


# Dynamic Programming Through the Lens of Semismooth Newton-Type Methods

**Journal Article****Author(s):**

Gargiani, Matilde; Zanelli, Andrea; Liao-McPherson, Dominic; Summers, Tyler H.; [Lygeros, John](#) 

**Publication date:**

2022

**Permanent link:**

<https://doi.org/10.3929/ethz-b-000558296>

**Rights / license:**

[In Copyright - Non-Commercial Use Permitted](#)

**Originally published in:**

IEEE Control Systems Letters 6, <https://doi.org/10.1109/LCSYS.2022.3181213>

# Dynamic Programming Through the Lens of Semismooth Newton-Type Methods

M. Gargiani, A. Zanelli, D. Liao-McPherson, T. H. Summers and J. Lygeros,

**Abstract**—Policy iteration and value iteration are at the core of many (approximate) dynamic programming methods. For Markov Decision Processes with finite state and action spaces, we show that they are instances of semismooth Newton-type methods for solving the Bellman equation. In particular, we prove that policy iteration is equivalent to the exact semismooth Newton method and enjoys a local quadratic convergence rate. This finding is corroborated by extensive numerical evidence in the fields of control and operations research, which confirms that policy iteration generally requires relatively few iterations to achieve convergence even in presence of a large number of admissible policies. We then show that value iteration is an instance of the fixed-point iteration method and develop a novel locally accelerated version of value iteration with global convergence guarantees and negligible extra computational costs.

## I. INTRODUCTION

Approximate dynamic programming (ADP) is a powerful algorithmic strategy to handle stochastic sequential decision making problems arising in a wide range of applications, from control to games and resource allocation, to name a few. At the core of some of the biggest success stories of ADP is an approximate version of policy iteration [17]. In particular, after an extensive offline training phase where an approximation of the optimal cost is produced, one iteration of an approximate version of policy iteration is performed (online learning). Empirical evidence suggests that this final step greatly enhances performance. In particular, Bertsekas links these success stories to the equivalence between policy iteration and Newton’s method [2].

The connection between policy iteration and Newton’s method dates back to the late 60’s [12]. Puterman and Brumelle [13] were among the first who exploited this connection to study the convergence properties of policy iteration for MDPs with continuous action spaces. More recently, Santos and Ruts [16] exploited this connection to analyze the asymptotic convergence of policy iteration for the discretization of a specific class of Markov Decision Processes (MDPs) with continuous spaces. Bertsekas in [2] provides a graphical analysis of the connection between policy iteration and Newton’s method and mathematically formalizes these visual insights by proving local quadratic convergence of policy iteration for MDPs with finite state and action spaces. These theoretical results are corroborated by numerous computational examples which demonstrate that policy iteration achieves convergence in a remarkably small number of iterations even in presence of rounding errors and a large number of potential policies. We refer to [2] for an extensive review of the related works.

In this work, we consider MDPs with finite state and action spaces and formally show that policy iteration and value iteration are both instances of semismooth Newton-type methods. The main differences between our analysis and that in [2] are that the latter focus only on policy iteration and does not deploy tools from generalized differentiation, but instead works in a neighborhood of the solution where the iterates can be expressed as the Newton iterates for an auxiliary continuously differentiable mapping. We then exploit this connection to develop a novel variant of value iteration inspired by the fixed-point iteration method. In particular, our main contributions are: *i*) we develop a unified theoretical analysis for the local convergence of semismooth Newton-type methods based on the so-called *kappa condition* [4]; *ii*) we formalize the connection of policy iteration and value iteration with semismooth Newton-type methods using tools from generalized differentiation and results from Section II. After discussing the significant algorithmic and theoretical implications of this connection, *iii*) we design a globally convergent and locally accelerated variant of value iteration with negligible additional computational cost per iteration and superior numerical performance.

**Notation.** In the following, we use  $\|\cdot\| : \mathbb{R}^d \rightarrow \mathbb{R}$  to denote an arbitrary vector norm,  $\|\cdot\| : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$  for its induced matrix norm,  $\mathcal{B}(c, \delta)$  for the Euclidean ball with center  $c \in \mathbb{R}^d$  and radius  $\delta > 0$ ,  $\rho$  for the spectral radius of a matrix,  $r'$  for the Jacobian operator of a differentiable function  $r : \mathbb{R}^d \rightarrow \mathbb{R}^d$ ,  $\text{cl}(\mathcal{T})$  and  $\text{int}(\mathcal{T})$  for the closure and the interior of a set  $\mathcal{T} \subseteq \mathbb{R}^d$ , respectively.

## II. BACKGROUND

We consider infinite horizon discounted cost problems for MDPs  $\{\mathcal{S}, \mathcal{A}, P, g, \gamma\}$  comprising a finite state space  $\mathcal{S} = \{1, \dots, n\}$ , a finite action space  $\mathcal{A} = \{1, \dots, m\}$ , a transition probability function  $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$  that defines the probability of ending in state  $s'$  when applying action  $a$  in state  $s$ , a stage-cost function  $g : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  that associates to each state-action pair a bounded cost, and a discount factor  $\gamma \in (0, 1)$ . Throughout the paper, with a slight abuse of notation we use  $\mathcal{A}(s)$  to denote the nonempty subset of actions that are allowed at state  $s$ ,  $p_{ss'}(a) = P(s, a, s')$  for the probability of transitioning to state  $s'$  when the system is in state  $s$  and action  $a \in \mathcal{A}(s)$  is selected with  $\sum_{s' \in \mathcal{S}} p_{ss'}(a) = 1$  for all  $s \in \mathcal{S}$  and  $a \in \mathcal{A}(s)$ .

