Generalized Second Order Value Iteration in Markov Decision Processes

Chandramouli Kamanchi, Raghuram Bharadwaj Diddigi, Shalabh Bhatnagar

Friday Seminar

Li Chen

June 2022

Markov Decision Processes

Recall the infinite-horizon discounted Markov Decision Processes (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$

- state space $\mathcal{S} = \{s_1, s_2, ..., s_m\}$
- action space $\mathcal{A} = \{a_1, a_2, ..., a_n\}$
- transition kernel: $p(j|i, a)$ is the state transition probability from i to j conditioning on action a
- reward: $r(i, a)$ obtained taking action a at state i
- discount factor $\gamma \in [0, 1)$

The goal is to find a stationary (deterministic) policy $\pi : \mathcal{S} \to \mathcal{A}$ to maximize the expected reward

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t))\right]$$

where the expectation is w.r.t. the randomness of the states.

Solving MDP

Bellman optimality condition:

$$V^* = TV^*$$

where the operator $T : \mathbb{R}^m \to \mathbb{R}^m$ is defined as

$$\begin{aligned}(TV)_i \quad &\triangleq \max_{a \in \mathcal{A}} \sum_{j \in \mathcal{S}} p(j|i,a)(r(i,a) + \gamma V(j)) \\ &= \max_{a \in \mathcal{A}} r(i,a) + \gamma \sum_{j \in \mathcal{S}} p(j|i,a) V(j)\end{aligned} \qquad \forall i \in \mathcal{S}.$$

Solving MDP

Bellman optimality condition:

$$V^* = TV^*$$

- Value Iteration (VI): fixed point iteration $V_{k+1} \leftarrow TV_k$, which converges linearly since T is γ-contractive

$$\|TV - TV'\|_\infty \leq \gamma \|V - V'\|_\infty$$

Solving MDP

Bellman optimality condition:

$$V^* = TV^*$$

- Value Iteration (VI): fixed point iteration $V_{k+1} \leftarrow TV_k$, which converges linearly since T is γ-contractive

$$\|TV - TV'\|_\infty \leq \gamma \|V - V'\|_\infty$$

- Policy Iteration (PI): policy evaluation + policy improvement
  - Given a policy π_k, solve a linear system

    $$T^{\pi_k} V = V$$

    to obtain V^{π_k} where $(T^\pi V)_i \triangleq r(i, \pi(i)) + \gamma \sum_{j \in \mathcal{S}} p(j|i, \pi(i)) V(j)$.
  - Find an improved policy π_{k+1} by greedy strategy such that

    $$T^{\pi_{k+1}} V^{\pi_k} = TV^{\pi_k},$$

    i.e., $\pi_{k+1}(i) \in \arg\max_{a \in \mathcal{A}} r(i, a) + \gamma \sum_{j \in \mathcal{S}} p(j|i, a) V^{\pi_k}(j)$.

Solving MDP

Bellman optimality condition:

$$V^* = TV^*$$

- Value Iteration (VI): fixed point iteration $V_{k+1} \leftarrow TV_k$, which converges linearly since T is γ-contractive

$$\|TV - TV'\|_\infty \leq \gamma \|V - V'\|_\infty$$

- Policy Iteration (PI): policy evaluation + policy improvement
- Linear Programming (LP): growing interests, e.g., [BSCKN21]

$$\begin{aligned} \min \quad & \boldsymbol{e}^\top V \\ \text{s.t.} \quad & V_i \geq r(i,a) + \gamma \sum_{j \in \mathcal{S}} p(j|i,a) V_j \quad \forall i \in \mathcal{S}, a \in \mathcal{A} \end{aligned} \qquad \text{(dual)}$$

$$\begin{aligned} \max \quad & \sum_{i \in \mathcal{S}, a \in \mathcal{A}} r_{ia} u_{ia} \\ \text{s.t.} \quad & \sum_{a \in \mathcal{A}} u_{ia} - \gamma \sum_{s \in \mathcal{S}, a \in \mathcal{A}} p(i|s,a) u_{sa} = 1 \quad \forall i \in \mathcal{S} \\ & u_{ia} \geq 0 \qquad \forall i \in \mathcal{S}, a \in \mathcal{A} \end{aligned} \qquad \text{(primal)}$$

