Rank-One Modified Value Iteration

Arman S. Kolarijani [1]
Tolga Ok [1]
Peyman Mohajerin Esfahani [1, 2]
Mohamad Amin Sharif Kolarijani [1]

[1] Delft University of Technology, **DCSC,** The Netherlands

[2] University of Toronto, **MIE,** Canada

ICML International Conference On Machine Learning 2025

Generalized Policy Iteration

$$v_{k+1} = v_k + G_k \left(\mathbf{T}(v_k) - v_k \right)$$

Policy Iteration

$$G_k = (I - \gamma P_k)^{-1}$$

Value Iteration

$$G_k = I$$

$\mathbf{T}$: Bellman optimality operator

π_v : Greedy policy w.r.t v

v_k : Value function at the k-th step

$P_k := P^{\pi_{v_k}}$: Transition matrix of the greedy policy π_{v_k}

Generalized Policy Iteration

$$v_{k+1} = v_k + G_k \left(\mathbf{T}(v_k) - v_k\right)$$

Policy Iteration

$$G_k = (I - \gamma P_k)^{-1}$$

Value Iteration (VI)

$$G_k = I$$

Modified Policy Iteration

$$G_k = \sum_{i=0}^{n-1} (\gamma P_k)^i$$

λ-Policy Iteration

$$G_k = (I - \lambda\gamma P_k)^{-1}$$

Acceleration Methods

	Planning	Learning
	$v \in \mathbb{R}^n$ $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ $P^\pi \in \mathbb{R}^{n \times n}$	$q \in \mathbb{R}^{nm}$ $[\widehat{\mathbf{T}}(q, \hat{s}^+)](s,a)$ $\hat{s}^+ \sim P(\cdot \vert s,a),\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$
→	Anderson VI	
→	Nesterov VI	Speedy Q-learning
→		Zap Q-learning
→	Operator Splitting VI	
→	Rank-One VI (R1VI)	Rank-One QL (R1QL)

Rank-One Value Iteration (R1VI)

$$v_{k+1} = v_k + G_k\left(\mathbf{T}(v_k) - v_k\right)$$

Policy Iteration

$$G_k = (I - \gamma P_k)^{-1}$$

Approximate P_k

$$G_k = (I - \gamma \tilde{P}_k)^{-1}$$

Lemma 1: If the MDP is ergodic, then

Spectral radius

$$\tilde{P}_k := \mathbf{1}\boldsymbol{d}_k^\top = \arg\min_{P \in \mathbb{R}^{n\times n}} \rho(P - P_k)$$
$$\text{s.t. } P \geq 0,\ P\mathbf{1} = \mathbf{1},\ \text{rank}(P) = 1$$

Rank-one approximation

Woodbury's formula

$$G_k = (I - \gamma \tilde{P}_k)^{-1} \longrightarrow (I - \gamma \mathbf{1}\boldsymbol{d}_k^\top)^{-1} \longrightarrow I + \frac{\gamma}{1-\gamma}\mathbf{1}\boldsymbol{d}_k^\top$$

$$v_{k+1} = \mathbf{T}^{\text{R1}}(v_k) = \mathbf{T}(v_k) + \frac{\gamma}{1-\gamma}\left\langle \boldsymbol{d}_k^\top, \mathbf{T}(v_k) - v_k \right\rangle \mathbf{1}$$

Rank-One Value Iteration (R1VI)

Algorithm R1VI

initialize : $v_0 \in \mathbb{R}^n,\ \boldsymbol{d}_{-1} \in \Delta(\mathrm{R}^n)$

for $k = 0, 1, \ldots$

  compute π_{v_k} and P_k

  $\boldsymbol{d}_k = P_k^\top \boldsymbol{d}_{k-1}$

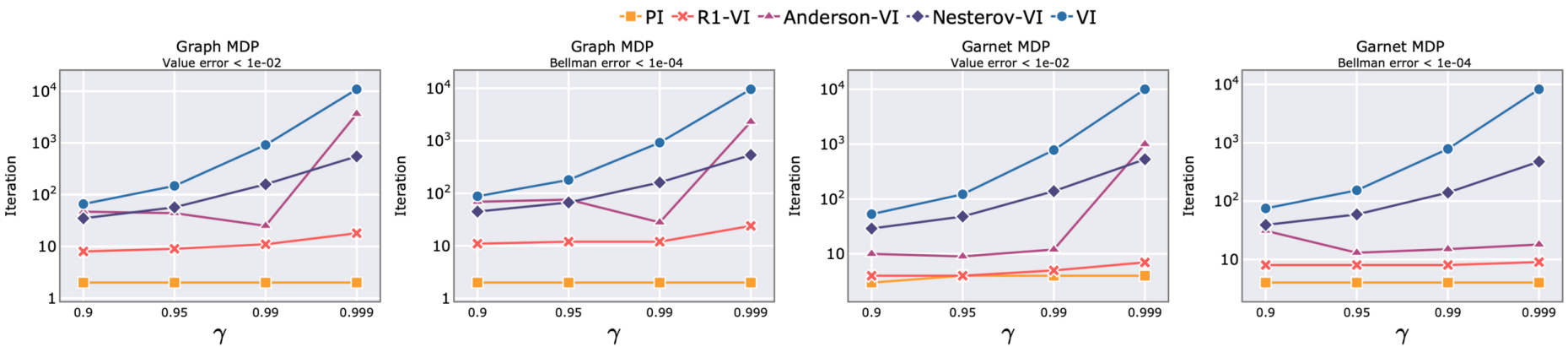
  $v_{k+1} = \mathbf{T}^{\mathrm{R1}}(v_k) = \mathbf{T}(v_k) + \dfrac{\gamma}{1-\gamma} \left\langle \boldsymbol{d}_k^\top, \mathbf{T}(v_k) - v_k \right\rangle \mathbf{1}$