A *deterministic stationary control policy*  $\pi : \mathcal{S} \rightarrow \mathcal{A}$  is a function that maps states to actions, with  $\pi(s) \in \mathcal{A}(s)$ . We use  $\Pi$  to denote the set of all deterministic stationary control policies, from now on simply *policies*. At step

$t$  of the decision process under the policy  $\pi \in \Pi$ , the system is in some state  $s_t$  and the action  $a_t = \pi(s_t)$  is applied. The discounted cost  $\gamma^t g(s_t, a_t)$  is accrued and the system transitions to a state  $s_{t+1}$  according to the probability distribution  $P(s_t, a_t, \cdot)$ . This process is repeated leading to the following cumulative discounted cost

$$V^\pi(s) = \lim_{T \rightarrow \infty} \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma^t g(s_t, \pi(s_t)) \mid s_0 = s \right], \quad (1)$$

where  $\{s_0, \pi(s_0), s_1, \pi(s_1), \dots, s_t, \pi(s_t), \dots\}$  is the state-action sequence generated by the MDP under policy  $\pi$  with initial state  $s_0$ , and the expected value is taken with respect to the corresponding probability measure over the space of sequences. The transition probability distributions induced by policy  $\pi$  can be compactly represented by the rows of an  $n \times n$  row-stochastic matrix  $[P^\pi]_{ss'} = p_{ss'}(\pi(s))$  for all  $s, s' \in \mathcal{S}$  and the costs induced by policy  $\pi$  by the vector  $g^\pi = [g(1, \pi(1)) \ \dots \ g(n, \pi(n))]^\top \in \mathbb{R}^n$ . The optimal cost is defined as

$$V^*(s) := \min_{\pi \in \Pi} V^\pi(s) \quad \forall s \in \mathcal{S}. \quad (2)$$

Any policy  $\pi^* \in \Pi$  that attains the optimal cost is called an optimal policy. Notice that in (2) we restrict our attention to stationary deterministic policies as in our setting there exists a policy in this class that attains  $V^*$  [1]. The optimal cost admits a *recursive* definition known as the *Bellman equation*

$$V^*(s) = \min_{a \in \mathcal{A}(s)} \left\{ g(s, a) + \gamma \sum_{s' \in \mathcal{S}} p_{ss'}(a) V^*(s') \right\} \quad \forall s \in \mathcal{S}. \quad (3)$$

Equation (1) admits an analogous recursive definition known as the Bellman equation associated with policy  $\pi$ . In the considered setting, the cost function associated with policy  $\pi$  and the optimal cost function can be represented by  $V^\pi \in \mathbb{R}^n$  and  $V^* \in \mathbb{R}^n$ , where the  $s$ -th element is given by (1) and (2) evaluated at  $s$ , respectively.

### A. Dynamic Programming

Dynamic Programming (DP) comprises methods for solving stochastic optimal control problems by solving the Bellman equation [1]. We are specifically interested in the value iteration (VI) and policy iteration (PI) algorithms. Starting from Equation (3), we define a nonsmooth mapping  $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ , known as the *Bellman operator*, by

$$(TV)(s) = \min_{a \in \mathcal{A}(s)} \left\{ g(s, a) + \gamma \sum_{s' \in \mathcal{S}} p_{ss'}(a) V(s') \right\} \quad \forall s \in \mathcal{S}.$$

An analogous linear operator  $T^\pi : \mathbb{R}^n \rightarrow \mathbb{R}^n$  can be defined for the Bellman equation associated with policy  $\pi$  as

$$(T^\pi V)(s) = g(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p_{ss'}(\pi(s)) V(s') \quad \forall s \in \mathcal{S}.$$

Given the cost vector  $V$ , any policy  $\pi$  such that  $\forall s \in \mathcal{S}$

$$\pi(s) \in \arg \min_{a \in \mathcal{A}(s)} \left\{ g(s, a) + \gamma \sum_{s' \in \mathcal{S}} p_{ss'}(a) V(s') \right\} \quad (4)$$

is called *greedy* with respect to the cost  $V$ . It can be shown [1] that the Bellman operator is contractive and, thanks to the Banach Theorem [15], admits a unique fixed point  $V^*$ . Moreover, the corresponding Picard-Banach iteration  $V_{k+1} = TV_k$  converges asymptotically to the fixed point from any initial value  $V_0$ , i.e.,  $\lim_{k \rightarrow \infty} T^k V_0 = V^*$ . This is at the core of VI, which repeatedly applies the  $T$  operator starting from an arbitrary finite cost. The generated sequence linearly converges to  $V^*$  with a  $\gamma$ -contraction rate.

An alternative method to solve Equation (3) is PI. With PI, we start from an arbitrary initial policy and alternate *policy evaluation* and *policy improvement* until convergence. The policy evaluation step at iteration  $k$  computes the cost  $V^{\pi_k}$  associated with the current policy  $\pi_k$ , i.e.,  $V^{\pi_k} = (I - \gamma P^{\pi_k})^{-1} g^{\pi_k}$ . This requires the solution of a system with  $n$  linear equations, which is generally computationally demanding for MDPs with large state spaces. The policy  $\pi_{k+1}$  is then updated by extracting a greedy policy associated with  $V^{\pi_k}$  in the policy improvement step (see Equation (4)). Unlike VI, PI converges in a finite number of iterations since the policy, and therefore also its cost, are improved at each iteration and since, by the finiteness of  $\mathcal{S}$  and  $\mathcal{A}$ , there only exists a finite number of policies. It is nonetheless important to characterize its convergence rate and asymptotic behavior since, for large state and action spaces, the number of iterations could be prohibitive (exponential in  $n$  and  $m$ ). By exploiting the properties of the Bellman operator, we can show that PI is globally  $\gamma$ -contractive, which is similar to VI. Extensive empirical evidence, however, suggests that PI has superior convergence properties and generally requires considerably fewer iterations than VI. From a computational viewpoint, the per-iteration costs of PI with direct inversion amount to  $\mathcal{O}(n^3 + m \cdot n^2)$  versus the  $\mathcal{O}(m \cdot n^2)$  of VI.