Q-value functions

- Q-value function is the function of state-action pairs

$$Q(i,a) = r(i,a) + \gamma \sum_{j \in \mathcal{S}} p(j|i,a) V(i)$$

- The optimality condition (Q-Bellman equation) is

$$Q^*(i,a) = r(i,a) + \gamma \sum_{j \in \mathcal{S}} p(j|i,a) \max_{a \in \mathcal{A}} Q^*(j,a)$$

  Then the optimality policy is $\pi^*(i) \in \arg\max_{a \in \mathcal{A}} Q^*(i,a)$.

- Note Q-Bellman equation is linear in probability transition, which makes it popular for model-free settings (reinforcement learning).
- Preserve γ-contraction

Generalized Q-Bellman equation

- This paper mainly focus on a generalized Q-Bellman equation

$$Q_w(i,a) = w\left(r(i,a) + \gamma \sum_{j\in\mathcal{S}} p(j|i,a) \max_{a\in\mathcal{A}} Q_w(j,a)\right) + (1-w) \max_{a\in\mathcal{A}} Q_w(i,a)$$

  where $w \in (0, w^*]$ and $w^* = \dfrac{1}{1-\gamma \min_{i\in\mathcal{S},a\in\mathcal{A}} p(i|i,a)}$ based on the idea of successive over-relaxation (SOR).

- The optimal Q_w^* may be different from Q^*, but the optimal value functions are the same, i.e.,

$$\max_{a\in\mathcal{A}} Q_w^*(i,a) = \max_{a\in\mathcal{A}} Q^*(i,a), \quad \forall i \in \mathcal{S}$$

- The goal of this paper is to apply Newton's method to solve the generalized Q-Bellman equation with smoothing.

Smoothing

Basic idea: Approximate $\max_{i \in [m]} \{x_i\}$ by $\frac{1}{N} \log \sum_{i=1}^{m} \exp(N x_i)$ with $N > 0$

- can be understood as entropy regularization for the dual

Lemma 1

Let $f(x) = \max_{i \in [m]} \{x_i\}$ *and* $g_N(x) = \frac{1}{N} \log \sum_{i=1}^{m} \exp(N x_i)$*, then*

$$\sup_{x \in \mathbb{R}^m} |f(x) - g_N(x)| \to 0 \text{ as } N \to \infty$$

- Indeed, $\sup_{x \in \mathbb{R}^m} |f(x) - g_N(x)| \leq \left| \frac{\log m}{N} \right|$
- Note that $\frac{\partial g_N}{\partial x_i} = \frac{\exp(N x_i)}{\sum_{\ell=1}^{m} \exp(N x_\ell)}$, so $\|\nabla g_N(x)\|_1 \leq 1$ and g_N is non-expansive w.r.t. $\|\cdot\|_\infty$

Contractive properties

Given $w \in (0, w^*]$ and $N > 0$, define the modified Successive Q-Bellman (SQB) operator $U : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ as

$$(UQ)(i,a) = w\left(r(i,a) + \gamma \sum_{j\in\mathcal{S}} p(j|i,a) g_N(Q(j,:))\right) + (1-w) g_N(Q(i,:)),\ \forall i, a$$

where $Q(i,:) = [Q(i,a)]_{a\in\mathcal{A}} \in \mathbb{R}^n$.

Lemma 2

The operator U is a $(1 - w + w\gamma)$-contraction under $\|\cdot\|_\infty$-norm.

