Bellman Residual Minimization for Control: Geometry, Stationarity, and Convergence

Donghwan Lee and Hyukjun Yang
School of Electrical Engineering
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea 34141
donghwan@kaist.ac.kr

Abstract

Markov decision problems are most commonly solved via dynamic programming. Another approach is Bellman residual minimization, which directly minimizes the squared Bellman residual objective function. However, compared to dynamic programming, this approach has received relatively less attention, mainly because it is often less efficient in practice and can be more difficult to extend to model-free settings such as reinforcement learning. Nonetheless, Bellman residual minimization provides an optimization objective that remains well defined under function approximation even when a projected Bellman operator is not a contraction. The control setting introduces additional challenges because maximization over actions can make the residual objective nonconvex and nonsmooth. In this paper, we establish foundational results for control Bellman residual minimization for policy optimization.

1 Introduction

A standard approach to solving Markov decision problems [23] is dynamic programming [7, 3, 4]. Another approach is Bellman residual minimization [2, 15, 27], which directly minimizes the squared Bellman residual objective. However, compared to dynamic programming, this approach has received relatively less attention, mainly because it is often less efficient in practice and can be more difficult to extend to model-free settings such as reinforcement learning [31]. Nevertheless, Bellman residual minimization has several advantages that make it worth investigating [15, 27]. For example, Bellman-residual objectives remain well defined under function approximation even when properties such as contraction of a projected Bellman operator [7, 3, 4] are unavailable; the quality of a stationary or minimizing solution must, however, be assessed separately. In contrast, dynamic-programming-based methods may lose the contraction property under function approximation, in which case convergence is hard to guarantee without additional conditions [19]. Building on prior work on policy evaluation [27] and policy optimization [15, 26], we establish results on the geometry, stationarity, and convergence of control Bellman residual (CBR) minimization and its soft counterpart.

In particular, we first consider the following control Bellman residual (CBR) objective:

$$f(\theta) = \frac{1}{2} \left\| TQ_\theta - Q_\theta \right\|_2^2,$$

where T denotes the standard control Bellman operator [7], Q_θ is a linear function approximator of the Q-function, and θ is its parameter vector. In policy evaluation, the Bellman residual objective is a strongly convex quadratic function of the parameter vector, which makes the analysis relatively straightforward. In contrast, the max operator generally makes the CBR objective nonconvex and can make it nonsmooth, although degenerate instances may be convex or smooth. As a result, the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

analysis becomes more challenging, and the optimization problem is also more difficult to solve. In this paper, we study various mathematical properties of the CBR objective. For instance, we show that although the objective is generally nonconvex and may be nonsmooth, it is locally Lipschitz and admits a piecewise quadratic structure. Building on these properties, we derive an explicit characterization of the Clarke subdifferential [11, 1], and use it to investigate the applicability and convergence of generalized gradient descent methods [10]. Next, we consider the soft control Bellman residual (SCBR) defined using the softmax and soft Bellman operator [16, 13, 17] as follows:

$$f(\theta) = \frac{1}{2} \left\| F_\lambda Q_\theta - Q_\theta \right\|_2^2 ,$$

where F_λ denotes the soft control Bellman operator [16, 13, 17] and $\lambda > 0$ is the temperature parameter. Compared to the CBR case, the SCBR objective is differentiable, and thus standard gradient descent [22] can be applied. As in the CBR case, we establish several properties of SCBR. In the tabular setting, these results yield geometric convergence of the objective and of the iterates to the soft Bellman fixed point; with linear function approximation, they yield vanishing gradient norms and stationary accumulation points under the stated step-size assumptions. Based on these findings, this paper investigates gradient-based algorithms for minimizing the CBR/SCBR objectives.

2 Related work

Baird [2] first proposed Bellman residual gradient methods, which perform gradient descent on the mean squared Bellman residual to guarantee convergence (albeit potentially slowly), and supported the approach with both analysis and simulations. Schoknecht [28] showed that several classic policy-evaluation algorithms with linear function approximation (e.g., TD(0) [30], LSTD [8], and residual-gradient methods [2]) each converge to the minimizer of a particular quadratic objective. The resulting solutions can be interpreted as different projections of the true value function under the corresponding projection operators. The original gradient temporal-difference (GTD) algorithm minimizes the squared norm of the expected TD update (NEU) [32]. The subsequent GTD2 and TDC algorithms target the mean-squared projected Bellman error (MSPBE), using an auxiliary parameter to estimate the required gradient terms [33]. These methods provide convergent off-policy policy evaluation with linear function approximation under the assumptions stated in those works. Scherrer [27] studies policy evaluation with linear function approximation, compares the projected TD(0) fixed point (satisfying the projected Bellman equation) with mean-squared Bellman residual minimization, and gives small MDP examples where either can perform better. Their key contribution is a unified oblique-projection framework that views both solutions as oblique projected Bellman equations, which leads to simplified characterizations and tight, interpretable error bounds. Geist et al. [15] compare Bellman-residual minimization with direct value maximization for policy search, derive proxy/performance bounds, and empirically show that the Bellman residual can be a poor proxy unless the sampling distribution matches the optimal occupancy measure. Although their focus is policy optimization, the Bellman-residual objective they study differs from the one considered in this paper. Fujimoto et al. [14] show that the Bellman error can be minimized without improving value accuracy, due to error cancellation and non-uniqueness in finite or incomplete data, making it a poor proxy for value error. Sun and Bagnell [29] study squared Bellman-residual minimization for policy evaluation as an online learning problem and show that (under an online-stability condition) no-regret updates yield small long-term prediction error. They also generalize residual-gradient methods as a special case and propose a broader family of online residual algorithms with theoretical guarantees and empirical validation. Maillard et al. [20] study policy evaluation via Bellman residual minimization with linear function approximation (using a generative model and double-sampling) and provide finite-sample generalization bounds showing the true Bellman residual of the empirical minimizer is controlled by the empirical residual plus an $O(1/\sqrt{n})$ term. Saleh and Jiang [26] propose deterministic Bellman residual minimization for policy optimization in deterministic environments, where the double-sampling issue [2] in squared Bellman-residual minimization disappears because two conditionally independent joint transition samples coincide. They provide theoretical and empirical results, and discuss limitations when stochasticity or distribution mismatch is present. Building on these studies, we analyze the piecewise quadratic geometry and Clarke-stationary points of the control Bellman residual, its oblique-projection characterizations, and the convergence properties of both CBR and SCBR under explicitly stated assumptions.

3 Preliminaries

3.1 Markov decision problem

We consider infinite-horizon discounted Markov decision problems [23], where the agent sequentially takes actions to maximize cumulative discounted rewards. In a Markov decision process with the state space $\mathcal{S} := \{1, 2, \ldots, |\mathcal{S}|\}$ and action space $\mathcal{A} := \{1, 2, \ldots, |\mathcal{A}|\}$, where $|\mathcal{S}|$ and $|\mathcal{A}|$ denote cardinalities of the respective sets, the decision maker selects an action $a \in \mathcal{A}$ at the current state $s \in \mathcal{S}$, then the state transitions to the next state $s' \in \mathcal{S}$ with probability $P(s'|s, a)$, and the transition incurs a reward $r(s, a, s') \in \mathbb{R}$, where $P(s'|s, a)$ is the state transition probability from the current state $s \in \mathcal{S}$ to the next state $s' \in \mathcal{S}$ under action $a \in \mathcal{A}$, and $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition reward function. We use $R(s, a) := \mathbb{E}[r(s, a, s')|s, a]$ for the expected one-step reward; when the reward is deterministic we simply write $r(s_k, a_k, s_{k+1}) =: r_{k+1}$, where $k \in \{0, 1, \ldots\}$ is the time step. We write $R_{\max} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |R(s, a)|$. Throughout the paper, we adopt the convention $\min \varnothing := +\infty$; in particular, the action gap below is $+\infty$ when there is only one action. A deterministic policy, $\pi : \mathcal{S} \to \mathcal{A}$, maps a state $s \in \mathcal{S}$ to an action $\pi(s) \in \mathcal{A}$. For any such policy, its Q-function is $Q^\pi(s, a) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} \middle| s_0 = s, a_0 = a, \pi\right]$. The optimal Q-function is defined statewise by $Q^*(s, a) := \max_{\pi\in\Theta} Q^\pi(s, a)$, where Θ is the set of stationary deterministic policies. For every finite discounted MDP, there exists a stationary deterministic policy π^* satisfying $Q^{\pi^*} = Q^*$ and $\pi^*(s) \in \arg\max_{a\in\mathcal{A}} Q^*(s, a)$ for every $s \in \mathcal{S}$. Thus optimality is statewise and does not depend on an unspecified initial-state distribution. Throughout, the MDP has finite state and action spaces and a discount factor $\gamma \in (0, 1)$; no ergodicity assumption is imposed. Additional assumptions on the feature matrix, oblique projections, and optimization algorithms are stated separately where required.

3.2 Definitions and notation

Throughout the paper, we will use the following notation for compact matrix-vector representations:

$$P := \begin{bmatrix} P_1 \\ \vdots \\ P_{|\mathcal{S}|} \end{bmatrix}, \; R := \begin{bmatrix} R(1, \cdot) \\ \vdots \\ R(|\mathcal{S}|, \cdot) \end{bmatrix}, \; Q := \begin{bmatrix} Q(1, \cdot) \\ \vdots \\ Q(|\mathcal{S}|, \cdot) \end{bmatrix},$$

where $P_s \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{S}|}$ is the matrix whose a-th row is $P(\cdot|s, a)^\top$, $Q(s, \cdot) \in \mathbb{R}^{|\mathcal{A}|}$, and $R(s, \cdot) \in \mathbb{R}^{|\mathcal{A}|}$. Note that $P \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|}$, $R \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, and $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$. In this notation, the Q-function is encoded as a single vector $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ using the *state-major* ordering, i.e., the block corresponding to a fixed state s contains all actions $a \in \mathcal{A}$. In particular, the single value $Q(s, a)$ can be written as

$$Q(s, a) = (e_s \otimes e_a)^\top Q,$$

where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ and $e_a \in \mathbb{R}^{|\mathcal{A}|}$ are the s-th basis vector (all components are 0 except for the s-th component which is 1) and the a-th basis vector, respectively. This convention is used throughout, including the SCBR Hessian formulas in Lemmas A.32 and A.41.

For any stochastic policy, $\pi : \mathcal{S} \to \Delta_{|\mathcal{A}|}$, where $\Delta_{|\mathcal{A}|}$ is the set of all probability distributions over $\mathcal{A}$, we define the corresponding action transition matrix [34] as

$$\Pi^\pi := \begin{bmatrix} e_1^\top \otimes \pi(1)^\top \\ e_2^\top \otimes \pi(2)^\top \\ \vdots \\ e_{|\mathcal{S}|}^\top \otimes \pi(|\mathcal{S}|)^\top \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}\times\mathcal{A}|}, \tag{1}$$

where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ is the basis vector defined above. Then, it is well known that $P\Pi^\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$ is the transition probability matrix of the state-action pair under policy π [34]. If we consider a deterministic policy, $\pi : \mathcal{S} \to \mathcal{A}$, the stochastic policy can be replaced by the corresponding one-hot encoding vector $\vec{\pi}(s) := e_{\pi(s)} \in \Delta_{|\mathcal{A}|}$, where $e_a \in \mathbb{R}^{|\mathcal{A}|}$, and the corresponding action transition matrix is identical to (1) with π replaced by $\vec{\pi}$. Given preselected state-action feature functions $\phi_1, \ldots, \phi_m : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the matrix, $\Phi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times m}$, called the feature matrix, is defined as a matrix whose (s, a)-th row vector is $\phi(s, a) := [\phi_1(s, a) \quad \cdots \quad \phi_m(s, a)]$. Throughout the paper, we assume that $\Phi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times m}$ has full column rank. All proofs in this paper are provided in the Appendix.

4 Control Bellman residual (CBR)

In this paper, we first consider the control Bellman residual (CBR) objective [2] with linear function approximation (LFA), $Q_\theta = \Phi\theta$, and the corresponding CBR objective function

$$f(\theta) = \frac{1}{2} \|T(\Phi\theta) - \Phi\theta\|_2^2 ,$$

where $\theta \in \mathbb{R}^m$ is the parameter vector and T is the control Bellman operator [7] defined as

$$(TQ_\theta)(s,a) := R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} Q_\theta(s',a').$$

Finding a solution that minimizes the above objective function constitutes the most direct and intuitive approach to solving the control Bellman equation (CBE), $Q_\theta = TQ_\theta$. However, since the above f is generally nonconvex and may be nonsmooth (due to the max operator), finding its minimizer is generally a challenging problem. Therefore, in this paper we examine the properties of f in greater depth. First, we can see that f is a piecewise quadratic function (Proposition A.2).

Proposition 1. *f is piecewise quadratic (so piecewise smooth) and continuous with the finite polyhedral cover S_π of $\mathbb{R}^m$ defined as*

$$S_\pi := \{\theta \in \mathbb{R}^m : \pi(s) \in \arg\max_{a \in \mathcal{A}} Q_\theta(s,a), \forall s \in \mathcal{S}\}$$

for each $\pi \in \Theta$, where Θ is the set of all deterministic policies, and $\arg\max$ *is interpreted as a set-valued map, i.e., $\pi(s)$ is one of the maximizers, and ties are allowed.*

In other words, S_π is the set of all $\theta \in \mathbb{R}^m$ such that $\pi(s) \in \arg\max_{a \in \mathcal{A}} Q_\theta(s,a)$. In general, the interiors of the sets S_π need not be disjoint under linear function approximation because persistent ties can occur. Within each region, f reduces to a single quadratic function, and the cover exhibits an interesting geometric property (Proposition A.3) summarized below.

Proposition 2. *For each $\pi \in \Theta$, the set S_π is an intersection of half-spaces and is a convex cone.*

Another geometric property that helps characterize optimality is that f can be upper- and lower-bounded by two quadratic functions, respectively (Proposition A.4).

Proposition 3. *f is bounded by strongly convex quadratic functions as $q_1(\theta) \le f(\theta) \le q_2(\theta)$, where $q_1(\theta) := \frac{(1-\gamma)^2}{2|\mathcal{S}\times\mathcal{A}|} \|Q_\theta - Q^*\|_2^2$ and $q_2(\theta) := \frac{(1+\gamma)^2|\mathcal{S}\times\mathcal{A}|}{2} \|Q_\theta - Q^*\|_2^2$.*

In this case, the two quadratic functions q_1 and q_2 share the same minimizer set, $\arg\min_{\theta \in \mathbb{R}^m} \|Q_\theta - Q^*\|_2$, but their minimum values are generally different. These properties help us understand the geometric structure of f and will be useful later for deriving several results related to optimality. Next, we may attempt to solve the above optimization problem by using the (Clarke) subdifferential [11, 1], which is a standard tool for optimizing nonconvex and nonsmooth functions.

4.1 Subdifferential

To minimize f, conventional methods such as gradient descent [22] cannot be directly applied at all points, because f is generally nonsmooth and may fail to be differentiable on switching sets. Instead, we can leverage the (Clarke) subdifferential [11, 1], which extends the notion of gradients to locally Lipschitz nonsmooth functions. To this end, we first establish the subdifferential of the CBR objective f (Theorem A.5) as follows. In the LFA setting, not every greedy policy at Q_θ needs to be reachable from differentiability regions in the parameter space. Therefore, we use the active set

$$\Lambda_\Phi(\theta) := \left\{ \pi \in \Theta : \theta \in \overline{\operatorname{int}(S_\pi)} \right\},$$

where the interior and closure are taken in $\mathbb{R}^m$.

Theorem 1. *The subdifferential of f is given by*

$$\partial f(\theta) = \{\Phi^\top (\gamma P \Pi^\beta - I)^\top (TQ_\theta - Q_\theta) : \beta \in \operatorname{conv}(\Lambda_\Phi(\theta))\},$$

where conv *denotes the convex hull.*

We can observe that the subdifferential of f admits a simple and intuitive form. We can also derive several useful properties of f. For instance, f is locally Lipschitz, and its subdifferential is nonempty,

convex, and compact (Lemma A.27). Using the subdifferential derived above, one can see that the stationary point $\bar{\theta}$, i.e. $0 \in \partial f(\bar{\theta})$, satisfies

$$\Phi^\top(\gamma P\Pi^{\bar{\beta}} - I)^\top(TQ_{\bar{\theta}} - Q_{\bar{\theta}}) = 0 \tag{2}$$

for some $\bar{\beta} \in \mathrm{conv}(\Lambda_\Phi(\bar{\theta}))$. Specifically, $\bar{\beta}$ satisfies

$$\bar{\beta} \in \underset{\beta \in \mathrm{conv}(\Lambda_\Phi(\bar{\theta}))}{\arg\min} \left\| \Phi^\top(\gamma P\Pi^{\beta} - I)^\top(TQ_{\bar{\theta}} - Q_{\bar{\theta}}) \right\|_2. \tag{3}$$

We can easily see that (2) can be equivalently written as

$$\mathcal{R}((\gamma P\Pi^{\bar{\beta}} - I)\Phi) \perp (TQ_{\bar{\theta}} - Q_{\bar{\theta}}), \tag{4}$$

where $\mathcal{R}$ denotes the range space. Equivalently,

$$TQ_{\bar{\theta}} - Q_{\bar{\theta}} \in \mathcal{N}\left(\left((\gamma P\Pi^{\bar{\beta}} - I)\Phi\right)^\top\right) = \mathcal{R}((\gamma P\Pi^{\bar{\beta}} - I)\Phi)^\perp.$$

Thus the Bellman error lies in the orthogonal complement of the range of $(\gamma P\Pi^{\bar{\beta}} - I)\Phi$, not in the null space of that matrix itself. Following the ideas in Scherrer [27], (4) can be viewed as an oblique-projection condition, provided that the relevant oblique projector is well defined. The Bellman error $TQ_{\bar{\theta}} - Q_{\bar{\theta}}$ then lies in the orthogonal complement of $\mathcal{R}((\gamma P\Pi^{\bar{\beta}} - I)\Phi)$. For any $\Psi \in \mathbb{R}^{n\times m}$, define the set-valued projected-equation map

$$\mathcal{G}_{\Phi|\Psi}(y) := \left\{\Phi z : z \in \mathbb{R}^m,\ \Psi^\top(\Phi z - y) = 0\right\}. \tag{5}$$

Unlike the matrix formula for an oblique projector, this map does not require $\Psi^\top\Phi$ to be nonsingular. It may be empty for an arbitrary y, but the stationarity results below guarantee membership for the Bellman images occurring at stationary points. If $\Psi^\top\Phi$ is nonsingular, then $\mathcal{G}_{\Phi|\Psi}(y) = \{\Gamma_{\Phi|\Psi}(y)\}$.

Assumption 1 (Oblique-projection regularity)**.** *Whenever an oblique projection $\Gamma_{\Phi|\Psi}$ is used, we assume that $\Psi^\top\Phi$ is nonsingular. Whenever an oblique projector $\Gamma_{\Phi|\Omega}$ is used in an error bound, we likewise assume that $\Omega^\top\Phi$ is nonsingular. In particular, for CBR stationary points we assume $\Phi^\top(\gamma P\Pi^{\bar{\beta}} - I)^\top\Phi$ is nonsingular, where $\bar{\beta} \in \mathrm{conv}(\Lambda_\Phi(\bar{\theta}))$ satisfies* (2)*.*

Under Assumption 1, $\Psi_{\bar{\beta}}^\top\Phi = \Phi^\top(\gamma P\Pi^{\bar{\beta}} - I)^\top\Phi$ is invertible, where $\Psi_{\bar{\beta}} := (\gamma P\Pi^{\bar{\beta}} - I)\Phi$, and the oblique projector onto $\mathcal{R}(\Phi)$ along $\mathcal{N}(\Psi_{\bar{\beta}}^\top)$ is given by

$$\Gamma_{\Phi|\Psi_{\bar{\beta}}}(x) = \Phi(\Psi_{\bar{\beta}}^\top\Phi)^{-1}\Psi_{\bar{\beta}}^\top x,$$

where $\mathcal{N}$ denotes the null space and this operator is called the oblique projection [27]. Then, (4) can be written as the following compact form:

$$Q_{\bar{\theta}} = \Gamma_{\Phi|\Psi_{\bar{\beta}}} TQ_{\bar{\theta}}, \tag{6}$$

which is called the oblique projected control Bellman equation (OP-CBE). This oblique projection generalizes the standard orthogonal projection: when $\Psi_{\bar{\beta}}$ is replaced by Φ, it reduces to the usual orthogonal projection, $\Gamma_{\Phi|\Phi}$, onto the range space of Φ. Scherrer [27] developed a unified oblique-projection view of policy evaluation with linear function approximation. In this paper, we extend this policy-evaluation framework to policy optimization. The following theorem formalizes this fact.

Theorem 2. *A point $\bar{\theta}$ is Clarke-stationary if and only if there exists $\bar{\beta} \in \mathrm{conv}(\Lambda_\Phi(\bar{\theta}))$ such that*

$$Q_{\bar{\theta}} \in \mathcal{G}_{\Phi|\Psi_{\bar{\beta}}}(TQ_{\bar{\theta}}), \qquad \Psi_{\bar{\beta}} = (\gamma P\Pi^{\bar{\beta}} - I)\Phi.$$

If Assumption 1 holds for this $\bar{\beta}$, the set on the right is a singleton and the condition is equivalently

$$Q_{\bar{\theta}} = \Gamma_{\Phi|\Psi_{\bar{\beta}}} TQ_{\bar{\theta}}.$$

Since a stationary point always exists, the set-valued OP-CBE always admits a solution. Whenever Assumption 1 holds at a stationary pair, that solution reduces to the single-valued OP-CBE above (Lemma A.28). This conclusion does not require the composite operator $\Gamma_{\Phi|\Psi_\beta}T$ to be a contraction.

4.2 Generalized gradient descent method

To minimize f, which is generally nonconvex and may be nonsmooth, we can apply the following generalized gradient descent method [10]:

$$\theta_{k+1} = \theta_k - \alpha_k g_k, \tag{7}$$

where $g_k \in \arg\min_{g \in \partial f(\theta_k)} \|g\|_2$ and $\alpha_k > 0$ is a step size generated by backtracking search with the Armijo rule [9]. A key issue in nonsmooth nonconvex optimization is that an arbitrary choice $g_k \in \partial f(\theta_k)$ does not necessarily yield a descent direction. A classical remedy is to select the minimum-norm subgradient [10], $g_k \in \arg\min_{g \in \partial f(\theta_k)} \|g\|_2 = \mathrm{Proj}_{\partial f(\theta_k)}(0)$, where Proj denotes the projection. Note that by construction, $\|g_k\|_2 = \mathrm{dist}\big(0, \partial f(\theta_k)\big)$, where dist denotes the Euclidean distance between a point and a convex set. The minimum-norm element can be computed by the explicit convex quadratic program in Proposition A.6. Under the active-set nondegeneracy condition stated below, the Armijo method has stationary limit points (Theorem A.8).

Theorem 3. *Let $(\theta_k)_{k\geq 0}$ be generated by $\theta_{k+1} = \theta_k - \alpha_k g_k$, where $g_k \in \arg\min_{g \in \partial f(\theta_k)} \|g\|_2$ and $\alpha_k > 0$ is generated by Armijo backtracking. Assume that every nonstationary limit point $\bar{\theta}$ has a unique greedy action at every state, and hence f agrees with one quadratic policy branch on a neighborhood of $\bar{\theta}$. Then $(\theta_k)_{k\geq 0}$ admits at least one limit point, and every limit point $\bar{\theta}$ satisfies $0 \in \partial f(\bar{\theta})$.*

A stationary point always satisfies the set-valued OP-CBE in Theorem 2; under oblique regularity, this reduces to (6). With the full-column-rank feature matrix assumed here, orthogonally projected value iteration is well defined, but its convergence requires additional conditions [6]. Under the stated active-set nondegeneracy condition, the generalized-gradient method has Clarke-stationary limit points, and every such point satisfies the set-valued OP-CBE; oblique regularity reduces it to the single-valued equation.

4.3 Tabular case

In the tabular case, $\Phi = I, \theta = Q$, and $m = |\mathcal{S} \times \mathcal{A}|$, the corresponding CBR objective is reduced to $f(Q) := \frac{1}{2}\|TQ - Q\|_2^2$ and its subdifferential is given by

$$\partial f(Q) = \{(\gamma P \Pi^\beta - I)^\top (TQ - Q) : \beta \in \mathrm{conv}(\Lambda(Q))\},$$

where $\Lambda(Q) := \{\pi \in \Theta : \pi(s) \in \arg\max_{a \in \mathcal{A}} Q(s,a),\ \forall s \in \mathcal{S}\}$ is the set of all possible greedy policies for Q. Based on the above result, we can show that, in the tabular case, there exists a unique stationary solution Q satisfying $0 \in \partial f(Q)$, and this solution coincides with the optimal Q^* (Theorem A.3).

Theorem 4. *The stationary point $\bar{Q}$ with $0 \in \partial f(\bar{Q})$ in the tabular case is unique and is given by $\bar{Q} = Q^*$.*

This function f also satisfies the properties in Proposition 3 for the LFA case. In this case, the two quadratic functions q_1 and q_2 not only share the same minimizer Q^* but also attain the same minimum value. This implies that f likewise has a unique minimizer, which is the same one. From the above result, we can observe that in the tabular case, although the CBR objective need not be convex, its stationary point is unique and coincides with the solution, Q^*, of the optimal control Bellman equation, $Q = TQ$, which is a favorable structure. Therefore, obtaining the stationary point amounts to solving the optimal control Bellman equation. Writing $n := |\mathcal{S} \times \mathcal{A}|$, the tabular CBR also satisfies the explicit global nonsmooth PL inequality

$$f(Q) \leq \frac{n}{2(1-\gamma)^2}\,\mathrm{dist}(0, \partial f(Q))^2, \qquad Q \in \mathbb{R}^n.$$

To minimize f, consider the generalized gradient update $Q_{k+1} = Q_k - \alpha_k g_k$. Assume that every nonstationary limit point has a unique greedy action at every state. Under this active-set nondegeneracy condition, every limit point is stationary and therefore equals the unique stationary point Q^*; consequently, the entire sequence converges to Q^*.

Theorem 5. *Let $(Q_k)_{k\geq 0}$ be generated by $Q_{k+1} = Q_k - \alpha_k g_k$, where $g_k \in \arg\min_{g \in \partial f(Q_k)} \|g\|_2$ and $\alpha_k > 0$ is generated by Armijo backtracking. If every nonstationary limit point of (Q_k) has a unique greedy action at every state, then $(Q_k)_{k\geq 0}$ converges to Q^*.*

Generalized gradient descent can be slower and less favorable in practice than value iteration. In the tabular setting, value iteration, $Q_{k+1} = TQ_k$, leverages the contraction property of the Bellman operator and enjoys geometric convergence, whereas Theorem 5 does not provide a comparable global rate for the generalized-gradient method. Nevertheless, this approach remains appealing and worthy of further investigation.

4.4 Error bounds

For a fixed policy and a full-column-rank feature matrix, the squared Bellman residual objective is strongly convex and has a unique stationary point; its oblique-projection interpretation is developed in Scherrer [27] and adapted to state-action values in Lemma A.9. In the control setting, there may exist multiple stationary points, and it is generally difficult to derive a relationship between the policy associated with each stationary point and the optimal policy. However, by leveraging the following general bound (Lemma A.29), we can infer, to some extent, the relationship between the solution obtained from a stationary point and the optimal policy as well as the optimal Q-function:

Proposition 4. *For any $\theta \in \mathbb{R}^m$, we have*

$$\|Q_\theta - Q^*\|_\infty \leq \frac{\sqrt{2}}{1-\gamma}\sqrt{f(\theta)}, \tag{8}$$

$$\|Q^{\pi_\theta} - Q^*\|_\infty \leq \frac{2\sqrt{2}\gamma}{1-\gamma}\sqrt{f(\theta)}, \tag{9}$$

where π_θ is any deterministic greedy policy satisfying $\pi_\theta(s) \in \arg\max_{a\in\mathcal{A}} Q_\theta(s,a)$ for every $s \in \mathcal{S}$.

That is, let us suppose we select a minimizer, θ, of f among the stationary points. Then, this amounts to finding a parameter θ that at least minimizes the upper bounds on the right-hand sides of (8) and (9). In particular, the second bound (9) provides an upper bound on the distance between the Q-function, Q^{π_θ}, induced by the greedy policy corresponding to our computed Q_θ and Q^*. By minimizing f, we can find the best θ in terms of the upper bounds given above. Moreover, the following result shows that if Q^* is sufficiently close to the function representation space $\mathcal{R}(\Phi)$, then Q_{θ^*} is close to Q^*, where θ^* is a minimizer of f (Proposition A.7).

Proposition 5. *Suppose that $\theta^* \in \arg\min_{\theta\in\mathbb{R}^m} f(\theta)$. Then, we have*

$$\|Q_{\theta^*} - Q^*\|_\infty \leq \frac{(1+\gamma)\sqrt{|\mathcal{S}\times\mathcal{A}|}}{1-\gamma}\left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2. \tag{10}$$

In (10), the term $\|\Gamma_{\Phi|\Phi}Q^* - Q^*\|_2$ represents the approximation error of Q^*, i.e., the distance between Q^* and its best possible function approximation $Q_\theta = \Gamma_{\Phi|\Phi}Q^*$, where $\Gamma_{\Phi|\Phi}$ is the orthogonal projection onto $\mathcal{R}(\Phi)$. Hereafter, we refer to this as the approximation error. Therefore, (10) implies that if the approximation error is arbitrarily small, then the error between Q^* and Q_{θ^*} can also be made arbitrarily small. Several other related error bounds are summarized below (Proposition A.5).

Proposition 6. *Let us define the minimizers*

$$\theta_1^* \in \arg\min_{\theta\in\mathbb{R}^m} \|Q_\theta - Q^*\|_2, \qquad \theta_2^* \in \arg\min_{\theta\in\mathbb{R}^m} f(\theta).$$

Let $e_\Phi := \left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2$ and $n := |\mathcal{S}\times\mathcal{A}|$. Then, we have

$$0 \leq f(\theta_1^*) - f(\theta_2^*) \leq \left\{\frac{(1+\gamma)^2 n}{2} - \frac{(1-\gamma)^2}{2n}\right\} e_\Phi^2,$$

$$\left\|\Gamma_{\Phi|\Phi}Q^* - Q_{\theta_2^*}\right\|_2 \leq \left\{1 + \frac{(1+\gamma)n}{1-\gamma}\right\} e_\Phi.$$

The exact-recovery result in Appendix Theorem A.7 shows that, under the stated approximation and action-gap conditions, the CBR objective is strongly convex on the convex parameter set $K = \{\theta : \|\Phi\theta - Q^*\|_\infty \leq \varepsilon\}$, and its unique global minimizer lies in K and induces the optimal greedy policy.

