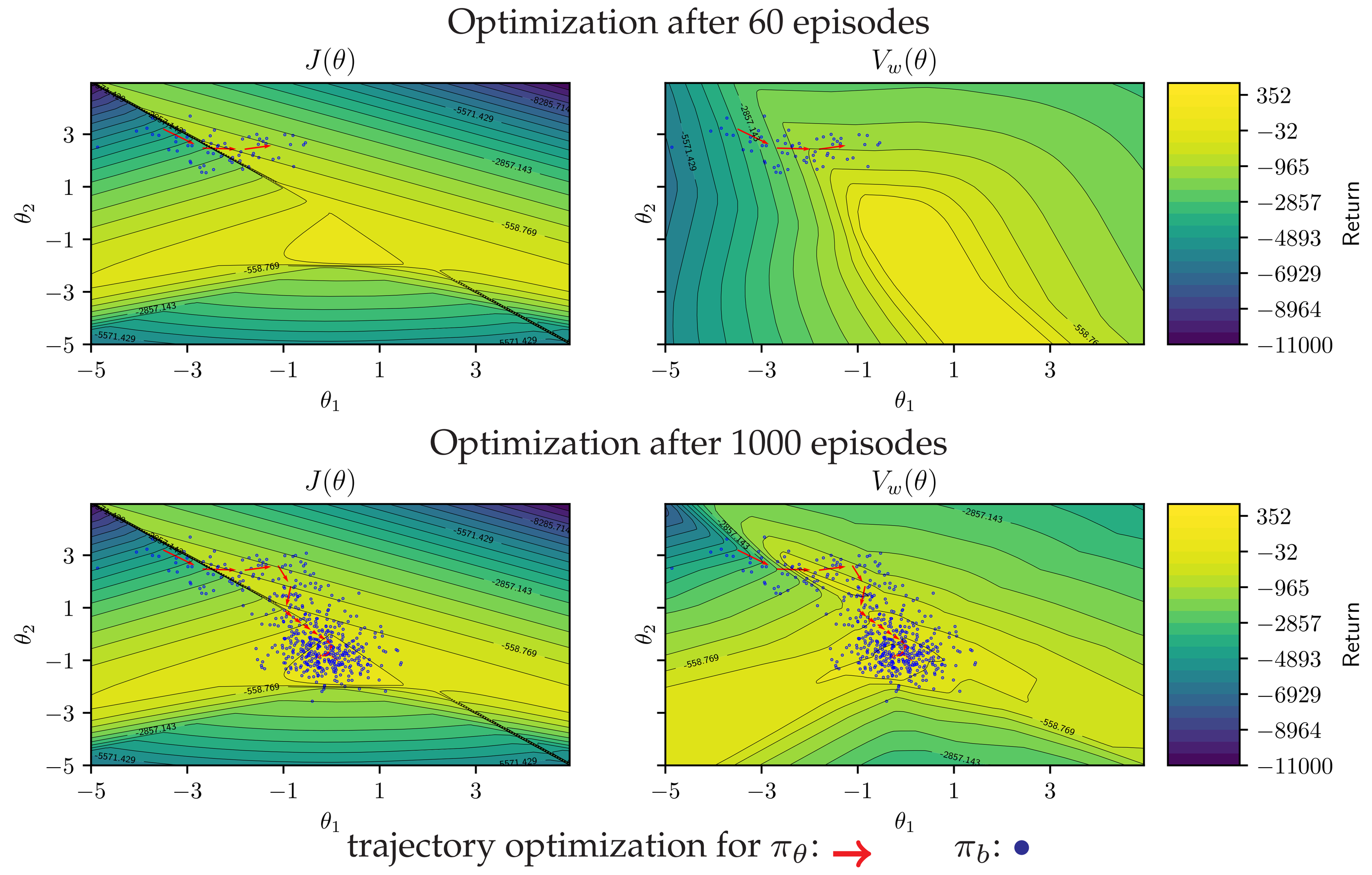


- Taking the gradient of $J(\pi_\theta)$ we obtain:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim d^{\pi_b}(s)} \left[\nabla_a Q(s,a,\theta)|_{a=\pi_\theta(s)} \nabla_\theta \pi_\theta(s) + \nabla_\theta Q(s,a,\theta)|_{a=\pi_\theta(s)} \right]$$