Contractive properties

Given $w \in (0, w^*]$ and $N > 0$, define the modified Successive Q-Bellman (SQB) operator $U : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ as

$$(UQ)(i,a) = w\left(r(i,a) + \gamma \sum_{j \in \mathcal{S}} p(j|i,a) g_N(Q(j,:))\right) + (1-w) g_N(Q(i,:)),\ \forall i, a$$

where $Q(i,:) = [Q(i,a)]_{a \in \mathcal{A}} \in \mathbb{R}^n$.

Lemma 2

The operator U is a $(1 - w + w\gamma)$-contraction under $\|\cdot\|_\infty$-norm.

Proof. For any P, Q, calculate

$$\begin{aligned}
& |UP(i,a) - UQ(i,a)| \\
= & \left| w\gamma \sum_{j \in \mathcal{S}} p(j|i,a)[g_N(Q(j,:)) - g_N(P(j,:))] + (1-w)[g_N(Q(i,:)) - g_N(P(i,:))] \right| \\
= & (1 - w + w\gamma) \left| \mathbb{E}_{\mathbb{Q}} \left[g_N(Q(\tilde{j},:)) - g_N(P(\tilde{j},:)) \right] \right| \\
\leq & (1 - w + w\gamma) \mathbb{E}_{\mathbb{Q}} \left[\left| g_N(Q(\tilde{j},:)) - g_N(P(\tilde{j},:)) \right| \right] \\
\leq & (1 - w + w\gamma) \mathbb{E}_{\mathbb{Q}} \left[\max_{a \in \mathcal{A}} \left| (Q(\tilde{j},a)) - P(\tilde{j},a) \right| \right] \\
\leq & (1 - w + w\gamma) \max_{j \in \mathcal{S}, a \in \mathcal{A}} |(Q(j,a)) - P(j,a)|
\end{aligned}$$

Contractive properties

Given $w \in (0, w^*]$ and $N > 0$, define the modified Successive Q-Bellman (SQB) operator $U : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ as

$$(UQ)(i,a) = w \left(r(i,a) + \gamma \sum_{j \in \mathcal{S}} p(j|i,a) g_N(Q(j,:)) \right) + (1-w) g_N(Q(i,:)), \ \forall i, a$$

where $Q(i,:) = [Q(i,a)]_{a \in \mathcal{A}} \in \mathbb{R}^n$.

Lemma 2

The operator U is a $(1 - w + w\gamma)$-contraction under $\| \cdot \|_\infty$-norm.

- Benefit of SOR: $1 - w + w\gamma < \gamma$ whenever $w > 1$.
- U has a unique fixed point

Error from smoothing

Lemma 3

Let Q_w^* be the solution of the generalized Q-Bellman equation, Q' be the fixed point of U, then

$$\|Q_w^* - Q'\|_\infty \leq \frac{1 - w + w\gamma}{Nw(1-\gamma)} \log n.$$

- Note $\dfrac{1 - w + w\gamma}{w} < \gamma$ whenever $w > 1$.

Error from smoothing

Lemma 3

Let Q_w^* be the solution of the generalized Q-Bellman equation, Q' be the fixed point of U, then

$$\|Q_w^* - Q'\|_\infty \leq \frac{1 - w + w\gamma}{Nw(1-\gamma)} \log n.$$

Proof. By def we have

$$Q'(i,a) = wr(i,a) + (1 - w + w\gamma)\mathbb{E}_{\mathbb{Q}}\left[g_N(Q'(\tilde{j},:))\right], \ \forall i, a$$
$$Q_w^*(i,a) = wr(i,a) + (1 - w + w\gamma)\mathbb{E}_{\mathbb{Q}}\left[\max_{b\in\mathcal{A}}\left\{Q_w^*(\tilde{j},b)\right\}\right], \ \forall i, a$$