Corollary 1 (Exact policy recovery from approximation error)**.** *Let $n := |\mathcal{S} \times \mathcal{A}|$, let Q^* have a unique greedy policy with action gap $\Delta > 0$, and define $e_\Phi := \|\Gamma_{\Phi|\Phi} Q^* - Q^*\|_2$. If*

$$e_\Phi < \frac{(1-\gamma)\Delta}{2(1+\gamma)\sqrt{n}},$$

then the CBR objective has a unique global minimizer θ^* and every hard greedy policy of Q_{θ^*} equals π^*. Moreover,

$$\|Q_{\theta^*} - Q^*\|_\infty \leq \frac{(1+\gamma)\sqrt{n}}{1-\gamma} e_\Phi < \frac{\Delta}{2}.$$

In the next section, we focus on the soft control Bellman equation [16, 13]. Because the soft control Bellman residual yields a differentiable objective, one can apply the more standard gradient descent methods [22] for differentiable functions. Accordingly, the next section presents a theoretical analysis of solutions to the soft Bellman residual, and we then discuss in greater detail gradient descent algorithms for solving it.

5 Soft control Bellman residual (SCBR)

In this section, we consider the so-called soft control Bellman residual (SCBR) objective function

$$f(\theta) = \frac{1}{2} \|F_\lambda(\Phi\theta) - \Phi\theta\|_2^2,$$

where F_λ is the soft control Bellman operator [16, 13, 17] defined by

$$(F_\lambda Q_\theta)(s,a) := R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \lambda \ln \left(\sum_{u \in \mathcal{A}} \exp\left\{ \frac{Q_\theta(s',u)}{\lambda} \right\} \right).$$

Here, $\lambda > 0$ is the temperature parameter. In the SCBR objective above, the soft Bellman equation replaces the nonsmooth max operator by its log-sum-exp approximation. Thus, it produces a differentiable Bellman operator (actually, C^∞). This construction provides the theoretical basis for soft Q-learning and the soft actor-critic framework [16, 13, 17]. Introducing this soft Bellman operator resolves the differentiability issue of the CBR in the previous section. However, the SCBR objective is generally nonconvex, and hence, finding a global minimum is still not straightforward. Moreover, analogous quadratic upper and lower bounds hold for the SCBR objective after replacing the hard optimal solution Q^* by the soft fixed point Q^*_λ of F_λ.

5.1 Gradient

To further understand the properties of the SCBR, we first compute its gradient as follows (Theorem A.11).

Theorem 6. *The gradient of f is given by*

$$\nabla_\theta f(\theta) = \Phi^\top (\gamma P \Pi^{\pi_\theta} - I)^\top (F_\lambda(\Phi\theta) - \Phi\theta)$$

where $\pi_\theta(j|i) := \frac{\exp(Q_\theta(i,j)/\lambda)}{\sum_{u \in \mathcal{A}} \exp(Q_\theta(i,u)/\lambda)}$ is the Boltzmann policy of Q_θ.

This expression is similar to the result for CBR. One difference is that CBR uses the subdifferential, whereas here we can use the standard gradient for SCBR. From the gradient given above, we can see that the stationary point $\bar{\theta}$, i.e., $\nabla_\theta f(\bar{\theta}) = 0$, satisfies $\Phi^\top (\gamma P \Pi^{\pi_{\bar{\theta}}} - I)^\top (F_\lambda(\Phi\bar{\theta}) - \Phi\bar{\theta}) = 0$, which also implies $\mathcal{R}((\gamma P \Pi^{\pi_{\bar{\theta}}} - I)\Phi) \perp (F_\lambda(\Phi\bar{\theta}) - \Phi\bar{\theta})$, where $\mathcal{R}$ is the range space. This result is analogous to that for the CBR case, and is also closely related to the result in Scherrer [27] for policy evaluation. As before, stationarity admits a projected-equation characterization even when the oblique regularity matrix is singular.

Theorem 7. *For every $\bar{\theta}$,*

$$\nabla f(\bar{\theta}) = 0 \quad \Longleftrightarrow \quad Q_{\bar{\theta}} \in \mathcal{G}_{\Phi|\Psi_{\bar{\theta}}}(F_\lambda Q_{\bar{\theta}}), \qquad \Psi_\theta = (\gamma P \Pi^{\pi_\theta} - I)\Phi.$$

Under Assumption 1, the set is a singleton and this condition reduces to

$$Q_{\bar{\theta}} = \Gamma_{\Phi|\Psi_{\bar{\theta}}} F_\lambda Q_{\bar{\theta}}.$$

Since a stationary point always exists (Lemma A.43), the set-valued OP-SCBE always has a solution; oblique regularity makes it single-valued. Moreover, similarly to the CBR objective, f is locally

smooth (Lemma A.44). The above result plays an important role in establishing the convergence of the gradient descent method.

5.2 Gradient descent method

In the SCBR case, since the objective function is C^∞, we can apply standard gradient descent [22]. Using the local smoothness and coercivity of f, we show that the gradient norms converge to zero and that every accumulation point is stationary (Theorem A.13); convergence of the entire parameter sequence to a single point is not asserted in the general LFA setting.

Theorem 8. *Let us consider the gradient descent iterates, $\theta_{k+1} = \theta_k - \alpha\nabla_\theta f(\theta_k)$, for $k = 0, 1, \ldots$ with any initial point $\theta_0 \in \mathbb{R}^m$. Fix any $c > f(\theta_0)$ and set $\mathcal{L}_c := \{\theta : f(\theta) \le c\}$. For the chosen step size α, define the compact convex set*

$$D_\alpha := \operatorname{conv}\left(\mathcal{L}_c \cup \{\theta - \alpha\nabla_\theta f(\theta) : \theta \in \mathcal{L}_c\}\right),$$

and let L_α be a Lipschitz constant of $\nabla_\theta f$ on D_α. If $0 < \alpha < 2/L_\alpha$, then the iterates remain in the compact set $\mathcal{L}_c$, every accumulation point is stationary, $\lim_{k\to\infty} \|\nabla_\theta f(\theta_k)\|_2 = 0$, and

$$\min_{0\le i\le N} \|\nabla_\theta f(\theta_i)\|_2^2 \le \frac{f(\theta_0) - \min_{\theta\in\mathbb{R}^m} f(\theta)}{(N+1)\alpha\left(1 - \frac{\alpha L_\alpha}{2}\right)}.$$

Because D_α and hence L_α depend on the selected α, the displayed step-size requirement is an implicit sufficient condition rather than a closed-form a priori rule. The result gives asymptotic stationarity without the active-set condition needed by the nonsmooth CBR analysis. It does not, in the general LFA setting, imply convergence of the whole sequence or guarantee that the attained residual or policy error is small. Every accumulation point satisfies the set-valued OP-SCBE, and under oblique regularity this reduces to the single-valued OP-SCBE. Orthogonally projected soft value iteration is also well defined under the full-column-rank assumption, but its convergence requires additional conditions.

5.3 Tabular case

In the tabular case, $\Phi = I, \theta = Q$, and $m = |\mathcal{S}\times\mathcal{A}|$, the SCBR objective function is reduced to $f(Q) := \frac{1}{2}\|F_\lambda Q - Q\|_2^2$. In this case, the gradient of f is given by

$$\nabla_Q f(Q) = (\gamma P\Pi^{\pi_Q} - I)^\top (F_\lambda Q - Q),$$

where $\pi_Q(j|i) := \frac{\exp(Q(i,j)/\lambda)}{\sum_{u\in\mathcal{A}}\exp(Q(i,u)/\lambda)}$ is the Boltzmann policy of Q. As in the previous section for the CBR, the SCBR has a unique stationary point that coincides with the solution of the soft control Bellman equation (SCBE), Q_λ^* satisfying $Q_\lambda^* = F_\lambda Q_\lambda^*$.

Theorem 9. *The stationary point $\bar{Q}$ with $\nabla_Q f(\bar{Q}) = 0$ is the unique solution $\bar{Q} = Q_\lambda^*$ to the soft control Bellman equation $Q_\lambda^* = F_\lambda Q_\lambda^*$.*

This is a useful fact: in the tabular case, although the SCBR need not be convex, if we use gradient descent to locate a stationary point, that point will be the unique solution of the SCBE. Because the tabular setting is a special case of the LFA, the SCBR objective function is locally smooth. Moreover, the tabular setting admits several stronger properties than the linear function approximation setting. In particular, we can prove that f is locally strongly convex (Lemma A.37) in a neighborhood of the unique solution, Q_λ^*, to the soft Bellman equation, $Q_\lambda^* = F_\lambda Q_\lambda^*$. Furthermore, writing $n := |\mathcal{S}\times\mathcal{A}|$, the tabular SCBR satisfies the global explicit PL inequality

$$f(Q) - f(Q_\lambda^*) \le \ell_{\mathrm{PL}}\|\nabla_Q f(Q)\|_2^2, \qquad \ell_{\mathrm{PL}} := \frac{n}{2(1-\gamma)^2}, \qquad Q \in \mathbb{R}^n$$

(Lemma A.38). This yields a geometric rate once a step size satisfying the condition below is fixed.

Theorem 10. *Let $Q_{k+1} = Q_k - \alpha\nabla_Q f(Q_k)$, fix $c > f(Q_0)$, and set $\mathcal{L}_c := \{Q : f(Q) \le c\}$. Define*

$$D_\alpha := \operatorname{conv}\left(\mathcal{L}_c \cup \{Q - \alpha\nabla_Q f(Q) : Q \in \mathcal{L}_c\}\right),$$

and let L_α be a Lipschitz constant of $\nabla_Q f$ on D_α. If $0 < \alpha < 2/L_\alpha$ and $\ell_{\mathrm{PL}} > \alpha(1 - \alpha L_\alpha/2)$, then

$$f(Q_k) - f(Q_\lambda^*) \le \left(1 - \frac{\alpha(1-\alpha L_\alpha/2)}{\ell_{\mathrm{PL}}}\right)^k \left(f(Q_0) - f(Q_\lambda^*)\right).$$

Here too, because D_α and L_α depend on α, these are implicit sufficient conditions rather than a closed-form a priori step-size rule. Compared with the general CBR result, the tabular SCBR result gives a geometric decrease of the objective under a constant step size. Together with $\|Q - Q^*_\lambda\|_\infty \le \sqrt{2f(Q)}/(1-\gamma)$, it also yields geometric convergence of the iterates to the unique soft Bellman fixed point. This stronger conclusion is specific to the tabular setting and the stated step-size condition.

5.4 Error bounds

The CBR error bounds that compare Q_θ with the Bellman fixed point carry over to the SCBR setting after replacing Q^* by the soft fixed point Q^*_λ and T by F_λ. For example, $\|Q_\theta - Q^*_\lambda\|_\infty \le \frac{\sqrt{2}}{1-\gamma}\sqrt{f(\theta)}$. If $\theta^* \in \arg\min_{\theta\in\mathbb{R}^m} f(\theta)$, then $\|Q_{\theta^*} - Q^*_\lambda\|_\infty \le \frac{(1+\gamma)\sqrt{|\mathcal{S}\times\mathcal{A}|}}{1-\gamma}\|\Gamma_{\Phi|\Phi}Q^*_\lambda - Q^*_\lambda\|_2$.

Proposition 7 (SCBR guarantees for the original hard-control problem)**.** *Let $f_\lambda(\theta) := \frac{1}{2}\|F_\lambda Q_\theta - Q_\theta\|_2^2$, let Q^*_λ be the fixed point of F_λ, and let π_θ be any hard greedy policy of Q_θ. Then*

$$\|Q^*_\lambda - Q^*\|_\infty \le \frac{\gamma\lambda\log|\mathcal{A}|}{1-\gamma},$$

$$\|Q_\theta - Q^*\|_\infty \le \frac{\sqrt{2f_\lambda(\theta)} + \gamma\lambda\log|\mathcal{A}|}{1-\gamma},$$

$$\|Q^{\pi_\theta} - Q^*\|_\infty \le \frac{2\gamma}{1-\gamma}\left(\sqrt{2f_\lambda(\theta)} + \gamma\lambda\log|\mathcal{A}|\right).$$

If Q^* has a unique greedy policy with action gap $\Delta > 0$ and

$$\sqrt{2f_\lambda(\theta)} + \gamma\lambda\log|\mathcal{A}| < \frac{(1-\gamma)\Delta}{2},$$

then $\pi_\theta = \pi^*$. In particular, if $|\mathcal{A}| \ge 2$, every hard greedy policy of Q^*_λ is optimal whenever

$$\lambda < \frac{(1-\gamma)\Delta}{2\gamma\log|\mathcal{A}|}.$$

The corresponding local convexity result for SCBR is provided in Appendix Theorem A.14.

As illustrated in Example 1 of the Appendix, both the SCBR gradient method and the implemented projected value-iteration trajectory remained stable and reduced the hard-Q error overall in the reported instance. This observation is instance-specific and does not establish a general performance ordering.

6 Conclusion

This paper develops a convergence and structural analysis of control Bellman residual (CBR) minimization for policy optimization. Despite the generally nonconvex and potentially nonsmooth structure induced by the max operator, we show CBR is locally Lipschitz and piecewise quadratic, derive its Clarke subdifferential, and prove stationary-limit-point convergence of the Armijo method under the stated active-set nondegeneracy condition. We also analyze a differentiable soft variant (SCBR) and relate its stationarity to a projected/oblique Bellman equation perspective.

Acknowledgement

This work was partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25510069), and partially supported by the Institute of Information Communications Technology Planning Evaluation (IITP) funded by the Korea government under Grant 2022-0-00469.

Declaration of LLM usage

During manuscript preparation, large language models (LLMs) were used to assist with English-language editing, proofreading, LaTeX formatting, checking mathematical derivations and proof

consistency, identifying potential errors, and checking citation accuracy. This assistance supported manuscript preparation and review. LLMs are not components of the proposed CBR/SCBR algorithms or experimental procedures. The authors take responsibility for the accuracy and integrity of all mathematical statements, proofs, experimental results, and conclusions presented in this paper.

References

[1] Adil Bagirov, Napsu Karmitsa, and Marko M Mäkelä. *Introduction to Nonsmooth Optimization: theory, practice and software*. Springer, 2014.

[2] Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the twelfth international conference on machine learning*, pages 30–37, 1995.

[3] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*. Athena scientific, 2012.

[4] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume II*. Athena scientific, 2012.

[5] Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*, volume 1. Athena Scientific Belmont, 2003.

[6] Dimitri P Bertsekas. Temporal difference methods for general projected equations. *IEEE Transactions on Automatic Control*, 56(9):2128–2139, 2011.

[7] Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific Belmont, MA, 1996.

[8] Justin A Boyan. Least-squares temporal difference learning. In *Proceedings of the 16th annual international conference on machine learning*, pages 49–56, 1999.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[10] James V Burke, Adrian S Lewis, and Michael L Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

[11] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[12] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International conference on machine learning*, pages 1125–1134, 2018.

[13] Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Uncertainty in Artificial Intelligence*, pages 202–211, 2016.

[14] Scott Fujimoto, David Meger, Doina Precup, Ofir Nachum, and Shixiang Shane Gu. Why should i trust you, bellman? the bellman error is a poor replacement for value error. In *International Conference on Machine Learning*, pages 6918–6943, 2022.

[15] Matthieu Geist, Bilal Piot, and Olivier Pietquin. Is the bellman residual a bad proxy? *Advances in Neural Information Processing Systems*, 30, 2017.

[16] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361, 2017.

[17] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870, 2018.

[18] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[19] Han-Dong Lim and Donghwan Lee. Regularized q-learning. *Advances in Neural Information Processing Systems*, 37:129855–129887, 2024.

[20] Odalric-Ambrym Maillard, Rémi Munos, Alessandro Lazaric, and Mohammad Ghavamzadeh. Finite-sample analysis of bellman residual minimization. In *Proceedings of 2nd Asian Conference on Machine Learning*, pages 299–314. JMLR Workshop and Conference Proceedings, 2010.

[21] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[22] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[23] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[24] R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

[25] Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill, 3 edition, 1976.

[26] Ehsan Saleh and Nan Jiang. Deterministic bellman residual minimization. In *Proceedings of Optimization Foundations for Reinforcement Learning Workshop at NeurIPS*, 2019.

[27] Bruno Scherrer. Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view. In *Proceedings of the 27th International Conference on Machine Learning*, pages 959–966, 2010.

[28] Ralf Schoknecht. Optimality of reinforcement learning algorithms with linear function approximation. *Advances in neural information processing systems*, 15, 2002.

[29] Wen Sun and J Andrew Bagnell. Online bellman residual algorithms with predictive error guarantees. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 852–861, 2015.

[30] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[31] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

[32] Richard S Sutton, Hamid Maei, and Csaba Szepesvári. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. *Advances in neural information processing systems*, 21, 2008.

[33] Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th annual international conference on machine learning*, pages 993–1000, 2009.

[34] Tao Wang, Michael Bowling, and Dale Schuurmans. Dual representations for dynamic programming and reinforcement learning. In *2007 IEEE International symposium on approximate dynamic programming and reinforcement learning*, pages 44–51, 2007.

A.1 Definitions, lemmas, and notation

A.1.1 Oblique projection

Here, we briefly describe the concept of oblique projection [27] before proceeding. First, the oblique projection is defined as follows.

Definition A.1 (Oblique projection)**.** *Let $\Psi \in \mathbb{R}^{n\times m}$ be a constant matrix and let $C \subseteq \mathbb{R}^n$ be nonempty, closed, and convex. Define the (possibly set-valued) oblique projection on the domain*

$$\mathrm{dom}(\Gamma_\Psi) := \left\{ y \in \mathbb{R}^n : \arg\min_{x\in C} \frac{1}{2}\|\Psi^\top (x-y)\|_2^2 \neq \varnothing \right\}$$

by

$$\Gamma_\Psi(y) := \arg\min_{x\in C} \frac{1}{2}\|\Psi^\top (x-y)\|_2^2, \qquad y \in \mathrm{dom}(\Gamma_\Psi).$$

When C is the column space $\mathcal{R}(\Phi)$ of the feature matrix $\Phi \in \mathbb{R}^{n\times m}$, we write

$$\Gamma_{\Phi|\Psi}(y) := \arg\min_{x\in \mathcal{R}(\Phi)} \frac{1}{2}\|\Psi^\top (x-y)\|_2^2$$

whenever this argmin is nonempty. Under the nonsingularity condition used below, it is nonempty and single-valued for every y. The geometric meaning of this oblique projection is illustrated in Figure 1. Although the oblique projection above is formulated as an optimization problem, under

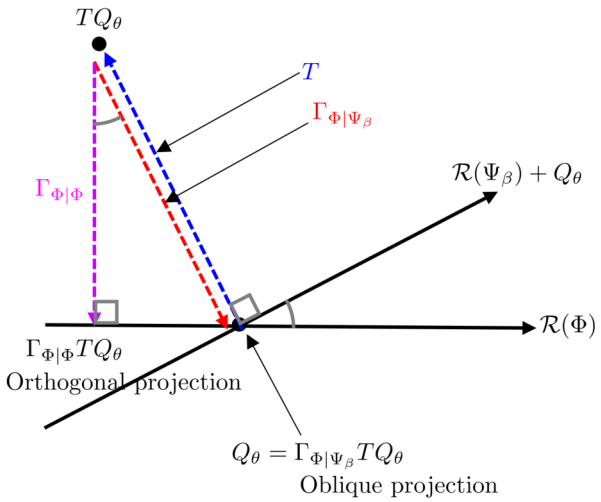


Figure 1: Illustration of the oblique projection

certain conditions it can be expressed simply as a matrix product.

Lemma A.1. *Let $\Psi, \Phi \in \mathbb{R}^{n\times m}$ be some constant matrices. If $\Psi^\top \Phi \in \mathbb{R}^{m\times m}$ is invertible, then*

$$\Gamma_{\Phi|\Psi}(y) = \Phi(\Psi^\top \Phi)^{-1}\Psi^\top y.$$

Proof. First, the parameter defining $\Gamma_{\Phi|\Psi}(y)$ is the unique element of

$$\arg\min_{\theta\in\mathbb{R}^m} \frac{1}{2}\left\|\Psi^\top(\Phi\theta - y)\right\|_2^2.$$

Since the objective function of the above optimization is a strongly convex quadratic function, by the first-order optimality condition, the unique global optimizer satisfies

$$\nabla_\theta \frac{1}{2}\left\|\Psi^\top(\Phi\theta - y)\right\|_2^2 = \Phi^\top\Psi\Psi^\top\Phi\theta - \Phi^\top\Psi\Psi^\top y = 0,$$

Equivalently, with $B := \Psi^\top\Phi$, the normal equation is $B^\top B\theta = B^\top\Psi^\top y$, and since B is invertible, $\theta = B^{-1}\Psi^\top y$. Hence $\Phi\theta = \Phi(\Psi^\top\Phi)^{-1}\Psi^\top y$. ∎

A.1.2 Markov decision problem

Here, we briefly summarize several standard properties related to Markov decision processes and Markov decision problems. These results will be used later to prove various statements in this paper. We first introduce a lemma to bound the norm of a matrix inverse:

Lemma A.2 (Page 351 in Horn and Johnson [18])**.** *If the square matrix $M \in \mathbb{R}^{n\times n}$ satisfies $||M|| < 1$ for some induced matrix norm $||\cdot||$, then $I - M$ is nonsingular, and*

$$\left\|(I-M)^{-1}\right\| \leq \frac{1}{1-\|M\|}.$$

The above lemma can be used to prove the following result, which is frequently used in MDPs.

Lemma A.3. *For any stochastic policy $\pi : \mathcal{S} \to \Delta_{|\mathcal{A}|}$, where $\Delta_{|\mathcal{A}|}$ is the set of all probability distributions over $\mathcal{A}$, the matrix $\gamma P\Pi^\pi - I$ is always nonsingular.*

Proof. Since $\|\gamma P\Pi^\pi\|_\infty = \gamma\|P\Pi^\pi\|_\infty = \gamma < 1$, the proof is completed by directly applying Lemma A.2. □

Lemma A.4 (Automatic nonsingularity of the normal-equation oblique direction)**.** *For any stochastic policy π, define*

$$A_\pi := \gamma P\Pi^\pi - I, \qquad \Omega_\pi := A_\pi^\top A_\pi \Phi.$$

Then

$$\Omega_\pi^\top \Phi = \Phi^\top A_\pi^\top A_\pi \Phi$$

is symmetric positive definite and hence nonsingular.

Proof. By Lemma A.3, A_π is nonsingular. For every nonzero $x \in \mathbb{R}^m$, full column rank of Φ gives $\Phi x \neq 0$, and hence $A_\pi \Phi x \neq 0$. Therefore,

$$x^\top \Omega_\pi^\top \Phi x = x^\top \Phi^\top A_\pi^\top A_\pi \Phi x = \|A_\pi \Phi x\|_2^2 > 0.$$

Thus $\Omega_\pi^\top \Phi$ is symmetric positive definite. □

The following two lemmas are also results that are used very frequently in our theoretical analysis throughout this paper.

Lemma A.5. *For any $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, we have $\|Q - Q^\pi\|_\infty \leq \frac{1}{1-\gamma}\|T^\pi Q - Q\|_\infty \leq \frac{1}{1-\gamma}\|T^\pi Q - Q\|_2$.*

Proof. Using the contraction property of T^π, one gets

$$\begin{aligned}\|Q - Q^\pi\|_\infty =& \|T^\pi Q - T^\pi Q^\pi + Q - T^\pi Q\|_\infty \\ \leq& \|T^\pi Q - T^\pi Q^\pi\|_\infty + \|Q - T^\pi Q\|_\infty \\ \leq& \gamma\|Q - Q^\pi\|_\infty + \|Q - T^\pi Q\|_\infty.\end{aligned}$$

Rearranging terms leads to $(1-\gamma)\|Q - Q^\pi\|_\infty \leq \|T^\pi Q - Q\|_\infty \leq \|T^\pi Q - Q\|_2$. This completes the proof. □

Lemma A.6. *For any $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, we have $\|T^\pi Q - Q\|_\infty \leq (1+\gamma)\|Q - Q^\pi\|_\infty$.*

Proof. Using the contraction property of T^π, one gets

$$\begin{aligned}\|T^\pi Q - Q\|_\infty =& \|T^\pi Q - T^\pi Q^\pi - Q + Q^\pi\|_\infty \\ \leq& \|T^\pi Q - T^\pi Q^\pi\|_\infty + \|Q - Q^\pi\|_\infty \\ \leq& (1+\gamma)\|Q - Q^\pi\|_\infty.\end{aligned}$$

This completes the proof. □

Similar to the two lemmas above, the following two lemmas, applied to the optimal Bellman operator T, also play an important role in this paper.

Lemma A.7. *For any $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, we have*

$$\|Q - Q^*\|_\infty \leq \frac{1}{1-\gamma}\|TQ - Q\|_\infty \leq \frac{1}{1-\gamma}\|TQ - Q\|_2.$$

Proof. Using the contraction property of T, one gets

$$\begin{aligned}\|Q - Q^*\|_\infty =& \|TQ - TQ^* + Q - TQ\|_\infty \\ \leq& \|TQ - TQ^*\|_\infty + \|Q - TQ\|_\infty \\ \leq& \gamma\|Q - Q^*\|_\infty + \|Q - TQ\|_\infty.\end{aligned}$$

Rearranging terms leads to $(1-\gamma)\|Q - Q^*\|_\infty \leq \|TQ - Q\|_\infty \leq \|TQ - Q\|_2$. This completes the proof. □

Lemma A.8. *For any $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, we have $\|TQ - Q\|_\infty \leq (1+\gamma)\|Q - Q^*\|_\infty$.*

Proof. Using the contraction property of T, one gets

$$\begin{aligned}\|TQ - Q\|_\infty =& \|TQ - TQ^* - Q + Q^*\|_\infty \\ \leq& \|TQ - TQ^*\|_\infty + \|Q - Q^*\|_\infty \\ \leq& (1+\gamma)\|Q - Q^*\|_\infty.\end{aligned}$$

This completes the proof. □

The following policy-evaluation result adapts the oblique-projection characterization and error bound of Scherrer [27, Propositions 2 and 3] to state-action values with uniform weighting. We include a proof to keep the presentation self-contained.

Lemma A.9 (Adapted from Scherrer [27])**.** *For a given policy π, let us consider the Bellman residual objective*

$$f(\theta) = \frac{1}{2}\|T^\pi\Phi\theta - \Phi\theta\|_2^2$$

for policy evaluation, and let $\Psi := (\gamma P\Pi^\pi - I)\Phi$. Its stationary point satisfying $\nabla_\theta f(\bar\theta) = \Phi^\top(\gamma P\Pi^\pi - I)^\top(T^\pi Q_{\bar\theta} - Q_{\bar\theta}) = 0$ is unique. If $\Psi^\top\Phi$ is nonsingular, then this stationary point also satisfies the oblique projected Bellman equation

$$Q_{\bar\theta} = \Gamma_{\Phi|\Psi}T^\pi Q_{\bar\theta}.$$

Moreover, we have

$$\|Q^\pi - Q_{\bar\theta}\|_2 \leq \left\|I - \Gamma_{\Phi|\Omega}\right\|_2 \left\|Q^\pi - \Gamma_{\Phi|\Phi}Q^\pi\right\|_2$$

where $\Omega := (\gamma P\Pi^\pi - I)^\top(\gamma P\Pi^\pi - I)\Phi$.

Proof. The uniqueness of the stationary point comes from the strong convexity of f; equivalently, the normal-equation matrix is positive definite by Lemma A.4. The stationary point equation is given as

$$\Phi^\top(\gamma P\Pi^\pi - I)^\top(R + \gamma P\Pi^\pi\Phi\bar\theta - \Phi\bar\theta) = 0.$$

Equivalently, $\Psi^\top(\Phi\bar\theta - T^\pi Q_{\bar\theta}) = 0$. If $\Psi^\top\Phi$ is nonsingular, this gives

$$Q_{\bar\theta} = \Phi(\Psi^\top\Phi)^{-1}\Psi^\top T^\pi Q_{\bar\theta} = \Gamma_{\Phi|\Psi}T^\pi Q_{\bar\theta}.$$

After algebraic manipulations, it can be seen that the above equation is equivalent to

$$\bar\theta = -\Big\{\Phi^\top(\gamma P\Pi^\pi - I)^\top(\gamma P\Pi^\pi - I)\Phi\Big\}^{-1}\Phi^\top(\gamma P\Pi^\pi - I)^\top R,$$

which can be further expressed as

$$\bar\theta = -\Big\{\Phi^\top(\gamma P\Pi^\pi - I)^\top(\gamma P\Pi^\pi - I)\Phi\Big\}^{-1}\Phi^\top(\gamma P\Pi^\pi - I)^\top(\gamma P\Pi^\pi - I)(\gamma P\Pi^\pi - I)^{-1}R$$

Equivalently, using $Q^\pi = (I - \gamma P\Pi^\pi)^{-1}R$, we have

$$\Phi\bar\theta = \Phi\Big\{\Phi^\top(\gamma P\Pi^\pi - I)^\top(\gamma P\Pi^\pi - I)\Phi\Big\}^{-1}\Phi^\top(\gamma P\Pi^\pi - I)^\top(\gamma P\Pi^\pi - I)Q^\pi = \Gamma_{\Phi|\Omega}Q^\pi$$

Using this equation, we can derive

$$\begin{aligned}Q^\pi - Q_{\bar\theta} =& Q^\pi - \Gamma_{\Phi|\Omega}Q^\pi \\ =& (I - \Gamma_{\Phi|\Omega})Q^\pi \\ =& (I - \Gamma_{\Phi|\Omega})(I - \Gamma_{\Phi|\Phi})Q^\pi \\ =& (I - \Gamma_{\Phi|\Omega})(Q^\pi - \Gamma_{\Phi|\Phi}Q^\pi),\end{aligned}$$

where the third line is due to $(I - \Gamma_{\Phi|\Omega})(I - \Gamma_{\Phi|\Phi}) = I - \Gamma_{\Phi|\Omega}$. Finally, we obtain the desired conclusion: $\|Q^\pi - Q_{\bar{\theta}}\|_2 \leq \left\|I - \Gamma_{\Phi|\Omega}\right\|_2 \left\|Q^\pi - \Gamma_{\Phi|\Phi} Q^\pi\right\|_2$. □

A.1.3 Clarke subdifferential

To analyze and minimize the control Bellman residual (CBR) objective function in this paper, we need to understand the notion of subdifferentials, which serve as a counterpart to derivatives for nonconvex and nonsmooth functions, as well as several of their key properties. Accordingly, in this section we briefly review some well-established basic concepts related to such subdifferentials. Throughout, let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and consider the optimization problem

$$x^* \in \arg\min_{x \in \mathbb{R}^n} f(x),$$

where f is potentially nonconvex and nonsmooth. First, we define the generalized directional derivative.

