POWER MIRROR: OFFLINE POLICY IMPROVEMENT WITH BOUNDED REWEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline reinforcement learning (RL) must balance policy improvement with reliable use of logged data. We introduce *Power Mirror*, a policy-improvement framework derived from shifted power-mean value backups. With a suitable shift, the induced updates satisfy explicit two-sided bounds on actionwise probability ratios, together with population improvement guarantees and sensitivity bounds for critic error. We develop two complementary algorithms: fixed-anchor value iteration maintains a permanent bound relative to the behavior policy, while moving-anchor policy iteration improves monotonically and converges in value to the support-constrained optimum under exact evaluation and persistent updates. We also construct a ratio-preserving actor for continuous actions and a held-out certification rule that deploys a learned policy only when improvement is certified, otherwise retaining the behavior policy. Together, these results connect value estimation, policy optimization, and deployment through explicit control of policy reweighting. Experiments support our theoretical findings, with Power Mirror achieving strong policy improvement while remaining conservative under limited data coverage

1 INTRODUCTION

Offline RL learns from a fixed dataset collected by a *behavior policy* π_β. Unlike online RL, the learner cannot collect more data when an action is poorly estimated. As a result, a greedy update can place too much probability on a rare action with an overestimated value, shifting the policy away from the data distribution and amplifying critic error through subsequent Bellman backups.

Behavior cloning avoids this shift by returning π_β, but does not improve the policy, while greedy value iteration can move arbitrarily far from behavior. Existing offline-RL methods therefore constrain support, regularize toward behavior, learn pessimistic values, or certify improvement over a baseline (Fujimoto et al., 2019; Kumar et al., 2019; Wu et al., 2019; Kumar et al., 2020; Laroche et al., 2019; Petrik et al., 2016). However, an average constraint such as KL divergence does not directly control the probability ratio of every action, and an unconstrained actor can lose a ratio guarantee satisfied by the underlying policy update.

Our starting question. Can we build the behavior constraint directly into the policy-improvement rule, so that the same update both favors actions with larger estimated value and places a certified limit on every probability change? Power Mirror answers this question by writing the new policy as

$$\pi_{\mathrm{PM}}(a|s) = \pi_\beta(a|s)\, r_{\mathrm{PM}}(s,a),$$

where the multiplier r_{PM} is normalized to have mean one under π_β and is bounded above and below. The multiplier is obtained by differentiating a shifted, behavior-weighted power-mean Bellman backup. Power means and regularized Bellman operators are classical objects (Dam et al., 2020; Geist et al., 2019; Neu et al., 2017); our contribution is to show that this particular behavior-relative construction connects policy improvement, pointwise distribution-shift control, coverage, critic robustness, and safe deployment.

Two population algorithms, with different goals. *Fixed-anchor Power-Mirror value iteration* (PMVI) uses π_β as the reference in every Bellman backup. With fixed reference, powers, and shifts, its population operator is a monotone contraction. Starting from V^{π_β}, the iterates improve monotonically, and every extracted policy remains in the same actionwise ratio box around π_β. This permanent control introduces bias, so PMVI generally does not reach the best policy supported by π_β.

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.

Moving-anchor Power-Mirror policy iteration (PMPI) instead uses the current policy as the next reference. Each exact step improves value, and sufficiently persistent updates converge to the best policy supported by π_β. Its ratio guarantee is only one-step, so ratios accumulate across iterations and the final policy may leave the original behavior box. Thus fixed anchoring gives permanent proximity, while moving anchoring gives support-constrained optimality.

Moving-anchor Power-Mirror policy iteration (PMPI) instead evaluates the current policy and uses it as the reference for the next update. Every exact step improves the value, and sufficiently persistent updates converge to the best policy supported by π_β. Its ratio guarantee, however, is only local: consecutive ratios multiply across iterations, so the final policy need not remain inside the original box around π_β. Thus fixed anchoring provides permanent behavior-relative control, whereas moving anchoring trades this guarantee for eventual support-constrained optimality.

From population results to an offline deployment decision. Exact action values are unavailable in a finite dataset. We therefore do not present population monotonicity as a finite-sample safety theorem. We train the critic and candidate on one data split, then use independent behavior-policy trajectories to estimate the candidate's return improvement with per-decision importance sampling (PDIS). The pointwise policy-ratio bounds give deterministic bounds on every PDIS prefix weight. The method deploys the candidate only if a high-probability lower confidence bound is nonnegative (or exceeds a user-chosen margin); otherwise it falls back exactly to π_β.

Continuous actions without losing the certificate. In continuous action spaces, weighted maximum-likelihood fitting controls only an average objective and need not preserve pointwise density ratios. We instead parameterize the actor as a bounded multiplier of the behavior density and normalize the multiplier by construction. The ratio box therefore holds for every actor parameter, even before convergence, and the resulting policy can be sampled exactly by rejection sampling.

In summary, we contribute: (i) a sharp pointwise behavior-ratio theorem, an exact population-improvement identity, a critic-error transfer bound, and a coverage-dependent gap to greedy improvement; (ii) fixed-anchor PMVI and moving-anchor PMPI, with distinct convergence and behavior-drift guarantees; and (iii) a high-probability deployment test and a continuous-action actor that preserve the ratio certificate. We evaluate the resulting return–coverage tradeoff in tabular and continuous-control tasks. Full proofs and experimental details appear after the references.

2 The Power-Mirror update and its behavior-ratio guarantee

We first study one state and omit s to keep the notation readable. The behavior policy is the probability vector $\pi_\beta \in \Delta_K$, and $Q(a)$ is the action value used by the policy update. We consider only actions with $\pi_\beta(a) > 0$: no method that stays within behavior support can learn about an action that never appears in the data. Let

$$D = \max_a Q(a) - \min_a Q(a)$$

be the range (or span) of the action values. Choose a finite power $1 < p < \infty$ and a shift $b > D$. Define

$$\bar{Q}_\beta = \mathbb{E}_{A\sim\pi_\beta} Q(A), \qquad z(a) = b + Q(a) - \bar{Q}_\beta, \qquad m_p = (\mathbb{E}_{A\sim\pi_\beta} z(A)^p)^{1/p}, \tag{1}$$

$$C_{p,\pi_\beta,b}(Q) = \bar{Q}_\beta + m_p - b. \tag{2}$$

The condition $b > D$ makes every shifted value $z(a)$ positive. It also gives the useful identity $\mathbb{E}_{\pi_\beta} z = b$. The scalar $C_{p,\pi_\beta,b}(Q)$ is the Power-Mirror backup: it lies between the behavior-policy average and the largest supported action value. Its gradient with respect to the vector of action values is a policy. In explicit form, let

$$R = \mathbb{E}_{A\sim\pi_\beta}(z(A)/m_p)^{p-1}, \qquad \pi_{\mathrm{PM}}(a) = \pi_\beta(a)\left[(z(a)/m_p)^{p-1} + 1 - R\right]. \tag{3}$$

The first term favors above-average actions. The correction $1 - R$ is not an arbitrary baseline: it makes the probabilities sum exactly to one. Power-mean monotonicity gives $R \le 1$, so the correction is nonnegative and the update cannot assign a negative probability.

Theorem 1 (Induced policy and sharp ratio envelope)**.** *Equation* (3) *defines the normalized policy* $\pi_{\mathrm{PM}} = \nabla C_{p,\pi_\beta,b}(Q)$. *On* $\mathrm{supp}(\pi_\beta)$, *let* $r(a) = \pi_{\mathrm{PM}}(a)/\pi_\beta(a)$ *and* $\eta = D/b$. *Then*

$$L_{p,\eta} := (1-\eta)^{p-1} \le r(a) \le (1+\eta)^{p-1} =: U_{p,\eta}, \tag{4}$$

and

$$\max_a |r(a) - 1| \le H_{p,\eta} := \max\{1 - L_{p,\eta},\, U_{p,\eta} - 1\}. \tag{5}$$

Moreover, $\max_a r(a)/\min_a r(a) \le L_{p,\eta}^{-1}$*. All constants are uniformly sharp, already for two actions.*

The theorem says more than "the new policy is close on average." It limits the change of *each* supported action probability. The bound is controlled by two quantities. Increasing p makes the update more aggressive. Increasing b decreases $\eta = D/b$ and keeps the policy closer to behavior. Supplementary Theorem 7 proves sharpness and shows why no smaller uniform endpoints are possible. For $\eta > 0$ and $\rho > 1$, if a practitioner requests the upper ratio cap $\pi_{\mathrm{PM}}(a)/\pi_\beta(a) \le \rho$, it is sufficient—and uniformly exact for this construction—to choose

$$p \le 1 + \frac{\log \rho}{\log(1+\eta)}, \tag{6}$$

while a symmetric interval $[1/\rho, \rho]$ uses $p \le 1 + \log \rho/[-\log(1-\eta)]$. When $\eta = 0$, the policy equals behavior for every finite p, so no power restriction is needed. Thus p and b can be selected from an interpretable probability-ratio budget rather than only tuned as opaque regularization coefficients. Supplementary Corollary 1 converts the same interval into bounds on total variation, KL and chi-square divergence; Corollary 2 gives trajectory likelihood-ratio bounds.

Why the power update? An exact exponential update can also satisfy pointwise ratio bounds when its temperature is calibrated to the critic range. Our rationale therefore concerns the geometry of improvement within a behavior constraint. The case $p = 2$ makes this explicit. At one state, write $w = \pi_\beta(\cdot \mid s)$ and define

$$\widehat{A} = \widehat{Q} - \mathbb{E}_w \widehat{Q}, \qquad \widehat{\sigma}^2 = \mathbb{E}_w \widehat{A}^2, \qquad \widehat{m}_2 = \sqrt{b^2 + \widehat{\sigma}^2}. \tag{7}$$

Here $\widehat{\sigma}$ describes variation across estimated action values, not estimation uncertainty. For a fixed shift $b > \mathrm{span}(\widehat{Q})$, the induced ratio is

$$\widehat{r}_2(a) = \frac{\widehat{\pi}_2(a)}{w(a)} = 1 + \frac{\widehat{A}(a)}{\widehat{m}_2}. \tag{8}$$

For any supported policy $\pi = rw$, its predicted gain and squared ratio movement are $\widehat{G}(\pi) = \mathbb{E}_w[(r-1)\widehat{Q}]$ and $\chi^2(\pi\|w) = \mathbb{E}_w[(r-1)^2]$.

Proposition 1 (Quadratic improvement efficiency)**.** *Let* $c_2 = \widehat{\sigma}^2/(b^2 + \widehat{\sigma}^2)$ *and* $\varepsilon_Q = \sqrt{\mathrm{Var}_w(Q - \widehat{Q})}$ *for the true critic* Q*. Maximizing over probability policies,*

$$\chi^2(\widehat{\pi}_2\|w) = c_2, \qquad \widehat{G}(\widehat{\pi}_2) = \max_{\substack{\pi \ll w \\ \chi^2(\pi\|w) \le c_2}} \widehat{G}(\pi) = \frac{\widehat{\sigma}^2}{\widehat{m}_2}. \tag{9}$$

Moreover,

$$\mathbb{E}_{\widehat{\pi}_2} Q - \mathbb{E}_w Q \ge \frac{\widehat{\sigma}(\widehat{\sigma} - \varepsilon_Q)}{\widehat{m}_2}. \tag{10}$$

The proof is Cauchy–Schwarz: predicted gain is at most $\widehat{\sigma}\sqrt{\chi^2(\pi\|w)}$, and the quadratic update attains equality. The error contribution is at most $\varepsilon_Q \sqrt{\chi^2(\pi\|w)}$. Thus $p = 2$ maximizes predicted gain for its achieved critic-error amplification budget. This is a characterization of classical quadratic trust-region geometry, rather than a claim that such geometry is new (Belousov & Peters, 2018). In particular, a χ^2-constrained optimizer at radius c_2 reproduces the same policy.

The fixed shift also controls sensitivity. If $r_2[Q]$ and $r_2[Q']$ use the same w and b, with both critic spans below b, then

$$\|r_2[Q] - r_2[Q']\|_{L^2(w)} \le \frac{1}{b} \left\|(Q - Q') - \mathbb{E}_w(Q - Q')\right\|_{L^2(w)}. \tag{11}$$

This policy changes continuously near action-value ties; a direct hard ratio-box optimizer need not share this continuity. Exponential updates are also smooth, so stability alone does not distinguish Power Mirror from KL updates. Appendix I.2 gives the proofs, an exact worst-case critic-error interpretation, and a comparison with box-constrained improvement. These are statewise properties of the ideal $p = 2$ policy; they neither establish universal return superiority nor replace the separate finite-data deployment test.

What happens when the critic is wrong? Let Q denote the true action-value function for the policy-improvement step and let $\widehat{Q}$ be its estimate. Define the critic error explicitly by $\Delta_Q(a) = \widehat{Q}(a) - Q(a)$. Assume that the estimated values used by the update satisfy $\mathrm{span}(\widehat{Q}) \le D < b$, so $\eta = D/b$ controls the ratio of the learned candidate. Let $\widehat{\pi}_{\mathrm{PM}}$ be the policy constructed from $\widehat{Q}$, and write its behavior ratio as $r(a) = \widehat{\pi}_{\mathrm{PM}}(a)/\pi_\beta(a)$. The change in expected critic error caused by changing policies satisfies

$$\left|\mathbb{E}_{\widehat{\pi}_{\mathrm{PM}}}\Delta_Q - \mathbb{E}_{\pi_\beta}\Delta_Q\right| = \left|\mathbb{E}_{\pi_\beta}[(r-1)\Delta_Q]\right| \le H_{p,\eta}\mathbb{E}_{\pi_\beta}|\Delta_Q|, \tag{12}$$

and the multiplier $H_{p,\eta}$ converges to zero as $p \downarrow 1$. Therefore, if the improvement predicted by $\widehat{Q}$ exceeds the right-hand side, the update also improves under the true Q. If it does not, the test is inconclusive rather than evidence of harm. The important coupling is that a more conservative policy update automatically reduces the amount of critic error that can be reweighted.

How behavior coverage controls the price of safety. For $\varepsilon \ge 0$, call an action ε-good if its value is within ε of the best supported action. Let $Q_{\max} = \max_a Q(a)$, $M = b + Q_{\max} - \mathbb{E}_{\pi_\beta}Q$, and let $W_\varepsilon = \Pr_{A\sim\pi_\beta}\{Q(A) \ge Q_{\max} - \varepsilon\}$ be the total behavior probability of these good actions. Then

$$0 \le Q_{\max} - C_{p,\pi_\beta,b}(Q) \le \min\Big\{D,\ \varepsilon W_\varepsilon + D(1-W_\varepsilon),\ M(1-W_\varepsilon^{1/p}) + \varepsilon W_\varepsilon^{1/p}\Big\}. \tag{13}$$

When W_ε is large, behavior already samples useful actions often and the Power-Mirror backup can approach the greedy maximum. When W_ε is small, the bound deliberately becomes conservative. Merely saying that behavior has nonzero support is not enough: an action observed with probability 10^{-6} should not be treated like an action observed with probability 0.2.

3 Fixed-anchor value iteration and coverage-dependent optimality

The one-state construction has an exact improvement identity:

$$\mathbb{E}_{\pi_{\mathrm{PM}}}Q - \mathbb{E}_{\pi_\beta}Q = m_p - bR = (m_p - b) + b(1-R) \ge 0.$$

The two terms are nonnegative by the power-mean inequalities $m_p \ge b$ and $R \le 1$. The inequality is strict whenever Q is nonconstant on behavior support. We now show how this local fact becomes a complete population algorithm.

Consider a finite discounted MDP with discount $0 \le \gamma < 1$, rewards in $[r_{\min}, r_{\max}]$, and behavior policy π_β. The supported action set is $\mathcal{A}_\beta(s) = \{a : \pi_\beta(a|s) > 0\}$. The largest possible range of a discounted action-value function is $\bar{D} = (r_{\max} - r_{\min})/(1-\gamma)$. At each state, choose $1 < p_s < \infty$ and $b_s > \bar{D}$ before starting value iteration. The reference policy, power, and shift remain fixed throughout the recursion. These parameters may differ across states but do not depend on the current value iterate.

Let $\mathcal{V} = [r_{\min}/(1-\gamma), r_{\max}/(1-\gamma)]^{|\mathcal{S}|}$. For $V \in \mathcal{V}$, define its one-step action values in the usual way, $Q_V(s,a) = r(s,a) + \gamma\mathbb{E}[V(S') \mid s,a]$. Fixed-anchor PMVI applies the Power-Mirror backup using the *same behavior policy at every iteration*:

$$(\mathcal{T}_C V)(s) = C_{p_s,\pi_\beta(\cdot|s),b_s}(Q_V(s,\cdot)), \qquad V_{k+1} = \mathcal{T}_C V_k, \qquad V_0 = V^{\pi_\beta}, \tag{14}$$

and extracts $\pi_k^{\mathrm{fix}}(\cdot|s) = \nabla C_{p_s,\pi_\beta(\cdot|s),b_s}(Q_{V_k}(s,\cdot))$. The quantity

$$\mathrm{Res}_k^{\mathrm{PM}}(s) := (\mathcal{T}_C V_k)(s) - V_k(s) = V_{k+1}(s) - V_k(s)$$

is the *Power-Mirror Bellman residual*. It measures how far the current value is from the fixed point of the Power-Mirror operator; it is not the error to the unrestricted optimal Bellman equation.

Theorem 2 (Fixed-anchor Power-Mirror value iteration)**.** *On the bounded invariant value box, $\mathcal{T}_C$ is monotone and γ-contractive, with $\mathcal{T}^{\pi_\beta}V \le \mathcal{T}_C V \le \mathcal{T}_\beta^\star V$, where $\mathcal{T}_\beta^\star$ maximizes only over $\mathcal{A}_\beta(s)$. It has a unique fixed point V_C, and every finite iterate satisfies*

$$V^{\pi_\beta} \le V_k \le V_{k+1} \le V^{\pi_k^{\mathrm{fix}}} \le V_\beta^\star, \qquad \|V_k - V_C\|_\infty \le \gamma^k \|V^{\pi_\beta} - V_C\|_\infty. \tag{15}$$

The limiting policy $\pi_C = \nabla C(Q_{V_C})$ satisfies

$$V^{\pi_\beta} \le V_C \le V^{\pi_C} \le V_\beta^\star, \tag{16}$$

and both $\pi_k^{\mathrm{fix}}/\pi_\beta$ and π_C/π_β satisfy Theorem 1 statewise with $D = \bar{D}$ (or the smaller action-value range actually observed at that state).

Here $V_\beta^\star$ is the value of the best policy restricted to actions supported by π_β. Equation (15) gives an anytime population statement: stopping PMVI after any number of exact iterations still produces an extracted policy whose value is no worse than behavior. It also clarifies what the contraction does and does not provide. The contraction gives stable convergence to V_C; it does not claim that $V_C = V_\beta^\star$ or $V_C = V^\star$. Exact gain decompositions, Bellman-residual stopping bounds, and the proof based on the Bellman sandwich and convexity are in supplementary Theorem 9 and Corollary 4.

How far can the fixed-anchor solution be from unrestricted optimality? There are two possible losses. First, the optimal policy may use actions outside behavior support. Second, even when good actions are supported, π_β may assign them very little probability. The following bound makes the second loss quantitative and keeps it separate from state-distribution coverage.

Let $Q^\star$ and $V^\star$ be the unrestricted optimal action-value and value functions, and let $\pi^\star$ be an optimal policy. For a chosen tolerance $\varepsilon_s \geq 0$, define the behavior probability of unrestricted near-optimal actions, their shifted value scale, and the resulting local coverage defect by

$$W^\beta_{\star,\varepsilon}(s) = \Pr_{A\sim\pi_\beta(\cdot|s)}\{Q^\star(s,A) \geq V^\star(s) - \varepsilon_s\}, \tag{17}$$

$$X^\beta_\star(s) = b_s + V^\star(s) - \mathbb{E}_{A\sim\pi_\beta(\cdot|s)}Q^\star(s,A), \tag{18}$$

$$\bar{\Delta}^C_{\star,\varepsilon}(s) = \min\Big\{D^\beta_\star(s),\ X^\beta_\star(s)(1-(W^\beta_{\star,\varepsilon}(s))^{1/p_s}) + \varepsilon_s(W^\beta_{\star,\varepsilon}(s))^{1/p_s}\Big\}, \tag{19}$$

where $D^\beta_\star(s) = V^\star(s) - \min_{a\in\mathcal{A}_\beta(s)} Q^\star(s,a)$. The defect is small when π_β places substantial probability on actions whose unrestricted optimal value is near $V^\star(s)$. For an initial-state distribution ν, write

$$J_\nu(\pi) = \mathbb{E}_{S_0\sim\nu}V^\pi(S_0), \qquad d^\pi_\nu(s) = (1-\gamma)\sum_{t\geq 0}\gamma^t \Pr_{\nu,\pi}(S_t = s).$$

There exists a comparison policy π^{cmp} inside the same ratio box such that

$$J_\nu(\pi^\star) - J_\nu(\pi_C) \leq \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi^{\mathrm{cmp}}}_\nu}[\bar{\Delta}^C_{\star,\varepsilon}]. \tag{20}$$

If this occupancy is replaced by a dataset state distribution μ, Hölder's inequality introduces the explicit state-coverage factor $\|d^{\pi^{\mathrm{cmp}}}_\nu/\mu\|_{L^r(\mu)}$. This separation is intentional: bounding the action ratio at states seen in the data does not create coverage of states that the behavior policy rarely visits.

4 Monotone moving-anchor policy iteration

Fixed-anchor PMVI always compares a candidate with π_β, which preserves the original ratio box but generally leaves a permanent conservative bias. Moving-anchor PMPI removes that bias gradually. It starts from $\pi_0 = \pi_\beta$, exactly evaluates the current action-value function Q^{π_k}, and applies Power Mirror relative to the *current* policy. Let the full proposed update be $\widetilde{\pi}_{k+1}(\cdot|s) = \nabla C_{p_s,\pi_k(\cdot|s),b_s}(Q^{\pi_k}(s,\cdot))$. The algorithm may use a damping coefficient $\lambda_{k,s}$:

$$\pi_{k+1}(\cdot|s) = (1-\lambda_{k,s})\pi_k(\cdot|s) + \lambda_{k,s}\widetilde{\pi}_{k+1}(\cdot|s), \qquad \lambda_{k,s}\in[0,1]. \tag{21}$$

For the theorem, define

$$z_{k,s}(a) = b_s + Q^{\pi_k}(s,a) - \mathbb{E}_{A\sim\pi_k(\cdot|s)}Q^{\pi_k}(s,A),$$
$$m_{k,s} = \left(\mathbb{E}_{A\sim\pi_k(\cdot|s)}z_{k,s}(A)^{p_s}\right)^{1/p_s},$$
$$R_{k,s} = \mathbb{E}_{A\sim\pi_k(\cdot|s)}\left(z_{k,s}(A)/m_{k,s}\right)^{p_s-1}.$$

Theorem 3 (Monotone Power-Mirror policy iteration)**.** *Every exact step preserves support and has nonnegative local gain*

$$G_k(s) = \mathbb{E}_{A\sim\pi_{k+1}(\cdot|s)}Q^{\pi_k}(s,A) - \mathbb{E}_{A\sim\pi_k(\cdot|s)}Q^{\pi_k}(s,A) = \lambda_{k,s}\{m_{k,s} - b_s R_{k,s}\} \geq 0. \tag{22}$$

Consequently

$$V^{\pi_{k+1}} - V^{\pi_k} = (I-\gamma P^{\pi_{k+1}})^{-1}G_k \geq 0. \tag{23}$$

For fixed $1 < p_s < \infty$, fixed $b_s > \bar{D}$, and damping bounded away from zero, $V^{\pi_k} \uparrow V^\star_\beta$. Policies converge when the behavior-supported optimum is unique; under ties, every limit point is support-optimal.

Equation (23) is the standard policy-improvement identity written as a resolvent: nonnegative one-step advantage at every state produces nonnegative improvement in the full discounted value function. The theorem requires exact evaluation. With a learned critic, $G_k \geq 0$ for the estimated values does not by itself imply true improvement; the critic-error bound or the independent deployment shield is then needed.

The ratio theorem applies at each step to π_{k+1}/π_k, not directly to π_{k+1}/π_β. If the realized damped one-step bounds are $\ell_{j,s}$ and $u_{j,s}$, then

$$\prod_{j<k} \ell_{j,s} \leq \frac{\pi_k(a|s)}{\pi_\beta(a|s)} \leq \prod_{j<k} u_{j,s}. \tag{24}$$

This product is the algorithm's behavior-drift ledger. It makes the central tradeoff transparent: PMPI can eventually reach $V^\star_\beta$, but the price is an expanding cumulative ratio relative to the dataset-generating policy. Therefore, the deployment test must use the *final cumulative ratio*, not only the last one-step ratio. A line search that forces every iterate to remain in one permanent box around π_β preserves monotonicity, but it may stall at the box boundary. Supplementary Theorem 11 also shows that PMPI can stall if the damping coefficients are summable, if $p_k \downarrow 1$ too quickly, or if $b_k \uparrow \infty$ too quickly: every finite step may improve while the total available movement remains insufficient.

5 Strict high-probability deployment shield

The preceding theorems describe exact population updates. An offline learner instead estimates Q from finite data, so we add a separate accept-or-fallback decision. Split the trajectory data before training. Use the training split to choose the critic, all hyperparameters, and a single candidate policy π. Reserve n independent trajectories generated by the known behavior policy π_β only for certification.

Assume that the certification trajectories have length H and a common initial-state distribution ν. Early termination may be represented by an absorbing state with zero reward, provided $0 \in [r_-, r_+]$. Rewards satisfy $R_t \in [r_-, r_+]$. Assume that the candidate satisfies $\pi(\cdot|s) \ll \pi_\beta(\cdot|s)$ at every state reachable under the behavior policy. Thus the candidate assigns no probability to actions outside logging support. Its likelihood ratio is exactly evaluable and satisfies the following bounds $\pi_\beta(\cdot|s)$-almost surely:

$$0 < L \leq \pi(a|s)/\pi_\beta(a|s) \leq U < \infty \tag{25}$$

on every state–action pair reachable under π_β. Let $c = (r_- + r_+)/2$ be the midpoint of the reward interval and let $R_\Delta = r_+ - r_-$ be its width. For certification episode j, define the usual PDIS prefix likelihood ratio and the paired improvement estimate

$$W_{j,t} = \prod_{h=0}^{t} \frac{\pi(A_{j,h}|S_{j,h})}{\pi_\beta(A_{j,h}|S_{j,h})}, \qquad Z_j = \sum_{t=0}^{H-1} \gamma^t (W_{j,t} - 1)(R_{j,t} - c). \tag{26}$$

Subtracting the behavior term on the *same trajectory* makes Z_j a paired estimate rather than the difference of two separately estimated returns. The change-of-measure identity gives $\mathbb{E}_{\pi_\beta} Z_j = J_{\nu,H}(\pi) - J_{\nu,H}(\pi_\beta)$.

Because each action ratio lies in $[L, U]$, every prefix ratio lies in $[L^{t+1}, U^{t+1}]$. Therefore a deterministic range width for Z_j and an upper bound on the omitted infinite-horizon tail are

$$\mathcal{B}_H = R_\Delta \sum_{t=0}^{H-1} \gamma^t \max\{1 - L^{t+1}, U^{t+1} - 1\}, \tag{27}$$

$$\mathcal{T}_H = R_\Delta \frac{\gamma^H}{1-\gamma}. \tag{28}$$

Let $\bar{Z} = n^{-1} \sum_{j=1}^{n} Z_j$. For a confidence level $1 - \delta$ and a prespecified improvement margin $m \geq 0$, define the lower confidence bound and deployment rule

$$\mathrm{LCB}_H(\pi) = \bar{Z} - \mathcal{B}_H \sqrt{\frac{\log(1/\delta)}{2n}} - \mathcal{T}_H, \qquad \pi_{\mathrm{dep}} = \begin{cases} \pi, & \mathrm{LCB}_H(\pi) \geq m, \\ \pi_\beta, & \text{otherwise.} \end{cases} \tag{29}$$

Theorem 4 (Strict deployment safety)**.** *Conditional on the training data, with probability at least $1 - \delta$ over the certification fold, $J_\nu(\pi_{\mathrm{dep}}) \geq J_\nu(\pi_\beta)$. Whenever the candidate branch is selected, the same event gives $J_\nu(\pi) - J_\nu(\pi_\beta) \geq m$.*

Taking $m = 0$ certifies non-degradation; $m > 0$ certifies strict improvement whenever the candidate is accepted. Fallback deploys π_β and gives equality. The theorem follows from the PDIS identity, the deterministic prefix interval, Hoeffding's inequality, and the one-sided tail bound. It does not require a separate state-concentrability coefficient, because evaluation is performed directly under behavior trajectories. However, $\mathcal{B}_H$ makes the familiar horizon–overlap problem explicit: even a moderate $U > 1$ can make long-horizon importance weights large, causing the method to fall back frequently. The independence conditions are part of the theorem, not implementation details. The candidate, horizon, and form of the confidence bound must be selected before looking at the certification fold, unless simultaneous confidence correction is used. Reusing training trajectories, substituting estimated behavior probabilities without an additional error analysis, replacing the theoretical ratio bound by an observed sample maximum, clipping or self-normalizing the weights, or using an actor with an unknown normalizer breaks the theorem as stated. Supplementary Corollaries 9–10 give an empirical-Bernstein version and a corrected test for a finite candidate family.

6 Continuous ratio-preserving actor and pipeline

For actor fitting, let $\pi_* = r_*\bar{\pi}_\beta$ denote the ideal Power-Mirror target constructed relative to the actor's actual base distribution $\bar{\pi}_\beta$. For continuous actions, $\bar{\pi}_\beta(da|s)$ is a probability measure rather than a finite vector. Replacing the sums in (2) by integrals produces the same ideal Power-Mirror policy. Its density ratio (formally, its Radon–Nikodym derivative) with respect to this same base is

$$\frac{d\pi_*}{d\bar{\pi}_\beta}(a|s) = \left(\frac{b + Q(s,a) - \mathbb{E}_{A\sim\bar{\pi}_\beta(\cdot|s)}Q(s,A)}{m_p(s)}\right)^{p-1} + 1 - R(s), \tag{30}$$

where $m_p(s)$ and $R(s)$ are also computed under $\bar{\pi}_\beta$. Theorem 1 applies relative to this base, up to base-probability-zero sets. When the moments defining the ideal ratio can be evaluated exactly and the base policy can be sampled, this ideal policy can be sampled by rejection sampling. In large-scale learning, however, it is useful to fit an amortized actor. An unconstrained actor trained by weighted maximum likelihood need not preserve either the exact normalization of the ideal ratio or its pointwise endpoints.

