BELLMAN-CONVEX REINFORCEMENT LEARNING IN CONTINUOUS MARKOV DECISION PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning in continuous state and action spaces faces two fundamental challenges: sample complexity grows exponentially with dimension, while planning may require high-dimensional nonconvex optimization over continuous actions. Yet many real-world problems in operations and control possess exploitable convex structure. We introduce *Bellman convexity*, a framework under which Bellman updates preserve convexity and Lipschitz continuity of both Q functions and value functions, capturing a broad family of convex sequential decision making problems arising in robotics, autonomous systems, and energy systems. Under generative-model access, we develop Bellman envelope regression (BER), which adaptively samples proximal anchors of noisy Bellman targets and converts the anchor values into global convex envelopes. Embedding BER in fitted Q-iteration yields fixed-discount sample complexity $\widetilde{O}(\varepsilon^{-2-(d_s+d_a)/2})$ for uniform recovery of $Q^\star$, and we prove a matching lower bound in the ε-accuracy exponent. Counterintuitively, we uncover a sharp separation from standard RL theory: Bellman convexity makes policy learning statistically easier than uniform $Q^\star$ recovery by removing the action dimension from the accuracy exponent. An ε-optimal policy can be learned with $\widetilde{O}(\varepsilon^{-2-d_s/2})$ samples, matching the fixed-discount lower bound. Finally, we show that the deterministic planning steps used by our algorithms are convex programs and, under standard linear or conic descriptions, admit LP, SOCP, or SDP formulations.

Keywords. Reinforcement learning, continuous-state MDPs, convex value functions, minimax-optimal sample complexity.

1 INTRODUCTION

Reinforcement learning provides a general framework for sequential decision-making and has been widely used in resource allocation, financial markets, robotics, and control (Powell, 2011; Moody and Saffell, 2001; Kober et al., 2013). Mathematically, reinforcement learning is commonly modeled through Markov decision processes (MDPs) (Puterman, 1994). Learning and planning in MDPs face a severe curse of dimensionality, as modern applications can involve continuous and extremely high-dimensional state and action spaces. Statistically, generic continuous-space MDP methods require a number of samples that grows exponentially with the problem dimension (Shah and Xie, 2018; Maran et al., 2024). Computationally, generic dynamic programming and greedy-action computation over continuous actions often lead to high-dimensional global optimization problems, which can be intractable without additional structure (Bertsekas, 2017; Yang, 2020). These statistical and computational barriers make high-dimensional continuous RL difficult.

Many real-world decision-making problems can be naturally modeled as continuous MDPs, and many possess exploitable convex structure. In robotic control, linear dynamics with convex costs and constraints lead to tractable formulations, including LQR and convex model predictive control (Bertsekas, 2017; Rawlings et al., 2017). In AI computing systems, workload scheduling is a central sequential decision problem (Banerjee et al., 2020), while dynamic capacity allocation admits structured online optimization formulations (Lin et al., 2011). In these settings, convexity can make control and resource-allocation decisions computationally tractable (Boyd and Vandenberghe, 2004).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, these benefits are usually established separately within application-specific model classes through assumptions on their dynamics, costs, and constraints (Agrawal et al., 2020; Cassel et al., 2022; Lin et al., 2011). These lines of work do not provide a unified MDP-level condition that captures their shared convex structure. It therefore remains unclear how convexity can be exploited directly at the continuous-MDP level to improve computation and, more importantly, whether it can reduce the sample complexity of learning high-dimensional continuous MDPs.

To answer this question, we introduce *Bellman convexity*, a framework under which Bellman updates preserve convexity and Lipschitz continuity of both Q functions and value functions. Building on this framework, we develop a unified learning approach that exploits convexity directly in Bellman updates. Its central statistical component is *Bellman envelope regression*, which adaptively identifies informative proximal anchors and reconstructs Bellman updates as global convex envelopes. This framework yields improved sample complexity for both $Q^\star$ recovery and policy learning, while reducing the associated planning steps to convex optimization problems.

Our contributions can be summarized as follows:

Bellman convexity and Bellman envelope regression. We introduce *Bellman convexity*, a framework under which Bellman updates preserve convexity and Lipschitz continuity of both Q functions and value functions. Building on this structure, we develop Bellman envelope regression (BER), which adaptively samples proximal anchors and reconstructs Bellman targets through convex envelope computation. This yields both tractable convex computation and improved sample efficiency, while ensuring that every fitted function remains bounded, Lipschitz, and convex.

Minimax-optimal $Q^\star$ learning with sample complexity guarantees. We develop BER-FQI by embedding BER into fitted Q-iteration for learning the optimal Q function over continuous state and action spaces. For a fixed discount factor, BER-FQI achieves sample complexity $\widetilde{O}(\varepsilon^{-2-(d_s+d_a)/2})$ for uniform recovery of $Q^\star$, and we prove a matching lower bound in the ε-accuracy exponent. This improves on the generic Lipschitz continuous-MDP policy-learning benchmark $\widetilde{O}(\varepsilon^{-2-(d_s+d_a)})$ under generative-model access (Maran et al., 2024). Thus, Bellman convexity halves the dimensional contribution to the accuracy exponent for uniform Q learning. The technical core combines adaptive search for proximal anchors with a convex envelope construction that controls the range and Lipschitz constant throughout fitted Q-iteration.

Minimax-optimal policy learning with sample complexity guarantees. Perhaps counterintuitively, we show that policy learning under Bellman convexity can be statistically easier than uniform recovery of $Q^\star$. We develop a direct value-learning algorithm that learns an ε-optimal policy without uniformly estimating the full $Q^\star$ function. For a fixed discount factor, the algorithm achieves sample complexity $\widetilde{O}(\varepsilon^{-2-d_s/2})$, together with a matching lower bound in the ε-accuracy exponent. Thus, direct policy learning removes the action dimension from the accuracy exponent, reducing the dimension term from $(d_s + d_a)/2$ to $d_s/2$.

2 Bellman-Convex Continuous MDPs

Consider a discounted continuous MDP $M = (\mathcal{S}, \mathcal{A}, P, c, \gamma)$ in cost-minimization form. The known state space $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ is compact and convex with bounded geometry, and the continuous action space has a compact feasible set $\mathcal{A}(s) \subseteq \mathbb{R}^{d_a}$ at each state s. The resulting feasible state–action space $\mathcal{X} := \{(s, a) : s \in \mathcal{S},\ a \in \mathcal{A}(s)\}$ is known, compact, and convex in its original coordinates. We impose standard measurability conditions and Hausdorff-Lipschitz regularity of the feasible-action sets (Aubin and Frankowska, 2009; Bonnans and Shapiro, 2000): $d_{\mathrm{H}}(\mathcal{A}(s), \mathcal{A}(s')) \le L_A \|s - s'\|_2$ for all $s, s' \in \mathcal{S}$. The transition kernel $P(\cdot \mid s, a)$ and expected one-step cost $c : \mathcal{X} \to [0, 1]$ are unknown, while $\gamma \in [0, 1)$ is the discount factor. A generative-model query at any $(s, a) \in \mathcal{X}$ returns an independent sample (S', C), where $S' \sim P(\cdot \mid s, a)$, $C \in [0, 1]$, and $\mathbb{E}[C \mid s, a] = c(s, a)$.

Let Π denote the policy class. For $\pi \in \Pi$, let V^π and Q^π be its discounted cost functions. Define $V^\star(s) = \inf_{\pi\in\Pi} V^\pi(s)$ and $Q^\star(s, a) = \inf_{\pi\in\Pi} Q^\pi(s, a)$. For any continuous function $q : \mathcal{X} \to [0, \frac{1}{1-\gamma}]$, define the Bellman optimality operator $\mathcal{T}$ on Q functions by

$$(\mathcal{T}q)(s, a) := c(s, a) + \gamma \mathbb{E}_{S'\sim P(\cdot|s,a)} \left[\min_{a' \in \mathcal{A}(S')} q(S', a') \right].$$

The operator $\mathcal{T}$ is a γ-contraction under the sup norm, and therefore has the unique fixed point $Q^\star$ (Puterman, 1994; Bertsekas, 2017). Moreover, $V^\star(s) = \min_{a\in\mathcal{A}(s)} Q^\star(s,a)$, and any policy satisfying $\pi^\star(s) \in \arg\min_{a\in\mathcal{A}(s)} Q^\star(s,a)$ is an optimal policy.

For a value function v, define the one-step target $F_v(s,a) := \mathbb{E}[C + \gamma v(S') \mid s,a]$ and the Bellman optimality operator $\mathcal{B}$ by

$$(\mathcal{B}v)(s) := \min_{a\in\mathcal{A}(s)} F_v(s,a).$$

The operator $\mathcal{B}$ is also a γ-contraction under the sup norm, with unique fixed point $V^\star$.

We denote by $\mathcal{Q}_L$ the class of convex L-Lipschitz functions $q : \mathcal{X} \to [0, (1-\gamma)^{-1}]$, and by $\mathcal{V}_{L_V}$ the class of convex L_V-Lipschitz functions $v : \mathcal{S} \to [0, (1-\gamma)^{-1}]$.

Assumption 2.1 (Q-Bellman Convexity)**.** *There exists a finite constant L such that $\mathcal{T}(\mathcal{Q}_L) \subseteq \mathcal{Q}_L$.*

Assumption 2.2 (V-Bellman Convexity)**.** *There exist finite constants L_V and L_F such that $\mathcal{B}(\mathcal{V}_{L_V}) \subseteq \mathcal{V}_{L_V}$. In addition, F_v is convex and L_F-Lipschitz on $\mathcal{X}$ for every $v \in \mathcal{V}_{L_V}$.*

Together, Assumptions 2.1 and 2.2 formalize Bellman convexity for updates of Q functions and value functions, respectively. In both cases, convexity and Lipschitz continuity are preserved throughout Bellman iteration. We assume the algorithms are supplied with valid upper bounds on the constants appearing in the corresponding conditions.

Examples. The corresponding Bellman-convexity condition holds in many multistage convex models with convex feasible sets and costs and affine dynamics. Standard examples include inventory control, energy storage, resource allocation, workload control, and constrained linear-system control. Formal sufficient conditions and further examples are provided in section J.

3 Minimax-Optimal $Q^\star$ Learning via Bellman Envelope Regression

In this section, we introduce a customized fitted Q-learning algorithm for leveraging Bellman convexity. We first describe the one-step Bellman envelope regression approach, then embed it in fitted Q-iteration, and finally state matching upper and lower sample-complexity bounds.

3.1 Standard Fitted Q-Learning

Fitted Q-iteration learns $Q^\star$ by repeatedly applying approximate Bellman updates. Given a current fitted Q function q, one Bellman update requires estimating the function $f_q = \mathcal{T}q$ over the feasible state–action space $\mathcal{X}$. The standard continuous-space approach is to place a grid over $\mathcal{X}$, sample the generative model at each grid point, and form empirical Bellman labels. Specifically, at a grid point $x_\ell = (s_\ell, a_\ell)$, one sample gives

$$Y_{\ell,i} = C_{\ell,i} + \gamma \min_{a'\in\mathcal{A}(S'_{\ell,i})} q(S'_{\ell,i}, a').$$

This Bellman label $Y_{\ell,i}$ is an unbiased sample of the Bellman target at x_ℓ because $\mathbb{E}[Y_{\ell,i} \mid x_\ell] = (\mathcal{T}q)(x_\ell)$. Averaging these labels estimates $(\mathcal{T}q)(x_\ell)$, and the estimates can be extended from the grid to the whole space using Lipschitz continuity. To obtain one-step uniform error η, this route needs mesh size of order η, hence about $\eta^{-(d_s+d_a)}$ grid points (Maran et al., 2024). This is the continuous analogue of uniform sampling in tabular generative-model methods (Azar et al., 2013), which does not exploit convexity and can be sample inefficient.

3.2 Bellman Envelope Regression

To address this challenge, we introduce Bellman envelope regression (BER) to implement one approximate Bellman update. Under Assumption 2.1, the target $f_q = \mathcal{T}q$ associated with any current fitted Q function $q \in \mathcal{Q}_L$ is convex, bounded, and Lipschitz. BER exploits this structure by adaptively sampling noisy Bellman labels and constructing a global convex approximation $\widehat{q}^+$, which can be used in the next fitted Q-iteration. The procedure consists of three steps.

Algorithm 1 Bellman Envelope Regression, $\mathrm{BER}(q, \eta, \delta_{\mathrm{tot}})$

Require: Fitted Q function $q \in \mathcal{Q}_L$, target accuracy η, failure probability δ_{tot}.
Ensure: Convex, bounded, and Lipschitz estimate $\widehat{q}^{+}$ on $\mathcal{X}$.

```
if L = 0 then
    Fix any feasible point x_0 ∈ X, take n_0 = ⌈2(1 − γ)^{−2} η^{−2} log(2/δ_tot)⌉ fresh samples of Y_q(x_0), and return the constant function equal to the clipped empirical average.
end if
Set the internal proximal scale, center spacing, optimization tolerance, and validation budget according to the schedule in section C.1.
Construct the uniform centers specified in section C.1.
for each uniform center u_j do
    Use samples of the Bellman label to find an approximate minimizer x̂_j of (Tq)(x) + (2μ)^{−1} ‖x − u_j‖_2^2 over x ∈ X.
    Validate x̂_j with fresh samples and form a clipped upper-confidence estimate t_j.
end for
Construct the convex envelope E_B from (10) and set q̂^+ = max{0, E_B − e} with the offset e from section C.1.
return q̂^+.
```

Adaptive proximal anchor search. The first step determines where BER should collect informative Bellman samples. It begins with uniform centers $u_1, \ldots, u_m$, which are used to search for final sampling locations. For each center u_j, BER seeks the proximal anchor

$$x^{\mu}_{j,q} \in \arg\min_{x \in \mathcal{X}} \left\{ (\mathcal{T}q)(x) + \frac{1}{2\mu} \|x - u_j\|_2^2 \right\}.$$

This point identifies the part of the convex Bellman target that supports the envelope near u_j. Since $\mathcal{T}q$ is unknown, BER approximates this minimizer using noisy Bellman labels and a stochastic zeroth-order convex-optimization routine.

Bellman label validation. The second step estimates the Bellman target at each selected anchor. After finding $\widehat{x}_j$, BER draws an independent validation batch at this state–action pair. It uses the empirical average to form an upper-confidence estimate t_j, clipped to $[0, (1-\gamma)^{-1}]$. Each center therefore produces one labeled proximal anchor $(\widehat{x}_j, t_j)$.

Convex Bellman envelope construction. The final step reconstructs the Bellman target over the entire state–action space. BER combines the labeled proximal anchors through a convex envelope that enforces the known value range and Lipschitz bound. After correcting for the anchor-search and validation errors, the procedure returns a function $\widehat{q}^{+} \in \mathcal{Q}_L$ that uniformly approximates f_q. The exact envelope program and parameter choices are provided in section C.1.

Intuitively, the uniform centers provide global coverage, the search for proximal anchors moves them toward informative points, and the convex envelope extends their validated values to the entire state–action space.

3.3 BER-FQI Algorithm

BER-FQI uses the one-step BER module inside fitted Q-iteration. Starting from $Q_0 \equiv 0$, round k applies BER to the Bellman target $f_k = \mathcal{T}Q_k$ and sets the returned envelope to be Q_{k+1}. At a sampled feasible pair (s, a), BER uses the Bellman label

$$Y_k(s,a) = C + \gamma \min_{a' \in \mathcal{A}(S')} Q_k(S', a'), \qquad \mathbb{E}[Y_k(s,a) \mid s, a] = (\mathcal{T}Q_k)(s,a).$$

This identity remains valid for adaptively chosen sampling locations because every sampled pair uses a fresh generative-model draw. After K rounds, the algorithm outputs Q_K and a greedy policy. Algorithm 2 summarizes the full procedure.

Algorithm 2 Uniform $Q^\star$ Learning via BER-FQI

Require: Number of Bellman iterations K, accuracies $(\eta_k)_{k=0}^{K-1}$, failure probabilities $(\delta_k)_{k=0}^{K-1}$.
Ensure: Fitted Q function Q_K and greedy policy $\widehat{\pi}$.
1: Initialize $Q_0 \equiv 0$.
2: **for** $k = 0, \dots, K-1$ **do**
3: Define $\mathsf{Sample}_k(s,a)$ by drawing one fresh generative-model sample (S', C) at feasible state–action pair (s,a) and returning $Y_k(s,a) = C + \gamma \min_{a' \in \mathcal{A}(S')} Q_k(S', a')$.
4: Run $\mathrm{BER}(Q_k, \eta_k, \delta_k)$ using Sample_k as its sampling oracle, and set its output to be Q_{k+1}.
5: **end for**
6: Choose a greedy policy $\widehat{\pi}(s) \in \arg\min_{a \in \mathcal{A}(s)} Q_K(s,a)$.
7: **return** $(Q_K, \widehat{\pi})$.