Definition A.2 (Clarke [11], Bagirov et al. [1]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at x. The generalized directional derivative of f at $x \in \mathbb{R}^n$ in the direction $d \in \mathbb{R}^n$ is defined by*

$$f^\circ(x; d) := \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y + td) - f(y)}{t}.$$

Then, based on the above definition, the Clarke subdifferential (Clarke generalized gradient) is given as follows.

Definition A.3 (Clarke [11], Bagirov et al. [1]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at x. The subdifferential of f at $x \in \mathbb{R}^n$ is the set $\partial f(x)$ of vectors such that*

$$\partial f(x) := \left\{ v \in \mathbb{R}^n : \ \langle v, d \rangle \leq f^\circ(x; d) \ \ \forall d \in \mathbb{R}^n \right\}.$$

It is standard that $\partial f(x)$ is a nonempty, convex, and compact set [1, Theorem 3.3, page 64].

Lemma A.10 (Bagirov et al. [1], Theorem 3.3). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at x with a Lipschitz constant $L > 0$. Then, the subdifferential $\partial f(x)$ is a nonempty, convex, and compact set such that*

$$\partial f(x) \subseteq B(0; L),$$

where $B(z; r)$ is a ball centered at z with radius $r > 0$.

Moreover, the generalized directional derivative above can be expressed equivalently as follows.

Lemma A.11 (Bagirov et al. [1], Theorem 3.4). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at x. Then, for every $d \in \mathbb{R}^n$, one has*

$$f^\circ(x; d) = \max_{v \in \partial f(x)} \langle v, d \rangle.$$

Next, we summarize several properties related to continuity.

Lemma A.12 (Bagirov et al. [1], Theorem 3.2). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at x. Then, the function $(x, d) \mapsto f^\circ(x; d)$ is upper semicontinuous.*

Lemma A.13 (Bagirov et al. [1], Theorem 3.5). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at x. Then, the set-valued function $x \mapsto \partial f(x)$ is upper semicontinuous.*

The following result is essential when calculating subgradients in practice. Namely, the subdifferential can be constructed as a convex hull of all possible limits of gradients at point x_k converging to x.

Lemma A.14 (Bagirov et al. [1], Theorem 3.9). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at $x \in \mathbb{R}^n$. Then,*

$$\partial f(x) := \operatorname{conv}\left(\left\{ g \in \mathbb{R}^n : \exists (x_i)_{i=1}^\infty, \lim_{i \to \infty} x_i = x, \lim_{i \to \infty} \nabla_x f(x)|_{x = x_i} = g \right\} \right).$$

Next, we define the notions of a generalized stationary point and a descent direction.

Definition A.4 (Bagirov et al. [1], Definition 4.3). *A point x satisfying $0 \in \partial f(x)$ is called a stationary point of f.*

Definition A.5 (Bagirov et al. [1], Definition 4.5)**.** *Direction $d \in \mathbb{R}^n$ is called a descent direction for $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$, if there exists $\varepsilon > 0$ such that for all $t \in (0, \varepsilon]$*

$$f(x + td) < f(x).$$

Using the subdifferential, we can obtain the following generalized first-order necessary conditions for a point to be an extremum.

Lemma A.15 (Bagirov et al. [1], Theorem 3.17)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at x. If it attains its extremum at x, then*

$$0 \in \partial f(x).$$

Lemma A.16 (Bagirov et al. [1], Theorem 4.1)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at x^*. If f attains its local minimum at x^*, then*

$$0 \in \partial f(x^*), \quad f^\circ(x^*; d) \geq 0, \quad \forall d \in \mathbb{R}^n.$$

Lemma A.17 (Bagirov et al. [1], Theorem 4.5)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at $x \in \mathbb{R}^n$. The direction $d \in \mathbb{R}^n$ is a descent direction for f at x if*

$$g^\top d < 0, \quad \forall g \in \partial f(x)$$

or equivalently, $f^\circ(x; d) < 0$.

Lemma A.18 (Bagirov et al. [1], Corollary 4.1)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at $x \in \mathbb{R}^n$. Then either $0 \in \partial f(x)$ or there exists a descent direction $d \in \mathbb{R}^n$ for f at $x \in \mathbb{R}^n$.*

Lemma A.19 (Burke et al. [10])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function at $x \in \mathbb{R}^n$. Fix $x \in \mathbb{R}^n$ and let $g = \mathrm{Proj}_{\partial f(x)}(0)$ be the Euclidean projection of 0 onto $\partial f(x)$ or the minimum-norm subgradient*

$$g = \mathrm{Proj}_{\partial f(x)}(0) \in \operatorname*{arg\,min}_{v \in \partial f(x)} \|v\|_2.$$

Then $f^\circ(x; -g) \leq -\|g\|_2^2$. In particular, if $g \neq 0$, then $-g$ is a strict descent direction.

Proof. Since $\partial f(x)$ is closed and convex, the projection optimality condition [5, Proposition 2.2.1, page 88] gives

$$\langle v - g,\, 0 - g \rangle \leq 0 \qquad \forall v \in \partial f(x),$$

equivalently, it can be rewritten as follows:

$$\langle v, g \rangle \geq \|g\|^2 \qquad \forall v \in \partial f(x).$$

Let $d = -g$. Then for all $v \in \partial f(x)$,

$$\langle v, d \rangle = -\langle v, g \rangle \leq -\|g\|^2.$$

Using the support-function representation of the Clarke directional derivative,

$$f^\circ(x; d) = \max_{v \in \partial f(x)} \langle v, d \rangle,$$

we obtain $f^\circ(x; -g) \leq -\|g\|^2$. This completes the proof. □

In particular, when f is a piecewise quadratic function, we can obtain several special-case properties that are useful in practice. Before proceeding, we formally define the class of piecewise quadratic functions considered in this paper.

Definition A.6. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called (finite) piecewise quadratic if there exist a finite index set $\mathcal{I} = \{1, \ldots, N\}$, a polyhedral partition $\{U_i\}_{i \in \mathcal{I}}$ of $\mathbb{R}^n$ (i.e., U_i are polyhedra with pairwise disjoint interiors and $\bigcup_i U_i = \mathbb{R}^n$), and quadratic polynomials*

$$f_i(x) = \tfrac{1}{2} x^\top H_i x + b_i^\top x + c_i, \qquad H_i = H_i^\top,$$

such that

$$f(x) = f_i(x) \qquad \forall x \in U_i.$$

In particular, each f_i is $\mathcal{C}^\infty$. Therefore, f is piecewise $\mathcal{C}^\infty$ with finitely many pieces.

It can be shown that any piecewise quadratic function f defined as above is always locally Lipschitz.

Lemma A.20. *Every finite piecewise quadratic f is locally Lipschitz on $\mathbb{R}^n$.*

Proof. Let us consider a piecewise quadratic function defined in Definition A.6. Let us fix any $x_0 \in \mathbb{R}^n$ and radius $r > 0$, and consider the closed ball

$$B(x_0, r) = \{x \in \mathbb{R}^n : \|x - x_0\|_2 \leq r\}.$$

Because each f_i is quadratic, its gradient is affine: $\nabla f_i(x) = H_i x + b_i$. Hence on the compact set $B(x_0, r)$,

$$M_i := \max_{x \in B(x_0,r)} \|\nabla f_i(x)\|_2 < \infty.$$

Let $M := \max_{i \in \mathcal{I}} M_i < \infty$. Now, let us take any $x, y \in B(x_0, r)$ and consider the line segment

$$\ell(t) := y + t(x - y), \qquad t \in [0, 1].$$

Restricting the finite polyhedral arrangement to the segment yields a finite subdivision into intervals on which the active cell is constant: each defining affine inequality either changes sign at at most one point of the segment or is identically zero on the segment. Therefore, there exist breakpoints

$$0 = t_0 < t_1 < \cdots < t_m = 1$$

and indices $i_1, \ldots, i_m \in \mathcal{I}$ such that $\ell(t) \in U_{i_j}$ for all $t \in [t_{j-1}, t_j]$. Let us define $z_j := \ell(t_j)$ so that $z_0 = y$ and $z_m = x$. On each subsegment $[z_{j-1}, z_j] \subset B(x_0, r)$, we have $f = f_{i_j}$, and by the mean value theorem,

$$|f(z_j) - f(z_{j-1})| = |f_{i_j}(z_j) - f_{i_j}(z_{j-1})| \leq \sup_{\xi \in [z_{j-1}, z_j]} \|\nabla f_{i_j}(\xi)\|_2 \, \|z_j - z_{j-1}\|_2 \leq M \, \|z_j - z_{j-1}\|_2.$$

Summing over $j = 1, \ldots, m$ yields

$$|f(x) - f(y)| \leq \sum_{j=1}^{m} |f(z_j) - f(z_{j-1})| \leq M \sum_{j=1}^{m} \|z_j - z_{j-1}\|_2.$$

But all z_j lie on the same segment from y to x, so the subsegment lengths add up to the total length:

$$\sum_{j=1}^{m} \|z_j - z_{j-1}\|_2 = \|x - y\|_2.$$

Therefore, $|f(x) - f(y)| \leq M\|x - y\|_2$ for all $x, y \in B(x_0, r)$, which implies that f is Lipschitz on $B(x_0, r)$; since x_0 and r were arbitrary, f is locally Lipschitz on $\mathbb{R}^n$. This concludes the proof. □

A.2 Generalized gradient descent method

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function (not necessarily convex or smooth). We study a generalized gradient (subgradient) descent method [10] based on the Clarke subdifferential and present a convergence analysis showing, under the active-set nondegeneracy condition introduced below, that every limit point is Clarke-stationary. A key issue in nonsmooth nonconvex optimization is that an *arbitrary* choice $g \in \partial f(x)$ does not necessarily yield a descent direction. A classical remedy [10] is to select the minimum-norm Clarke subgradient $g_k \in \arg\min_{v \in \partial f(x_k)} \|v\|_2 = \mathrm{Proj}_{\partial f(x_k)}(0)$, and update

$$x_{k+1} = x_k - \alpha_k g_k,$$

with some step size $\alpha_k > 0$. Note that by construction, we can easily see

$$\|g_k\|_2 = \mathrm{dist}\big(0, \partial f(x_k)\big).$$

The step sizes are selected based on the Armijo rule [9]. In particular, we first fix parameters $\bar{\alpha} > 0$, $\beta \in (0, 1)$, and $\sigma \in (0, 1)$. If $g_k = 0$, stop. Otherwise, set $d_k := -g_k$ and choose the step size $\alpha_k = \beta^{m_k} \bar{\alpha}$ with the smallest integer $m_k \geq 0$ such that the Armijo condition with Clarke derivative holds:

$$f(x_k - \alpha_k g_k) \;\leq\; f(x_k) \;+\; \sigma \, \alpha_k \, f^{\circ}(x_k; d_k). \tag{11}$$

Then, we update the parameter by

$$x_{k+1} = x_k - \alpha_k g_k.$$

Based on the above discussion and preliminary developments, we can now consider the algorithm presented in Algorithm 1.

The following lemma provides a guarantee on the existence of a step size that satisfies the Armijo rule.

Algorithm 1 Generalized gradient method

1: Initialize x_0
2: **for** $k \in \{0, 1, 2, \ldots\}$ **do**
3: Calculate $g_k \in \arg\min_{v \in \partial f(x_k)} \|v\|_2^2$
4: Perform a subgradient descent step $x_{k+1} = x_k - \alpha_k g_k$, where the step size $\alpha_k > 0$ is selected based on the Armijo rule
5: **end for**

Lemma A.21 (Existence of an Armijo step)**.** *Suppose f is locally Lipschitz and $f^\circ(x; d) < 0$. Fix any $\sigma \in (0, 1)$. Then there exists $\delta > 0$ such that for all $\alpha \in (0, \delta)$,*

$$f(x + \alpha d) \;\le\; f(x) + \sigma\alpha f^\circ(x; d).$$

Consequently, the backtracking rule (11) *terminates in finitely many reductions.*

Proof. First of all, by the definition of the generalized directional derivative, one has

$$f^\circ(x; d) := \limsup_{y \to x, \alpha \downarrow 0} \frac{f(y + \alpha d) - f(y)}{\alpha} \ge \limsup_{\alpha \downarrow 0} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

By definition of $\limsup$, for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all $\alpha \in (0, \delta)$,

$$\limsup_{t \downarrow 0} \frac{f(x + td) - f(x)}{t} + \varepsilon \ge \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

Combining the last two inequalities leads to

$$\frac{f(x + \alpha d) - f(x)}{\alpha} \le \limsup_{t \downarrow 0} \frac{f(x + td) - f(x)}{t} + \varepsilon \le f^\circ(x; d) + \varepsilon.$$

Now, let us choose $\varepsilon = (1-\sigma)(-f^\circ(x; d)) > 0$ (possible since $f^\circ(x; d) < 0$), so that $f^\circ(x; d) + \varepsilon = \sigma f^\circ(x; d)$. Multiplying by α yields the desired inequality. □

Definition A.7 (Local active-set stability)**.** *A finite piecewise quadratic function f is called locally active-set stable at x if there exist an open neighborhood U_x and one quadratic piece $f_{i(x)}$ such that $f = f_{i(x)}$ throughout U_x.*

Lemma A.22 (Uniform Armijo step under active-set stability)**.** *Let f be finite piecewise quadratic, let $K \subset \mathbb{R}^n$ be compact, and let $g(x) = \mathrm{Proj}_{\partial f(x)}(0)$. Suppose that f is locally active-set stable at every point of an open neighborhood of K and that $\|g(x)\|_2 \ge \eta > 0$ for all $x \in K$. Then, for every $\sigma \in (0, 1)$, there exists $\tau > 0$ such that, for every $x \in K$ and every $\alpha \in (0, \tau]$,*

$$f(x - \alpha g(x)) \le f(x) + \sigma\alpha f^\circ(x; -g(x)).$$

Proof. For every $x \in K$, choose a bounded open set U_x on which f equals one quadratic piece, and then choose an open set V_x with $x \in V_x$ and compact closure $\overline{V}_x \subset U_x$. By compactness, finitely many such sets V_j cover K; write U_j for their corresponding branch neighborhoods, where $f = f_{i_j}$ on U_j. Set

$$\rho := \min_j \mathrm{dist}(\overline{V}_j, U_j^c) > 0, \qquad M := \max_j \|H_{i_j}\|_2,$$

where H_{i_j} is the Hessian of f_{i_j}. Since f is C^1 on an open neighborhood of K, $g(x) = \nabla f(x) = \nabla f_{i_j}(x)$ whenever $x \in V_j$, and $G := \max_{x \in K} \|g(x)\|_2 < \infty$. Choose

$$0 < \tau \le \min\left\{ \frac{\rho}{2G}, \frac{2(1-\sigma)}{M} \right\},$$

with the second term omitted when $M = 0$. If $x \in K$, choose j with $x \in V_j$. For $0 < \alpha \le \tau$, every point on the segment from x to $x - \alpha g(x)$ lies within distance at most $\rho/2$ of $\overline{V}_j$ and hence remains in U_j, and the exact quadratic expansion gives

$$f(x - \alpha g(x)) \le f(x) - \alpha\|g(x)\|_2^2 + \frac{M}{2}\alpha^2\|g(x)\|_2^2 \le f(x) - \sigma\alpha\|g(x)\|_2^2.$$

On U_j, $f^\circ(x; -g(x)) = -\|g(x)\|_2^2$, proving the claim. □

We are now ready to establish monotone decrease and stationarity of the limit points of Algorithm 1. The required assumptions are summarized below.

Assumption A.1. *We assume*

1. *f is a finite piecewise quadratic function, locally Lipschitz, and bounded below:* $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$.
2. *The initial sublevel set $\mathcal{L}_c := \{x : \ f(x) \le c\}$ with $c = f(x_0)$ is bounded (hence compact).*

Assumption A.2 (Active-set nondegeneracy)**.** *Every limit point $\bar{x}$ of the generated sequence that is not Clarke-stationary is locally active-set stable in the sense of Definition A.7. In particular, a genuine switching boundary may be a limit point only if it is already Clarke-stationary.*

The following result establishes the convergence of the algorithm under the above assumptions.

Theorem A.1. *Let $(x_k)_{k\ge0}$ be generated by Algorithm 1 under Assumptions A.1 and A.2. Then the following results hold:*

1. *(Monotone decrease) $f(x_{k+1}) \le f(x_k)$ for all $k \ge 0$, and $(f(x_k))_{k\ge0}$ converges.*
2. *(Sufficient decrease) For all $k \ge 0$,*
$$f(x_{k+1}) \le f(x_k) - \sigma\alpha_k\|g_k\|_2^2, \tag{12}$$
and hence $\sum_{k=0}^{\infty} \alpha_k\|g_k\|_2^2 < \infty$.
3. *(Stationarity of limit points) The sequence $\{x_k\} \subset \mathcal{L}_c$ with $c = f(x_0)$ admits at least one limit point. Every limit point $\bar{x}$ satisfies*
$$0 \in \partial f(\bar{x}),$$
i.e., $\bar{x}$ is a stationary point.

Proof. Let us fix $k \ge 0$. If $g_k = 0$, then $0 \in \partial f(x_k)$ and the algorithm stops at a stationary point. Therefore, let us assume $g_k \ne 0$. By Lemma A.19, $f^\circ(x_k; d_k) = f^\circ(x_k; -g_k) \le -\|g_k\|_2^2 < 0$. Moreover, by Lemma A.21, since the Armijo step size exists, the backtracking rule terminates and yields $\alpha_k > 0$ satisfying (11). Combining (11) with $f^\circ(x_k; d_k) \le -\|g_k\|_2^2$ gives
$$f(x_{k+1}) = f(x_k + \alpha_k d_k) \le f(x_k) + \sigma\alpha_k f^\circ(x_k; d_k) \le f(x_k) - \sigma\alpha_k\|g_k\|_2^2,$$
which is (12) and implies $f(x_{k+1}) \le f(x_k)$. Since f is bounded below, the monotone sequence $(f(x_k))_{k\ge0}$ converges. Summing (12) over $k = 0, \dots, N$ yields
$$\sum_{k=0}^{N} \sigma\alpha_k\|g_k\|_2^2 \le f(x_0) - f(x_{N+1}) \le f(x_0) - \inf f < \infty,$$
and letting $N \to \infty$ gives $\sum_{k=0}^{\infty} \alpha_k\|g_k\|_2^2 < \infty$. By (12), $f(x_k) \le f(x_0)$ for all k, hence $x_k \in \mathcal{L}_c$ with $c = f(x_0)$ for all $k \ge 0$. By Assumption A.1(2), $\mathcal{L}_c$ is compact, and hence, $(x_k)_{k\ge0}$ has at least one limit point [25, Theorem 3.6]. Let $\bar{x}$ be a limit point and take a subsequence $x_{k_j} \to \bar{x}$. Now, let us assume for contradiction that $\bar{x}$ is not stationary, i.e., $0 \notin \partial f(\bar{x})$. Let us define the distance $\delta := \operatorname{dist}(0, \partial f(\bar{x})) > 0$. For locally Lipschitz f, the set-valued mapping $x \mapsto \partial f(x)$ is upper semicontinuous. Therefore, by the definition of the upper semicontinuous set-valued mapping, there exists a neighborhood U of $\bar{x}$ such that
$$\operatorname{dist}(0, \partial f(x)) \ge \delta/2 \qquad \forall x \in U. \tag{13}$$
For each $x \in U$, let $g(x)$ denote the minimum-norm element of $\partial f(x)$. Then, (13) yields $\|g(x)\|_2 \ge \delta/2$ for all $x \in U$. By Lemma A.19, we have
$$f^\circ(x; -g(x)) \le -\|g(x)\|_2^2 \le -\delta^2/4 \qquad \forall x \in U.$$
By Assumption A.2, there is an open neighborhood $V \subset U$ of $\bar{x}$ on which f agrees with one quadratic piece. Choose a compact neighborhood $K \subset V$ of $\bar{x}$. Since $x_{k_j} \to \bar{x}$, we have $x_{k_j} \in K$ for all sufficiently large j. By Lemma A.22 with $\eta = \delta/2$, there exists $\tau > 0$ such that the Armijo condition holds for every $x_k \in K$ and every trial step $t \in (0, \tau]$. Because the trial steps are $\{\bar{\alpha}, \beta\bar{\alpha}, \beta^2\bar{\alpha}, \dots\}$, the accepted step satisfies $\alpha_k \ge \alpha_{\min} := \min\{\bar{\alpha}, \beta\tau\} > 0$ whenever $x_k \in K$. Therefore, for such k,
$$f(x_{k+1}) \le f(x_k) - \sigma\alpha_k\|g_k\|_2^2 \le f(x_k) - \sigma\alpha_{\min}(\delta/2)^2 = f(x_k) - c,$$

where $c := \sigma\alpha_{\min}\delta^2/4 > 0$ is a constant. Since $x_{k_j} \to \bar{x}$, we have $x_{k_j} \in K$ for all sufficiently large $j \geq 0$. Therefore, the inequality above would force $f(x_k)$ to decrease by at least $c > 0$ infinitely often, which contradicts the convergence of $f(x_k)$. Therefore, the contradiction proves $0 \in \partial f(\bar{x})$. This shows that every limit point is Clarke stationary. □

A.3 CBR tabular case

Let us consider the CBR objective

$$f(Q) := \frac{1}{2} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} ((TQ)(s,a) - Q(s,a))^2 = \frac{1}{2}\|TQ - Q\|_2^2.$$

Let Θ denote the set of deterministic policies $\pi : \mathcal{S} \to \mathcal{A}$, so $|\Theta| = |\mathcal{A}|^{|\mathcal{S}|} < \infty$. Let us define the set

$$S_\pi := \left\{ Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} : \pi(s) \in \arg\max_{a\in\mathcal{A}} Q(s,a),\ \forall s \in \mathcal{S} \right\}$$

for each $\pi \in \Theta$, where $\arg\max$ is interpreted as a set-valued map, i.e., $\pi(s)$ is one of the maximizers; ties are allowed.

Proposition A.1. *f is piecewise quadratic (so piecewise smooth) and continuous.*

Proof. For each deterministic policy $\pi \in \Theta$, define the quadratic polynomial

$$f_\pi(Q) := \frac{1}{2}\,\|R + (\gamma P\Pi^\pi - I)Q\|_2^2\,.$$

If $Q \in S_\pi$, then $\pi(s)$ is a maximizing action at every state, so $\Pi^\pi Q$ is the vector of statewise maxima and $TQ = T^\pi Q = R + \gamma P\Pi^\pi Q$. Hence $f(Q) = f_\pi(Q)$ on S_π. The finitely many polyhedra $\{S_\pi\}_{\pi\in\Theta}$ cover the whole space, and the arrangement of their defining action-comparison hyperplanes yields a finite polyhedral partition refining this cover. On each cell, f therefore agrees with one of the quadratic polynomials f_π. No selector is required to remain constant on a tie set. Indeed, if $Q \in S_\pi \cap S_{\tilde{\pi}}$, then both policies select maximizing actions, so $\Pi^\pi Q = \Pi^{\tilde{\pi}} Q$, $T^\pi Q = T^{\tilde{\pi}} Q = TQ$, and $f_\pi(Q) = f_{\tilde{\pi}}(Q)$. Thus the quadratic descriptions are compatible on overlaps, and f is finite piecewise quadratic in the sense of Definition A.6. Finally, TQ is continuous because each statewise maximum is the maximum of finitely many linear coordinate maps; consequently, $f(Q) = \frac{1}{2}\|TQ - Q\|_2^2$ is continuous. □

From the above results, we know that f is piecewise quadratic. However, we can obtain further insights into the regions S_π over which it is quadratic.

Lemma A.23. *For each $\pi \in \Theta$, the set S_π is an intersection of homogeneous half-spaces.*

Proof. For each $s \in \mathcal{S}$, the condition $\pi(s) \in \arg\max_{a\in\mathcal{A}} Q(s,a)$ is equivalent to the family of linear inequalities

$$Q(s,\pi(s)) \ \geq\ Q(s,a), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Hence, the set S_π can be written as

$$S_\pi = \bigcap_{s\in\mathcal{S}} \bigcap_{a\in\mathcal{A}} \left\{ Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} :\ Q(s,\pi(s)) - Q(s,a) \geq 0 \right\} = \left\{ Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} :\ L\Pi^\pi Q - Q \geq 0 \right\},$$

where $L \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|}$ is a lifting matrix defined as $L := I_{|\mathcal{S}|} \otimes \mathbf{1}_{|\mathcal{A}|}$, where $I_{|\mathcal{S}|}$ is the identity matrix of dimension $|\mathcal{S}|$, $\otimes$ denotes the Kronecker product, $\mathbf{1}_{|\mathcal{A}|}$ is the all-ones vector of dimension $|\mathcal{A}|$. Each set in the intersection is a closed half-space containing the origin. Therefore, it is an intersection of homogeneous half-spaces. □

Lemma A.24. *For each $\pi \in \Theta$, the set S_π is a convex cone.*

Proof. First of all, let us fix any $\pi \in \Theta$ and consider any $Q \in S_\pi$. For every $\alpha > 0$, multiplication by α preserves the maximizing actions, so $\alpha Q \in S_\pi$. For $\alpha = 0$, all actions are tied and therefore $\pi(s)$ is a maximizer at every state, which gives $0 \in S_\pi$. Hence S_π is a cone. Moreover, let $Q_1, Q_2 \in S_\pi$ and $\alpha \in [0,1]$. For all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$Q_1(s,\pi(s)) \geq Q_1(s,a), \qquad Q_2(s,\pi(s)) \geq Q_2(s,a).$$

Multiplying the first inequality by $\alpha \in [0,1]$ and the second by $1-\alpha$, and adding, yields

$$\alpha Q_1(s,\pi(s)) + (1-\alpha)Q_2(s,\pi(s)) \geq \alpha Q_1(s,a) + (1-\alpha)Q_2(s,a),$$

which shows that $\alpha Q_1 + (1-\alpha)Q_2 \in S_\pi$. Hence S_π is convex. Combining the two parts, we conclude that S_π is a convex cone. □

Lemma A.25. *The following statements hold:*

1. *f is bounded below.*
2. *f is locally Lipschitz continuous.*
3. *The Clarke subdifferential of f given by $\partial f(Q) = \{(\gamma P\Pi^\beta - I)^\top (TQ - Q) : \beta \in \operatorname{conv}(\Lambda(Q))\}$ is nonempty, convex, and bounded.*
4. *We have $\|TQ\|_\infty \leq R_{\max} + \gamma\|Q\|_\infty$, and $\|TQ - Q\|_\infty \geq (1-\gamma)\|Q\|_\infty - R_{\max}$.*
5. *f is coercive.*
6. *Every sublevel set, $\mathcal{L}_c := \{Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : f(Q) \leq c\}$, is bounded for any $c \geq 0$.*
7. *$Q \in \mathcal{L}_c$ implies $\|Q\|_\infty \leq \frac{R_{\max} + \sqrt{2c}}{1-\gamma}$.*
8. *Writing $n := |\mathcal{S} \times \mathcal{A}|$, the following explicit global nonsmooth PL inequality holds:*

$$f(Q) - f^* = f(Q) \leq \frac{n}{2(1-\gamma)^2} \operatorname{dist}(0, \partial f(Q))^2, \qquad Q \in \mathbb{R}^n.$$

Equivalently, one may take $\mu = (1-\gamma)^2/n$.

Proof.

1. Since f is a nonnegative sum of squares, f is bounded below by 0.
2. Moreover, by Proposition A.1, f is piecewise quadratic with finitely many pieces. Therefore, it is locally Lipschitz by Lemma A.20.
3. Since f is locally Lipschitz by the second statement, the desired conclusion is directly obtained using Lemma A.10.
4. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we can derive the following bounds:

$$|(TQ)(s,a)| \leq |R(s,a)| + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} |Q(s',a')| \leq R_{\max} + \gamma\|Q\|_\infty.$$

   This implies $\|TQ\|_\infty \leq R_{\max} + \gamma\|Q\|_\infty$. Moreover, we have

$$\|TQ - Q\|_\infty \geq \|Q\|_\infty - \|TQ\|_\infty \geq (1-\gamma)\|Q\|_\infty - R_{\max},$$

   where the last inequality uses $\|TQ\|_\infty \leq R_{\max} + \gamma\|Q\|_\infty$.
5. Using the bounds in the fourth statement, whenever $\|Q\|_\infty \geq R_{\max}/(1-\gamma)$, the quantity $(1-\gamma)\|Q\|_\infty - R_{\max}$ is nonnegative and hence

$$f(Q) = \frac{1}{2}\|TQ - Q\|_2^2 \geq \frac{1}{2}\|TQ - Q\|_\infty^2 \geq \frac{1}{2}\{(1-\gamma)\|Q\|_\infty - R_{\max}\}^2.$$

   The last expression tends to infinity as $\|Q\|_\infty \to \infty$, so f is coercive.
6. The coercivity in the fifth statement implies that the level sets are bounded.
7. For the seventh statement, fix $c \geq 0$ and let $Q \in \mathcal{L}_c$. By the fourth statement,

$$(1-\gamma)\|Q\|_\infty - R_{\max} \leq \|TQ - Q\|_\infty \leq \|TQ - Q\|_2 = \sqrt{2f(Q)} \leq \sqrt{2c}.$$

   Rearranging gives $\|Q\|_\infty \leq (R_{\max} + \sqrt{2c})/(1-\gamma)$, as desired.
8. Let $r := TQ - Q$ and $M_\beta := P\Pi^\beta$. For every admissible stochastic policy mixture β, M_β is row-stochastic and

$$(I - \gamma M_\beta)^{-1} = \sum_{k=0}^{\infty} \gamma^k M_\beta^k, \qquad \|(I - \gamma M_\beta)^{-1}\|_\infty \leq \frac{1}{1-\gamma}.$$

Since $\|A\|_2 \le \sqrt{n}\|A\|_\infty$ for $A \in \mathbb{R}^{n\times n}$,

$$\sigma_{\min}(I - \gamma M_\beta) \ge \frac{1-\gamma}{\sqrt{n}}.$$

Therefore,

$$\begin{aligned}\operatorname{dist}(0, \partial f(Q)) &= \min_{\beta \in \operatorname{conv}(\Lambda(Q))} \|(I - \gamma M_\beta)^\top r\|_2 \\ &\ge \frac{1-\gamma}{\sqrt{n}}\|r\|_2.\end{aligned}$$

Using $f(Q) = \frac{1}{2}\|r\|_2^2$ and $f^* = 0$ gives the stated bound.

□

Theorem A.2. *The Clarke subdifferential of f is given by*

$$\partial f(Q) = \{(\gamma P \Pi^\beta - I)^\top (TQ - Q) : \beta \in \operatorname{conv}(\Lambda(Q))\}$$

where $\Lambda(Q) := \{\pi \in \Theta : \pi(s) \in \arg\max_{a\in\mathcal{A}} Q(s,a),\ \forall s \in \mathcal{S}\}$ is the set of all possible greedy policies for Q.