### B. Semismooth Newton-Type Methods

Consider the following nonlinear root finding problem

$$r(\theta) = 0, \quad (5)$$

where  $r : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is a locally Lipschitz-continuous vector-valued function. A vector  $\theta^* \in \mathbb{R}^d$  that satisfies (5) is called *root* or *solution* of (5). In general, we can not rely on smooth root finding methods [9] to solve (5) since  $r$  can be nonsmooth, so its Jacobian  $r'(\theta) \in \mathbb{R}^{d \times d}$  might not exist. We therefore need to employ some notions of generalized differentiability from nonsmooth analysis [3], such as the *B-differential* and Clarke's *generalized Jacobian*. Since  $r$  is a locally Lipschitz-continuous map, the Rademacher Theorem [14] implies that it is differentiable almost everywhere and we denote by  $\mathcal{M}_r$  the set of all points where  $r$  is differentiable. Another fundamental implication of the Rademacher Theorem is the definition of the B-differential of  $r$  at  $\theta \in \mathbb{R}^d$  as the set

$$\partial_B r(\theta) = \{J \in \mathbb{R}^{d \times d} \mid \exists \{\theta_k\} \subset \mathcal{M}_r : \{\theta_k\} \rightarrow \theta, \{r'(\theta_k)\} \rightarrow J\}.$$

We denote with  $\partial r(\theta)$  Clarke's generalized Jacobian of  $r$  at  $\theta \in \mathbb{R}^d$ , which is defined as the convex hull of  $\partial_B r(\theta)$ . Consequently,  $\partial_B r(\theta) \subseteq \partial r(\theta)$ . These sets are

always nonempty when evaluated at points where the function is Lipschitz continuous [9, Proposition 1.51]. If  $r$  is continuously differentiable at  $\theta$ , then  $\partial r(\theta) = \partial_{BR}(\theta) = \{r'(\theta)\}$ . Otherwise,  $\partial_{BR}(\theta)$  and, consequently,  $\partial r(\theta)$  are not necessarily singletons.

The B-differential and Clarke's generalized Jacobian are of practical interest only if we can compute at least some of their elements. Because they lack sharp calculus rules, this can be done only in few cases, depending on the structure of  $r$ . For instance, consider the class of piecewise continuously differentiable functions on  $\mathbb{R}^d$  [10], which is formally characterized by the following definition.

*Definition 2.1 (PC<sup>1</sup> Functions):* Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}^o$  be a continuous vector-valued function and  $n_p$  be some positive integer. The function  $f$  is said to be *piecewise continuously differentiable* of order 1 (PC<sup>1</sup>) if there exist finitely many continuously differentiable functions  $\{f_i\}_{i=1}^{n_p}$  on  $\mathbb{R}^d$ , called *selection functions*, such that  $f(\theta) \in \{f_i(\theta)\}_{i=1}^{n_p}$  for all  $\theta \in \mathbb{R}^d$ . In addition,  $f_i$  is *active* at  $\bar{\theta} \in \mathbb{R}^n$  if  $f(\bar{\theta}) = f_i(\bar{\theta})$  and *essentially active* if  $\bar{\theta} \in \text{cl}(\text{int}(\{\theta \in \mathbb{R}^d : f(\theta) = f_i(\theta)\}))$ .

We denote with  $\mathcal{F}_f(\bar{\theta})$  the collection of essentially active functions at  $\bar{\theta}$ . Piecewise affine functions are an example of PC<sup>1</sup> functions with affine selection functions and are particularly relevant in the context of DP as will be discussed in Section III.

The following proposition [10, Lemma 2.10] gives a representation of the B-differential for PC<sup>1</sup> functions. This representation can be used to determine a  $J \in \partial_B f(\theta)$  in cases where we can compute the Jacobian matrix of at least one of the essentially active selection functions at  $\theta \in \mathbb{R}^d$ .

*Proposition 2.2:* Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}^o$  be a PC<sup>1</sup> function. The B-differential of  $f$  at  $\theta \in \mathbb{R}^d$  is  $\partial_B f(\theta) = \{f'_i(\theta) : f_i \in \mathcal{F}_f(\theta)\}$ .

*Example 2.3:* Consider the following piecewise affine function:  $f(\theta) = 2\theta - 5$  if  $\theta > 5$ ,  $f(\theta) = \theta$  if  $\theta = 5$  and  $f(\theta) = -2\theta + 15$  if  $\theta < 5$ . Then  $\partial_B f(5) = \{2, -2\}$  since  $\text{int}(\{\theta \in \mathbb{R} : f(\theta) = \theta\}) = \emptyset$  and  $\partial_B f(\theta) = f'(\theta)$  for all  $\theta \in \mathbb{R} \setminus \{5\}$ .

We refer to [10] for more details on the computation of elements in Clarke's generalized Jacobian for piecewise continuous functions and to [9, Ch. 1] for functions with different structures.

Newton's method [9] is not directly applicable to solve (5) when  $r$  is nonsmooth. One notable extension of Newton's method to nonsmooth equations dates back to [11] and is known as the semismooth Newton method [9]. Similarly to Newton's method, instead of solving directly (5), the semismooth Newton method solves a series of linear equations that locally approximate (5), but the Jacobian matrix in the Newtonian iteration system is replaced by an element from Clarke's generalized Jacobian. In particular, the semismooth Newton method generates a sequence of iterates  $\{\theta_k\}$  where  $\theta_0 \in \mathbb{R}^d$  is the initial approximation of the root and, for any  $k \geq 0$ ,  $\theta_{k+1}$  satisfies the linear equation  $r(\theta_k) + J_k(\theta_{k+1} - \theta_k) = 0$ , with  $J_k \in \partial r(\theta_k)$ . When  $J_k$  is nonsingular, the iterate  $\theta_{k+1}$  can be computed in closed-form

as follows

$$\theta_{k+1} = \theta_k - J_k^{-1}r(\theta_k). \quad (6)$$

Under certain assumptions, the semismooth Newton method enjoys fast local quadratic convergence, but the cost per iteration with direct inversion is in the order of  $\mathcal{O}(d^3)$ . In addition, as discussed previously, it may be difficult to obtain an element from Clarke's generalized Jacobian. These are some of the main motivations behind the design of different variants of the semismooth Newton method of the form

$$r(\theta_k) + B_k(\theta_{k+1} - \theta_k) = 0, \quad (7)$$

where  $B_k \in \mathbb{R}^{d \times d}$  somehow approximate  $J_k$ . These variants, collectively known as semismooth Newton-type methods [9], can lead to lower computational costs while maintaining acceptable convergence rates. Clearly, if  $B_k \in \partial r(\theta_k)$ , then we recover the semismooth Newton method. One of the most frequently used semismooth Newton-type methods is the *fixed-point iteration method* where  $B_k = \alpha_k I$  with  $\alpha_k \neq 0$  [5].