3.4 Sample Complexity Guarantees

This subsection gives the sample-complexity guarantees for BER-FQI with matching upper and lower bounds on the ε-dependence up to logarithmic terms. We use $\widetilde{O}(\cdot)$ only to suppress logarithmic factors in $1/\varepsilon$, $1/\delta$, and the dimensions.

The theorem below gives the sample complexity of BER-FQI. The constant $C_{\mathrm{rec}}(\gamma)$ collects the one-step BER recovery constants. An explicit expression is given in equation (7).

Theorem 3.1 (Uniform fitted-Q upper bound)**.** *Fix $0 < \varepsilon \le 1/(2(1-\gamma))$, $\delta \in (0, 1/2)$, and an MDP satisfying Assumption 2.1 and the model conditions in section 2. Run BER-FQI for $K = 1$ if $\gamma = 0$, and otherwise for $K = \lceil \log(2/((1-\gamma)\varepsilon))/(-\log\gamma) \rceil$ rounds, with the per-round sample budgets $N_0, \dots, N_{K-1}$ specified in section C. Then, with probability at least $1-\delta$, the output Q_K satisfies $\|Q_K - Q^\star\|_\infty \le \varepsilon$, and the total number of fresh generative-model samples satisfies*

$$N_{\mathrm{env}} = \widetilde{O}\left(C_{\mathrm{rec}}(\gamma) \frac{\varepsilon^{-2-(d_s+d_a)/2}}{\left(1 - \gamma^{(d_s+d_a+4)/(d_s+d_a+6)}\right)^{(d_s+d_a)/2+3}} \right). \tag{1}$$

Discussion. For a fixed discount factor and other problem-dependent constants, Theorem 3.1 gives $N_{\mathrm{env}} = \widetilde{O}(\varepsilon^{-2-(d_s+d_a)/2})$ for uniform $Q^\star$ recovery. This improves on the generic Lipschitz continuous-MDP policy-learning benchmark $\widetilde{O}(\varepsilon^{-2-(d_s+d_a)})$ (Maran et al., 2024) by replacing the dimension term $d_s + d_a$ with $(d_s + d_a)/2$. BER obtains this improvement by finding fewer proximal anchors and building a convex envelope for each Bellman target instead of filling a tensor grid. Thus the theorem realizes within fitted Q-iteration the statistical benefit of convex structure studied in convex-function estimation (Seijo and Sen, 2011; Guntuboyina and Sen, 2013; Kur and Putterman, 2022), even though the target changes across rounds and every fitted function must remain valid for the next Bellman update. Unlike smooth continuous-MDP rates, which rely on high-order Bellman regularity (Maran et al., 2024), this convex rate allows nonsmooth Bellman targets.

We then reveal that the established upper bound is essentially tight.

Theorem 3.2 (Uniform fitted-Q lower bound)**.** *Fix $d_s, d_a \ge 1$, the canonical product space $\mathcal{S} = [-1,1]^{d_s}$ and $\mathcal{A} = [-1,1]^{d_a}$, a discount factor $\gamma \in [0,1)$, and $\delta_0 \in (0, 1/2)$. There is a constant $c_0 > 0$, depending only on the domain geometry, Lipschitz constants, discount factor, and δ_0, but not on ε, such that the following holds for every sufficiently small ε. For every generative-model algorithm with sample budget satisfying*

$$N \le c_0 \varepsilon^{-2-(d_s+d_a)/2} \tag{2}$$

there exists an MDP instance with discount factor γ satisfying both Assumptions 2.1 and 2.2 and the model conditions in section 2 such that, with probability at least δ_0, the output $\widehat{Q}$ satisfies $\|\widehat{Q} - Q^\star\|_\infty \ge \varepsilon$.

The proof is given in section D. Its main idea is to place many separated local convex bumps on a product state–action space. Each bump is constructed by raising a tangent plane. A bump of height ε has radius of order $\sqrt{\varepsilon}$, so convexity permits about $\varepsilon^{-(d_s+d_a)/2}$ independent hard locations. Since the learner may sample state–action pairs adaptively, the proof counts how many samples fall near each

hidden bump rather than relying on a fixed grid. If the total budget is too small, at least one bump receives too little data, and two MDPs that differ only there remain statistically hard to distinguish.

4 Minimax-Optimal Policy Learning

Standard worst-case policy-learning guarantees for tabular and generic Lipschitz continuous MDPs retain the full state–action dependence (Azar et al., 2013; Agarwal et al., 2020; Maran et al., 2024). Counterintuitively, this is unnecessary under Bellman convexity. We show that a near-optimal policy can be learned from a convex approximation to $V^\star$ and a set of state–action anchors, without uniformly estimating $Q^\star$. This reduces the sample complexity from $\widetilde{O}(\varepsilon^{-2-(d_s+d_a)/2})$ to $\widetilde{O}(\varepsilon^{-2-d_s/2})$ samples with a matching lower bound.

4.1 Direct Policy Learning with Fitted Value Functions

Instead of recovering the full $Q^\star$ function, State-BER-VI learns a convex approximation to $V^\star$ together with a collection of state–action anchors. The value approximation records the optimal cost at representative states, while the anchors retain the corresponding actions. Because the centers cover only the state space, their number depends on d_s rather than $d_s + d_a$.

For notational clarity, let G_s and G_a be Lipschitz bounds for F_v in the state and action coordinates, respectively, and define $L_{\mathrm{pol}} := \max\{L_V, G_s + L_A G_a\}$. Under Assumption 2.2, one may take $G_s = G_a = L_F$. We use L_{pol} only to construct the state centers and analyze action extraction. The fitted value function remains L_V-Lipschitz.

Learning the value function. The first step constructs the Bellman target associated with the current value estimate v. At a feasible state–action pair, this target is $F_v(s,a) = \mathbb{E}[C + \gamma v(S') \mid s, a]$, and minimizing it over feasible actions gives

$$(\mathcal{B}v)(s) = \min_{a \in \mathcal{A}(s)} F_v(s,a).$$

The algorithm approximates the state function $\mathcal{B}v$, rather than the entire function F_v over the state–action space. Actions are handled inside the Bellman minimization and are therefore not covered by a statistical action grid.

Finding state–action anchors. At each state anchor, the second step finds an action whose Bellman value is close to $(\mathcal{B}v)(s)$. The uniform centers cover $\mathcal{S} + \mu L_{\mathrm{pol}} \mathbb{B}_2^{d_s}$. For each center u_j, the algorithm searches over the feasible state–action space for

$$(s^\mu_{j,v}, a^\mu_{j,v}) \in \operatorname*{arg\,min}_{(s,a)\in\mathcal{X}} \left\{ F_v(s,a) + \frac{1}{2\mu} \|s - u_j\|_2^2 \right\}.$$

The regularization acts only on the state coordinate, while the action is optimized jointly with the state. After finding an approximate anchor $(\widehat{s}_j, \widehat{a}_j)$, the algorithm uses fresh samples to form an upper-confidence estimate t_j of the Bellman target. These estimates construct the next fitted value function, and the associated actions are retained for convex action extraction.

Convex action extraction. The learned value function does not by itself specify an action. Let $(\widehat{s}_j, \widehat{a}_j, t_j)_{j=1}^m$ be the anchors and validated labels returned by the final iteration. At a state s, $\widehat{\pi}(s)$ is the a-component of an optimizer of

$$\begin{aligned} \min_{a,\alpha,\rho,w,\bar{s},\bar{a}} \quad & \sum_{j=1}^m \alpha_j t_j + \frac{\rho}{1-\gamma} + G_s \|s - \bar{s}\|_2 + G_a \|a - \bar{a}\|_2 \\ \text{s.t.} \quad & a \in \mathcal{A}(s), \quad \alpha \in \mathbb{R}^m_+, \quad \rho \geq 0, \quad \mathbf{1}^\top \alpha + \rho = 1, \\ & w \in \rho\mathcal{X}, \quad (\bar{s}, \bar{a}) = \sum_{j=1}^m \alpha_j (\widehat{s}_j, \widehat{a}_j) + w. \end{aligned}$$

Here $\rho\mathcal{X} = \{\rho x : x \in \mathcal{X}\}$. The program combines the stored anchors and uses the Lipschitz terms to account for their distance from the query state. Its constraints ensure that the returned action

belongs to $\mathcal{A}(s)$. Convex action extraction uses only the stored anchors and requires no additional generative-model samples. It preserves the one-step $O(\eta)$ learning error, as shown in Lemma E.2.

Combining these steps gives STATE-BER-VI: each round updates the fitted value function and stores the state–action anchors and their validated labels, and the final policy solves the convex program for action extraction above. Algorithm 3 summarizes the procedure.

Algorithm 3 Direct Policy Learning via STATE-BER-VI

Require: Number of value-iteration rounds K, accuracies $(\eta_k)_{k=0}^{K-1}$, and failure probabilities $(\delta_k)_{k=0}^{K-1}$.
Ensure: Policy $\widehat{\pi}$.
1: Initialize $v_0 \equiv 0$.
2: **for** $k = 0, \dots, K-1$ **do**
3: At every sampled (s, a), draw a fresh (S', C) and form the label $C + \gamma v_k(S')$.
4: Apply BER to the value update with accuracy η_k and failure probability δ_k.
5: Set the returned fitted value function to v_{k+1} and store its state–action anchors and validated labels.
6: **end for**
7: **return** the policy that, at state s, solves the displayed convex extraction problem using the final anchors and returns its a-component.

4.2 SAMPLE COMPLEXITY GUARANTEES

This subsection gives the sample-complexity guarantees for direct policy learning with matching upper and lower bounds on the ε-dependence up to logarithmic terms. The constant C_{pol} collects the problem-dependent one-step recovery factors defined in Lemma E.2.

Theorem 4.1 (Policy-learning upper bound). *Fix $0 < \varepsilon \le 1/(2(1-\gamma))$, $\delta \in (0, 1/2)$, and an MDP satisfying Assumption 2.2 and the model conditions in section 2. Run* STATE-BER-VI *for $K = 1$ if $\gamma = 0$, and otherwise for $K = \lceil \log(4/((1-\gamma)^2\varepsilon))/(-\log\gamma) \rceil$ rounds, with the per-round sample budgets $N_0, \dots, N_{K-1}$ specified in section E. Then, with probability at least $1-\delta$, the returned policy $\widehat{\pi}$ satisfies $\|V^{\widehat{\pi}} - V^\star\|_\infty \le \varepsilon$, and the total number of fresh generative-model samples satisfies*

$$N_{\mathrm{env}} = \widetilde{O}\left(C_{\mathrm{pol}} \frac{\{(1-\gamma)\varepsilon\}^{-2-d_s/2}}{\left(1 - \gamma^{(d_s+4)/(d_s+6)}\right)^{d_s/2+3}} \right). \tag{3}$$

Discussion. For fixed discount and problem-dependent constants, STATE-BER-VI admits an $\widetilde{O}(\varepsilon^{-2-d_s/2})$ sample complexity, which significantly improves on the generic Lipschitz continuous-MDP policy-learning benchmark $\widetilde{O}(\varepsilon^{-2-(d_s+d_a)})$ (Maran et al., 2024). Moreover, compared with the uniform $Q^\star$ recovery sample complexity $\widetilde{O}(\varepsilon^{-2-(d_s+d_a)/2})$ in Theorem 3.1, we eliminate the d_a component. The gain comes from approximating $V^\star$ over state centers and recovering actions by convex extraction, which avoids an action grid. The action dimension may still affect computation and problem-dependent constants, but not the ε-exponent.

Theorem 4.2 (Matching lower bound for direct policy learning). *Fix $d_s, d_a \ge 1$, the canonical product space $\mathcal{S} = [-1,1]^{d_s}$ and $\mathcal{A} = [-1,1]^{d_a}$, a discount factor $\gamma \in [0,1)$, and $\delta_0 \in (0, 1/2)$. There is a constant $c_0 > 0$, depending only on the domain geometry, Lipschitz constants, discount factor, and δ_0, but not on ε, such that the following holds for every sufficiently small ε. For every generative-model algorithm returning a policy with sample budget satisfying*

$$N \le c_0 \{(1-\gamma)\varepsilon\}^{-2-d_s/2}, \tag{4}$$

there exists an MDP instance with discount factor γ satisfying both Assumptions 2.1 and 2.2 and the model conditions in section 2 such that, with probability at least δ_0, the output $\widehat{\pi}$ satisfies $\|V^{\widehat{\pi}} - V^\star\|_\infty \ge \varepsilon$.

The proofs are in sections E and F. Together, Theorems 4.1 and 4.2 show that policy learning has the minimax-optimal dependence $\varepsilon^{-2-d_s/2}$ for fixed γ. This is better than the $\varepsilon^{-2-(d_s+d_a)/2}$ dependence required for uniform $Q^\star$ recovery. The difference comes from learning one good action at each state instead of uniformly recovering the full $Q^\star$ function.

5 Convex Planning and Solver Forms

This section describes the computational implications of Bellman convexity. The proximal searches in BER have convex expected objectives, and the remaining planning steps in BER-FQI and State-BER-VI are deterministic convex programs. We then identify when these programs admit standard LP, SOCP, or SDP formulations.

5.1 Convexity of the Algorithmic Subproblems

For each uniform center, BER uses stochastic zeroth-order optimization to minimize a convex Bellman target plus a quadratic proximal term. Once the anchors have been obtained, evaluating the fitted function is a finite-dimensional convex program. The next-state action minimizations and the final BER-FQI policy are also computed by minimizing a fitted convex function over a convex action set. In particular,

$$\widehat{\pi}(s) \in \arg\min_{a \in \mathcal{A}(s)} Q_K(s, a).$$

State-BER-VI obtains its policy from the convex program for action extraction in section 4.

Proposition 5.1 (Convexity of the algorithmic subproblems)**.** *Under the corresponding condition in Assumptions 2.1 and 2.2, each search for a proximal anchor has a convex expected objective. For every realization of the sampled data, evaluating a fitted function, computing continuation values or greedy actions in* BER-FQI*, and extracting actions in* State-BER-VI *are convex optimization problems over compact feasible sets.*

5.2 Solver-Ready LP, SOCP, and SDP Formulations

The previous subsection requires no additional assumption beyond the convex structure used by the learning theory. In many robotics, energy-system, and resource-allocation models, the constraints and cost functions also have explicit linear or conic descriptions (Boyd and Vandenberghe, 2004; Rawlings et al., 2017; Low, 2014; Powell, 2011). These descriptions determine which standard solver can be used:

- If the feasible-set constraints and all objective terms are piecewise linear, the subproblems are LPs.
- Convex quadratic terms, Euclidean norms, or second-order-cone constraints give convex QPs or SOCPs. In particular, the Euclidean norm used by BER is second-order-cone representable, so otherwise linear models generally lead to SOCPs rather than LPs.
- Linear matrix inequalities give SDPs.

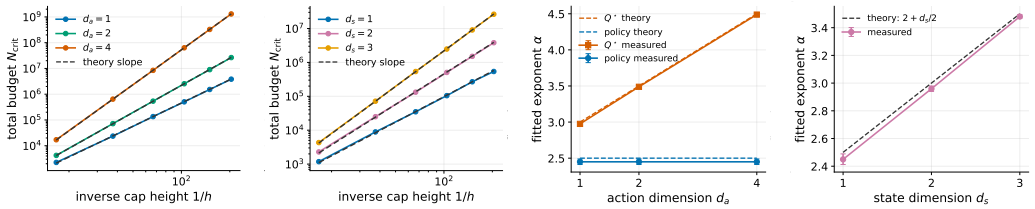
Proposition 5.2 (LP, SOCP, and SDP formulations)**.** *Suppose the feasible sets, costs, and terms defining the fitted functions in BER-FQI or State-BER-VI can be represented by finitely many linear, second-order-cone, or semidefinite constraints. Then every deterministic computation of continuation values, greedy actions, fitted functions, and extracted actions admits the corresponding LP, SOCP, or SDP formulation.*

Thus BER replaces action-grid enumeration with convex optimization over continuous actions. The statistical guarantees treat these deterministic subproblems as solved exactly. Approximate solutions that preserve the required convexity add the optimization error described in section I. Proofs are provided in section H.

6 Numerical Illustration

We report two numerical studies. The first examines the finite hard-family scaling underlying the lower bounds using a single total sampling budget. The second compares the full iterative algorithms in wind-integrated energy-storage control using 2023 EIA-930 data for training and 2024 data for evaluation (U.S. Energy Information Administration, 2026).