Let $Q'(Z,c) = \max_{b\in\mathcal{A}} Q'(Z,b)$ where $Z \sim \mathbb{Q}$, then

$$\begin{aligned}
& |Q_w(i,a) - Q'(i,a)| \\
= \ & (1 - w + w\gamma)\left|\mathbb{E}_{\mathbb{Q}}\left[\max_{b\in\mathcal{A}}\left\{Q_w^*(\tilde{j},b)\right\} - g_N(Q'(\tilde{j},:))\right]\right| \\
= \ & (1 - w + w\gamma)\left|\mathbb{E}_{\mathbb{Q}}\left[\max_{b\in\mathcal{A}}\left\{Q_w^*(\tilde{j},b)\right\} - \max_{b\in\mathcal{A}}\left\{Q'(\tilde{j},b)\right\} - g_N(Q'(\tilde{j},:) - Q'(\tilde{j},c))\right]\right. \\
\leq \ & (1 - w + w\gamma)\mathbb{E}_{\mathbb{Q}}\left[\left|\max_{b\in\mathcal{A}}\left\{Q_w^*(\tilde{j},b)\right\} - \max_{b\in\mathcal{A}}\left\{Q'(\tilde{j},b)\right\}\right| + \left|g_N(Q'(\tilde{j},:) - Q'(\tilde{j},c))\right|\right.
\end{aligned}$$

The Algorithm: G-SOVI

The algorithm applies Newton's method to solve the smooth equation

$$F(Q) = 0 \text{ with } F \triangleq I - U$$

by

$$Q_{k+1} \leftarrow Q_k - (I - J_U(Q_k))^{-1}(Q_k - UQ_k)$$

where the Jacobian of U is

$$\begin{aligned} J_U(Q)_{ia,jc} &= (w\gamma p(j|i,a) + (1-w)\delta_{i,j}) \frac{\exp(NQ(j,c))}{\sum_{b\in\mathcal{A}} \exp(NQ(j,b)} \\ &= (1 - w + w\gamma) q(j|i,a) \frac{\exp(NQ(j,c))}{\sum_{b\in\mathcal{A}} \exp(NQ(j,b)} \end{aligned}$$

where $q(j|i,a) = \dfrac{w\gamma p(j|i,a) + (1-w)\delta_{i,j}}{1 - w + w\gamma}$

The Algorithm: G-SOVI

The algorithm applies Newton's method to solve the smooth equation

$$F(Q) = 0 \text{ with } F \triangleq I - U$$

by

$$Q_{k+1} \leftarrow Q_k - (I - J_U(Q_k))^{-1}(Q_k - UQ_k)$$

where the Jacobian of U is

$$\begin{aligned} J_U(Q)_{ia,jc} &= (w\gamma p(j|i,a) + (1-w)\delta_{i,j}) \frac{\exp(NQ(j,c))}{\sum_{b\in\mathcal{A}} \exp(NQ(j,b)} \\ &= (1 - w + w\gamma) q(j|i,a) \frac{\exp(NQ(j,c))}{\sum_{b\in\mathcal{A}} \exp(NQ(j,b)} \end{aligned}$$

where $q(j|i,a) = \dfrac{w\gamma p(j|i,a) + (1-w)\delta_{i,j}}{1 - w + w\gamma}$

- Pure Newton step
- Stopping condition: maximal iteration
- Inverting $I - J_U(Q_k)$ is related to the policy evaluation step in policy iteration since $J_U(Q_k)$ is a (row) stochastic matrix scaled by $w\gamma + 1 - w \in (0,1)$. (Newton's method is equivalent to policy iteration [PB79]).

Global convergence

Theorem 1 (Global Newton Theorem [OR00])

Suppose

- $F : \mathbb{R}^d \to \mathbb{R}^d$ *is continuous differentiable, component-wise concave on* $\mathbb{R}^d$*,*
- $F'(x)$ *is non-singular and* $F'(x)^{-1} \geq 0$ *(non-negative) for all* $x \in \mathbb{R}^d$*,*
- $F(x) = 0$ *has a unique solution* x^**.*

Then for any $x_0 \in \mathbb{R}^d$ *the Newtons iterates converges to* x^**.*

Global convergence

Theorem 1 (Global Newton Theorem [OR00])

Suppose

- $F : \mathbb{R}^d \to \mathbb{R}^d$ *is continuous differentiable, component-wise concave on* $\mathbb{R}^d$,
- $F'(x)$ *is non-singular and* $F'(x)^{-1} \geq 0$ *(non-negative) for all* $x \in \mathbb{R}^d$,
- $F(x) = 0$ *has a unique solution* x^*.