Before proceeding with the formal characterization of the local convergence of semismooth Newton-type methods, we have to introduce the notions of strong semismoothness [9, Subsection 1.4.2] and CD-regularity [9, Remark 1.65].

*Definition 2.4 (Strong Semismoothness):* A function  $f : \mathbb{R}^d \rightarrow \mathbb{R}^o$  is strongly semismooth at  $\theta \in \mathbb{R}^d$  if it is locally Lipschitz-continuous at  $\theta$ , directionally differentiable at  $\theta$  in every direction, and the following estimate holds as  $\xi \in \mathbb{R}^d$  tends to zero

$$\sup_{J \in \partial f(\theta + \xi)} \|f(\theta + \xi) - f(\theta) - J\xi\| = \mathcal{O}(\|\xi\|^2).$$

*Definition 2.5 (CD/BD-Regularity):* A function  $f : \mathbb{R}^d \rightarrow \mathbb{R}^o$  is CD-regular (BD-regular) at  $\theta \in \mathbb{R}^d$  if each matrix  $J \in \partial f(\theta)$  ( $J \in \partial_B f(\theta)$ ) is nonsingular.

The function in Example 2.3 is strongly semismooth and BD-regular everywhere, but not CD-regular at  $\theta = 5$ .

The following theorem characterizes the local contraction of a semismooth Newton-type sequence generated by Algorithm 1. Similar *a-posteriori* results based on perturbation analysis can be found in [9].

*Theorem 2.6:* Let  $r : \mathbb{R}^d \rightarrow \mathbb{R}^d$  be strongly semismooth at  $\theta^* \in \mathbb{R}^d$ ,  $L > 0$  and  $\kappa \in [0, 1)$  a constant. Then there exist an open neighborhood of  $\theta^*$  such that, for any  $\theta_0$  in the neighborhood and any sequence of nonsingular matrices  $\{B_k\} \subseteq \mathbb{R}^{d \times d}$  such that, for all  $k$ ,  $\|B_k^{-1}\| \leq L$  and  $\exists J_k \in \partial r(\theta_k)$  for which the kappa condition

$$\|B_k^{-1}(B_k - J_k)\| \leq \kappa_k \leq \kappa \quad (8)$$

is verified, the sequence  $\{\theta_k\} \subseteq \mathbb{R}^d$  generated by Algorithm 1 converges to  $\theta^*$  and

$$\|\theta_{k+1} - \theta^*\| \leq \kappa_k \|\theta_k - \theta^*\| + \mathcal{O}(\|\theta_k - \theta^*\|^2). \quad (9)$$

*Proof:* See the extended version [7].

Theorem 2.6 shows that the local convergence rate of semismooth Newton-type methods strongly depends on the choice of  $\{B_k\}$ . In particular, we obtain quadratic convergence if  $\kappa = 0$ , superlinear convergence if  $\kappa_k \rightarrow 0$  as  $k \rightarrow \infty$  and linear convergence if  $\kappa_k = \kappa$  for all  $k$  with  $\kappa \in (0, 1)$ .

The following corollary characterizes the local convergence of the exact semismooth Newton method [9, Theorem 2.42].

*Corollary 2.7:* Let  $r$  be strongly semismooth and CD-regular at  $\theta^*$ . Provided that  $\theta_0$  is close enough to  $\theta^*$ , the sequence  $\{\theta_k\}$  generated by the semismooth Newton method iteration (6) with starting point  $\theta_0$  converges to  $\theta^*$  and

$$\|\theta_{k+1} - \theta^*\| = \mathcal{O}(\|\theta_k - \theta^*\|^2).$$

*Proof:* See the extended version [7].

---

### Algorithm 1 Semismooth Newton-Type Method

---

- 1: **Initialization:** select  $\theta_0 \in \mathbb{R}^d$ ,  $tol \geq 0$  and set  $k = 0$
  - 2: **while**  $\|r(\theta_k)\| > tol$  **do**
  - 3:     select  $B_k \in \mathbb{R}^{d \times d}$  nonsingular and compute
 
$$\theta_{k+1} = \theta_k - B_k^{-1}r(\theta_k) \quad (10)$$
  - 4:      $k \leftarrow k + 1$
  - 5: **end while**
- 

### III. SEMISMOOTH NEWTON-TYPE DYNAMIC PROGRAMMING

In this section, we formalize the connection between PI, VI and semismooth Newton-type methods.

We start by looking at the Bellman equation (3) as a nonlinear root finding problem, where  $r(\theta) = \theta - T\theta$  and  $r : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . We call  $r$  the *Bellman residual function*.

Looking at the set of the admissible policies and based on the relation between  $T$  and  $T^\pi$ , we can rewrite the Bellman residual function as follows

$$r(\theta) = \theta - \min_{\pi \in \Pi} \{T^\pi \theta\} = \theta - \min_{\pi \in \Pi} \{g^\pi + \gamma P^\pi \theta\}, \quad (11)$$

where  $T^\pi \theta = g^\pi + \gamma P^\pi \theta$  is an affine function of  $\theta$ . The Bellman residual function is piecewise affine since it is continuous and there exist  $|\Pi|$  affine selection functions  $\{\theta - T^\pi \theta\}_{\pi \in \Pi}$  such that  $r(\theta) \in \{\theta - T^\pi \theta\}_{\pi \in \Pi}$  for all  $\theta \in \mathbb{R}^n$ . Because of its piecewise affine structure, the Bellman residual function is globally Lipschitz continuous (Proposition 4.2.2 in [6]) and strongly semismooth everywhere (Proposition 7.4.7 in [5]).