Proof. Fix $\bar{Q}$. For each $\pi \in \Lambda(\bar{Q})$, define

$$Q_k^\pi(s,a) := \bar{Q}(s,a) + \frac{1}{k}\mathbf{1}\{a = \pi(s)\}.$$

Because $\pi(s)$ is greedy for $\bar{Q}$ at every state, the perturbation makes $\pi(s)$ the unique greedy action for Q_k^π at every state. Hence $Q_k^\pi \to \bar{Q}$, f is differentiable at every Q_k^π, and the unique greedy policy there is π.

At every differentiability point Q with unique greedy policy π, one has $TQ = T^\pi Q$ and

$$\nabla_Q f(Q) = (\gamma P \Pi^\pi - I)^\top (T^\pi Q - Q) = (\gamma P \Pi^\pi - I)^\top (TQ - Q).$$

Consequently,

$$\lim_{k\to\infty} \nabla_Q f(Q_k^\pi) = (\gamma P \Pi^\pi - I)^\top (T\bar{Q} - \bar{Q}).$$

Thus every policy in $\Lambda(\bar{Q})$ supplies a limiting gradient.

Conversely, consider any sequence of differentiability points $Q_k \to \bar{Q}$ whose gradients converge, and choose any $\pi_k \in \Lambda(Q_k)$. The perturbation constructed above shows that the smooth branch associated with π_k has a full-dimensional unique-greedy region whose closure contains Q_k. Since f is differentiable at Q_k, the gradients of all such incident smooth branches must agree with $\nabla f(Q_k)$; hence

$$\nabla f(Q_k) = (\gamma P \Pi^{\pi_k} - I)^\top (TQ_k - Q_k).$$

Since there are finitely many deterministic policies, a subsequence has one fixed policy π. Passing the greedy inequalities to the limit gives $\pi \in \Lambda(\bar{Q})$, and continuity of the displayed branch gradient gives the corresponding limiting vector above. Therefore, by the limiting-gradient characterization of the Clarke subdifferential,

$$\partial f(\bar{Q}) = \operatorname{conv}\left\{(\gamma P \Pi^\pi - I)^\top (T\bar{Q} - \bar{Q}) : \pi \in \Lambda(\bar{Q})\right\}.$$

Since the displayed vector is affine in the policy matrix, this convex hull is exactly

$$\left\{(\gamma P \Pi^\beta - I)^\top (T\bar{Q} - \bar{Q}) : \beta \in \operatorname{conv}(\Lambda(\bar{Q}))\right\}.$$

□

Theorem A.3. *The stationary point $0 \in \partial f(Q)$ in the tabular case is unique and is given by $Q = Q^*$, which is the optimal Q-function.*

Proof. Suppose that $\bar{Q}$ is a stationary point. Then, $(\gamma P \Pi^\beta - I)^\top (T\bar{Q} - \bar{Q}) = 0$ for some $\beta \in \operatorname{conv}(\Lambda(\bar{Q}))$. Since $\gamma P \Pi^\beta - I$ is invertible by Lemma A.3, we have $T\bar{Q} = \bar{Q}$. The Bellman operator T is a γ-contraction, so its fixed point is unique and therefore $\bar{Q} = Q^*$. Conversely, $TQ^* = Q^*$, and hence every vector in the subdifferential formula of Theorem A.2 is zero. Thus $0 \in \partial f(Q^*)$. Therefore, the Clarke-stationary set is exactly $\{Q^*\}$. □

Theorem A.4. *Let $(Q_k)_{k\geq 0}$ be generated by $Q_{k+1} = Q_k - \alpha_k g_k$, where $g_k = \mathrm{Proj}_{\partial f(Q_k)}(0)$ and $\alpha_k > 0$ is generated by Armijo backtracking. Assume that every nonstationary limit point $\bar{Q}$ has a unique greedy action at every state. Then $(Q_k) \subset \mathcal{L}_{f(Q_0)}$ and $Q_k \to Q^*$.*

Proof. The strict greedy-action gap at every nonstationary limit point makes the active policy locally constant, so Assumption A.2 holds. The remaining hypotheses of Theorem A.1 follow from Lemma A.25. Hence every limit point is stationary. By Theorem A.3, the only stationary point is Q^*, and compactness of the level set then implies convergence of the whole sequence to Q^*. □

A.3.1 Examples

We provide a simple MDP example for the Clarke subdifferential of the CBR. Let us consider a discounted MDP with $\mathcal{S} = \{1\}$, $\mathcal{A} = \{1, 2\}$, and $\gamma = 0.9$. The transition is deterministic and self-looping for both actions: $P(1 \mid 1, 1) = P(1 \mid 1, 2) = 1$. Moreover, rewards satisfy $R(1, 1) = 0, R(1, 2) = 1$. Let

$$Q := \begin{bmatrix} Q(1,1) \\ Q(1,2) \end{bmatrix} \in \mathbb{R}^2$$

The control Bellman operator is $(TQ)(1, a) = R(1, a) + \gamma \max_{a' \in \{1,2\}} Q(1, a'), a \in \{1, 2\}$, i.e.,

$$TQ = \begin{bmatrix} R(1,1) + \gamma \max_{a' \in \{1,2\}} Q(1,a') \\ R(1,2) + \gamma \max_{a' \in \{1,2\}} Q(1,a') \end{bmatrix}$$

and the CBR objective is $f(Q) := \frac{1}{2}\|Q - TQ\|_2^2$. The surface of $f(Q)$ is illustrated in Figure 2.

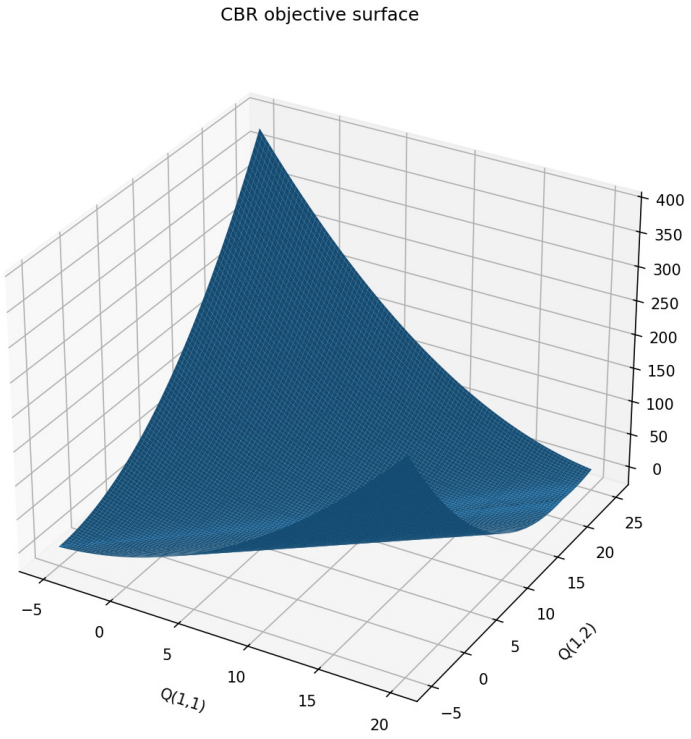


Figure 2: The surface of $f(Q)$

Nondifferentiability can occur only on the tie set $\{Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} : Q(1, 1) = Q(1, 2)\}$. A tie point is nevertheless differentiable when the gradients of all adjacent quadratic pieces coincide; in this

example, this occurs at $Q(1,1) = Q(1,2) = 1/[2(1-\gamma)] = 5$. For a (possibly stochastic) policy β, we can represent

$$\Pi^\beta = e_s^\top \otimes \beta(\cdot|1)^\top = \begin{bmatrix} \beta(1|1) & 1-\beta(1|1) \end{bmatrix}, \qquad \beta(1|1) \in [0,1],$$

where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ is the standard basis vector whose s-th element is one and whose other elements are zero. In the Clarke formula below, however, β is restricted to $\mathrm{conv}(\Lambda(Q))$; equivalently,

$$\beta(1|1) \in \begin{cases} \{1\}, & Q(1,1) > Q(1,2), \\ \{0\}, & Q(1,1) < Q(1,2), \\ [0,1], & Q(1,1) = Q(1,2). \end{cases}$$

Moreover, since both state-action pairs transition back to 1, the tabular transition matrix from state-action to next-state is

$$P = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \in \mathbb{R}^{2\times 1}.$$

Therefore, $P\Pi^\beta$ is given as

$$P\Pi^\beta = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} \beta(1|1) & 1-\beta(1|1) \end{bmatrix} = \begin{bmatrix} \beta(1|1) & 1-\beta(1|1) \\ \beta(1|1) & 1-\beta(1|1) \end{bmatrix} \in \mathbb{R}^{2\times 2},$$

which is row-stochastic for every $\beta(1|1) \in [0,1]$, and

$$\gamma P\Pi^\beta - I = \begin{bmatrix} \gamma\beta(1|1) - 1 & \gamma(1-\beta(1|1)) \\ \gamma\beta(1|1) & \gamma(1-\beta(1|1)) - 1 \end{bmatrix},$$

The Clarke subdifferential of the CBR objective is

$$\partial f(Q) = \left\{ (\gamma P\Pi^\beta - I)^\top (TQ - Q) \ : \ \beta \in \mathrm{conv}(\Lambda(Q)) \right\},$$

where

$$\begin{aligned} (\gamma P\Pi^\beta - I)^\top (TQ - Q) = & \begin{bmatrix} \gamma\beta(1|1) - 1 & \gamma(1-\beta(1|1)) \\ \gamma\beta(1|1) & \gamma(1-\beta(1|1)) - 1 \end{bmatrix}^\top \\ & \times \begin{bmatrix} R(1,1) + \gamma \max_{a' \in \{1,2\}} Q(1,a') - Q(1,1) \\ R(1,2) + \gamma \max_{a' \in \{1,2\}} Q(1,a') - Q(1,2) \end{bmatrix}. \end{aligned}$$

Case 1: $Q(1,2) > Q(1,1)$**.** The greedy policy selects 2, and $\beta(1|1) = 0, \beta(2|1) = 1$. In this case,

$$\begin{aligned} (\gamma P\Pi^\beta - I)^\top (TQ - Q) = & \begin{bmatrix} -1 & \gamma \\ 0 & \gamma - 1 \end{bmatrix}^\top \begin{bmatrix} \gamma Q(1,2) - Q(1,1) \\ 1 + \gamma Q(1,2) - Q(1,2) \end{bmatrix} \\ = & \begin{bmatrix} 1 & -\gamma \\ -\gamma & \gamma^2 + (\gamma-1)^2 \end{bmatrix} Q + \begin{bmatrix} 0 \\ \gamma - 1 \end{bmatrix} \\ =: & g_1. \end{aligned}$$

Therefore, the subdifferential is the singleton, $\partial f(Q)|_{Q(1,2)>Q(1,1)} = \{g_1\}$.

Case 2: $Q(1,1) > Q(1,2)$**.** The greedy policy selects 1, and hence, $\beta(1|1) = 1, \beta(2|1) = 0$. In this case, the subdifferential is given as

$$\begin{aligned} (\gamma P\Pi^\beta - I)^\top (TQ - Q) = & \begin{bmatrix} \gamma - 1 & 0 \\ \gamma & -1 \end{bmatrix}^\top \begin{bmatrix} \gamma Q(1,1) - Q(1,1) \\ 1 + \gamma Q(1,1) - Q(1,2) \end{bmatrix} \\ = & \begin{bmatrix} \gamma - 1 & \gamma \\ 0 & -1 \end{bmatrix} \left\{ \begin{bmatrix} \gamma - 1 & 0 \\ \gamma & -1 \end{bmatrix} Q + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\} \\ = & \begin{bmatrix} (\gamma-1)^2 + \gamma^2 & -\gamma \\ -\gamma & 1 \end{bmatrix} Q + \begin{bmatrix} \gamma \\ -1 \end{bmatrix} \\ =: & g_2. \end{aligned}$$

Therefore, the subdifferential is the singleton, $\partial f(Q)|_{Q(1,1)>Q(1,2)} = \{g_2\}$.

Case 3: $Q(1,1) = Q(1,2)$**.** Both actions are greedy, and the set of greedy policies given below contains all mixtures

$$\beta(\cdot|1) = \begin{bmatrix} \beta(1|1) \\ \beta(2|1) \end{bmatrix} = \begin{bmatrix} \beta(1|1) \\ 1-\beta(1|1) \end{bmatrix}, \qquad \beta(1|1) \in [0,1].$$

In this case, an element of the subdifferential is given as

$$\begin{aligned}(\gamma P\Pi^\beta - I)^\top (TQ - Q) &= \begin{bmatrix} \gamma\beta(1|1) - 1 & \gamma(1-\beta(1|1)) \\ \gamma\beta(1|1) & \gamma(1-\beta(1|1)) - 1 \end{bmatrix}^\top \\ &\quad \times \begin{bmatrix} \gamma Q(1,1) - Q(1,1) \\ 1 + \gamma Q(1,1) - Q(1,2) \end{bmatrix} \\ &= \beta(1|1) g_2 + [1-\beta(1|1)] g_1.\end{aligned}$$

for all $\beta(1|1) \in [0,1]$. Therefore,

$$\partial f(Q)|_{Q(1,1)=Q(1,2)} = \{\beta(1|1) g_2 + [1-\beta(1|1)] g_1 : \beta(1|1) \in [0,1]\}.$$

Next, let us consider the tie point

$$Q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Then, $\max_{a\in\{1,2\}} Q(1,a) = 0$ and thus

$$TQ = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad TQ - Q = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Using Theorem A.2, the subdifferential of the CBR objective can be written as

$$\partial f(Q) = \left\{ (\gamma P\Pi^\beta - I)^\top (TQ - Q) \ : \ \beta(1|1) \in [0,1] \right\},$$

where

$$\begin{aligned}(\gamma P\Pi^\beta - I)^\top (TQ - Q) &= \begin{bmatrix} \gamma\beta(1|1) - 1 & \gamma(1-\beta(1|1)) \\ \gamma\beta(1|1) & \gamma(1-\beta(1|1)) - 1 \end{bmatrix}^\top \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \gamma\beta(1|1) \\ \gamma - 1 - \gamma\beta(1|1) \end{bmatrix}.\end{aligned}$$

Hence, the subdifferential at $Q = (0,0)$ is the following line segment:

$$\begin{aligned}\partial f(0) &= \left\{ \begin{bmatrix} \gamma\beta(1|1) \\ \gamma - 1 - \gamma\beta(1|1) \end{bmatrix} : \ \beta(1|1) \in [0,1] \right\} \\ &= \operatorname{conv}\left\{ \begin{bmatrix} 0 \\ -0.1 \end{bmatrix}, \begin{bmatrix} 0.9 \\ -1 \end{bmatrix} \right\}.\end{aligned}$$

In particular, $0 \notin \partial f(0)$, and hence, 0 is not a stationary point.

For any $\beta \in [0,1]$, $I - \gamma P\Pi^\beta$ is invertible by Lemma A.3. Therefore, $0 \in \partial f(\bar{Q})$ for some $\bar{Q}$ implies that there exists a $\beta(1|1) \in [0,1]$ such that $(\gamma P\Pi^\beta - I)^\top (T\bar{Q} - \bar{Q}) = 0$. This implies that $T\bar{Q} = \bar{Q}$. Therefore, any stationary point must satisfy the Bellman optimality equation $TQ = Q$.

Solving $TQ = Q$, which is equivalent to

$$Q(1,1) = \gamma \max\{Q(1,1), Q(1,2)\}, \qquad Q(1,2) = 1 + \gamma \max\{Q(1,1), Q(1,2)\},$$

yields the unique fixed point:

$$Q^* = \begin{bmatrix} \frac{\gamma}{1-\gamma} \\ \frac{1}{1-\gamma} \end{bmatrix} = \begin{bmatrix} 9 \\ 10 \end{bmatrix}.$$

We can apply a subgradient descent method to the above objective, and the results are shown in Figure 3 and Figure 4. Under the active-set condition of Theorem A.4, convergence is guaranteed, and the plots confirm that the iterates converge to Q^*.

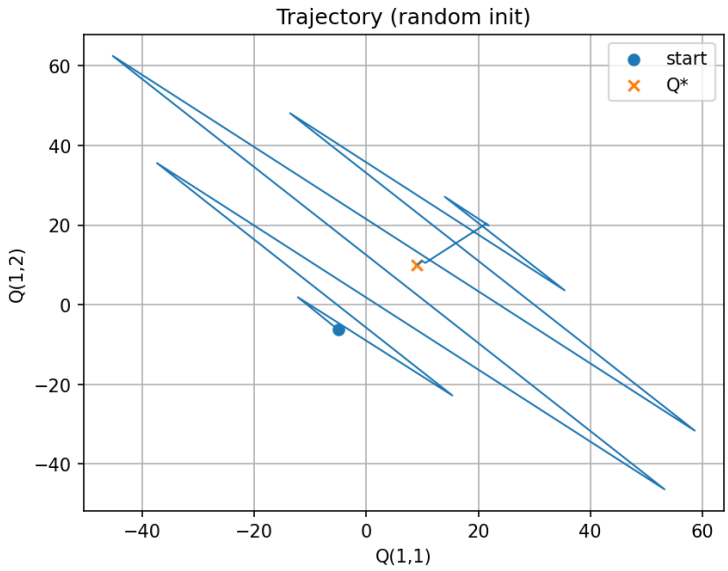


Figure 3: Convergence trajectory of Q_k with the subgradient descent method.

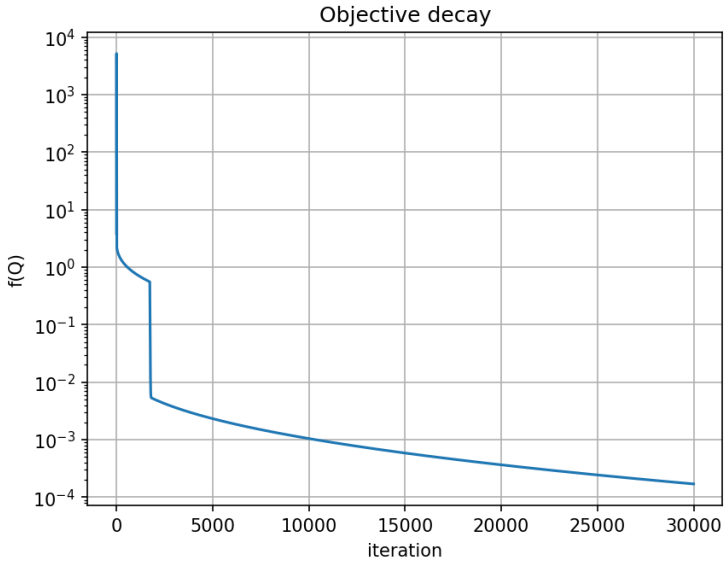


Figure 4: CBR objective function value $f(Q_k)$ with Q_k generated by the subgradient descent method.

A.4 CBR with linear function approximation

Let Θ denote the set of deterministic policies $\pi : \mathcal{S} \to \mathcal{A}$, so $|\Theta| = |\mathcal{A}|^{|\mathcal{S}|} < \infty$. Let us consider the CBR objective with LFA

$$f(\theta) := \frac{1}{2} \left\| T\Phi\theta - \Phi\theta \right\|_2^2 ,$$

Moreover, let us define the set

$$S_\pi := \left\{ \theta \in \mathbb{R}^m : \pi(s) \in \arg\max_{a \in \mathcal{A}} Q_\theta(s,a), \forall s \in \mathcal{S} \right\}$$

for each $\pi \in \Theta$, where $\arg\max$ is interpreted as a set-valued map, i.e., $\pi(s)$ is one of the maximizers; ties are allowed.

Proposition A.2. *f is piecewise quadratic (so piecewise smooth) and continuous.*

Proof. For each deterministic policy $\pi \in \Theta$, define

$$f_\pi(\theta) := \frac{1}{2} \|R + (\gamma P \Pi^\pi - I)\Phi\theta\|_2^2 .$$

If $\theta \in S_\pi$, then $\pi(s)$ is a maximizing action for Q_θ at every state, and therefore $TQ_\theta = T^\pi Q_\theta = R + \gamma P \Pi^\pi \Phi\theta$. Hence $f(\theta) = f_\pi(\theta)$ on S_π. The finite family of polyhedra $\{S_\pi\}_{\pi\in\Theta}$ covers $\mathbb{R}^m$, and the arrangement of their defining action-comparison hyperplanes yields a finite polyhedral partition refining this cover. On every cell, f agrees with one of the quadratic polynomials f_π. This argument does not require a selector to be constant when ties occur. If $\theta \in S_\pi \cap S_{\tilde\pi}$, then both policies select statewise maximizers of Q_θ, so $\Pi^\pi\Phi\theta = \Pi^{\tilde\pi}\Phi\theta$, $T^\pi Q_\theta = T^{\tilde\pi} Q_\theta = TQ_\theta$, and $f_\pi(\theta) = f_{\tilde\pi}(\theta)$. Thus the quadratic descriptions agree on overlaps, and f is finite piecewise quadratic in the sense of Definition A.6. Finally, the map $\theta \mapsto T(\Phi\theta)$ is continuous because it is formed from finitely many maxima of linear functions; hence f is continuous. □

Proposition A.3. *For each $\pi \in \Theta$, the set S_π is an intersection of half-spaces and is a convex cone.*

Proof. For each $s \in \mathcal{S}$, the condition $\pi(s) \in \arg\max_{a\in\mathcal{A}} Q_\theta(s,a)$ is equivalent to the family of linear inequalities $Q_\theta(s,\pi(s)) \geq Q_\theta(s,a), \forall (s,a) \in \mathcal{S}\times\mathcal{A}$. Hence, S_π can be written as

$$S_\pi = \bigcap_{s\in\mathcal{S}} \bigcap_{a\in\mathcal{A}} \{\theta \in \mathbb{R}^m :\ Q_\theta(s,\pi(s)) - Q_\theta(s,a) \geq 0\} = \{\theta \in \mathbb{R}^m :\ L\Pi^\pi\Phi\theta - \Phi\theta \geq 0\},$$

where $L \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|}$ is a lifting matrix defined as $L := I_{|\mathcal{S}|} \otimes \mathbf{1}_{|\mathcal{A}|}$, where $I_{|\mathcal{S}|}$ is the identity matrix of dimension $|\mathcal{S}|$, $\otimes$ denotes the Kronecker product, $\mathbf{1}_{|\mathcal{A}|}$ is the all-ones vector of dimension $|\mathcal{A}|$. Each set in the intersection is a closed half-space containing the origin. Therefore, it is a polyhedral set.

To prove the second statement, let us fix any $\pi \in \Theta$ and consider any $\theta \in S_\pi$. For every $\alpha > 0$, $Q_{\alpha\theta} = \alpha Q_\theta$ preserves the maximizing actions, and hence $\alpha\theta \in S_\pi$. For $\alpha = 0$, $Q_0 = 0$ ties all actions, so $0 \in S_\pi$. Therefore, S_π is a cone. Moreover, let $\theta_1, \theta_2 \in S_\pi$ and $\alpha \in [0,1]$. For all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$Q_{\theta_1}(s,\pi(s)) \geq Q_{\theta_1}(s,a), \qquad Q_{\theta_2}(s,\pi(s)) \geq Q_{\theta_2}(s,a).$$

Multiplying the first inequality by $\alpha \in [0,1]$ and the second by $1-\alpha$, and adding, yields

$$\alpha Q_{\theta_1}(s,\pi(s)) + (1-\alpha) Q_{\theta_2}(s,\pi(s)) \geq \alpha Q_{\theta_1}(s,a) + (1-\alpha) Q_{\theta_2}(s,a),$$

Because $Q_{\alpha\theta_1+(1-\alpha)\theta_2} = \alpha Q_{\theta_1} + (1-\alpha)Q_{\theta_2}$ by linearity of the approximator, this shows that $\alpha\theta_1 + (1-\alpha)\theta_2 \in S_\pi$. Hence S_π is convex. Combining the two parts, we conclude that S_π is a convex cone. □

Proposition A.4. *f is bounded by strongly convex quadratic functions as $q_1(\theta) \leq f(\theta) \leq q_2(\theta)$, where*

$$q_1(\theta) := \frac{(1-\gamma)^2}{2|\mathcal{S}\times\mathcal{A}|} \|Q_\theta - Q^*\|_2^2, \quad q_2(\theta) := \frac{(1+\gamma)^2|\mathcal{S}\times\mathcal{A}|}{2} \|Q_\theta - Q^*\|_2^2 .$$

Proof. Using Lemma A.7, we have

$$\begin{aligned}\frac{1}{\sqrt{|\mathcal{S}\times\mathcal{A}|}}\|Q_\theta - Q^*\|_2 \leq \|Q_\theta - Q^*\|_\infty &\leq \frac{1}{1-\gamma}\|TQ_\theta - Q_\theta\|_\infty \\ &\leq \frac{1}{1-\gamma}\|TQ_\theta - Q_\theta\|_2.\end{aligned}$$

Algebraic manipulations lead to $\frac{(1-\gamma)^2}{2|\mathcal{S}\times\mathcal{A}|}\|Q_\theta - Q^*\|_2^2 \leq f(\theta)$. Similarly, using Lemma A.8, one gets

$$\begin{aligned}\|TQ_\theta - Q_\theta\|_2 &\leq \sqrt{|\mathcal{S}\times\mathcal{A}|}\|TQ_\theta - Q_\theta\|_\infty \\ &\leq (1+\gamma)\sqrt{|\mathcal{S}\times\mathcal{A}|}\|Q_\theta - Q^*\|_\infty \\ &\leq (1+\gamma)\sqrt{|\mathcal{S}\times\mathcal{A}|}\|Q_\theta - Q^*\|_2,\end{aligned}$$

which leads to $f(\theta) \leq \frac{(1+\gamma)^2|\mathcal{S}\times\mathcal{A}|}{2}\|Q_\theta - Q^*\|_2^2$. This completes the proof. □

Proposition A.5. *Let us define*

$$\theta_1^* \in \arg\min_{\theta\in\mathbb{R}^m} \|Q_\theta - Q^*\|_2, \quad \theta_2^* \in \arg\min_{\theta\in\mathbb{R}^m} f(\theta).$$

Let $e_\Phi := \left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2$ *and* $n := |\mathcal{S}\times\mathcal{A}|$*. Then, we have*

$$\begin{aligned}
\frac{(1-\gamma)^2}{2n}e_\Phi^2 \le f(\theta_1^*) &\le \frac{(1+\gamma)^2 n}{2}e_\Phi^2,\\
\frac{(1-\gamma)^2}{2n}e_\Phi^2 \le f(\theta_2^*) &\le \frac{(1+\gamma)^2 n}{2}e_\Phi^2,\\
0 \le f(\theta_1^*) - f(\theta_2^*) &\le \left\{\frac{(1+\gamma)^2 n}{2} - \frac{(1-\gamma)^2}{2n}\right\}e_\Phi^2,\\
\left\|Q_{\theta_1^*} - Q_{\theta_2^*}\right\|_2 &\le \left\{1 + \frac{(1+\gamma)n}{1-\gamma}\right\}e_\Phi.
\end{aligned}$$

Proof. By plugging θ_1^* and θ_2^* into the quadratic bounds in Proposition A.4, one has

$$\begin{aligned}
\frac{(1-\gamma)^2}{2n}e_\Phi^2 = q_1(\theta_1^*) \le f(\theta_1^*) &\le q_2(\theta_1^*) = \frac{(1+\gamma)^2 n}{2}e_\Phi^2,\\
\frac{(1-\gamma)^2}{2n}\left\|Q_{\theta_2^*} - Q^*\right\|_2^2 = q_1(\theta_2^*) \le f(\theta_2^*) &\le q_2(\theta_2^*)\\
&= \frac{(1+\gamma)^2 n}{2}\left\|Q_{\theta_2^*} - Q^*\right\|_2^2,
\end{aligned}$$

respectively. The first inequality above proves the first statement. Moreover, since $f(\theta_2^*) \le f(\theta_1^*)$ and $\left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2^2 = \left\|Q_{\theta_1^*} - Q^*\right\|_2^2 \le \left\|Q_{\theta_2^*} - Q^*\right\|_2^2$, it follows that

$$\begin{aligned}
\frac{(1-\gamma)^2}{2|\mathcal{S}\times\mathcal{A}|}\left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2^2 &\le \frac{(1-\gamma)^2}{2|\mathcal{S}\times\mathcal{A}|}\left\|Q_{\theta_2^*} - Q^*\right\|_2^2\\
&\le f(\theta_2^*)\\
&\le f(\theta_1^*)\\
&\le \frac{(1+\gamma)^2|\mathcal{S}\times\mathcal{A}|}{2}\left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2^2, \qquad (14)
\end{aligned}$$

which leads to

$$\frac{(1-\gamma)^2}{2|\mathcal{S}\times\mathcal{A}|}\left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2^2 \le f(\theta_2^*) \le \frac{(1+\gamma)^2|\mathcal{S}\times\mathcal{A}|}{2}\left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2^2,$$

and it proves the second statement. From (14) again, we obtain the following two inequalities:

$$\begin{aligned}
\frac{(1-\gamma)^2}{2|\mathcal{S}\times\mathcal{A}|}\left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2^2 &\le f(\theta_2^*) \le f(\theta_1^*),\\
f(\theta_1^*) - \frac{(1+\gamma)^2|\mathcal{S}\times\mathcal{A}|}{2}\left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2^2 &\le 0.
\end{aligned}$$

Adding the two inequalities yields the third statement. For the last statement, we derive

$$\begin{aligned}
\left\|Q_{\theta_1^*} - Q_{\theta_2^*}\right\|_2 &= \left\|Q_{\theta_1^*} - Q^* + Q^* - Q_{\theta_2^*}\right\|_2\\
&\le \left\|Q_{\theta_1^*} - Q^*\right\|_2 + \left\|Q_{\theta_2^*} - Q^*\right\|_2\\
&\le \left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2 + \left\|Q_{\theta_2^*} - Q^*\right\|_2. \qquad (15)
\end{aligned}$$

To bound the second term in (15), (14) is used to get

$$\left\|Q_{\theta_2^*} - Q^*\right\|_2 \le \frac{(1+\gamma)|\mathcal{S}\times\mathcal{A}|}{1-\gamma}\left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2. \qquad (16)$$

Now, combining (15) and (16) leads to the last statement. This completes the proof. □

For the LFA case, define the set of policies that are active from full-dimensional differentiability regions by

$$\Lambda_\Phi(\theta) := \left\{\pi \in \Theta : \theta \in \overline{\mathrm{int}(S_\pi)}\right\}.$$

Lemma A.26 (Persistent ties under LFA)**.** *Let $U \subset \mathbb{R}^m$ be a nonempty open set. If $Q_\theta(s,a) = Q_\theta(s,b)$ for every $\theta \in U$, then $\phi(s,a) = \phi(s,b)$. Consequently, if $U \subset S_\pi \cap S_{\widetilde{\pi}}$ for two deterministic policies π and $\widetilde{\pi}$, then $\Pi^\pi \Phi = \Pi^{\widetilde{\pi}} \Phi$, and the corresponding quadratic branch functions and their gradients coincide.*

Proof. The equality on U is

$$(\phi(s,a) - \phi(s,b))\theta = 0, \qquad \theta \in U.$$

A linear function that vanishes on a nonempty open set is identically zero, so $\phi(s,a) = \phi(s,b)$. If $U \subset S_\pi \cap S_{\widetilde{\pi}}$, then at each state the actions selected by π and $\widetilde{\pi}$ are both maximizers throughout U and hence are persistently tied whenever they differ. Applying the first conclusion statewise gives $\Pi^\pi \Phi = \Pi^{\widetilde{\pi}} \Phi$. The branch functions $\frac{1}{2}\|R + (\gamma P \Pi^\pi - I)\Phi\theta\|_2^2$ and $\frac{1}{2}\|R + (\gamma P \Pi^{\widetilde{\pi}} - I)\Phi\theta\|_2^2$ are therefore identical, as are their gradients. □

Theorem A.5. *The subdifferential of f is given by*

$$\partial f(\theta) = \{\Phi^\top (\gamma P \Pi^\beta - I)^\top (TQ_\theta - Q_\theta) : \beta \in \mathrm{conv}(\Lambda_\Phi(\theta))\},$$

where only the greedy policies that are reachable from differentiability regions in the parameter space are included.