Minimax rate study. For each cap height, we instantiate all hidden cap bits jointly. An oracle-aware decoder knows the cap locations and observation laws, but not their bits, and adaptively allocates one total budget toward the locations with greatest posterior uncertainty. Figure 1 reports the critical total budget at which its average bit error falls below 15%.

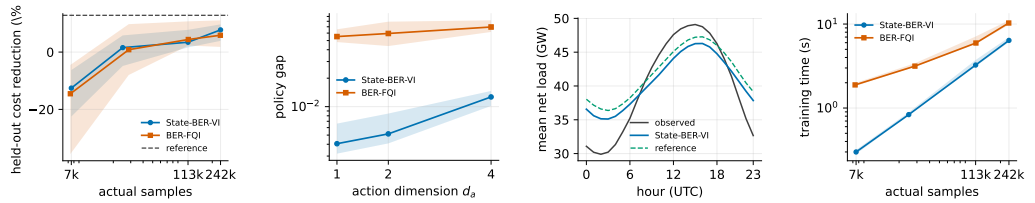


(a) $Q^\star$ recovery (b) Policy learning (c) Action scaling (d) State scaling

Figure 1: Total-budget scaling on the raised-cap hard families. Solid lines report measured critical budgets, with 95% bootstrap intervals, and dashed lines show the theoretical slopes normalized at the geometric center of each measured curve. Panels (c) and (d) compare fitted exponents with theory.

The measured exponents closely track the predicted joint state–action dependence for uniform $Q^\star$ recovery and the state-only dependence for policy learning. This corroborates the finite-family mechanism used by the lower bounds. A separate smooth synthetic-MDP study is reported in section K.

Wind-integrated energy-storage control. We compare State-BER-VI and BER-FQI using battery charge and normalized net load as the state, with battery and flexible-load controls as the action. Figure 2 reports held-out cost reduction, sensitivity to action dimension, 2024 daily operation, and training time. A dense dynamic program is used only for evaluation.



(a) Held-out operation (b) Action dimension (c) Daily operation (d) Training time

Figure 2: Wind-integrated energy-storage control using EIA-930 data. Lines and bands report final-iterate medians and interquartile ranges over independent runs. The dashed line in panel (a) is the dense-program reference.

State-BER-VI improves steadily with the sample budget and remains more accurate than BER-FQI as the action dimension grows, while achieving positive held-out cost reduction. Further settings are provided in section K.

7 Conclusion

This paper develops a finite-sample theory for Bellman-convex continuous MDPs under generative-model access. Bellman envelope regression exploits convexity in Bellman updates, yielding minimax-optimal dependence on ε, for fixed γ, with exponents $2 + (d_s + d_a)/2$ for uniform $Q^\star$ learning and $2 + d_s/2$ for direct policy learning on the product classes used in the lower bounds. The results show that Bellman convexity can improve both statistical scaling and planning: the fitted functions are convex, and the greedy and action-extraction steps are convex programs. We do not claim optimal dependence on $(1-\gamma)^{-1}$. Extending the theory to trajectory data, sharper discount dependence, and broader models is a natural next step.

AI USE STATEMENT

Generative AI tools were used to assist with literature search, experimental code, LaTeX editing, polish proofs, and prose revision. The authors are responsible for reviewing all AI-assisted text, mathematical statements, proofs, and code before submission.

REFERENCES

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed. Springer, 2017.

M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*. Birkhäuser, 2009.

J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer, 2000.

D. Maran, A. M. Metelli, M. Papini, and M. Restelli. Projection by convolution: Optimal sample complexity for reinforcement learning in continuous-space MDPs. In *Proceedings of COLT*, PMLR 247:3743–3774, 2024.

D. Shah and Q. Xie. Q-learning with nearest neighbors. In *Advances in Neural Information Processing Systems 31*, 2018.

E. Seijo and B. Sen. Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3):1633–1657, 2011.

A. Guntuboyina and B. Sen. Covering numbers for convex functions. *IEEE Transactions on Information Theory*, 59(4):1957–1965, 2013.

G. Balázs, A. György, and C. Szepesvári. Near-optimal max-affine estimators for convex regression. In *Proceedings of AISTATS*, PMLR 38:56–64, 2015.

G. Kur and E. Putterman. An efficient minimax optimal estimator for multivariate convex regression. In *Proceedings of COLT*, PMLR 178:1510–1546, 2022.

M. Simchowitz, K. Jamieson, J. W. Suchow, and T. Griffiths. Adaptive sampling for convex regression. arXiv:1808.04523, 2018.

A. Bonito, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and J. W. Siegel. Sampling and reconstruction of convex functions. arXiv:2606.04190, 2026.

I. Yang. A convex optimization approach to dynamic programming in continuous state and action spaces. *Journal of Optimization Theory and Applications*, 187(1):133–157, 2020.

H. Fokkema, D. van der Hoeven, T. Lattimore, and J. J. Mayo. Online Newton method for bandit convex optimisation extended abstract. In *Proceedings of COLT*, PMLR 247:1713–1714, 2024. Full version: arXiv:2406.06506.

M. Kearns and S. Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems 11*, pages 996–1002, 1998.

M. G. Azar, R. Munos, and H. J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91:325–349, 2013.

A. Agarwal, S. M. Kakade, and L. F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Proceedings of COLT*, PMLR 125:67–83, 2020.

A. Sidford, M. Wang, X. Wu, L. F. Yang, and Y. Ye. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems 31*, 2018.

G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems 33*, 2020.

A. Ayoub, Z. Jia, C. Szepesvári, M. Wang, and L. F. Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of ICML*, PMLR 119:463–474, 2020.

D. Zhou, Q. Gu, and C. Szepesvári. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Proceedings of COLT*, PMLR 134:4532–4576, 2021.

C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of COLT*, PMLR 125:2137–2143, 2020.

J. He, H. Zhao, D. Zhou, and Q. Gu. Nearly minimax optimal reinforcement learning for linear Markov decision processes. In *Proceedings of ICML*, PMLR 202:12790–12822, 2023.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

A. Agrawal, S. Barratt, S. Boyd, and B. Stellato. Learning convex optimization control policies. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, PMLR 120:361–373, 2020.

A. B. Cassel, A. Cohen, and T. Koren. Efficient online linear control with stochastic convex costs and unknown dynamics. In *Proceedings of the 35th Conference on Learning Theory*, PMLR 178:3589–3604, 2022.

S. Banerjee, S. Jha, Z. Kalbarczyk, and R. Iyer. Inductive-bias-driven reinforcement learning for efficient schedules in heterogeneous clusters. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:629–641, 2020.

M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska. Dynamic right-sizing for power-proportional data centers. In *Proceedings of IEEE INFOCOM*, pages 1098–1106, 2011.

J. Moody and M. Saffell. Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889, 2001.

H. Scarf. The optimality of (S, s) policies in the dynamic inventory problem. In K. J. Arrow, S. Karlin, and P. Suppes, editors, *Mathematical Methods in the Social Sciences*, pages 196–202. Stanford University Press, 1960.

E. L. Porteus. *Foundations of Stochastic Inventory Theory*. Stanford University Press, 2002.

P. H. Zipkin. *Foundations of Inventory Management*. McGraw-Hill, 2000.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume I, 4th ed. Athena Scientific, 2017.

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

S. H. Low. Convex relaxation of optimal power flow—Part I: Formulations and equivalence. *IEEE Transactions on Control of Network Systems*, 1(1):15–27, 2014.

J. B. Rawlings, D. Q. Mayne, and M. M. Diehl. *Model Predictive Control: Theory, Computation, and Design*, 2nd ed. Nob Hill Publishing, 2017.

W. B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, 2nd ed. Wiley, 2011.

U.S. Energy Information Administration. Hourly Electric Grid Monitor (Form EIA-930). https://www.eia.gov/electricity/gridmonitor/databrowser/, accessed 2026.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

A Appendix Roadmap

The appendix is organized around the main theorems. Each proof first explains its main steps and is completed before the supporting results are proved in dedicated subsections. Short algebraic consequences remain inside the main proof, while only substantive reusable results are stated separately.

Appendix part	Role
Section B	Review of model-based reinforcement learning, convex regression and adaptive recovery, and convex dynamic programming.
Section C	Proof of the BER-FQI upper bound, followed by the one-step BER result and its deterministic and stochastic supporting results.
Section D	Proof of the uniform $Q^\star$ lower bound through a raised-cap construction and adaptive testing argument.
Sections E and F	Separate proofs of the direct policy-learning upper and lower bounds.
Sections H and I	Convex-program formulations and the effect of approximate optimization.
Sections J and K	Sufficient model conditions and numerical protocol details.

B Literature Review

This appendix reviews four lines of work most closely related to our results: model-based reinforcement learning in finite and continuous MDPs, convex regression, and convex dynamic programming.

Model-Based Reinforcement Learning in Finite MDPs. Model-based reinforcement learning separates learning into model estimation and planning. Finite-sample learning and planning in finite MDPs have been extensively studied (Kearns and Singh, 1998; Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2020; Li et al., 2020). Under generative-model access, empirical-model planning attains the minimax sample complexity $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\varepsilon^{-2})$ for learning an ε-optimal policy in discounted finite MDPs (Agarwal et al., 2020; Li et al., 2020). This worst-case rate depends linearly on the number of state–action pairs and therefore does not extend directly to continuous spaces without discretization or additional structure.

Model-Based Reinforcement Learning in Continuous MDPs. For large or continuous state spaces, a related line of structured reinforcement-learning work imposes a finite-dimensional transition representation. Representative examples include linear MDPs (Jin et al., 2020; He et al., 2023) and linear-mixture models learned through value-targeted regression (Ayoub et al., 2020; Zhou et al., 2021). These methods replace tabular dependence by a feature or model dimension. For general nonparametric continuous MDPs, Maran et al. (2024) establish rate-optimal generative-model guarantees under high-order Bellman smoothness. BER instead learns Bellman targets under Bellman convexity, so its complexity is governed by convex geometry without assuming a finite-dimensional transition parameterization or high-order differentiability.

Convex Regression. Multivariate convex regression studies shape-constrained estimation under fixed or random designs and several loss criteria (Seijo and Sen, 2011; Guntuboyina and Sen, 2013; Balazs et al., 2015; Kur and Putterman, 2022). Adaptive sampling and optimal-recovery methods further exploit convex geometry to select informative evaluation points (Simchowitz et al., 2018; Bonito et al., 2026). BER brings these ideas into a sequential Bellman setting: its sampling locations are selected adaptively, its target changes across fitted Q-iterations, and every fitted output must remain convex, bounded, and Lipschitz so that it can be used in the next Bellman update. The

search for proximal anchors uses the high-probability stochastic bandit convex optimization result of Fokkema et al. (2024) as an optimization input, while the envelope construction supplies the global uniform approximation required by Bellman iteration.

Convex Dynamic Programming. Convex dynamic programming exploits affine dynamics, convex costs, and convex constraints to evaluate Bellman updates through convex programs, often on prescribed grids (Yang, 2020). This perspective is central to inventory control, resource allocation, and approximate dynamic programming, where convexity makes planning computationally tractable (Scarf, 1960; Porteus, 2002; Zipkin, 2000; Bertsekas, 2017; Powell, 2011). Our work studies the statistical counterpart with unknown transitions and noisy generative-model samples. It gives upper and lower sample-complexity bounds for learning convex Bellman functions and identifies when direct policy learning is statistically easier than uniform $Q^\star$ recovery.

C Proof of Theorem 3.1

This section proves Theorem 3.1 in four steps. We first show that Bellman updates stay in $\mathcal{Q}_L$, then construct an event on which every one-step recovery succeeds, propagate the resulting errors through fitted iteration, and finally sum the sample budgets. The one-step lemma used by the main argument is stated next and proved after the theorem, followed by its deterministic and stochastic supporting results.

Lemma C.1 (One-step Bellman envelope recovery)**.** *Assume Assumption 2.1. Fix any $q \in \mathcal{Q}_L$, accuracy $\eta \in (0, (1-\gamma)^{-1}]$, and confidence level $\delta_{\mathrm{tot}} \in (0, 1/2)$. Run $\mathrm{BER}(q, \eta, \delta_{\mathrm{tot}})$ using generative-model samples of the Bellman label. There is a universal constant $c_{\log} < \infty$ such that the returned function $\widehat{q}^{+}$ belongs to $\mathcal{Q}_L$ for every sample realization and*

$$\mathbb{P}\big(\big\|\widehat{q}^{+} - \mathcal{T}q\big\|_\infty \le \eta \big) \ge 1 - \delta_{\mathrm{tot}}, \qquad N_{\mathrm{BER}} \le C_{\mathrm{rec}} \eta^{-((d_s+d_a)/2+2)} \left[\log\left(\frac{e C_{\mathrm{rec}}}{\eta \delta_{\mathrm{tot}}} \right) \right]^{c_{\log}}. \quad (5)$$

Here C_{rec} depends only on the fixed geometry, boundedness, Lipschitz, and stochastic convex-optimization constants. An explicit value is given in (7).

Step 1: Show that Bellman updates stay in $\mathcal{Q}_L$.

The zero function belongs to $\mathcal{Q}_L$, and Assumption 2.1 implies inductively that $\mathcal{T}^k 0 \in \mathcal{Q}_L$ for every k. Since $\mathcal{T}$ is a γ-contraction in the sup norm, $\mathcal{T}^k 0 \to Q^\star$ uniformly. Uniform limits preserve the range, convexity, and the common Lipschitz constant, so $Q^\star \in \mathcal{Q}_L$. By Lemma C.1, every fitted iterate Q_k also belongs to $\mathcal{Q}_L$ and can be used in the next BER call.

Step 2: Ensure that every one-step recovery succeeds.

Write $p_Q = d_s + d_a$. For $\gamma = 0$, take $K = 1$ and $\eta_0 = \varepsilon/2$. For $\gamma \in (0,1)$, use the value of K in Theorem 3.1 and define

$$G_K := \sum_{j=0}^{K-1} \gamma^{(p_Q+4)j/(p_Q+6)}, \qquad \eta_{K-1-j} = \frac{\varepsilon \gamma^{-2j/(p_Q+6)}}{2G_K}, \qquad j = 0, \dots, K-1. \quad (6)$$

If an assigned accuracy is at least $(1-\gamma)^{-1}$, the corresponding round may be skipped and $Q_{k+1} = 0$. This only reduces the sample count. At every remaining round, apply Lemma C.1 conditionally on the preceding history with failure probability at most δ/K. A union bound gives one event of probability at least $1-\delta$ on which all sampled rounds satisfy $\|Q_{k+1} - \mathcal{T}Q_k\|_\infty \le \eta_k$.

Step 3: Propagate the one-step errors.

On this event, the one-step recovery error and Bellman contraction give the recursion

$$\|Q_{k+1} - Q^\star\|_\infty \le \eta_k + \gamma \|Q_k - Q^\star\|_\infty.$$

Iterating this inequality and using $\|Q_0 - Q^\star\|_\infty \le (1-\gamma)^{-1}$ yields

$$\|Q_K - Q^\star\|_\infty \le \frac{\gamma^K}{1-\gamma} + \sum_{k=0}^{K-1} \gamma^{K-1-k}\eta_k$$
$$\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2G_K} \sum_{j=0}^{K-1} \gamma^{(p_Q+4)j/(p_Q+6)} \le \varepsilon.$$

The first term is at most $\varepsilon/2$ by the choice of K, and the second equality follows directly from (6). The same calculation is immediate when $\gamma = 0$.

Step 4: Sum the sample budgets.

Let $r_Q = 2 + p_Q/2$. Ignoring only the logarithmic factors already suppressed in Theorem 3.1, Lemma C.1 and (6) give

$$\sum_{k=0}^{K-1} N_k \le \widetilde{O}\left(C_{\mathrm{rec}}(\gamma) \sum_{k=0}^{K-1} \eta_k^{-r_Q}\right)$$
$$\le \widetilde{O}\left(C_{\mathrm{rec}}(\gamma)\, 2^{r_Q} \varepsilon^{-r_Q} G_K^{r_Q+1}\right)$$
$$\le \widetilde{O}\left(C_{\mathrm{rec}}(\gamma) \frac{\varepsilon^{-2-p_Q/2}}{\left(1-\gamma^{(p_Q+4)/(p_Q+6)}\right)^{p_Q/2+3}}\right).$$

The last line uses the geometric-series bound on G_K. Substituting $p_Q = d_s + d_a$ proves (1) and completes the proof. □

C.1 Proof of Lemma C.1

The proof has three steps. We first identify the Bellman target and its sampling law, then combine stochastic proximal search with deterministic envelope recovery, and finally count the required samples.

Step 1: Identify the Bellman target and its samples.