We therefore build the constraint into the actor parameterization. Let $\bar{\pi}_\beta$ denote the base behavior distribution actually used by the actor; it equals π_β when the true behavior density and sampler are available, and otherwise may be a fitted behavior model. Choose desired endpoints $0 < L < 1 < U$ and set $\tau = (1 - L)/(U - L)$. A neural network first produces a bounded score $h_\theta(s, a) \in [0, 1]$. A statewise calibration step converts it to $g_\theta(s, a) \in [0, 1]$ satisfying $\mathbb{E}_{A\sim\bar{\pi}_\beta(\cdot|s)}g_\theta(s, A) = \tau$. Define

$$r_\theta(s,a) = L + (U - L)g_\theta(s,a), \qquad \pi_\theta(da|s) = r_\theta(s,a)\bar{\pi}_\beta(da|s). \tag{31}$$

The calibration condition implies $\mathbb{E}_{\bar{\pi}_\beta} r_\theta = 1$, so the actor is already normalized; there is no unknown state-dependent partition function.

Theorem 5 (Ratio-preserving continuous actor)**.** *For every actor parameter,* (31) *is normalized and $L \le d\pi_\theta/d\bar{\pi}_\beta \le U$. The unrestricted calibrated-multiplier family represents every normalized target whose density ratio to $\bar{\pi}_\beta$ lies in $[L, U]$. In particular, it represents the ideal Power-Mirror target constructed relative to the same base whenever the chosen endpoints contain that target's ratio range. Propose $A \sim \bar{\pi}_\beta(\cdot|s)$ and accept it with probability $r_\theta(s, A)/U$; accepted actions are exactly distributed as $\pi_\theta(\cdot|s)$, the mean acceptance probability is $1/U$, and the expected number of proposals is U.*

The ratio guarantee holds for every θ, including an early-stopped or imperfectly optimized actor. Fitting error can reduce policy improvement, but it cannot create an arbitrarily large density-ratio spike. We fit the multiplier within this constrained family by minimizing $\mathrm{KL}(\pi_*\|\pi_\theta) = \mathbb{E}_{\bar{\pi}_\beta}[r_* \log(r_*/r_\theta)]$. If $Q(s, \cdot)$ has range D and the ideal policy improves its expected value by G_*, Pinsker's inequality gives

$$\mathbb{E}_{\pi_\theta}Q - \mathbb{E}_{\bar{\pi}_\beta}Q \ge G_* - D\sqrt{\mathrm{KL}(\pi_*\|\pi_\theta)/2}. \tag{32}$$

Thus the actor retains positive improvement whenever its approximation error is small enough relative to G_*. Exact calibration is required for the sharp endpoints; supplementary Lemma 1 gives enlarged, still certified endpoints under bounded calibration error. If $\bar{\pi}_\beta$ is learned rather than equal to the true behavior policy, then the certificate is only relative to $\bar{\pi}_\beta$. Transferring it to π_β requires an additional pointwise bound on $d\bar{\pi}_\beta/d\pi_\beta$.

End-to-end certified pipeline. The complete offline procedure is: (1) divide full behavior-policy trajectories into training and independent certification folds; (2) on training data, estimate a bounded-range critic and choose p and b from a desired behavior-ratio budget; (3) run fitted PMVI when permanent proximity to π_β is required, or fitted PMPI when gradual movement toward the support-constrained optimum is preferred; (4) extract the exact finite-action policy or fit the calibrated continuous actor; (5) audit the *actual final* likelihood ratio relative to the true π_β; and (6) apply Theorem 4. Critic learning constructs a promising candidate. The independent accept-or-fallback test, not the training objective, supplies the finite-sample non-degradation statement.

7 Experiments

Experimental questions. We investigate how behavior-relative update constraints affect empirical return and cost through three questions:

1. **How does data coverage affect relative performance?** How do fixed-anchor, moving-anchor, and greedy updates compare as consequential transitions become better represented in the offline data?
2. **How does update strength affect return and cost?** At fixed η, how does the power p change the return–cost tradeoff, and how does this tradeoff depend on data coverage?
3. **How effective are the fitted continuous-action implementations?** How do their returns and cost diagnostics compare with offline-RL baselines as dataset size and composition vary?

Wet Chicken addresses the first two questions, while NeoRL and offline MetaDrive address the third. These experiments evaluate learned policies without applying the independent deployment shield of Theorem 4.

Experimental setup. We evaluate whether PMVI improves on the behavior policy when offline data provide limited coverage of alternative actions. Baselines include behavior, AWAC (Nair et al., 2020), IQL (Kostrikov et al., 2022), KL-fixed, KL-moving, and PMPI, together with greedy improvement in Wet Chicken. Curves report means and pointwise 95% confidence intervals over 100 offline datasets for Wet Chicken and ten training seeds for continuous control. Costs are evaluation metrics, not training penalties, and we evaluate learned policies without the certification test. Appendix L provides implementation and evaluation details.

Wet Chicken: return and cost across dataset sizes. Wet Chicken rewards a canoeist for approaching a waterfall, while stochastic currents can cause a fall and reset. We use a discrete version (Tresp, 1994; Hans & Udluft, 2009) that permits exact evaluation of return and waterfall cost. Across dataset sizes, PMVI remains close to behavior in both return and cost, with larger p allowing progressively more improvement at the price of higher cost (Figure 1). In contrast, PMPI can move farther from behavior and achieves substantially higher return, while still keeping waterfall cost below greedy improvement and KL-moving over much of the sweep. This illustrates the intended distinction between the two variants: PMVI provides conservative behavior-relative improvement, whereas PMPI trades the permanent behavior constraint for greater policy improvement.

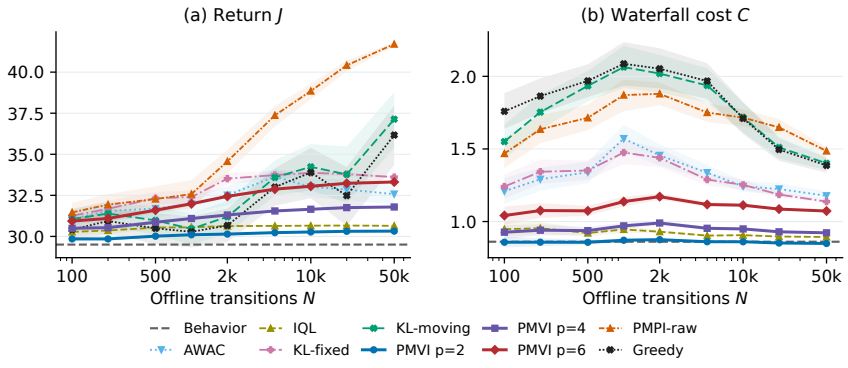

Figure 1: Wet Chicken: return J and waterfall cost C versus dataset size, with exploration $\epsilon = 0.2$. Dashed gray lines indicate behavior. PMVI uses $p = 2, 4, 6$.

NeoRL: improvement across dataset sizes. Hopper and HalfCheetah require coordinated control to move forward efficiently, with Hopper additionally requiring balance. We use NeoRL's medium-quality datasets (Qin et al., 2022), which involve less exploration than the stochastic-policy and replay-buffer datasets in D4RL MuJoCo (Fu et al., 2020), and vary $N \in \{100, 1000, 10000\}$ trajectories at fixed dataset quality. Across the sweep, all PMVI powers outperform behavior and the other plotted methods in Hopper, while remaining competitive with the strongest baselines in HalfCheetah (Figure 2). PMVI also avoids the severe losses of PMPI and KL-moving in HalfCheetah, including at the largest dataset size. Although IQL attains higher mean return there, PMVI remains strong across both tasks and data sizes, consistent with fixed-reference updates trading some improvement for protection against the larger losses induced by repeatedly updated references.

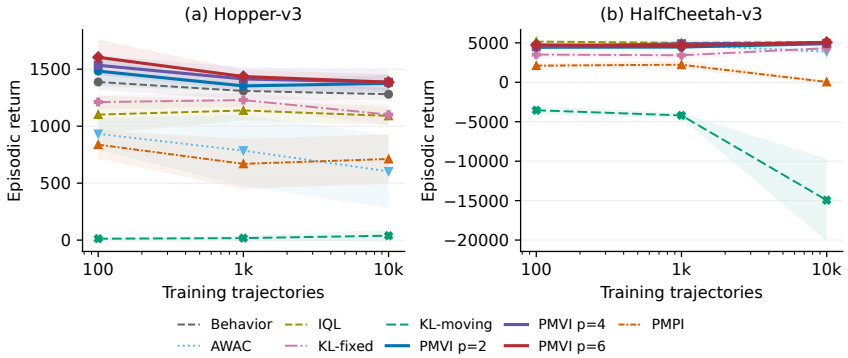

Figure 2: NeoRL: episodic return versus training trajectories on Hopper-v3 and HalfCheetah-v3. PMVI uses $p = 2, 4, 6$; PMPI uses $p = 2$.

MetaDrive: task success and off-road failures. MetaDrive rewards progress toward a destination, but successful driving also requires avoiding collisions and leaving the road. Using DSRL's `hardmean` dataset (Liu et al., 2024), we retain all zero-cost episodes and increase the retained fraction q_{risk} of positive-cost episodes, thereby varying both dataset size and coverage of risky behavior. As this coverage improves, PMVI achieves substantially higher return and task success while keeping its off-road rate far below PMPI and KL-moving (Figure 3). When positive-cost episodes are scarce, PMVI instead remains conservative and rarely completes the task. This pattern highlights the intended behavior of the fixed anchor: PMVI exploits additional coverage when the data support improvement, while limiting policy changes in regions where the offline data remain weak.

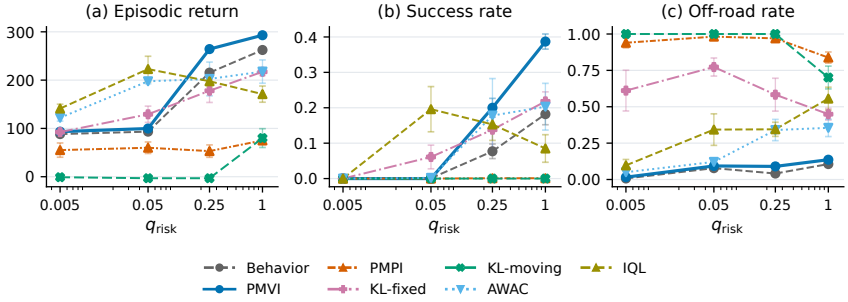

Figure 3: MetaDrive: return, success rate, and off-road rate as the retained fraction of positive-cost episodes increases. PMVI and PMPI use $p = 2$.

8 Limitations and Conclusion

Our guarantees rely on population and certification assumptions, while practical performance depends on data coverage and estimation accuracy; conservative updates may therefore limit improvement. Power Mirror links policy improvement and deployment through behavior-relative reweighting: the critic range, power, and shift determine sharp ratio bounds controlling critic-error transfer and importance weights, while bounded actors preserve them. Fixed-anchor PMVI maintains behavior proximity, whereas under exact evaluation and persistent updates, PMPI approaches the support-optimal value while accumulating behavior drift. Overall, the framework makes explicit the tradeoff in departing from the behavior policy.

AI USE STATEMENT

Generative AI tools were used to help develop and organize theoretical formulations, formulate and stress-test mathematical claims, assist with proof drafting, critique methodology and experimental design, identify and summarize related literature, edit prose, format references, and prepare the LaTeX source. AI-assisted derivations and empirical summaries were checked against proofs and source artifacts where applicable. The authors take responsibility for the final content, including all AI-assisted text, claims, proofs, and artifacts.

ETHICS STATEMENT

This is a theoretical and methodological study using no human-subject or sensitive personal data. Its safety claims are conditional on explicit assumptions, especially known logging propensities, independent certification episodes, bounded rewards, and an executable behavior fallback. Deploying the method without these conditions could create unjustified confidence and is outside the claimed guarantee.

REPRODUCIBILITY STATEMENT

All assumptions and main constructions are stated in the main paper. The post-reference supplement contains complete derivations, proofs, fitted-algorithm details, limitations, and experimental settings. Appendix L documents dataset construction, training configurations, evaluation procedures, and uncertainty calculations.

REFERENCES

A. Abdolmaleki, J. T. Springenberg, Y. Tassa, R. Munos, N. Heess, and M. Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018.

R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Boris Belousov and Jan Peters. f-divergence constrained policy improvement, 2018. URL `https://arxiv.org/abs/1801.00056`. Version 2, revised April 4, 2018.

M. Blondel, A. F. T. Martins, and V. Niculae. Learning with Fenchel–Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

J. Chen and N. Jiang. Offline reinforcement learning under value and density-ratio realizability: The power of gaps. In *Conference on Uncertainty in Artificial Intelligence*, 2022.

T. Dam, P. Klink, C. D'Eramo, J. Peters, and J. Pajarinen. Generalized mean estimation in Monte-Carlo tree search. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 2397–2404, 2020.

T. Dam, O.-A. Maillard, and E. Kaufmann. Power mean estimation in stochastic Monte-Carlo tree search. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 894–918, 2024.

J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 49(3):1378–1406, 2021.

J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

S. Fujimoto and S. S. Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

S. Fujimoto, D. Meger, and D. Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, 2019.

M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169, 2019.

Alexander Hans and Steffen Udluft. Efficient uncertainty propagation for reinforcement learning with limited data. In *International Conference on Artificial Neural Networks*, pp. 70–79. Springer, 2009.

I. Kostrikov, A. Nair, and S. Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022.

A. Kumar, J. Fu, G. Tucker, and S. Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, 2019.

A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, 198:1059–1106, 2023.

R. Laroche, P. Trichelair, and R. Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661, 2019.

Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill. Provably good batch off-policy reinforcement learning without great exploration. In *Advances in Neural Information Processing Systems*, 2020.

Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, and Ding Zhao. Datasets and benchmarks for offline safe reinforcement learning. *Journal of Data-centric Machine Learning Research*, 1(12):1–29, 2024. URL https://openreview.net/forum?id=ajf5PE3Z7w.

A. F. T. Martins, M. Treviso, A. Farinhas, P. M. Q. Aguiar, M. A. T. Figueiredo, M. Blondel, and V. Niculae. Sparse continuous distributions and Fenchel–Young losses. *Journal of Machine Learning Research*, 23(257):1–74, 2022.

A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory*, 2009.

A. Nair, A. Gupta, M. Dalal, and S. Levine. AWAC: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

H. Namkoong and J. C. Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, 2017.

G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

X. B. Peng, A. Kumar, G. Zhang, and S. Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

J. Peters, K. Mülling, and Y. Altun. Relative entropy policy search. In *AAAI Conference on Artificial Intelligence*, 2010.

M. Petrik, M. Ghavamzadeh, and Y. Chow. Safe policy improvement by minimizing robust baseline regret. In *Advances in Neural Information Processing Systems*, 2016.

Rong-Jun Qin, Xingyuan Zhang, Songyi Gao, Xiong-Hui Chen, Zewen Li, Weinan Zhang, and Yang Yu. NeoRL: A near real-world benchmark for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24753–24765, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9cd828eb8dc81a84fb6bf89a94263e1b-Abstract-Datasets_and_Benchmarks.html.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, 2015.

Volker Tresp. The wet game of chicken. Technical report, Siemens AG, CT IC 4, 1994.

R. Wang, D. P. Foster, and S. M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2021.

Y. Wu, G. Tucker, and O. Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

W. Zhan, B. Huang, A. Huang, N. Jiang, and J. D. Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2730–2775, 2022.

W. Zhan, S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023.

Appendix of "Power Mirror: Offline Policy Improvement with Bounded Reweighting"

TABLE OF CONTENTS

A Background and related work **15**

B Power-mirror backup and fixed-shift domain **16**

B.1 Why the shift matters 17

B.2 How to interpret the two parameters 17

B.3 Bellman-compatible properties 17

C Induced policy and density-ratio envelope **17**

D Coverage-aware approximation to greedy improvement **19**

E Behavior-anchored population control **21**

E.1 Population Power-Mirror Value Iteration 21

E.2 End-to-end population improvement 22

E.3 Support-relative coverage-dependent optimality 23

E.4 Unrestricted near-optimality: action and state coverage 24

F Moving-anchor Power-Mirror Policy Iteration **25**

F.1 Population updates, monotone improvement, and ratio control 25

F.2 Finite-iterate coverage-dependent near-optimality 28

G Offline safe-improvement template **31**

H Strict high-probability deployment shield **32**

I An interpretable special case: $p = 2$ **36**

I.1 Quadratic form and dual interpretation 36

I.2 Quadratic improvement efficiency and critic sensitivity 37

J Continuous-action power mirror **39**

J.1 Continuous-action formulation and guarantees 39

J.2 Ratio-preserving continuous actors 40

K Algorithmic instantiations **44**

K.1 Fitted PMVI for finite actions 44

K.2 Fitted PMPI for finite actions 44

K.3 Continuous ratio-preserving Power-Mirror actor 45

K.4 Certified end-to-end deployment 47

L Experimental settings **47**

L.1 Tabular Wet Chicken 48

L.2 NeoRL 48

L.3 MetaDrive 49

M Positioning and limitations **49**

N Conclusion **50**

O Proofs **50**

O.1 Proof of Bellman-compatible scaffolding 50

O.2 Proof of the induced policy formula 51

O.3 Proof of the sharp density-ratio theorem and its corollaries 51

O.4 Proof of the near-optimal coverage bound 53

O.5 Proof of the population convergence theorem 53

O.6 Proof of end-to-end population control 54

O.7 Proof of coverage-dependent support optimality 54

O.8 Proof of coverage-dependent unrestricted near-optimality 55

O.9 Proof of monotone moving-anchor policy iteration 56

O.10 Proof of finite-iterate PMPI coverage near-optimality 57

O.11 Proof of exact reference-policy improvement 58

O.12 Proof of the behavior-ratio critic-error bound 58

O.13 Proof of strict high-probability deployment safety 59

O.14 Proof of the quadratic dual 60

O.15 Continuous-action proofs 60

O.16 Proofs for ratio-preserving continuous actors 61

A Background and related work

Regularized Bellman operators. Many RL algorithms replace the hard maximum in the Bellman optimality operator by a smoother policy optimization problem of the form

$$S(q) = \max_{\pi \in \Delta_K} \{\langle \pi, q \rangle - \Omega(\pi)\},$$

where the vector $q = (q_1, \ldots, q_K)$ contains the action values at one state, π is a candidate action distribution, and Ω penalizes aggressive policies. The general theory covers modified policy iteration, Fenchel duality, error propagation, soft Q-learning, and SAC (Geist et al., 2019; Neu et al., 2017). We use this theory for standard Bellman properties. Our contribution is different: the gradient of our backup is an improved policy whose probability ratio to a chosen reference policy is available in closed form. We denote an arbitrary reference by π_{ref} and use the offline behavior policy π_β as the reference in all offline-RL results.

Moving-reference and mirror-descent policy iteration. Policy mirror descent and generalized policy mirror descent already analyze policy optimization in which the current policy becomes the next reference (Lan, 2023; Zhan et al., 2023); regularized modified policy iteration is also classical (Geist et al., 2019). We therefore do not claim generic monotone policy iteration as new. Our moving-reference result is specific to the power-mirror map: it gives the exact improvement at each state, sharp ratios between consecutive policies, a condition for reaching the best policy on behavior support, and an explicit ledger showing how local ratios multiply relative to the original data-generating policy. This last issue separates moving-anchor PMPI from the fixed-behavior PMVI target studied first.

Offline RL: pessimism and behavior constraints. Offline RL controls distribution shift in several ways. BCQ restricts action selection toward the batch support (Fujimoto et al., 2019); BEAR constrains the learned policy to reduce out-of-distribution bootstrapping error (Kumar et al., 2019); BRAC studies behavior-regularized actor–critic methods (Wu et al., 2019); CQL learns pessimistic Q-values (Kumar et al., 2020); and TD3+BC adds behavior cloning to a standard actor objective (Fujimoto & Gu, 2021). Power mirror is closest to behavior-constrained policy improvement. The difference is that the constraint is not only a penalty in a training loss: the resulting policy itself satisfies a closed-form, action-by-action probability-ratio bound.

KL, REPS, MPO, and advantage-weighted regression. REPS and MPO improve a policy subject to a relative-entropy constraint (Peters et al., 2010; Abdolmaleki et al., 2018). AWR and AWAC fit an actor by advantage-weighted regression, which typically produces exponential target weights (Peng et al., 2019; Nair et al., 2020). If π_{ref} denotes the reference policy, an ideal exponential update has the form

$$\pi_\tau(i) \propto \pi_{\mathrm{ref}}(i) \exp(q_i/\tau).$$

Its probability ratio depends exponentially on the Q-value range divided by the temperature τ. Power mirror instead applies a power law to positive shifted Q-values, and its ratio bound depends on the power p and the dimensionless quantity D/b, where D bounds the Q-value range and b is the positive shift. This does not imply that KL methods lack safety mechanisms. Rather, power mirror provides a particular sharp, two-sided ratio interval that can be selected directly through p once b has been fixed. Moreover, a small average KL or weighted-regression error does not by itself guarantee a pointwise density-ratio bound. This is why our continuous actor keeps the reference distribution inside the deployed policy parameterization.

Calibrated exponential updates and quadratic geometry. For an exact exponential update $\pi_\tau(a) = w(a)e^{Q(a)/\tau}/\mathbb{E}_w e^{Q/\tau}$, a critic-span bound $\mathrm{span}(Q) \le D$ implies

$$e^{-D/\tau} \le \frac{\pi_\tau(a)}{w(a)} \le e^{D/\tau}. \tag{33}$$

Indeed, the normalizer lies between the minimum and maximum of $e^{Q/\tau}$. Thus calibrated exponential updates also provide pointwise guarantees. Pearson-χ^2 and broader f-divergence policy updates have also been studied previously (Belousov & Peters, 2018). The quadratic Power-Mirror policy coincides with a χ^2-constrained improvement step at its induced radius, and we use this equivalence

as an interpretation. Our analysis concerns the centered power family, sharp ratio calibration, and the consequences of fixed and moving references. At $p = 2$, the update attains maximal predicted gain for its achieved χ^2 movement and has a controlled response to critic perturbations. These facts establish neither universal dominance over exponential updates nor a new quadratic trust-region geometry. Bounded actors and deployment certification can be applied to several improvement rules; their benefits must be separated from those of the target update itself.

Safe policy improvement and density ratios. Safe policy-improvement methods aim to beat, or at least not degrade, a known baseline using fixed data. SPIBB returns to the baseline when action counts are insufficient (Laroche et al., 2019); robust baseline-regret methods provide related guarantees (Petrik et al., 2016); and high-confidence policy improvement deploys a candidate only when an off-policy confidence interval certifies it (Thomas et al., 2015). Density ratios and state concentrability are also standard tools in offline RL theory (Liu et al., 2020; Zhan et al., 2022; Chen & Jiang, 2022; Wang et al., 2021). Our contribution is to separate two kinds of coverage that are often conflated: which actions the behavior policy takes at a given state, and which states a comparator policy visits. The power-mirror ratio theorem controls the first quantity and transfers Q-function error from the behavior distribution to the improved policy. The deployment shield then uses the same ratio constants inside a classical high-confidence accept-or-fallback procedure.

Power means in planning. Power-UCT uses power means as tree backups that interpolate between an average and a maximum (Dam et al., 2020); stochastic Power-UCT studies the corresponding planning estimator (Dam et al., 2024). We do not claim that power means are new. We ask a different question: what policy is obtained by differentiating a reference-weighted power backup, and what can be guaranteed when the reference is the behavior policy that generated the offline dataset?

Continuous regularized prediction maps. Fenchel–Young and regularized-prediction frameworks include many discrete and continuous prediction maps (Blondel et al., 2020; Martins et al., 2022). Continuous regularized prediction is therefore not itself new. Our result is the particular continuous density-ratio formula for the reference-weighted power backup and the way it is preserved by the deployed offline-RL actor.

B Power-mirror backup and fixed-shift domain

We begin at a single state s with K available actions $a_1, \dots, a_K$. Throughout this section,

$$q_i := Q(s, a_i)$$

is the Q-value of action a_i at that state. Thus q is a vector of statewise Q-values, not a separate value function or an abstract score. Let $\pi_{\mathrm{ref}}(\cdot|s)$ be any reference action distribution. To keep the formulas short, write

$$w_i := \pi_{\mathrm{ref}}(a_i|s), \qquad w = (w_1, \dots, w_K) \in \Delta_K^\circ,$$

where $\Delta_K = \{\pi \in \mathbb{R}_+^K : \sum_i \pi_i = 1\}$ and the superscript $\circ$ means that every reference probability is positive. In offline RL, we set $\pi_{\mathrm{ref}} = \pi_\beta$, the behavior policy that generated the dataset. Define

$$\mu_w(q) = \sum_i w_i q_i, \qquad \delta_i(q) = q_i - \mu_w(q), \qquad z_i(q) = b + \delta_i(q).$$

Here $\mu_w(q) = \mathbb{E}_{a\sim\pi_{\mathrm{ref}}}[Q(s, a)]$ is the reference policy's average Q-value, $\delta_i(q)$ is the centered Q-value of action a_i, and $z_i(q)$ is that centered value shifted upward by $b > 0$. For a power $1 \le p < \infty$, define the power-mirror backup

$$C_{p,w,b}(q) = \mu_w(q) + \|z(q)\|_{p,w} - b, \qquad \|z\|_{p,w} = \left(\sum_i w_i z_i^p\right)^{1/p}. \tag{34}$$

For $p = \infty$, set $\|z\|_{\infty,w} = \max_i z_i$. The two endpoints have an immediate RL meaning: $p = 1$ returns the Q-value obtained by sampling from π_{ref}, while $p = \infty$ returns the largest Q-value.

B.1 WHY THE SHIFT MATTERS

Let

$$\operatorname{span}(q) := \max_i q_i - \min_i q_i \le D < b. \tag{35}$$

The quantity $\operatorname{span}(q)$ is the range of the statewise Q-values, and D is a known upper bound on this range. Requiring the fixed shift b to exceed D gives

$$b - D \le z_i(q) \le b + D, \qquad i = 1, \dots, K. \tag{36}$$

so every shifted Q-value is strictly positive. The word "fixed" matters: if b is recomputed from the current Q-vector inside each Bellman update, the algorithm no longer iterates the fixed operator in (34). The contraction and monotonicity results below require b to be selected in advance and held fixed throughout the value-iteration recursion being analyzed.

Remark 1 (How to choose a conservative fixed shift). If rewards are bounded, $r(s, a) \in [r_{\min}, r_{\max}]$, then a crude global action-value span bound is

$$\bar{D} = \frac{r_{\max} - r_{\min}}{1 - \gamma}.$$

One can set $b = \kappa \bar{D}$ with $\kappa > 1$. A larger b keeps the improved policy closer to the reference policy, but also makes the backup more conservative. An implementation may estimate a high quantile of the observed Q-value ranges and add a safety margin. The population theorem, however, is stated using a fixed a priori bound.

B.2 HOW TO INTERPRET THE TWO PARAMETERS

Both p and b control how far the backup moves from reference-policy evaluation toward greedy maximization. For the especially simple case $p = 2$,

$$C_{2,w,b}(q) = \mu_w(q) + \sqrt{b^2 + \operatorname{Var}_w(q)} - b, \tag{37}$$

where $\operatorname{Var}_w(q)$ is the variance of the statewise Q-values under π_{ref}. Sending $b \to \infty$ collapses the update to the reference-policy average. We therefore use the following interpretation:

fix b from a span margin, and use p as the behavior-ratio budget.

Define the dimensionless Q-range-to-shift ratio

$$\eta = D/b \in [0, 1). \tag{38}$$

All policy-ratio bounds below depend on the Q-values only through η.

B.3 BELLMAN-COMPATIBLE PROPERTIES

Proposition 2 (Bellman-compatible scaffolding)**.** *On the fixed domain* (35)*, $C_{p,w,b}$ is convex, translation equivariant, monotone, and ℓ_∞ nonexpansive for $1 \le p \le \infty$. The endpoints are*

$$C_{1,w,b}(q) = \mu_w(q), \qquad C_{\infty,w,b}(q) = \max_i q_i,$$

and $C_{p,w,b}(q)$ is nondecreasing in p.

These properties ensure that the backup can replace a standard Bellman backup without losing monotonicity or contraction. They are standard for convex regularized Bellman operators; the new result begins with the policy obtained by differentiating this particular backup.

C INDUCED POLICY AND DENSITY-RATIO ENVELOPE

The backup is differentiable for a finite power $1 < p < \infty$. Its gradient with respect to the Q-vector is a probability distribution, which we use as the improved policy. To state this policy compactly, define

$$m_p(q) = \|z(q)\|_{p,w}, \qquad a_i(q) = \left(\frac{z_i(q)}{m_p(q)}\right)^{p-1}, \qquad R(q) = \sum_i w_i a_i(q). \tag{39}$$

Here $m_p(q)$ is the power mean already used in the backup, $a_i(q)$ is the raw multiplicative power weight for action i (not the action itself), and $R(q)$ is the average of these weights under π_{ref}.

Theorem 6 (Induced power-mirror policy)**.** *For $1 < p < \infty$ and $z_i(q) > 0$, $C_{p,w,b}$ is differentiable and*

$$\pi_{p,i}(q) := \nabla_i C_{p,w,b}(q) = w_i\big(a_i(q) + 1 - R(q)\big). \tag{40}$$

Moreover $0 < R(q) \le 1$, and hence $\pi_p(q) \in \Delta_K$.

Equation (40) has a useful interpretation. Start with the reference probability $w_i = \pi_{\mathrm{ref}}(a_i|s)$, multiply it by the Q-dependent power weight $a_i(q)$, and then add the same correction $1 - R(q)$ inside every likelihood ratio. The correction makes the policy sum to one. Because $R(q) \le 1$, it also keeps every improved probability nonnegative. Normalization and positivity are therefore consequences of the backup, not a separate actor projection.