The following lemma characterizes the relation between greedy policies and active selection functions at  $\theta \in \mathbb{R}^n$ .

*Lemma 3.1:* Let  $\tilde{\Pi}_\theta \subseteq \Pi$  denote the set of the greedy policies with respect to the cost-vector  $\theta \in \mathbb{R}^n$ . Then  $r(\theta) = \theta - T^\pi \theta$  for all  $\pi \in \tilde{\Pi}_\theta$ . In other terms,  $\{\theta - T^\pi \theta\}_{\pi \in \tilde{\Pi}_\theta}$  is the collection of the active selection functions of  $r$  at  $\theta$ .

*Proof:* The proof follows directly from the definition of greedy policy (4). In particular, a policy  $\pi$  is greedy with respect to the cost-vector  $\theta \in \mathbb{R}^n$  if  $T^\pi \theta = T\theta$ . ■

The next definition introduces the concept of *spurious greedy policy*, which will later be used together with Proposition 2.2 to characterize the B-differential of the Bellman residual function.

*Definition 3.2 (Spurious Greedy Policy):* Let  $\bar{\theta} \in \mathbb{R}^n$ .  $\pi \in \tilde{\Pi}_{\bar{\theta}}$  is a spurious greedy policy for the cost-vector  $\bar{\theta}$  if  $\text{int}(\{\theta \in \mathbb{R}^n : r(\theta) = \theta - T^\pi \theta\}) = \emptyset$ .

In other terms, a greedy policy  $\pi \in \tilde{\Pi}_\theta$  is spurious if there exist  $s \in \mathcal{S}$  for which for all  $\epsilon > 0$ ,  $\pi(s)$  is not greedy with respect to any  $\tilde{\theta}_s \neq \theta_s$  with  $|\theta_s - \tilde{\theta}_s| \leq \epsilon$ . We denote with  $\tilde{\Pi}_\theta^S$  the subset of  $\tilde{\Pi}_\theta$  comprising the spurious greedy policies.

The next proposition characterizes the B-differential of the Bellman residual function.

*Proposition 3.3:* Let  $r : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be the Bellman residual function. The B-differential of  $r$  at  $\theta \in \mathbb{R}^n$  is the set

$$\partial_{Br}(\theta) = \left\{ I - \gamma P^\pi \mid \forall \pi \in \tilde{\Pi}_\theta \setminus \tilde{\Pi}_\theta^S \right\}. \quad (12)$$

In addition,  $r$  is globally CD-regular.

*Proof:* From the definition of essentially active selection functions and spurious greedy policies, it follows that  $\mathcal{F}_r(\theta) = \left\{ \theta - T^\pi \theta \mid \forall \pi \in \tilde{\Pi}_\theta \setminus \tilde{\Pi}_\theta^S \right\}$ . From Proposition 2.2 and since  $(\theta - T^\pi \theta)' = I - \gamma P^\pi$  for any  $\pi \in \Pi$ , we conclude that the B-differential of  $r$  is given by the set in (12). Since  $P^\pi$  is a row-stochastic matrix, its eigenvalues lie within the unit circle of the complex plane. Thus  $I - \gamma P^\pi$  with  $\gamma \in (0, 1)$  has no eigenvalue equal to zero. We can therefore conclude that all the matrices in the B-differential of  $r$  are nonsingular and therefore  $r$  is BD-regular. Finally, since the convex combination of row stochastic matrices is a row stochastic matrix, we can conclude that  $r$  is CD-regular. ■

#### A. Policy Iteration

We start by introducing an assumption on the sets of the spurious greedy policies, which excludes the presence of selection functions that are active but not essentially active.

*Assumption 3.4:* We assume that  $\tilde{\Pi}_\theta^S = \emptyset$  for all  $\theta \in \mathbb{R}^n$ .

The following proposition characterizes the connection between PI and the semismooth Newton method.

*Proposition 3.5:* Under Assumption 3.4, PI is an instance of the semismooth Newton method to solve the Bellman residual function. Hence, the local contraction is quadratic.

*Proof:* Let  $\{\theta_k^{\text{PI}}\}$  denote the iterates of PI. We show by induction that, through an appropriate choice of  $J_k$ , we can generate iterates  $\{\theta_k^{\text{N}}\}$  of the semismooth Newton method for the Bellman residual function such that  $\theta_k^{\text{PI}} = \theta_k^{\text{N}}$  for all  $k$ . Assume that  $\theta_k^{\text{PI}} = \theta_k^{\text{N}} = \theta_k$  and let  $\pi_{k+1} \in \tilde{\Pi}_{\theta_k}$  be the greedy policy selected by PI at the  $k$ -th policy improvement step. Then it follows that  $\theta_{k+1}^{\text{PI}} = (I - \gamma P^{\pi_{k+1}})^{-1} g^{\pi_{k+1}}$ . From Assumption 3.4 and Proposition 3.3, we have that  $I - \gamma P^{\pi_{k+1}}$  is invertible and belongs to  $\partial_{Br}(\theta_k)$ . Recall in addition that, from the definition of greedy policy,  $T^{\pi_{k+1}} \theta_k = T\theta_k$ . Therefore, the  $(k+1)$ -th semismooth Newton iterate with  $J_k = I - \gamma P^{\pi_{k+1}}$  is

$$\begin{aligned} \theta_{k+1}^{\text{N}} &= \theta_k - (I - \gamma P^{\pi_{k+1}})^{-1} r(\theta_k) \\ &= \theta_k - (I - \gamma P^{\pi_{k+1}}) (\theta_k - g^{\pi_{k+1}} - \gamma P^{\pi_{k+1}} \theta_k) \\ &= \theta_k - (I - \gamma P^{\pi_{k+1}})^{-1} ((I - \gamma P^{\pi_{k+1}}) \theta_k - g^{\pi_{k+1}}) \\ &= (I - \gamma P^{\pi_{k+1}})^{-1} g^{\pi_{k+1}} \\ &= \theta_{k+1}^{\text{PI}}. \end{aligned}$$