Fix $q \in \mathcal{Q}_L$, and write $f = \mathcal{T}q$, $p = d_s + d_a$, and $B = B_Y = (1-\gamma)^{-1}$. By Assumption 2.1, f is convex, bounded by B, and L-Lipschitz on $\mathcal{X}$. Every sample of the Bellman label satisfies

$$Y_q(s,a) := C + \gamma \min_{a' \in \mathcal{A}(S')} q(S', a'), \qquad \mathbb{E}[Y_q(s,a) \mid s, a, \text{past}] = f(s,a).$$

For the nonconstant case $L > 0$, use $\mu = D/L$. One explicit choice of the constant in Lemma C.1 is

$$C_{\mathrm{rec}} = C_p \left(\frac{D + \mu L}{\sqrt{\mu}}\right)^p \left[\kappa_X^2 p^4 \bar{\Lambda}^2 + B_Y^2\right]. \tag{7}$$

Here κ_X is the geometric factor in the stochastic convex-optimization interface of section C.3, after restricting $\mathcal{X}$ to its affine hull and applying a fixed affine rounding. It is independent of η and δ_{tot}. The rounding and geometric-oracle calls are deterministic computations and are not counted as generative-model samples. When $L = 0$, the constant-target branch uses only $O(B_Y^2 \eta^{-2} \log(1/\delta_{\mathrm{tot}}))$ samples and is absorbed by the same display after enlarging C_{rec}.

If $L = 0$, then f is constant on the convex set $\mathcal{X}$. Hoeffding's inequality for the empirical average at x_0 gives an error at most η with probability at least $1 - \delta_{\mathrm{tot}}$. Clipping keeps the estimate in $[0, B]$ for every sample realization. We henceforth assume $L > 0$.

Step 2: Recover the proximal anchors and the envelope.

By Lemma C.5, the number of uniform centers satisfies

$$M \le C_p (D + \mu L)^p (\mu\eta)^{-p/2}.$$

Apply Proposition C.8 independently at these centers with failure probability δ_{tot}/M. A union bound gives the required proximal gaps and upper-confidence estimates at all centers with probability at least $1-\delta_{\mathrm{tot}}$. On this event, Lemma C.5 gives $\|\widehat{q}^{+}-\mathcal{T}q\|_\infty \le \eta$, while Lemma C.4 ensures that $\widehat{q}^{+} \in \mathcal{Q}_L$ for every sample realization.

Step 3: Count the samples.

With $\beta, \tau \asymp \eta$, multiplying the center count by the per-center bound in (14) yields

$$N_{\mathrm{BER}} \le C_{\mathrm{rec}} \eta^{-2-p/2} \left[\log\left(\frac{eC_{\mathrm{rec}}}{\eta\delta_{\mathrm{tot}}}\right)\right]^{c_{\log}}.$$

This is (5). The confidence and validation factors are absorbed into the displayed logarithm, completing the proof. □

C.2 Deterministic Ingredients for Lemma C.1

This subsection proves the deterministic ingredient of Lemma C.1. The construction is specific to BER, while its analysis uses standard Fenchel duality and proximal-map geometry (Rockafellar, 1970; Bauschke and Combettes, 2017). All statements are relative to the affine hull of the feasible set. If X has positive intrinsic dimension, we identify its affine hull with a Euclidean space and apply a fixed nonsingular affine rounding. In the resulting coordinates, let $X \subset \mathbb{R}^p$ be nonempty, compact, and convex, with

$$\mathbb{B}_2^p \subseteq X \subseteq D\mathbb{B}_2^p. \tag{8}$$

The induced changes in D and the Lipschitz constants are fixed geometry factors. If X is lower dimensional in its original coordinates, the argument uses $p = \dim(\operatorname{aff} X)$; replacing this intrinsic dimension by the ambient dimension only weakens the final bound. A singleton feasible set requires no optimization. Let $f : X \to [0, B]$ be convex and L-Lipschitz. We use the extended-valued function, bounded slope set, and conjugate

$$h := f + I_X, \qquad G_L := \{g \in \mathbb{R}^p : \|g\|_2 \le L\}, \qquad h^*(g) := \sup_{x\in X}\{\langle g, x\rangle - f(x)\}.$$

Lemma C.2 (Bounded-slope Fenchel representation)**.** *For every $z \in X$, the conjugate $h^*(g) = \max_{x\in X}\{\langle g, x\rangle - f(x)\}$ satisfies*

$$f(z) = \sup_{g\in G_L}\{\langle g, z\rangle - h^*(g)\}.$$

Proof. The proof first extends f to the ambient Euclidean space, then identifies the conjugate of this extension, and finally applies Fenchel–Moreau duality.

Step 1: Construct a global extension.

Define the infimal extension

$$\bar{f}(x) := \inf_{y\in X}\{f(y) + L\|x-y\|_2\}, \qquad x \in \mathbb{R}^p.$$

Compactness ensures that the minimum is finite and attained. If $x \in X$, choosing $y = x$ gives $\bar{f}(x) \le f(x)$, while Lipschitzness gives the reverse inequality. The objective is jointly convex in (x, y), so minimizing over y makes $\bar{f}$ convex, and the triangle inequality gives $|\bar{f}(x) - \bar{f}(x')| \le L\|x - x'\|_2$.

Step 2: Identify the conjugate.

Direct calculation and the substitution $v = x - y$ give

$$\bar{f}^*(g) = \sup_{y,v}\{\langle g, y\rangle - h(y) + \langle g, v\rangle - L\|v\|_2\} = h^*(g) + I_{G_L}(g).$$

Indeed, the supremum over v is zero for $\|g\|_2 \le L$ and $+\infty$ otherwise.

Step 3: Apply Fenchel–Moreau.

The extension $\bar{f}$ is finite and continuous, hence proper, closed, and convex. Fenchel–Moreau and Step 2 give

$$f(z) = \bar{f}(z) = \sup_{g \in G_L} \{\langle g, z\rangle - h^*(g)\}, \qquad z \in X,$$

which proves the claim. □

Fix $\mu > 0$. For each center $u \in \mathbb{R}^p$, define

$$\Phi_u(x) := f(x) + \frac{1}{2\mu}\|x - u\|_2^2, \qquad x_u := \arg\min_{x \in X} \Phi_u(x), \qquad g_u := \frac{u - x_u}{\mu} \in \partial h(x_u).$$

Lemma C.3 (Approximate proximal-anchor support)**.** *Fix $g \in G_L$, choose $x_g \in \arg\max_{x \in X}\{\langle g, x\rangle - f(x)\}$, and set $u_g = x_g + \mu g$. Suppose $\|u - u_g\|_2 \le \delta$ and $\widehat{x} \in X$ satisfies $\Phi_u(\widehat{x}) \le \Phi_u(x_u) + \beta$. Then*

$$0 \le h^*(g) - \{\langle g, \widehat{x}\rangle - f(\widehat{x})\} \le \frac{\delta^2}{\mu} + \beta + \delta\sqrt{\frac{2\beta}{\mu}}. \tag{9}$$

Proof. The proof identifies the exact proximal point associated with a target slope, controls the effect of perturbing its center, and then accounts for approximate optimization.

Step 1: Locate the ideal proximal point.

The maximizing property of x_g gives $f(x) \ge f(x_g) + \langle g, x - x_g\rangle$. Expanding the quadratic around $u_g = x_g + \mu g$ yields

$$\Phi_{u_g}(x) - \Phi_{u_g}(x_g) = f(x) - f(x_g) - \langle g, x - x_g\rangle + \frac{\|x - x_g\|_2^2}{2\mu} \ge 0,$$

so x_g is the unique proximal point at center u_g.

Step 2: Control center perturbation.

Subgradient monotonicity and the proximal optimality identities give

$$\langle g_u - g_v, x_u - x_v\rangle \ge 0, \qquad u - v = (x_u - x_v) + \mu(g_u - g_v).$$

Expanding the squared norm therefore yields

$$\|x_u - x_v\|_2^2 + \mu^2\|g_u - g_v\|_2^2 \le \|u - v\|_2^2.$$

With $v = u_g$, both $\|x_u - x_g\|_2 \le \delta$ and $\|g_u - g\|_2 \le \delta/\mu$. The subgradient inequality at x_u then gives

$$0 \le h^*(g) - \{\langle g, x_u\rangle - f(x_u)\} \le \langle g_u - g, x_u - x_g\rangle \le \delta^2/\mu.$$

Step 3: Account for approximate optimization.

Let $d_x = \widehat{x} - x_u$. Expanding the proximal objective gives

$$\Phi_u(\widehat{x}) - \Phi_u(x_u) = f(\widehat{x}) - f(x_u) - \langle g_u, d_x\rangle + \frac{\|d_x\|_2^2}{2\mu}.$$

The first three terms are nonnegative by the subgradient inequality. Hence

$$\|d_x\|_2 \le \sqrt{2\mu\beta}, \qquad 0 \le f(\widehat{x}) - f(x_u) - \langle g_u, d_x\rangle \le \beta.$$

Adding and subtracting the support at x_u, and then applying Step 2, proves the displayed bound. Nonnegativity follows from the definition of h^*. □

Given points $x_1, \ldots, x_M \in X$ and labels $t_j \in [0, B]$, define the bounded Lipschitz convex envelope

$$\begin{aligned} E_B(z) = \min_{\alpha, \rho, w} \quad & t^\top \alpha + \rho B + L \left\| z - \sum_j \alpha_j x_j - w \right\|_2 \\ \text{s.t.} \quad & \alpha \ge 0, \quad \rho \ge 0, \quad \mathbf{1}^\top \alpha + \rho = 1, \quad w \in \rho X. \end{aligned} \tag{10}$$

Lemma C.4 (Bounded Lipschitz convex-envelope recovery). *Assume $t_j \geq f(x_j)$ for all j and, for some $e \geq 0$, $h^*(g) - e \leq \max_j\{\langle g, x_j\rangle - t_j\} \leq h^*(g)$ for every $g \in G_L$. Then E_B is convex, L-Lipschitz, takes values in $[0, B]$, and*

$$f \leq E_B \leq f + e, \qquad \widehat{f} := \max\{0, E_B - e\} \in \mathcal{F}_{B,L}(X), \qquad \|\widehat{f} - f\|_{\infty,X} \leq e.$$

Proof. The proof establishes the convexity and bounds of the envelope, derives its dual representation, and then applies the assumed conjugate bounds.

Step 1: Establish convexity and bounds.

The perspective constraint $w \in \rho X$ is convex, and the objective is jointly convex in z and the decision variables, so minimizing over the decision variables makes E_B convex. Reusing an optimizer at a second point and applying the triangle inequality proves L-Lipschitzness. Finally, every objective term is nonnegative, while $\alpha = 0, \rho = 1, w = z$ is feasible with value B. Hence $0 \leq E_B(z) \leq B$ for every $z \in X$.

Step 2: Dualize the envelope.

Use $L\|v\|_2 = \sup_{\|g\|_2 \leq L}\langle g, v\rangle$. Compact convexity and Sion's minimax theorem (Sion, 1958) permit exchanging the minimum and supremum. Minimizing first over $w \in \rho X$, and then over the convex weights (α, ρ), gives

$$E_B(z) = \sup_{g \in G_L}\left[\langle g, z\rangle - \max\left\{\max_j(\langle g, x_j\rangle - t_j),\, \sigma_X(g) - B\right\}\right],$$

where $\sigma_X(g) = \sup_{x \in X}\langle g, x\rangle$.

Step 3: Use the two-sided conjugate bounds.

Since $t_j \geq f(x_j)$ and $f \leq B$, both terms inside the maximum in Step 2 are at most $h^*(g)$. The assumed lower bound makes the same maximum at least $h^*(g) - e$. Applying Lemma C.2 gives

$$f(z) \leq E_B(z) \leq f(z) + e, \qquad z \in X.$$

Finally, $\widehat{f} = \max\{0, E_B - e\}$ shifts the envelope and clips it at zero. The positive-part map is convex, nondecreasing, and one-Lipschitz, so this adjustment preserves convexity and the Lipschitz constant and gives the claimed uniform error. □

Lemma C.5 (Deterministic convex-envelope recovery). *Let $U_\mu = X + \mu L\mathbb{B}_2^p$, let $\widetilde{U} \subseteq D_{\widetilde{U}}\mathbb{B}_2^p$ be a compact set containing U_μ, and let $\{u_j\}_{j=1}^M$ be a Cartesian δ-net of $\widetilde{U}$. Suppose $\Phi_{u_j}(\widehat{x}_j) \leq \min_X \Phi_{u_j} + \beta$ and $f(\widehat{x}_j) \leq t_j \leq f(\widehat{x}_j) + \tau$ for every j. With $e = \delta^2/\mu + \beta + \delta\sqrt{2\beta/\mu} + \tau$, the adjusted convex envelope satisfies*

$$\|\widehat{f} - f\|_{\infty,X} \leq e, \qquad M \leq \left(1 + \frac{2\sqrt{p}D_{\widetilde{U}}}{\delta}\right)^p.$$

In particular, $\delta^2 = \mu\eta/16$ and $\beta = \tau = \eta/16$ give $e < \eta$ and $M \leq C(p, D_{\widetilde{U}}, \mu)\eta^{-p/2}$.

Proof. The set $\widetilde{U}$ lies in the box $[-D_{\widetilde{U}}, D_{\widetilde{U}}]^p$. A Cartesian grid with coordinate spacing $\delta/\sqrt{p}$ is a Euclidean δ-net, and counting its points gives the displayed cardinality.

Fix $g \in G_L$ and its ideal center $u_g = x_g + \mu g$. Choose u_j within δ of u_g. By Lemma C.3 and $t_j - f(\widehat{x}_j) \leq \tau$,

$$h^*(g) - e \leq \langle g, \widehat{x}_j\rangle - t_j \leq \max_i\{\langle g, \widehat{x}_i\rangle - t_i\} \leq h^*(g),$$

where the last inequality uses $t_i \geq f(\widehat{x}_i)$. Thus the two-sided conjugate bounds in Lemma C.4 hold and give the recovery claim. Under the stated parameter choice,

$$e = \frac{3 + \sqrt{2}}{16}\eta < \eta, \qquad \delta = \sqrt{\mu\eta/16}, \qquad M \leq C(p, D, L, \mu)\eta^{-p/2}.$$

For the direct application $\widetilde{U} = U_\mu$, one may take $D_{\widetilde{U}} = D + \mu L$. When $L > 0$, we use $\mu = D/L$, which gives $((D + \mu L)/\sqrt{\mu})^p = (4DL)^{p/2}$. If $L = 0$, every function in the class is constant on X, so one center suffices. □

C.3 Stochastic Ingredients for Lemma C.1

C.3.1 Stochastic convex-optimization interface

The proximal search uses Theorem 2 in the full version of Fokkema et al. (2024), which gives high-probability regret for a fixed convex loss observed through conditionally centered sub-Gaussian noise. We record only the simple-regret consequence needed below. After an affine change of coordinates, the rounded position satisfies $\mathbb{B}_2^p \subseteq K \subseteq 2(p+1)\mathbb{B}_2^p$. Let κ_K denote the geometric factor in the cited regret bound; its value lies between $p^{-1/2}$ and $p^{-1/4}$ according to the positioning and computational guarantees. An efficient construction of the rounded position uses membership and uniform-sampling oracles, while the number of noisy function evaluations is unchanged.

Corollary C.6 (Stochastic convex-optimization interface)**.** *Let $K \subset \mathbb{R}^p$ be a rounded compact convex body with nonempty interior and let $\ell : K \to [0,1]$ be convex. On round t, the algorithm chooses a feasible point $X_t \in K$ and observes $\ell(X_t) + \xi_t$, where $\mathbb{E}[\xi_t \mid \mathscr{G}_{t-1}, X_t] = 0$ and $\mathbb{E}[\exp(\xi_t^2) \mid \mathscr{G}_{t-1}, X_t] \le 2$ almost surely. There are universal constants $C_{\mathrm{F}}, c_{\log} < \infty$ such that, for every $\alpha, \delta \in (0, 1/2)$, the algorithm returns $\bar{X} \in K$ after*

$$N_{\mathrm{opt}} \le C_{\mathrm{F}} \kappa_K^2 p^4 \alpha^{-2} \left[\log\left(\frac{eC_{\mathrm{F}}\kappa_K p}{\alpha\delta}\right)\right]^{c_{\log}}, \qquad \mathbb{P}\left(\ell(\bar{X}) - \min_{x\in K} \ell(x) \le \alpha\right) \ge 1 - \delta. \tag{11}$$

To derive the corollary, let $X_1, \dots, X_N \in K$ be the feasible iterates of the stochastic algorithm and set $\bar{X} = N^{-1}\sum_t X_t$. Convexity of K keeps $\bar{X}$ feasible, and convexity of ℓ gives

$$\ell(\bar{X}) - \min_{x\in K} \ell(x) \le \frac{1}{N}\sum_{t=1}^{N}\left(\ell(X_t) - \min_{x\in K}\ell(x)\right).$$

Theorem 2 in the full version of Fokkema et al. (2024) bounds the numerator by $C\kappa_K p^2\sqrt{N}$ up to logarithmic factors with probability at least $1-\delta$. Choosing N so that the average regret is at most α gives (11). If K is lower dimensional, we first restrict to its affine hull and then apply the same rounding. If its intrinsic dimension is zero, no stochastic optimization is needed. Proposition C.7 verifies the noise normalization for bounded Bellman labels and adds the independent validation samples used to construct upper-confidence labels.