Proof. Fix $\bar{\theta} \in \mathbb{R}^m$. For a deterministic policy π, let

$$S_\pi := \left\{ \theta \in \mathbb{R}^m : \pi(s) \in \arg\max_{a \in \mathcal{A}} Q_\theta(s,a),\ \forall s \in \mathcal{S} \right\}.$$

This is a closed polyhedral set. On S_π one has $TQ_\theta = T^\pi Q_\theta$, and hence the corresponding quadratic branch has gradient

$$g_\pi(\theta) := \Phi^\top (\gamma P \Pi^\pi - I)^\top (TQ_\theta - Q_\theta).$$

Let $\pi \in \Lambda_\Phi(\bar{\theta})$. There exists a sequence $\theta_k \in \mathrm{int}(S_\pi)$ with $\theta_k \to \bar{\theta}$. Since a neighborhood of each θ_k is contained in S_π, the function f agrees there with the quadratic branch associated with π and is differentiable at θ_k. If another policy is active throughout such a neighborhood because of a persistent tie, Lemma A.26 shows that its branch and gradient coincide with those of π. Therefore,

$$\lim_{k \to \infty} \nabla f(\theta_k) = \lim_{k \to \infty} g_\pi(\theta_k) = g_\pi(\bar{\theta}).$$

Thus every policy in $\Lambda_\Phi(\bar{\theta})$ contributes the displayed limiting gradient.

Conversely, consider any sequence of differentiability points converging to $\bar{\theta}$ whose gradients have a limit. By the finite polyhedral piecewise-quadratic representation in Proposition A.2, at each such point its gradient equals the gradient of a full-dimensional adjacent quadratic branch. There are finitely many deterministic policies, so after passing to a subsequence one fixed policy π supplies these branch gradients. The corresponding full-dimensional regions approach $\bar{\theta}$, which gives $\bar{\theta} \in \overline{\mathrm{int}(S_\pi)}$ and hence $\pi \in \Lambda_\Phi(\bar{\theta})$. Differentiable boundary points add no further limiting gradients, because differentiability forces the gradients of all adjacent branches to coincide.

By Lemma A.14 and Lemma A.20, the Clarke subdifferential is therefore

$$\mathrm{conv}\left\{ g_\pi(\bar{\theta}) : \pi \in \Lambda_\Phi(\bar{\theta}) \right\}.$$

Since the displayed gradient is affine in Π^π, this convex hull is exactly the set obtained by allowing $\beta \in \mathrm{conv}(\Lambda_\Phi(\bar{\theta}))$. This proves the result. □

Proposition A.6 (Convex QP for the minimum-norm Clarke subgradient)**.** *Fix θ and enumerate the reachable active policies as $\Lambda_\Phi(\theta) = \{\pi_1, \ldots, \pi_L\}$. Define*

$$r_\theta := TQ_\theta - Q_\theta, \qquad g_i := \Phi^\top (\gamma P \Pi^{\pi_i} - I)^\top r_\theta, \qquad G_\theta := [g_1 \quad \cdots \quad g_L].$$

Then the minimum-norm Clarke subgradient is $g^* = G_\theta w^*$, where

$$w^* \in \arg\min_{w \geq 0,\ \mathbf{1}^\top w = 1} \frac{1}{2} w^\top G_\theta^\top G_\theta w.$$

The vector g^* is unique, although the minimizing weight vector need not be. In the tabular case, the same vector is obtained from the explicit product-simplex QP

$$\beta^* \in \arg\min_{\beta} \frac{1}{2} \left\| (\gamma P \Pi^{\beta} - I)^{\top} (TQ - Q) \right\|_2^2,$$

$$\text{subject to } \beta(\cdot|s) \in \Delta_{|\mathcal{A}|}, \qquad s \in \mathcal{S},$$
$$\beta(a|s) = 0 \quad \text{whenever } a \notin \arg\max_{u \in \mathcal{A}} Q(s,u).$$

At any minimizer β^*, $g^* = (\gamma P \Pi^{\beta^*} - I)^{\top}(TQ - Q)$. Under LFA, only policies in $\Lambda_{\Phi}(\theta)$ may be used, because pointwise greedy policies that are not reachable from full-dimensional parameter regions do not contribute to the Clarke subdifferential.

Proof. By Theorem A.5, $\partial f(\theta) = \operatorname{conv}\{g_1, \ldots, g_L\} = \{G_\theta w : w \geq 0, \mathbf{1}^{\top} w = 1\}$. Projecting the origin onto this compact convex set is therefore exactly the first displayed positive-semidefinite quadratic program. In the tabular case, $\operatorname{conv}(\Lambda(Q))$ is precisely the product of the statewise simplexes supported on tied greedy actions, and the displayed subgradient is affine in β, which gives the second QP. Uniqueness of g^* follows from uniqueness of Euclidean projection onto a closed convex set. □

Theorem A.6. *A point $\bar{\theta}$ is Clarke-stationary if and only if there exists $\bar{\beta} \in \operatorname{conv}(\Lambda_{\Phi}(\bar{\theta}))$ such that*

$$Q_{\bar{\theta}} \in \mathcal{G}_{\Phi|\Psi_{\bar{\beta}}}(TQ_{\bar{\theta}}), \qquad \Psi_{\bar{\beta}} := (\gamma P \Pi^{\bar{\beta}} - I)\Phi.$$

If $\Psi_{\bar{\beta}}^{\top}\Phi$ is nonsingular, then $\mathcal{G}_{\Phi|\Psi_{\bar{\beta}}}(TQ_{\bar{\theta}})$ is the singleton $\{\Gamma_{\Phi|\Psi_{\bar{\beta}}} TQ_{\bar{\theta}}\}$, and the condition reduces to the usual OP-CBE.

Proof. By Theorem A.5, Clarke stationarity is equivalent to the existence of $\bar{\beta} \in \operatorname{conv}(\Lambda_{\Phi}(\bar{\theta}))$ such that

$$\Psi_{\bar{\beta}}^{\top}(TQ_{\bar{\theta}} - Q_{\bar{\theta}}) = 0.$$

Since $Q_{\bar{\theta}} = \Phi\bar{\theta}$, this is exactly $Q_{\bar{\theta}} \in \mathcal{G}_{\Phi|\Psi_{\bar{\beta}}}(TQ_{\bar{\theta}})$. If $\Psi_{\bar{\beta}}^{\top}\Phi$ is nonsingular, the defining linear equation has the unique solution

$$\bar{\theta} = (\Psi_{\bar{\beta}}^{\top}\Phi)^{-1}\Psi_{\bar{\beta}}^{\top} TQ_{\bar{\theta}},$$

which gives the single-valued oblique-projector formula. The reverse implication follows by applying $\Psi_{\bar{\beta}}^{\top}$ to the projected equation and then using Theorem A.5. □

Many properties of the CBR objective function continue to hold even when linear function approximation is employed. The following summarizes these properties.

Lemma A.27. *1. f is bounded below, locally Lipschitz continuous, and is coercive.*

2. *The Clarke subdifferential of f given by $\partial f(\theta) = \{\Phi^{\top}(\gamma P \Pi^{\beta} - I)^{\top}(TQ_\theta - Q_\theta) : \beta \in \operatorname{conv}(\Lambda_{\Phi}(\theta))\}$ is nonempty, convex, and bounded.*
3. *We have $\|TQ_\theta\|_\infty \leq R_{\max} + \gamma\|Q_\theta\|_\infty$, and $\|TQ_\theta - Q_\theta\|_\infty \geq (1-\gamma)\|Q_\theta\|_\infty - R_{\max}$.*
4. *Every sublevel set $\mathcal{L}_c := \{\theta \in \mathbb{R}^m : f(\theta) \leq c\}$ is bounded for any $c \geq 0$.*
5. *$\theta \in \mathcal{L}_c$ implies*

$$\|Q_\theta\|_\infty \leq \frac{R_{\max} + \sqrt{2c}}{1-\gamma}.$$

Proof. The lower boundedness and the estimate in item 3 follow exactly as in the tabular case, with Q replaced by $Q_\theta = \Phi\theta$. Moreover, by Proposition A.2, f is piecewise quadratic with finitely many pieces, and hence it is locally Lipschitz by Lemma A.20. If $\theta \in \mathcal{L}_c$, then item 3 yields

$$\|Q_\theta\|_\infty \leq \frac{R_{\max} + \sqrt{2c}}{1-\gamma}.$$

Therefore,

$$\|Q_\theta\|_2 \leq \sqrt{|\mathcal{S} \times \mathcal{A}|}\, \|Q_\theta\|_\infty \leq \sqrt{|\mathcal{S} \times \mathcal{A}|}\, \frac{R_{\max} + \sqrt{2c}}{1-\gamma}.$$

Since Φ has full column rank,

$$\sigma_{\min}(\Phi)\,\|\theta\|_2 \le \|\Phi\theta\|_2 = \|Q_\theta\|_2,$$

and thus

$$\|\theta\|_2 \le \frac{\sqrt{|\mathcal{S}\times\mathcal{A}|}}{\sigma_{\min}(\Phi)}\,\frac{R_{\max}+\sqrt{2c}}{1-\gamma}.$$

This proves item 4. Item 5 is exactly the displayed bound above. Finally, the same bound in item 3 together with the full column rank of Φ implies coercivity: if $\|\theta\|_2 \to \infty$, then $\|Q_\theta\|_2 = \|\Phi\theta\|_2 \to \infty$, hence $\|Q_\theta\|_\infty \to \infty$, and therefore $f(\theta) \to \infty$. The subdifferential claim in item 2 follows from local Lipschitz continuity and Lemma A.10. □

The PL-type inequality that holds in the tabular case does not, in general, hold when linear function approximation is used.

Lemma A.28. *A stationary point of f always exists.*

Proof. Since f is continuous and coercive by Lemma A.27, it attains a global minimum. Since the subdifferential of f always exists at any $\theta \in \mathbb{R}^m$ by Lemma A.27, at the minimizer θ, we have $0 \in \partial f(\theta)$ by Lemma A.15, i.e., θ is a stationary point. This completes the proof. □

Lemma A.29. *For any $\theta \in \mathbb{R}^m$, we have*

$$\begin{aligned}\|Q_\theta - Q^*\|_\infty \le& \frac{1}{1-\gamma}\|TQ_\theta - Q_\theta\|_2,\\ \|Q_\theta - Q^*\|_\infty^2 \le& 2\left(\frac{1}{1-\gamma}\right)^2 f(\theta).\end{aligned}$$

Moreover, we have

$$\begin{aligned}\|Q^{\pi_\theta} - Q^*\|_\infty \le& \frac{2\gamma}{1-\gamma}\|TQ_\theta - Q_\theta\|_2,\\ \|Q^{\pi_\theta} - Q^*\|_\infty^2 \le& \frac{8\gamma^2}{(1-\gamma)^2} f(\theta).\end{aligned}$$

where π_θ is any deterministic greedy policy satisfying $\pi_\theta(s) \in \arg\max_{a\in\mathcal{A}} Q_\theta(s,a)$ for every $s \in \mathcal{S}$.

Proof. The first statement can be directly obtained from Lemma A.7. The second statement can be obtained via simple calculations from the first statement. For the third statement, using $TQ_\theta = T^{\pi_\theta}Q_\theta$, the triangle inequality, and the contraction of T and T^{π_θ}, we obtain

$$\begin{aligned}\|Q^{\pi_\theta} - Q^*\|_\infty =& \|T^{\pi_\theta}Q^{\pi_\theta} - TQ^*\|_\infty\\ \le& \|T^{\pi_\theta}Q^{\pi_\theta} - T^{\pi_\theta}Q_\theta\|_\infty + \|TQ_\theta - TQ^*\|_\infty\\ \le& \gamma\|Q^{\pi_\theta} - Q_\theta\|_\infty + \gamma\|Q_\theta - Q^*\|_\infty.\end{aligned}$$

By Lemma A.5 applied to π_θ and by Lemma A.7,

$$\begin{aligned}\|Q^{\pi_\theta} - Q_\theta\|_\infty \le& \frac{1}{1-\gamma}\|T^{\pi_\theta}Q_\theta - Q_\theta\|_\infty = \frac{1}{1-\gamma}\|TQ_\theta - Q_\theta\|_\infty,\\ \|Q_\theta - Q^*\|_\infty \le& \frac{1}{1-\gamma}\|TQ_\theta - Q_\theta\|_\infty.\end{aligned}$$

Combining the last three inequalities yields

$$\|Q^{\pi_\theta} - Q^*\|_\infty \le \frac{2\gamma}{1-\gamma}\|TQ_\theta - Q_\theta\|_\infty \le \frac{2\gamma}{1-\gamma}\|TQ_\theta - Q_\theta\|_2.$$

The last statement follows by squaring this bound and using $\|TQ_\theta - Q_\theta\|_2^2 = 2f(\theta)$. This completes the proof. □

Proposition A.7. *Suppose that $\theta_2^* \in \arg\min_{\theta\in\mathbb{R}^m} f(\theta)$. Then, we have*

$$\left\|Q_{\theta_2^*} - Q^*\right\|_\infty \le \frac{(1+\gamma)\sqrt{|\mathcal{S}\times\mathcal{A}|}}{1-\gamma}\left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2.$$

Proof. From the first statement of Lemma A.29, we have

$$\left\|Q_{\theta_2^*} - Q^*\right\|_\infty \le \frac{1}{1-\gamma}\left\|TQ_{\theta_2^*} - Q_{\theta_2^*}\right\|_2.$$

Moreover, Proposition A.5 leads to

$$\left\|TQ_{\theta_2^*} - Q_{\theta_2^*}\right\|_2 \le (1+\gamma)\sqrt{|\mathcal{S}\times\mathcal{A}|}\left\|\Gamma_{\Phi|\Phi}Q^* - Q^*\right\|_2.$$

Combining the two inequalities concludes the proof. □

Proposition A.8. *Suppose that $\bar{\theta}$ is a stationary point satisfying Assumption 1. Let $\bar{\pi} := \pi_{\bar{\theta}}$ and let $\bar{\beta} \in \mathrm{conv}(\Lambda_\Phi(\bar{\theta}))$ satisfy the stationary condition. Define*

$$A_{\bar{\pi}} := \gamma P\Pi^{\bar{\pi}} - I, \qquad A_{\bar{\beta}} := \gamma P\Pi^{\bar{\beta}} - I, \qquad \Omega_{\bar{\beta},\bar{\pi}} := A_{\bar{\pi}}^\top A_{\bar{\beta}}\Phi.$$

Assume that $\Omega_{\bar{\beta},\bar{\pi}}^\top\Phi$ is nonsingular. Then,

$$\left\|Q^{\bar{\pi}} - Q^*\right\|_\infty \le \min\left\{\frac{2\gamma}{1-\gamma}\left\|I - \Gamma_{\Phi|\Omega_{\bar{\beta},\bar{\pi}}}\right\|_2\left\|Q^{\bar{\pi}} - \Gamma_{\Phi|\Phi}Q^{\bar{\pi}}\right\|_2, \frac{\|T^{\bar{\pi}}Q^{\bar{\pi}} - TQ^{\bar{\pi}}\|_\infty}{1-\gamma}\right\}.$$

Proof. For any stochastic policy $\pi : \mathcal{S} \to \Delta_{|\mathcal{A}|}$, using the triangle inequality, we obtain

$$\|Q^\pi - Q^*\|_\infty \le \frac{1}{1-\gamma}\|T^\pi Q^\pi - TQ^\pi\|_\infty. \tag{17}$$

For $\bar{\pi} = \pi_{\bar{\theta}}$, using $TQ_{\bar{\theta}} = T^{\bar{\pi}}Q_{\bar{\theta}}$, $TQ^* = Q^*$, and the contraction of T and $T^{\bar{\pi}}$, we have

$$\begin{aligned}\left\|Q^{\bar{\pi}} - Q^*\right\|_\infty &= \left\|T^{\bar{\pi}}Q^{\bar{\pi}} - TQ^*\right\|_\infty \\ &\le \gamma\left\|Q^{\bar{\pi}} - Q_{\bar{\theta}}\right\|_\infty + \gamma\|Q_{\bar{\theta}} - Q^*\|_\infty \\ &\le \gamma\left\|Q^{\bar{\pi}} - Q_{\bar{\theta}}\right\|_\infty + \gamma\Big(\left\|Q_{\bar{\theta}} - Q^{\bar{\pi}}\right\|_\infty + \left\|Q^{\bar{\pi}} - Q^*\right\|_\infty\Big),\end{aligned} \tag{18}$$

which implies

$$\left\|Q^{\bar{\pi}} - Q^*\right\|_\infty \le \frac{2\gamma}{1-\gamma}\left\|Q^{\bar{\pi}} - Q_{\bar{\theta}}\right\|_\infty. \tag{19}$$

Since $\bar{\theta}$ is stationary, $Q_{\bar{\theta}}$ satisfies

$$\Phi^\top A_{\bar{\beta}}^\top(T^{\bar{\pi}}Q_{\bar{\theta}} - Q_{\bar{\theta}}) = 0.$$

Equivalently, $Q_{\bar{\theta}} = \Gamma_{\Phi|\Omega_{\bar{\beta},\bar{\pi}}}Q^{\bar{\pi}}$, and hence

$$\left\|Q^{\bar{\pi}} - Q_{\bar{\theta}}\right\|_2 \le \left\|I - \Gamma_{\Phi|\Omega_{\bar{\beta},\bar{\pi}}}\right\|_2\left\|Q^{\bar{\pi}} - \Gamma_{\Phi|\Phi}Q^{\bar{\pi}}\right\|_2.$$

Combining this inequality with (19) and also using (17) with $\pi = \bar{\pi}$ gives the desired bound. □

Theorem A.7. *Let $\pi^*(s)$ be the unique element of $\arg\max_{a\in\mathcal{A}} Q^*(s,a)$ for each $s \in \mathcal{S}$, and define*

$$\Delta := \min_{s\in\mathcal{S}}\min_{a\neq\pi^*(s)}\{Q^*(s,\pi^*(s)) - Q^*(s,a)\} > 0.$$

Assume $0 < \varepsilon < \Delta/2$. If $\mathcal{L}_{c_0}$ is nonempty with $c_0 = \frac{(1-\gamma)^2}{2}\varepsilon^2$, then the convex set

$$K := \{\theta \in \mathbb{R}^m : \|Q_\theta - Q^*\|_\infty \le \varepsilon\}$$

is nonempty, $K \subset S_{\pi^*}$, f is strongly convex on K, and the global minimizer of f is unique and lies in K. This minimizer is the unique solution of

$$\{\theta^*\} = \arg\min_{\theta\in\mathbb{R}^m}\frac{1}{2}\left\|T^{\pi^*}Q_\theta - Q_\theta\right\|_2^2.$$

Moreover, $\pi_{\theta^*} = \pi^*$, so $\|Q^{\pi_{\theta^*}} - Q^*\|_\infty = 0$. Defining $\Omega_{\pi^*} := (\gamma P\Pi^{\pi^*} - I)^\top(\gamma P\Pi^{\pi^*} - I)\Phi$, whose normal-equation matrix is automatically nonsingular by Lemma A.4, we also have

$$\|Q_{\theta^*} - Q^*\|_2 \le \left\|I - \Gamma_{\Phi|\Omega_{\pi^*}}\right\|_2\left\|Q^* - \Gamma_{\Phi|\Phi}Q^*\right\|_2.$$

Proof. The action-gap condition and $0 < \varepsilon < \Delta/2$ imply that every Q satisfying $\|Q - Q^*\|_\infty \le \varepsilon$ has the same greedy policy π^* as Q^*. Hence $K \subset S_{\pi^*}$. Moreover, K is convex because it is the

inverse image, under the affine map $\theta \mapsto \Phi\theta - Q^*$, of the convex ℓ_∞-ball of radius ε. By Lemma A.7,

$$(1-\gamma)\|Q_\theta - Q^*\|_\infty \le \|TQ_\theta - Q_\theta\|_\infty \le \|TQ_\theta - Q_\theta\|_2,$$

so $\mathcal{L}_{c_0} \subseteq K$. If $\mathcal{L}_{c_0}$ is nonempty, then K is nonempty. Moreover, any global minimizer θ_g of f satisfies $f(\theta_g) \le c_0$ and therefore lies in K. On K, the control Bellman operator agrees with T^{π^*}, so f coincides with the strongly convex quadratic

$$f^{\pi^*}(\theta) := \frac{1}{2}\|T^{\pi^*}Q_\theta - Q_\theta\|_2^2.$$

Since $\gamma P\Pi^{\pi^*} - I$ is nonsingular and Φ has full column rank, the Hessian of f^{π^*} is positive definite. Therefore, f^{π^*} has a unique global minimizer on $\mathbb{R}^m$; denote it by θ_{π^*}. Because $\mathcal{L}_{c_0}$ is nonempty, choose $\theta_0 \in \mathcal{L}_{c_0} \subseteq K$. Then $f^{\pi^*}(\theta_0) = f(\theta_0) \le c_0$, and by optimality of θ_{π^*},

$$f^{\pi^*}(\theta_{\pi^*}) \le f^{\pi^*}(\theta_0) \le c_0.$$

Applying Lemma A.5 to the policy π^* yields

$$(1-\gamma)\|Q_{\theta_{\pi^*}} - Q^*\|_\infty \le \|T^{\pi^*}Q_{\theta_{\pi^*}} - Q_{\theta_{\pi^*}}\|_2 = \sqrt{2f^{\pi^*}(\theta_{\pi^*})} \le \sqrt{2c_0} = (1-\gamma)\varepsilon.$$

Hence $\theta_{\pi^*} \in K$. Since every global minimizer of f lies in K and $f = f^{\pi^*}$ on K, the unique global minimizer of f coincides with θ_{π^*} and therefore equals the unconstrained minimizer of f^{π^*} displayed above. The action-gap condition also gives $\pi_{\theta^*} = \pi^*$, hence $Q^{\pi_{\theta^*}} = Q^*$. The final approximation bound follows from Lemma A.9 applied to the policy π^*; the required normal-equation inverse exists by Lemma A.4. □

Proof of Corollary 1. Let θ_1^* satisfy $Q_{\theta_1^*} = \Gamma_{\Phi|\Phi}Q^*$. By Proposition A.5,

$$f(\theta_1^*) \le \frac{(1+\gamma)^2 n}{2} e_\Phi^2.$$

The strict assumption on e_Φ permits a choice of ε such that

$$\frac{(1+\gamma)\sqrt{n}}{1-\gamma} e_\Phi < \varepsilon < \frac{\Delta}{2}.$$

For $c_0 = (1-\gamma)^2\varepsilon^2/2$, the preceding bound gives $f(\theta_1^*) < c_0$, so $\mathcal{L}_{c_0}$ is nonempty. Applying Theorem A.7 proves uniqueness and exact policy recovery. The displayed approximation bound follows from Proposition A.7. □

Theorem A.8. *Let $(\theta_k)_{k\ge 0}$ be generated by $\theta_{k+1} = \theta_k - \alpha_k g_k$, where $g_k = \mathrm{Proj}_{\partial f(\theta_k)}(0)$ and $\alpha_k > 0$ is generated by Armijo backtracking. Assume that every nonstationary limit point $\bar\theta$ has a unique greedy action at every state. Then $(\theta_k) \subset \mathcal{L}_{f(\theta_0)}$ admits at least one limit point and every limit point is Clarke-stationary.*

Proof. A unique greedy action at every state gives a strictly positive finite action gap and hence a neighborhood on which the active deterministic policy is constant. Thus Assumption A.2 holds. The level-set and regularity hypotheses of Theorem A.1 follow from Lemma A.27, so every limit point is Clarke-stationary. □

A.4.1 Examples

In this example, we verify Theorem A.6 on a simple linear function approximation example. Let us consider an MDP with $\mathcal{S} = \{1\}$, $\mathcal{A} = \{1, 2\}$, and $\gamma = 0.9$ with deterministic self-loop transitions $P(1|1,1) = P(1|1,2) = 1$, and rewards with $R(1,1) = 0, R(1,2) = 1$. Let $Q := \begin{bmatrix} Q(1,1) \\ Q(1,2) \end{bmatrix} \in \mathbb{R}^2$. We use a one-dimensional linear function approximation

$$Q_\theta = \Phi\theta, \qquad \Phi = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^{2\times 1}, \qquad \theta \in \mathbb{R}.$$

Hence, $Q_\theta = \begin{bmatrix} \theta \\ 0 \end{bmatrix}$.

A.4.1.1 Subdifferential and stationary point

Since the transition is a self-loop, the control Bellman operator is

$$TQ_\theta = \begin{bmatrix} \gamma \max_{a' \in \{1,2\}} Q_\theta(1,a') \\ 1 + \gamma \max_{a' \in \{1,2\}} Q_\theta(1,a') \end{bmatrix} = \begin{bmatrix} \gamma \max\{\theta, 0\} \\ 1 + \gamma \max\{\theta, 0\} \end{bmatrix},$$

and the residual is

$$TQ_\theta - Q_\theta = \begin{bmatrix} \gamma \max\{\theta, 0\} - \theta \\ 1 + \gamma \max\{\theta, 0\} \end{bmatrix}.$$

The CBR objective, $f(\theta) = \frac{1}{2}\|T(\Phi\theta) - \Phi\theta\|_2^2$, is nondifferentiable at $\theta = 0$ because of the max operator. The graph of f is depicted in Figure 5. For any θ and any $\beta \in \text{conv}(\Lambda_\Phi(\theta))$, we can

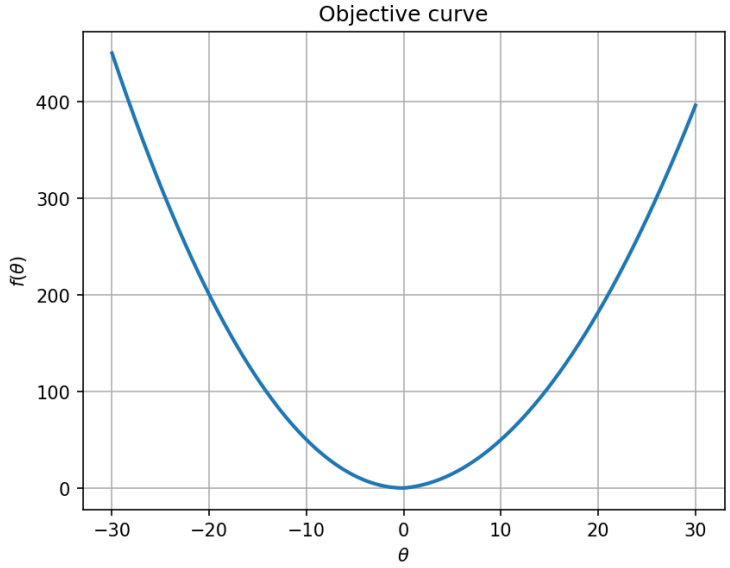


Figure 5: The graph of the CBR objective, $f(\theta) = \frac{1}{2}\|T(\Phi\theta) - \Phi\theta\|_2^2$.

represent

$$\Pi^\beta = e_s^\top \otimes \beta(\cdot|1)^\top = \begin{bmatrix} \beta(1|1) & \beta(2|1) \end{bmatrix}$$

Moreover, since both state-action pairs transition back to 1, the tabular transition matrix from state-action to next-state is

$$P = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \in \mathbb{R}^{2\times 1}.$$

Therefore,

$$P\Pi^\beta = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} \beta(1|1) & 1 - \beta(1|1) \end{bmatrix} = \begin{bmatrix} \beta(1|1) & 1 - \beta(1|1) \\ \beta(1|1) & 1 - \beta(1|1) \end{bmatrix} \in \mathbb{R}^{2\times 2},$$

which is row-stochastic. Using Theorem A.5, we have

$$\partial f(\theta) = \left\{ \Phi^\top (\gamma P \Pi^\beta - I)^\top (TQ_\theta - Q_\theta) \ : \ \beta \in \text{conv}(\Lambda_\Phi(\theta)) \right\},$$

and an element of this set can be parameterized by $\beta(1|1) \in [0,1]$ as

$$\begin{aligned}
\Phi^\top (\gamma P \Pi^\beta - I)^\top (TQ_\theta - Q_\theta) =& \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \gamma\beta(1|1) - 1 & \gamma(1 - \beta(1|1)) \\ \gamma\beta(1|1) & \gamma(1 - \beta(1|1)) - 1 \end{bmatrix}^\top \begin{bmatrix} \gamma \max\{\theta, 0\} - \theta \\ 1 + \gamma \max\{\theta, 0\} \end{bmatrix} \\
=& \begin{bmatrix} \gamma\beta(1|1) - 1 & \gamma\beta(1|1) \end{bmatrix} \begin{bmatrix} \gamma \max\{\theta, 0\} - \theta \\ 1 + \gamma \max\{\theta, 0\} \end{bmatrix} \\
=& (\gamma\beta(1|1) - 1)(\gamma \max\{\theta, 0\} - \theta) \\
& + \gamma\beta(1|1)(1 + \gamma \max\{\theta, 0\})
\end{aligned}$$

Case 1: $\theta > 0$**.** $Q_\theta = [\theta, 0]^\top$ and the greedy action is uniquely $a = 1$. Therefore, $\beta(1|1) = 1, \beta(2|1) = 0$ and $\mathrm{conv}(\Lambda_\Phi(\theta)) = \{\beta\}$. A direct calculation yields the singleton subdifferential

$$\partial f(\theta) = \{\gamma + (\gamma^2 + (1-\gamma)^2)\theta\}, \qquad \theta > 0.$$

Case 2: $\theta < 0$**.** The greedy action is uniquely $a = 2$, and hence $\beta(1|1) = 0, \beta(2|1) = 1$ and $\mathrm{conv}(\Lambda_\Phi(\theta)) = \{\beta\}$. A direct calculation yields

$$\partial f(\theta) = \{\theta\}, \qquad \theta < 0.$$

Case 3: $\theta = 0$ In this case, $Q_\theta = [0, 0]^\top$, and hence, both actions $a = 1, 2$ are greedy. Therefore, $\mathrm{conv}(\Lambda_\Phi(\theta))$ contains all mixtures

$$\beta(\cdot|1) = \begin{bmatrix} \beta(1|1) \\ \beta(2|1) \end{bmatrix} = \begin{bmatrix} \beta(1|1) \\ 1 - \beta(1|1) \end{bmatrix}, \qquad \beta(1|1) \in [0, 1].$$

Moreover, the control Bellman residual is

$$TQ_\theta - Q_\theta = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Substituting the above formulations into the subdifferential formula gives

$$\Phi^\top (\gamma P \Pi^\beta - I)^\top (TQ_\theta - Q_\theta) = \gamma \beta(1|1).$$

Therefore, the Clarke subdifferential at $\theta = 0$ satisfies

$$\partial f(0) = \{\gamma\beta(1|1) : \beta(1|1) \in [0, 1]\} = [0, \gamma] = [0, 0.9].$$

In summary, we have

$$\partial f(\theta) = \begin{cases} \{\gamma + (\gamma^2 + (1-\gamma)^2)\theta\}, & \theta > 0, \\ [0, \gamma], & \theta = 0, \\ \{\theta\}, & \theta < 0. \end{cases}$$

Since $0 \in \partial f(0)$, choosing $\beta(1|1) = 0 \in [0, 1]$ selects a zero subgradient. Therefore, $\theta = 0$ is a stationary point. At the stationary point $\theta = 0$, the CBR objective value is $f(0) = \frac{1}{2} \|TQ_\theta - Q_\theta\|_2^2 = \frac{1}{2}$.