Then for any $x_0 \in \mathbb{R}^d$ *the Newtons iterates converges to* x^*.

Theorem 2

Let Q' *be the unique fixed point of* U, *then the G-SOVI algorithm converges to* Q' *for any choice of initial point* Q_0.

Proof of Theorem 2

We need to verify the conditions of Theorem 1 for $F = I - U$. Recall

$$(UQ)(i,a) = wr(i,a) + (1 - w + w\gamma)\mathbb{E}_{\mathbb{Q}}\left[g_N(Q(\tilde{j},:))\right], \ \forall i,a$$

$$J_U(Q)_{ia,jc} = (1 - w + w\gamma)q(j|i,a)\frac{\exp(NQ(j,c))}{\sum_{b\in\mathcal{A}}\exp(NQ(j,b))} \ \forall i,a,j,c.$$

- $I - U$ is continuous differentiable, component-wise concave
- Note $J_U(Q) \in \mathbb{R}^{mn\times mn}$ is $(1 - w + w\gamma)\Phi$ with Φ as a row stochastic matrix and $1 - w + w\gamma \in (0,1)$. Then $(I - J_U(Q))^{-1}$ exists (Proof later) and

$$(I - J_U(Q))^{-1} = \sum_{\ell=0}^{\infty}(1 - w + w\gamma)^{\ell}\Phi^{\ell}$$

  so that $(I - J_U(Q))^{-1} \geq 0$.
- $F(Q) = 0$ has unique solution as U is $(1 - w + w\gamma)$-contractive.

Quadratic convergence

Lemma 4

The inverse Jacobian is bounded: $\|(I - J_U(Q))^{-1}\| \leq \dfrac{1}{w(1-\gamma)}$.

Quadratic convergence

Lemma 4

The inverse Jacobian is bounded: $\|(I - J_U(Q))^{-1}\| \leq \dfrac{1}{w(1-\gamma)}$.

Proof. Note that $I - J_U(Q) = I - (1 - w + \gamma w)\Phi$ where Φ is a row stochastic matrix ($\Phi \boldsymbol{e} = \boldsymbol{e}$). So its eigenvalue λ is bounded by

$$0 < 1 - (1 - w + \gamma w) \leq |\lambda|.$$

Hence, $(I - J_U(Q))^{-1}$ exists and $\|(I - J_U(Q))^{-1}\| \leq \dfrac{1}{1 - (1 - w + \gamma w)}$.

Quadratic convergence

Lemma 4

The inverse Jacobian is bounded: $\|(I - J_U(Q))^{-1}\| \leq \dfrac{1}{w(1-\gamma)}$.

Theorem 3

The G-SOVI algorithm converges quadratically.

Quadratic convergence

Lemma 4

The inverse Jacobian is bounded: $\|(I - J_U(Q))^{-1}\| \leq \dfrac{1}{w(1-\gamma)}$.

Theorem 3

The G-SOVI algorithm converges quadratically.

Proof. Let Q^* be the fixed point of $F(Q^*) = 0$, then

$$\begin{aligned}
\|Q_{k+1} - Q^*\| &= \|Q_k - F'(Q_k)^{-1}F(Q_k) - Q^*\| \\
&= \|Q_k - Q^* - F'(Q_k)^{-1}(F(Q_k) - F(Q^*))\| \\
&= \|F'(Q_k)^{-1}[F'(Q_k)(Q_k - Q^*) + F(Q^*) - F(Q_k)]\| \\
&\leq \|F'(Q_k)^{-1}\| \cdot \|F(Q^*) - F(Q_k) - F'(Q_k)(Q^* - Q_k)\| \\
&\leq \frac{1}{w(1-\gamma)} \cdot \frac{L}{2}\|Q^* - Q_k\|^2
\end{aligned}$$

where L is the Lipschitz constant of the mapping $F'(\cdot)$.