C.3.2 Regularized stochastic optimization and validation

Proposition C.7 (Regularized stochastic optimization with validation)**.** *Condition on past data under which $f : X \to [0, B]$ and a known regularizer $r : X \to [0, R_r]$ are fixed, $\psi = f + r$ is convex, and sampling at an adaptively selected $x \in X$ returns $Y_f(x) \in [0, B_Y]$ with conditional mean $f(x)$. Let $\Lambda_\psi = \max\{B + R_r, B_Y/\sqrt{\log 2}\}$. For every $\beta, \tau \in (0, B]$ and $\delta' \in (0, 1/2)$, there are universal constants $C, c_{\log} < \infty$ and an algorithm that samples only feasible points and returns $(\widehat{x}, t)$. Let $\mathcal{E}_{\mathrm{opt,val}}$ be the event that $\psi(\widehat{x}) - \min_X \psi \le \beta$, $f(\widehat{x}) \le t \le f(\widehat{x}) + \tau$, and $t \in [0, B]$. Conditionally on these past data,*

$$\mathbb{P}(\mathcal{E}_{\mathrm{opt,val}}) \ge 1 - \delta', \qquad N \le C\kappa_X^2 p^4 \Lambda_\psi^2 \beta^{-2}\left\{\log\left(\frac{eC\kappa_X p \Lambda_\psi}{\beta\delta'}\right)\right\}^{c_{\log}} + CB_Y^2\tau^{-2}\log\frac{4}{\delta'}. \tag{12}$$

Proof. Set $\ell = \psi/\Lambda_\psi$. Then $0 \le \ell \le 1$. If the optimizer selects x for sampling, the normalized observation decomposes as

$$\frac{Y_f(x) + r(x)}{\Lambda_\psi} = \ell(x) + \frac{\xi}{\Lambda_\psi}, \qquad \xi := Y_f(x) - f(x).$$

Conditional centering follows from the conditional mean property of Y_f, and $\exp(\xi^2/\Lambda_\psi^2) \le 2$ because $|\xi| \le B_Y \le \Lambda_\psi\sqrt{\log 2}$. Therefore Corollary C.6, applied conditionally on the past data with

failure probability $\delta'/2$ and accuracy β/Λ_ψ, returns a feasible $\widehat{x}$ satisfying $\psi(\widehat{x}) - \min_X \psi \le \beta$ with conditional probability at least $1 - \delta'/2$.

It remains to produce an upper-confidence label for f. Take $n_{\mathrm{val}} := \lceil 2B_Y^2 \tau^{-2} \log(4/\delta') \rceil$ fresh samples at $\widehat{x}$, and let $\bar{Y}$ be their average. Hoeffding's inequality, conditioned on the stochastic-optimization history and on $\widehat{x}$, gives

$$\mathbb{P}\Big(|\bar{Y} - f(\widehat{x})| \le \frac{\tau}{2} \,\Big|\, \widehat{x},\ \text{optimization history} \Big) \ge 1 - \frac{\delta'}{2}.$$

Set $t = \min\{B, \bar{Y} + \tau/2\}$. On the validation event, both B and $\bar{Y} + \tau/2$ dominate $f(\widehat{x})$, while clipping preserves the upper bound $t \le f(\widehat{x}) + \tau$. A union bound and the sample count from Corollary C.6 complete the proof. □

C.3.3 Application to proximal Bellman targets

Fix a target $f : X \to [0, B]$ and a center u in the explicit center box $\mathcal{U}_{\mathrm{box}} = [-D - \mu L, D + \mu L]^p$. Define $\Phi_u(z) = f(z) + (2\mu)^{-1}\|z - u\|_2^2$. For the stochastic sampling step and the uniform sample bound, use

$$\begin{aligned} R_u &:= \sup_{z \in X} \|z - u\|_2, \qquad \Lambda_u := \max\left\{ B + \frac{R_u^2}{2\mu}, \frac{B_Y}{\sqrt{\log 2}} \right\}, \\ \bar{R} &:= \sup_{\substack{z \in X \\ u \in \mathcal{U}_{\mathrm{box}}}} \|z - u\|_2, \qquad \bar{\Lambda} := \max\left\{ B + \frac{\bar{R}^2}{2\mu}, \frac{B_Y}{\sqrt{\log 2}} \right\}. \end{aligned} \tag{13}$$

Then $\Lambda_u \le \bar{\Lambda} < \infty$ for every center used by the algorithm.

Proposition C.8 (Search and validation of proximal anchors). *Fix u, accuracy parameters $\beta, \tau \in (0, B]$, and failure probability $\delta' \in (0, 1/2)$. Condition on any past data for which f is fixed and can be sampled through Sample_f. There are universal constants $C, c_{\log} < \infty$ such that the one-center search-and-validation procedure used in* BER *returns $(\widehat{x}, t)$ with conditional probability at least $1 - \delta'$ and*

$$\begin{aligned} &\Phi_u(\widehat{x}) - \Phi_u^\star \le \beta, \qquad f(\widehat{x}) \le t \le f(\widehat{x}) + \tau, \qquad t \in [0, B], \\ &N_{\mathrm{prox}} \le C\left[\kappa_X^2 p^4 \bar{\Lambda}^2 \beta^{-2} \left\{ \log\left(\frac{eC\kappa_X p \bar{\Lambda}}{\beta \delta'} \right) \right\}^{c_{\log}} + B_Y^2 \tau^{-2} \log \frac{4}{\delta'} \right]. \end{aligned} \tag{14}$$

Proof. Apply Proposition C.7 with $r_u(z) := (2\mu)^{-1}\|z - u\|_2^2$, $\psi = f + r_u = \Phi_u$, and $\Lambda_\psi = \Lambda_u \le \bar{\Lambda}$. To verify the external noise condition directly, put $\ell_u = \Phi_u/\Lambda_u$ and $\xi = Y - f$. Then

$$0 \le \ell_u \le 1, \qquad \frac{Y_f(z) + r_u(z)}{\Lambda_u} = \ell_u(z) + \frac{\xi}{\Lambda_u}, \qquad \frac{|\xi|}{\Lambda_u} \le \sqrt{\log 2}.$$

The noise is conditionally centered and satisfies the required exponential-moment bound. The returned point therefore has the desired proximal gap, while the validation part of Proposition C.7 gives the clipped upper-confidence estimate. Replacing Λ_u by $\bar{\Lambda}$ in the sample bound proves (14). □

Remark C.9 (Why the Bellman noise fits exactly). *Within one proximal run, the sampling locations are adaptive. Nevertheless, a fresh generative-model sample at the selected state–action pair (s, a) has conditional mean $(\mathcal{T}Q)(s, a)$ and is bounded in $[0, (1 - \gamma)^{-1}]$. Therefore the centered error satisfies the filtration assumptions of Corollary C.6. The deterministic quadratic regularizer changes the mean objective but not the noise.*

Theorem C.10 (Adaptive convex-function recovery by BER). *Let $X \subset \mathbb{R}^p$ be a known compact convex set satisfying (8), let $f : X \to [0, B]$ be convex and L-Lipschitz, and fix the BER proximal parameter $\mu > 0$. Suppose that, whenever the algorithm selects $x \in X$, it receives a fresh bounded observation $Y(x) \in [0, B_Y]$ satisfying $\mathbb{E}[Y(x) \mid \mathscr{H}, x] = f(x)$. For every $\eta \in (0, B]$ and $\delta \in (0, 1/2)$, the BER envelope construction always returns $\widehat{f} \in \mathcal{F}_{B,L}(X)$, where $\mathcal{F}_{B,L}(X)$ is the class of convex L-Lipschitz functions from X to $[0, B]$, and there are constants $C_{\mathrm{cvx}}, c_{\log} < \infty$, depending only on p, D, L, B, B_Y, μ and κ_X, such that*

$$\mathbb{P}\left(\left\| \widehat{f} - f \right\|_{\infty, X} \le \eta \right) \ge 1 - \delta, \qquad N \le C_{\mathrm{cvx}} \eta^{-2-p/2} \left[\log\left(\frac{eC_{\mathrm{cvx}}}{\eta\delta} \right) \right]^{c_{\log}}. \tag{15}$$

Proof. If $L = 0$, the target is constant on the convex set X. Averaging fresh observations at one feasible point and clipping gives the claim after enlarging C_{cvx}. Assume $L > 0$. For $L > 0$, the proof chooses the center resolution, recovers and validates all proximal anchors on one event, and then constructs the envelope and counts samples.

Step 1: Choose the center resolution.

Use the center net from Lemma C.5 with $\delta_{\mathrm{net}}^2 = \mu\eta/16$ and $\beta = \tau = \eta/16$. The deterministic center count is $M = O(\eta^{-p/2})$.

Step 2: Recover and validate the proximal anchors.

Apply Proposition C.8 at each center with failure probability δ/M. A union bound gives, simultaneously over all centers,

$$\Phi_{u_j}(\widehat{x}_j) - \min_{x \in X} \Phi_{u_j}(x) \le \beta, \qquad f(\widehat{x}_j) \le t_j \le f(\widehat{x}_j) + \tau.$$

Step 3: Construct the envelope and count samples.

On the preceding event, Lemma C.5 gives $\|\widehat{f} - f\|_{\infty,X} \le \eta$, and Lemma C.4 ensures that the output belongs to $\mathcal{F}_{B,L}(X)$ for every sample realization. Each center uses $O(\eta^{-2})$ samples up to the logarithmic factor in (14). Therefore

$$N \le C_{\mathrm{cvx}} \eta^{-2-p/2} \left[\log\left(\frac{e C_{\mathrm{cvx}}}{\eta\delta} \right) \right]^{c_{\log}},$$

which is (15). □

C.4 Greedy-policy consequence

Let $\xi \ge 0$ and suppose $\widehat{\pi}$ is ξ-greedy for Q_K, namely $Q_K(s, \widehat{\pi}(s)) \le \inf_{a \in \mathcal{A}(s)} Q_K(s, a) + \xi$ for every state s. If $\|Q_K - Q^\star\|_\infty \le \varepsilon$, then the standard performance argument (Puterman, 1994; Bertsekas, 2017) gives, for every initial distribution ρ,

$$J_\rho(\widehat{\pi}) - J_\rho^\star \le \frac{2\varepsilon + \xi}{1 - \gamma}. \tag{16}$$

Indeed, let $\pi^\star$ be an optimal policy. Greediness of $\widehat{\pi}$ for Q_K gives, for every state s,

$$\begin{aligned} Q^\star(s, \widehat{\pi}(s)) - V^\star(s) &\le |Q^\star(s, \widehat{\pi}(s)) - Q_K(s, \widehat{\pi}(s))| \\ &\quad + Q_K(s, \widehat{\pi}(s)) - Q_K(s, \pi^\star(s)) \\ &\quad + |Q_K(s, \pi^\star(s)) - Q^\star(s, \pi^\star(s))| \\ &\le 2\varepsilon + \xi. \end{aligned}$$

The discounted performance-difference identity for costs yields (16).

D Proof of Theorem 3.2

This section proves Theorem 3.2. The raised-cap family and its Bellman-convex MDP embedding are specific to our setting, while the testing step uses standard adaptive change-of-measure and Assouad tools (Kaufmann et al., 2016; Tsybakov, 2009). The proof follows five steps: construct many local convex alternatives, embed them as Bellman-convex MDPs, bound the information gathered by adaptive sampling, decode the alternatives from a uniformly accurate estimate, and verify that the structural constants remain fixed.

Step 1: Construct a product-space cap family.

Let $\mathcal{S} = [-1, 1]^{d_s}$, $\mathcal{A} = [-1, 1]^{d_a}$, and choose an interior action a_0. Fix compact cubes $D_s \Subset (-1, 1)^{d_s}$ and $D_a \Subset (-1, 1)^{d_a}$ whose action coordinates are separated from a_0 by a positive constant. For a fixed $\lambda > 0$ and cap height $h > 0$, define

$$\phi_0(s, a) := \lambda\big(\|s\|_2^2 + \|a - a_0\|_2^2\big), \qquad r := \sqrt{h/\lambda}.$$

Choose centers $z_j = (s_j, a_j) \in D_s \times D_a$, $j = 1, \dots, m$, at pairwise distance at least $6r$. Standard packing gives

$$m \geq c\, r^{-(d_s+d_a)} = c\, h^{-(d_s+d_a)/2}.$$

Let ℓ_j be the tangent plane to ϕ_0 at z_j, raised by h, and define the hard family by

$$\ell_j(z) := \phi_0(z_j) + \langle \nabla\phi_0(z_j), z - z_j \rangle + h,$$
$$f_\theta(z) := \max\{\phi_0(z), \ell_j(z) : \theta_j = 1\}, \qquad \theta \in \{0,1\}^m.$$

The quadratic identity

$$\phi_0(z) - \ell_j(z) = \lambda \|z - z_j\|_2^2 - h$$

shows that the jth cap is active exactly on $U_j = B(z_j, r)$. These active balls are disjoint, neighboring hypotheses agree outside U_j, and their values at z_j differ by h. For sufficiently small h, all caps are inactive on the action slice $\{(s, a_0) : s \in \mathcal{S}\}$. Consequently, $\min_a f_\theta(s, a) = \lambda \|s\|_2^2$ for every θ.

Step 2: Embed the family as MDPs.

Choose a fixed $\alpha > 0$ small enough that all Bernoulli means below lie in $[1/4, 1/2]$, and use a deterministic self-loop transition. The resulting MDP satisfies

$$p_\theta(s, a) = \frac{1}{4} + \alpha f_\theta(s, a), \qquad S' = s,$$
$$V_\theta^\star(s) = \frac{1/4 + \alpha\lambda\|s\|_2^2}{1-\gamma}, \quad Q_\theta^\star(s, a) = p_\theta(s, a) + \gamma V_\theta^\star(s).$$

The value function is common to all hypotheses. Hence two hypotheses differing only in bit j have identical $Q^\star$ functions outside U_j, differ by at most αh everywhere, and differ by exactly αh at z_j.

Step 3: Control information under adaptive sampling.

Let P_θ be the distribution of all data generated by an arbitrary generative-model algorithm, and let $N_j := \sum_{t=1}^N \mathbf{1}\{(S_t, A_t) \in U_j\}$ be the adaptive number of samples in the jth cap. The adaptive KL change-of-measure identity (cf. Kaufmann et al., 2016, Lemma 1) and the scalar Bernoulli bound on $[1/4, 1/2]$ give

$$\mathrm{KL}(P_\theta \| P_{\theta^{(j)}}) \leq C\alpha^2 h^2\, \mathbb{E}_\theta N_j.$$

For bit j, mix these distributions uniformly over hypotheses with $\theta_j = 0$ and $\theta_j = 1$, and denote the resulting distributions by $\bar{P}_{j,0}$ and $\bar{P}_{j,1}$. Binary testing, Pinsker's inequality, and the Assouad reduction (Tsybakov, 2009, Chapter 2) imply that every decoder $\widehat{\theta}$ satisfies

$$\mathbb{E}\,\mathrm{Ham}(\widehat{\theta}, \theta) \geq \frac{1}{2}\sum_{j=1}^m \big(1 - \mathrm{TV}(\bar{P}_{j,0}, \bar{P}_{j,1})\big)$$
$$\geq \frac{m}{2}\left(1 - \sqrt{\frac{C\alpha^2 h^2 N}{m}}\right).$$

Here we used $\sum_j N_j \leq N$, which holds for every sample sequence because the cap regions are disjoint. Taking the leading constant in $N \leq c_{\delta_0} m h^{-2}$ sufficiently small makes the last display at least $\delta_0 m$. Since $\mathrm{Ham}(\widehat{\theta}, \theta) \leq m$,

$$\mathbb{P}(\widehat{\theta} \neq \theta) \geq \frac{\mathbb{E}\,\mathrm{Ham}(\widehat{\theta}, \theta)}{m} \geq \delta_0$$

under the uniform mixture, and therefore for at least one hypothesis.

Step 4: Convert testing error into uniform $Q^\star$ error.