A.4.1.2 Oblique projected Bellman equation

With the choice $\beta(1|1) = 0$, Ψ_β is given by

$$\Psi_\beta := (\gamma P \Pi^\beta - I)\Phi = \begin{bmatrix} -1 & \gamma \\ 0 & \gamma - 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

Moreover, the corresponding $\Psi_\beta^\top \Phi$ is

$$\Psi_\beta^\top \Phi = [-1 \quad 0] \begin{bmatrix} 1 \\ 0 \end{bmatrix} = -1.$$

Therefore, $(\Psi_\beta^\top \Phi)^{-1}$ exists and equals -1. Under the nonsingularity condition verified above, Theorem A.6 gives the following single-valued oblique projected Bellman equation at the stationary point:

$$Q_{\bar\theta} = \Gamma_{\Phi|\Psi_\beta} T Q_{\bar\theta}, \quad \Gamma_{\Phi|\Psi_\beta}(x) = \Phi(\Psi_\beta^\top \Phi)^{-1} \Psi_\beta^\top x, \qquad \Psi_\beta = (\gamma P \Pi^\beta - I)\Phi$$

for some $\beta \in \mathrm{conv}(\Lambda_\Phi(\bar\theta))$. In our example with $\bar\theta = 0$ and $\beta(1|1) = 0$, for any $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$,

$$\Gamma_{\Phi|\Psi_\beta}(x) = \Phi(\Psi_\beta^\top \Phi)^{-1} \Psi_\beta^\top x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} (-1) \begin{bmatrix} -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}.$$

Moreover, we have

$$Q_{\bar\theta} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad TQ_{\bar\theta} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Therefore, the OP-BE is computed as

$$\Gamma_{\Phi|\Psi_\beta}(TQ_0) = \Gamma_{\Phi|\Psi_\beta} \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = Q_{\bar\theta}.$$

Therefore, the Clarke-stationary point $\bar{\theta} = 0$ indeed satisfies the OP-BE.

A.4.1.3 Subgradient method

We can apply the subgradient descent method described in Theorem A.8 to this example, and the results are presented in Figure 6, which shows the evolution of the iterates of the subgradient descent method converging to the stationary point.

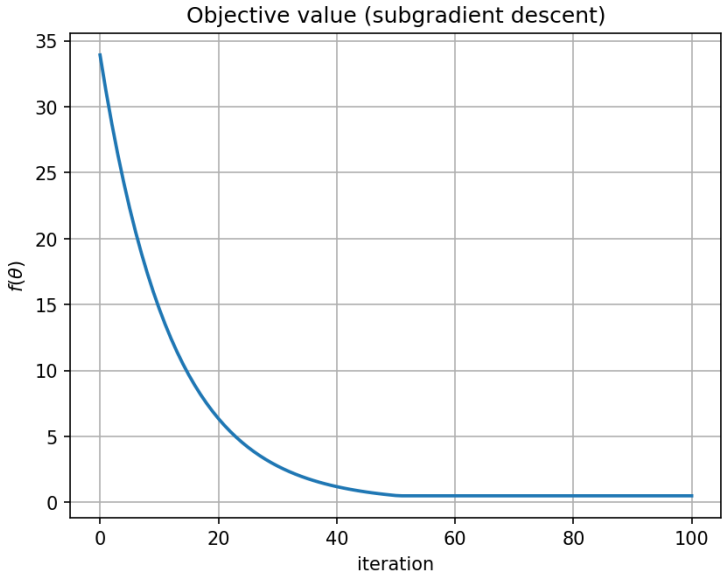


Figure 6: The evolution of the iterates of the subgradient descent method converging to the stationary point.

A.5 SCBR tabular case

Proof of Proposition 7. For any vector $q \in \mathbb{R}^{|\mathcal{A}|}$,

$$\max_a q_a \le \lambda \log \sum_a e^{q_a/\lambda} \le \max_a q_a + \lambda \log |\mathcal{A}|.$$

Consequently, componentwise,

$$0 \le F_\lambda Q - TQ \le \gamma\lambda \log |\mathcal{A}|\, \mathbf{1}.$$

Using the contraction of F_λ and $TQ^* = Q^*$ gives

$$\|Q^*_\lambda - Q^*\|_\infty \le \gamma \|Q^*_\lambda - Q^*\|_\infty + \gamma\lambda \log |\mathcal{A}|,$$

which proves the first bound. For arbitrary Q_θ,

$$\begin{aligned}\|TQ_\theta - Q_\theta\|_\infty &\le \|F_\lambda Q_\theta - Q_\theta\|_\infty + \|F_\lambda Q_\theta - TQ_\theta\|_\infty \\ &\le \sqrt{2 f_\lambda(\theta)} + \gamma\lambda \log |\mathcal{A}|.\end{aligned}$$

The second bound follows from Lemma A.7, and the third follows from the hard-greedy policy bound in Lemma A.29. If the displayed quantity is less than $(1-\gamma)\Delta/2$, then the second bound gives $\|Q_\theta - Q^*\|_\infty < \Delta/2$, so every hard greedy action agrees with the unique optimal action. Finally, the first bound gives $\|Q^*_\lambda - Q^*\|_\infty < \Delta/2$ under the stated temperature condition, so every hard greedy policy of Q^*_λ is optimal; this argument does not require Q^*_λ to lie in $\mathcal{R}(\Phi)$. ☐

Proposition A.9. *The level set $\mathcal{L}_c = \{Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} : f(Q) \le c\}$ is bounded and closed (compact).*

Proof. First of all, we can derive the following bounds:

$$\begin{aligned}\|Q - F_\lambda Q\|_\infty =& \|Q - Q^*_\lambda + F_\lambda Q^*_\lambda - F_\lambda Q\|_\infty \\ \ge& \|Q - Q^*_\lambda\|_\infty - \|F_\lambda Q^*_\lambda - F_\lambda Q\|_\infty \\ \ge& \|Q - Q^*_\lambda\|_\infty - \gamma \|Q^*_\lambda - Q\|_\infty \\ =& (1-\gamma)\|Q - Q^*_\lambda\|_\infty \\ \ge& (1-\gamma)\frac{1}{\sqrt{|\mathcal{S}\times\mathcal{A}|}}\|Q - Q^*_\lambda\|_2,\end{aligned}$$

where the first inequality comes from the reverse triangle inequality and the second inequality is due to the contraction property of the soft control Bellman operator F_λ [12]. Using the above bound, the objective f can be bounded as

$$f(Q) = \frac{1}{2}\|Q - F_\lambda Q\|_2^2 \ge \frac{1}{2}\|Q - F_\lambda Q\|_\infty^2 \ge \frac{(1-\gamma)^2}{2|\mathcal{S}\times\mathcal{A}|}\|Q - Q^*_\lambda\|_2^2.$$

Therefore, $f(Q) \le c$ implies $\|Q - Q^*_\lambda\|_2 \le \sqrt{\frac{2|\mathcal{S}\times\mathcal{A}|c}{(1-\gamma)^2}}$. This implies the following inclusion relation:

$$\mathcal{L}_c \subseteq \left\{ Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} : \|Q - Q^*_\lambda\|_2 \le \sqrt{\frac{2|\mathcal{S}\times\mathcal{A}|c}{(1-\gamma)^2}} \right\}.$$

Since the right-hand side above is bounded, so is $\mathcal{L}_c$. Moreover, since f is continuous, its level set is closed [24, Theorem 1.6]. This completes the proof. ☐

Lemma A.30. *f is C^∞.*

Proof. The log-sum-exp map is C^∞, and hence $Q \mapsto F_\lambda Q - Q$ is C^∞. Since $x \mapsto \frac{1}{2}\|x\|_2^2$ is also C^∞, their composition $f(Q) = \frac{1}{2}\|F_\lambda Q - Q\|_2^2$ is C^∞. ☐

Lemma A.31. *The gradient of f is given by*

$$\nabla_Q f(Q) = (\gamma P \Pi^{\pi_Q} - I)^\top (F_\lambda Q - Q),$$

where $\pi_Q(j|i) := \frac{\exp(Q(i,j)/\lambda)}{\sum_{u\in\mathcal{A}}\exp(Q(i,u)/\lambda)}$ is the Boltzmann policy of Q.

Proof. By direct calculations, we can derive

$$\frac{\partial f(Q)}{\partial Q(i,j)} = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \delta(s,a)\frac{\partial}{\partial Q(i,j)}\left[(F_\lambda Q)(s,a) - Q(s,a)\right],$$

where $\delta(s,a) := (F_\lambda Q)(s,a) - Q(s,a)$. In the equation, the term $\frac{\partial}{\partial Q(i,j)}[(F_\lambda Q)(s,a) - Q(s,a)]$ can be further expressed as

$$\begin{aligned}
&\frac{\partial}{\partial Q(i,j)}[(F_\lambda Q)(s,a) - Q(s,a)] \\
=&\gamma \sum_{s'\in\mathcal{S}} P(s'|s,a)\frac{\partial}{\partial Q(i,j)}\lambda \ln\left(\sum_{u\in\mathcal{A}} \exp(Q(s',u)/\lambda)\right) - \mathbf{1}[(i,j)=(s,a)] \\
=&\gamma \sum_{s'\in\mathcal{S}} P(s'|s,a)\mathbf{1}[s'=i]\pi_Q(j|i) - \mathbf{1}[(i,j)=(s,a)] \\
=&\gamma P(i|s,a)\pi_Q(j|i) - \mathbf{1}[(i,j)=(s,a)],
\end{aligned}$$

where $\mathbf{1}[s'=i]$ and $\mathbf{1}[(i,j)=(s,a)]$ are indicator functions, and

$$\pi_Q(j|i) := \frac{\exp(Q(i,j)/\lambda)}{\sum_{u\in\mathcal{A}} \exp(Q(i,u)/\lambda)}$$

is the Boltzmann policy corresponding to Q. Combining the last results leads to

$$\begin{aligned}
\frac{\partial f(Q)}{\partial Q(i,j)} &= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \delta(s,a)[\gamma P(i|s,a)\pi_Q(j|i) - \mathbf{1}[(i,j)=(s,a)]] \\
&= \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \delta(s,a)P(i|s,a)\pi_Q(j|i) - \delta(i,j).
\end{aligned}$$

Now, the gradient can be written as

$$\nabla_Q f(Q) = \begin{bmatrix} \frac{\partial f(Q)}{\partial Q(1,1)} \\ \frac{\partial f(Q)}{\partial Q(1,2)} \\ \vdots \\ \frac{\partial f(Q)}{\partial Q(|\mathcal{S}|,|\mathcal{A}|)} \end{bmatrix} = (\gamma P\Pi^{\pi_Q} - I)^\top (F_\lambda Q - Q) \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}.$$

This completes the proof. □

Theorem A.9. *The stationary point, $\nabla_Q f(\bar{Q}) = 0$, is the unique solution $\bar{Q} = Q^*_\lambda$ to the soft control Bellman equation $F_\lambda Q^*_\lambda = Q^*_\lambda$.*

Proof. Suppose that $\bar{Q}$ is a stationary point, which implies $\nabla_Q f(\bar{Q}) = (\gamma P\Pi^{\pi_{\bar{Q}}} - I)^\top (F_\lambda \bar{Q} - \bar{Q}) = 0$. Since $\gamma P\Pi^{\pi_{\bar{Q}}} - I$ is invertible by Lemma A.3, we have $F_\lambda \bar{Q} = \bar{Q}$. The soft Bellman operator F_λ is a γ-contraction, so this fixed point is unique and $\bar{Q} = Q^*_\lambda$. Conversely, $F_\lambda Q^*_\lambda = Q^*_\lambda$ makes the residual, and therefore the gradient in Lemma A.31, equal to zero. Thus the stationary-point set is exactly $\{Q^*_\lambda\}$. □

Lemma A.32. *The Hessian of f under the state-major ordering defined in the preliminaries is*

$$\nabla^2_Q f(Q) = (\gamma P\Pi^{\pi_Q} - I)^\top (\gamma P\Pi^{\pi_Q} - I) + \frac{\gamma}{\lambda}\begin{bmatrix} \Omega_1(Q) & 0 & \cdots & 0 \\ 0 & \Omega_2(Q) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Omega_{|\mathcal{S}|}(Q) \end{bmatrix},$$

where

$$\begin{aligned}
\Omega_s(Q) &:= m_s\left(\mathrm{diag}(\pi_Q(\cdot|s)) - \pi_Q(\cdot|s)\pi_Q(\cdot|s)^\top\right), \\
m_s &:= \sum_{(x,a)\in\mathcal{S}\times\mathcal{A}} \delta(x,a)P(s|x,a), \\
\delta(x,a) &:= (F_\lambda Q)(x,a) - Q(x,a).
\end{aligned}$$

Proof. Let $r(Q) := F_\lambda Q - Q$ and $J(Q) := Dr(Q) = \gamma P \Pi^{\pi_Q} - I$. Since $f(Q) = \frac{1}{2} r(Q)^\top r(Q)$, the second-order chain rule gives

$$\nabla_Q^2 f(Q) = J(Q)^\top J(Q) + \sum_{(x,a)\in\mathcal{S}\times\mathcal{A}} \delta(x,a) \nabla_Q^2 r_{x,a}(Q).$$

For each state s, let $E_s := e_s \otimes I_{|\mathcal{A}|}$, so that $E_s^\top Q = Q(s,\cdot)$. The Hessian of the log-sum-exp term is

$$\nabla_Q^2 \left[\lambda \log \sum_{u\in\mathcal{A}} e^{Q(s,u)/\lambda} \right] = \frac{1}{\lambda} E_s \left(\operatorname{diag}(\pi_Q(\cdot|s)) - \pi_Q(\cdot|s)\pi_Q(\cdot|s)^\top \right) E_s^\top.$$

Consequently,

$$\nabla_Q^2 r_{x,a}(Q) = \frac{\gamma}{\lambda} \sum_{s\in\mathcal{S}} P(s|x,a) E_s \left(\operatorname{diag}(\pi_Q(\cdot|s)) - \pi_Q(\cdot|s)\pi_Q(\cdot|s)^\top \right) E_s^\top.$$

Multiplying by $\delta(x,a)$, summing over (x,a), and using the definition of m_s yields a block-diagonal matrix whose s-th action block is $(\gamma/\lambda)\Omega_s(Q)$. Substitution into the second-order chain rule gives the stated formula. □

Lemma A.33. *For every probability vector p,*

$$C(p) := \operatorname{diag}(p) - pp^\top \succeq 0, \qquad C(p)\mathbf{1} = 0, \qquad \|C(p)\|_2 \le \frac{1}{2}.$$

If $|\mathcal{A}| \ge 2$, the constant $1/2$ is sharp.

Proof. For every x, $x^\top C(p) x = \mathbb{E}_p[x_j^2] - (\mathbb{E}_p[x_j])^2 \ge 0$, so $C(p)$ is positive semidefinite, and $C(p)\mathbf{1} = 0$. Since $C(p)$ is symmetric, its spectral norm is bounded by its maximum absolute row sum. Row i has absolute sum

$$p_i(1-p_i) + \sum_{j\ne i} p_i p_j = 2p_i(1-p_i) \le \frac{1}{2}.$$

Thus $\|C(p)\|_2 \le 1/2$. When $|\mathcal{A}| \ge 2$, equality holds for $p = (1/2, 1/2, 0, \ldots, 0)$; when $|\mathcal{A}| = 1$, $C(p) = 0$. □

Lemma A.34. *The set $M := \{Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} : F_\lambda Q - Q \ge 0\}$ satisfies the following properties:*

1. *M is closed*
2. *$Q_\lambda^* \in M$*
3. *$M \subseteq \{Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} : Q_\lambda^* \ge Q\}$*

Proof. Since F_λ is continuous and $K := \{Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} : Q \ge 0\}$ is a closed set, M is closed. In particular, let us define the mapping $g(Q) := F_\lambda(Q) - Q$, which is continuous. Then, $M = g^{-1}(K)$, which is the preimage of the closed set K. This implies that M is closed [25, Theorem 4.8]. Since $F_\lambda Q_\lambda^* = Q_\lambda^*$, we have $F_\lambda Q_\lambda^* - Q_\lambda^* = 0 \ge 0$, and hence $Q_\lambda^* \in M$. Moreover, if $Q \in M$, i.e., $F_\lambda Q \ge Q$, then $F_\lambda^k Q \ge \cdots \ge F_\lambda^3 Q \ge F_\lambda Q \ge Q$ because F_λ is monotone [7]. Therefore, $\lim_{k\to\infty} F_\lambda^k Q = Q_\lambda^* \ge Q$, and $Q \in \{Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} : Q_\lambda^* \ge Q\}$. □

Lemma A.35. *For $M := \{Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} : F_\lambda Q - Q \ge 0\}$, the Hessian is uniformly positive definite on M. Consequently, f is strongly convex on every convex subset of M.*

Proof. Let $J(Q) := \gamma P \Pi^{\pi_Q} - I$. By Lemma A.3, $J(Q)$ is nonsingular for every Q. Moreover, the inverse is uniformly bounded because $\|(I - \gamma P \Pi^{\pi_Q})^{-1}\|_\infty \le (1-\gamma)^{-1}$, and hence $J(Q)^\top J(Q) \succeq c_0 I$ for some $c_0 > 0$ independent of Q. In M, we have $m_s \ge 0$ for every s, so the covariance term in Lemma A.32 is positive semidefinite by Lemma A.33. Therefore, $\nabla_Q^2 f(Q) \succeq c_0 I$ on M. The Hessian criterion for strong convexity then applies on any convex subset of M. This completes the proof. □

Lemma A.36. *f is locally L-smooth, i.e., for any compact C, there exists $L(C) > 0$ such that*

$$\|\nabla_Q f(Q_1) - \nabla_Q f(Q_2)\|_2 \le L(C)\|Q_1 - Q_2\|_2$$

for all $Q_1, Q_2 \in C$, where one may take

$$L(C) = (\gamma\sqrt{|\mathcal{S}\times\mathcal{A}|}+1)^2 + \frac{\gamma}{2\lambda}\|P^\top\|_\infty(\gamma+1)\max_{Q\in C}\|Q^*_\lambda - Q\|_\infty.$$

Proof. Let $n := |\mathcal{S}\times\mathcal{A}|$ and let $r_C := \max_{Q\in C}\|Q - Q^*_\lambda\|_\infty$. For any $Q_1, Q_2 \in C$ and $t \in [0,1]$, convexity of the norm gives

$$\|Q^*_\lambda - [tQ_1 + (1-t)Q_2]\|_\infty \le t\|Q^*_\lambda - Q_1\|_\infty + (1-t)\|Q^*_\lambda - Q_2\|_\infty \le r_C.$$

Thus the entire segment joining Q_1 and Q_2 lies in the compact convex ball $\widehat{C} := \{Q : \|Q - Q^*_\lambda\|_\infty \le r_C\}$, even when C itself is nonconvex.

For every $Q \in \widehat{C}$,

$$\|(\gamma P\Pi^{\pi_Q} - I)^\top(\gamma P\Pi^{\pi_Q} - I)\|_2 \le (\gamma\sqrt{n}+1)^2.$$

Moreover, with m_s as in Lemma A.32,

$$\begin{aligned}\max_s |m_s| &= \|P^\top(F_\lambda Q - Q)\|_\infty \\ &\le \|P^\top\|_\infty\|F_\lambda Q - Q\|_\infty \\ &\le \|P^\top\|_\infty(1+\gamma)\|Q - Q^*_\lambda\|_\infty \\ &\le \|P^\top\|_\infty(1+\gamma)r_C.\end{aligned}$$

Using Lemma A.33 in the Hessian formula therefore gives

$$\|\nabla^2_Q f(Q)\|_2 \le (\gamma\sqrt{n}+1)^2 + \frac{\gamma}{2\lambda}\|P^\top\|_\infty(1+\gamma)r_C = L(C) \qquad (Q \in \widehat{C}).$$

Finally, integrating the Hessian along the segment from Q_2 to Q_1 yields

$$\nabla_Q f(Q_1) - \nabla_Q f(Q_2) = \int_0^1 \nabla^2_Q f(Q_2 + t(Q_1 - Q_2))(Q_1 - Q_2)\,dt,$$

and hence the asserted Lipschitz bound. □

Lemma A.37. *f is locally strongly convex around Q^*_λ, i.e., there exists a ball $B_r(Q^*_\lambda)$ centered at Q^*_λ with radius $r > 0$ and a number $\mu > 0$ such that*

$$\nabla^2_Q f(Q) \succeq \mu I, \quad \forall Q \in B_r(Q^*_\lambda)$$

Proof. Let us consider the Hessian given in Lemma A.32. Since

$$\lim_{Q\to Q^*_\lambda} m_s = \lim_{Q\to Q^*_\lambda} \sum_{(x,a)\in\mathcal{S}\times\mathcal{A}} \delta(x,a)P(s|x,a) = 0,$$

we have

$$\lim_{Q\to Q^*_\lambda} \nabla^2_Q f(Q) = (\gamma P\Pi^{\pi_{Q^*_\lambda}} - I)^\top(\gamma P\Pi^{\pi_{Q^*_\lambda}} - I) \succ 0$$

By Nesterov [22, Theorem 2.1.11], this implies that f is locally strongly convex around Q^*_λ. □

Lemma A.38. *Writing $n := |\mathcal{S}\times\mathcal{A}|$, the tabular SCBR satisfies the global PL inequality*

$$f(Q) - f(Q^*_\lambda) \le \frac{n}{2(1-\gamma)^2}\|\nabla_Q f(Q)\|_2^2, \qquad Q \in \mathbb{R}^n.$$

Proof. Let $M_Q := P\Pi^{\pi_Q}$ and $r_Q := F_\lambda Q - Q$. Since M_Q is row-stochastic,

$$\|(I - \gamma M_Q)^{-1}\|_\infty \le \frac{1}{1-\gamma}.$$

Using $\|A\|_2 \le \sqrt{n}\|A\|_\infty$ for $A \in \mathbb{R}^{n\times n}$ therefore gives

$$\sigma_{\min}(I - \gamma M_Q) \ge \frac{1-\gamma}{\sqrt{n}}.$$

Hence

$$\|\nabla_Q f(Q)\|_2 = \|(I - \gamma M_Q)^\top r_Q\|_2 \geq \frac{1-\gamma}{\sqrt{n}} \|r_Q\|_2.$$

Since $f(Q^*_\lambda) = 0$ and $f(Q) = \frac{1}{2}\|r_Q\|_2^2$, the result follows. □

Before proving the convergence of the gradient descent algorithm, we first present the following descent lemma, which is essential for our analysis.

Lemma A.39 (Lemma 1.2.3 in Nesterov [22])**.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth, then*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^n.$$

We now present the convergence result for the gradient descent algorithm.

Theorem A.10. *Let $Q_{k+1} = Q_k - \alpha \nabla_Q f(Q_k)$ and define $\ell_{\mathrm{PL}} := n/[2(1-\gamma)^2]$, where $n = |\mathcal{S} \times \mathcal{A}|$. Fix $c > f(Q_0)$, set $\mathcal{L}_c := \{Q : f(Q) \leq c\}$, define*

$$D_\alpha := \operatorname{conv}\left(\mathcal{L}_c \cup \{Q - \alpha \nabla_Q f(Q) : Q \in \mathcal{L}_c\}\right),$$

and let L_α be a Lipschitz constant of $\nabla_Q f$ on D_α. If $0 < \alpha < 2/L_\alpha$ and $\ell_{\mathrm{PL}} > \alpha(1 - \alpha L_\alpha/2)$, then

$$f(Q_k) - f(Q^*_\lambda) \leq \left(1 - \frac{\alpha(1 - \alpha L_\alpha/2)}{\ell_{\mathrm{PL}}}\right)^k \big(f(Q_0) - f(Q^*_\lambda)\big).$$

Proof. The descent lemma on D_α gives

$$f(Q^+) \leq f(Q) - \alpha \left(1 - \frac{\alpha L_\alpha}{2}\right) \|\nabla_Q f(Q)\|_2^2, \qquad Q^+ := Q - \alpha \nabla_Q f(Q).$$

Thus all iterates remain in $\mathcal{L}_c$. Applying the global PL inequality in Lemma A.38 and iterating gives the stated rate. □

A.5.1 Examples

Let us consider a discounted MDP with a single state and two actions: $\mathcal{S} = \{1\}, \mathcal{A} = \{1, 2\}, 0 < \gamma < 1, \lambda > 0$. Moreover, let us assume that the transition is self-looping for both actions: $P(1 \mid 1, 1) = P(1 \mid 1, 2) = 1$, and

$$Q := \begin{bmatrix} Q(1,1) \\ Q(1,2) \end{bmatrix}.$$

For this MDP, the soft Bellman operator F_λ satisfies, for $a \in \{1, 2\}$,

$$(F_\lambda Q)(1, a) = R(1, a) + \gamma\lambda \ln(\exp(Q(1,1)/\lambda) + \exp(Q(1,2)/\lambda)).$$

Define the soft control Bellman residual (SCBR)

$$\begin{aligned} \delta(Q) :=& F_\lambda Q - Q \\ =& \begin{bmatrix} R(1,1) + \gamma\lambda \ln\left(\exp(Q(1,1)/\lambda) + \exp(Q(1,2)/\lambda)\right) - Q(1,1) \\ R(1,2) + \gamma\lambda \ln\left(\exp(Q(1,1)/\lambda) + \exp(Q(1,2)/\lambda)\right) - Q(1,2) \end{bmatrix} \\ =& \begin{bmatrix} \delta_1(Q) \\ \delta_2(Q) \end{bmatrix}, \end{aligned}$$

and the SCBR objective

$$f(Q) := \frac{1}{2} \|\delta(Q)\|_2^2 .$$

The Boltzmann (softmax) policy induced by Q is

$$\pi_\lambda(1|1) = \frac{\exp(Q(1,1)/\lambda)}{\exp(Q(1,1)/\lambda) + \exp(Q(1,2)/\lambda)}, \qquad \pi_\lambda(2|1) = \frac{\exp(Q(1,2)/\lambda)}{\exp(Q(1,1)/\lambda) + \exp(Q(1,2)/\lambda)}$$

Let $\pi_1 := \pi_\lambda(1 \mid 1)$ and $\pi_2 := \pi_\lambda(2 \mid 1)$. Then the Jacobian of F_λ is

$$\nabla_Q(F_\lambda Q) = \gamma \begin{bmatrix} \pi_\lambda(1|1) & \pi_\lambda(2|1) \\ \pi_\lambda(1|1) & \pi_\lambda(2|1) \end{bmatrix}.$$

Using the identity $\nabla_Q f(Q) = (\nabla_Q (F_\lambda Q) - I)^\top \delta(Q)$, we obtain the explicit gradient

$$\nabla_Q f(Q) = \begin{bmatrix} (\gamma\pi_\lambda(1|1) - 1)\delta_1(Q) + \gamma\pi_\lambda(1|1)\delta_2(Q) \\ \gamma\pi_\lambda(2|1)\delta_1(Q) + (\gamma\pi_\lambda(2|1) - 1)\delta_2(Q) \end{bmatrix}.$$

A stationary point satisfies $\nabla_Q f(Q) = 0$. Since

$$\begin{aligned} \det(\nabla_Q (F_\lambda Q) - I) =& \det\left(\begin{bmatrix} \gamma\pi_\lambda(1|1) - 1 & \gamma\pi_\lambda(2|1) \\ \gamma\pi_\lambda(1|1) & \gamma\pi_\lambda(2|1) - 1 \end{bmatrix}\right) \\ =& (\gamma\pi_\lambda(1|1) - 1)(\gamma\pi_\lambda(2|1) - 1) - \gamma^2\pi_\lambda(1|1)\pi_\lambda(2|1) \\ =& 1 - \gamma \\ \neq& 0 \end{aligned}$$

we conclude $\nabla_Q f(Q) = 0 \Leftrightarrow \delta(Q) = 0 \Leftrightarrow F_\lambda Q = Q$, i.e., the stationary point coincides with the unique fixed point of the soft Bellman operator. Solving $F_\lambda Q = Q$ yields a closed-form expression. From $\delta_1(Q) = \delta_2(Q) = 0$, we obtain

$$Q(1,1) - Q(1,2) = R(1,1) - R(1,2)$$

and substituting $Q(1,2) = Q(1,1) - \{R(1,1) - R(1,2)\}$ into $\delta_1(Q) = 0$ gives

$$\begin{aligned} & R(1,1) + \gamma\lambda \ln(\exp(Q(1,1)/\lambda) + \exp(Q(1,2)/\lambda)) - Q(1,1) \\ =& R(1,1) + \gamma\lambda \ln(\exp(Q(1,1)/\lambda) + \exp(Q(1,1)/\lambda - \{R(1,1) - R(1,2)\}/\lambda)) - Q(1,1) \\ =& R(1,1) + \gamma\lambda \ln(\exp(Q(1,1)/\lambda)[1 + \exp(\{R(1,2) - R(1,1)\}/\lambda)]) - Q(1,1) \\ =& R(1,1) + \gamma Q(1,1) + \gamma\lambda \ln(1 + \exp(\{R(1,2) - R(1,1)\}/\lambda)) - Q(1,1) \\ =& 0 \end{aligned}$$

which is equivalently written as $(1-\gamma)Q(1,1) = R(1,1) + \gamma\lambda \ln(1 + \exp(-(R(1,1) - R(1,2))/\lambda))$. Therefore, the solution of the SCBE can be calculated as

$$\begin{aligned} Q^*_\lambda(1,1) =& \frac{R(1,1) + \gamma\lambda \ln(1 + \exp(-(R(1,1) - R(1,2))/\lambda))}{1 - \gamma}, \\ Q^*_\lambda(1,2) =& Q^*_\lambda(1,1) - \{R(1,1) - R(1,2)\}. \end{aligned}$$

Moreover, the induced Boltzmann policy at the stationary point has the following form:

$$\pi^*_\lambda(1|1) = \frac{1}{1 + \exp(-(R(1,1) - R(1,2))/\lambda)}, \qquad \pi^*_\lambda(2|1) = 1 - \pi^*_\lambda(1|1).$$

As a concrete example, let us choose $\gamma = 0.9, \lambda = 1,\ R(1,1) = 1, R(1,2) = 0$. Then, we obtain

$$\begin{aligned} Q^*_\lambda(1,1) =& \frac{R(1,1) + \gamma\lambda \ln(1 + \exp(-(R(1,1) - R(1,2))/\lambda))}{1 - \gamma} \\ =& \frac{1 + 0.9 \log(1 + e^{-1})}{0.1} \\ \cong& 12.8194 \\ Q^*_\lambda(1,2) =& Q^*_\lambda(1,1) - (R(1,1) - R(1,2)) \cong 11.8194 \end{aligned}$$

The corresponding Boltzmann policy is computed as

$$\pi^*_\lambda(1|1) = \frac{1}{1 + e^{-1}} \cong 0.7311, \quad \pi^*_\lambda(2|1) = \frac{e^{-1}}{1 + e^{-1}} \cong 0.2689.$$

For this example, we can apply gradient descent, and the results are illustrated in Figure 7 and Figure 8. Figure 7 shows the evolution of the SCBR objective function value over iterations on a log scale. Over the plotted range, the objective exhibits an approximately geometric decrease. Figure 8 shows the convergence of $Q_k(1,1)$ and $Q_k(1,2)$, where the dashed lines indicate the optimal Q-values, $Q^*_\lambda(1,1)$ and $Q^*_\lambda(1,2)$. As observed, both $Q_k(1,1)$ and $Q_k(1,2)$ converge well to their optimal Q-values.