The quadratic local contraction follows from Corollary 2.7. ■

The theoretical results of Proposition 3.5 are corroborated by extensive empirical evidence that suggests that, in practice, PI leads to faster convergence in terms of number of iterations than VI [1], [8]. Despite its simplicity, the algorithmic implications of Proposition 3.5 are significant, especially in light of the results in Theorem 2.6. We can develop novel DP methods in the spirit of semismooth Newton-type methods, where the elements in the B-differential are approximated with non-singular matrices that verify the kappa condition (8). Assumption 3.4 allows to directly employ Proposition 2.2 and could be further relaxed by considering only the iterates  $\theta_k$ . In addition, despite its technicality and limited intuitiveness, empirical evidence seems to suggest that it is realistic to assume that  $\tilde{\Pi}_{\theta_k}^S = \emptyset$  for all  $\theta_k$ . A more detailed discussion on Assumption 3.4 and Proposition 3.5 is available in the extended version [7].

### B. Value Iteration

In light of the equivalence between PI and the semismooth Newton method to solve (11), we investigate the connection between VI and semismooth Newton-type methods. In particular, with the following proposition we show that VI is a semismooth Newton-type method where the elements in Clarke's generalized Jacobian are approximated with the identity matrix.

*Proposition 3.6:* VI is a semismooth Newton-type method to solve the Bellman residual function with  $\{B_k\} = \{I\}$ .

*Proof:* Let  $\theta_{k+1}^{\text{VI}}$  and  $\theta_{k+1}^{\text{N-type}}$  denote the  $(k+1)$ -th iterate of VI and the semismooth Newton-type method with  $\{B_k\} = \{I\}$ , respectively. Assume that  $\theta_k^{\text{VI}} = \theta_k^{\text{N-type}} = \theta_k$ . Then, from the definition of VI, it follows that  $\theta_{k+1}^{\text{VI}} = T\theta_k$ . From the definition of semismooth Newton-type iterate in (10) and with the specific choice of  $B_k = I$ , we obtain that  $\theta_{k+1}^{\text{N-type}} = \theta_k - I^{-1}r(\theta_k) = \theta_k - (\theta_k - T\theta_k) = T\theta_k = \theta_{k+1}^{\text{VI}}$ . ■

The classical DP convergence analysis of VI based on the properties of the Bellman operator indicates that VI enjoys a global linear rate of convergence with a  $\gamma$ -contraction rate. In light of this novel connection between VI and the fixed-point iteration method, we can adopt the semismooth Newton-type theory perspective to study the local convergence of VI. In particular, from the results of Theorem 2.6, we obtain that VI has a local linear contraction rate given by the discount factor as  $\|I^{-1}(I - (I - \gamma P^\pi))\|_\infty = \gamma\|P^\pi\|_\infty = \gamma < 1$  for all  $\pi \in \Pi$ .

### C. $\alpha$ -Value Iteration

Proposition 3.6 shows that VI is also an instance of the fixed-point iteration method with  $\alpha_k = 1$  for all  $k$ . The question that naturally arises is what do the iterates of the fixed-point iteration method correspond to if we allow  $\alpha_k \neq 1$ . In this spirit, we propose to use  $\alpha I$  with  $\alpha > 0$  to approximate the elements in Clarke's generalized Jacobian.

The following lemma characterizes the iterates of this method, which we call  $\alpha$ -Value Iteration ( $\alpha$ -VI).

*Lemma 3.7:* Consider the semismooth Newton-type iteration for the Bellman residual function with  $B_k = \alpha I$  and  $\alpha > 0$ . Then  $\theta_{k+1} = \frac{\alpha-1}{\alpha}\theta_k + \frac{1}{\alpha}T\theta_k$ .

*Proof:* See the extended version [7].

Starting from Lemma 3.7, we can define the operator  $T_\alpha = \frac{\alpha-1}{\alpha}I + \frac{1}{\alpha}T$ , where  $I$  is the identity map and  $T$  is the Bellman operator. Notice that when  $\alpha = 1$  we recover the Bellman operator and therefore 1-VI is simply VI.

In the following, we are interested in studying the global and local convergence of  $\alpha$ -VI. We start by studying the contractivity of the  $T_\alpha$  operator and its fixed-points.

*Proposition 3.8:* For any  $\theta, \bar{\theta} \in \mathbb{R}^n$  and  $\alpha > \frac{1+\gamma}{2}$ ,

$$\|T_\alpha\theta - T_\alpha\bar{\theta}\|_\infty \leq \beta\|\theta - \bar{\theta}\|_\infty,$$

where  $\beta = \frac{|\alpha-1|}{\alpha} + \frac{\gamma}{\alpha} < 1$ . In addition, the optimal cost  $\theta^*$  is the unique fixed-point of  $T_\alpha$ .