Theorem 7 (Sharp density-ratio envelope and oscillation)**.** *For $1 < p < \infty$, assume that the statewise Q-value range satisfies $\mathrm{span}(q) \le D < b$, let $\eta = D/b$, and write $\alpha = p - 1$. Define*

$$L_{p,\eta} = (1-\eta)^\alpha, \qquad U_{p,\eta} = (1+\eta)^\alpha, \qquad K_{p,\eta} = L_{p,\eta}^{-1}, \tag{41}$$

and

$$H_{p,\eta} = \begin{cases} 1 - (1-\eta)^\alpha, & 0 < \alpha \le 1, \\ (1+\eta)^\alpha - 1, & \alpha \ge 1. \end{cases} \tag{42}$$

For the actionwise probability ratio $r_i = \pi_{p,i}(q)/\pi_{\mathrm{ref}}(a_i|s) = \pi_{p,i}(q)/w_i$, every action satisfies

$$L_{p,\eta} \le r_i \le U_{p,\eta}. \tag{43}$$

More strongly,

$$\frac{\max_i r_i}{\min_i r_i} \le K_{p,\eta}, \qquad \max_i r_i - \min_i r_i \le H_{p,\eta}, \tag{44}$$

$$\max_i |r_i - 1| \le H_{p,\eta}, \qquad \max_i \log r_i - \min_i \log r_i \le \log K_{p,\eta}. \tag{45}$$

The constants $L_{p,\eta}$, $U_{p,\eta}$, $H_{p,\eta}$, and $K_{p,\eta}$ are each optimal as uniform bounds: no strictly stronger constant is valid over all positive reference policies and all Q-value vectors satisfying the range condition, even when there are only two actions.

The theorem gives more than an average trust region. It bounds the change in probability separately for every action. The lower endpoint $L_{p,\eta}$ prevents the update from deleting an action that the reference can take, while the upper endpoint $U_{p,\eta}$ prevents it from placing too much probability on an action that is rare under the reference. The identity $\sum_i w_i z_i = b$ is essential for obtaining these sharp constants; bounding each z_i and the power mean separately would ignore their shared mean and yield looser bounds. Both endpoints remain unimprovable even in a two-action problem.

Corollary 1 (Trust-region and divergence consequences)**.** *If $\eta = 0$, then $\pi_p = w$ and all divergences below vanish. Suppose $\eta > 0$, let $L = L_{p,\eta}$, $U = U_{p,\eta}$, $H = H_{p,\eta}$, and let f be convex with $f(1) = 0$. Then*

$$D_f(\pi_p\|w) := \mathbb{E}_w f\left(\frac{\pi_p}{w}\right) \le \frac{U-1}{U-L} f(L) + \frac{1-L}{U-L} f(U). \tag{46}$$

In particular,

$$D_\infty(\pi_p\|w) \le \log U, \qquad D_\infty(w\|\pi_p) \le -\log L, \tag{47}$$

$$\mathrm{TV}(\pi_p, w) \le \min\left\{\frac{(1-L)(U-1)}{U-L}, \frac{H}{4}\right\}, \qquad \chi^2(\pi_p\|w) \le \min\left\{(1-L)(U-1), \frac{H^2}{4}\right\}. \tag{48}$$

Writing $X_i = z_i/b$ and $B = (\mathbb{E}_w X^p)^{(p-1)/p}$, the latter two quantities have the exact map-specific identities

$$\mathrm{TV}(\pi_p, w) = \frac{\mathbb{E}_w|X^{p-1} - \mathbb{E}_w X^{p-1}|}{2B}, \qquad \chi^2(\pi_p\|w) = \frac{\mathrm{Var}_w(X^{p-1})}{B^2}. \tag{49}$$

Moreover,

$$\mathrm{KL}(\pi_p\|w) \le \frac{(U-1)L\log L + (1-L)U\log U}{U-L}, \tag{50}$$

$$\mathrm{KL}(w\|\pi_p) \le -\frac{(U-1)\log L + (1-L)\log U}{U-L}. \tag{51}$$

Corollary 1 translates the same pointwise interval into familiar policy distances. Here D_∞ is the worst-case log likelihood ratio, TV is total variation distance, χ^2 is chi-square divergence, and KL is Kullback–Leibler divergence. The first bound in (46) applies to any convex f-divergence. These are consequences of the stronger pointwise statement, not additional tuning constraints.

Offline-RL meaning. Set the reference equal to the behavior policy, that is, $w = \pi_{\mathrm{ref}}(\cdot|s) = \pi_\beta(\cdot|s)$. Then (43) says

$$L_{p,\eta_s}\pi_\beta(a|s) \le \pi_p(a|s) \le \rho_{p,\eta_s}\pi_\beta(a|s), \qquad \rho_{p,\eta_s} := U_{p,\eta_s} = (1+\eta_s)^{p-1}. \tag{52}$$

If an action occurs rarely in the offline data, its probability under the improved policy remains small up to the known factor ρ_{p,η_s}. Conversely, every action that the behavior policy can take keeps positive probability for every finite p. Thus the update cannot silently turn weak evidence for a rare action into near-deterministic deployment.

Choosing p from a target ratio cap. For an upper density-ratio budget $\rho_+ > 1$, the sharp uniform rule is

$$p \le 1 + \frac{\log \rho_+}{\log(1+\eta)}. \tag{53}$$

If a symmetric likelihood-ratio budget $\rho^{-1} \le \pi_p/w \le \rho$ is required, the lower side is more restrictive and the sharp uniform rule is

$$p \le 1 + \frac{\log \rho}{-\log(1-\eta)}. \tag{54}$$

Thus one can choose the largest power compatible with either a one-sided upper probability-ratio budget or a two-sided trust region, without solving another constrained optimization problem. When $\eta = 0$, all Q-values are equal on reference support and $\pi_p = \pi_{\mathrm{ref}}$ for every finite p, so no restriction is needed.

Corollary 2 (Finite-horizon pathwise ratio control)**.** *Suppose power mirror is applied at each state with lower and upper action-ratio bounds L_s and U_s. Let $\mathbb{P}_\pi^{(T)}$ and $\mathbb{P}_w^{(T)}$ denote the length-T trajectory distributions under the improved and reference policies, respectively, using the same initial state distribution and transition dynamics. A trajectory is*

$$\tau_T = (s_0, a_0, \ldots, s_{T-1}, a_{T-1}, s_T).$$

On every behavior-supported trajectory,

$$\prod_{t=0}^{T-1} L_{s_t} \le \frac{d\mathbb{P}_\pi^{(T)}}{d\mathbb{P}_w^{(T)}} \le \prod_{t=0}^{T-1} U_{s_t}. \tag{55}$$

In particular, uniform statewise bounds give $L^T \le d\mathbb{P}_\pi^{(T)}/d\mathbb{P}_w^{(T)} \le U^T$. This is useful for bounding finite-horizon importance weights, but it also exposes their exponential dependence on horizon. It is not a horizon-free guarantee on state-distribution shift.

Comparison with exponential tilting. A KL-to-reference or advantage-weighted policy often has the form

$$\pi_\tau(i) \propto w_i \exp(q_i/\tau).$$

Under a bounded Q-value range, this exact exponential update can also have a bounded ratio; its bound depends on the Q-value range divided by the temperature. Power mirror uses a power law in positive shifted Q-values, and (52) instead depends on D/b. This is the relevant structural difference from MPO/AWR/AWAC-style exponential weights (Abdolmaleki et al., 2018; Peng et al., 2019; Nair et al., 2020); we do not claim that exact KL updates can never be behavior constrained.

D Coverage-aware approximation to greedy improvement

Let

$$m(q) = \max_i q_i.$$

Thus $m(q)$ is the best Q-value at the current state (it should not be confused with the power mean $m_p(q)$). Because

$$m(q) = \mu_w(q) + \max_i z_i(q) - b,$$

the difference between a hard max backup and the power-mirror backup is

$$m(q) - C_{p,w,b}(q) = M(q) - m_p(q), \qquad M(q) = \max_i z_i(q). \tag{56}$$

where $M(q)$ is the largest shifted Q-value. This gap is small when π_{ref} already assigns substantial probability to high-value actions: a power mean can then approach the maximum without moving far from the reference distribution.

Definition 1 (Near-optimal reference mass)**.** For $\varepsilon \geq 0$, define

$$G_\varepsilon(q) = \{i : q_i \geq m(q) - \varepsilon\}, \qquad W_\varepsilon(q) = \sum_{i \in G_\varepsilon(q)} w_i. \tag{57}$$

$G_\varepsilon(q)$ is the set of actions whose Q-values are within ε of the best action, and $W_\varepsilon(q)$ is the total probability that π_{ref} assigns to this set. In offline RL with $\pi_{\mathrm{ref}} = \pi_\beta$, $W_\varepsilon(q)$ directly measures how often the behavior policy selects nearly greedy actions at this state.

Proposition 3 (Near-optimal coverage bound)**.** *If* $\mathrm{span}(q) \leq D \leq b$*, then for every* $\varepsilon \geq 0$*,*

$$0 \leq m(q) - C_{p,w,b}(q) \leq \min \begin{Bmatrix} D,\ \varepsilon W_\varepsilon(q) + D\big(1 - W_\varepsilon(q)\big), \\ M(q)\big(1 - W_\varepsilon(q)^{1/p}\big) + \varepsilon W_\varepsilon(q)^{1/p} \end{Bmatrix}. \tag{58}$$

The three upper bounds describe increasingly informative cases. The first uses only the Q-value range D. The second uses the average behavior mass on near-optimal actions. The third also uses the power geometry and is never worse—and is generally strictly better—than the coarser expression $\varepsilon + (b + D)(1 - W_\varepsilon^{1/p})$ obtained from the Q-value range alone.

Corollary 3 (Joint utility and ratio tuning)**.** *Fix* $0 < W_\varepsilon(q) < 1$*, choose a desired greedy-backup error* τ *satisfying* $\varepsilon < \tau < M(q)$*, assume* $D < b$*, and choose an upper action-probability-ratio budget* $\rho_+ > 1$*. The power-specific term in* (58) *is at most* τ *whenever*

$$p \geq p_{\mathrm{util}} := \frac{\log W_\varepsilon(q)}{\log\big((M(q) - \tau)/(M(q) - \varepsilon)\big)}. \tag{59}$$

If the likelihood-ratio budget $\pi_p/w \leq \rho_+$ *is also required, Theorem 7 gives*

$$p \leq p_{\mathrm{ratio}} := 1 + \frac{\log \rho_+}{\log(1 + D/b)}. \tag{60}$$

The first inequality asks for a sufficiently large p *to approach the greedy backup, whereas the second asks for a sufficiently small* p *to stay close to the reference policy. The two displayed sufficient inequalities can be met simultaneously exactly when* $\max\{1, p_{\mathrm{util}}\} \leq p_{\mathrm{ratio}}$*. Failure of this diagnostic does not imply incompatibility of the actual utility and ratio requirements for a particular* (q, w)*. The true quantities* $M(q)$ *and* $W_\varepsilon(q)$ *are generally unknown in offline data, so this is a population-level feasibility diagnostic rather than a finite-sample rule for selecting* p*.*

What the bound proves. Set $w = \pi_{\mathrm{ref}} = \pi_\beta$. If the behavior policy assigns large probability to actions whose true Q-values are within ε of the best value, finite-p power mirror can stay inside its behavior-ratio budget while remaining close to the hard greedy backup. This is the utility benefit of action coverage.

What the bound does not prove. The result does not say that power mirror discovers an optimal action that is absent from the dataset. If $W_\varepsilon(q)$ is tiny, the coverage-dependent terms become weak and the theorem falls back to the coarse Q-value-range bound. No offline method can safely infer the value of an entirely unsupported action without additional assumptions. The proposition makes this action-coverage requirement visible rather than hiding it behind a full-support assumption.

Combining (52) and (58) gives the core safety-utility tradeoff:

$$\boxed{\begin{aligned} \text{greedy approximation gap} &\le M(q)\big(1 - W_\varepsilon^{1/p}\big) + \varepsilon W_\varepsilon^{1/p}, \\ \frac{\pi_p(a|s)}{\pi_\beta(a|s)} &\in \left[(1 - D/b)^{p-1}, (1 + D/b)^{p-1}\right], \\ \text{critic-error deviation multiplier} &\le H_{p,D/b}. \end{aligned}} \tag{61}$$

For $0 \le \varepsilon \le D$, increasing p reduces the displayed power-specific greedy gap, but it also widens the behavior-ratio interval and increases the later Q-function-error multiplier $H_{p,D/b}$. In plain language, a larger power can exploit good covered actions more strongly but also trusts the learned Q-function more aggressively. This is the central safety–utility tradeoff.

E Behavior-anchored population control

The previous sections described one policy-improvement update at one state. We now apply the same update at every state and at every Bellman iteration. The reference is fixed permanently to the behavior policy π_β that generated the offline dataset. This produces one stationary Bellman operator and therefore a complete value-iteration algorithm.

We call the result a *population* algorithm because this section assumes that the reward and transition model are known exactly. It defines the target that a fitted offline algorithm should estimate; it is not yet a finite-sample guarantee. Section F later studies the different choice of re-anchoring at the current policy after every exact policy-evaluation step.

E.1 Population Power-Mirror Value Iteration

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ be a finite discounted MDP with discount factor $\gamma \in [0, 1)$. From this point onward, set

$$\pi_{\mathrm{ref}}(\cdot|s) = \pi_\beta(\cdot|s).$$

To keep the equations compact and consistent with the preceding generic derivation, we write $w := \pi_\beta$. Thus every occurrence of w in this section means the behavior policy, not a second policy. Its supported action set is

$$\mathcal{A}_w(s) = \{a \in \mathcal{A} : w(a|s) > 0\}.$$

Equivalently, $\mathcal{A}_w(s)$ is the set of actions observed with positive probability under π_β at state s. Write $w_s = \pi_\beta(\cdot|s)$. The behavior distribution w_s, power p_s, and shift b_s are fixed before value iteration begins. Every sum is over $\mathcal{A}_w(s)$, where $\pi_\beta(a|s) > 0$; the algorithm can therefore reweight covered actions but never introduces an unsupported action.

Assumption 1 (Population control domain)**.** Rewards satisfy $r(s, a) \in [r_{\min}, r_{\max}]$. Define

$$V_{\min} = \frac{r_{\min}}{1 - \gamma}, \qquad V_{\max} = \frac{r_{\max}}{1 - \gamma}, \qquad \overline{D} = V_{\max} - V_{\min}. \tag{62}$$

For every state, fix $1 < p_s < \infty$ and $b_s > \overline{D}$ independently of the value iterate. Write $\overline{\eta}_s = \overline{D}/b_s \in [0, 1)$.

The interval $[V_{\min}, V_{\max}]$ contains the value of every policy, and $\overline{D}$ is its width. It is a conservative uniform bound on the range of every statewise Q-vector encountered below. The strict condition $b_s > \overline{D}$ therefore keeps every shifted Q-value positive throughout the entire recursion, rather than only at a single iteration.

Let

$$\mathcal{V} = [V_{\min}, V_{\max}]^{|\mathcal{S}|}$$

and, for a candidate value function $V \in \mathcal{V}$, define the usual one-step Bellman Q-value

$$Q_V(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) V(s'). \tag{63}$$

Thus $Q_V(s,a)$ is the return from taking action a once and then using V for the next state. We compare three Bellman operators: ordinary evaluation of π_β, the proposed power-mirror backup, and a hard maximum restricted to behavior-supported actions:

$$(\mathcal{T}^w V)(s) = \sum_{a \in \mathcal{A}_w(s)} w(a|s) Q_V(s,a), \tag{64}$$

$$(\mathcal{T}_C V)(s) = C_{p_s,w_s,b_s}\big(Q_V(s,\cdot)\big), \tag{65}$$

$$(\mathcal{T}_w^\star V)(s) = \max_{a \in \mathcal{A}_w(s)} Q_V(s,a). \tag{66}$$

Population Power-Mirror Value Iteration (PMVI). The exact algorithm has three steps:

1. Compute the behavior value V^{π_β} and initialize $V_0 = V^{\pi_\beta}$ when every finite output must be safe. In the compact notation above, $V^{\pi_\beta} = V^w$. An arbitrary $V_0 \in \mathcal{V}$ is enough if only convergence to the fixed point is required.
2. For $k = 0, 1, \ldots$, compute the Bellman Q-function Q_{V_k} from (63) and update
$$V_{k+1} = \mathcal{T}_C V_k, \tag{67}$$
using the same behavior reference $w_s = \pi_\beta(\cdot|s)$ and the same p_s, b_s at every iteration.
3. At any iteration, extract a deployable policy by differentiating the same backup with respect to its Q-values:
$$\pi_k(\cdot|s) = \nabla C_{p_s,w_s,b_s}\big(Q_{V_k}(s,\cdot)\big). \tag{68}$$
Stop when the Bellman residual or $\|V_{k+1} - V_k\|_\infty$ is below a chosen tolerance. The limiting policy is obtained by applying this same gradient formula to the fixed-point Q-function.

Theorem 8 (Bellman sandwich and population convergence)**.** *Under Assumption 1, $\mathcal{T}_C$ maps $\mathcal{V}$ into itself, is monotone, and is a γ-contraction in ℓ_∞. Moreover, for every $V \in \mathcal{V}$,*

$$\mathcal{T}^w V \le \mathcal{T}_C V \le \mathcal{T}_w^\star V \qquad \textit{componentwise}. \tag{69}$$

Consequently, $\mathcal{T}_C$ has a unique fixed point $V_C \in \mathcal{V}$, and the iterates in (67) *satisfy*

$$\|V_k - V_C\|_\infty \le \gamma^k \|V_0 - V_C\|_\infty. \tag{70}$$

If $V_0 = V^w$, then V_k is componentwise nondecreasing and

$$V^w \le V_k \le V_C \qquad \textit{for every } k \ge 0. \tag{71}$$

The operator sandwich in (69) is the key fact. The lower bound says that each power-mirror backup is at least as large as ordinary evaluation of π_β. The upper bound says that it never exceeds a hard maximization over actions covered by π_β. Because $\mathcal{T}_C$ is a γ-contraction, these local inequalities define a unique global value target rather than only a one-step heuristic.

E.2 End-to-end population improvement

Let V_C be the unique fixed point of $\mathcal{T}_C$. Its Bellman Q-function and extracted policy are

$$Q_C(s,a) = Q_{V_C}(s,a), \qquad \pi_C(\cdot|s) = \nabla C_{p_s,w_s,b_s}\big(Q_C(s,\cdot)\big). \tag{72}$$

Here Q_C is a genuine Q-function: it uses V_C after the first action. Let $V^w = V^{\pi_\beta}$ be the ordinary value of the behavior policy, let V^{π_C} be the ordinary unregularized value of the extracted policy, and let $V_w^\star$ be the best value achievable without ever selecting an action outside behavior support. The subscript w in $V_w^\star$ means "restricted to the support of π_β." For a policy π, write

$$P^\pi(s,s') = \sum_a \pi(a|s) P(s'|s,a)$$

for its state-transition matrix.

Theorem 9 (End-to-end behavior-anchored population control)**.** *Under Assumption 1, π_C is well defined, has the same action support as w, and satisfies the sharp statewise bounds*

$$L_{p_s,\overline{\eta}_s} \le \frac{\pi_C(a|s)}{w(a|s)} \le U_{p_s,\overline{\eta}_s}, \qquad a \in \mathcal{A}_w(s), \tag{73}$$

together with Corollary 1. The tighter realized span of $Q_C(s,\cdot)$ may replace $\overline{D}$ a posteriori. Its value obeys

$$\boxed{V^w \le V_C \le V^{\pi_C} \le V_w^\star} \qquad \textit{componentwise}. \tag{74}$$

More precisely, define two nonnegative gains at each state. The backup gain

$$g_C(s) = C_{p_s,w_s,b_s}\big(Q_C(s,\cdot)\big) - \sum_a w(a|s) Q_C(s,a), \tag{75}$$

compares the power backup with behavior evaluation, while the policy-extraction gain

$$\zeta_C(s) = \sum_a \pi_C(a|s) Q_C(s,a) - C_{p_s,w_s,b_s}\big(Q_C(s,\cdot)\big). \tag{76}$$

compares the expected Q-value of the gradient policy with the scalar power backup. Then

$$V_C - V^w = (I - \gamma P^w)^{-1} g_C, \tag{77}$$

$$V^{\pi_C} - V_C = (I - \gamma P^{\pi_C})^{-1} \zeta_C. \tag{78}$$

Hence the total improvement over π_β has two exact components: the backup gain is propagated through future transitions under the behavior policy, and the policy-extraction gain is propagated through transitions under π_C. The matrices $(I - \gamma P^\pi)^{-1}$ are the usual discounted transition resolvents. In addition, if (67) *is initialized with $V_0 = V^w$, every finite-iterate policy in* (68) *is already behavior-safe:*

$$V^{\pi_k} \ge V_{k+1} \ge V_k \ge V^w \qquad \textit{for every } k \ge 0. \tag{79}$$

Every π_k has the same support as w and satisfies the bounds in (73)*, with π_C replaced by π_k; Corollary 1 also applies. For each π_k, the tighter realized span is available a posteriori.*

This theorem goes beyond applying a one-step improvement rule once to the behavior Q-function Q^{π_β}. It proves safety after the complete Bellman recursion and even for every intermediate extracted policy when the algorithm starts at V^{π_β}. If π_β has positive probability on every action, $V_w^\star$ equals the usual optimal value $V^\star$. Otherwise, $V_w^\star$ is necessarily weaker: no offline guarantee can remove the missing-action support defect without further assumptions.

E.3 Support-relative coverage-dependent optimality

The value sandwich guarantees improvement but does not yet say how close π_C is to the best behavior-supported policy. The statewise coverage bound answers this question. For each state, define the Q-value range $D_C(s)$, the behavior probability $W_\varepsilon^C(s)$ of actions within ε of the best supported action, the largest shifted Q-value $M_C(s)$, and the gap $\delta_C(s)$ between a hard supported maximum and the power backup:

$$D_C(s) = \max_{a \in \mathcal{A}_w(s)} Q_C(s,a) - \min_{a \in \mathcal{A}_w(s)} Q_C(s,a), \tag{80}$$

$$W_\varepsilon^C(s) = \sum_{a:\, Q_C(s,a) \ge \max_{a' \in \mathcal{A}_w(s)} Q_C(s,a') - \varepsilon} w(a|s), \tag{81}$$

$$M_C(s) = b_s + \max_{a \in \mathcal{A}_w(s)} Q_C(s,a) - \mathbb{E}_{a \sim w_s} Q_C(s,a), \tag{82}$$

$$\delta_C(s) = \max_{a \in \mathcal{A}_w(s)} Q_C(s,a) - C_{p_s,w_s,b_s}\big(Q_C(s,\cdot)\big). \tag{83}$$

Assumption 1 implies $0 \le D_C(s) \le \overline{D} < b_s$. Let $\pi_w^\star$ be an optimal policy among policies supported by π_β. For an initial-state distribution ν, define the usual expected discounted return $J_\nu(\pi)$ and normalized discounted state-visitation distribution d_ν^π by

$$J_\nu(\pi) = \sum_s \nu(s) V^\pi(s), \qquad d_\nu^\pi(s) = (1-\gamma) \sum_{t \ge 0} \gamma^t \Pr_{\nu,\pi}(s_t = s).$$

Corollary 4 (Coverage-dependent support optimality). *For arbitrary choices $\varepsilon_s \geq 0$,*

$$0 \leq V_w^\star - V^{\pi_C} \leq V_w^\star - V_C \leq (I - \gamma P^{\pi_w^\star})^{-1}\delta_C \qquad \textit{componentwise}, \tag{84}$$

and

$$0 \leq J_\nu(\pi_w^\star) - J_\nu(\pi_C) \leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\nu^{\pi_w^\star}}\left[\delta_C(s)\right], \tag{85}$$

$$\delta_C(s) \leq \min\Big\{D_C(s),\ \varepsilon_s W_{\varepsilon_s}^C(s) + D_C(s)\left(1 - W_{\varepsilon_s}^C(s)\right), \\ M_C(s)\left(1 - W_{\varepsilon_s}^C(s)^{1/p_s}\right) + \varepsilon_s W_{\varepsilon_s}^C(s)^{1/p_s}\Big\}. \tag{86}$$

Thus π_C is close to the best behavior-supported policy when that comparator visits states where π_β assigns substantial probability to actions that are nearly greedy under Q_C. The expectation under $d_\nu^{\pi_w^\star}$ is important: poor action coverage matters most at states that the comparator visits often. Moreover, when PMVI is initialized with $V_0 = V^w$, its finite-iterate output has the explicit stopping bound

$$0 \leq J_\nu(\pi_w^\star) - J_\nu(\pi_k) \leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\nu^{\pi_w^\star}}\left[\delta_C(s)\right] + \min\left\{\gamma^{k+1}\overline{D},\ \frac{\gamma}{1-\gamma}\|V_{k+1} - V_k\|_\infty\right\}. \tag{87}$$

The quantity $W_\varepsilon^C(s)$ is a population diagnostic because it depends on the exact fixed-point Q-function Q_C. The theorem uses it to characterize when the method is near-optimal on behavior support; it does not assume that an offline learner knows it exactly or can use it directly to tune hyperparameters.

E.4 Unrestricted near-optimality: action and state coverage

Support-relative optimality does not quantify the additional loss caused by optimal actions that are missing or rare in the offline data. We therefore compare directly with the unrestricted optimum. Let $V^\star$ and $Q^\star = Q_{V^\star}$ be the usual optimal value and optimal Q-function, and let $\pi^\star$ be any optimal policy. At each state, define:

- $D_\star^w(s)$: the gap from the optimal value to the worst behavior-supported action;
- $W_{\star,\varepsilon}^w(s)$: the probability that π_β assigns to globally ε-optimal actions;
- $X_\star^w(s)$: the largest shifted value used in the power bound; and
- $\Delta_\star^C(s)$: the one-step loss of the power backup relative to the unrestricted optimal Bellman value.

Formally,

$$D_\star^w(s) = V^\star(s) - \min_{a\in\mathcal{A}_w(s)} Q^\star(s,a), \tag{88}$$

$$W_{\star,\varepsilon}^w(s) = \sum_{a:\, Q^\star(s,a)\geq V^\star(s)-\varepsilon} w(a|s), \tag{89}$$

$$X_\star^w(s) = b_s + V^\star(s) - \mathbb{E}_{a\sim w_s} Q^\star(s,a), \tag{90}$$

$$\Delta_\star^C(s) = V^\star(s) - (\mathcal{T}_C V^\star)(s). \tag{91}$$

For a statewise tolerance $\varepsilon_s \geq 0$, set

$$\overline{\Delta}_{\star,\varepsilon}^C(s) = \min\Big\{D_\star^w(s),\ \varepsilon_s W_{\star,\varepsilon_s}^w(s) + D_\star^w(s)\big(1 - W_{\star,\varepsilon_s}^w(s)\big), \\ X_\star^w(s)\big(1 - W_{\star,\varepsilon_s}^w(s)^{1/p_s}\big) + \varepsilon_s W_{\star,\varepsilon_s}^w(s)^{1/p_s}\Big\}. \tag{92}$$

Theorem 10 (Coverage-dependent unrestricted near-optimality). *Under Assumption 1,*

$$0 \leq \Delta_\star^C \leq \overline{\Delta}_{\star,\varepsilon}^C \qquad \textit{componentwise}. \tag{93}$$

Moreover, there is a mean-value policy π^{mv} with the same support as w and

$$L_{p_s,\overline{\eta}_s} \leq \frac{\pi^{\mathrm{mv}}(a|s)}{w(a|s)} \leq U_{p_s,\overline{\eta}_s} \tag{94}$$

for every state s and $a \in \mathcal{A}_w(s)$. This policy satisfies

$$0 \le V^\star - V^{\pi_C} \le V^\star - V_C = (I - \gamma P^{\pi^{\mathrm{mv}}})^{-1}\Delta_\star^C \le (I - \gamma P^{\pi^{\mathrm{mv}}})^{-1}\overline{\Delta}_{\star,\varepsilon}^C. \tag{95}$$

Consequently,

$$J_\nu(\pi^\star) - J_\nu(\pi_C) \le \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\nu^{\pi^{\mathrm{mv}}}}\left[\overline{\Delta}_{\star,\varepsilon}^C(s)\right], \tag{96}$$

$$\|V^\star - V^{\pi_C}\|_\infty \le \frac{\|\overline{\Delta}_{\star,\varepsilon}^C\|_\infty}{1-\gamma}. \tag{97}$$

If PMVI starts from $V_0 = V^w$, every finite output also obeys

$$J_\nu(\pi^\star) - J_\nu(\pi_k) \le \frac{1}{1-\gamma}\mathbb{E}_{d_\nu^{\pi^{\mathrm{mv}}}}[\overline{\Delta}_{\star,\varepsilon}^C] + \min\left\{\gamma^{k+1}\overline{D},\ \frac{\gamma}{1-\gamma}\|V_{k+1} - V_k\|_\infty\right\}. \tag{98}$$

The interpolation policy π^{mv} is a proof device, not an additional policy that the algorithm must deploy.

The theorem exposes two different coverage requirements. First, $W_{\star,\varepsilon}^w(s)$ measures *action coverage*: whether the behavior policy π_β chooses globally near-optimal actions at state s. Second, the expectation in (96) measures *state coverage*: how often a ratio-admissible interpolation policy visits states with a large local defect. To compare this occupancy with a data-state distribution μ, assume $d_\nu^{\pi^{\mathrm{mv}}} \ll \mu$ and let $r, r' \in [1,\infty]$ be conjugate exponents. Hölder's inequality gives

$$J_\nu(\pi^\star) - J_\nu(\pi_C) \le \frac{1}{1-\gamma}\left\|\frac{d_\nu^{\pi^{\mathrm{mv}}}}{\mu}\right\|_{L^r(\mu)}\left\|\overline{\Delta}_{\star,\varepsilon}^C\right\|_{L^{r'}(\mu)}. \tag{99}$$

For $\mu = d_\nu^w = d_\nu^{\pi_\beta}$, choosing $r = \infty$ gives the usual worst-case concentrability coefficient, whereas $r = 2$ gives the factor $\sqrt{1 + \chi^2(d_\nu^{\pi^{\mathrm{mv}}}\|d_\nu^w)}$. The pointwise action-ratio theorem guarantees overlap between action distributions at the same state. It does not by itself guarantee that the behavior policy visits all states needed by an improved policy.