Given $\widehat{Q}$, decode bit j by thresholding $\widehat{Q}(z_j)$ at the midpoint of the two possible values of $Q_\theta^\star(z_j)$. Uniform error strictly below $\alpha h/2$ recovers every bit, so the preceding testing bound yields, for at least one θ,

$$\mathbb{P}_\theta\Big(\|\widehat{Q} - Q_\theta^\star\|_\infty \geq \alpha h/2\Big) \geq \delta_0 \quad \text{whenever} \quad N \leq c_{\delta_0} h^{-2-(d_s+d_a)/2}.$$

Set $h = 2\varepsilon/\alpha$. For every sufficiently small ε, the construction is valid and

$$N \le c_{\delta_0} h^{-2-(d_s+d_a)/2} = c_0 \varepsilon^{-2-(d_s+d_a)/2},$$

which proves the statistical conclusion in Theorem 3.2.

Step 5: Verify membership in the model class.

The state–action space is a known compact product convex set with the same feasible action set at every state. Every f_θ is a maximum of convex functions and has range and Lipschitz constants bounded uniformly over all sufficiently small h and all θ. For any $q \in \mathcal{Q}_L$, minimizing over this convex action set preserves convexity in s, and

$$(\mathcal{T}_\theta q)(s,a) = p_\theta(s,a) + \gamma \min_{a' \in \mathcal{A}} q(s,a')$$

is convex with Lipschitz constant at most $L_p + \gamma L$. Thus $L \ge L_p/(1-\gamma)$ ensures $\mathcal{T}_\theta(\mathcal{Q}_L) \subseteq \mathcal{Q}_L$, while the target used to update value functions is $F_v(s,a) = p_\theta(s,a) + \gamma v(s)$. For the state operator, the caps are inactive on the slice $a = a_0$, so

$$(\mathcal{B}_\theta v)(s) = \min_{a \in \mathcal{A}} F_v(s,a) = \frac{1}{4} + \alpha\lambda\|s\|_2^2 + \gamma v(s).$$

The right-hand side is convex whenever v is convex. Let L_0 be the Lipschitz constant of $s \mapsto \frac{1}{4} + \alpha\lambda\|s\|_2^2$ on $\mathcal{S}$. Choosing $L_V \ge L_0/(1-\gamma)$ gives $\mathrm{Lip}(\mathcal{B}_\theta v) \le L_0 + \gamma L_V \le L_V$. Moreover, $F_v = p_\theta + \gamma v$ is jointly convex, with state and action Lipschitz constants bounded uniformly over h and θ. Thus the whole hard family satisfies both Assumptions 2.1 and 2.2 and the model conditions in section 2, completing the proof.

E Proof of Theorem 4.1

Throughout this section, write $B = (1-\gamma)^{-1}$. For $v \in \mathcal{V}_{L_V}$, recall the Bellman target F_v and its state minimum $\mathcal{B}v$, and define the Bellman update under policy π by $\mathcal{B}^\pi v(s) := F_v(s, \pi(s))$. The proof uses two ingredients. The first keeps the exact update $\mathcal{B}v$ inside $\mathcal{V}_{L_V}$. The second gives one-step statistical recovery together with a policy whose Bellman residual is controlled. We then construct an event on which every round succeeds, propagate the fitted value error, convert the final Bellman residual into policy performance, and sum the sample budgets.

Lemma E.1 (Bellman properties on $\mathcal{V}_{L_V}$)**.** *The operator $\mathcal{B}$ maps $\mathcal{V}_{L_V}$ into itself, is monotone, and satisfies*

$$\|\mathcal{B}v - \mathcal{B}w\|_\infty \le \gamma \|v - w\|_\infty, \qquad v, w \in \mathcal{V}_{L_V}.$$

Lemma E.2 (One-step value recovery and convex action extraction)**.** *Under the model conditions in section 2 and Assumption 2.2, the BER subroutine for value functions applied to $v \in \mathcal{V}_{L_V}$ returns $\widehat{g} \in \mathcal{V}_{L_V}$ and a policy $\widehat{\pi}$. Let $U_v(s)$ denote the optimal value of the corresponding program for action extraction. There are constants $C_{\mathrm{pol}}, c_{\log} < \infty$, depending on the fixed geometry and on $L_{\mathrm{pol}} := \max\{L_V, G_s + L_A G_a\}$, such that, with probability at least $1-\delta$,*

$$\begin{gathered} \mathcal{B}v - \eta \le \widehat{g} \le \mathcal{B}v, \qquad \mathcal{B}^{\widehat{\pi}} v \le \mathcal{B}v + \eta, \\ 0 \le F_v(s, \widehat{\pi}(s)) - (\mathcal{B}v)(s) \le U_v(s) - \widehat{g}(s) \le 2\eta, \qquad s \in \mathcal{S}, \\ N_{\mathrm{state}} \le C_{\mathrm{pol}} \eta^{-2-d_s/2} \left[\log\left(\frac{e C_{\mathrm{pol}}}{\eta\delta} \right) \right]^{c_{\log}}. \end{gathered}$$

All functional inequalities are pointwise on $\mathcal{S}$.

Step 1: Ensure that every one-step update succeeds.

Set $\varepsilon_V = (1-\gamma)\varepsilon/2$. For $\gamma = 0$, take $K = 1$ and $\eta_0 = \varepsilon_V/2$. For $\gamma \in (0,1)$, the value of K in Theorem 4.1 ensures $\gamma^K B \le \varepsilon_V/2$. Define the normalization and schedule by

$$G_K^V := \sum_{j=0}^{K-1} \gamma^{(d_s+4)j/(d_s+6)}, \qquad \eta_{K-1-j} = \frac{\varepsilon_V \gamma^{-2j/(d_s+6)}}{2 G_K^V}, \qquad j = 0, \dots, K-1.$$

Starting from $v_0 = 0$, Lemma E.1 and Lemma E.2 ensure that every fitted iterate belongs to $\mathcal{V}_{L_V}$ and can be used in the next round. Apply Lemma E.2 conditionally at each round with failure probabilities summing to at most δ. A union bound gives an event of probability at least $1 - \delta$ on which every one-step conclusion holds.

Step 2: Propagate the fitted value error.

On this event, set $E_0 = B$. If $0 \le V^\star - v_k \le E_k$, monotonicity and contraction from Lemma E.1, together with the one-step envelope bound, give

$$0 \le V^\star - v_{k+1} \le V^\star - \mathcal{B}v_k + \eta_k \le \gamma E_k + \eta_k.$$

Consequently,

$$E_K \le \gamma^K B + \sum_{k=0}^{K-1} \gamma^{K-1-k}\eta_k \le \varepsilon_V.$$

Step 3: Convert the final Bellman residual into policy performance.

For the policy π_{k+1} returned at round k, monotonicity also gives

$$\mathcal{B}^{\pi_{k+1}}V^\star \le \mathcal{B}^{\pi_{k+1}}v_k + \gamma E_k \le V^\star + \eta_k + \gamma E_k.$$

Iterating the policy Bellman contraction therefore yields

$$V^{\pi_{k+1}} - V^\star \le \frac{\eta_k + \gamma E_k}{1-\gamma}\mathbf{1}.$$

At the last round, the numerator is at most E_K, and hence

$$\|V^{\pi_K} - V^\star\|_\infty \le \frac{\varepsilon_V}{1-\gamma} \le \varepsilon.$$

Step 4: Sum the sample budgets.

The schedule calculation from the proof of Theorem 3.1, now with dimension d_s and target ε_V, gives

$$\sum_{k=0}^{K-1} \eta_k^{-2-d_s/2} \le C\frac{\{(1-\gamma)\varepsilon\}^{-2-d_s/2}}{\left(1-\gamma^{(d_s+4)/(d_s+6)}\right)^{d_s/2+3}}.$$

Multiplying by the one-step constant C_{pol} from Lemma E.2 and absorbing the confidence and accuracy logarithms into $\widetilde{O}$ proves (3).

E.1 Proof of Lemma E.1

The inclusion $\mathcal{B}v \in \mathcal{V}_{L_V}$ is part of Assumption 2.2. Monotonicity follows directly from the definition of F_v and minimization over actions. Finally,

$$\|\mathcal{B}v - \mathcal{B}w\|_\infty \le \|F_v - F_w\|_\infty \le \gamma\|v - w\|_\infty,$$

which proves the claim.

E.2 Proof of Lemma E.2

Fix $v \in \mathcal{V}_{L_V}$, write $f = F_v$, $g = \mathcal{B}v$, and $L_{\mathrm{pol}} = \max\{L_V, G_s + L_A G_a\}$. In the rounded state-space coordinates, let D_s satisfy $\mathcal{S} \subseteq D_s\mathbb{B}_2^{d_s}$. When $L_{\mathrm{pol}} > 0$, set $\mu = D_s/L_{\mathrm{pol}}$; when $L_{\mathrm{pol}} = 0$, fix $\mu = 1$. For a state center u, the joint and state-only proximal objectives are

$$\Psi_u(s,a) := f(s,a) + \frac{1}{2\mu}\|s-u\|_2^2, \qquad \Phi_u(s) := g(s) + \frac{1}{2\mu}\|s-u\|_2^2.$$

The proof finds and validates joint anchors, uses them to fit the value function, extracts a feasible action, and finally counts samples.

Step 1: Find and validate state–action anchors.

The objective Ψ_u is convex on $\mathcal{X}$. A fresh generative-model sample at an adaptively selected pair is $Y = C + \gamma v(S') \in [0, B]$, with $\mathbb{E}[Y \mid s, a, \text{past}] = f(s, a)$. Hence Proposition C.7, applied with the state-only quadratic regularizer, returns a feasible pair $(\widehat{s}, \widehat{a})$ and an independently validated label t such that

$$\Psi_u(\widehat{s}, \widehat{a}) \le \inf_{(s,a)\in\mathcal{X}} \Psi_u(s, a) + \beta, \qquad f(\widehat{s}, \widehat{a}) \le t \le f(\widehat{s}, \widehat{a}) + \tau.$$

Its sample cost is $\widetilde{O}(\beta^{-2} + \tau^{-2})$, with the fixed state–action geometry and stochastic-optimization factors absorbed into C_{pol}.

Because the regularizer does not depend on a, $\inf_{\mathcal{X}} \Psi_u = \inf_{\mathcal{S}} \Phi_u$. Moreover, $g(\widehat{s}) \le f(\widehat{s}, \widehat{a})$, and therefore

$$\Phi_u(\widehat{s}) \le \inf_{\mathcal{S}} \Phi_u + \beta, \qquad g(\widehat{s}) \le t \le g(\widehat{s}) + \beta + \tau.$$

Thus every state center produces both an approximate state proximal point and a compatible action anchor.

Step 2: Fit the value function.

Use a δ_{net}-net of $\mathcal{S} + \mu L_{\mathrm{pol}} \mathbb{B}_2^{d_s}$ and run Step 1 independently at each center with $\beta = \tau = \eta/16$, $\delta_{\mathrm{net}}^2 = \mu\eta/16$, and per-center failure probabilities summing to at most δ. For $R \in \{L_V, L_{\mathrm{pol}}\}$, let E_R be the bounded Lipschitz convex envelope built from the same anchors and labels, with Lipschitz constant R. On the event that all centerwise searches succeed, Lemma C.5 applies to both envelopes because g is L_V-Lipschitz and hence also L_{pol}-Lipschitz. The deterministic error is

$$e = \frac{\delta_{\mathrm{net}}^2}{\mu} + 2\beta + \delta_{\mathrm{net}}\sqrt{\frac{2\beta}{\mu}} + \tau = \frac{4 + \sqrt{2}}{16}\eta < \eta.$$

For every fixed feasible solution, the objective is nondecreasing in the Lipschitz constant. Consequently, $g \le E_{L_V} \le E_{L_{\mathrm{pol}}} \le g + e$. The algorithm returns the adjusted envelope $\widehat{g} = \max\{0, E_{L_V} - e\}$. Therefore $\widehat{g} \in \mathcal{V}_{L_V}$ for every sample realization and $g - \eta \le \widehat{g} \le g$.

Step 3: Extract a feasible action.

Let $U_v(s) := E_{\mathrm{ext}}(s)$ denote the optimal value of the convex program for action extraction in section 4.1. For every feasible solution, convexity of $\mathcal{X}$ and f gives

$$f(\bar{s}, \bar{a}) \le \sum_j \alpha_j f(\widehat{s}_j, \widehat{a}_j) + \rho B \le \sum_j \alpha_j t_j + \rho B.$$

The separate Lipschitz bounds in state and action then imply $f(s, \widehat{\pi}(s)) \le E_{\mathrm{ext}}(s)$. Conversely, fix a feasible solution (α, ρ, w_s) for $E_{L_{\mathrm{pol}}}(s)$ and write $\bar{s} = \sum_j \alpha_j \widehat{s}_j + w_s$. If $\rho > 0$, the condition $w_s \in \rho\mathcal{S}$ and the definition of $\mathcal{S}$ as the state projection of $\mathcal{X}$ provide an action $a_w \in \mathcal{A}(w_s/\rho)$. Set $w_a = \rho a_w$. If $\rho = 0$, set $w_a = 0$. Then $(w_s, w_a) \in \rho\mathcal{X}$, and convexity of $\mathcal{X}$ gives $(\bar{s}, \bar{a}) := \sum_j \alpha_j(\widehat{s}_j, \widehat{a}_j) + (w_s, w_a) \in \mathcal{X}$. The Hausdorff-Lipschitz condition on the feasible action sets provides $a \in \mathcal{A}(s)$ with $\|a - \bar{a}\|_2 \le L_A \|s - \bar{s}\|_2$. Hence

$$G_s\|s - \bar{s}\|_2 + G_a\|a - \bar{a}\|_2 \le (G_s + L_A G_a)\|s - \bar{s}\|_2 \le L_{\mathrm{pol}}\|s - \bar{s}\|_2,$$

so this solution is feasible for action extraction with no larger objective. Combining this comparison with Step 2 gives

$$f(s, \widehat{\pi}(s)) \le E_{\mathrm{ext}}(s) \le E_{L_{\mathrm{pol}}}(s) \le g(s) + e \le g(s) + \eta,$$

which proves $\mathcal{B}^{\widehat{\pi}} v \le \mathcal{B} v + \eta$. Moreover, $g(s) \le f(s, \widehat{\pi}(s)) \le U_v(s)$ by the definition of g and the first inequality above. Combining this relation with $g - \eta \le \widehat{g} \le g$ and $U_v \le g + \eta$ gives

$$0 \le f(s, \widehat{\pi}(s)) - g(s) \le U_v(s) - \widehat{g}(s) \le 2\eta,$$

which gives the computable upper bound on the one-step action error. Compactness and the standard measurability assumptions ensure that these minimizers define a policy.

Step 4: Count samples.

For $L_{\mathrm{pol}} > 0$, the choice $\mu = D_s/L_{\mathrm{pol}}$ and the calculation in Lemma C.5 give at most $C_{d_s}(4D_sL_{\mathrm{pol}})^{d_s/2}\eta^{-d_s/2}$ state centers. If $L_{\mathrm{pol}} = 0$, compactness of $\mathcal{S}$ and the fixed choice $\mu = 1$ give the same order. By Proposition C.7, each center uses $\widetilde{O}(\eta^{-2})$ samples. Its fixed prefactor includes the state–action optimization factor $\kappa_X^2(d_s + d_a)^4$ and the bounded-label and regularizer scales. Therefore

$$N_{\mathrm{state}} = \widetilde{O}\Big(\eta^{-d_s/2}\eta^{-2}\Big) = \widetilde{O}\Big(\eta^{-2-d_s/2}\Big),$$

and the union bound over centers gives the probability statement in Lemma E.2. Thus the action dimension is absent from the exponent of η, but it remains in the polynomial prefactor through the stochastic optimization over $\mathcal{X}$.

F Proof of Theorem 4.2

This section proves Theorem 4.2 directly. The proof constructs a state-indexed cap family, converts each incorrectly decoded cap into policy loss, controls the probability of such errors under adaptive sampling, and finally chooses the cap height and verifies the model conditions.

Step 1: Build a state-indexed hard family.

Let $\mathcal{S} = [-1, 1]^{d_s}$ and $\mathcal{A} = [-1, 1]^{d_a}$. Only the first action coordinate will be active. Choose state centers $s_1, \ldots, s_m$ separated by $6r$, where

$$r := \sqrt{h/\lambda}, \qquad m \geq c\, h^{-d_s/2}, \qquad b := r/2.$$

Define the baseline, raised tangents, and hard family by

$$\phi_0(s, a) := \lambda(\|s\|_2^2 + a_1^2), \qquad x_j := (s_j, be_1), \qquad f_\theta := \max\{\phi_0, \ell_j : \theta_j = 1\},$$

where ℓ_j is the tangent plane to ϕ_0 at x_j, raised by h. Use the deterministic self-loop transition and Bernoulli costs with means $p_\theta(s, a) := 1/4 + \alpha f_\theta(s, a)$, where the fixed $\alpha > 0$ is small enough that all means lie in $[1/4, 1/2]$.