*Proof:* We start by showing that, if  $\alpha > \frac{1+\gamma}{2}$ , the operator is  $\beta$ -contractive with respect to the infinity norm. For any  $\theta, \bar{\theta} \in \mathbb{R}^n$

$$\begin{aligned} \|T_\alpha\theta - T_\alpha\bar{\theta}\|_\infty &= \max_{s \in \mathcal{S}} \left| \frac{\alpha-1}{\alpha}(\theta_s - \bar{\theta}_s) + \frac{1}{\alpha}(T(\theta - \bar{\theta}))(s) \right| \\ &\stackrel{(a)}{\leq} \left| \frac{\alpha-1}{\alpha} \right| \max_{s \in \mathcal{S}} |\theta_s - \bar{\theta}_s| + \frac{1}{|\alpha|} \max_{s \in \mathcal{S}} |(T(\theta - \bar{\theta}))(s)| \\ &\stackrel{(b)}{\leq} \left( \left| \frac{\alpha-1}{\alpha} \right| + \frac{\gamma}{|\alpha|} \right) \max_{s \in \mathcal{S}} |\theta_s - \bar{\theta}_s| \\ &= \left( \left| \frac{\alpha-1}{\alpha} \right| + \frac{\gamma}{|\alpha|} \right) \|\theta - \bar{\theta}\|_\infty, \end{aligned}$$

where (a) follows from the triangle inequality and (b) from the fact that the Bellman operator is  $\gamma$ -contractive in the infinity norm. In order for  $T_\alpha$  to be contractive, we need  $\left( \left| \frac{\alpha-1}{\alpha} \right| + \frac{\gamma}{|\alpha|} \right) < 1$ . For  $\alpha \geq 1$ , since  $\gamma \in (0, 1)$ ,  $T_\alpha$  is contractive with rate  $(\alpha-1)/\alpha + \gamma/\alpha$ . For  $\alpha \in (0, 1)$ ,  $\left| \frac{\alpha-1}{\alpha} \right| + \frac{\gamma}{|\alpha|} = \frac{1-\alpha}{\alpha} + \frac{\gamma}{\alpha}$  and  $\frac{1-\alpha}{\alpha} + \frac{\gamma}{\alpha} < 1$  if and only if  $\alpha > \frac{1+\gamma}{2}$ . For  $\alpha < 0$ ,  $\left| \frac{\alpha-1}{\alpha} \right| + \frac{\gamma}{|\alpha|} = \frac{\alpha-1}{\alpha} - \frac{\gamma}{\alpha}$  and the inequality  $\frac{\alpha-1}{\alpha} - \frac{\gamma}{\alpha} < 1$  is never satisfied since  $\gamma \in (0, 1)$ . We can therefore conclude that if  $\alpha > \frac{1+\gamma}{2}$  then  $T_\alpha$  is  $\beta$ -contractive in the infinity norm with  $\beta = |\alpha-1|/\alpha + \gamma/\alpha$ . To verify that  $\theta^*$  is a fixed-point of  $T_\alpha$ , we exploit the definition of  $T_\alpha$  and the fact that  $\theta^*$  is the unique fixed-point of  $T$ . In particular,  $T_\alpha\theta^* = \frac{\alpha-1}{\alpha}\theta^* + \frac{1}{\alpha}T\theta^* = \frac{\alpha-1}{\alpha}\theta^* + \frac{1}{\alpha}\theta^* = \theta^*$ . Uniqueness follows directly from the Banach Theorem [15]. ■

The main implication of Proposition 3.8 is that, if  $\alpha > (1+\gamma)/2$ , then  $\alpha$ -VI converges globally to the optimal cost  $\theta^*$  with linear rate  $\beta$ . Results similar to Proposition 3.8 can be derived for the local contraction rate by considering Theorem 2.6 and evaluating the kappa condition with the infinity norm. Unfortunately, using this type of analysis it is not possible to conclude that  $\alpha$ -VI improves over VI in terms of convergence rate. Instead, we introduce the following proposition, which analyses the asymptotic rate of convergence of  $\alpha$ -VI via local stability analysis of nonlinear systems. For the sake of simplicity and interpretability, we consider a simplified setting in which the transition probability matrix at the solution has only real and positive eigenvalues. Notice that similar considerations can be made in a more general setting. This approach provides a tighter bound on the local rate of convergence, but is only applicable

in a neighborhood of the root where the Bellman residual function is continuously differentiable.

*Proposition 3.9:* Assume that  $r(\theta^*)$  is continuously differentiable in a neighborhood of  $\theta^*$  and that  $P\pi^*$  has only real and positive eigenvalues. Let  $\alpha \in (1/(1+\gamma), 1)$  and  $\tilde{\beta} = 1 - \frac{1-\gamma}{\alpha}$  if  $\alpha \in [1-\gamma/2, 1)$  and  $\tilde{\beta} = \frac{1}{\alpha} - 1$  if  $\alpha \in (1/(1+\gamma), 1-\gamma/2)$ .  $\alpha$ -VI converges linearly to  $\theta^*$  with asymptotic contraction rate  $\tilde{\beta} < \gamma$ .

*Proof:* We start by linearizing  $\theta_{k+1} = T_\alpha \theta_k$  at  $\theta^*$  via the first-order Taylor expansion

$$\theta_{k+1} - \theta^* = T_\alpha \theta_k - \theta^* = T_\alpha \theta^* + (T_\alpha \theta^*)' (\theta_k - \theta^*) + \mathcal{O}(\|\theta_k - \theta^*\|^2).$$

Since  $\theta^* = T_\alpha \theta^*$  and  $(T_\alpha \theta^*)' = \frac{(\alpha-1)}{\alpha} I + \frac{\gamma}{\alpha} P\pi^*$  for any optimal policy  $\pi^*$ , then

$$\theta_{k+1} - \theta^* = \left( I - \frac{1}{\alpha} (I - \gamma P\pi^*) \right) (\theta_k - \theta^*) + \mathcal{O}(\|\theta_k - \theta^*\|^2).$$