Corollary 4 and Theorem 10 answer different questions. The corollary measures only the smoothing loss inside behavior support using Q_C. The theorem also charges the missing-support loss by using globally optimal $Q^\star$. If π_β gives zero probability to every globally near-optimal action at an important state, the bound correctly falls back to its coarse Q-range term. The exact $Q^\star$, near-optimal behavior masses, and occupancy ratios are theoretical diagnostics, not statistics assumed known to a fitted offline algorithm.

F Moving-anchor Power-Mirror Policy Iteration

F.1 Population updates, monotone improvement, and ratio control

Fixed-anchor PMVI always compares actions with the original behavior policy w. This gives a single contracting Bellman operator and keeps every extracted policy inside a fixed likelihood-ratio range relative to the data-collection policy. Moving-anchor Power-Mirror Policy Iteration (PMPI) makes a different choice: after evaluating the current policy, it uses that current policy as the reference for the next improvement step. In standard RL language, PMPI alternates exact policy evaluation and conservative policy improvement.

This change creates a useful but important tradeoff. PMPI improves the ordinary, unregularized value at every population step and, if the updates remain sufficiently strong, converges to the best policy whose actions are supported by w. However, its likelihood-ratio guarantee is local: it controls the new policy relative to the previous policy. Ratios relative to the original behavior policy therefore multiply over iterations. The table summarizes the distinction before the formal result.

Property	Fixed-anchor PMVI	Moving-anchor PMPI
Policy used as reference	Original behavior w at every backup	Current policy σ_k after exact evaluation
Update structure	One stationary γ-contracting value iteration	Standard evaluation–improvement loop
Population guarantee	Every extracted policy improves on w	Every new policy improves on the previous one
Likelihood-ratio control	Directly relative to w	One step at a time; products relative to w
Convergence target	A behavior-smoothed fixed point	$V_w^\star$ under persistent updates

Initialize the policy with the behavior policy, $\sigma_0 = w$, and never assign probability to actions outside $\mathcal{A}_w(s)$. Fix $1 < p_s < \infty$ and $b_s > \overline{D}$ as in Assumption 1. One population PMPI iteration consists of the following familiar RL operations:

1. **Policy evaluation.** Evaluate σ_k exactly to obtain its state-value function V^{σ_k} and action-value function. We write that action-value function locally as $Q_k(s,a)$:
$$Q_k(s,a) = Q^{\sigma_k}(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^{\sigma_k}(s'). \tag{100}$$
Thus $Q_k(s,a)$ is simply $Q^{\sigma_k}(s,a)$, not a new value concept.
2. **Power-Mirror improvement.** Reweight the current policy according to these action values to form the full-step candidate
$$\widetilde{\sigma}_{k+1}(\cdot|s) = \nabla C_{p_s,\sigma_k(\cdot|s),b_s}\big(Q_k(s,\cdot)\big). \tag{101}$$
3. **Optional damping.** Choose a statewise step size $\lambda_{k,s} \in [0,1]$ and mix the current policy with the full-step candidate:
$$\sigma_{k+1}(\cdot|s) = (1-\lambda_{k,s})\sigma_k(\cdot|s) + \lambda_{k,s}\widetilde{\sigma}_{k+1}(\cdot|s). \tag{102}$$
The undamped update uses $\lambda_{k,s} = 1$. Smaller values make each improvement more conservative and reduce how quickly the policy moves away from the data distribution. The behavior-box line search below chooses $\lambda_{k,s}$ to enforce a permanent ratio budget relative to w. To retain convergence, the sequence of step sizes cannot shrink so quickly that the total amount of policy improvement is finite.

For each state and iteration define
$$D_{k,s} = \operatorname{span}_{a\in\mathcal{A}_w(s)} Q_k(s,a), \qquad \eta_{k,s} = D_{k,s}/b_s, \tag{103}$$
$$L_{k,s} = L_{p_s,\eta_{k,s}}, \qquad U_{k,s} = U_{p_s,\eta_{k,s}}. \tag{104}$$
Here $D_{k,s}$ is the range of the current action values at state s, and $\eta_{k,s}$ is that range divided by the shift b_s. Assumption 1 gives $0 \le D_{k,s} \le \overline{D} < b_s$, so $\eta_{k,s} < 1$ and the lower and upper ratio bounds are well defined.

Theorem 11 (Monotone moving-anchor power-mirror policy iteration)**.** *Under Assumption 1, every finite PMPI iterate has exactly the same support as w. The raw and damped updates satisfy*
$$L_{k,s} \le \frac{\widetilde{\sigma}_{k+1}(a|s)}{\sigma_k(a|s)} \le U_{k,s}, \tag{105}$$
$$1 + \lambda_{k,s}(L_{k,s}-1) \le \frac{\sigma_{k+1}(a|s)}{\sigma_k(a|s)} \le 1 + \lambda_{k,s}(U_{k,s}-1) \tag{106}$$
for every $a \in \mathcal{A}_w(s)$.

Let
$$z_{k,s}(a) = b_s + Q_k(s,a) - V^{\sigma_k}(s), \qquad m_{k,s} = \left(\sum_a \sigma_k(a|s) z_{k,s}(a)^{p_s}\right)^{1/p_s}, \tag{107}$$
$$R_{k,s} = \sum_a \sigma_k(a|s)\left(\frac{z_{k,s}(a)}{m_{k,s}}\right)^{p_s-1}. \tag{108}$$

The raw and damped one-step gains are exactly

$$\widetilde{G}_k(s) = \langle \widetilde{\sigma}_{k+1}(\cdot|s) - \sigma_k(\cdot|s), Q_k(s,\cdot)\rangle = m_{k,s} - b_s R_{k,s}$$
$$= (m_{k,s} - b_s) + b_s(1 - R_{k,s}) \geq 0, \tag{109}$$

$$G_k(s) = \langle \sigma_{k+1}(\cdot|s) - \sigma_k(\cdot|s), Q_k(s,\cdot)\rangle = \lambda_{k,s}\widetilde{G}_k(s). \tag{110}$$

Consequently,

$$V^{\sigma_{k+1}} - V^{\sigma_k} = (I - \gamma P^{\sigma_{k+1}})^{-1} G_k \geq 0 \qquad \textit{componentwise}, \tag{111}$$

and hence

$$V^w = V^{\sigma_0} \leq V^{\sigma_1} \leq \cdots \leq V_w^\star. \tag{112}$$

If $\lambda_{k,s} > 0$, the local gain is strict exactly when $Q_k(s,\cdot)$ is nonconstant on $\mathcal{A}_w(s)$.

If the statewise parameters p_s, b_s are fixed and $\inf_{k,s} \lambda_{k,s} > 0$, then

$$V^{\sigma_k} \longrightarrow V_w^\star \qquad \textit{componentwise}. \tag{113}$$

Every accumulation point of σ_k is optimal among policies supported on $\mathcal{A}_w$. If that support-constrained optimal policy is unique, then σ_k itself converges to it.

The theorem makes three points. First, the candidate changes each action probability by a controlled multiplicative factor, given in (105)–(106). Second, the expected advantage $G_k(s)$ is nonnegative at every state. The standard policy-improvement identity then propagates these local advantages through future state visits, yielding (111). Third, repeated nonvanishing steps converge to the best policy available on behavior support.

The two terms in (109) explain where the local improvement comes from. The term $m_{k,s} - b_s$ is the improvement produced by replacing the ordinary expectation under σ_k with the power backup. The term $b_s(1 - R_{k,s})$ is the additional improvement obtained when the gradient of that backup is converted into a policy. Both are nonnegative. Unlike greedy policy iteration, a finite-p update never collapses immediately onto one action: every finite iterate keeps exactly the original action support and normally reaches a deterministic boundary optimum only in the limit.

The parameters may also vary across iterations. The following condition formalizes the intuitive requirement that the algorithm must keep making enough progress. Suppose $1 < p_{k,s} < \infty$, $b_{k,s} > D_{k,s}$, and $\lambda_{k,s} \in [0,1]$ are chosen at iteration k and then held fixed while differentiating the backup. Redefine $\eta_{k,s} = D_{k,s}/b_{k,s}$, put $\alpha_{k,s} = p_{k,s} - 1$, and let

$$\vartheta_{k,s} = \min\left\{(1-\eta_{k,s})^{\alpha_{k,s}-1}, (1+\eta_{k,s})^{\alpha_{k,s}-1}\right\}, \tag{114}$$

$$\xi_{k,s} = \frac{\lambda_{k,s}\alpha_{k,s}\vartheta_{k,s}}{b_{k,s}(1+\eta_{k,s})^{2\alpha_{k,s}}}. \tag{115}$$

The conclusion (113) still holds whenever $\sum_k \xi_{k,s} = \infty$ for every state with more than one supported action. Under a uniform span margin $\eta_{k,s} \leq \overline{\eta} < 1$ and powers bounded above, the simpler condition $\sum_k \lambda_{k,s}(p_{k,s} - 1)/b_{k,s} = \infty$ is sufficient. In practical terms, convergence can fail if the step size $\lambda_{k,s}$ or the power strength $p_{k,s} - 1$ decays too quickly, or if the shift $b_{k,s}$ grows too quickly.

Corollary 5 (Cumulative behavior drift)**.** *Define*

$$\ell_{k,s} = 1 + \lambda_{k,s}(L_{k,s} - 1), \qquad u_{k,s} = 1 + \lambda_{k,s}(U_{k,s} - 1). \tag{116}$$

Then, for every $K \geq 1$ and $a \in \mathcal{A}_w(s)$,

$$\prod_{k=0}^{K-1} \ell_{k,s} \leq \frac{\sigma_K(a|s)}{w(a|s)} \leq \prod_{k=0}^{K-1} u_{k,s}. \tag{117}$$

Consequently,

$$D_\infty(\sigma_K(\cdot|s)\|w_s) \leq \sum_{k<K} \log u_{k,s}, \tag{118}$$

$$D_\infty(w_s\|\sigma_K(\cdot|s)) \leq \sum_{k<K} -\log \ell_{k,s}. \tag{119}$$

The product in (117) *is essential in offline RL. A small change relative to the previous policy is not necessarily a small change relative to the policy that generated the dataset. For fixed p_s, b_s, define the conservative iteration-uniform constants*

$$L_s^{\mathrm{unif}} = L_{p_s,\overline{\eta}_s}, \qquad U_s^{\mathrm{unif}} = U_{p_s,\overline{\eta}_s}, \qquad U_{\max}^{\mathrm{unif}} = \max_s U_s^{\mathrm{unif}}.$$

The logger-relative action ratio is bounded by $[(L_s^{\mathrm{unif}})^K, (U_s^{\mathrm{unif}})^K]$. Along a length-$H$ state sequence, its trajectory upper product satisfies

$$\prod_{t=0}^{H-1} (U_{s_t}^{\mathrm{unif}})^K \le (U_{\max}^{\mathrm{unif}})^{KH}.$$

This compounding is the main offline caveat of moving-anchor updates. The final held-out test must evaluate the likelihood ratio σ_K/w of the policy that will actually be deployed, or use the full product in (117). The ratio bound from only the last PMPI step is not a valid certificate relative to the logged data.

Keeping every iterate inside a fixed behavior ratio. Fix $0 < \underline{\rho}_s \le 1 \le \overline{\rho}_s$ and define

$$\mathfrak{B}_s(w) = \left\{ \pi \in \Delta(\mathcal{A}_w(s)) : \underline{\rho}_s w(a|s) \le \pi(a|s) \le \overline{\rho}_s w(a|s) \text{ for every } a \in \mathcal{A}_w(s) \right\}. \tag{120}$$

At iteration k, choose any $\lambda_{k,s}$ no larger than

$$\lambda_{k,s}^{\max} = \sup\left\{ \lambda \in [0,1] : (1-\lambda)\sigma_k(\cdot|s) + \lambda \widetilde{\sigma}_{k+1}(\cdot|s) \in \mathfrak{B}_s(w) \right\}. \tag{121}$$

At each state, (121) takes the largest permitted fraction of the Power-Mirror update that remains inside the fixed behavior box. Starting from w, it preserves the direct logger-relative interval

$$\underline{\rho}_s \le \frac{\sigma_k(a|s)}{w(a|s)} \le \overline{\rho}_s \qquad \text{for every } k, \tag{122}$$

while (111) still proves componentwise value improvement. The tradeoff is unavoidable: when the policy reaches the boundary of this box, the line search may return $\lambda_{k,s}^{\max} = 0$ and stop improving. A positive lower ratio also rules out deterministic policies. Thus this variant offers a permanent, auditable behavior-relative guarantee but may not reach the support-constrained optimum.

Corollary 6 (Supported Bellman stopping certificate)**.** *Let*

$$\mathrm{Res}_k(s) = \max_{a \in \mathcal{A}_w(s)} Q_k(s,a) - V^{\sigma_k}(s). \tag{123}$$

The quantity $\mathrm{Res}_k(s)$ is the standard optimal Bellman residual restricted to behavior-supported actions: it is the gap between the best supported one-step action value and the value of the current policy. It is not a critic-estimation error. A zero residual at every state means that σ_k is already optimal on behavior support. Then

$$0 \le V_w^\star - V^{\sigma_k} \le (I - \gamma P^{\pi_w^\star})^{-1} \mathrm{Res}_k \le \frac{\| \mathrm{Res}_k \|_\infty}{1-\gamma} \mathbf{1}. \tag{124}$$

Hence $\| \mathrm{Res}_k \|_\infty \le (1-\gamma)\varepsilon$ certifies ε-optimality on behavior support. At a finite interior iterate, $\widetilde{G}_k \equiv 0$ also certifies exact support optimality. A merely small gain or a small change in policy parameters is not enough: an improving action with very little current probability can produce a tiny update even when the Bellman residual remains large.

F.2 Finite-iterate coverage-dependent near-optimality

The preceding residual can be read directly in terms of action coverage. At state s, define the current action-value gap $g_{k,s}(a)$ as the loss from choosing action a instead of the best behavior-supported action. The set $\mathcal{G}_{k,t}(s)$ contains actions whose current Q^{σ_k} value is within t of that best value. Its

probability under the current policy and under the original behavior policy is denoted by $W^{\sigma}_{k,t}(s)$ and $W^{w}_{k,t}(s)$, respectively:

$$g_{k,s}(a) = \max_{c \in \mathcal{A}_w(s)} Q_k(s,c) - Q_k(s,a), \tag{125}$$

$$\mathcal{G}_{k,t}(s) = \{a \in \mathcal{A}_w(s) : g_{k,s}(a) \le t\}, \tag{126}$$

$$W^{\sigma}_{k,t}(s) = \sum_{a \in \mathcal{G}_{k,t}(s)} \sigma_k(a|s), \qquad W^{w}_{k,t}(s) = \sum_{a \in \mathcal{G}_{k,t}(s)} w(a|s). \tag{127}$$

For a tolerance $0 \le \varepsilon \le D_{k,s}$, the following simple upper bound becomes small when σ_k assigns large probability to ε-near-greedy actions:

$$B^{\sigma}_{k,\varepsilon}(s) = D_{k,s} - (D_{k,s} - \varepsilon) W^{\sigma}_{k,\varepsilon}(s). \tag{128}$$

We also separate the gap caused by the smoothed Power-Mirror backup itself:

$$\delta^{\mathrm{PM}}_k(s) = \max_{a \in \mathcal{A}_w(s)} Q_k(s,a) - C_{p_s,\sigma_k(\cdot|s),b_s}(Q_k(s,\cdot)), \tag{129}$$

$$\overline{\delta}^{\mathrm{PM}}_{k,\varepsilon}(s) = \min\Big\{ D_{k,s},\ B^{\sigma}_{k,\varepsilon}(s), \\ (b_s + \mathrm{Res}_k(s))\big(1 - W^{\sigma}_{k,\varepsilon}(s)^{1/p_s}\big) + \varepsilon W^{\sigma}_{k,\varepsilon}(s)^{1/p_s} \Big\}. \tag{130}$$

Theorem 12 (Finite-iterate PMPI coverage recursion)**.** *Under Assumption 1, consider any raw, damped, or behavior-box PMPI step in* (102) *with the fixed p_s, b_s construction above. The supported residual has the exact layer-cake representation*

$$\mathrm{Res}_k(s) = \int_0^{D_{k,s}} \big(1 - W^{\sigma}_{k,t}(s)\big)\, dt \le B^{\sigma}_{k,\varepsilon}(s). \tag{131}$$

Define the next policy's approximate-greedy error at the current value by

$$\mathrm{GreedyErr}_k(s) = (\mathcal{T}^{\star}_w V^{\sigma_k})(s) - (\mathcal{T}^{\sigma_{k+1}} V^{\sigma_k})(s). \tag{132}$$

Here $\mathrm{GreedyErr}_k(s)$ *denotes the exact one-step greedy gap of the new policy at the current value function; it is not a Q-function estimation error. Then* $\mathrm{GreedyErr}_k \ge 0$ *and*

$$\mathrm{GreedyErr}_k(s) = \mathrm{Res}_k(s) - G_k(s), \tag{133}$$

$$= (1 - \lambda_{k,s})\, \mathrm{Res}_k(s) + \lambda_{k,s} \left[\delta^{\mathrm{PM}}_k(s) - b_s(1 - R_{k,s})\right]. \tag{134}$$

Consequently,

$$0 \le \mathrm{GreedyErr}_k(s) \le B^{\sigma}_{k,\varepsilon}(s) - G_k(s), \tag{135}$$

$$0 \le \mathrm{GreedyErr}_k(s) \le (1 - \lambda_{k,s}) B^{\sigma}_{k,\varepsilon}(s) + \lambda_{k,s} \left[\overline{\delta}^{\mathrm{PM}}_{k,\varepsilon}(s) - b_s(1 - R_{k,s})\right]. \tag{136}$$

The value gap satisfies the componentwise recursion

$$0 \le V^{\star}_w - V^{\sigma_{k+1}} \le \gamma P^{\pi^{\star}_w}\big(V^{\star}_w - V^{\sigma_k}\big) + \mathrm{GreedyErr}_k\,. \tag{137}$$

Writing $e_j := \mathrm{GreedyErr}_j$, for every $K \ge 1$,

$$0 \le V^{\star}_w - V^{\sigma_K} \le (\gamma P^{\pi^{\star}_w})^K (V^{\star}_w - V^{w}) + \sum_{j=0}^{K-1} (\gamma P^{\pi^{\star}_w})^{K-1-j} e_j, \tag{138}$$

$$\|V^{\star}_w - V^{\sigma_K}\|_{\infty} \le \gamma^K \overline{D} + \sum_{j=0}^{K-1} \gamma^{K-1-j} \|e_j\|_{\infty}. \tag{139}$$

The bounds in (135)–(136)*, optimized over statewise thresholds, may replace each e_j. In particular,* (136) *retains the exact gradient-extraction credit $b_s(1 - R_{k,s})$ instead of treating the PMPI step as an arbitrary approximate-greedy update.*

In Theorem 12, $\mathrm{GreedyErr}_k(s)$ is *not* an error in an estimated critic. It is the approximate-greedy error of the exact next policy: the difference between the best supported Bellman backup and the backup obtained by following σ_{k+1} for one step. Equation (133) says that this remaining error equals the current supported Bellman residual minus the improvement actually taken by PMPI. The recursion (137) is then the standard approximate-policy-iteration recursion.

The layer-cake identity and Hölder's inequality are standard analysis tools. The Power-Mirror-specific content is the explicit coverage envelope for δ_k^{PM}, the extra policy-extraction gain $b_s(1 - R_{k,s})$, and the ability to transfer action coverage from the logger through the cumulative likelihood-ratio bounds.

Corollary 7 (Logger transfer, state occupancy, and coverage stopping)**.** *Let*

$$\rho^-_{k,s} = \min_{a \in \mathcal{A}_w(s)} \frac{\sigma_k(a|s)}{w(a|s)}. \tag{140}$$

Then

$$\rho^-_{k,s} \geq \prod_{j=0}^{k-1} \ell_{j,s}, \qquad W^\sigma_{k,t}(s) \geq \rho^-_{k,s} W^w_{k,t}(s). \tag{141}$$

Define

$$\widetilde{W}^w_{k,\varepsilon}(s) = \rho^-_{k,s} W^w_{k,\varepsilon}(s), \tag{142}$$

$$B^w_{k,\varepsilon}(s) = D_{k,s} - (D_{k,s} - \varepsilon)\widetilde{W}^w_{k,\varepsilon}(s), \tag{143}$$

$$\overline{\delta}^{w,\mathrm{PM}}_{k,\varepsilon}(s) = \min\Big\{ D_{k,s},\ B^w_{k,\varepsilon}(s), \\ (b_s + \mathrm{Res}_k(s))\big(1 - \widetilde{W}^w_{k,\varepsilon}(s)^{1/p_s}\big) + \varepsilon \widetilde{W}^w_{k,\varepsilon}(s)^{1/p_s} \Big\}. \tag{144}$$

The upper bounds (135)–(136)*, and their ensuing recursion and stopping consequences, remain valid with $B^\sigma, \overline{\delta}^{\mathrm{PM}}$ replaced by the conservative logger versions $B^w, \overline{\delta}^{w,\mathrm{PM}}$. The exact layer-cake identity itself remains expressed under W^σ. For globally behavior-capped PMPI, the fixed lower cap $\underline{\rho}_s$ may replace $\rho^-_{k,s}$.*

For any statewise thresholds $0 \leq \varepsilon_s \leq D_{k,s}$,

$$0 \leq J_\nu(\pi^\star_w) - J_\nu(\sigma_k) \leq \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\nu^{\pi^\star_w}}[B^\sigma_{k,\varepsilon_s}(s)]. \tag{145}$$

If μ is a state probability distribution with $d_\nu^{\pi^\star_w} \ll \mu$, Hölder's inequality gives, for conjugate r, r',

$$J_\nu(\pi^\star_w) - J_\nu(\sigma_k) \leq \frac{1}{1-\gamma} \left\| \frac{d_\nu^{\pi^\star_w}}{\mu} \right\|_{L^r(\mu)} \|B^\sigma_{k,\varepsilon_s}\|_{L^{r'}(\mu)}. \tag{146}$$

Thus $r = \infty$ gives an L^∞ concentrability coefficient, while $r = 2$ gives $\sqrt{1 + \chi^2(d_\nu^{\pi^\star_w} \| \mu)}$. Finally,

$$\max_s \inf_{0 \leq \varepsilon \leq D_{k,s}} B^\sigma_{k,\varepsilon}(s) \leq (1-\gamma)\tau \quad \Longrightarrow \quad \|V^\star_w - V^{\sigma_k}\|_\infty \leq \tau, \tag{147}$$

and the same implication holds with B^w in place of B^σ. The realized gain sharpens the next-iterate scalar bound to

$$J_\nu(\pi^\star_w) - J_\nu(\sigma_{k+1}) \leq \frac{1}{1-\gamma} \left\{ \mathbb{E}_{d_\nu^{\pi^\star_w}}[B^\sigma_{k,\varepsilon_s}] - \mathbb{E}_{d_\nu^{\sigma_{k+1}}}[G_k] \right\}. \tag{148}$$

The action masses above are defined using the exact action-value function Q^{σ_k}, written as $Q_k(s, \cdot)$ in the equations, because this is the quantity that determines whether the next update is close to greedy. These masses are population quantities and cannot generally be read directly from a finite replay buffer. The lower likelihood-ratio bound transfers *action* probability from w to σ_k, but it does not guarantee that the improved policy visits the same *states* as the dataset. Replacing the optimal-policy occupancy in (145) by a data-state distribution therefore requires the separate concentrability factor in (146). A finite-sample theorem would additionally need valid confidence bounds for the estimated Q values.

Corollary 8 (Unrestricted PMPI near-optimality)**.** *Define the one-step behavior-support defect*

$$\chi_w(s) = V^\star(s) - \max_{a \in \mathcal{A}_w(s)} Q^\star(s,a), \tag{149}$$

and let $\overline{\pi}_w$ select a supported maximizer of $Q^\star(s,\cdot)$ at every state. Then

$$0 \le V^\star - V_w^\star \le (I - \gamma P^{\overline{\pi}_w})^{-1} \chi_w. \tag{150}$$

For every $K \ge 1$, combining this with Theorem 12 gives

$$\begin{aligned} 0 \le V^\star - V^{\sigma_K} &\le (I - \gamma P^{\overline{\pi}_w})^{-1} \chi_w + (\gamma P^{\pi_w^\star})^K (V_w^\star - V^w) \\ &\quad + \sum_{j=0}^{K-1} (\gamma P^{\pi_w^\star})^{K-1-j} e_j, \end{aligned} \tag{151}$$

$$\|V^\star - V^{\sigma_K}\|_\infty \le \frac{\|\chi_w\|_\infty}{1-\gamma} + \gamma^K \overline{D} + \sum_{j=0}^{K-1} \gamma^{K-1-j} \|e_j\|_\infty. \tag{152}$$

If $W^w_{\star,\varepsilon_s}(s) > 0$, then $\chi_w(s) \le \varepsilon_s$. The result separates three familiar offline-RL difficulties: coverage of globally good actions controls the permanent support bias; coverage of actions that are good under the current Q^{σ_k} controls how far a finite PMPI iterate is from the supported optimum; and state occupancy determines how the local gaps accumulate into return loss.

Why monotone improvement alone is insufficient. Value improvement at every step does not by itself imply convergence to the optimum. Consider a one-state, two-action bandit with rewards 1 and 0. With $p = 2$ and any initial policy that gives positive probability to both actions, the summable schedule $\lambda_k = 2^{-k-1}$ makes every nonzero update strictly better, but the total movement is too small to eliminate the inferior action. The policy converges to a suboptimal mixture. The same problem occurs if $p_k \downarrow 1$ or $b_k \uparrow \infty$ too quickly: the update becomes indistinguishable from behavior before reaching the optimum. The divergent-series condition associated with (115), or fixed parameters with nonvanishing damping, is sufficient for convergence; these examples do not establish necessity of either condition.

Population versus offline estimation. Theorems 8–11 are population results: they assume that the transition model P, reward function r, and behavior policy w are known, so every Q^{σ_k} can be evaluated exactly. This distinction is especially important for PMPI because its reference policy changes at every outer iteration. If Q^{σ_k} is replaced by an estimate $\widehat{Q}_k$, the monotone chain (112) and convergence to $V_w^\star$ do not follow automatically. The ratio envelope limits how action-value estimation error affects each update, but a strict finite-data guarantee needs a separate statistical argument. The next section states an action-conditional Q-error bound, and the following section gives an independent held-out deployment test.

G Offline safe-improvement template

We now leave the known-model population setting and isolate the effect of estimating an action-value function from offline data. Let $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ be a discounted MDP. Let $w_s = \pi_\beta(\cdot|s)$ be the behavior policy and let V^{π_β} be its value. All updates below use only $\mathcal{A}_w(s) = \{a : w_s(a) > 0\}$; the improved policy assigns zero probability to actions that never occur under behavior. The true behavior-policy action value is

$$Q^{\pi_\beta}(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^{\pi_\beta}(s'). \tag{153}$$

If this action-value function were known exactly, the one-step Power-Mirror policy would be

$$\pi_p(\cdot|s) = \nabla C_{p,w_s,b_s}(Q^{\pi_\beta}(s,\cdot)). \tag{154}$$

Proposition 4 (Exact reference-policy improvement)**.** *If Q^{π_β} is known exactly and the fixed shifts make the backups well defined, then*

$$V^{\pi_p} \ge V^{\pi_\beta}.$$

This is ordinary policy improvement: the new policy has nonnegative expected advantage under the true Q^{π_β} at every state. It is the one-step counterpart of Theorem 9, which repeats the same idea inside a complete behavior-anchored Bellman recursion.

In offline RL, Q^{π_β} is unknown. Let the critic output $\widehat{Q}(s,a)$ and define its action-value estimation error explicitly as

$$\Delta_Q(s,a) = \widehat{Q}(s,a) - Q^{\pi_\beta}(s,a), \qquad \widehat{Q}(s,a) = Q^{\pi_\beta}(s,a) + \Delta_Q(s,a). \tag{155}$$

The implemented policy is constructed from the estimated action values:

$$\widehat{\pi}_p(\cdot|s) = \nabla C_{p,w_s,b_s}(\widehat{Q}(s,\cdot)). \tag{156}$$

Theorem 13 (Behavior-ratio critic-error bound)**.** *Assume* $1 < p < \infty$ *and* $\operatorname{span}_a \widehat{Q}(s,a) \le D_s < b_s$ *for every state. Set*

$$\eta_s = D_s/b_s, \qquad H_s = H_{p,\eta_s}, \qquad U_s = U_{p,\eta_s}.$$

Define the improvement predicted by the learned Q function

$$\widehat{\Delta}_s = \langle \widehat{\pi}_p(\cdot|s), \widehat{Q}(s,\cdot)\rangle - \langle w_s, \widehat{Q}(s,\cdot)\rangle \ge 0.$$

Then the one-step improvement measured using the true Q^{π_β} *satisfies*

$$\langle \widehat{\pi}_p(\cdot|s), Q^{\pi_\beta}(s,\cdot)\rangle - \langle w_s, Q^{\pi_\beta}(s,\cdot)\rangle \ge \widehat{\Delta}_s - H_s \mathbb{E}_{a\sim w_s}|\Delta_Q(s,a)| \ge -H_s \mathbb{E}_{a\sim w_s}|\Delta_Q(s,a)|. \tag{157}$$

Consequently,

$$V^{\widehat{\pi}_p}(s) \ge V^{\pi_\beta}(s) - \frac{1}{1-\gamma} \max_{s'} H_{s'} \mathbb{E}_{a\sim w_{s'}}|\Delta_Q(s',a)|. \tag{158}$$

In particular, if

$$\widehat{\Delta}_s \ge H_s \mathbb{E}_{a\sim w_s}|\Delta_Q(s,a)| \qquad \textit{for every state } s, \tag{159}$$

then $V^{\widehat{\pi}_p} \ge V^{\pi_\beta}$ *componentwise. The same deviation also admits the* $L_2(w_s)$ *bound*

$$|\langle \widehat{\pi}_p - w_s, \Delta_Q(s,\cdot)\rangle| \le \sqrt{\chi^2(\widehat{\pi}_p \| w_s)}\, \|\Delta_Q(s,\cdot)\|_{2,w_s} \le \frac{H_s}{2} \|\Delta_Q(s,\cdot)\|_{2,w_s}. \tag{160}$$

Why this matters. Offline data can usually support an average Q-error guarantee under behavior actions, but not a uniform guarantee over every possible action. The pointwise policy-ratio cap converts the behavior-weighted Q error into error under the candidate policy:

$$\mathbb{E}_{a\sim\widehat{\pi}_p}|\Delta_Q(s,a)| \le U_s \mathbb{E}_{a\sim w_s}|\Delta_Q(s,a)|. \tag{161}$$

The upper ratio U_s controls how much the candidate can amplify average action-value error. The smaller constant H_s controls the specific error that enters the comparison between the candidate and behavior. Both are explicit functions of p and the normalized Q-value range D_s/b_s. As $p \downarrow 1$, the policy returns to behavior and $H_s \to 0$, so the error penalty vanishes together with the size of the policy update. Condition (159) is therefore easy to interpret: the improvement predicted by $\widehat{Q}$ must exceed the worst behavior-weighted Q-estimation error induced by moving the policy.