Step 2: Convert an incorrect bit into policy loss.

The active-region geometry follows from

$$\phi_0(s, a) - \ell_j(s, a) = \lambda\big(\|s - s_j\|_2^2 + (a_1 - b)^2\big) - h.$$

Thus the cap regions are disjoint cylinders. At s_j, the optimal first action satisfies

$$a_1^\star(s_j) = \begin{cases} 0, & \theta_j = 0, \\ -r/2, & \theta_j = 1. \end{cases}$$

If $\theta_j = 0$, every action with $a_1 < -r/4$ has one-step excess at least $h/16$. If $\theta_j = 1$, every action with $a_1 \geq -r/4$ has one-step excess at least $h/4$. Hence a randomized policy placing at least half of its mass on the wrong side satisfies

$$\|V^{\widehat{\pi}} - V^\star\|_\infty \geq \frac{\alpha h}{32(1 - \gamma)}.$$

Step 3: Lower-bound the probability of a wrong bit.

Decode the jth bit according to

$$\widehat{\theta}_j := \mathbf{1}\left\{\widehat{\pi}(\{a : a_1 < -r/4\} \mid s_j) \geq \frac{1}{2}\right\}.$$

Neighboring hypotheses differ only in the jth cylinder and their Bernoulli means differ by at most αh. If N_j is the adaptive number of samples in that cylinder, then

$$\mathrm{KL}(P_\theta \| P_{\theta^{(j)}}) \leq C\alpha^2 h^2 \mathbb{E}_\theta N_j, \qquad \sum_{j=1}^m N_j \leq N.$$

Repeating the binary-mixture and Pinsker calculation from Step 3 of the $Q^\star$ lower bound yields

$$\mathbb{E}\,\mathrm{Ham}(\widehat{\theta},\theta) \geq \frac{m}{2}\left(1-\sqrt{\frac{C\alpha^2h^2N}{m}}\right).$$

Taking $N \leq c_{\delta_0} m h^{-2}$ makes the probability of at least one incorrectly decoded bit at least δ_0.

Step 4: Choose the cap height and verify the class.

Choose $h := 32(1-\gamma)\varepsilon/\alpha$. Then one wrong bit forces $\|V^{\widehat{\pi}} - V^\star\|_\infty \geq \varepsilon$, while $m \asymp h^{-d_s/2}$ converts the testing threshold into

$$N \leq c_0\{(1-\gamma)\varepsilon\}^{-2-d_s/2}.$$

The state–action space is a fixed compact product convex set with the same feasible action set at every state. The functions f_θ are maxima of affine and quadratic convex functions with uniformly bounded ranges and slopes. Write $m_\theta(s) := \min_{a\in\mathcal{A}} f_\theta(s,a)$. Minimizing over the convex action set makes m_θ convex, and the common Lipschitz bound of f_θ makes m_θ uniformly L_m-Lipschitz. Hence $(\mathcal{B}_\theta v)(s) = 1/4 + \alpha m_\theta(s) + \gamma v(s)$ is convex and has Lipschitz constant at most $\alpha L_m + \gamma L_V$ for every $v \in \mathcal{V}_{L_V}$. Choosing $L_V \geq \alpha L_m/(1-\gamma)$ shows that $\mathcal{B}$ preserves $\mathcal{V}_{L_V}$. The same self-loop calculation shows that $\mathcal{T}$ preserves $\mathcal{Q}_L$ for $L \geq L_p/(1-\gamma)$, and $F_v(s,a) = p_\theta(s,a) + \gamma v(s)$ has uniformly bounded state and action Lipschitz constants. All constants are independent of h and θ. The family therefore satisfies both Assumptions 2.1 and 2.2 and the model conditions in section 2, completing the proof.

G Generic Lipschitz Benchmark

The following rate is the classical sup-norm minimax benchmark for first-order Hölder, or Lipschitz, regression (Stone, 1982; Tsybakov, 2009). Because our oracle permits adaptive point queries, we include a short proof specialized to this sampling model.

Proposition G.1 (Generic Lipschitz uniform-recovery benchmark)**.** *Let $X \subset \mathbb{R}^p$ be a fixed well-rounded compact domain and let $\mathcal{L}_{B,L}(X)$ be the class of functions $f : X \to [0,B]$ with Lipschitz constant at most L. Suppose each adaptively selected query x_t returns a fresh observation $Y_t \in [0, B_Y]$ satisfying $\mathbb{E}[Y_t \mid \mathscr{H}_{t-1}, x_t] = f(x_t)$, where $B_Y \geq B$ is fixed. For any fixed confidence level $\delta_0 \in (0,1/2)$, the minimax sample complexity of uniform recovery over this function-and-oracle class is $\widetilde{\Theta}(\eta^{-p-2})$ for fixed B, B_Y, L, p and geometry.*

Proof. For the upper bound, take a Euclidean grid $\{z_j\}_{j=1}^M$ and average m independent observations at each grid point, where

$$h := \frac{\eta}{8L}, \qquad M \leq C(L/\eta)^p, \qquad m \geq CB_Y^2\eta^{-2}\log\frac{2M}{\delta_0}.$$

With probability at least $1-\delta_0$, all grid averages $\widehat{f}_j$ are within $\eta/4$ of their means. On this event, define the lower and upper Lipschitz envelopes

$$\underline{f}(x) := \max_j\left\{\widehat{f}_j - \frac{\eta}{4} - L\|x-z_j\|_2\right\}, \qquad \overline{f}(x) := \min_j\left\{\widehat{f}_j + \frac{\eta}{4} + L\|x-z_j\|_2\right\}.$$

The construction brackets f between $\underline{f}$ and $\overline{f}$, and the grid spacing makes their gap at most η uniformly. The clipped midpoint therefore achieves the upper bound with

$$Mm = \widetilde{O}(\eta^{-p-2})$$

samples.

For the lower bound, pack $M \asymp (L/\eta)^p$ disjoint Euclidean balls of radius $c\eta/L$ inside X. On the jth ball, place a Lipschitz tent ψ_j of height η and slope at most L, with disjoint supports, and define the Bernoulli observation means

$$p_\theta(x) := \frac{1}{4} + \alpha\sum_{j=1}^M \theta_j\psi_j(x), \qquad \theta \in \{0,1\}^M.$$

If N_j is the adaptive number of samples falling in the support of ψ_j, then neighboring hypotheses satisfy

$$\mathrm{KL}(P_\theta \| P_{\theta^{(j)}}) \le C\alpha^2\eta^2 \mathbb{E}_\theta N_j, \qquad \sum_{j=1}^{M} N_j \le N.$$

Mixture Assouad therefore gives expected Hamming error of order M whenever $N \lesssim M\eta^{-2}$. Uniform error below a small constant multiple of η decodes all bits, so any such estimator fails with constant probability. Hence

$$N \gtrsim M\eta^{-2} \asymp \eta^{-p-2}.$$

□

Thus, on a p-dimensional domain, Bellman convexity reduces the dimension term in the exponent from p to $p/2$ while retaining the usual η^{-2} sampling cost for noisy scalar values. In the fitted-Q application, $p = d_s + d_a$. The main lower bound remains within the convex Bellman class. This benchmark is only a comparison.

H Proof of Proposition 5.1

Consider first BER-FQI. By construction, every output of BER is convex on the feasible state–action space $\mathcal{X}$ for every sample realization. For any fixed state s, the feasible action set $\mathcal{A}(s) = \{a : (s,a) \in \mathcal{X}\}$ is compact and convex because it is obtained by fixing the state coordinate in the compact convex set $\mathcal{X}$. Thus computing continuation values and final greedy actions has the convex form $\min_{a\in\mathcal{A}(s)} Q_k(s,a)$.

For STATE-BER-VI, fix the stored fitted value labels and state–action anchors. The feasible set for convex action extraction is described by

$$\begin{aligned} &\alpha \ge 0, \qquad \rho \ge 0, \qquad \mathbf{1}^\top\alpha + \rho = 1, \\ &(\bar{s},\bar{a}) = \sum_j \alpha_j (s_j, a_j) + w, \qquad w \in \rho\mathcal{X}, \qquad a \in \mathcal{A}(s). \end{aligned}$$

The first constraints are affine or conic, $w \in \rho\mathcal{X}$ is the perspective of a convex set, and $\mathcal{A}(s)$ is convex. The objective is a sum of a linear term and norm terms. Hence convex action extraction is a compact convex optimization problem. Compactness follows from compactness of $\mathcal{X}$, bounded state–action anchors, and compact feasible action sets.

Proposition H.1 (Convex programs linear in the number of anchors)**.** *Fix a BER envelope built from M proximal anchors. Let N_{rep} and M_{rep} be the largest numbers of auxiliary variables and scalar constraints needed to describe the feasible state–action space, its perspective set, and one feasible action set in the chosen convex representation. Then evaluating a fitted function, computing a continuation value or final greedy action, and extracting an action with State-BER-VI can each be written as a convex program with size bounded by*

$$O(M + d_s + d_a + N_{\mathrm{rep}}) \text{ variables and } O(M + M_{\mathrm{rep}}) \text{ scalar constraints},$$

up to constant-size epigraph variables for the norm terms.

H.1 Proof of Proposition H.1

For a fixed fitted BER function, evaluating its convex-envelope representation at a state–action pair uses the variables $\alpha \in \mathbb{R}^M$, $\rho \in \mathbb{R}_+$, and $w \in \mathbb{R}^{d_s+d_a}$ from (10), plus a constant number of epigraph variables for the Euclidean norm and for the final positive part. The constraints are the $M+1$ nonnegativity constraints, one simplex equality, one membership constraint $w \in \rho\mathcal{X}$, and the norm epigraph. By definition of N_{rep} and M_{rep}, this has $O(M + d_s + d_a + N_{\mathrm{rep}})$ variables and $O(M + M_{\mathrm{rep}})$ scalar constraints.

Continuation and final greedy computations minimize this same epigraph form over the feasible action set $\mathcal{A}(s)$. They add only the action variable and one representation of $\mathcal{A}(s)$, so the same asymptotic size bound holds. For convex action extraction in State-BER-VI, the variables are

$\alpha, \rho, w = (w_s, w_a), \bar{s}, \bar{a}$, and the returned action a. The constraints consist of the same simplex and perspective-set constraints, affine definitions of $(\bar{s}, \bar{a})$, one feasible-action constraint, and a constant number of norm epigraphs. Hence the program for action extraction also has the displayed size, linear in the number of anchors.

H.2 Proof of Proposition 5.2

Standard conic-representability rules (Boyd and Vandenberghe, 2004) give closure under finite products, affine images, affine inverse images, intersections, and projections after adding auxiliary variables. Therefore the feasible action set $\mathcal{A}(s)$ obtained by fixing the state coordinate in a conic-representable state–action space is also conic-representable. If the epigraph of a fitted function is conic-representable, then next-state and greedy computations have the epigraph form

$$\min_{a,t} \; t \qquad \text{s.t.} \quad (s, a, t) \in \operatorname{epi}(Q_k), \quad a \in \mathcal{A}(s),$$

which is a finite conic program.

The convex envelope and action extraction computations add only the variables associated with stored anchors, simplex constraints, perspective-set constraints, feasible-action constraints, and epigraph variables for the relevant norms. Under the assumed conic descriptions of $\mathcal{X}$, $\operatorname{persp}(\mathcal{X})$, the feasible action sets, the epigraphs of fitted functions, and the norm epigraphs, this is again a finite conic program. The number of variables and constraints is polynomial in the description sizes and in the number of stored anchors because each anchor contributes only a constant number of scalar weights and affine terms, plus the auxiliary variables required by the fixed conic descriptions. In particular, polyhedral sets and piecewise-linear epigraphs give linear programs, Euclidean norm epigraphs give second-order cone programs when the remaining data are polyhedral, and semidefinite constraints give semidefinite programs.

I Numerical Optimization Tolerance

The statistical results assume that continuation values are computed exactly. Arbitrary pointwise numerical errors cannot simply be treated as harmless label bias: the perturbed Bellman mean need not be convex, and the stochastic proximal search relies on a convex expected objective. The following result covers approximate continuation values that preserve convexity.

Assumption I.1 (Convexity-preserving continuation approximation)**.** *At Bellman round k, the implementation uses an approximate continuation value $\widetilde{V}_k$ such that $\left\|\widetilde{V}_k - V_{Q_k}\right\|_\infty \le \zeta_k$, the induced mean $\widetilde{\mathcal{T}}_k Q_k(s,a) = \mathbb{E}[C + \gamma \widetilde{V}_k(S') \mid s, a]$ belongs to $\mathcal{Q}_L$, and the observation $C + \gamma \widetilde{V}_k(S')$ remains bounded by the same range used in the one-step theorem.*

Proposition I.2 (Effect of convexity-preserving continuation error)**.** *Suppose Assumption I.1 holds and the one-step recovery algorithm returns Q_{k+1} satisfying $\left\|Q_{k+1} - \widetilde{\mathcal{T}}_k Q_k\right\|_\infty \le \eta_k$. Then*

$$\|Q_K - Q^\star\|_\infty \le \gamma^K (1-\gamma)^{-1} + \sum_{k=0}^{K-1} \gamma^{K-1-k}(\eta_k + \gamma \zeta_k).$$

Proof. For every state–action pair $(s, a) \in \mathcal{X}$,

$$\begin{aligned} |\widetilde{\mathcal{T}}_k Q_k(s,a) - \mathcal{T} Q_k(s,a)| &\le \gamma \mathbb{E}\Big[|\widetilde{V}_k(S') - V_{Q_k}(S')| \mid s, a\Big] \\ &\le \gamma \zeta_k. \end{aligned}$$

Consequently,

$$\|Q_{k+1} - \mathcal{T} Q_k\|_\infty \le \eta_k + \gamma \zeta_k.$$

Iterating the usual Bellman contraction recursion proves the proposition. □

Thus the same theorem applies to numerical approximations that keep the Bellman mean in $\mathcal{Q}_L$, with the additional errors shown above. Floating-point bit complexity is outside the present scope.

J Sufficient Conditions: Affine Dynamics with Exogenous Noise

Definition J.1 (Affine dynamics with exogenous noise)**.** *The transition kernel has affine dynamics with exogenous noise if $S' = F(s, a, \omega)$, where $\omega \sim \nu$ has a distribution independent of (s, a), the map F is measurable, and $F(\cdot, \omega)$ is affine for almost every ω. Equivalently, $P(B \mid s, a) = \nu\{\omega : F(s, a, \omega) \in B\}$ for every Borel set $B \subseteq \mathcal{S}$.*

Proposition J.2 (Sufficient condition for convex value updates)**.** *Suppose the model geometry in section 2 holds with Hausdorff constant L_A, the expected cost c is convex and Lipschitz with constants c_s, c_a in the state and action coordinates, and the transition kernel satisfies Definition J.1. If F is Lipschitz with constants F_s, F_a, uniformly over ω, and $\gamma(F_s + L_A F_a) < 1$, then Assumption 2.2 holds for any $L_V \geq (c_s + L_A c_a)/(1 - \gamma(F_s + L_A F_a))$. The targets have separate state and action Lipschitz constants $G_s = c_s + \gamma F_s L_V$ and $G_a = c_a + \gamma F_a L_V$, and one may take $L_F = \sqrt{G_s^2 + G_a^2}$.*

Proposition J.3 (Sufficient condition for convex Q updates)**.** *Suppose the model conditions in section 2 hold, the expected cost c is convex and c_X-Lipschitz on $\mathcal{X}$, and the transition kernel satisfies Definition J.1. If F is F_X-Lipschitz on $\mathcal{X}$, uniformly over ω, and $\gamma F_X \sqrt{1 + L_A^2} < 1$, then Assumption 2.1 holds for any $L \geq c_X/(1 - \gamma F_X \sqrt{1 + L_A^2})$.*

These propositions are only sufficient conditions. State–action-independent noise ensures that taking expectation preserves convexity for every noise realization, but either Bellman-convexity condition may hold for more general transition models.

J.1 Proof of Proposition J.2

Fix $v \in \mathcal{V}_{L_V}$. The expected cost c is convex by assumption. For each noise realization ω, the map $(s, a) \mapsto F(s, a, \omega)$ is affine, so $(s, a) \mapsto v(F(s, a, \omega))$ is convex. Integration over the fixed noise distribution ν preserves convexity, hence $F_v(s, a) = c(s, a) + \gamma \mathbb{E}_\nu[v(F(s, a, \omega))]$ is convex on $\mathcal{X}$.