Therefore the asymptotic convergence rate is determined by the spectral radius of  $I - \frac{1}{\alpha} (I - \gamma P\pi^*)$ . In particular, since  $\rho\left(I - \frac{1}{\alpha} (I - \gamma P\pi^*)\right) \leq \max\left\{\left|1 - \frac{1-\gamma}{\alpha}\right|, \left|1 - \frac{1}{\alpha}\right|\right\}$ , we study different cases based on the values of  $\alpha$ . When  $\alpha \geq 1 - \gamma/2$ , then  $\max\left\{\left|1 - \frac{1-\gamma}{\alpha}\right|, \left|1 - \frac{1}{\alpha}\right|\right\} = 1 - \frac{1-\gamma}{\alpha}$ . In this case we get a contraction for any  $\alpha \geq 1 - \gamma/2$  since the inequality  $1 - \frac{1-\gamma}{\alpha} < 1$  is verified for any  $\alpha > 0$ . In addition, if  $\alpha \in [1 - \gamma/2, 1]$ , then we improve over the rate of VI since  $1 - \frac{1-\gamma}{\alpha} \leq \gamma$ . For  $\alpha < 1 - \gamma/2$ ,  $\max\left\{\left|1 - \frac{1-\gamma}{\alpha}\right|, \left|1 - \frac{1}{\alpha}\right|\right\} = \frac{1}{\alpha} - 1$  and we get a contraction if  $\alpha \in (1/2, 1 - \gamma/2)$ . In addition, if  $\alpha \in [1/(1+\gamma), 1 - \gamma/2)$ , then  $\frac{1}{\alpha} - 1 \leq \gamma$  and therefore we improve over the rate of VI. ■

By combining the results of Propositions 3.8 and 3.9 we obtain that, by setting  $\max\left\{\frac{1}{1+\gamma}, \frac{1+\gamma}{2}\right\} < \alpha < 1$ ,  $\alpha$ -VI converges globally with a linear rate and its asymptotic linear rate of convergence is strictly better than that of VI. The numerical experiments in Figure 1 corroborate our theoretical findings. Since our analysis is not tight, in practice we obtain convergence for a wider range of  $\alpha$ . We refer to the extended version [7] for additional numerical results.

#### IV. CONCLUSIONS & FUTURE WORK

We have shown that PI and VI are semismooth Newton-type methods. In particular, Propositions 3.5 and 3.6 reveal that PI and VI sit at the two opposite ends of the spectrum of semismooth Newton-type methods: PI enjoys local quadratic contraction but its costs per iteration are demanding; instead, VI is based on a coarse approximation of the elements in Clarke’s generalized Jacobian which allows to drastically reduce the costs per iteration at the price of downgrading the local quadratic convergence to a linear one. In the spirit of semismooth Newton-type methods, we proposed an extension of VI with negligible additional computational costs, global convergence guarantees and asymptotically faster contraction rate than VI. Finally, another promising future direction consists in formalizing and exploiting the connection between inexact semismooth Newton methods and optimistic policy iteration-type algorithms.

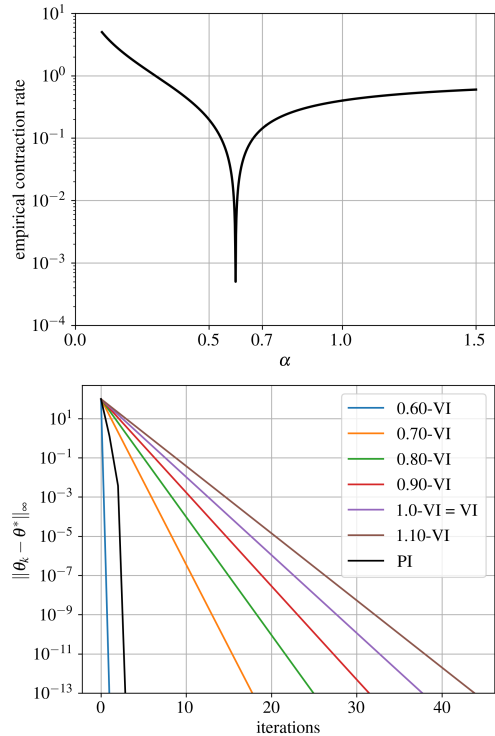


Fig. 1: Empirical global contraction rate of  $\alpha$ -VI for different values of  $\alpha$  and comparison of  $\alpha$ -VI and PI for a randomly generated MDP with 500 states, 10 actions and  $\gamma = 0.4$ . The maximum acceleration is dramatic and is obtained for  $\alpha \approx 0.6$ . Code at <https://gitlab.ethz.ch/gmatilde/alphaVI>.

#### REFERENCES

- [1] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2012.
- [2] D. P. Bertsekas. *Lessons from AlphaZero for Optimal, Model Predictive, and Adaptive Control*. Athena Scientific, 2022.
- [3] F. H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.
- [4] M. Diehl. Lecture notes on numerical optimization. Leuven-Freiburg 2007-2015 (last update: 02.02.2016).
- [5] F. Facchinei and J. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*, volume 2. Springer, 2003.
- [6] F. Facchinei and J. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*, volume 1. Springer, 2003.
- [7] Gargiani et al. Dynamic programming through the lens of semismooth Newton-type methods (extended version). arXiv, 2022.
- [8] M. Gargiani et al. Parallel and flexible dynamic programming via the randomized mini-batch operator. arXiv:2110.02901, 2021.
- [9] A. Izmailov and M. Solodov. *Newton-Type Methods for Optimization and Variational Problems*. Springer, 2014.
- [10] K. A. Khan and P. I. Barton. Evaluating an element of the Clarke generalized jacobian of a composite piecewise differentiable function. *ACM Trans. Math. Softw.*, 39(4), 2013.
- [11] B. Kummer. Newton’s method for non-differentiable functions. *Advances in Math. Optimization.*, 45:114–125, 12 1988.
- [12] M. Pollatschek and B. Avi-Itzhak. Algorithms for stochastic games with geometrical interpretation. *Management Science*, 15:399–413, 1969.
- [13] M. L. Puterman and S. L. Brumelle. On the convergence of policy iteration in stationary dynamic programming. *Mathematics of Operations Research*, 4(1):60–69, 1979.
- [14] H. Rademacher. Über partielle und totale Differenzierbarkeit von Funktionen mehrerer Variablen und über die Transformation der Doppelintegrale. *Mathematische Annalen*, 79(4):340–359, 1919.
- [15] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *Mathematics of Operations Research*, 14(5):877–898, 1996.
- [16] M. S. Santos and J. Rust. Convergence properties of policy iteration. *SIAM J. on Control and Optimization*, 42:2094–2115, 2004.
- [17] D. Silver et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016.