State-distribution caveat. The theorem compares actions at a fixed state. It does not by itself guarantee that the candidate visits only well-covered states. The next section addresses the end-to-end return directly by using held-out behavior trajectories and exact trajectory likelihood ratios. The two guarantees play different roles: Theorem 13 explains how Q-estimation error affects one policy-improvement step, whereas the deployment shield certifies the final policy without assuming that this unknown error bound can be computed from the replay buffer.

H STRICT HIGH-PROBABILITY DEPLOYMENT SHIELD

The population theorems assume exact Q functions. A practical offline algorithm uses estimated Q functions, so we add a separate final decision rule: deploy the learned candidate only when independent data provide statistically significant evidence that its return exceeds the behavior policy's return; otherwise deploy behavior. This accept-or-fallback rule is the *deployment shield*.

The candidate can come from fitted PMVI, fitted PMPI, one-step Power-Mirror improvement, or any other training method. The theorem needs only three verifiable inputs:

1. the candidate is chosen without using the certification trajectories;
2. its likelihood ratio relative to the *true* behavior policy can be evaluated and has known lower and upper bounds; and
3. the certification set contains independent trajectories collected by that behavior policy.

For PMPI, training includes all outer iterations, hyperparameter choices, and the choice of stopping iterate. Certification must use the final ratio σ_K/w, not only the ratio between the last two PMPI policies. The construction is a high-confidence policy-improvement test in the spirit of Thomas et al. (2015); the Power-Mirror contribution is the explicit two-sided ratio range that controls its importance weights.

Fix the deployment initial-state distribution ν. The expected discounted return of a policy is

$$J_\nu(\pi) = \mathbb{E}_{\nu,\pi}\left[\sum_{t=0}^{\infty} \gamma^t Y_t\right],$$

where $Y_t \in [r_{\min}, r_{\max}]$ almost surely. Let $R_\Delta = r_{\max} - r_{\min}$ and center the rewards at the midpoint of their known range: $c_r = (r_{\min} + r_{\max})/2$ and $\widetilde{Y}_t = Y_t - c_r \in [-R_\Delta/2, R_\Delta/2]$. This does not change the return difference $J_\nu(\pi) - J_\nu(w)$ and gives the tightest worst-case range obtainable from the reward bounds alone. If episodes terminate early, pad them to the common horizon with an absorbing state and zero rewards, assuming $0 \in [r_{\min}, r_{\max}]$. This padding must precede centering; otherwise policies with different stopping times receive different constant shifts.

Split the offline data *before* training. Use the training part to construct a candidate π. Keep the certification part untouched; it consists of n independent behavior trajectories of length H,

$$\tau_j = (S_{j,0}, A_{j,0}, Y_{j,0}, \ldots, S_{j,H}), \qquad j = 1, \ldots, n,$$

drawn from the true behavior policy w, with the same initial-state distribution $S_{j,0} \sim \nu$ that will be used at deployment. Once training is complete, treat π and its constants as fixed. Assume π never selects an action that behavior cannot select and that known bounds $0 \le L \le 1 \le U < \infty$ satisfy

$$L \le \frac{\pi(a|s)}{w(a|s)} \le U \qquad \text{for every behavior-supported } (s,a), \text{ or } w\text{-a.e. in continuous actions.} \quad (162)$$

For an exact Power-Mirror candidate, use the most conservative bounds across states,

$$L = \inf_s L_{p_s,\eta_s}, \qquad U = \sup_s U_{p_s,\eta_s}.$$

For a learned actor, (162) must hold for the density that will actually be deployed. A bound for an ideal target before actor fitting is insufficient, as is a maximum observed only over sampled state–action pairs. If the ratio bounds are themselves statistical rather than deterministic, their failure probability must be added to the shield's total failure probability.

For trajectory j, define the prefix importance weight and the paired per-decision importance-sampling (PDIS) return difference

$$W_{j,t} = \prod_{u=0}^{t} \frac{\pi(A_{j,u}|S_{j,u})}{w(A_{j,u}|S_{j,u})}, \quad (163)$$

$$Z_j(\pi) = \sum_{t=0}^{H-1} \gamma^t \big(W_{j,t} - 1\big)\widetilde{Y}_{j,t}, \qquad \widehat{\Delta}_H(\pi) = \frac{1}{n}\sum_{j=1}^{n} Z_j(\pi). \quad (164)$$

Here $W_{j,t}$ reweights the trajectory prefix through time t from behavior to the candidate. The term "-1" evaluates the behavior baseline on the same trajectory, reducing unnecessary variance compared with two independent return estimates. The standard per-decision change-of-measure identity gives

$$\mathbb{E}_w Z_j(\pi) = J_{\nu,H}(\pi) - J_{\nu,H}(w), \quad (165)$$

where $J_{\nu,H}$ is the return truncated after time $H-1$.

The pointwise action-ratio bounds imply $L^{t+1} \le W_{j,t} \le U^{t+1}$. These prefix bounds give a deterministic range for every trajectory statistic. Define

$$M_t(L,U) = \max\left\{1 - L^{t+1},\ U^{t+1} - 1\right\}, \tag{166}$$

$$\mathcal{B}_H(L,U) = R_\Delta \sum_{t=0}^{H-1} \gamma^t M_t(L,U), \qquad \mathcal{T}_H = \frac{R_\Delta \gamma^H}{1-\gamma}. \tag{167}$$

Then

$$-\frac{\mathcal{B}_H(L,U)}{2} \le Z_j(\pi) \le \frac{\mathcal{B}_H(L,U)}{2} \qquad \text{almost surely.} \tag{168}$$

The width $\mathcal{B}_H(L,U)$ measures the worst possible variation of the paired PDIS statistic over the first H steps. The term $\mathcal{T}_H$ is the largest return difference that could remain after truncation. Both are known before inspecting the certification returns.

Safety-shield algorithm. Before opening the certification fold, choose the truncation horizon H, confidence-bound family (Hoeffding or empirical Bernstein), failure probability $\delta \in (0,1)$, and required return margin $m \ge 0$. If several choices will be compared, allocate confidence across all of them as in Corollary 10.

1. **Train and audit.** Train the candidate without the certification fold. If its support, exact likelihood ratio, or deterministic bounds (L,U) cannot be verified relative to the true behavior policy, reject it and deploy w.
2. **Estimate improvement.** Otherwise compute the paired PDIS estimate in (163)–(164) and its Hoeffding lower confidence bound

$$\mathrm{LCB}_H^{\mathrm{Hfd}}(\pi) = \widehat{\Delta}_H(\pi) - \mathcal{B}_H(L,U)\sqrt{\frac{\log(1/\delta)}{2n}} - \mathcal{T}_H. \tag{169}$$

3. **Accept or fall back.** Deploy

$$\pi_{\mathrm{dep}} = \begin{cases} \pi, & \mathrm{LCB}_H^{\mathrm{Hfd}}(\pi) \ge m, \\ w, & \text{otherwise.} \end{cases} \tag{170}$$

For a genuinely strict improvement certificate, choose $m > 0$. Taking $m = 0$ certifies non-degradation. Notice that a rejected candidate is not declared harmful; the data are simply insufficient to distinguish it safely from behavior.

Theorem 14 (Strict high-probability deployment safety)**.** *Under the preceding assumptions, conditional on the training fold and hence also unconditionally,*

$$\Pr\left\{J_\nu(\pi) - J_\nu(w) \ge \mathrm{LCB}_H^{\mathrm{Hfd}}(\pi)\right\} \ge 1 - \delta. \tag{171}$$

Consequently, with probability at least $1-\delta$*, the output of* (170) *satisfies*

$$J_\nu(\pi_{\mathrm{dep}}) \ge J_\nu(w). \tag{172}$$

Moreover, whenever the shield deploys the candidate on this event,

$$J_\nu(\pi) \ge J_\nu(w) + m. \tag{173}$$

For a truly finite-horizon objective $J_{\nu,H}$*, the same statements hold with* $\mathcal{T}_H = 0$*.*

The safety statement is about the policy that the rule actually returns. On the $1-\delta$ event, fallback gives exactly $J_\nu(w)$, while acceptance gives at least $J_\nu(w) + m$. The positive lower ratio bound helps because it controls negative as well as positive deviations of each prefix weight. If only an upper bound is available, setting $L = 0$ is still valid but more conservative. For Power Mirror, $L > 0$, and $\mathcal{B}_H(L,U)$ decreases continuously to zero as $p \downarrow 1$ and the candidate approaches behavior.

When the shield is nontrivial. Safety is automatic because rejection falls back to behavior, but a useful shield should also accept genuinely better candidates with enough data. Let $\Delta_H = J_{\nu,H}(\pi) - J_{\nu,H}(w)$ be the candidate's true truncated-horizon advantage. A second one-sided Hoeffding argument shows that the candidate is accepted with probability at least $1-\delta$ whenever

$$\Delta_H \geq m + \mathcal{T}_H + 2\mathcal{B}_H(L,U)\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{174}$$

Equivalently, if $g_H = \Delta_H - m - \mathcal{T}_H > 0$, the sufficient sample size

$$n \geq \frac{2\mathcal{B}_H(L,U)^2 \log(1/\delta)}{g_H^2} \tag{175}$$

makes acceptance likely. The scaling is intuitive: more trajectories are required when importance weights have a wider range, the confidence level is higher, or the candidate's improvement above the requested margin is smaller. This is only an acceptance condition, not a condition for safety; when it fails, fallback remains valid.

Corollary 9 (Empirical-Bernstein shield)**.** *Let $n \geq 2$ and define*

$$\widehat{v}_Z = \frac{1}{n-1}\sum_{j=1}^{n}\big(Z_j(\pi) - \widehat{\Delta}_H(\pi)\big)^2.$$

If the empirical-Bernstein rule is chosen before inspecting the certification fold, then the alternative lower bound

$$\mathrm{LCB}_H^{\mathrm{EB}}(\pi) = \widehat{\Delta}_H(\pi) - \sqrt{\frac{2\widehat{v}_Z \log(2/\delta)}{n}} - \frac{7\mathcal{B}_H(L,U)\log(2/\delta)}{3(n-1)} - \mathcal{T}_H \tag{176}$$

satisfies

$$\Pr\left\{J_\nu(\pi) - J_\nu(w) \geq \mathrm{LCB}_H^{\mathrm{EB}}(\pi)\right\} \geq 1-\delta. \tag{177}$$

This is the one-sided empirical-Bernstein inequality of Maurer & Pontil (2009). It can be tighter than Hoeffding when the observed paired PDIS differences have low variance. Replacing (169) *by* (176) *in the deployment rule preserves Theorem 14. One may not inspect the data and then report whichever bound is larger without a simultaneous confidence correction.*

Corollary 10 (Simultaneous candidate selection)**.** *Suppose M candidates and their ratio bounds are fixed without the certification fold. Assign levels $\delta_1, \ldots, \delta_M$ with $\sum_{k=1}^{M}\delta_k \leq \delta$ and construct one valid lower confidence bound for each candidate at level $1-\delta_k$. With probability at least $1-\delta$, every accepted candidate satisfies its certified margin simultaneously. The shield may then select any accepted candidate, such as the one with the largest lower confidence bound. This correction is required when the certification fold is used to compare powers p, shifts b, actor architectures, or PMPI stopping iterates; unrestricted tuning on the same fold invalidates the nominal guarantee.*

What makes the certificate strict. The theorem certifies ordinary infinite-horizon return, not a regularized objective or a one-step Q surrogate. Because the likelihood ratio covers the entire trajectory prefix, the test includes both action and state-distribution shift. It also does not require the Q-estimation error in Theorem 13 to be observable. The price is visible in $\mathcal{B}_H$: even a modest one-step ratio can compound over a long horizon, so a mathematically valid shield may often fall back to behavior. Empirical Bernstein can improve acceptance in low-variance problems but cannot remove this worst-case horizon dependence.

Non-negotiable audit conditions. The shield is valid only under the following auditable conditions:

- Certification trajectories are independent at the episode level. A single dependent trajectory needs a different martingale or mixing analysis.
- Certification and deployment use the same initial-state distribution, dynamics, and reward law.
- The denominator in (163) is the true logging probability, not only a fitted behavior-model probability, and w remains executable as the fallback policy.
- The ratio of the *deployed* actor to the true logger is evaluable. Section J.2 gives a bound relative to the actor's base distribution. This becomes a logger-relative bound only if that base is w or the base-to-logger ratio in (221) is also certified.

- A sampler-only actor with unknown normalizer can be executable and ratio-bounded but still fail this likelihood-evaluation requirement. Weighted-MLE distillation, sampled ratio maxima, and average behavior-model likelihood also do not establish it.
- The reward bounds hold almost surely; empirical minima and maxima are not substitutes. Clipped or self-normalized importance weights change the estimator and require a separate bias analysis.

The conclusion concerns expected return under the specified initial distribution ν. It does not claim componentwise improvement at every starting state or improvement on every realized deployment trajectory. If any audit condition fails, the shield returns w rather than making an unsupported safety claim.

I AN INTERPRETABLE SPECIAL CASE: $p = 2$

I.1 QUADRATIC FORM AND DUAL INTERPRETATION

The choice $p = 2$ is especially easy to interpret because it depends on the behavior-weighted mean and variance of the action values. At one state, write $q_i = Q(s, a_i)$ and let w_i be the behavior probability. The backup becomes

$$C_{2,w,b}(q) = \mu_w(q) + \sqrt{b^2 + \mathrm{Var}_w(q)} - b. \tag{178}$$

The corresponding policy likelihood ratio is

$$\frac{\pi_{2,i}(q)}{w_i} = 1 + \frac{q_i - \mu_w(q)}{\sqrt{b^2 + \mathrm{Var}_w(q)}}. \tag{179}$$

An action receives more probability than under behavior exactly when its Q value is above the behavior-weighted mean; below-average actions lose probability. The denominator combines the shift b and the behavior-weighted variance, preventing an extreme update. The sharp ratio theorem reduces to the linear interval

$$1 - \eta \le \frac{\pi_{2,i}(q)}{w_i} \le 1 + \eta, \qquad H_{2,\eta} = \eta, \tag{180}$$

which follows directly from (179): the centered Q values have range at most D, while the denominator is at least b. Thus D/b is the maximum additive change in any behavior likelihood ratio at $p = 2$.

Proposition 5 ($p = 2$ Fenchel dual)**.** *For $p = 2$ and $\mathrm{span}(q) < b$,*

$$C_{2,w,b}(q) = \max_{\pi \in \Delta_K} \left\{ \langle \pi, q \rangle - \Omega_{2,w,b}(\pi) \right\}, \tag{181}$$

where

$$\Omega_{2,w,b}(\pi) = \begin{cases} b\left(1 - \sqrt{1 - \chi^2(\pi\|w)}\right), & \chi^2(\pi\|w) \le 1, \\ +\infty, & \textit{otherwise,} \end{cases} \tag{182}$$

and

$$\chi^2(\pi\|w) = \sum_i \frac{(\pi_i - w_i)^2}{w_i}.$$

Near w,

$$\Omega_{2,w,b}(\pi) = \frac{b}{2}\chi^2(\pi\|w) + O(\chi^2(\pi\|w)^2). \tag{183}$$

The dual form gives a second interpretation. Equation (181) chooses the policy with the largest expected Q value after subtracting the penalty in (182). That penalty depends on the Pearson χ^2 divergence from behavior and becomes approximately $(b/2)\chi^2(\pi\|w)$ near w. Consequently, $p = 2$ connects Power Mirror to familiar χ^2 trust regions, variance regularization, and distributionally robust optimization (Namkoong & Duchi, 2017; Duchi & Namkoong, 2021). We use it as the clearest special case, not as a claim that χ^2 duality itself is new.

I.2 Quadratic improvement efficiency and critic sensitivity

We work at one state and restrict the action set to the support of the reference policy. Thus $w = (w_1, \dots, w_K)$ has $w_i > 0$ and $\sum_i w_i = 1$. Define

$$\langle f, g\rangle_w = \sum_i w_i f_i g_i, \qquad \|f\|_w = \sqrt{\langle f, f\rangle_w}, \qquad P_w f = f - \mathbb{E}_w f.$$

Every supported probability policy is uniquely written as $\pi_i = w_i r_i$, where $r_i \geq 0$ and $\mathbb{E}_w r = 1$. Its Pearson divergence is

$$\chi^2(\pi\|w) = \|r - 1\|_w^2. \tag{184}$$

Fix a critic $\widehat{Q}$ and a shift $b > \operatorname{span}(\widehat{Q})$. Throughout this subsection,

$$\widehat{A} = P_w\widehat{Q}, \qquad \widehat{\sigma} = \|\widehat{A}\|_w, \qquad m = \sqrt{b^2 + \widehat{\sigma}^2}, \qquad t = \frac{\widehat{\sigma}}{m}, \qquad r_2 = 1 + \frac{\widehat{A}}{m}. \tag{185}$$

The shorthand m denotes the quadratic shifted moment $\widehat{m}_2$ in the main text. The quantity t^2 is the movement induced by the fixed-shift construction; the results below do not instruct the algorithm to recompute b from a desired radius at every Bellman iteration.

Proof of Proposition 1. Since $\mathbb{E}_w\widehat{A} = 0$, we have $\mathbb{E}_w r_2 = 1$. Also $|\widehat{A}_i| \leq \operatorname{span}(\widehat{Q}) < b \leq m$, so $r_{2,i} > 0$. Hence $\widehat{\pi}_2 = w r_2$ is a valid probability policy. Direct calculation gives

$$\|r_2 - 1\|_w^2 = \frac{\widehat{\sigma}^2}{m^2} = t^2, \qquad \widehat{G}(\widehat{\pi}_2) = \left\langle \frac{\widehat{A}}{m}, \widehat{A} \right\rangle_w = \frac{\widehat{\sigma}^2}{m} = \widehat{\sigma} t. \tag{186}$$

For any supported probability policy $\pi = wr$, normalization gives $\mathbb{E}_w(r - 1) = 0$. Consequently,

$$\begin{aligned} \widehat{G}(\pi) &= \langle r - 1, \widehat{Q}\rangle_w = \langle r - 1, \widehat{A}\rangle_w \\ &\leq \|r - 1\|_w\, \widehat{\sigma}. \end{aligned} \tag{187}$$

If $\chi^2(\pi\|w) \leq t^2$, the last expression is at most $\widehat{\sigma} t$, which is attained by r_2. If $\widehat{\sigma} > 0$, equality at this maximal gain forces $r - 1 = (t/\widehat{\sigma})\widehat{A}$ by the equality condition in Cauchy–Schwarz. The optimizer is therefore unique. If $\widehat{\sigma} = 0$, then $t = 0$ and the feasible set contains only the behavior policy, so the statement also holds.

For the true-score bound, let $e = P_w(Q - \widehat{Q})$ and $\varepsilon_Q = \|e\|_w$. Normalization again removes the constant part of the critic error, yielding

$$\begin{aligned} \mathbb{E}_{\widehat{\pi}_2} Q - \mathbb{E}_w Q &= \widehat{G}(\widehat{\pi}_2) + \langle r_2 - 1, e\rangle_w \\ &\geq \widehat{\sigma} t - t\varepsilon_Q = \frac{\widehat{\sigma}(\widehat{\sigma} - \varepsilon_Q)}{m}. \end{aligned} \tag{188}$$

This proves all statements of the proposition. □

Equivalent minimum-movement interpretation. If $\widehat{\sigma} > 0$ and a policy achieves predicted gain at least $\widehat{G}(\widehat{\pi}_2) = \widehat{\sigma} t$, then (187) implies $\chi^2(\pi\|w) \geq t^2$. Thus the quadratic update minimizes the squared ratio movement needed to attain its predicted gain. Moreover, $\mathbb{E}_w r^2 = 1 + \chi^2(\pi\|w)$, so it also minimizes the second moment of the one-step importance ratio at that gain. This identity does not give a horizon-independent bound on trajectory importance weights.

Corollary 11 (Exact worst-case gain over a centered critic-error ball)**.** *For $\varepsilon \geq 0$, define*

$$\mathcal{U}_\varepsilon = \{e : \mathbb{E}_w e = 0,\ \|e\|_w \leq \varepsilon\}, \qquad \mathcal{R}_w(t) = \{r \geq 0 : \mathbb{E}_w r = 1,\ \|r - 1\|_w \leq t\}.$$

With the definitions in (185)*,*

$$\max_{r \in \mathcal{R}_w(t)} \min_{e \in \mathcal{U}_\varepsilon} \langle r - 1, \widehat{A} + e\rangle_w = t(\widehat{\sigma} - \varepsilon)_+, \tag{189}$$

where $(x)_+ = \max\{x, 0\}$. If $\widehat{\sigma} \geq \varepsilon$, r_2 attains the maximum. If $\widehat{\sigma} < \varepsilon$, the behavior ratio $r = 1$ attains the maximum; when $\widehat{\sigma} > 0$, the undamped quadratic update has strictly negative worst-case gain.

Proof. Fix a feasible ratio and put $u = r - 1$. Since $\mathbb{E}_w u = 0$, Cauchy–Schwarz gives

$$\min_{e \in \mathcal{U}_\varepsilon} \langle u, \widehat{A} + e \rangle_w = \langle u, \widehat{A} \rangle_w - \varepsilon \|u\|_w. \tag{190}$$

For $u \neq 0$, equality is attained by $e = -\varepsilon u / \|u\|_w$, which is centered and has norm ε. For $u = 0$, both sides are zero. Applying Cauchy–Schwarz once more gives

$$\langle u, \widehat{A} \rangle_w - \varepsilon \|u\|_w \leq (\widehat{\sigma} - \varepsilon) \|u\|_w.$$

If $\widehat{\sigma} \geq \varepsilon$, this is at most $t(\widehat{\sigma} - \varepsilon)$. The valid ratio r_2 attains the bound by (186). If $\widehat{\sigma} < \varepsilon$, every nonzero u has strictly negative worst-case gain, while $u = 0$ achieves zero. This proves (189), including the degenerate case $\widehat{\sigma} = t = 0$. □

Scope of the error interpretation. The uncertainty set in Corollary 11 is a ball of centered score errors; it is not a claim that every error in this ball is realized by an MDP. No error radius is inferred from the data here. When $Q = Q^{\pi_\beta}$ and the bound in (188) is nonnegative at every state, the ordinary policy-improvement argument gives $V^{\widehat{\pi}_2} \geq V^{\pi_\beta}$. A bound at only one state does not establish that global conclusion. The statement also does not imply that reducing the step size repairs an uninformative critic: the sign of the displayed bound is controlled by $\widehat{\sigma} - \varepsilon_Q$. Actor approximation and an estimated behavior policy require their own error accounting.

Proposition 6 (Stability for a fixed behavior policy and shift)**.** *Fix w and $b > 0$. Let Q, Q' satisfy $\operatorname{span}(Q) < b$ and $\operatorname{span}(Q') < b$, and let $r_2[Q], r_2[Q']$ be their quadratic Power-Mirror ratios with this same reference and shift. Then*

$$\|r_2[Q] - r_2[Q']\|_w \leq \frac{1}{b} \|P_w(Q - Q')\|_w, \tag{191}$$

$$\mathrm{TV}(w r_2[Q], w r_2[Q']) \leq \frac{1}{2b} \|P_w(Q - Q')\|_w. \tag{192}$$

The inequalities have no inverse minimum-probability factor.

Proof. Work in the weighted Euclidean space with inner product $\langle \cdot, \cdot \rangle_w$, and define

$$F(A) = \frac{A}{\sqrt{b^2 + \|A\|_w^2}}.$$

For any direction H, differentiation gives

$$DF(A)[H] = \frac{H}{\sqrt{b^2 + \|A\|_w^2}} - \frac{A \langle A, H \rangle_w}{(b^2 + \|A\|_w^2)^{3/2}}. \tag{193}$$

If $A \neq 0$, decompose H into its component parallel to A and its orthogonal component. On the parallel component, the derivative multiplies by $b^2/(b^2 + \|A\|_w^2)^{3/2}$; on the orthogonal component, it multiplies by $1/\sqrt{b^2 + \|A\|_w^2}$. Both factors lie in $(0, 1/b]$. At $A = 0$, the derivative is H/b. Therefore $\|DF(A)\|_{\mathrm{op},w} \leq 1/b$ for every A.

Set $A = P_w Q$ and $A' = P_w Q'$. Integrating the derivative along the line segment between them gives

$$F(A) - F(A') = \int_0^1 DF(A' + u(A - A'))[A - A']\, du,$$

and hence $\|F(A) - F(A')\|_w \leq \|A - A'\|_w / b$. Since $r_2[Q] = 1 + F(P_w Q)$, this proves (191). Finally,

$$\mathrm{TV}(w r_2[Q], w r_2[Q']) = \tfrac{1}{2} \mathbb{E}_w |r_2[Q] - r_2[Q']| \leq \tfrac{1}{2} \|r_2[Q] - r_2[Q']\|_w,$$

which proves (192). □

A comparison with direct ratio-box optimization. Fix $w = (1/2, 1/2)$, $b > 0$, and $\eta \in (0, 1)$. Use the common permitted interval $[1 - \eta, 1 + \eta]$. For $0 < h \leq \eta b/2$, consider $Q^+ = (h, -h)$ and $Q^- = (-h, h)$. Both have span at most $\eta b < b$, so their quadratic Power-Mirror policies satisfy this

same interval. Writing a normalized ratio as $r = (1 + x, 1 - x)$, the box constraint is $|x| \le \eta$. The objective $\mathbb{E}_w[rQ^+]$ equals hx, while $\mathbb{E}_w[rQ^-] = -hx$. Thus the unique box optimizers are

$$\pi^+_{\mathrm{box}} = \left(\frac{1+\eta}{2}, \frac{1-\eta}{2}\right), \qquad \pi^-_{\mathrm{box}} = \left(\frac{1-\eta}{2}, \frac{1+\eta}{2}\right),$$

and

$$\mathrm{TV}(\pi^+_{\mathrm{box}}, \pi^-_{\mathrm{box}}) = \eta. \tag{194}$$

In contrast, the quadratic ratios are

$$r^\pm_2 = \left(1 \pm \frac{h}{\sqrt{b^2+h^2}}, 1 \mp \frac{h}{\sqrt{b^2+h^2}}\right),$$

so

$$\mathrm{TV}(wr^+_2, wr^-_2) = \frac{h}{\sqrt{b^2+h^2}} \longrightarrow 0 \quad \text{as } h \downarrow 0. \tag{195}$$

This example matches permitted ratio endpoints, not realized movement. It establishes a continuity distinction near ties, not a return-dominance result. The box optimizer achieves the larger predicted gain within its box. KL updates are also continuous for a fixed positive temperature.

What the characterization contributes. The quadratic update is the classical χ^2 trust-region direction at the radius it induces; this geometry is already part of divergence-constrained policy improvement (Belousov & Peters, 2018). Its role here is to explain the centered power construction and provide explicit sensitivity constants. At other powers, the sharp pointwise ratio bounds still apply, but the χ^2 efficiency statement is not asserted. All stability bounds above compare critics using the same w and the same fixed b. Recomputing the shift from each perturbed critic defines a different map and is not covered by this proposition.

J Continuous-action power mirror

J.1 Continuous-action formulation and guarantees

We now replace sums over discrete actions by expectations under a continuous behavior density. Let $w(\cdot|s)$ be the behavior density on a continuous action space $\mathcal{A}$, and let $Q_s(a) = Q(s, a)$ be a bounded action-value function. At each state, define its behavior-weighted mean and positive shifted value by

$$\mu_w(Q_s) = \int_{\mathcal{A}} Q_s(a) w(a|s)\, da, \qquad z_s(a) = b_s + Q_s(a) - \mu_w(Q_s).$$

For $1 \le p < \infty$,

$$C_{p,w,b}(Q_s) = \mu_w(Q_s) + \left(\int_{\mathcal{A}} z_s(a)^p w(a|s)\, da\right)^{1/p} - b_s. \tag{196}$$

As before, choose b_s larger than the essential range of $Q(s, \cdot)$, which guarantees $z_s(a) > 0$ for w-almost every action.

Theorem 15 (Continuous induced density and ratio envelope)**.** *For $1 < p < \infty$, let*

$$m_p = \left(\int z(a)^p\, dw(a)\right)^{1/p}, \qquad A(a) = \left(\frac{z(a)}{m_p}\right)^{p-1}, \qquad R = \int A(a)\, dw(a).$$

The functional derivative of $C_{p,w,b}$ induces a policy $\pi_p \ll w$ with

$$\frac{d\pi_p}{dw}(a) = A(a) + 1 - R. \tag{197}$$

Moreover $R \le 1$. If

$$\operatorname*{ess\,sup}_a Q(a) - \operatorname*{ess\,inf}_a Q(a) \le D < b,$$

then, with $\eta = D/b$,

$$L_{p,\eta} \le \frac{d\pi_p}{dw}(a) \le U_{p,\eta} \qquad w\text{-a.s.} \tag{198}$$

and the essential additive and log-ratio oscillations are bounded by $H_{p,\eta}$ and $\log K_{p,\eta}$, respectively.

Equation (197) is the continuous analogue of the finite-action policy formula: it gives the candidate-to-behavior density ratio directly. The ratio integrates to one, is positive, and lies in the same explicit interval $[L_{p,\eta}, U_{p,\eta}]$. Hence finite-p Power-Mirror improvement and w have exactly the same support, with

$$\frac{dw}{d\pi_p} \le K_{p,\eta} \qquad \pi_p\text{-a.s.}$$

Corollary 1 transfers with sums replaced by integrals. In contrast, a hard greedy policy in a continuous action space often becomes a Dirac mass at one maximizing action and is therefore singular relative to a smooth behavior density. Power Mirror remains an ordinary density with explicit forward and reverse likelihood-ratio bounds.