The separate state and action Lipschitz bound is

$$\begin{aligned} |F_v(s, a) - F_v(s', a')| \leq (c_s + \gamma F_s L_V)\|s - s'\|_2 \\ + (c_a + \gamma F_a L_V)\|a - a'\|_2. \end{aligned}$$

Thus $G_s = c_s + \gamma F_s L_V$ and $G_a = c_a + \gamma F_a L_V$. The assumed lower bound on L_V gives $c_s + L_A c_a + \gamma(F_s + L_A F_a) L_V \leq L_V$. For any states s, s', transfer an approximately minimizing action at s to $\mathcal{A}(s')$ by the Hausdorff condition and apply this bound. This gives $|\mathcal{B}v(s) - \mathcal{B}v(s')| \leq (G_s + L_A G_a)\|s - s'\|_2 \leq L_V \|s - s'\|_2$. Minimizing the convex function F_v over actions preserves convexity in the state, and the range condition follows from $0 \leq c \leq 1$, $F(s, a, \omega) \in \mathcal{S}$, and $0 \leq v \leq (1 - \gamma)^{-1}$. Hence Assumption 2.2 is verified.

J.2 Proof of Proposition J.3

Let $q : \mathcal{X} \to [0, (1 - \gamma)^{-1}]$ be convex and L-Lipschitz, and define $V_q(s) = \min_{a \in \mathcal{A}(s)} q(s, a)$. For any s_0, s_1, $\lambda \in [0, 1]$, and $\epsilon > 0$, choose $a_i \in \mathcal{A}(s_i)$ with $q(s_i, a_i) \leq V_q(s_i) + \epsilon$. Convexity of $\mathcal{X}$ and q gives

$$\begin{aligned} V_q(\lambda s_0 + (1 - \lambda)s_1) &\leq q\big(\lambda(s_0, a_0) + (1 - \lambda)(s_1, a_1)\big) \\ &\leq \lambda V_q(s_0) + (1 - \lambda)V_q(s_1) + \epsilon. \end{aligned}$$

Letting $\epsilon \downarrow 0$ proves convexity of V_q.

The Hausdorff-Lipschitz condition on the feasible action sets gives the state Lipschitz bound for V_q. Indeed, if $a \in \mathcal{A}(s)$ is ϵ-optimal and s' is another state, there is $a' \in \mathcal{A}(s')$ with $\|a - a'\|_2 \leq L_A \|s - s'\|_2$. Hence

$$\begin{aligned} V_q(s') - V_q(s) &\leq q(s', a') - q(s, a) + \epsilon \\ &\leq L\sqrt{1 + L_A^2}\|s - s'\|_2 + \epsilon. \end{aligned}$$

Symmetry and $\epsilon \downarrow 0$ give $\mathrm{Lip}(V_q) \leq L\sqrt{1 + L_A^2}$.

The expected cost c is convex. For each noise realization ω, $(s, a) \mapsto F(s, a, \omega)$ is affine, so $(s, a) \mapsto V_q(F(s, a, \omega))$ is convex. Integration over the fixed noise distribution preserves convexity,

and therefore $\mathcal{T}q$ is convex. Moreover,

$$\mathrm{Lip}(\mathcal{T}q) \le c_X + \gamma F_X L\sqrt{1+L_A^2} \le L.$$

Finally, $0 \le c \le 1$ and $0 \le q \le (1-\gamma)^{-1}$ imply $0 \le \mathcal{T}q \le (1-\gamma)^{-1}$, so Assumption 2.1 holds.

J.3 Storage control at high discount factors

The next example shows directly that $\mathcal{B}v$ remains convex and Lipschitz without requiring $\gamma(F_s + L_A F_a) < 1$. Thus the value update can satisfy Bellman convexity even when the sufficient contraction condition above fails.

Proposition J.4 (Bellman convexity for storage control)**.** *Let $\mathcal{S} = [0,1]$, let $\mathcal{A}(s) = [-r, r] \cap [-s, 1-s]$ for some $r \in (0,1]$, and let the next state be $S' = s + a$. Suppose c is jointly convex, takes values in $[0,1]$, and satisfies $|c(s,a) - c(s',a')| \le c_s|s-s'| + c_a|a-a'|$. Then, for every $\gamma \in [0,1)$, Assumption 2.2 holds with $L_A = 1$ and $L_V = (c_s + c_a)/(1-\gamma)$. The targets have separate state and action Lipschitz constants $G_s = c_s + \gamma L_V$ and $G_a = c_a + \gamma L_V$, and one may take $L_F = \sqrt{G_s^2 + G_a^2}$.*

J.4 Proof of Proposition J.4

The feasible state–action space is the polytope defined by $0 \le s \le 1$, $-r \le a \le r$, and $0 \le s+a \le 1$. Its feasible action sets are Hausdorff-Lipschitz with constant one. For $v \in \mathcal{V}_{L_V}$, the function $F_v(s,a) = c(s,a) + \gamma v(s+a)$ is convex and has separate state and action Lipschitz constants $G_s = c_s + \gamma L_V$ and $G_a = c_a + \gamma L_V$.

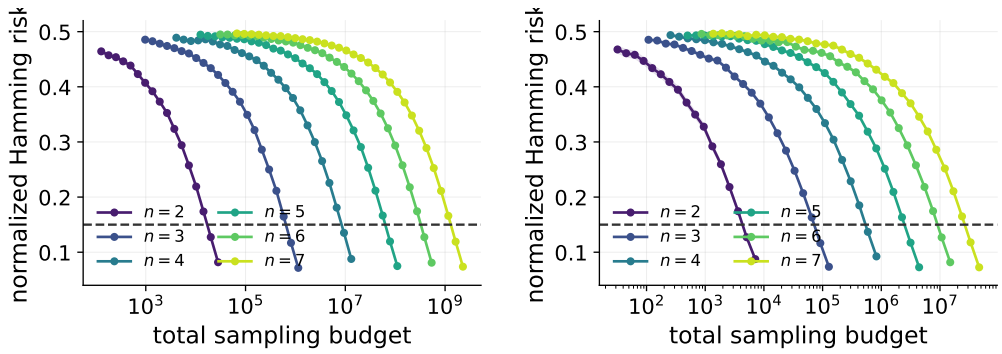
It remains to verify the sharper state regularity after minimizing over actions. Fix $s < s'$, write $d = s' - s$, and take $a \in \mathcal{A}(s)$. With $y = s + a$, the feasible next-state interval is $I(s) = [\max\{0, s-r\}, \min\{1, s+r\}]$. Its two endpoints are nondecreasing and one-Lipschitz in s. Therefore $y' := \max\{y, \max\{0, s'-r\}\}$ belongs to $I(s')$ and satisfies $0 \le y' - y \le d$. Set $a' = y' - s'$. Then $a' \in \mathcal{A}(s')$, $|y' - y| \le d$, and $|a' - a| = |(y'-y) - d| \le d$. Applying this transfer to an approximately minimizing action, and then reversing the roles of s and s', yields

$$|\mathcal{B}v(s') - \mathcal{B}v(s)| \le (c_s + c_a + \gamma L_V)|s' - s| = L_V|s' - s|.$$

Minimizing over the convex feasible state–action space preserves convexity, and the usual cost range gives $0 \le \mathcal{B}v \le (1-\gamma)^{-1}$. Thus $\mathcal{B}v \in \mathcal{V}_{L_V}$ for every $v \in \mathcal{V}_{L_V}$, which proves the proposition. For example, this conclusion applies to the convex charging cost $c(s,a) = \frac{1}{2}s + \frac{1}{2}a^2$ for every discount factor below one.

K Additional Numerical Illustration

This section gives the full protocol for the experiments in section 6 and an additional smooth synthetic-MDP study. The raised-cap experiment uses between 64 and 2,048 independent hard-family draws, with more repetitions for smaller families. The algorithmic experiments use fresh transition-cost samples at the state–action pairs selected during proximal search, and exact expected costs or dense dynamic programs only for evaluation. Every algorithmic curve reports the final iterate at the stated budget, with no best-so-far checkpoint selection. Lines and shaded regions show medians and interquartile ranges over sixteen seeds in the smooth synthetic experiment and twenty-four in the EIA-930 experiment.



(a) Uniform $Q^\star$ hard family, $d_s = 1, d_a = 4$

(b) Policy hard family, $d_s = 3$

Figure 3: Raw total-budget decoding trajectories. Each curve reports the mean normalized Hamming risk at the stated cumulative budget, with 95% intervals over independent full-family draws. The curves are not monotonized. The dashed line marks the 15% threshold used in Figure 1.

Experiment	Training protocol	Reported quantities
Raised-cap total-budget study	Six lattice widths $n \in \{2, \ldots, 7\}$, cap height $h = 0.24n^{-2}$, $d_s \in \{1, 2, 3\}$, and $d_a \in \{1, 2, 4\}$. The number of independent hard-family draws is adjusted to the family size.	Critical total budgets and fitted exponents for uniform $Q^\star$ recovery and direct policy learning.
Synthetic Bellman-convex MDP	$d_s = 1$, $d_a \in \{1, 2, 3\}$, $\gamma = 0$, and sample-budget caps $\{4, 8, 16, 32, 64\} \times 10^3$. BER-FQI and State-BER-VI are trained independently on the same MDP class.	Mean policy gap, dependence on d_a, training time, per-state action-extraction time, and actual sample count.
Wind-integrated energy-storage control	EIA-930 ERCOT data, $d_s = 2$, $d_a \in \{1, 2, 4\}$, $\gamma = 0.8$, eighteen Bellman rounds, and sample-budget caps $\{8, 32, 128, 256\} \times 10^3$.	Held-out operating cost, policy gap, action-dimension sensitivity, mean daily net load, training time, and actual sample count.

Table 1: Numerical protocol. Each budget is an upper bound on fresh generative-model samples. Deterministic optimization and evaluation are not charged to that budget. The action-dimension comparison uses a common budget of 128,000 samples.

Raised-cap minimax study. This experiment implements the binary statistical problem induced at the cap tips by the constructions in sections D and F. For lattice width n and dimension p, the experiment instantiates all $m = n^p$ cap bits jointly and sets their height to $h = 0.24n^{-2}$. Here $p = d_s + d_a$ for uniform $Q^\star$ recovery and $p = d_s$ for policy learning. The known Bernoulli baseline varies mildly across the quadratic cap-tip locations, and an active bit raises its mean by $0.6h$. The decoder knows these locations and laws but not the independently sampled bits. Starting from a balanced allocation, it assigns each subsequent batch from one total budget according to posterior bit uncertainty, with a small exploration component, and then applies the likelihood-ratio decoder. The critical budget is the interpolated point at which the estimated normalized Hamming error falls below 15%, with confidence intervals obtained by bootstrapping independent hard-family draws. Thus every plotted budget is measured from a jointly sampled family rather than formed by multiplying a separate per-cap estimate by the packing size. The dashed lines use the theoretical slopes and are normalized at the geometric center of each measured curve. Representative raw risk trajectories are shown in figure 3.

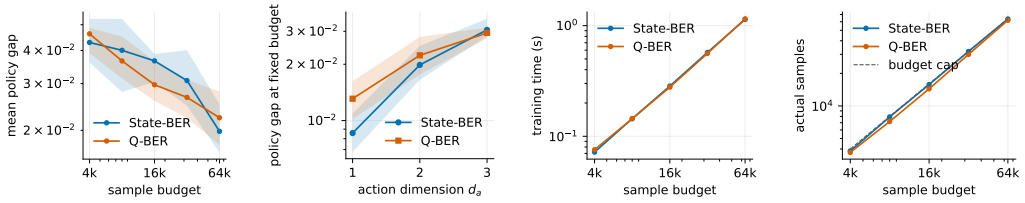
Practical BER implementation. The theoretical results permit any high-probability stochastic convex optimization routine satisfying the interface in Corollary C.6. In the experiments, we use a local-search implementation that observes only function values. Starting from each uniform center, the routine evaluates the proximal objective along positive and negative coordinate directions and two random directions, keeps the best sampled point, and shrinks the search radius geometrically. These sampling locations are adaptive and all objective evaluations use fresh generative-model draws. The selected proximal anchor is then evaluated with an independent validation batch. The synthetic experiment assigns approximately 42% of each center's budget to search, while the storage experiment assigns 74%. The remainder is used for validation. Storage budgets increase by a factor of 1.035 across Bellman rounds. The fitted max-affine envelopes and programs for action extraction are solved as finite convex programs. Thus the implementation follows the sampling, validation, and construction of convex envelopes in BER, although its practical local search is not the worst-case stochastic optimizer used in the theorem.

Synthetic Bellman-convex MDP. The state space is $[-1, 1]$, the action space is $[-1, 1]^{d_a}$, and the discount factor is zero, so this experiment isolates the one-step statistical distinction without Bellman truncation error. Let w_i be equally spaced between 0.35 and 0.75. At (s, a), the generative model returns a Bernoulli cost whose mean is

$$c(s,a) = 0.12 + 0.10(s + 0.25)^2 + \frac{0.15}{d_a} \sum_{i=1}^{d_a} |a_i - w_i s| + 0.035 \left(d_a^{-1/2} \sum_{i=1}^{d_a} a_i - 0.30 s \right)^2 .$$

This function is jointly convex and bounded strictly between zero and one on the feasible set. BER-FQI covers the joint state–action space and fits a convex Q function. State-BER-VI places centers only in the state coordinate, performs the same adaptive proximal search over states and actions with a quadratic regularizer on the state coordinate, and extracts actions by the convex interpolation program in section 4.1. Both methods receive the same budget cap. Policy gaps are evaluated on twenty-five held-out states using the exact mean cost, which is unavailable to the learners. This experiment compares the finite-sample behavior of the two learning objectives on the same problem.

Under matched generative-model budgets, State-BER-VI learns the value function and associated actions, while BER-FQI learns the full Q function. In figure 4, panels (a) and (b) report policy quality as the sample budget and action dimension vary, while panels (c) and (d) report training time and actual sample use. Both methods improve with the sample budget and achieve comparable policy gaps. Their training times remain similar, and the actual sample counts closely track the matched budget caps.



(a) Policy learning (b) Action dimension (c) Training time (d) Samples used

Figure 4: Smooth synthetic Bellman-convex MDP. Both methods improve with the sample budget and achieve comparable policy quality under matched actual sample counts. Curves and bands summarize sixteen independent runs.

Wind-integrated energy-storage control. We use hourly demand and wind generation reported for ERCOT in the EIA-930 Hourly Electric Grid Monitor (U.S. Energy Information Administration, 2026). Net load is demand minus wind generation. The first $8{,}760$ observations, corresponding to 2023, determine the empirical transition distribution and the 1st–99th percentile normalization. The remaining $8{,}784$ observations from 2024 are held out until evaluation.

The state is $s = (e, z) \in [0, 1]^2$, where e is battery charge and z is normalized net load. The action is $a = (b, u_1, \ldots, u_m)$, where $|b| \le 0.18$, $0 \le e + b \le 1$, and $|u_i| \le 0.04$. Here b is battery charging

and the u_i's are flexible-load adjustments. A transition sets $e' = e + b$ and draws z' independently from the 2023 empirical net-load distribution. With $\gamma = 0.8$, the expected one-step cost is

$$c(s,a) = 0.055 + 0.65\left(z + b + \sum_{i=1}^{m} u_i - 0.48\right)^2 + 0.25(e - 0.50)^2 + 0.12b^2 + 0.25\sum_{i=1}^{m} u_i^2.$$

Over the feasible set, the expected cost satisfies $0.055 \le c(s,a) \le 0.559648$. Since the observed cost adds an independent centered bounded perturbation of magnitude 0.006, it satisfies $C \in [0.049, 0.565648] \subset [0,1]$, as required by the model conditions in section 2. The main experiment uses $m = 3$, hence $d_a = 4$, and both algorithms run for eighteen fitted Bellman rounds. The reference is computed by value iteration on a 25×25 state grid with twenty-one empirical net-load quantiles. Model policy gaps use this stochastic reference. Held-out costs instead roll each learned final policy through the complete 2024 sequence. Panel (c) averages the observed and controlled net load by hour of day for the run with median final State-BER-VI policy gap. Neither the reference nor the 2024 observations are used during training.

Reference-grid refinement. We recomputed the storage reference on 25×25, 37×37, and 49×49 state grids while retaining twenty-one net-load quantiles. The corresponding held-out reference cost reductions were 12.833%, 12.720%, and 12.723%. For the representative final policy at the $256{,}000$-sample budget, the grid-estimated policy gaps were $(0.009630, 0.009626, 0.009615)$ for State-BER-VI and $(0.072026, 0.071285, 0.070992)$ for BER-FQI. Thus the reported comparison is stable under refinement of the evaluation grid.