Actor-critic algorithms

- PSSVF on LQR using deterministic shallow policies



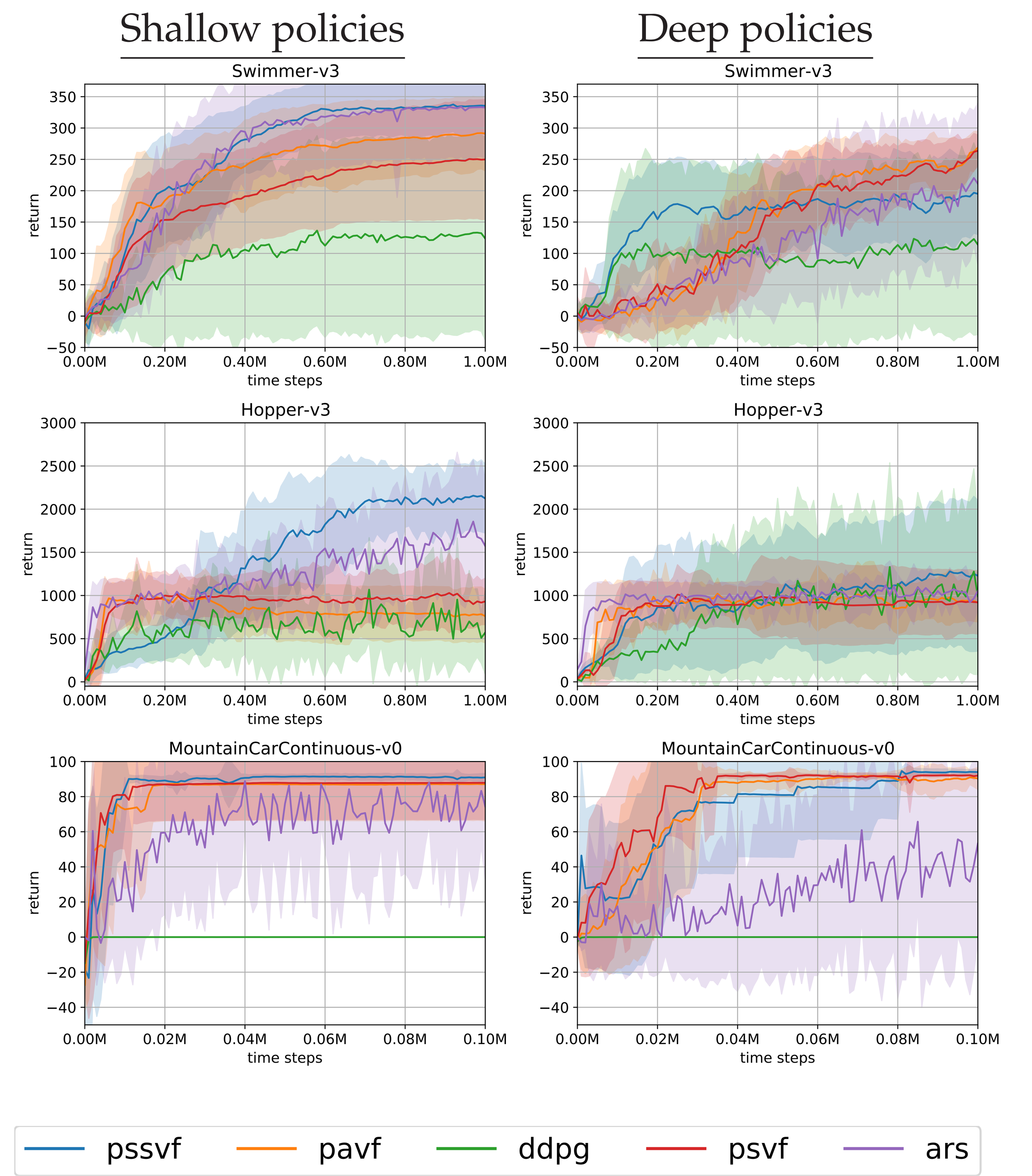
Off-policy actor-critic with PBVFs

Given the behavioral π_b, find π_θ **maximizing** $J(\theta)$:

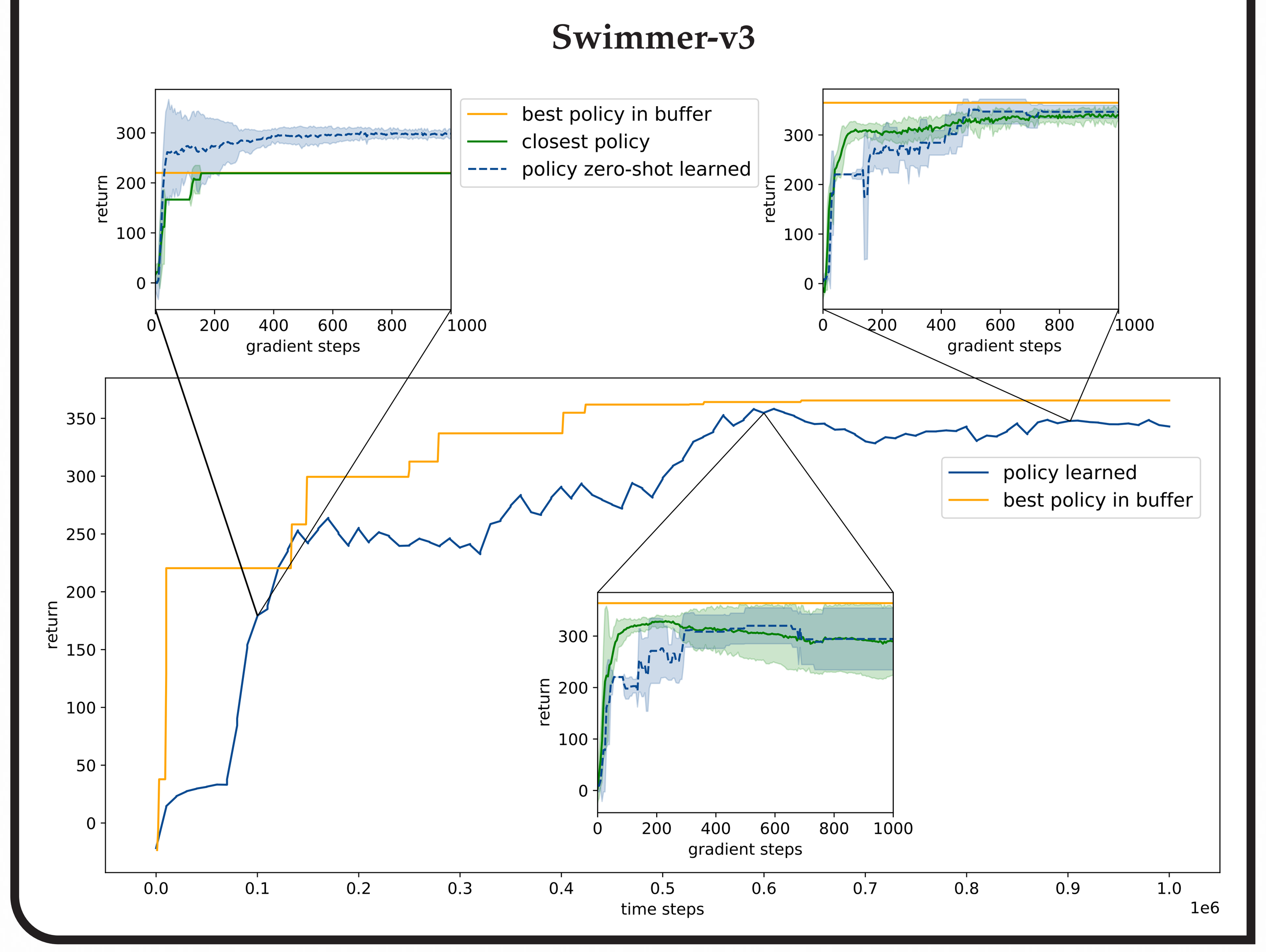
1. Collect data with π_b (expensive in RL)
2. Use data to train $V(\theta)$, $V(s,\theta)$ or $Q(s,a,\theta)$
3. Find π_θ following $\nabla_\theta J(\pi_\theta)$ (offline optimization)
4. Set new behavioral $\pi_\theta \leftarrow \pi_b$
5. Repeat until convergence

Experiments

- Comparison with DDPG (Lillicrap et al., 2015) and ARS (Mania et al., 2018) using deterministic policies



- Zero-shot learning performance of PSSVF using deterministic shallow policies



References

T. Degris, M. White, and R. S. Sutton. Off-policy actor-critic. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 179–186, USA, 2012. Omnipress. ISBN 978-1-4503-1285-1.

T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

H. Mania, A. Guy, and B. Recht. Simple random search of static linear policies is competitive for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1800–1809, 2018.

D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages I–387–I–395. JMLR.org, 2014.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 1057–1063, Cambridge, MA, USA, 1999. MIT Press.