The exact one-step PMPI argument also transfers when the current policy replaces w: with an exact Q function and the same positive-shift condition, the continuous candidate has nonnegative expected advantage and therefore improves the ordinary policy value. We do not claim the finite-action convergence theorem for arbitrary continuous spaces; its proof uses the compact finite-dimensional policy simplex and finite greedy action sets. The actors below solve a narrower but practically central problem: they ensure that function approximation cannot violate a chosen pointwise density-ratio interval. They preserve the ratio certificate automatically, but they preserve value improvement only when the ideal gain is larger than the actor-fitting and Q-estimation errors in Theorem 18.

Proposition 7 (Continuous near-optimal coverage)**.** *Let*

$$G_\varepsilon(Q) = \{a : Q(a) \ge \operatorname*{ess\,sup}_{a'} Q(a') - \varepsilon\}, \quad W_\varepsilon = w(G_\varepsilon(Q)),$$
$$M_Q = b + \operatorname*{ess\,sup}_a Q(a) - \mathbb{E}_w Q.$$

If $\operatorname{ess\,sup} Q - \operatorname{ess\,inf} Q \le D \le b$*, then*

$$0 \le \operatorname*{ess\,sup}_a Q(a) - C_{p,w,b}(Q) \le \min\left\{D,\ \varepsilon W_\varepsilon + D(1-W_\varepsilon),\ M_Q(1-W_\varepsilon^{1/p}) + \varepsilon W_\varepsilon^{1/p}\right\}. \tag{199}$$

J.2 Ratio-preserving continuous actors

The ideal density in (197) will rarely be exactly Gaussian or belong to a chosen normalizing-flow family. Fitting a free Gaussian or flow by weighted maximum likelihood does not preserve the pointwise ratio theorem: a small average KL error can coexist with an unbounded density ratio in the tails. We therefore keep a samplable base policy $\bar{w}(\cdot|s)$ inside the deployed actor and learn only a bounded multiplicative reweighting. In ordinary probability notation,

$$\pi_\theta(da|s) = r_\theta(s,a)\bar{w}(da|s), \qquad \mathbb{E}_{\bar{w}} r_\theta(s,\cdot) = 1.$$

The multiplier $r_\theta = d\pi_\theta/d\bar{w}$ is simply the actor-to-base likelihood ratio. The base may be the true executable behavior policy w or a fitted behavior model; only the former immediately gives a certificate relative to the logged data.

Definition 2 (Mass-calibrated bounded gate)**.** Fix a state, $0 < L < 1 < U < \infty$, and

$$\tau = \frac{1-L}{U-L}.$$

For any measurable raw gate $h : \mathcal{A} \to [0,1]$, write $\mu_h = \mathbb{E}_{\bar{w}} h$ and define

$$\mathfrak{C}_\tau[h](a) = \begin{cases} \dfrac{\tau}{\mu_h} h(a), & \mu_h > \tau, \\ h(a), & \mu_h = \tau, \\ 1 - \dfrac{1-\tau}{1-\mu_h}\big(1-h(a)\big), & \mu_h < \tau. \end{cases} \tag{200}$$

When $\mu_h = 0$, only the third branch applies; when $\mu_h = 1$, only the first applies, so every active denominator is nonzero.

The raw gate $h_\theta(s,a) \in [0,1]$ can be any neural network with a bounded output. The calibration map in (200) changes its mean to the single required value τ_s. The affine transformation in the next theorem then turns it into a normalized likelihood ratio between L_s and U_s.

Theorem 16 (Sharp-box ratio-preserving actor)**.** *Let $h_\theta(s,\cdot) : \mathcal{A} \to [0,1]$ be any measurable parametric gate and define, statewise,*

$$g_\theta(s,a) = \mathfrak{C}_{\tau_s}[h_\theta(s,\cdot)](a), \tag{201}$$

$$r_\theta(s,a) = L_s + (U_s - L_s) g_\theta(s,a), \tag{202}$$

$$\pi_\theta(da|s) = r_\theta(s,a)\bar{w}(da|s), \tag{203}$$

where $0 < L_s < 1 < U_s$ and $\tau_s = (1-L_s)/(U_s - L_s)$. Then, for every parameter value,

$$\mathbb{E}_{\bar{w}} g_\theta(s,\cdot) = \tau_s, \qquad \mathbb{E}_{\bar{w}} r_\theta(s,\cdot) = 1, \qquad L_s \le \frac{d\pi_\theta}{d\bar{w}}(a|s) \le U_s \quad \bar{w}\text{-a.s.} \tag{204}$$

Thus π_θ is a probability measure mutually absolutely continuous with $\bar{w}$, and the certificate is unaffected by critic error, optimization error, early stopping, or actor-class approximation.

The class is ratio-universal on this box. If a target $r_\star$ obeys $L_s \le r_\star \le U_s$ and $\mathbb{E}_{\bar{w}} r_\star = 1$, then

$$g_\star(s,a) = \frac{r_\star(s,a) - L_s}{U_s - L_s} \tag{205}$$

has mean τ_s and is fixed by the calibration. Hence any raw-gate class containing $g_\star$ represents the target exactly. In particular, choosing $L_s = L_{p_s,\eta_s}$ and $U_s = U_{p_s,\eta_s}$ includes the ideal continuous power-mirror density constructed relative to the same base $\bar{w}$. In the degenerate case $L_s = U_s = 1$, define $\pi_\theta = \bar{w}$.

An exact draw from (203) *is obtained by proposing $A \sim \bar{w}(\cdot|s)$ and accepting with probability $r_\theta(s,A)/U_s$. The acceptance probability is exactly $1/U_s$, so the expected number of proposals is U_s. If N failures are followed by a fresh independent base draw, the returned law is*

$$\pi_{\theta,N} = (1 - \xi^{\mathrm{sh}}_{N,s})\pi_\theta + \xi^{\mathrm{sh}}_{N,s}\bar{w}, \qquad \xi^{\mathrm{sh}}_{N,s} = (1 - 1/U_s)^N, \tag{206}$$

and its ratio remains in $[L_s, U_s]$ because it is a convex combination of r_θ and 1.

Theorem 16 makes the ratio constraint architectural: it holds for every parameter value, even before training converges. The actor has the same support as $\bar{w}$, obeys $L_s \le d\pi_\theta/d\bar{w} \le U_s$, and can be sampled exactly by rejection sampling. Its acceptance rate is $1/U_s$, so the expected number of base-policy proposals is U_s. Truncating rejection sampling and then drawing a fresh base action produces the mixture in (206); because both component ratios lie in $[L_s, U_s]$, the fallback implementation remains certified.

This actor inherits the support, endpoint-based divergence, and trajectory-ratio consequences of the box. For Power-Mirror endpoints, $\max|r_\theta - 1| \le H_{p,\eta}$. However, an arbitrary learned gate inside the box need not be the exact gradient of the Power-Mirror backup. It therefore does not automatically inherit the ideal policy's sharper pairwise oscillation bounds, exact TV and χ^2 identities, coverage formula, or exact improvement gain. Those additional statements hold when the ideal density itself is represented or when actor-fitting error is controlled as below.

Exact conditional expectations under $\bar{w}$ may be unavailable, especially when $\bar{w}$ is an implicit behavior generator. The following alternative avoids calibration. It constrains an unnormalized positive weight between 1 and κ_s; after normalization, the policy ratio is guaranteed to lie in the symmetric interval $[1/\kappa_s, \kappa_s]$. Sampling does not require evaluating the normalizing constant.

Theorem 17 (Normalizer-free-for-sampling bounded gate)**.** *Let $\kappa_s \ge 1$, let $h_\theta(s,a) \in [0,1]$, and set*

$$\widetilde{g}_\theta(s,a) = 1 + (\kappa_s - 1)h_\theta(s,a), \qquad Z_\theta(s) = \mathbb{E}_{\bar{w}}\widetilde{g}_\theta(s,\cdot), \tag{207}$$

$$\pi^{\mathrm{g}}_\theta(da|s) = \frac{\widetilde{g}_\theta(s,a)}{Z_\theta(s)}\bar{w}(da|s). \tag{208}$$

For every θ,

$$1 \le Z_\theta(s) \le \kappa_s, \qquad \frac{1}{\kappa_s} \le \frac{d\pi^{\mathrm{g}}_\theta}{d\bar{w}}(a|s) \le \kappa_s. \tag{209}$$

If $\kappa_s \le \min\{U_s, L_s^{-1}\}$, this symmetric interval lies inside a desired sharp box $[L_s, U_s]$, possibly at the cost of target-approximation bias. Conversely, every normalized target $r_\star$ with $0 < m_s := \operatorname{ess\,inf} r_\star$ and

$$\frac{\operatorname{ess\,sup} r_\star}{\operatorname{ess\,inf} r_\star} \le \kappa_s \tag{210}$$

belongs to the unrestricted measurable gate family: take $\widetilde{g}_\star = r_\star/m_s$, or equivalently

$$h_\star = \frac{r_\star/m_s - 1}{\kappa_s - 1}$$

when $\kappa_s > 1$. A parametric class represents the target only if it contains this gate; for $\kappa_s = 1$, the only feasible target is $r_\star = 1$. Thus $\kappa_s = K_{p_s,\eta_s}$ removes the structural gate-range exclusion for every ideal power-mirror target, but its symmetric upper certificate K_{p_s,η_s} is looser than the ideal upper endpoint U_{p_s,η_s} and function-class error may remain.

Rejection sampling proposes $A \sim \bar{w}$ and accepts with probability $\widetilde{g}_\theta(s,A)/\kappa_s$. It returns exactly (208)*, has acceptance probability $Z_\theta(s)/\kappa_s$, and uses at most κ_s proposals in expectation without evaluating Z_θ. If sampling is truncated after N failures and followed by a* fresh independent *draw from $\bar{w}$, the returned law is*

$$\pi^{\mathrm{g}}_{\theta,N} = (1-\xi_{N,s})\pi^{\mathrm{g}}_\theta + \xi_{N,s}\bar{w}, \qquad \xi_{N,s} = \left(1 - \frac{Z_\theta(s)}{\kappa_s}\right)^N \le \left(1 - \frac{1}{\kappa_s}\right)^N, \tag{211}$$

and it retains (209)*. The last rejected proposal is not a valid substitute for the fresh fallback in this statement.*

For either bounded-attempt construction, write its returned distribution as $\pi_N = (1-\xi_{N,s})\pi + \xi_{N,s}\bar{w}$. For any action-value function $q(a) = Q(s,a)$, fallback reduces the base-relative one-step gain by exactly the probability of successful rejection sampling:

$$\mathbb{E}_{\pi_N} q - \mathbb{E}_{\bar{w}} q = (1-\xi_{N,s})\left(\mathbb{E}_\pi q - \mathbb{E}_{\bar{w}} q\right), \tag{212}$$

with the corresponding sharp or normalized-gate failure probability. Thus bounded latency preserves the sign of an exact one-step gain while reducing its magnitude. For $\epsilon_{\mathrm{lat}} \in (0,1)$, a calibrated actor reaches failure probability at most ϵ_{lat} once

$$N \ge \left\lceil \frac{\log \epsilon_{\mathrm{lat}}}{\log(1 - 1/U_s)} \right\rceil,$$

and the same worst-case rule uses κ_s for the normalized gate when $\kappa_s > 1$. If $\kappa_s = 1$, the actor equals the base and one proposal suffices.

The two actors address different implementation settings. The calibrated actor preserves the sharp asymmetric interval $[L_s, U_s]$ but requires the conditional base expectation in (200). The normalized gate can be sampled without computing its normalizer, but its simplest certificate is the looser symmetric interval $[1/\kappa_s, \kappa_s]$. Distilling either composite policy into a second unconstrained Gaussian or flow would discard the structural guarantee.

Theorem 18 (Actor-fitting inheritance)**.** *Fix a state and let $\pi_\star = r_\star\bar{w}$ be an ideal target and $\pi_\theta = r_\theta\bar{w}$ a ratio-preserving actor. Let $Q(s,a)$ be the true action value, let $\widehat{Q}(s,a)$ be its estimate, and define the critic error by $\Delta_Q(s,a) = \widehat{Q}(s,a) - Q(s,a)$. Assume that $\widehat{Q}(s,\cdot)$ has essential span at most D, with $\Delta_Q(s,\cdot) \in L_1(\bar{w})$, and define*

$$\widehat{G}_\star = \mathbb{E}_{\pi_\star}\widehat{Q}(s,\cdot) - \mathbb{E}_{\bar{w}}\widehat{Q}(s,\cdot), \qquad \varepsilon_{\mathrm{KL}} = \mathrm{KL}(\pi_\star\|\pi_\theta), \qquad \varepsilon_2 = \mathbb{E}_{\bar{w}}(r_\theta - r_\star)^2.$$

Then

$$\mathbb{E}_{\pi_\theta}\widehat{Q}(s,\cdot) - \mathbb{E}_{\bar{w}}\widehat{Q}(s,\cdot) \ge \widehat{G}_\star - D\sqrt{\varepsilon_{\mathrm{KL}}/2}, \tag{213}$$

$$\mathbb{E}_{\pi_\theta}\widehat{Q}(s,\cdot) - \mathbb{E}_{\bar{w}}\widehat{Q}(s,\cdot) \ge \widehat{G}_\star - \sqrt{\mathrm{Var}_{\bar{w}}(\widehat{Q}(s,\cdot))\varepsilon_2} \ge \widehat{G}_\star - \frac{D}{2}\sqrt{\varepsilon_2}. \tag{214}$$

If $B_\theta = \operatorname{ess\,sup}_a |r_\theta(a) - 1|$, the true-score gain satisfies

$$\mathbb{E}_{\pi_\theta} Q(s,\cdot) - \mathbb{E}_{\bar{w}} Q(s,\cdot) \ge \widehat{G}_\star - D\sqrt{\varepsilon_{\mathrm{KL}}/2} - B_\theta \mathbb{E}_{\bar{w}}|\Delta_Q(s,\cdot)|. \tag{215}$$

For the sharp-box actor, $B_\theta \le \max\{1-L, U-1\}$; for the normalized gate, $B_\theta \le \kappa - 1$. If both $r_\star, r_\theta \in [L,U]$, then

$$\frac{\varepsilon_2}{2U} \le \mathrm{KL}(\pi_\star\|\pi_\theta) \le \frac{\varepsilon_2}{2L}. \tag{216}$$

When the true action values are $Q^{\bar{w}}(s,\cdot)$, assume the displayed statewise quantities are bounded and measurable and define

$$d_\theta(s) = D_s\sqrt{\varepsilon_{\mathrm{KL},s}/2} + B_{\theta,s}\mathbb{E}_{\bar{w}_s}|\Delta_Q(s,\cdot)| - \widehat{G}_\star(s). \tag{217}$$

If $d_\theta \le 0$ at every state, then $V^{\pi_\theta} \ge V^{\bar{w}}$; in general,

$$\left\|[V^{\bar{w}} - V^{\pi_\theta}]_+\right\|_\infty \le \frac{\|[d_\theta]_+\|_\infty}{1-\gamma}. \tag{218}$$

The two losses in (213)–(214) have a direct interpretation. Actor fitting loses at most the Q-value range D times a distance between the learned and ideal actors. Equation (215) then adds the behavior-weighted Q-estimation error, multiplied by the maximum amount B_θ that the actor moves away from its base. The quantity $d_\theta(s)$ in (217) is the resulting statewise performance shortfall: if it is nonpositive at every state, ordinary policy improvement still holds; otherwise (218) bounds the possible value loss.

Accordingly, actor training should optimize the bounded ratio directly, for example through

$$\min_\theta \mathrm{KL}(\pi_\star\|\pi_\theta) \quad\Longleftrightarrow\quad \max_\theta \mathbb{E}_{\bar{w}}[r_\star \log r_\theta], \qquad \min_\theta \mathbb{E}_{\bar{w}}(r_\theta - r_\star)^2. \tag{219}$$

For (208), $\log r_\theta = \log \widetilde{g}_\theta - \log Z_\theta$; under standard dominated-differentiation conditions, the normalizer gradient can be estimated from exact rejection samples. Because the constraint is built into r_θ, every optimizer iterate remains ratio-safe. Optimization error affects predicted improvement, as quantified above, but cannot create an out-of-range density ratio.

Proposition 8 (Why unconstrained actor fitting is insufficient)**.** *Let the target and reference be $w = \mathcal{N}(0,1)$, the ideal power-mirror policy for a constant critic, and let $\pi_\delta = \mathcal{N}(\delta,1)$. Then both KL directions equal $\delta^2/2$ and*

$$\mathbb{E}_w\left(\frac{d\pi_\delta}{dw} - 1\right)^2 = e^{\delta^2} - 1,$$

so all three errors vanish as $\delta \to 0$. Nevertheless, for every $\delta \neq 0$,

$$\frac{d\pi_\delta}{dw}(a) = \exp(\delta a - \delta^2/2)$$

has essential infimum zero and essential supremum infinity. Average actor-fit accuracy therefore cannot replace a structural ratio layer.

The Gaussian example isolates the problem with ordinary weighted maximum likelihood. The shifted Gaussian can be arbitrarily close to the target in either KL direction and in behavior-weighted squared error, yet its likelihood ratio tends to zero in one tail and infinity in the other. Average actor accuracy therefore cannot certify a uniform offline-RL trust region.

Lemma 1 (Calibration and behavior-model uncertainty)**.** *Suppose an approximate calibrated gate $\widetilde{g} \in [0,1]$ obeys $|\mathbb{E}_{\bar{w}}\widetilde{g} - \tau| \le \zeta$. Let $\widetilde{r} = L + (U-L)\widetilde{g}$, normalize it by $\overline{r} = \widetilde{r}/\mathbb{E}_{\bar{w}}\widetilde{r}$, and put $d = (U-L)\zeta < 1$. Then*

$$\frac{L}{1+d} \le \overline{r} \le \frac{U}{1-d}. \tag{220}$$

If instead an exactly normalized actor is built relative to a behavior model $\bar{w}$ satisfying

$$0 < c_- \le \frac{d\bar{w}}{dw} \le c_+ < \infty, \tag{221}$$

then a model-relative interval $[L,U]$ transfers to the true-logger interval $[Lc_-, Uc_+]$; the normalized gate interval transfers to $[c_-/\kappa, c_+\kappa]$. A uniform log-density error at most ϵ_β gives $c_- = e^{-\epsilon_\beta}$ and $c_+ = e^{\epsilon_\beta}$. Average KL, TV, or held-out likelihood accuracy of $\bar{w}$ does not imply (221)*.*

Thus the guarantee always refers first to the actor's actual base distribution. To claim a bound relative to the logged data, either use the true executable logger as that base or certify the pointwise base-model error in (221). Average behavior-model KL, TV, or likelihood does not suffice. The normalizer-free gate can be sampled exactly even when Z_θ is unknown, but the PDIS shield still needs the numerical likelihood ratio $\widetilde{g}_\theta/Z_\theta$ at every certification action. A structural interval alone does not provide an unbiased importance weight.

K ALGORITHMIC INSTANTIATIONS

This section translates the population constructions into training procedures. These procedures produce candidates; finite data and function approximation do not automatically inherit the exact population monotonicity theorems. A strict deployment claim requires the independent shield in Section H.

K.1 FITTED PMVI FOR FINITE ACTIONS

Population PMVI assumes that the MDP is known. With only a transition dataset

$$\mathcal{D} = \{(s_j, a_j, r_j, s'_j)\}_{j=1}^{n},$$

we instead fit a Q function with Power-Mirror Bellman targets:

1. **Fix the reference policy.** Estimate the behavior policy once, $w_s = \widehat{\pi}_\beta(\cdot|s)$, and fix its supported action set. Unlike PMPI, do not replace this reference after a Bellman iteration.
2. **Choose the ratio budget.** Use the known reward range to obtain a valid bound on the range of $Q(s,\cdot)$. Choose fixed powers p_s and shifts b_s larger than this bound. Equations (53) and (54) translate a desired behavior-likelihood-ratio cap into p_s. With a neural Q function, enforce the assumed range with a bounded output head or an explicit projection; observing a small range only on the training actions is not a uniform certificate.
3. **Fit Power-Mirror Bellman targets.** Start from a bounded estimate $\widehat{Q}_0$. Let $\overline{Q}_k$ be a frozen target-network copy of $\widehat{Q}_k$ and form

$$y_j = r_j + \gamma C_{p_{s'_j}, w_{s'_j}, b_{s'_j}}\big(\overline{Q}_k(s'_j, \cdot)\big), \tag{222}$$

   Regress $\widehat{Q}_{k+1}(s_j, a_j)$ toward y_j, as in fitted Q iteration. Cross-fitting, pessimistic confidence corrections, or a tabular empirical model can be inserted at this step.
4. **Extract the policy.** At the selected stopping iterate K, compute the behavior-centered shifted Q values, their power mean, and the normalization correction:

$$z_i(s) = b_s + \widehat{Q}_K(s,i) - \mu_{w_s}(\widehat{Q}_K(s,\cdot)),$$
$$m_{p_s}(s) = \left(\sum_i w_{s,i} z_i(s)^{p_s}\right)^{1/p_s},$$
$$R_s = \sum_i w_{s,i} \left(\frac{z_i(s)}{m_{p_s}(s)}\right)^{p_s - 1},$$

   Then form the policy

$$\widehat{\pi}_K(i|s) = w_{s,i}\left[\left(\frac{z_i(s)}{m_{p_s}(s)}\right)^{p_s-1} + 1 - R_s\right].$$

5. **Audit the candidate.** Report the fitted Bellman residual, realized lower and upper policy ratios, additive ratio deviation, behavior-weighted Q-error estimate or confidence radius, and return. Compare the measured ratios with the theoretical (L, U, H) values.

With exact regression and population expectations, (222) has the PMVI fixed point. With finite data and function approximation, regression error can break the anytime population-safety chain (79). The fitted algorithm should therefore be described as a candidate generator unless the extra Q-error conditions in Theorem 13 or the independent trajectory shield are satisfied.

K.2 FITTED PMPI FOR FINITE ACTIONS

A fitted PMPI implementation follows the familiar approximate policy-iteration template: estimate the current policy's Q function, improve the policy, and repeat. Unlike fitted PMVI, the evaluation target changes whenever the policy changes.

1. **Initialize and split data.** Set $\sigma_0 = \widehat{\pi}_\beta$ on the estimated behavior support. Reserve complete certification trajectories before choosing hyperparameters or a stopping iterate.

2. **Evaluate the current policy.** At outer iteration k, estimate Q^{σ_k} using fitted policy evaluation, cross-fitting, or a pessimistic tabular model. Denote this estimate directly by $\widehat{Q}_k(s,a)$. Freeze both $\widehat{Q}_k$ and the selected $p_{k,s}, b_{k,s}$ before computing the policy update.

3. **Improve conservatively.** Apply (101) to $\widehat{Q}_k$, then use either a prespecified damping schedule or the logger-relative line search (121). Record both the one-step ratio σ_{k+1}/σ_k and the cumulative ratio $\sigma_{k+1}/\widehat{\pi}_\beta$.

4. **Repeat and stop.** Continue evaluation and improvement on training data. Choose the outer stopping iterate without unrestricted reuse of the certification fold.

5. **Certify before deployment.** Audit the final ratio against the *true* behavior policy and apply Section H. A ratio relative only to the fitted model $\widehat{\pi}_\beta$ is not a strict logger-relative certificate.

To state the effect of evaluation error clearly, define $\Delta_{Q,k}(s,a) = \widehat{Q}_k(s,a) - Q^{\sigma_k}(s,a)$. If $\widehat{\eta}_{k,s} = \operatorname{span}_a \widehat{Q}_k(s,a)/b_{k,s} < 1$, the true local gain of the full-step candidate obeys the analogue of Theorem 13,

$$\langle \widetilde{\sigma}_{k+1} - \sigma_k, Q^{\sigma_k}(s,\cdot)\rangle \geq \widehat{G}_k(s) - H_{p_{k,s},\widehat{\eta}_{k,s}} \mathbb{E}_{a\sim\sigma_k(\cdot|s)}|\Delta_{Q,k}(s,a)|, \tag{223}$$

where $\widehat{G}_k(s) = \langle \widetilde{\sigma}_{k+1} - \sigma_k, \widehat{Q}_k(s,\cdot)\rangle \geq 0$ is the gain predicted by the learned critic. Damping multiplies both the predicted gain and the error penalty by $\lambda_{k,s}$. A nonnegative lower bound at every state certifies a truly monotone population step. A negative lower bound is inconclusive rather than evidence of degradation. In ordinary offline data, the true Q error is unknown and later PMPI policies may be progressively harder to evaluate as their cumulative ratio grows. Fitted PMPI is therefore a candidate-generation procedure, not an automatic finite-data extension of Theorem 11.

For continuous PMPI, the base of the gate at iteration k is σ_k, so its architectural guarantee controls only σ_{k+1}/σ_k. Logger-relative ratios and rejection-sampling costs can both compound across nested actors. Damping or a direct logger-relative gate can limit this growth. Replacing the nested actor by an unconstrained Gaussian or flow does not preserve the ratio history unless the new projection is itself certified.

K.3 Continuous ratio-preserving Power-Mirror actor

The continuous algorithm deploys a composite policy: sample from a base policy and accept or reweight that action with a bounded gate. It does not replace the certified composite policy by a freely fitted density.

1. **Choose the base and split data.** Separate training from certification, then choose a samplable base $\bar{w}(\cdot|s)$. Use the true behavior policy whenever it remains executable. A fitted generator can be used to construct a candidate, but the resulting ratio guarantee is relative to that generator unless the pointwise generator-to-logger bound (221) is established.

2. **Fit a bounded Q function.** Train $\widehat{Q}$ with a globally bounded action head or another mechanism enforcing

$$\operatorname*{ess\,sup}_a \widehat{Q}(s,a) - \operatorname*{ess\,inf}_a \widehat{Q}(s,a) \leq D_s < b_s.$$

   A range measured only over a finite set of sampled actions is not a uniform certificate. Choose p_s and compute the theoretical constants (L_s, U_s, H_s, K_s) from $(p_s, D_s/b_s)$.

3. **Construct the ideal training ratios.** For a training state s, draw $A_1, \dots, A_M \overset{\text{iid}}{\sim} \bar{w}(\cdot|s)$ and compute

$$\begin{aligned} \widehat{\mu}_s &= \frac{1}{M}\sum_{m=1}^{M} \widehat{Q}(s, A_m), & \widehat{z}_m &= b_s + \widehat{Q}(s, A_m) - \widehat{\mu}_s, \\ \widehat{m}_{p,s} &= \left(\frac{1}{M}\sum_{m=1}^{M} \widehat{z}_m^{p_s}\right)^{1/p_s}, & \widehat{A}_m &= \left(\frac{\widehat{z}_m}{\widehat{m}_{p,s}}\right)^{p_s - 1}, \\ \widehat{R}_s &= \frac{1}{M}\sum_{m=1}^{M} \widehat{A}_m, & \widehat{r}_m &= \widehat{A}_m + 1 - \widehat{R}_s. \end{aligned} \tag{224}$$

   The empirical weights average to one exactly, and positivity gives $\widehat{R}_s \le 1$. They are supervised targets for fitting a continuous ratio function. They are not themselves a deployable continuous policy: selecting only from this finite action pool would create a discrete distribution that is singular relative to a continuous base.

4. **Choose the normalization mechanism.** If $\mu_{h_\theta}(s) = \mathbb{E}_{\bar{w}} h_\theta(s, \cdot)$ is exactly evaluable at every relevant state, use the calibrated actor (203) to retain the asymmetric interval $[L_s, U_s]$. If only a certified calibration error ζ_s is available, normalize the approximate gate by rejection sampling and use the enlarged interval

$$\left[\frac{L_s}{1 + (U_s - L_s)\zeta_s}, \frac{U_s}{1 - (U_s - L_s)\zeta_s}\right]$$

   from Lemma 1, provided $(U_s - L_s)\zeta_s < 1$ and the deployed likelihood can be evaluated. Otherwise use the normalized gate (208). For Power-Mirror endpoints, choosing $\kappa_s \le U_s$ gives a conservative symmetric interval inside $[L_s, U_s]$. Choosing $\kappa_s = K_s$ can represent the full ideal ratio range but uses the looser certificate $[1/K_s, K_s]$ and does not eliminate function-class error.

5. **Fit and audit the actor.** Estimate actor-target KL or L_2 error on an independent action batch. Q-estimation and Monte Carlo errors affect target quality and return, but the applicable model-relative ratio box holds at every optimization iterate. Empirical fit errors are diagnostics unless accompanied by confidence bounds. Do not clip and renormalize a free density after training: the unknown normalization can move the clipped ratio outside its intended interval.

6. **Deploy by rejection sampling.** Keep and execute the pair $(\bar{w}, r_\theta)$ or $(\bar{w}, \widetilde{g}_\theta)$. Equation (206) applies to exact calibration, and (211) applies to the normalized gate. For an approximately calibrated gate, let $d_s = (U_s - L_s)\zeta_s$ and $\widetilde{Z}_s = \mathbb{E}_{\bar{w}} \widetilde{r}_s \in [1 - d_s, 1 + d_s]$. Accept with probability $\widetilde{r}_s / U_s$; the accepted ratio is $\widetilde{r}_s / \widetilde{Z}_s$, the expected proposal count is at most $U_s/(1 - d_s)$, and

$$\widetilde{\pi}_{\theta,N} = (1 - \widetilde{\xi}_{N,s})\widetilde{\pi}_\theta + \widetilde{\xi}_{N,s}\bar{w}, \qquad \widetilde{\xi}_{N,s} = \left(1 - \frac{\widetilde{Z}_s}{U_s}\right)^N \le \left(1 - \frac{1 - d_s}{U_s}\right)^N.$$

   This mixture retains the enlarged interval from Step 4 and requires the actual $\widetilde{Z}_s$ for likelihood evaluation. After N rejected proposals, draw a fresh independent action from the base. Returning the last rejected proposal has a different distribution. Do not distill accepted actions into a second free Gaussian or flow.

7. **Report both utility and certification diagnostics.** Report theoretical and sampled ratios, actor-target KL and L_2 errors, ideal and retained gain, acceptance rate, proposal-count tails, fallback rate, calibration error, and any certified behavior-model error. Sampled ratio quantiles are useful diagnostics but are not proofs of a uniform bound.