Power Iteration

$\boldsymbol{d}_k^{(i+1)} = P_k \boldsymbol{d}_k^{(i)} \quad \text{for } i = 0, 1, \ldots, N$

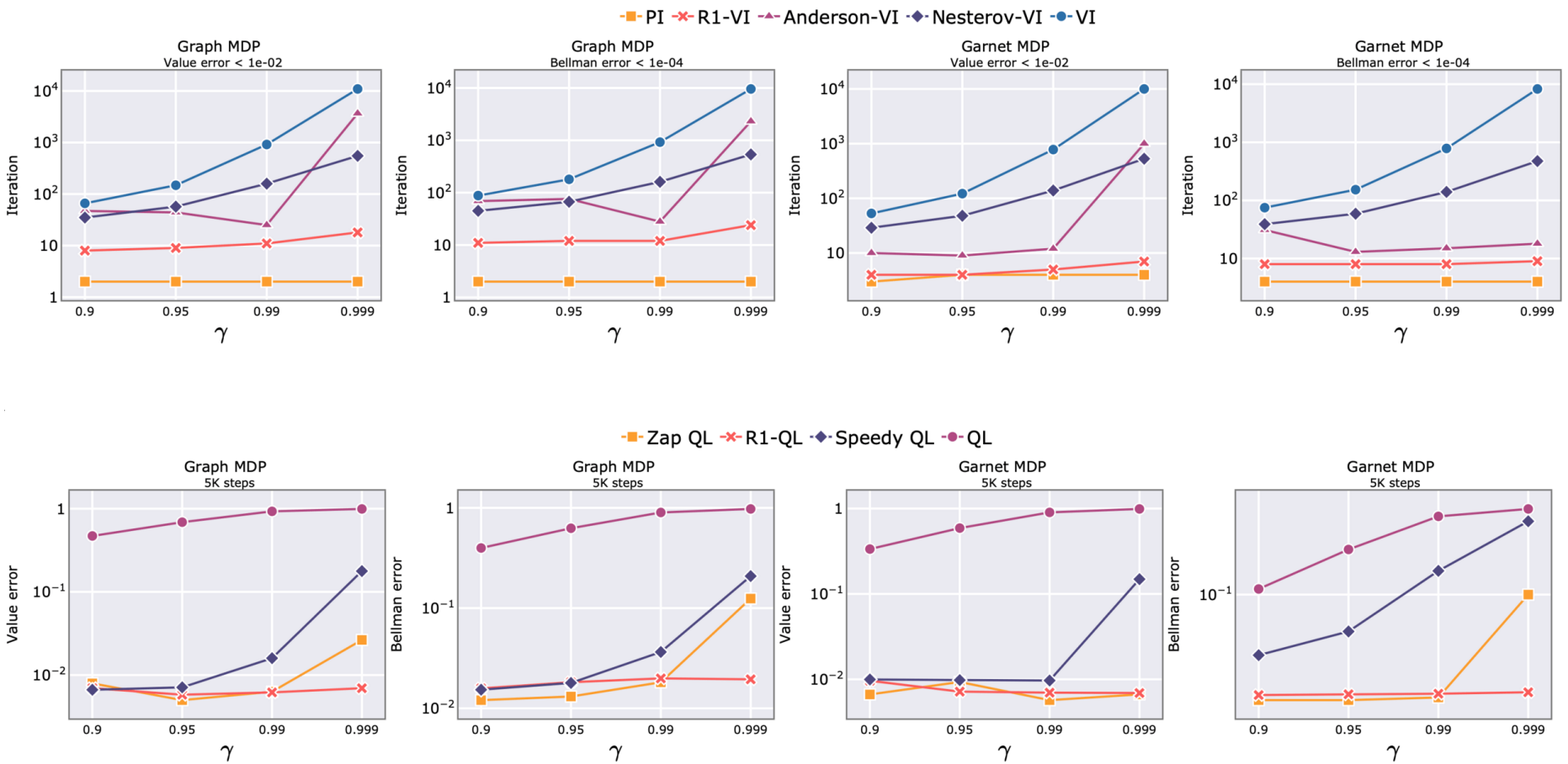
set $N \leftarrow 1$

Experiments



$\rightarrow$ R1-VI requires a similar number of iterations for larger γ values.

Experiments



Summary & Limitations

- R1-VI and R1-QL has the same computational complexity as VI and QL, respectively.
- Both R1-VI and R1-QL are convergent with rate γ, but the corresponding operators are not contractions (Theorems 3.3 & 4.2).
- Improvement in the value space, but no improvement in the policy space.
- Performance depends on the spectral gap of P_k.
- Adaptable to *asynchronous* learning (no convergence guarantee) and average reward setting.