- L may depend on N

Experiments

Compare G-SOVI, SOVI ($w = 1$), and standard VI.

- The result is averaged over 100 instances, $\gamma = 0.9$.
- The error at iteration k is $E(k) = \max_{i \in \mathcal{S}} |V^*(i) - \max_{a \in \mathcal{A}} Q_k(i, a)|$

Value of N	Standard Value Iteration	Standard SOVI	G-SOVI
N=20	0.1009 ± 0.0026	0.1205 ± 0.0372	0.1093 ± 0.0818
N=25		0.0822 ± 0.0273	0.0648 ± 0.0217
N=30		0.0611 ± 0.0211	0.0494 ± 0.017
N=35		0.0484 ± 0.0168	0.0397 ± 0.0136

Table I: Comparison of Average Error for different values of N on 10 states and 5 actions setting at the end of 50 iterations. For the G-SOVI algorithm, the relaxation parameter is chosen to be the optimal relaxation parameter w^*, i.e., $w = w^*$.

Experiments

Compare G-SOVI, SOVI ($w = 1$), and standard VI.

- The result is averaged over 100 instances, $\gamma = 0.9$.
- The error at iteration k is $E(k) = \max_{i \in \mathcal{S}} |V^*(i) - \max_{a \in \mathcal{A}} Q_k(i, a)|$

Value of w	**G-SOVI**
$w = 1$ (Standard SOVI)	0.04838 ± 0.017
$w = 1.00001$	0.04838 ± 0.017
$w = 1.0001$	0.04837 ± 0.017
$w = 1.001$	0.04830 ± 0.017
$w = 1.01$	0.0476 ± 0.017
$w = 1.05$	0.0448 ± 0.016
$w = 1.1$	0.0417 ± 0.014
$w = w^*$	0.0397 ± 0.014

Table II: Comparison of Average Error in G-SOVI for different values of w on 10 states and 5 actions setting at the end of 50 iterations. The value of N is 35.

Experiments

Compare G-SOVI, SOVI ($w = 1$), and standard VI.

- The result is averaged over 100 instances, $\gamma = 0.9$.
- The error at iteration k is $E(k) = \max_{i \in \mathcal{S}} |V^*(i) - \max_{a \in \mathcal{A}} Q_k(i, a)|$

Setting	Standard Value Iteration	Standard SOVI	G-SOVI
States = 30, Actions= 10	6.471 $\pm$ 0.07	0.087 $\pm$ 0.01	0.079 $\pm$ 0.01
States = 50, Actions = 10	6.587 $\pm$ 0.07	0.114 $\pm$ 0.01	0.108 $\pm$ 0.01
States = 80, Actions = 10	6.754 $\pm$ 0.03	0.141 $\pm$ 0.01	0.136 $\pm$ 0.01
States = 100, Actions = 10	6.772 $\pm$ 0.03	0.152 $\pm$ 0.01	0.148 $\pm$ 0.01

Table III: Comparison of Average Error across four settings at the end of 10 iterations with $N = 35$. For the G-SOVI alg
the relaxation parameter is chosen to be the optimal relaxation parameter w^*, i.e., $w = w^*$.

Experiments

Compare G-SOVI, SOVI ($w = 1$), and standard VI.