For the normalized gate, minimizing forward KL from the ideal actor is equivalent, up to constants independent of θ, to maximizing

$$\max_\theta \mathbb{E}_s \left[\mathbb{E}_{A \sim \bar{w}(\cdot|s)} r_\star(s, A) \log \widetilde{g}_\theta(s, A) - \log Z_\theta(s)\right]. \tag{225}$$

Monte Carlo versions for the calibrated and normalized actors are

$$\widehat{\mathcal{J}}_{\mathrm{cal}}(\theta; s) = \frac{1}{M} \sum_{m=1}^{M} \widehat{r}_m \log r_\theta(s, A_m), \tag{226}$$

$$\widehat{\mathcal{J}}_{\mathrm{gate}}(\theta; s) = \frac{1}{M} \sum_{m=1}^{M} \widehat{r}_m \log \widetilde{g}_\theta(s, A_m) - \log \left[\frac{1}{M_Z} \sum_{\ell=1}^{M_Z} \widetilde{g}_\theta(s, \widetilde{A}_\ell) \right], \tag{227}$$

where $\widetilde{A}_1, \ldots, \widetilde{A}_{M_Z}$ are an independent batch from $\bar{w}(\cdot|s)$ used to estimate the normalizer. The first action pool may be used both to construct and fit the discrete target, although cross-fitting is possible. Reported actor-fit error must use a separate audit batch. Under standard differentiability conditions, the normalizer gradient can alternatively be written as

$$\nabla_\theta \log Z_\theta(s) = \mathbb{E}_{A \sim \pi_\theta^{\mathrm{g}}(\cdot|s)} \nabla_\theta \log \widetilde{g}_\theta(s, A)$$

with exact rejection samples.

For a fixed trained actor and J queried states, Hoeffding's inequality gives the simultaneous diagnostic bound

$$\left| \widehat{Z}_\theta(s) - Z_\theta(s) \right| \le (\kappa_s - 1) \sqrt{\frac{\log(2J/\delta_Z)}{2M_Z}} \quad \text{for all queried states} \tag{228}$$

with probability at least $1 - \delta_Z$, using independent proposal batches and intersecting the resulting interval with $[1, \kappa_s]$. This produces high-probability ratio intervals at those queried states, but it is not an exact likelihood evaluator.

A Monte Carlo normalizer cannot be inserted into PDIS as though it were exact. The strict shield therefore accepts the normalized-gate actor only when $Z_\theta(s)$ and the base-to-logger ratio can be evaluated exactly at every certification pair. Otherwise it falls back to w. A shield based on confidence intervals for Z_θ would need a separate interval-importance-weight analysis; plug-in and self-normalized weights are outside the present theorem.

K.4 Certified end-to-end deployment

The complete offline-RL pipeline is "train a candidate, audit its exact ratio, then accept or fall back." Specifically:

1. reserve complete behavior trajectories before fitting models or choosing hyperparameters;
2. train one candidate, or a finite prespecified candidate family, on the remaining data;
3. evaluate and bound the likelihood ratio of the policy that will actually be deployed relative to the true behavior policy, including its actor normalizer and any behavior-model correction; reject candidates that fail this audit;
4. compute a Hoeffding or empirical-Bernstein lower confidence bound with the infinite-horizon tail correction; and
5. deploy a candidate only if its lower bound clears the requested return margin; otherwise deploy behavior.

If the certification fold is used to compare candidates, powers, or stopping iterates, apply Corollary 10. The certification fold is not an ordinary validation set: unrestricted repeated tuning on it destroys the stated high-probability guarantee.

L Experimental settings

This appendix specifies the three experiments in Section 7. The main-text sweeps contain the results; no evaluation return is used to select a power. Across tasks, KL-fixed uses a behavior reference and KL-moving updates that reference between outer iterations. AWAC uses advantage-weighted regression, while IQL uses expectile value fitting followed by advantage-weighted policy extraction. PMVI retains a fixed behavior anchor; PMPI uses a moving policy reference. The categorical versions

share an empirical transition model, whereas the continuous versions use function approximation. These comparisons do not apply the held-out deployment test of Theorem 4. The fitted PMVI runs use shifts recomputed from current or learned critic spans; they therefore differ from the fixed-shift operator analyzed in Theorem 2. Their normalized continuous gate uses the symmetric model-relative interval of Theorem 17, rather than the ideal target's sharp asymmetric interval.

L.1 Tabular Wet Chicken

Environment and data. We use a continuing 5×5 Wet Chicken grid with five actions and discount $\gamma = 0.95$. Reward is the next-state downstream coordinate; crossing the waterfall causes a reset and incurs evaluation cost 1. The logger mixes a heuristic policy with uniform exploration. Figure 1 fixes the exploration probability at $\epsilon = 0.2$ and uses $N \in \{100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000\}$ transitions, with 100 independently generated datasets per size. All methods and PMVI powers share each dataset. Policies are fitted only to the empirical model; the true transition kernel is used for evaluation.

Model and updates. At visited states, improvement is restricted to observed actions; at unvisited states, the implementation falls back to the known logger. Zero-count backups are set to zero. PMVI uses a statewise span shift with $\eta = 0.5$ and powers $p = 2, 4, 6$; its $p = 4, 6$ runs reuse the same datasets and empirical models as $p = 2$. Fitted value iteration uses tolerance 10^{-10} and at most 5,000 iterations. The categorical AWAC and IQL implementations use 2,000 updates, with AWAC temperature 1 and IQL expectile 0.8 and temperature 3. KL-fixed and KL-moving use temperature 1, with 20 outer updates for KL-moving. PMPI-raw also uses 20 outer updates, without damping or a permanent behavior box. Greedy improvement uses the same observed-action support.

Evaluation and uncertainty. We solve for true-MDP discounted return J and discounted waterfall cost C under each learned policy. Cost is absent from the training reward. The behavior curve is the exact logging policy, not its empirical estimate. Curves average 100 independent datasets; shaded bands are pointwise percentile-bootstrap 95% intervals using 5,000 resamples. The independent unit is the complete offline dataset. Thus the plotted uncertainty reflects offline sampling and policy fitting, while evaluation itself has no Monte Carlo error.

L.2 NeoRL

Data and preprocessing. We use NeoRL v1 medium-quality Hopper-v3 and HalfCheetah-v3 with $N \in \{100, 1000, 10000\}$ official training trajectories (Qin et al., 2022). Official validation trajectories are separated into policy-validation and behavior-audit folds by even/odd trajectory index. Training observations are standardized using training-fold statistics, and actions are scaled for the bounded actor representation. Rewards retain their native scale. The behavior reference is a fitted five-component conditional tanh-Gaussian mixture with two hidden layers of width 256, trained for 50,000 updates and frozen before policy fitting. Within each dataset cell, the frozen behavior density is shared across methods and policy-training seeds. The behavior curve evaluates this same model under the evaluation seeds; it is not the unknown policy that originally collected the data.

Policy fitting. Policies use $\gamma = 0.99$, batch size 256, learning rate 3×10^{-4}, Polyak coefficient 0.005, and 50,000 policy-training updates. The span-critic architecture and warmup match the MetaDrive configuration below. Validation diagnostics audit the fixed configuration and do not select a replacement using online return. PMVI uses the bounded normalized gate, with $p \in \{2, 4, 6\}$ and state-dependent shift $b(s) = 2D_\phi(s)$. The span critic parameterizes each head as $Q_i(s, a) = V_i(s) + \frac{1}{2} D_\phi(s) \tanh A_i(s, a)$, so the action span is bounded by the learned $D_\phi(s)$. PMPI is reported at $p = 2$ and is trained as a moving-reference actor; it is not assigned PMVI's permanent behavior-ratio box. KL-fixed and KL-moving use temperature 1, with 20 outer updates for KL-moving. AWAC uses temperature 1; IQL uses expectile 0.7 and inverse temperature 3. These settings are fixed across displayed dataset sizes.

Evaluation and uncertainty. Each completed policy is evaluated over ten episodes capped at 1,000 steps. A plotted observation is the per-seed mean native episodic return; means and pointwise 95% Student-t intervals are then computed across training seeds. We use ten seeds (11–20) per setting.

The independence unit is the training seed, not each evaluation episode; the intervals do not quantify uncertainty over new independently collected training datasets. This distinction follows the emphasis on replication and uncertainty in few-run RL evaluation (Agarwal et al., 2021).

L.3 METADRIVE

Dataset and filtering. We use DSRL's `OfflineMetadrive-hardmean-v0` (Liu et al., 2024), with 259-dimensional observations, two bounded action dimensions, and a 1,000-step horizon. For the coverage sweep, all zero-cost episodes are retained together with a deterministic hash sample of positive-cost episodes at $q_{\mathrm{risk}} \in \{0.005, 0.05, 0.25, 1\}$. After filtering, a fixed episode hash assigns the retained episodes to training, validation, and behavior-audit folds in proportions $80/10/10$. The unfiltered condition is $q_{\mathrm{risk}} = 1$. Filtering affects both dataset composition and size, while evaluation scenarios remain fixed. Native reward is used without adding the safety cost to the training objective.

Networks and optimization. The behavior reference is a five-component conditional tanh-Gaussian mixture with two hidden layers of width 256. It is fitted for 50,000 updates and then frozen; all methods share its checkpoint within a filtering condition. Policy training uses 50,000 updates, discount 0.99, batch size 256, learning rate 3×10^{-4}, and Polyak coefficient 0.005. The PMVI span critic uses five heads with value and advantage hidden width 256, and a two-layer width-128 span network. Its initial span is $0.9D_{\max}$, where $D_{\max}$ is the training reward range divided by $1 - \gamma$; span learning has a 5,000-step warmup. The state-dependent shift is $b(s) = 2D_\phi(s)$. PMVI and PMPI use $p = 2$; PMVI's nominal model-relative ratio budget is $[1/2, 2]$. Comparator temperatures, the IQL expectile, and the KL-moving outer-update count match the NeoRL settings. We evaluate the policy after the prescribed update budget; online returns do not select checkpoints or powers.

Metrics and replication. Every method at every filtering level uses ten training seeds (11–20), evaluated on the same 100 driving scenarios per seed. Return and cost are undiscounted episode sums; success, collision, and off-road rates are fractions of evaluation episodes. Each metric is first averaged within a training seed and then across seeds. Error bars are pointwise 95% Student-t intervals over the ten seed means, without clipping their endpoints to the range of the metric. The common scenarios reduce differences caused by evaluation conditions but do not estimate generalization to an independent scenario population. All likelihood-ratio diagnostics are relative to the fitted behavior model, not to known logging propensities.

M POSITIONING AND LIMITATIONS

What is new versus Power-UCT. Power-UCT uses power means as tree-backup estimators (Dam et al., 2020). Here the reference weights are an arbitrary policy, especially the offline behavior policy, and the derivative of the power mean is interpreted as a behavior-ratio constrained improvement policy. The novelty is not the backup value; it is the induced policy geometry.

What is new versus the matrix smoother. The matrix smoother is a finite-action local quadratic theory for choosing curvature from statistical uncertainty. Power mirror is a global reference-density policy-improvement map. Its strength is that it gives direct behavior-ratio control and extends to continuous actions without a finite-action curvature matrix.

What is new versus policy mirror descent. Exact moving-reference improvement and monotone policy iteration are classical ideas, and PMPI belongs to that family. The contribution claimed here is narrower: the power map gives a closed-form two-part gain, sharp stepwise density-ratio and odds bounds, an explicit persistent-update condition with a matching stalling mechanism, and a cumulative logger-drift analysis that is specific to offline deployment. We do not import a generic mirror-descent convergence rate or claim that the fitted algorithm is monotone.

What remains open. The fixed-anchor PMVI algorithm and moving-anchor PMPI algorithm are complete at the exact population level, and Section H gives a strict finite-sample *deployment* guarantee under auditable logging conditions. The coverage quantities in Theorems 10 and 12 depend on exact $Q^\star$ or exact changing-policy critics; converting them into high-probability learning

guarantees requires simultaneous confidence control on comparator-covered actions. Action-ratio control cannot replace state occupancy coverage, and a finite-sample near-optimality theorem still needs suitable realizability and concentrability assumptions. Theorem 11 assumes exact evaluation; without a valid lower-bound safeguard, fitted evaluation and actor approximation can reverse a step. Its policy-to-policy ratios also compound relative to the logger, while enforcing a permanent logger box can stall and supplies no state-occupancy guarantee by itself. Section J.2 closes the structural continuous actor-projection gap relative to a known base measure, but it does not solve target-quality, approximate calibration, or behavior-model mismatch. A complete finite-sample *learning* theory must combine critic estimation error across changing-policy evaluations or Bellman iterations, learned behavior-policy error, actor-fit utility loss, and Monte Carlo error in continuous actions. The shield certifies non-degradation, not near-optimality, and it does not cover a single dependent transition stream, an unknown logger, or an unevaluable actor normalizer without additional statistical assumptions. These issues are nontrivial and should not be hidden.

N Conclusion

Power mirror is best understood as a reference-anchored control construction, not as a new smooth maximum. Its local contribution is the sharp envelope

$$(1 - D/b)^{p-1} \le \pi_p/w \le (1 + D/b)^{p-1},$$

together with sharp additive and log-ratio oscillation, explicit divergence control, and a critic-error multiplier that vanishes as $p \downarrow 1$. Globally, keeping behavior fixed in every backup yields a contractive PMVI target whose policies retain direct logger-relative control. Its unrestricted near-optimality decomposes into global $Q^\star$ action coverage and state-occupancy coverage rather than assuming either implicitly. Re-anchoring at the exactly evaluated current policy instead yields PMPI: ordinary values improve monotonically and persistent updates converge to support-constrained optimality, but logger-relative ratios accumulate. At finite iterations, an exact coverage representation of the Bellman residual and the realized improvement gain give a recursive near-optimality certificate, while a separate support defect accounts for unrestricted optimality. A behavior-box line search preserves monotonicity and a permanent ratio cap at the possible cost of stalling. In continuous actions, calibrated or normalized behavior gates preserve a selected ratio box for every actor iterate and can be sampled exactly by rejection; imperfect target fitting then appears as an explicit utility term instead of an untracked safety violation. Finally, the two-sided ratio theorem calibrates a held-out per-decision importance-sampling shield: under known logging propensities, an evaluable candidate ratio, and independent certification trajectories, the deployed policy is no worse than behavior with probability at least $1 - \delta$. The reported tabular and continuous-control experiments provide finite-data evidence for the return–coverage and return–cost tradeoffs described in Section L. They do not evaluate the independent deployment shield or establish a true-logger ratio certificate on public datasets. A full finite-sample near-optimality analysis of fitted PMVI or PMPI remains open.

O Proofs

The proofs follow the same sequence as the algorithmic story. We first show that the shifted power mean is a valid Bellman backup and that its gradient is a probability distribution. We then prove the sharp actionwise likelihood-ratio bounds, which are the main reusable technical tool. The population-control, policy-iteration, critic-error, deployment-shield, and continuous-action results are subsequently obtained by combining that local result with standard Bellman, resolvent, concentration, and change-of-measure arguments. Whenever the proof uses a compact vector such as q, it denotes the action-value vector $Q(s, \cdot)$ at one fixed state; this convention avoids carrying the state index through a statewise calculation.

O.1 Proof of Bellman-compatible scaffolding

Convexity follows because $q \mapsto \mu_w(q)$ is linear and $q \mapsto b\mathbf{1} + (I - \mathbf{1}w^\top)q$ is affine, while $\| \cdot \|_{p,w}$ is convex for $p \ge 1$ on the positive orthant and extends as a norm for general signed vectors when $p \ge 1$. Translation equivariance follows from

$$\mu_w(q + c\mathbf{1}) = \mu_w(q) + c, \qquad (I - \mathbf{1}w^\top)(q + c\mathbf{1}) = (I - \mathbf{1}w^\top)q.$$

The endpoint $p = 1$ is

$$C_{1,w,b}(q) = \mu_w(q) + \sum_i w_i(b + q_i - \mu_w(q)) - b = \mu_w(q),$$

and $p = \infty$ gives

$$C_{\infty,w,b}(q) = \mu_w(q) + \max_i(b + q_i - \mu_w(q)) - b = \max_i q_i.$$

Monotonicity in p follows from monotonicity of $L_p(w)$ norms on a probability space. Monotonicity and nonexpansiveness in q follow from the gradient formula in Theorem 6: the gradient lies in the simplex, so

$$|C(q) - C(q')| \le \|q - q'\|_\infty.$$

O.2 Proof of the induced policy formula

Let $m_p = (\sum_i w_i z_i^p)^{1/p}$. Then

$$\frac{\partial m_p}{\partial z_j} = w_j z_j^{p-1} m_p^{1-p}.$$

Since $z_j = q_j - \sum_k w_k q_k + b$,

$$\frac{\partial z_j}{\partial q_i} = \mathbf{1}\{i = j\} - w_i.$$

Therefore

$$\begin{aligned}\frac{\partial C_{p,w,b}}{\partial q_i} &= w_i + \sum_j w_j z_j^{p-1} m_p^{1-p}(\mathbf{1}\{i = j\} - w_i) \\ &= w_i + w_i z_i^{p-1} m_p^{1-p} - w_i \sum_j w_j z_j^{p-1} m_p^{1-p} \\ &= w_i \left[\left(\frac{z_i}{m_p}\right)^{p-1} + 1 - R\right].\end{aligned}$$

The entries sum to one because

$$\sum_i w_i(a_i + 1 - R) = R + 1 - R = 1.$$

Moreover

$$R = m_p^{1-p} \sum_i w_i z_i^{p-1} \le m_p^{1-p} m_p^{p-1} = 1,$$

using $\|z\|_{p-1,w} \le \|z\|_{p,w}$. Thus the gradient is nonnegative and sums to one.

O.3 Proof of the sharp density-ratio theorem and its corollaries

Let $\alpha = p - 1$, normalize $X_i = z_i/b$, and define

$$A = \mathbb{E}_w X^\alpha, \qquad B = (\mathbb{E}_w X^{\alpha+1})^{\alpha/(\alpha+1)}.$$

Then $\mathbb{E}_w X = 1$ and $\operatorname{span}(X) \le \eta$. If $\eta = 0$, then $X = 1$ w-almost surely, $r = 1$, and every claim is immediate; henceforth assume $\eta > 0$. Power-mean monotonicity gives both $B \ge 1$ and $A^{1/\alpha} \le B^{1/\alpha}$, hence $A \le B$. Therefore

$$r_i := \frac{\pi_i}{w_i} = 1 + \frac{X_i^\alpha - A}{B} = \frac{X_i^\alpha}{B} + c, \qquad c = 1 - \frac{A}{B} \ge 0. \tag{229}$$

Also $\mathbb{E}_w r = 1$. Write $\ell = \min_i X_i$, $u = \max_i X_i$, and $h = u - \ell \le \eta$. Since $\ell \le 1 \le u$,

$$\max_i r_i - \min_i r_i = \frac{u^\alpha - \ell^\alpha}{B} \le H_{p,\eta}. \tag{230}$$

Indeed, for $0 < \alpha \le 1$, increments of x^α over an interval of fixed length decrease with the left endpoint. The interval contains 1, so

$$u^\alpha - \ell^\alpha \le 1 - (1 - h)^\alpha \le 1 - (1 - \eta)^\alpha.$$

For $\alpha \geq 1$, the increments increase with the left endpoint, giving

$$u^\alpha - \ell^\alpha \leq (1+h)^\alpha - 1 \leq (1+\eta)^\alpha - 1.$$

Because $\mathbb{E}_w r = 1$ lies between the minimum and maximum ratio, (230) also yields $\max_i |r_i - 1| \leq H_{p,\eta}$.

For the multiplicative oscillation, the common nonnegative correction c in (229) gives

$$\frac{\max_i r_i}{\min_i r_i} \leq \frac{u^\alpha}{\ell^\alpha}. \tag{231}$$

The mean-one and span constraints imply $\ell \geq 1 - \eta$ and $u \leq \ell + \eta$, hence

$$\frac{u}{\ell} \leq 1 + \frac{\eta}{\ell} \leq \frac{1}{1-\eta}.$$

Thus (231) is at most $K_{p,\eta}$. Since $\max_i r_i \geq 1$, it follows that

$$\min_i r_i \geq K_{p,\eta}^{-1} = L_{p,\eta}.$$

It remains to prove the sharp upper endpoint. If $\alpha \geq 1$, Jensen's inequality gives $A \geq 1$, so

$$\max_i r_i = 1 + \frac{u^\alpha - A}{B} \leq 1 + (u^\alpha - 1) \leq (1+\eta)^\alpha = U_{p,\eta}.$$

Now let $0 < \alpha < 1$ and put $d = u - 1 \in [0, \eta]$. Concavity of x^α on the containing interval $[u - \eta, u]$ gives the chord bound

$$A \geq \frac{d}{\eta}(u-\eta)^\alpha + \left(1 - \frac{d}{\eta}\right) u^\alpha.$$

Consequently,

$$\max_i r_i - 1 \leq F(d) := \frac{d}{\eta}\left[(1+d)^\alpha - (1+d-\eta)^\alpha\right].$$

The function F is nondecreasing on $[0, \eta]$. To see this, set $v = 1 + d - \eta > 0$ and $y = (1+d)/v \geq 1$. Since $d \leq v$ and $y^{\alpha-1} - 1 \leq 0$,

$$\begin{aligned}\eta F'(d) &= (1+d)^\alpha - v^\alpha + d\alpha\big((1+d)^{\alpha-1} - v^{\alpha-1}\big) \\ &\geq v^\alpha \left[y^\alpha - 1 + \alpha(y^{\alpha-1} - 1)\right] \geq 0.\end{aligned}$$

The last bracket vanishes at $y = 1$ and has derivative $\alpha y^{\alpha-2}(y + \alpha - 1) > 0$. Therefore

$$F(d) \leq F(\eta) = (1+\eta)^\alpha - 1,$$

which proves the upper bound for all $\alpha > 0$. Taking logarithms in the multiplicative oscillation proves the log-ratio statement.

We finally verify sharpness. Normalize $b = 1$ and fix two actions with weights $(\epsilon, 1-\epsilon)$. For the lower endpoint, take

$$z_1 = 1 - (1-\epsilon)\eta, \qquad z_2 = 1 + \epsilon\eta.$$

These shifted scores have mean 1 and span η, while as $\epsilon \downarrow 0$ the first action ratio converges to $(1-\eta)^\alpha$ and the second to 1. This family makes the lower endpoint, $K_{p,\eta}$, and the log oscillation sharp; it also makes $H_{p,\eta}$ sharp when $\alpha \leq 1$. For the upper endpoint, take

$$z_1 = 1 + (1-\epsilon)\eta, \qquad z_2 = 1 - \epsilon\eta.$$

Setting $q_i = z_i - 1$ realizes both families in the original shifted-score construction with $\mu_w(q) = 0$. The first action ratio in the second family converges to $(1+\eta)^\alpha$ and the second to 1, proving upper sharpness and sharpness of $H_{p,\eta}$ when $\alpha \geq 1$. Thus each claimed constant is optimal uniformly, although the limiting endpoint need not be attained for a fixed strictly positive reference policy.

For Corollary 1, the case $\eta = 0$ follows from $X = 1$ w-almost surely. For $\eta > 0$, convexity places $f(x)$ below the chord joining $(L, f(L))$ and $(U, f(U))$ for $x \in [L, U]$. Averaging that chord and using $\mathbb{E}_w r = 1$ gives (46). The forward and reverse D_∞ bounds are immediate from $L \leq r \leq U$. Substituting $f(x) = |x-1|/2$, $f(x) = (x-1)^2$, $f(x) = x \log x$, and $f(x) = -\log x$ yields the

endpoint-based TV, chi-square, and KL formulas. Equation (229) gives the exact identities (49). Finally, Popoviciu's inequality and Cauchy–Schwarz give

$$\mathrm{Var}_w(r) \le \frac{H^2}{4}, \qquad \mathrm{TV}(\pi_p, w) = \frac{1}{2}\mathbb{E}_w|r-1| \le \frac{1}{2}\sqrt{\mathrm{Var}_w(r)} \le \frac{H}{4},$$

which proves the remaining terms in (48).

For Corollary 2, factor the likelihood ratio of the state–action trajectory τ_T. The common initial-state law and transition factors cancel, leaving

$$\frac{d\mathbb{P}_\pi^{(T)}}{d\mathbb{P}_w^{(T)}} = \prod_{t=0}^{T-1} \frac{\pi(a_t|s_t)}{w(a_t|s_t)}.$$

Applying the statewise lower and upper bounds to every factor proves (55).

O.4 Proof of the near-optimal coverage bound

Let $M = \max_i z_i$. For $i \in G_\varepsilon(q)$,

$$z_i \ge M - \varepsilon.$$

First, $C(q) \ge \min_i q_i$ gives

$$m(q) - C(q) \le D.$$

Also $C(q) \ge \mu_w(q)$, and splitting the behavior expectation of $m(q) - q_i$ over $G_\varepsilon(q)$ and its complement gives

$$m(q) - C(q) \le m(q) - \mu_w(q) \le \varepsilon W_\varepsilon + D(1 - W_\varepsilon).$$

If $\varepsilon \ge M$, then

$$M - m_p \le M \le M(1 - W_\varepsilon^{1/p}) + \varepsilon W_\varepsilon^{1/p},$$

so the power-specific term is valid. It remains to consider $0 \le \varepsilon < M$, for which $M - \varepsilon > 0$. Then

$$m_p = \left(\sum_i w_i z_i^p\right)^{1/p} \ge W_\varepsilon^{1/p}(M - \varepsilon).$$

Using $m(q) - C_{p,w,b}(q) = M - m_p$,

$$\begin{aligned} M - m_p &\le M - W_\varepsilon^{1/p}(M - \varepsilon) \\ &= M(1 - W_\varepsilon^{1/p}) + \varepsilon W_\varepsilon^{1/p} \end{aligned}$$

Taking the minimum of the three valid upper bounds proves (58). Solving the last bound for p and intersecting it with the sharp upper-ratio rule proves Corollary 3.

O.5 Proof of the population convergence theorem

For $V \in \mathcal{V}$, the reward and value bounds give

$$V_{\min} = r_{\min} + \gamma V_{\min} \le Q_V(s,a) \le r_{\max} + \gamma V_{\max} = V_{\max}.$$

Hence $\mathrm{span}_a(Q_V(s,\cdot)) \le \overline{D} < b_s$, so every shifted coordinate is strictly positive. Writing $z_a = b_s + Q_V(s,a) - \mathbb{E}_{w_s} Q_V(s,\cdot)$, we have $\mathbb{E}_{w_s} z = b_s$. Monotonicity of weighted power means and the elementary bound of an L_p mean by its maximum give

$$\mathbb{E}_{w_s} Q_V(s,\cdot) \le C_{p_s,w_s,b_s}(Q_V(s,\cdot)) \le \max_{a \in \mathcal{A}_w(s)} Q_V(s,a).$$

This proves (69). It also shows that $\mathcal{T}_C V \in \mathcal{V}$.

For $U, V \in \mathcal{V}$, the segment joining $Q_U(s,\cdot)$ and $Q_V(s,\cdot)$ remains in the convex fixed-span domain. The gradient of C_{p_s,w_s,b_s} lies in the simplex throughout this segment. It follows both that C_{p_s,w_s,b_s} is coordinatewise monotone and that it is ℓ_∞ nonexpansive there. Consequently $\mathcal{T}_C$ is monotone, and

$$\begin{aligned} |(\mathcal{T}_C V)(s) - (\mathcal{T}_C U)(s)| &\le \|Q_V(s,\cdot) - Q_U(s,\cdot)\|_\infty \\ &\le \gamma \|V - U\|_\infty. \end{aligned}$$

Banach's fixed-point theorem yields the unique V_C and (70).

If $V_0 = V^w$, then

$$V_1 = \mathcal{T}_C V^w \geq \mathcal{T}^w V^w = V^w.$$

Monotonicity of $\mathcal{T}_C$ gives $V_{k+1} \geq V_k$ inductively. The fixed-point comparison $V^w \leq V_C$, or equivalently monotonicity applied to $V_0 \leq V_C$, gives $V_k \leq V_C$. This proves (71).

O.6 Proof of end-to-end population control

The fixed-point comparison induced by (69) gives

$$V^w \leq V_C \leq V_w^\star.$$

Because $Q_C(s, \cdot)$ lies in the strict fixed-span domain, Theorem 7 shows that π_C is a probability distribution supported on w and satisfies the stated density-ratio envelope.

For each state, convexity and $C_{p_s, w_s, b_s}(0) = 0$ imply

$$0 \geq C_{p_s, w_s, b_s}(Q_C(s, \cdot)) - \langle \pi_C(\cdot|s), Q_C(s, \cdot) \rangle .$$

Thus

$$\mathcal{T}^{\pi_C} V_C \geq \mathcal{T}_C V_C = V_C.$$

Iterating the monotone contraction $\mathcal{T}^{\pi_C}$ yields $V^{\pi_C} \geq V_C$. Since π_C is supported on w, $V^{\pi_C} \leq V_w^\star$, proving (74).

The definitions in (75)–(76) can be written as

$$g_C = V_C - \mathcal{T}^w V_C, \qquad \zeta_C = \mathcal{T}^{\pi_C} V_C - V_C.$$

The Bellman sandwich gives $g_C \geq 0$, while the preceding convexity argument gives $\zeta_C \geq 0$. Subtracting the appropriate Bellman equations yields

$$\begin{aligned} (I - \gamma P^w)(V_C - V^w) &= g_C, \\ (I - \gamma P^{\pi_C})(V^{\pi_C} - V_C) &= \zeta_C. \end{aligned}$$

The inverses exist and are entrywise nonnegative because

$$(I - \gamma P)^{-1} = \sum_{t \geq 0} \gamma^t P^t.$$

This proves (77)–(78).

Finally, when $V_0 = V^w$, convexity applied to Q_{V_k} gives

$$\mathcal{T}^{\pi_k} V_k \geq \mathcal{T}_C V_k = V_{k+1} \geq V_k.$$

Iteration of $\mathcal{T}^{\pi_k}$ from V_k therefore gives

$$V^{\pi_k} \geq \mathcal{T}^{\pi_k} V_k \geq V_{k+1} \geq V_k \geq V^w,$$

which establishes (79).