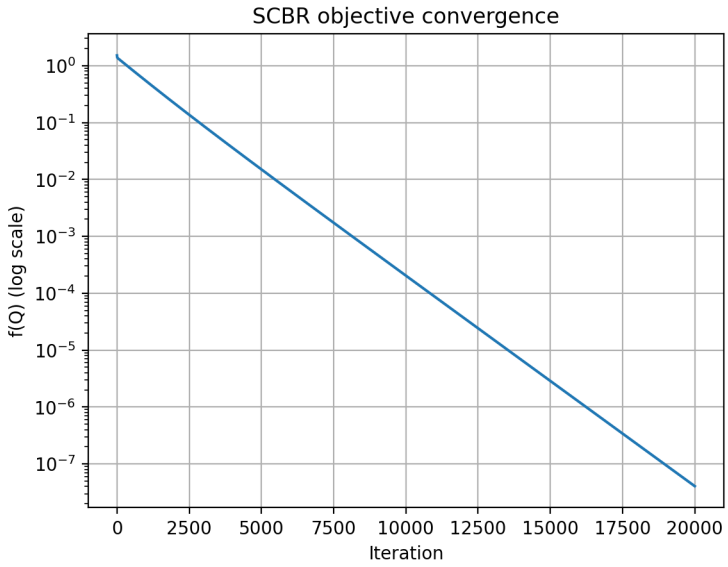


Figure 7: The evolution of the SCBR objective function value over iterations on a log scale.

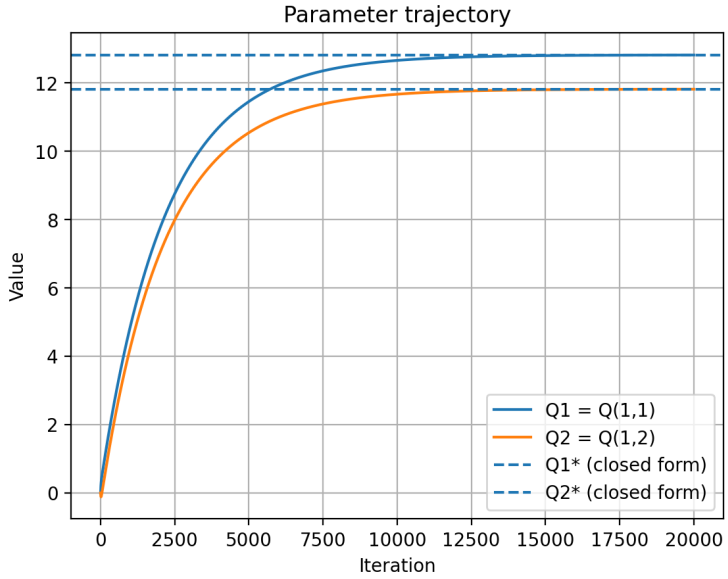


Figure 8: The convergence of $Q_k(1,1)$ and $Q_k(1,2)$ to $Q^*_\lambda(1,1)$ and $Q^*_\lambda(1,2)$, respectively, where the dashed lines indicate the optimal Q-values.

A.6 SCBR with linear function approximation

In this section, we derive and summarize several properties of the SCBR objective function under linear function approximation. Let us consider the SCBR objective function

$$f(\theta) = \frac{1}{2} \left\| F_\lambda \Phi\theta - \Phi\theta \right\|_2^2 .$$

Lemma A.40. *The level set $\mathcal{L}_c = \{\theta \in \mathbb{R}^m : f(\theta) \leq c\}$ is bounded and closed (compact).*

Proof. First of all, we have

$$\left\| Q_\theta - F_\lambda Q_\theta \right\|_\infty = \left\| Q_\theta - Q^*_\lambda + F_\lambda Q^*_\lambda - F_\lambda Q_\theta \right\|_\infty$$

$$
\begin{aligned}
&\geq \|Q_\theta - Q^*_\lambda\|_\infty - \|F_\lambda Q^*_\lambda - F_\lambda Q_\theta\|_\infty \\
&\geq \|Q_\theta - Q^*_\lambda\|_\infty - \gamma \|Q^*_\lambda - Q_\theta\|_\infty \\
&= (1-\gamma)\|Q_\theta - Q^*_\lambda\|_\infty \\
&\geq (1-\gamma)\frac{1}{\sqrt{|\mathcal{S}\times\mathcal{A}|}}\|Q_\theta - Q^*_\lambda\|_2,
\end{aligned}
$$

where the second line is due to the reverse triangle inequality and the third line is due to the fact that F_λ is a contraction. Using the above inequality, we can derive the lower bound

$$
f(\theta) = \frac{1}{2}\|Q_\theta - F_\lambda Q_\theta\|_2^2 \geq \frac{1}{2}\|Q_\theta - F_\lambda Q_\theta\|_\infty^2 \geq \frac{(1-\gamma)^2}{2|\mathcal{S}\times\mathcal{A}|}\|Q_\theta - Q^*_\lambda\|_2^2 .
$$

Therefore, one can conclude that $f(\theta) \leq c$ implies $\|Q_\theta - Q^*_\lambda\|_2 \leq \frac{\sqrt{2|\mathcal{S}\times\mathcal{A}|c}}{1-\gamma}$. Hence,

$$
\|Q_\theta\|_2 \leq \|Q_\theta - Q^*_\lambda\|_2 + \|Q^*_\lambda\|_2 \leq \frac{\sqrt{2|\mathcal{S}\times\mathcal{A}|c}}{1-\gamma} + \|Q^*_\lambda\|_2.
$$

Since Φ has full column rank,

$$
\sigma_{\min}(\Phi)\,\|\theta\|_2 \leq \|\Phi\theta\|_2 = \|Q_\theta\|_2,
$$

and therefore

$$
\|\theta\|_2 \leq \frac{1}{\sigma_{\min}(\Phi)}\left(\frac{\sqrt{2|\mathcal{S}\times\mathcal{A}|c}}{1-\gamma} + \|Q^*_\lambda\|_2\right).
$$

Therefore, $\mathcal{L}_c$ is bounded. Moreover, since f is continuous, its level set is closed [24, Theorem 1.6]. This completes the proof. □

Theorem A.11. *The gradient of f is given by*

$$
\nabla_\theta f(\theta) = \Phi^\top(\gamma P \Pi^{\pi_\theta} - I)^\top(F_\lambda(\Phi\theta) - \Phi\theta),
$$

where $\pi_\theta(j|i) := \frac{\exp(Q_\theta(i,j)/\lambda)}{\sum_{u\in\mathcal{A}}\exp(Q_\theta(i,u)/\lambda)}$ is the Boltzmann policy of Q_θ.

Proof. By direct calculation, we can prove that the gradient is given by

$$
\nabla_\theta f(\theta) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \delta(s,a)\nabla_\theta[(F_\lambda Q_\theta)(s,a) - Q_\theta(s,a)],
$$

where

$$
\begin{aligned}
\nabla_\theta[(F_\lambda Q_\theta)(s,a) - Q_\theta(s,a)] =& \gamma\sum_{s'\in\mathcal{S}} P(s'|s,a)\lambda\nabla_\theta \ln\left(\sum_{u\in\mathcal{A}}\exp(e_{s',u}^\top\Phi\theta/\lambda)\right) - \Phi^\top e_{s,a} \\
=& \gamma\sum_{s'\in\mathcal{S}} P(s'|s,a)\lambda\frac{\nabla_\theta\sum_{u\in\mathcal{A}}\exp(e_{s',u}^\top\Phi\theta/\lambda)}{\sum_{u\in\mathcal{A}}\exp(e_{s',u}^\top\Phi\theta/\lambda)} - \Phi^\top e_{s,a} \\
=& \gamma\sum_{s'\in\mathcal{S}} P(s'|s,a)\sum_{u\in\mathcal{A}}\pi_\theta(u|s')\Phi^\top e_{s',u} - \Phi^\top e_{s,a} \\
=& \Phi^\top\left\{\gamma\sum_{s'\in\mathcal{S}} P(s'|s,a)\sum_{u\in\mathcal{A}}\pi_\theta(u|s')e_{s',u} - e_{s,a}\right\},
\end{aligned}
$$

where $e_{s,a} := e_s \otimes e_a \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, $e_s \in \mathbb{R}^{|\mathcal{S}|}$ is the standard basis vector whose s-th element is one and whose other elements are zero, $e_a \in \mathbb{R}^{|\mathcal{A}|}$ is the standard basis vector whose a-th element is one and whose other elements are zero. Combining the two results leads to

$$
\begin{aligned}
\nabla_\theta f(\theta) =& \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Phi^\top\left\{\gamma\sum_{s'\in\mathcal{S}} P(s'|s,a)\sum_{u\in\mathcal{A}}\pi_\theta(u|s')e_{s',u} - e_{s,a}\right\}\delta(s,a) \\
=& \Phi^\top(\gamma P\Pi^{\pi_\theta} - I)^\top(F_\lambda(\Phi\theta) - \Phi\theta).
\end{aligned}
$$

This completes the proof. □

Theorem A.12. *For every $\bar{\theta}$,*

$$\nabla f(\bar{\theta}) = 0 \quad \Longleftrightarrow \quad Q_{\bar{\theta}} \in \mathcal{G}_{\Phi|\Psi_{\bar{\theta}}}(F_\lambda Q_{\bar{\theta}}), \qquad \Psi_\theta := (\gamma P \Pi^{\pi_\theta} - I)\Phi.$$

If $\Psi_{\bar{\theta}}^\top \Phi$ is nonsingular, this set is the singleton $\{\Gamma_{\Phi|\Psi_{\bar{\theta}}} F_\lambda Q_{\bar{\theta}}\}$.

Proof. By Theorem A.11, $\nabla f(\bar{\theta}) = 0$ is equivalent to

$$\Psi_{\bar{\theta}}^\top (F_\lambda Q_{\bar{\theta}} - Q_{\bar{\theta}}) = 0.$$

Since $Q_{\bar{\theta}} = \Phi\bar{\theta}$, this is exactly the defining condition for membership in $\mathcal{G}_{\Phi|\Psi_{\bar{\theta}}}(F_\lambda Q_{\bar{\theta}})$. Under nonsingularity, solving the linear equation gives the usual oblique-projector formula. □

Lemma A.41. *The Hessian under the state-major ordering defined in the preliminaries is*

$$\nabla_\theta^2 f(\theta) = \Phi^\top (\gamma P \Pi^{\pi_{Q_\theta}} - I)^\top (\gamma P \Pi^{\pi_{Q_\theta}} - I)\Phi + \frac{\gamma}{\lambda}\Phi^\top \begin{bmatrix} \Omega_1(Q_\theta) & 0 & \cdots & 0 \\ 0 & \Omega_2(Q_\theta) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Omega_{|\mathcal{S}|}(Q_\theta) \end{bmatrix} \Phi,$$

where

$$\begin{aligned} \Omega_s(Q_\theta) &:= m_s \left(\mathrm{diag}(\pi_{Q_\theta}(\cdot|s)) - \pi_{Q_\theta}(\cdot|s)\pi_{Q_\theta}(\cdot|s)^\top \right), \\ m_s &:= \sum_{(x,a)\in\mathcal{S}\times\mathcal{A}} \delta(x,a) P(s|x,a), \\ \delta(x,a) &:= (F_\lambda Q_\theta)(x,a) - Q_\theta(x,a). \end{aligned}$$

Proof. Let $\widehat{f}(Q) := \frac{1}{2}\|F_\lambda Q - Q\|_2^2$, so that $f(\theta) = \widehat{f}(\Phi\theta)$. Since $\theta \mapsto \Phi\theta$ is affine, the chain rule gives

$$\nabla_\theta f(\theta) = \Phi^\top \nabla_Q \widehat{f}(Q_\theta), \qquad \nabla_\theta^2 f(\theta) = \Phi^\top \nabla_Q^2 \widehat{f}(Q_\theta)\Phi;$$

there is no additional second-derivative term from the affine map. Substituting the tabular Hessian formula of Lemma A.32 at $Q = Q_\theta$ gives exactly the displayed expression. □

Lemma A.42. *f is bounded below and coercive.*

Proof. Clearly $f(\theta) \geq 0$. Writing $n := |\mathcal{S} \times \mathcal{A}|$, the contraction of F_λ and the fixed-point identity $F_\lambda Q_\lambda^* = Q_\lambda^*$ give

$$f(\theta) \geq \frac{(1-\gamma)^2}{2n}\|\Phi\theta - Q_\lambda^*\|_2^2.$$

Because Φ has full column rank,

$$\|\Phi\theta - Q_\lambda^*\|_2 \geq \sigma_{\min}(\Phi)\|\theta\|_2 - \|Q_\lambda^*\|_2 \longrightarrow \infty \quad \text{as } \|\theta\|_2 \to \infty.$$

The lower bound therefore tends to infinity, proving coercivity. □

Lemma A.43. *A stationary point of f always exists.*

Proof. Since f is continuous and coercive by Lemma A.42, it attains a global minimum. Then, the minimizer is a stationary point. This completes the proof. □

Lemma A.44. *f is locally L-smooth, i.e., for any compact C, there exists some $L(C) > 0$ such that*

$$\|\nabla_\theta f(\theta_1) - \nabla_\theta f(\theta_2)\|_2 \leq L(C)\|\theta_1 - \theta_2\|_2$$

for all $\theta_1, \theta_2 \in C$.

Proof. The log-sum-exp map is C^∞, so the Hessian in Lemma A.41 is continuous. For a compact set C, its convex hull $\mathrm{conv}(C)$ is compact. Define

$$L(C) := \max_{\theta\in\mathrm{conv}(C)} \|\nabla_\theta^2 f(\theta)\|_2 < \infty.$$

For any $\theta_1, \theta_2 \in C$, the segment joining them is contained in $\mathrm{conv}(C)$. Hence

$$\nabla f(\theta_1) - \nabla f(\theta_2) = \int_0^1 \nabla^2 f(\theta_2 + t(\theta_1 - \theta_2))(\theta_1 - \theta_2)\, dt,$$

which implies the stated Lipschitz bound. □

Lemma A.45. *Let us define*

$$H(\theta) := \Phi^\top(\gamma P\Pi^{\pi_{Q_\theta}} - I)^\top(\gamma P\Pi^{\pi_{Q_\theta}} - I)\Phi, \qquad \underline{\lambda}_H := \inf_{\theta\in\mathbb{R}^m} \lambda_{\min}(H(\theta)).$$

Under full column rank of Φ, we have $\underline{\lambda}_H > 0$. If there exists

$$0 < c \le \frac{1}{2}\left(\frac{\underline{\lambda}_H}{\frac{\gamma}{\lambda}\|\Phi\|_2^2\|P^\top\|_\infty}\right)^2 \tag{20}$$

such that the level set

$$\mathcal{L}_c := \left\{\theta \in \mathbb{R}^m : f(\theta) = \frac{1}{2}\|F_\lambda Q_\theta - Q_\theta\|_2^2 \le c\right\}$$

is nonempty, then

$$\nabla_\theta^2 f(\theta) \succeq \frac{1}{2}\underline{\lambda}_H I \qquad \textit{for every } \theta \in \mathcal{L}_c.$$

Consequently, f is $\underline{\lambda}_H/2$-strongly convex on every convex subset of $\mathcal{L}_c$.

Proof. Let $A(\theta) := \gamma P\Pi^{\pi_{Q_\theta}} - I$. By Lemma A.3, $A(\theta)$ is nonsingular for every θ. Moreover,

$$\|A(\theta)^{-1}\|_\infty = \|(I - \gamma P\Pi^{\pi_{Q_\theta}})^{-1}\|_\infty \le \frac{1}{1-\gamma},$$

so

$$\sigma_{\min}(A(\theta)) \ge \frac{1}{\|A(\theta)^{-1}\|_2} \ge \frac{1}{\sqrt{|\mathcal{S}\times\mathcal{A}|}\,\|A(\theta)^{-1}\|_\infty} \ge \frac{1-\gamma}{\sqrt{|\mathcal{S}\times\mathcal{A}|}}.$$

Since $H(\theta) = (A(\theta)\Phi)^\top(A(\theta)\Phi)$ and Φ has full column rank,

$$\lambda_{\min}(H(\theta)) = \sigma_{\min}(A(\theta)\Phi)^2 \ge \sigma_{\min}(A(\theta))^2\,\lambda_{\min}(\Phi^\top\Phi) \ge \frac{(1-\gamma)^2}{|\mathcal{S}\times\mathcal{A}|}\lambda_{\min}(\Phi^\top\Phi) > 0.$$

Hence $\underline{\lambda}_H > 0$.

Now write the Hessian in Lemma A.41 as

$$\nabla_\theta^2 f(\theta) = H(\theta) + E(\theta)$$

with

$$E(\theta) := \frac{\gamma}{\lambda}\Phi^\top \begin{bmatrix} \Omega_1(Q_\theta) & 0 & \cdots & 0 \\ 0 & \Omega_2(Q_\theta) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Omega_{|\mathcal{S}|}(Q_\theta) \end{bmatrix}\Phi.$$

Using Lemma A.33, we can obtain the following bounds:

$$\begin{aligned}
\|E(\theta)\|_2 &\le \frac{\gamma}{2\lambda}\|\Phi\|_2^2 \max_{s\in\mathcal{S}}|m_s| \\
&= \frac{\gamma}{2\lambda}\|\Phi\|_2^2\|P^\top(F_\lambda Q_\theta - Q_\theta)\|_\infty \\
&\le \frac{\gamma}{2\lambda}\|\Phi\|_2^2\|P^\top\|_\infty\|F_\lambda Q_\theta - Q_\theta\|_2 \\
&\le \frac{\gamma}{2\lambda}\|\Phi\|_2^2\|P^\top\|_\infty\sqrt{2c}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\nabla_\theta^2 f(\theta) &= H(\theta) + E(\theta) \\
&\succeq \underline{\lambda}_H I - \|E(\theta)\|_2 I \\
&\succeq \underline{\lambda}_H I - \left\{\frac{\gamma}{2\lambda}\|\Phi\|_2^2\left\|P^\top\right\|_\infty\sqrt{2c}\right\} I.
\end{aligned}$$

If $c > 0$ is chosen such that

$$\frac{\gamma}{2\lambda}\|\Phi\|_2^2\left\|P^\top\right\|_\infty\sqrt{2c} \le \frac{1}{2}\underline{\lambda}_H, \tag{21}$$

then

$$\nabla_\theta^2 f(\theta) \succeq \frac{1}{2}\underline{\lambda}_H I \succ 0.$$

Thus the asserted Hessian lower bound holds at every point of $\mathcal{L}_c$. On any convex subset of $\mathcal{L}_c$, the line segment between two of its points remains in $\mathcal{L}_c$, so integrating this Hessian bound along the segment gives $\underline{\lambda}_H/2$-strong convexity there. The condition in (21) can be equivalently written as (20). This completes the proof. □

Lemma A.46. *Under Assumption 1, $\mathcal{R}(\Phi)$ and $\mathcal{R}(\Psi_\theta)$ with $\Psi_\theta = (\gamma P \Pi^{\pi_\theta} - I)\Phi$ are not orthogonal at any θ for which the assumption holds.*

Proof. Note that $\mathcal{R}(\Phi) = \{\Phi x : x \in \mathbb{R}^m\}$ and $\mathcal{R}((\gamma P \Pi^{\pi_\theta} - I)\Phi) = \{(\gamma P \Pi^{\pi_\theta} - I)\Phi x : x \in \mathbb{R}^m\}$ for a fixed $\theta \in \mathbb{R}^m$. Orthogonality of the two subspaces would require

$$\langle u, v\rangle = u^\top v = 0, \quad \forall u \in \mathcal{R}(\Phi), \quad \forall v \in \mathcal{R}((\gamma P \Pi^{\pi_\theta} - I)\Phi),$$

which is equivalent to

$$x^\top \Phi^\top (\gamma P \Pi^{\pi_\theta} - I)\Phi y = 0, \quad \forall x, y \in \mathbb{R}^m.$$

However, under Assumption 1, $\Phi^\top(\gamma P \Pi^{\pi_\theta} - I)\Phi$ is nonsingular because it is the transpose of $\Psi_\theta^\top \Phi$. Hence it cannot be the zero matrix, and the two subspaces cannot be orthogonal. This completes the proof. □

Theorem A.13. *Let us consider the gradient descent iterates, $\theta_{k+1} = \theta_k - \alpha \nabla_\theta f(\theta_k)$, for $k = 0, 1, \ldots$ with any initial point $\theta_0 \in \mathbb{R}^m$. Fix $c > f(\theta_0)$ and set $\mathcal{L}_c := \{\theta : f(\theta) \le c\}$. For the chosen step size α, define*

$$D_\alpha := \operatorname{conv}\left(\mathcal{L}_c \cup \{\theta - \alpha \nabla_\theta f(\theta) : \theta \in \mathcal{L}_c\}\right),$$

and let $L_\alpha > 0$ be a Lipschitz constant of $\nabla_\theta f$ on D_α. If $0 < \alpha < 2/L_\alpha$, the iterates remain in the compact set $\mathcal{L}_c$, every accumulation point is stationary, $\lim_{k\to\infty} \|\nabla_\theta f(\theta_k)\|_2 = 0$, and

$$\min_{0 \le i \le N} \|\nabla_\theta f(\theta_i)\|_2^2 \le \frac{f(\theta_0) - \min_{\theta \in \mathbb{R}^m} f(\theta)}{(N+1)\alpha\left(1 - \frac{\alpha L_\alpha}{2}\right)}.$$

Proof. For any $\theta \in \mathcal{L}_c$, the point $\theta^+ := \theta - \alpha \nabla_\theta f(\theta)$ and the line segment from θ to θ^+ are contained in D_α. The descent lemma with L_α gives

$$f(\theta^+) \le f(\theta) - \alpha\left(1 - \frac{\alpha L_\alpha}{2}\right)\|\nabla_\theta f(\theta)\|_2^2.$$

Thus all iterates remain in $\mathcal{L}_c$. Summing the resulting inequality over $k = 0, \ldots, N$ yields

$$\alpha\left(1 - \frac{\alpha L_\alpha}{2}\right)\sum_{k=0}^{N} \|\nabla_\theta f(\theta_k)\|_2^2 \le f(\theta_0) - f(\theta_{N+1}) \le f(\theta_0) - \min_{\theta \in \mathbb{R}^m} f(\theta).$$

Letting $N \to \infty$ gives $\|\nabla_\theta f(\theta_k)\|_2 \to 0$, and taking the minimum term in the sum gives the displayed complexity bound. By Lemma A.40, $\mathcal{L}_c$ is compact, so the sequence has accumulation points. If $\theta_{k_j} \to \bar{\theta}$, continuity of the gradient yields $\nabla f(\bar{\theta}) = \lim_j \nabla f(\theta_{k_j}) = 0$. Thus every accumulation point is stationary. □

Theorem A.14. *Let us define*

$$H(\theta) := \Phi^\top (\gamma P \Pi^{\pi_{Q_\theta}} - I)^\top (\gamma P \Pi^{\pi_{Q_\theta}} - I)\Phi, \qquad \underline{\lambda}_H := \inf_{\theta \in \mathbb{R}^m} \lambda_{\min}(H(\theta)).$$

If there exists

$$0 < \varepsilon \le \frac{\lambda}{\sqrt{|\mathcal{S} \times \mathcal{A}|}\gamma(1+\gamma)} \frac{\underline{\lambda}_H}{\|\Phi\|_2^2 \left\|P^\top\right\|_\infty}$$

such that $\mathcal{L}_{c_0}$ is nonempty with $c_0 = \frac{(1-\gamma)^2}{2}\varepsilon^2$, then the following set

$$K := \{\theta \in \mathbb{R}^m : \|Q_\theta - Q_\lambda^*\|_\infty \le \varepsilon\}$$

is nonempty, f is strongly convex in K, and the global minimizer of f is unique and lies in K.

Proof. For any $\theta \in K$, using the contraction of F_λ,

$$f(\theta) = \frac{1}{2}\|F_\lambda Q_\theta - Q_\theta\|_2^2 \le \frac{1}{2}|\mathcal{S} \times \mathcal{A}|(1+\gamma)^2\varepsilon^2.$$

Thus K is contained in the sublevel set with threshold $c_1 = \frac{1}{2}|\mathcal{S} \times \mathcal{A}|(1+\gamma)^2\varepsilon^2$. Moreover, K is convex because it is the inverse image, under the affine map $\theta \mapsto \Phi\theta - Q_\lambda^*$, of the convex ℓ_∞-ball of radius ε. The displayed upper bound on ε implies that c_1 satisfies (20). Hence Lemma A.45 gives the Hessian lower bound on the whole level set containing K, and the convex-subset conclusion of that lemma shows that f is strongly convex on K. Next, the residual bound for the contraction F_λ gives

$$(1-\gamma)\|Q_\theta - Q_\lambda^*\|_\infty \le \|F_\lambda Q_\theta - Q_\theta\|_\infty \le \|F_\lambda Q_\theta - Q_\theta\|_2.$$

Therefore $\mathcal{L}_{c_0} \subseteq K$. Since $\mathcal{L}_{c_0}$ is nonempty, K is nonempty. Moreover, if θ_g is any global minimizer of f, then $f(\theta_g) \le c_0$, so $\theta_g \in K$. Strong convexity on K implies that such a global minimizer is unique. Since a global minimizer is an unconstrained minimizer, it is a stationary point. This completes the proof. □

A.6.1 Examples

In this example, we consider a simple Markov decision process with a single state and two actions: $\mathcal{S} = \{1\}$, $\mathcal{A} = \{1,2\}$, $\gamma \in (0,1)$ and $\lambda > 0$. The linear function approximator for the soft Q-function is given as

$$Q_\theta := \begin{bmatrix} Q_\theta(1,1) \\ Q_\theta(1,2) \end{bmatrix} = \begin{bmatrix} \theta \\ -\theta \end{bmatrix} = \Phi\theta, \qquad \Phi := \begin{bmatrix} 1 \\ -1 \end{bmatrix} \in \mathbb{R}^{2\times 1}, \quad \theta \in \mathbb{R}.$$

The model parameters are

$$R := \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad P = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \in \mathbb{R}^{2\times 1}.$$

The Boltzmann policy associated with Q_θ at state $s = 1$ is the vector

$$\pi_\theta(\cdot|1) := \begin{bmatrix} \pi_\theta(1|1) \\ \pi_\theta(2|1) \end{bmatrix} = \frac{1}{\exp(Q_\theta(1,1)/\lambda) + \exp(Q_\theta(1,2)/\lambda)} \begin{bmatrix} \exp(Q_\theta(1,1)/\lambda) \\ \exp(Q_\theta(1,2)/\lambda) \end{bmatrix}$$

and the corresponding policy matrix mapping the state to state-action pairs is

$$\Pi^{\pi_\theta} = \begin{bmatrix} \pi_\theta(1|1) & \pi_\theta(2|1) \end{bmatrix} \in \mathbb{R}^{1\times 2}.$$

The soft Bellman operator acting on Q_θ can then be written in vector form as

$$F_\lambda(Q_\theta) = R + \gamma P \lambda \ln\left(\exp\left(\frac{Q_\theta(1,1)}{\lambda}\right) + \exp\left(\frac{Q_\theta(1,2)}{\lambda}\right)\right) \in \mathbb{R}^2.$$

For convenience, let us define the following notation:

$$V(\theta) := \lambda \ln\left(\exp\left(\frac{Q_\theta(1,1)}{\lambda}\right) + \exp\left(\frac{Q_\theta(1,2)}{\lambda}\right)\right) = \lambda \ln\left(\exp\left(\frac{\theta}{\lambda}\right) + \exp\left(\frac{-\theta}{\lambda}\right)\right).$$

Moreover, let

$$\delta(\theta) := F_\lambda(\Phi\theta) - \Phi\theta \in \mathbb{R}^2$$

denote the residual vector. Then, the SCBR objective function can be written as

$$f(\theta) = \frac{1}{2}\,\delta(\theta)^\top \delta(\theta).$$

Following Theorem A.11, the gradient of f with respect to θ can be written in matrix-vector form as

$$\begin{aligned}
\nabla_\theta f(\theta) =& \Phi^\top (\gamma P \Pi^{\pi_\theta} - I)^\top (F_\lambda(\Phi\theta) - \Phi\theta) \\
=& \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \gamma\pi_\theta(1|1) - 1 & \gamma\pi_\theta(1|1) \\ \gamma\pi_\theta(2|1) & \gamma\pi_\theta(2|1) - 1 \end{bmatrix} \begin{bmatrix} \gamma V(\theta) - 1 - \theta \\ \gamma V(\theta) - 1 + \theta \end{bmatrix} \\
=& \begin{bmatrix} \gamma\tanh(\theta/\lambda) - 1 & \gamma\tanh(\theta/\lambda) + 1 \end{bmatrix} \begin{bmatrix} \gamma V(\theta) - 1 - \theta \\ \gamma V(\theta) - 1 + \theta \end{bmatrix} \\
=& 2\gamma\tanh(\theta/\lambda)(\gamma V(\theta) - 1) + 2\theta.
\end{aligned}$$

The stationary points should satisfy

$$\nabla_\theta f(\bar\theta) = 2\gamma\tanh(\bar\theta/\lambda)(\gamma V(\bar\theta) - 1) + 2\bar\theta = 0,$$

where $\bar\theta = 0$ is obviously a solution. The symmetry of the linear features and rewards does not by itself guarantee nonzero stationary points. It implies only that any nonzero root occurs together with

its negative. In particular, we can show that

$$\nabla_\theta f(\theta) = 2\gamma \tanh(\theta/\lambda)(\gamma V(\theta) - 1) + 2\theta =: g(\theta)$$

is an odd function $g(-\theta) = -g(\theta)$. Therefore, if $g(\theta) = 0$ admits a nonzero solution θ^*, then $-\theta^*$ should also be a solution. For the numerical values used in the figures below, $\gamma = 0.9$ and $\lambda = 0.5$. With these values, a fixed-point/root-finding computation gives the approximate solutions $\theta^* = 0, 0.2898, -0.2898$. Therefore, there are multiple stationary points. Substituting the numerical values into the oblique projected soft control Bellman equation, we can easily verify that they are solutions of the oblique projected soft control Bellman equation.