- The result is averaged over 100 instances, $\gamma = 0.9$.
- The error at iteration k is $E(k) = \max_{i \in \mathcal{S}} |V^*(i) - \max_{a \in \mathcal{A}} Q_k(i,a)|$

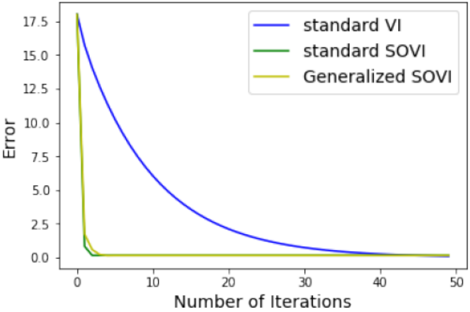

Figure 1: Error vs Number of iterations on setting with 100 states and 10 actions with $w = w^*$ for Generalized SOVI (G-SOVI).

Experiments

Compare G-SOVI, SOVI ($w = 1$), and standard VI.

- The result is averaged over 100 instances, $\gamma = 0.9$.
- The error at iteration k is $E(k) = \max_{i \in \mathcal{S}} |V^*(i) - \max_{a \in \mathcal{A}} Q_k(i, a)|$

Setting	Standard Value Iteration	Standard SOVI	G-SOVI
States = 30, Actions= 10	0.0008 ± 0.00	0.0154 ± 0.01	0.0267 ± 0.01
States = 50, Actions = 10	0.0009 ± 0.00	0.0242 ± 0.00	0.0488 ± 0.00
States = 80, Actions = 10	0.0011 ± 0.00	0.0532 ± 0.00	0.0988 ± 0.01
States = 100, Actions = 10	0.0026 ± 0.00	0.1202 ± 0.01	0.1343 ± 0.01

Table IV: Per-iteration Execution time of algorithms across four settings in seconds, with the relaxation parameter in chosen as $w = w^*$.

- Inverting Hessian seems expensive

Experiments

Compare G-SOVI, SOVI ($w = 1$), and standard VI.

- The result is averaged over 100 instances, $\gamma = 0.9$.
- The error at iteration k is $E(k) = \max_{i \in \mathcal{S}} |V^*(i) - \max_{a \in \mathcal{A}} Q_k(i, a)|$

Configuration	Computational Time (in seconds)	Standard Value Iteration	Standard SOVI	G-SOVI
10 States, 5 Actions	0.01	25.485 ± 2.21	3.930± 0.92	3.885 ± 0.94
20 States, 5 Actions	0.02	18.291 ± 0.77	5.444 ± 0.51	5.473 ± 0.50
30 States, 5 Actions	0.03	7.327 ± 0.20	7.111 ± 0.32	7.118 ± 0.33

Table V: Average Error vs Computational Time (rounded off to the nearest millisecond). Initial Q-values for algorit assigned random integers between 60 and 70. The discount factor is set to 0.99. G-SOVI is run with $w = 1.00001$.

- This comparison is chosen to make SOVI better (only 3 iterations)

Conclusion

- Smoothing + Newton's method works and has fast convergence
- Smoothing does not solve the original MDP
- Inverting the Jacobian ($mn \times mn$) directly seems expensive, working on value function might be better
- The (Generalized) Bellman operator is piecewise linear, smoothing may not be necessary
- Growing research in accelerating value iteration, see also [GC21, GGC22]

References I

- Joan Bas-Serrano, Sebastian Curi, Andreas Krause, and Gergely Neu, *Logistic q-learning*, International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 3610–3618.
- Julien Grand-Clément, *From convex optimization to mdps: A review of first-order, second-order and quasi-newton methods for mdps*, arXiv preprint arXiv:2104.10677 (2021).
- Vineet Goyal and Julien Grand-Clement, *A first-order approach to accelerated value iteration*, Operations Research (2022).
- James M Ortega and Werner C Rheinboldt, *Iterative solution of nonlinear equations in several variables*, SIAM, 2000.
- Martin L Puterman and Shelby L Brumelle, *On the convergence of policy iteration in stationary dynamic programming*, Mathematics of Operations Research **4** (1979), no. 1, 60–69.