Finally, Figures 9 to 11 illustrate $\mathcal{R}(\Phi)$, $F_\lambda(\Phi\bar{\theta})$, $\mathcal{R}(\Psi_{\bar{\theta}})$, and the corresponding projected point for the three stationary points obtained in this way. We can see that these figures are in exact agreement with all the results analyzed in this paper.

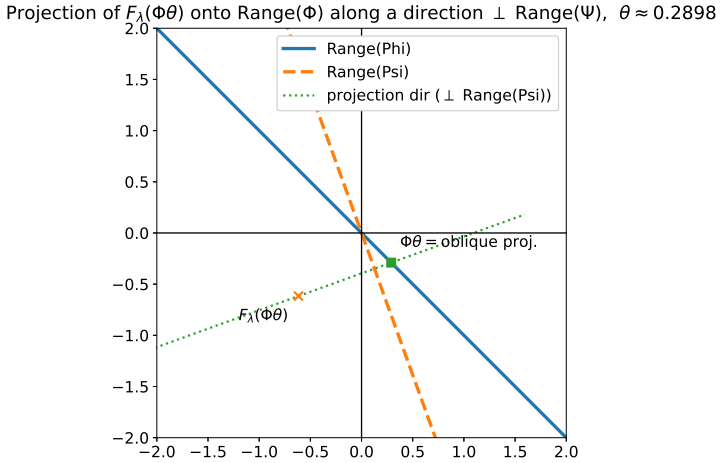


Figure 9: This figure illustrates $\mathcal{R}(\Phi)$, $F_\lambda(\Phi\bar{\theta})$, $\mathcal{R}(\Psi_{\bar{\theta}})$, and the corresponding projected point for the stationary point $\bar{\theta} \approx 0.2898$

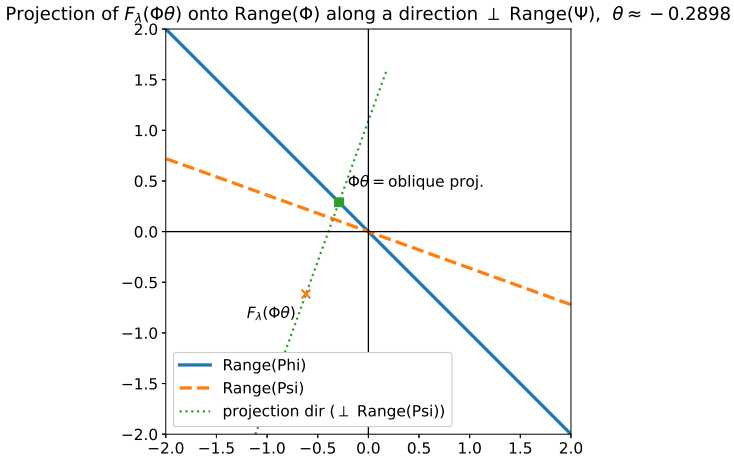


Figure 10: This figure illustrates $\mathcal{R}(\Phi)$, $F_\lambda(\Phi\bar{\theta})$, $\mathcal{R}(\Psi_{\bar{\theta}})$, and the corresponding projected point for the stationary point $\bar{\theta} \approx -0.2898$

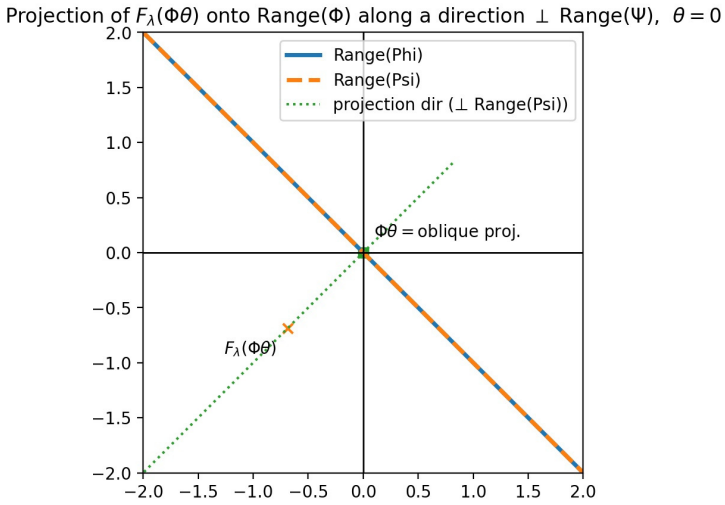


Figure 11: This figure illustrates $\mathcal{R}(\Phi)$, $F_\lambda(\Phi\bar{\theta})$, $\mathcal{R}(\Psi_{\bar{\theta}})$, and the corresponding projected point for the stationary point $\bar{\theta} = 0$

The first plot of Figure 12 shows the evolution of the objective function value f when gradient descent is applied to the objective function in this example. We observe that the objective value decreases and converges to a constant. The second plot of Figure 12 illustrates the convergence of θ_k under gradient descent. We can see that θ_k converges to the positive stationary point, approximately 0.2898.

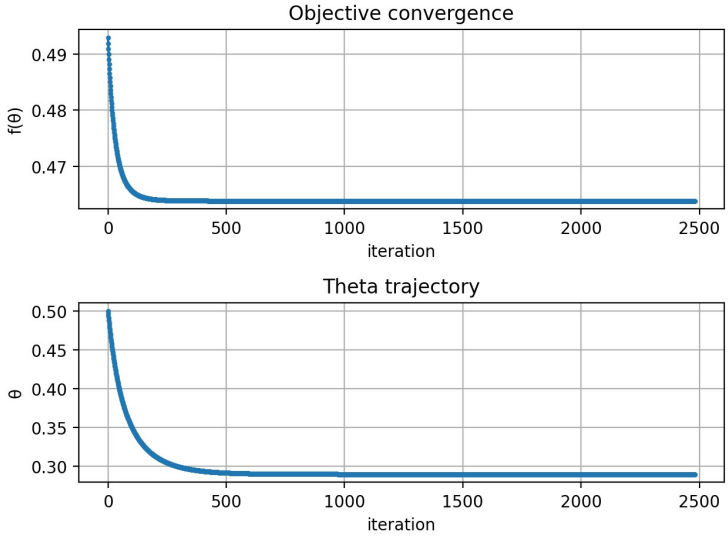


Figure 12: (1) The evolution of the objective function value f when gradient descent is applied to the objective function; (2) The convergence of θ_k under gradient descent

Example 1. *In this example, experiments are conducted on a deterministic* 8×8 `FrozenLake-v1` *map generated with frozen-tile probability* 0.8 *and seed* 101*, with* $|\mathcal{S} \times \mathcal{A}| = 256$*, where we use a random feature matrix with* $m = 120$*. The theoretical development above is stated for non-episodic infinite-horizon discounted MDPs.* `FrozenLake-v1` *is episodic; for terminal transitions, the one-step reward is retained and the continuation term is set to zero. This terminal-mask convention gives the standard direct extension of the same arguments to the episodic case, so we do not restate the parallel theory. We compared gradient descent on the SCBR objective with projected value iteration (P-VI),* $\Phi\theta_{k+1} = \Gamma_{\Phi|\Phi} T(\Phi\theta_k)$ *for* $k = 0, 1, \ldots$*. Unless the composite operator* $\Gamma_{\Phi|\Phi} T$ *satisfies*

additional conditions such as contraction, convergence of P-VI is not guaranteed. For the SCBR gradient method, the theorem guarantees vanishing gradient norms and stationary accumulation points under the stated step-size assumptions; it does not automatically guarantee a small residual or policy error. The comparative results are given in Figure 13, which compares the $\|Q^* - Q_k\|_\infty$ *error of the proposed SCBR against the baseline P-VI.*

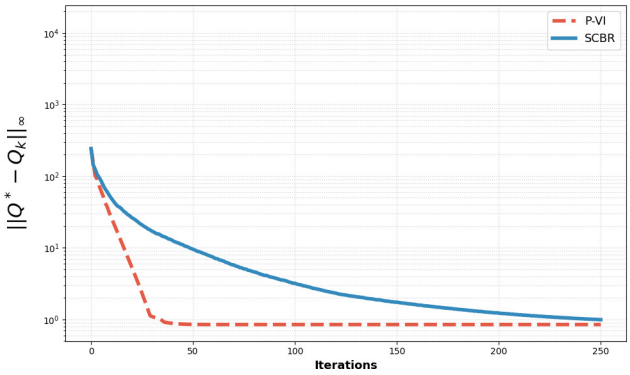


Figure 13: Comparison of the hard-Q errors for gradient descent on the SCBR (blue) and P-VI (red). Both trajectories decrease overall; the SCBR error has occasional nonmonotone fluctuations, while P-VI settles at a smaller final error in this instance.

As illustrated in Figure 13, both trajectories remained stable and substantially reduced the hard-Q error in this experiment. The SCBR curve has occasional nonmonotone fluctuations, whereas P-VI settles at a smaller final hard-Q error. The projected operator is not known to be contractive here, so this observed behavior is instance-specific and is not a general convergence guarantee for P-VI. Following the convergence analysis, we evaluated the quality of the policies derived from the trained Q-functions. Consistently with the corresponding Bellman operators, the code extracts a greedy policy from P-VI and a Boltzmann policy with temperature 10^{-4} *from SCBR. We conducted 2,000 simulation episodes, each with a maximum limit of 100 steps. The agent's starting position was randomly sampled from the* `8x8 FrozenLake-v1` *grid, excluding hole (H) and goal (G) states. In these simulations, the P-VI policy achieved an average success rate of* 23.0%*, while the SCBR policy achieved* 12.4%*.*

A.7 Deep SCBR version

In this section, we consider a deep reinforcement learning (RL) variant of SCBR directly: the proposed objective uses the soft Bellman residual from the outset, rather than first introducing a hard Bellman residual and then changing the Bellman operator. The theoretical guarantees established above for tabular and linear function approximation do not directly carry over to nonlinear neural-network approximation, so the purpose of this section is empirical.

Let Q_θ be a neural-network Q-function and define the soft state value

$$V_{\lambda,\theta}(s) := \lambda \log \sum_{a' \in \mathcal{A}} \exp\left(\frac{Q_\theta(s,a')}{\lambda}\right).$$

For a state–action pair (s,a), let

$$Z = (R, S^+, D) \sim \mathcal{P}(\cdot \mid s,a)$$

denote a joint conditional transition sample, where R is the realized reward, S^+ is the successor state, and $D \in \{0,1\}$ is the true terminal indicator. The single-transition soft Bellman residual is

$$\delta_\theta(s,a,Z) := R + \gamma(1-D)V_{\lambda,\theta}(S^+) - Q_\theta(s,a).$$

Thus the continuation term is masked at a true terminal transition. A time-limit truncation ends the rollout but is not a true terminal transition, and hence uses $D = 0$ in the Bellman target.

Uniform sampling of transitions from the current replay buffer induces an empirical state–action distribution μ_t. For a fixed update, write this distribution as μ and define the conditional mean

residual

$$m_\theta(s,a) := \mathbb{E}[\delta_\theta(s,a,Z) \mid s,a]\,.$$

The replay-weighted deep SCBR objective is

$$J_\mu(\theta) := \frac{1}{2}\mathbb{E}_{(S,A)\sim\mu}\left[m_\theta(S,A)^2\right]. \tag{22}$$

This is the stochastic, replay-distribution-weighted analogue of the model-based SCBR objective. It is not identical to the unweighted finite-state Euclidean objective analyzed in the tabular and linear-function-approximation sections, and μ_t changes as the replay buffer evolves. This distributional difference is another reason that the preceding convergence theory is not claimed for the deep implementation.

It is important to distinguish (22) from the expected squared sample residual

$$\widetilde{J}_\mu(\theta) := \frac{1}{2}\mathbb{E}_{(S,A)\sim\mu}\mathbb{E}\big[\delta_\theta(S,A,Z)^2 \mid S,A\big]\,.$$

By the conditional bias–variance identity,

$$\widetilde{J}_\mu(\theta) = J_\mu(\theta) + \frac{1}{2}\mathbb{E}_{(S,A)\sim\mu}\operatorname{Var}(\delta_\theta(S,A,Z) \mid S,A)\,.$$

Therefore, a single-sample squared-residual update is an unbiased stochastic-gradient update for $\widetilde{J}_\mu$, not generally for J_μ. The double-sampling issue concerns estimating the gradient of the square of the conditional mean in (22); it is not a bias in the exact gradient of a fixed mini-batch squared-error objective.

For each sampled state–action pair, let

$$Z^{(1)}, Z^{(2)} \overset{\text{i.i.d.}}{\sim} \mathcal{P}(\cdot \mid s,a)$$

be two conditionally independent joint transition samples. For a mini-batch B of such paired samples, define the scalar paired-sample estimator

$$\widehat{J}_B(\theta) := \frac{1}{2|B|}\sum_{i\in B}\delta_\theta(s_i,a_i,Z_i^{(1)})\,\delta_\theta(s_i,a_i,Z_i^{(2)}). \tag{23}$$

Conditional independence gives

$$\mathbb{E}\Big[\delta_\theta(s,a,Z^{(1)})\,\delta_\theta(s,a,Z^{(2)}) \mid s,a\Big] = m_\theta(s,a)^2.$$

Consequently, $\widehat{J}_B$ is an unbiased estimator of J_μ, and, under the standard interchange of differentiation and expectation,

$$\mathbb{E}[\nabla_\theta \widehat{J}_B(\theta)] = \nabla_\theta J_\mu(\theta),$$

where automatic differentiation computes the gradient of the same scalar estimator,

$$\nabla_\theta \widehat{J}_B(\theta) = \frac{1}{2|B|}\sum_{i\in B}\left(\delta_i^{(2)}\nabla_\theta\delta_i^{(1)} + \delta_i^{(1)}\nabla_\theta\delta_i^{(2)}\right).$$

No stop-gradient pseudo-loss is required.

The experiments below use the deterministic `CartPole-v1`, `Acrobot-v1`, `Pendulum-v1`, and `MountainCar-v0` environments. Their conditional joint transition laws are point masses, so $Z^{(1)} = Z^{(2)}$ almost surely and the replayed transition can be reused as the exact second conditional sample. In this case (23) reduces to the ordinary scalar loss $\frac{1}{2|B|}\sum_i \delta_i^2$. We exclude `LunarLander-v3`, for which the implementation does not provide an exact second conditional sample. For both SCBR and DQN, the continuous torque interval of `Pendulum-v1` is discretized into five evenly spaced actions; hence the reported Pendulum task is a five-action discretization of the original continuous-action environment. The SCBR behavior policy is Boltzmann with temperature $\tau_\pi = 0.1$, which is distinct from the residual temperature λ, whereas DQN uses fixed ε-greedy exploration with $\varepsilon = 0.1$.

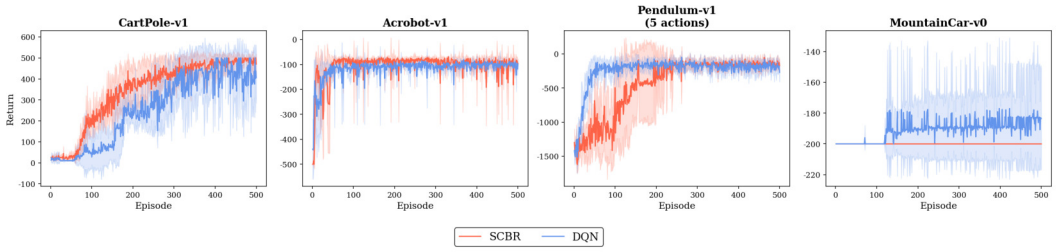


Figure 14: Deep learning experiments across four deterministic tasks from the `Gymnasium` control environment, comparing SCBR (red) and DQN (blue). The `Pendulum-v1` panel uses the five-action torque discretization described in the text. Each solid curve is the mean return over five random seeds for 500 training episodes. Shading shows the mean plus or minus one standard deviation across seeds at each episode, computed using `numpy.std` with `ddof=0`. Over the final 50 episodes, SCBR has the higher mean return on `CartPole-v1` and the discretized `Pendulum-v1`, the two methods are comparable on `Acrobot-v1`, and DQN has the higher mean return on `MountainCar-v0`.

Figure 14 compares the deep RL variant of the SCBR method with the baseline DQN [21]. We conducted each experiment for 500 episodes across five different random seeds. As mentioned above, the SCBR method does not use a target network, whereas DQN does. Averaging the returns over the final 50 episodes and all five seeds, SCBR and DQN obtain, respectively, 487.4 and 421.4 on `CartPole-v1`, -103.8 and -102.5 on `Acrobot-v1`, -146.8 and -182.7 on the five-action discretization of `Pendulum-v1`, and -200.0 and -185.3 on `MountainCar-v0`. Thus, SCBR has the higher late-stage mean return on `CartPole-v1` and the discretized `Pendulum-v1`, while the two methods are nearly tied on `Acrobot-v1`. On `MountainCar-v0`, SCBR remains at -200 in all five seeds; DQN improves in one seed, which raises its aggregate mean but also produces substantial variability. These mixed results do not support a general performance ordering between SCBR and DQN. A more comprehensive comparative analysis would be needed for a rigorous comparison; however, this is not the primary goal of the present paper. Therefore, we do not pursue additional experiments in this paper. Further investigation and experiments remain important directions for future work.

A.8 Additional mathematical results and counterexamples

A.8.1 Meaning and sufficient conditions for oblique regularity

Let $n = |\mathcal{S} \times \mathcal{A}|$, $M_\beta = P\Pi^\beta$, and

$$B_\beta := \Phi^\top (\gamma M_\beta - I)^\top \Phi.$$

Proposition A.10. *The following statements hold.*

1. *B_β is nonsingular if and only if*
$$\mathcal{R}(\Phi) \cap \mathcal{N}\big(((\gamma M_\beta - I)\Phi)^\top\big) = \{0\}.$$
2. *The condition holds when $\Phi = I$, when $\Phi \in \mathbb{R}^{n\times n}$ is invertible, and when Φ is a full-column-rank one-hot state-action aggregation matrix.*
3. *For any induced matrix norm, the condition holds if*
$$\gamma \left\|(\Phi^\top \Phi)^{-1} \Phi^\top M_\beta^\top \Phi\right\| < 1.$$

Proof. For item 1, $B_\beta x = 0$ is equivalent to $\Phi x \in \mathcal{R}(\Phi) \cap \mathcal{N}(((\gamma M_\beta - I)\Phi)^\top)$. Since Φ has full column rank, $x = 0$ if and only if $\Phi x = 0$, which proves the equivalence. For the tabular and invertible-square cases in item 2, $\gamma M_\beta - I$ is nonsingular because M_β is row-stochastic and $\gamma < 1$. In the square case, $B_\beta = \Phi^\top(\gamma M_\beta - I)^\top \Phi$ is therefore a product of nonsingular square matrices. For one-hot aggregation, let

$$D = \Phi^\top \Phi = \mathrm{diag}(d_1, \ldots, d_m), \qquad C = \Phi^\top M_\beta \Phi.$$

Since M_β is row-stochastic and every row of Φ contains exactly one nonzero entry equal to one, $C\mathbf{1} = D\mathbf{1}$. Hence $\gamma C - D$ is strictly row diagonally dominant because, in row i,

$$|d_i - \gamma C_{ii}| - \gamma \sum_{j \neq i} C_{ij} = (1-\gamma)d_i > 0.$$

Thus $B_\beta^\top = \gamma C - D$ is nonsingular. Finally,

$$B_\beta = -(\Phi^\top \Phi)\left[I - \gamma(\Phi^\top \Phi)^{-1}\Phi^\top M_\beta^\top \Phi\right],$$

and item 3 follows from the Neumann-series criterion. □

At a returned pair (θ, β), regularity and conditioning can be diagnosed directly from $\sigma_{\min}(B_\beta)$. A uniform pre-run guarantee can be more expensive because the active policy mixtures depend on the parameter and may form a large set.

Example 2 (Oblique regularity can fail)**.** *Consider one state, two self-loop actions, the deterministic policy selecting action 1, and*

$$M = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \qquad \Phi = \begin{bmatrix} 1 \\ t \end{bmatrix}, \qquad 0 < t < 1.$$

Then Φ has full column rank and, for every $\theta > 0$, action 1 is uniquely greedy. Nevertheless,

$$B = \Phi^\top(\gamma M - I)^\top \Phi = \gamma(1+t) - (1+t^2).$$

For $\gamma = 0.9$ and $t = (0.9 + \sqrt{0.41})/2 \approx 0.770156$, one has $B = 0$. Thus full column rank and a unique greedy policy do not by themselves imply oblique regularity. The set-valued equations in Theorems A.6 and A.12 remain valid in this singular case.

A.8.2 A global CBR minimizer can induce a poor policy

Example 3. *Consider one state, two self-loop actions, rewards $R = (0,1)^\top$, discount $\gamma \in (0,1)$, and the one-dimensional feature matrix*

$$\Phi = \begin{bmatrix} 1 & k \end{bmatrix}^\top, \qquad \gamma < k < 1.$$

For $\theta \geq 0$, action 1 is greedy and

$$f(\theta) = \frac{1}{2}\left[(1-\gamma)^2\theta^2 + \big(1 - (k-\gamma)\theta\big)^2\right].$$

The minimizer of this branch is

$$\theta^* = \frac{k-\gamma}{(1-\gamma)^2 + (k-\gamma)^2} > 0.$$

For $\theta \leq 0$, action 2 is greedy and

$$f(\theta) = \frac{1}{2}\left[(1-\gamma k)^2\theta^2 + \big(1 - k(1-\gamma)\theta\big)^2\right] \geq \frac{1}{2},$$

with equality only at $\theta = 0$. On the positive branch,

$$f(\theta^*) = \frac{(1-\gamma)^2}{2\big((1-\gamma)^2 + (k-\gamma)^2\big)} < \frac{1}{2},$$

so θ^* is the unique global CBR minimizer. Nevertheless, its greedy policy chooses action 1, whereas action 2 is optimal. Taking $k = (1+\gamma)/2$ gives

$$f(\theta^*) = \frac{2}{5}, \qquad \|Q^{\pi_{\theta^*}} - Q^*\|_\infty = \frac{\gamma}{1-\gamma}.$$

Thus, under misspecification, even a unique global CBR minimizer can induce a poor policy, and the $1/(1-\gamma)$ dependence in residual-based control certificates is consequential.

A.8.3 Local geometry of CBR stationary points

Proposition A.11. *Suppose that a deterministic policy π is the unique greedy policy of Q_θ throughout an open neighborhood of $\bar{\theta}$. If $0 \in \partial f(\bar{\theta})$, then $\bar{\theta}$ is a strict local minimizer of the CBR objective.*

Proof. On that neighborhood, $TQ_\theta = T^\pi Q_\theta$ and

$$f(\theta) = \frac{1}{2} \|R + (\gamma P \Pi^\pi - I)\Phi\theta\|_2^2 .$$

Consequently,

$$\nabla^2 f(\theta) = \Phi^\top (\gamma P \Pi^\pi - I)^\top (\gamma P \Pi^\pi - I)\Phi.$$

The matrix $P\Pi^\pi$ is row-stochastic, so $\rho(\gamma P\Pi^\pi) < 1$ and $\gamma P\Pi^\pi - I$ is nonsingular. Since Φ has full column rank, $(\gamma P\Pi^\pi - I)\Phi$ has full column rank and therefore $\nabla^2 f(\theta) \succ 0$. The local quadratic branch is thus strongly convex. At the interior point $\bar{\theta}$, Clarke stationarity is ordinary gradient stationarity, so $\bar{\theta}$ is the unique minimizer of this quadratic branch and hence a strict local minimizer of f. □

Proposition A.12 (Local linear convergence inside one greedy-policy region)**.** *Under the assumptions of Proposition A.11, define*

$$H_\pi := \Phi^\top (\gamma P \Pi^\pi - I)^\top (\gamma P \Pi^\pi - I)\Phi \succ 0.$$

For any fixed step size $0 < \alpha < 2/\lambda_{\max}(H_\pi)$, there exists a neighborhood V of $\bar{\theta}$ such that every generalized-gradient iterate initialized in V remains in the same greedy-policy region and satisfies

$$\|\theta_k - \bar{\theta}\|_2 \le \rho^k \|\theta_0 - \bar{\theta}\|_2, \qquad \rho := \max_i |1 - \alpha\lambda_i(H_\pi)| < 1.$$

Proof. On the unique-policy neighborhood, the minimum-norm Clarke subgradient is the ordinary gradient and

$$\nabla f(\theta) = H_\pi(\theta - \bar{\theta}).$$

Thus $\theta_{k+1} - \bar{\theta} = (I - \alpha H_\pi)(\theta_k - \bar{\theta})$ and $\|I - \alpha H_\pi\|_2 = \rho < 1$. Choose a closed ball centered at $\bar{\theta}$ contained in the unique-policy neighborhood. The contraction keeps every iterate initialized in that ball inside it and gives the displayed rate. □

Example 4 (A tie-boundary stationary point can be a strict local maximum)**.** *Let $0 < \gamma < 1$ and consider one state, two self-loop actions, rewards $R = (-1, -1)^\top$, and*

$$\Phi = \begin{bmatrix} 1 & -1 \end{bmatrix}^\top .$$

Then $Q_\theta = (\theta, -\theta)^\top$ and

$$\begin{aligned} f(\theta) &= \frac{1}{2}\left[\big(-1 + \gamma|\theta| - \theta\big)^2 + \big(-1 + \gamma|\theta| + \theta\big)^2\right] \\ &= 1 - 2\gamma|\theta| + (1 + \gamma^2)\theta^2. \end{aligned}$$

The one-sided derivatives satisfy

$$f'_-(0) = 2\gamma, \qquad f'_+(0) = -2\gamma,$$

so

$$\partial f(0) = [-2\gamma, 2\gamma]$$

and $0 \in \partial f(0)$. However, $f(\theta) < f(0)$ for every nonzero θ satisfying $|\theta| < 2\gamma/(1 + \gamma^2)$. Therefore, $\theta = 0$ is a strict local maximum. This shows that Clarke stationarity need not imply local minimality at a greedy-action tie boundary.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim foundational results for control Bellman residual minimization, including the geometry of CBR, Clarke subdifferential characterization, stationarity via oblique projected Bellman equations, and conditional convergence of generalized-gradient methods, tabular SCBR convergence, and LFA SCBR asymptotic stationarity. These claims match the theoretical results proved in the main text and Appendix. The empirical claims are presented as illustrative examples rather than as the central contribution.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses several limitations and scope restrictions, including the generally nonsmooth/nonconvex nature of CBR, the lack of global optimality guarantees for general linear function approximation, the double-sampling issue for model-free residual minimization, the dependence of some results on regularity assumptions, and the mixed empirical performance of the deep RL variant compared with DQN. The main theoretical guarantees are explicitly scoped to tabular and linear-function-approximation settings.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions used for the theoretical results, including finite discounted MDPs, full-column-rank linear features, compactness or level-set conditions for convergence, and oblique-projection regularity where needed. The Appendix provides proofs for the stated propositions, lemmas, and theorems, with standard external results cited when used.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the MDP definitions, reward and transition specifications, feature matrices, algorithmic updates, step-size rules, random-seed information where applicable, and environment descriptions needed to reproduce the illustrative tabular/LFA examples and the reported Gym-based experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper is primarily theoretical, and most experiments are simple illustrative examples that can be implemented directly from the mathematical definitions and algorithmic descriptions in the manuscript. We do not provide an open-source code release. The experiments use synthetic examples and standard simulation environments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The examples and Appendix specify the relevant experimental settings, including the MDP/environment definitions, feature dimensions and feature choices where used, optimization methods, step-size conditions, number of simulation episodes, maximum episode length, and the random-seed protocol for the deep RL experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper is primarily theoretical, and most experiments are simple, deterministic illustrations of the analysis. We do not provide formal statistical significance tests or confidence intervals for these examples. For the additional deep RL experiments, Figure 14 reports the mean and one standard deviation across five random seeds as descriptive information about variability. These experiments are illustrative, and we do not claim statistically significant performance differences.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The tabular and linear-function-approximation examples are lightweight and likely reproducible on standard CPU hardware, but the manuscript does not explicitly report the compute workers, memory, execution time, or total compute used for each experiment, especially for the deep RL comparisons.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is primarily theoretical and uses standard simulation environments. It does not involve human subjects, private or sensitive data, surveillance, deception, or safety-critical deployment.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is foundational work on Bellman residual minimization and convergence analysis. It does not introduce a deployed system, dataset, or model with a direct path to societal impact beyond general reinforcement-learning research. Therefore, a detailed broader-impact discussion is not applicable.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: The paper does not release high-risk models, scraped datasets, pretrained language models, image generators, or other assets requiring misuse safeguards.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The manuscript identifies the benchmark environments and cites the baseline algorithms, but the accompanying materials do not provide a complete account of the licenses and terms of use for existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce a new dataset, benchmark, released model, or other standalone asset. The algorithms and analysis are described in the manuscript, while the empirical results use standard or synthetic environments.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper involves neither crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper involves neither crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not components of the proposed algorithms or experimental methodology, so this item is not applicable to the core methods. Routine writing, editing, and formatting assistance does not require a methodology declaration. The separate Declaration of LLM usage section describes the manuscript-preparation assistance, including proof and citation checks. The authors remain responsible for all content.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to the NeurIPS 2026 handbook for the LLM policy.