O.7 Proof of coverage-dependent support optimality

Let $\pi_w^\star$ be support optimal. Since

$$V_w^\star = \mathcal{T}^{\pi_w^\star} V_w^\star, \qquad V_C = \mathcal{T}_C V_C,$$

we have

$$\begin{aligned} (I - \gamma P^{\pi_w^\star})(V_w^\star - V_C) &= \mathcal{T}^{\pi_w^\star} V_C - V_C \\ &\leq \mathcal{T}_w^\star V_C - \mathcal{T}_C V_C = \delta_C. \end{aligned}$$

Multiplication by the nonnegative resolvent gives

$$V_w^\star - V_C \leq (I - \gamma P^{\pi_w^\star})^{-1} \delta_C.$$

Theorem 9 gives $V^{\pi_C} \geq V_C$, while support optimality gives $V_w^\star \geq V^{\pi_C}$. Multiplying by ν and using

$$\nu^\top (I - \gamma P^\pi)^{-1} f = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\nu^\pi}[f(s)]$$

proves (85).

The trivial inequalities $C(Q_C) \geq \min_a Q_C(s,a)$ and $C(Q_C) \leq \max_a Q_C(s,a)$ give $0 \leq \delta_C(s) \leq D_C(s)$. Applying Proposition 3 to $Q_C(s,\cdot)$ on the restricted behavior support gives the remaining terms in (86). Taking their minimum completes the proof.

For the finite stopping statement, Theorem 9 gives $V^{\pi_k} \geq V_{k+1}$, so

$$V_w^\star - V^{\pi_k} \leq (V_w^\star - V_C) + (V_C - V_{k+1}).$$

The first term is controlled by (84). The contraction rate and the value box give

$$\|V_C - V_{k+1}\|_\infty \leq \gamma^{k+1} \|V_C - V^w\|_\infty \leq \gamma^{k+1} \overline{D}.$$

The standard a posteriori contraction estimate also gives

$$\|V_C - V_{k+1}\|_\infty \leq \frac{\gamma}{1-\gamma} \|V_{k+1} - V_k\|_\infty.$$

Multiplying by ν proves (87).

O.8 Proof of coverage-dependent unrestricted near-optimality

For $V_t = V_C + t(V^\star - V_C)$, define

$$\pi_t(\cdot|s) = \nabla C_{p_s, w_s, b_s}(Q_{V_t}(s,\cdot)), \qquad \pi^{\mathrm{mv}}(\cdot|s) = \int_0^1 \pi_t(\cdot|s)\, dt.$$

Every V_t lies in $\mathcal{V}$. Hence each π_t has the same support as w and satisfies the uniform sharp ratio interval in (73); averaging preserves support, normalization, and both pointwise endpoints. The fundamental theorem of calculus gives

$$\begin{aligned}(\mathcal{T}_C V^\star)(s) - (\mathcal{T}_C V_C)(s) &= \int_0^1 \langle \pi_t(\cdot|s), Q_{V^\star}(s,\cdot) - Q_{V_C}(s,\cdot) \rangle\, dt \\ &= \gamma (P^{\pi^{\mathrm{mv}}}(V^\star - V_C))(s).\end{aligned}$$

Since $V_C = \mathcal{T}_C V_C$, rearrangement yields the exact identity

$$(I - \gamma P^{\pi^{\mathrm{mv}}})(V^\star - V_C) = \Delta_\star^C.$$

The defect is nonnegative because $\mathcal{T}_C V^\star \leq \mathcal{T}^\star V^\star = V^\star$. The resolvent is nonnegative, Theorem 9 gives $V^{\pi_C} \geq V_C$, and unrestricted optimality gives $V^\star \geq V^{\pi_C}$. This proves the value ordering and exact resolvent statement in (95).

Fix a state and abbreviate $D = D_\star^w(s)$, $X = X_\star^w(s)$, $W = W_{\star,\varepsilon_s}^w(s)$, and $m_p = (\mathbb{E}_{w_s} z^{p_s})^{1/p_s}$ for the shifted $Q^\star$ scores. Then

$$\Delta_\star^C(s) = X - m_p.$$

The Bellman sandwich gives

$$\Delta_\star^C(s) \leq V^\star(s) - \mathbb{E}_{w_s} Q^\star(s,\cdot) \leq \varepsilon_s W + D(1-W),$$

and $\mathcal{T}_C V^\star \geq \min_{a \in \mathcal{A}_w(s)} Q^\star(s,a)$ gives $\Delta_\star^C(s) \leq D$. Every action counted by W has shifted score at least $X - \varepsilon_s$. If $\varepsilon_s < X$, then

$$m_p \geq W^{1/p_s}(X - \varepsilon_s),$$

so

$$\Delta_\star^C(s) \leq X(1 - W^{1/p_s}) + \varepsilon_s W^{1/p_s}.$$

If $\varepsilon_s \geq X$, the same inequality follows from $\Delta_\star^C(s) \leq X$ because its right-hand side is a convex combination of X and ε_s. Taking the minimum proves (93). The discounted-occupancy identity, the resolvent norm bound, and Hölder's inequality prove (96)–(99).

Finally, for $V_0 = V^w$, Theorem 9 gives $V^{\pi_k} \geq V_{k+1}$, and hence

$$V^\star - V^{\pi_k} \leq (V^\star - V_C) + (V_C - V_{k+1}).$$

Apply the global resolvent bound to the first term and the same a priori and a posteriori contraction estimates used in (87) to the second. This proves (98).

O.9 Proof of monotone moving-anchor policy iteration

Fix an iteration and state, and abbreviate

$$\sigma = \sigma_k(\cdot|s), \qquad q = Q_k(s,\cdot), \qquad p = p_s, \qquad b = b_s.$$

Exact policy evaluation gives $\langle \sigma, q\rangle = V^{\sigma_k}(s)$ and hence $\mathbb{E}_\sigma z = b$. Theorem 6 gives

$$\frac{\widetilde{\sigma}_{k+1}(a|s)}{\sigma_k(a|s)} = \left(\frac{z(a)}{m}\right)^{p-1} + 1 - R.$$

The strict span domain and Theorem 7 prove support invariance and (105). Since the damped ratio is one plus $\lambda_{k,s}$ times the raw ratio minus one, (106) follows immediately.

Put $A(a) = (z(a)/m)^{p-1}$. Because $q = V^{\sigma_k}(s) + z - b$, $\mathbb{E}_\sigma(A - R) = 0$, and

$$\mathbb{E}_\sigma[Az] = \frac{\mathbb{E}_\sigma z^p}{m^{p-1}} = m,$$

we obtain

$$\langle \widetilde{\sigma}_{k+1} - \sigma_k, q\rangle = \mathbb{E}_\sigma[(A-R)(z-b)] = m - bR.$$

Power-mean monotonicity gives $m \geq \mathbb{E}_\sigma z = b$, while Theorem 6 gives $R \leq 1$. This proves (109); strict power-mean monotonicity shows that its first term is positive exactly when q is nonconstant on the full current support. Linearity of (102) proves (110).

Subtracting the two policy Bellman equations yields

$$\begin{aligned}(I - \gamma P^{\sigma_{k+1}})(V^{\sigma_{k+1}} - V^{\sigma_k}) &= \mathcal{T}^{\sigma_{k+1}} V^{\sigma_k} - V^{\sigma_k} \\ &= G_k.\end{aligned}$$

The resolvent is entrywise nonnegative, proving (111). Every iterate is supported on $\mathcal{A}_w$, so support optimality gives the upper bound in (112).

We next prove the persistent-update convergence statement, including the adaptive extension. For this odds calculation, let $p = p_{k,s}$, $b = b_{k,s}$, $\alpha = \alpha_{k,s}$, and $\eta = \eta_{k,s}$ at iteration k. Let $r_a = \sigma_{k+1}(a|s)/\sigma_k(a|s)$ denote the damped ratio, and suppose $q(a) \geq q(c)$. From the update formula,

$$r_a - r_c = \lambda_{k,s}\frac{z(a)^\alpha - z(c)^\alpha}{m^\alpha}.$$

All shifted scores lie in $[b(1-\eta), b(1+\eta)]$. The mean-value theorem and $m \leq b(1+\eta)$ therefore give

$$r_a - r_c \geq \frac{\lambda_{k,s}\alpha\vartheta_{k,s}}{b(1+\eta)^\alpha}[q(a) - q(c)].$$

Moreover $r_a \leq U_{p,\eta} = (1+\eta)^\alpha$. Since $\log x - \log y \geq (x-y)/x$ for $x \geq y > 0$,

$$\log\frac{\sigma_{k+1}(a|s)}{\sigma_{k+1}(c|s)} - \log\frac{\sigma_k(a|s)}{\sigma_k(c|s)} \geq \xi_{k,s}[Q_k(s,a) - Q_k(s,c)]. \tag{232}$$

By (112), V^{σ_k} increases to some $\overline{V} \leq V_w^\star$, and hence $Q_k(s,\cdot) \to \overline{q}_s = Q_{\overline{V}}(s,\cdot)$. Take any subsequence along which σ_k converges to $\overline{\sigma}$; compactness of the finite policy simplex guarantees one. If $\overline{\sigma}(c|s) > 0$ while $\overline{q}_s(a) > \overline{q}_s(c)$ for some a, then the action-value gap is bounded below by a positive constant for all sufficiently large k. Summing (232) and using $\sum_k \xi_{k,s} = \infty$ forces $\sigma_k(a|s)/\sigma_k(c|s) \to \infty$, contradicting positive limiting mass on c. Thus every accumulation policy is supported on the maximizers of $\overline{q}_s$ at every state.

Policy evaluation is continuous on the finite policy simplex, so $\overline{V} = V^{\overline{\sigma}}$. The preceding greedy-support property gives

$$\overline{V} = \mathcal{T}^{\overline{\sigma}}\overline{V} = \mathcal{T}_w^\star \overline{V}.$$

Uniqueness of the support-constrained optimal Bellman fixed point yields $\overline{V} = V_w^\star$. Every accumulation policy is therefore support optimal; uniqueness of that policy implies convergence of the policy sequence. With fixed p_s, b_s and $\inf_{k,s}\lambda_{k,s} > 0$, the quantities in (115) have a positive statewise lower bound, so the required series diverges. More generally, if $\eta_{k,s} \leq \overline{\eta} < 1$ and $\alpha_{k,s}$ is bounded

above, the factor multiplying $\lambda_{k,s}\alpha_{k,s}/b_{k,s}$ in (115) is bounded below by a positive constant. This proves the stated adaptive sufficient condition.

Multiplying the stepwise inequalities (106) proves (117); taking logarithms proves (119). The behavior-box line search keeps every statewise mixture inside the convex box (120). Its gain remains $\lambda_{k,s}\widetilde{G}_k(s) \geq 0$, so the same resolvent proof gives monotonicity, including the case $\lambda_{k,s} = 0$.

For the stopping certificate, support optimality and the Bellman equations imply

$$\begin{aligned}(I - \gamma P^{\pi^\star_w})(V^\star_w - V^{\sigma_k}) &= \mathcal{T}^{\pi^\star_w} V^{\sigma_k} - V^{\sigma_k} \\ &\leq \mathcal{T}^\star_w V^{\sigma_k} - V^{\sigma_k} = \mathrm{Res}_k\,.\end{aligned}$$

Multiplication by the nonnegative resolvent proves (124). Because every finite iterate is positive on the original support, $\widetilde{G}_k \equiv 0$ holds exactly when every $Q_k(s,\cdot)$ is constant there, which is equivalent to $\mathrm{Res}_k \equiv 0$. No quantitative converse holds for a merely small gain.

Finally, monotone improvement alone does not ensure optimal convergence. In the one-state two-action bandit with rewards 1 and 0, write x_k for the good-action probability. For $p = 2$ and fixed $b > 1$, direct substitution gives

$$x_{k+1} = x_k + \lambda_k \frac{x_k(1-x_k)}{\sqrt{b^2 + x_k(1-x_k)}}.$$

Hence every $\lambda_k > 0$ strictly improves value, but

$$1 - x_{k+1} \geq (1 - x_k)\left(1 - \frac{\lambda_k}{b}\right).$$

For $\lambda_k = 2^{-k-1}$ the infinite product on the right is positive, so x_k converges to a value strictly below one. The other two stalling mechanisms follow from the sharp multiplicative ratio bound. With raw optimal-action odds $O_k = x_k/(1-x_k)$ and gap-one shifts $b_k > 1$,

$$0 \leq \log O_{k+1} - \log O_k \leq \log K_{p_k, 1/b_k} = -(p_k - 1)\log(1 - 1/b_k).$$

Additional damping can only decrease this odds increment. Hence fixed $b > 1$ with, for example, $p_k - 1 = 2^{-k-1}$ leaves $O_\infty < \infty$; so does $p_k = 2$ with $b_k = 2^{k+2}$, because $\sum_k -\log(1 - 1/b_k) < \infty$. In both cases the good-action probability remains strictly below one and the monotone limit is suboptimal. This proves the claimed stalling warning.

O.10 Proof of finite-iterate PMPI coverage near-optimality

Fix k, s and abbreviate $g(a) = g_{k,s}(a)$ and $D = D_{k,s}$. Exact evaluation gives $\mathbb{E}_{a\sim\sigma_k(\cdot|s)} Q_k(s,a) = V^{\sigma_k}(s)$, so the layer-cake formula for the nonnegative random variable $g \in [0, D]$ yields

$$\begin{aligned}\mathrm{Res}_k(s) &= \mathbb{E}_{a\sim\sigma_k}[g(a)] \\ &= \int_0^D \Pr_{a\sim\sigma_k}(g(a) > t)\, dt = \int_0^D (1 - W^\sigma_{k,t}(s))\, dt.\end{aligned}$$

Splitting according to $g \leq \varepsilon$ and $g > \varepsilon$ gives

$$\mathrm{Res}_k(s) \leq \varepsilon W^\sigma_{k,\varepsilon}(s) + D(1 - W^\sigma_{k,\varepsilon}(s)) = B^\sigma_{k,\varepsilon}(s).$$

Proposition 3, with reference $\sigma_k(\cdot|s)$, score $Q_k(s,\cdot)$, and $M = b_s + \mathrm{Res}_k(s)$, gives $\delta^{\mathrm{PM}}_k(s) \leq \overline{\delta}^{\mathrm{PM}}_{k,\varepsilon}(s)$.

Because $\mathcal{T}^{\sigma_{k+1}} V^{\sigma_k} = V^{\sigma_k} + G_k$ and $\mathcal{T}^\star_w V^{\sigma_k} = V^{\sigma_k} + \mathrm{Res}_k$, equation (133) follows. For the raw policy, the gain decomposition gives

$$\mathbb{E}_{\widetilde{\sigma}_{k+1}} Q_k(s,\cdot) - C_{p_s,\sigma_k,b_s}(Q_k(s,\cdot)) = b_s(1 - R_{k,s}).$$

Therefore its approximate-greedy error is

$$\max_{a\in\mathcal{A}_w(s)} Q_k(s,a) - \mathbb{E}_{\widetilde{\sigma}_{k+1}} Q_k(s,a) = \delta^{\mathrm{PM}}_k(s) - b_s(1 - R_{k,s}) \geq 0.$$

Linearity of the damped mixture proves (134); substitution of the coverage bounds proves (135)–(136).

Write $V_k = V^{\sigma_k}$ and $V_{k+1} = V^{\sigma_{k+1}}$. Adding and subtracting Bellman terms gives

$$\begin{aligned} V_w^\star - V_{k+1} &= \mathcal{T}^{\pi_w^\star} V_w^\star - \mathcal{T}^{\sigma_{k+1}} V_{k+1} \\ &= \gamma P^{\pi_w^\star}(V_w^\star - V_k) + [\mathcal{T}^{\pi_w^\star} V_k - \mathcal{T}^{\sigma_{k+1}} V_k] + \gamma P^{\sigma_{k+1}}(V_k - V_{k+1}). \end{aligned}$$

The middle term is at most $\mathrm{GreedyErr}_k$, and the last is nonpositive by PMPI monotonicity. This proves (137). Induction and $0 \le V_w^\star - V^w \le \overline{D}\mathbf{1}$ prove (138)–(139).

For Corollary 7, the definition of $\rho_{k,s}^-$ gives $\sigma_k(a|s) \ge \rho_{k,s}^- w(a|s)$ on every supported action. Summing over $\mathcal{G}_{k,t}(s)$ proves the coverage transfer, and Corollary 5 proves the product lower bound. All three terms defining the coverage envelope are nonincreasing in the transferred mass on $[0,1]$, because $0 \le \varepsilon \le D_{k,s} < b_s + \mathrm{Res}_k(s)$. Hence the conservative logger envelopes follow.

The performance-difference identity gives

$$\begin{aligned} J_\nu(\pi_w^\star) - J_\nu(\sigma_k) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\nu^{\pi_w^\star}}[\langle \pi_w^\star, Q_k(s,\cdot)\rangle - V^{\sigma_k}(s)] \\ &\le \frac{1}{1-\gamma}\mathbb{E}_{d_\nu^{\pi_w^\star}}[\mathrm{Res}_k] \le \frac{1}{1-\gamma}\mathbb{E}_{d_\nu^{\pi_w^\star}}[B_{k,\varepsilon_s}^\sigma]. \end{aligned}$$

Hölder's inequality proves (146); combining (124) with (131) proves the uniform stopping rule. Finally, subtract

$$J_\nu(\sigma_{k+1}) - J_\nu(\sigma_k) = \frac{1}{1-\gamma}\mathbb{E}_{d_\nu^{\sigma_{k+1}}}[G_k]$$

from the preceding bound to obtain (148).

For the unrestricted statement, the supported maximizer policy satisfies

$$\mathcal{T}^{\overline{\pi}_w} V^\star = V^\star - \chi_w.$$

Subtracting its Bellman equation gives

$$V^\star - V^{\overline{\pi}_w} = (I - \gamma P^{\overline{\pi}_w})^{-1}\chi_w.$$

Since $V_w^\star \ge V^{\overline{\pi}_w}$, this proves (150). Adding it to (138) proves (151)–(152). If $W_{\star,\varepsilon_s}^w(s) > 0$, some supported action has $Q^\star(s,a) \ge V^\star(s) - \varepsilon_s$, so $\chi_w(s) \le \varepsilon_s$.

O.11 Proof of exact reference-policy improvement

Since $C_{p,w,b}$ is convex and $C_{p,w,b}(0) = 0$, for $\pi = \nabla C_{p,w,b}(q)$,

$$0 = C(0) \ge C(q) + \langle \pi, 0 - q\rangle,$$

so

$$\langle \pi, q\rangle \ge C(q).$$

Since $C_{p,w,b}(q) \ge C_{1,w,b}(q) = \langle w, q\rangle$, we have

$$\langle \pi, q\rangle \ge \langle w, q\rangle.$$

Applying this statewise to $Q^{\pi_\beta}(s,\cdot)$ gives

$$\mathcal{T}^{\pi_p} V^{\pi_\beta} \ge \mathcal{T}^{\pi_\beta} V^{\pi_\beta} = V^{\pi_\beta}.$$

The standard policy-improvement theorem yields $V^{\pi_p} \ge V^{\pi_\beta}$.

O.12 Proof of the behavior-ratio critic-error bound

By exact improvement applied to $\widehat{Q}(s,\cdot)$,

$$\widehat{\Delta}_s = \langle \widehat{\pi}_p - w_s, \widehat{Q}(s,\cdot)\rangle \ge 0.$$

Let $r_s(a) = \widehat{\pi}_p(a|s)/w_s(a)$. Since $\widehat{Q}(s,a) = Q^{\pi_\beta}(s,a) + \Delta_Q(s,a)$,

$$\langle \widehat{\pi}_p - w_s, Q^{\pi_\beta}(s,\cdot)\rangle = \widehat{\Delta}_s - \mathbb{E}_{w_s}[(r_s-1)\Delta_Q(s,\cdot)].$$

Theorem 7 gives $\|r_s - 1\|_\infty \le H_s$, and therefore

$$|\mathbb{E}_{w_s}[(r_s-1)\Delta_Q(s,\cdot)]| \le H_s \mathbb{E}_{w_s}|\Delta_Q(s,\cdot)|.$$

This proves (157). If $\epsilon = \max_s H_s \mathbb{E}_{w_s}|\Delta_Q(s,\cdot)|$, then

$$\mathcal{T}^{\widehat{\pi}_p} V^{\pi_\beta} \ge V^{\pi_\beta} - \epsilon \mathbf{1}.$$

Since $\mathcal{T}^{\widehat{\pi}_p}$ is a γ-contraction and monotone,

$$V^{\widehat{\pi}_p} \ge V^{\pi_\beta} - \frac{\epsilon}{1-\gamma}\mathbf{1}.$$

Under (159), the true advantage is nonnegative at every state, so $\mathcal{T}^{\widehat{\pi}_p} V^{\pi_\beta} \ge V^{\pi_\beta}$ and ordinary policy improvement yields $V^{\widehat{\pi}_p} \ge V^{\pi_\beta}$.

For the L_2 statement, Cauchy–Schwarz and Corollary 1 give

$$|\mathbb{E}_{w_s}[(r_s-1)\Delta_Q(s,\cdot)]| \le \sqrt{\mathbb{E}_{w_s}(r_s-1)^2}\,\|\Delta_Q(s,\cdot)\|_{2,w_s} \le \frac{H_s}{2}\|\Delta_Q(s,\cdot)\|_{2,w_s}.$$

Finally, $r_s \le U_s$ directly gives

$$\mathbb{E}_{\widehat{\pi}_p}|\Delta_Q(s,\cdot)| = \mathbb{E}_{w_s}[r_s|\Delta_Q(s,\cdot)|] \le U_s \mathbb{E}_{w_s}|\Delta_Q(s,\cdot)|,$$

which proves (161).

O.13 Proof of strict high-probability deployment safety

Condition on the complete training fold. The candidate π and the constants L, U are then fixed, while the certification trajectories remain i.i.d. under w. For each t, the ratio certificate gives

$$L^{t+1} \le W_{j,t} \le U^{t+1}. \tag{233}$$

Because $|\widetilde{Y}_{j,t}| \le R_\Delta/2$,

$$-\frac{R_\Delta}{2} M_t(L,U) \le (W_{j,t}-1)\widetilde{Y}_{j,t} \le \frac{R_\Delta}{2} M_t(L,U).$$

Multiplying by γ^t and summing proves (168), whose interval width is (167).

For the expectation, the joint likelihood ratio of $(S_0, A_0, \dots, S_t, A_t, Y_t)$ under π relative to w is $W_{j,t}$: the common initial-state law, transition kernel, and conditional reward law cancel. Therefore

$$\mathbb{E}_w[W_{j,t}\widetilde{Y}_{j,t}] = \mathbb{E}_\pi[\widetilde{Y}_t].$$

Summing over $t = 0, \dots, H-1$ yields (165). One-sided Hoeffding concentration for the i.i.d. variables $Z_j(\pi)$ now gives, with probability at least $1-\delta$,

$$J_{\nu,H}(\pi) - J_{\nu,H}(w) \ge \widehat{\Delta}_H(\pi) - \mathcal{B}_H(L,U)\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{234}$$

The centered discounted tail of either policy lies in $[-\mathcal{T}_H/2, \mathcal{T}_H/2]$, so their tail difference is at least $-\mathcal{T}_H$. Consequently,

$$J_\nu(\pi) - J_\nu(w) \ge J_{\nu,H}(\pi) - J_{\nu,H}(w) - \mathcal{T}_H.$$

Combining this inequality with (234) proves (171). On the same event, candidate deployment implies that its true improvement is at least m; fallback deploys w itself. This proves (172)–(173). If the structural ratio audit fails, the algorithm falls back before examining the estimator, so safety is deterministic. For a finite-horizon objective there is no tail term.

For the acceptance statement, the opposite one-sided deviation gives

$$\widehat{\Delta}_H(\pi) \ge \Delta_H - \mathcal{B}_H(L,U)\sqrt{\frac{\log(1/\delta)}{2n}}$$

with probability at least $1-\delta$. On this event, the Hoeffding lower bound is at least $\Delta_H - 2\mathcal{B}_H\sqrt{\log(1/\delta)/(2n)} - \mathcal{T}_H$. Condition (174) therefore implies deployment, and rearranging it gives (175).

For Corollary 9, affinely rescale Z_j from the interval in (168) to $[0,1]$ and apply the one-sided empirical-Bernstein inequality of Maurer & Pontil (2009). Rescaling back multiplies the linear term by $\mathcal{B}_H$ and transforms the normalized sample variance into $\widehat{v}_Z/\mathcal{B}_H^2$, giving exactly (176). The case $\mathcal{B}_H = 0$ is deterministic and follows directly. Corollary 10 follows by applying the selected one-policy confidence bound at level δ_k and taking a union bound over the fixed candidate family.

O.14 Proof of the quadratic dual

Let $\delta = q - \mu_w(q)\mathbf{1}$ and consider the zero-mean tangent space $\{\delta : \sum_i w_i\delta_i = 0\}$. For $p = 2$,

$$C_{2,w,b}(q) = \mu_w(q) + f(\delta), \qquad f(\delta) = \sqrt{b^2 + \|\delta\|_{2,w}^2} - b.$$

The dual norm of $\|\delta\|_{2,w}$ on the tangent space is

$$\|u\|_{*,w} = \left(\sum_i \frac{u_i^2}{w_i}\right)^{1/2}.$$

A radial conjugacy calculation gives

$$f^*(u) = b\left(1 - \sqrt{1 - \|u\|_{*,w}^2}\right), \qquad \|u\|_{*,w} \le 1,$$

and $+\infty$ otherwise. Substituting $u = \pi - w$ yields (182). Indeed, the unconstrained tangent-space optimizer has

$$\pi_i = w_i\left(1 + \frac{q_i - \mu_w(q)}{\sqrt{b^2 + \mathrm{Var}_w(q)}}\right).$$

The condition $\mathrm{span}(q) < b$ implies $q_i - \mu_w(q) > -b$, while the denominator is at least b, so this optimizer is nonnegative and therefore lies in the simplex. This feasibility step is why the stated simplex dual is restricted to the strict fixed-span domain; without it, the tangent-space conjugate need not equal the simplex maximum.

O.15 Continuous-action proofs

The continuous derivative follows from the same calculation with sums replaced by integrals. Let $m_p = (\int z^p\,dw)^{1/p}$ and perturb Q by th. Then

$$\frac{d}{dt}C(Q+th)\Big|_{t=0} = \int h\,dw + \int z^{p-1}m_p^{1-p}(h - \int h\,dw)\,dw.$$

Thus the derivative is integration against the density

$$\frac{d\pi_p}{dw}(a) = \left(\frac{z(a)}{m_p}\right)^{p-1} + 1 - R.$$

The proof of $R \le 1$ uses $L_{p-1}(w) \le L_p(w)$ exactly as in the finite case. For the sharp envelope, apply the preceding proof to $X = z/b$ with essential infimum and supremum in place of the finite minimum and maximum; the mean, span, concave-chord, oscillation, and likelihood-ratio arguments are unchanged. Uniform sharpness over continuous-action models follows by taking an atomless behavior measure, partitioning the action space into two measurable regions of w-masses ϵ and $1-\epsilon$, and assigning the two piecewise-constant shifted scores from the finite sharpness construction. The strictly positive lower density bound gives mutual absolute continuity, and the divergence proof carries over by integration. The coverage argument is likewise identical with $w(G_\varepsilon)$ replacing the finite sum.

O.16 Proofs for ratio-preserving continuous actors

Proof of Theorem 16. Suppress the state index. If $\mu_h > \tau$, then $0 < \tau/\mu_h < 1$, so the first branch of (200) lies in $[0,1]$ and has mean τ. If $\mu_h < \tau$, then $0 < (1-\tau)/(1-\mu_h) < 1$, so the third branch also lies in $[0,1]$, and

$$\mathbb{E}_{\bar{w}}\mathfrak{C}_\tau[h] = 1 - \frac{1-\tau}{1-\mu_h}(1-\mu_h) = \tau.$$

The middle branch is immediate. Therefore

$$\mathbb{E}_{\bar{w}} r_\theta = L + (U-L)\tau = 1,$$

and $g_\theta \in [0,1]$ gives $L \le r_\theta \le U$. Since $L > 0$, the two measures are mutually absolutely continuous. For a feasible target, (205) lies in $[0,1]$ and normalization gives

$$\mathbb{E}_{\bar{w}} g_\star = \frac{1-L}{U-L} = \tau,$$

so calibration leaves it fixed. Finally, the joint measure of a proposal and acceptance in the rejection sampler is

$$\bar{w}(da)\frac{r_\theta(a)}{U}.$$

Its mass is $1/U$, and conditioning on acceptance gives $r_\theta \bar{w} = \pi_\theta$. The geometric proposal count therefore has mean U. All N proposals fail with probability $(1-1/U)^N$; conditioning on failure and success gives (206). Its ratio is $(1-\xi_N^{\mathrm{sh}})r_\theta + \xi_N^{\mathrm{sh}}$, which remains in $[L,U]$.

Proof of Theorem 17. The gate bound $1 \le \widetilde{g}_\theta \le \kappa$ implies $1 \le Z_\theta \le \kappa$, and division gives (209). The inner-box claim follows from $1/\kappa \ge L$ and $\kappa \le U$. For representability, let $m = \operatorname{ess\,inf} r_\star > 0$ and take $\widetilde{g}_\star = r_\star/m$. Then $1 \le \widetilde{g}_\star \le \kappa$ and $Z_\star = 1/m$, hence $\widetilde{g}_\star/Z_\star = r_\star$.

The accepted-proposal measure is $\bar{w}(da)\widetilde{g}_\theta(a)/\kappa$, of total mass Z_θ/κ; conditioning gives (208) and the mean proposal count is $\kappa/Z_\theta \le \kappa$. All N trials fail with probability $\xi_N = (1 - Z_\theta/\kappa)^N$. Conditional on at least one acceptance, the first accepted action has law π_θ^{g}; conditional on failure, the stipulated fresh draw has law $\bar{w}$. This proves (211). Its density ratio is a convex combination of 1 and a number in $[1/\kappa, \kappa]$, so it stays in the same interval.

Proof of Theorem 18. Pinsker's inequality and the variational characterization of total variation give

$$\left|\mathbb{E}_{\pi_\theta}\widehat{Q}(s,\cdot) - \mathbb{E}_{\pi_\star}\widehat{Q}(s,\cdot)\right| \le D\ \mathrm{TV}(\pi_\theta, \pi_\star) \le D\sqrt{\varepsilon_{\mathrm{KL}}/2},$$

which proves (213). Since both ratios integrate to one,

$$\begin{aligned}\left|\mathbb{E}_{\pi_\theta}\widehat{Q}(s,\cdot) - \mathbb{E}_{\pi_\star}\widehat{Q}(s,\cdot)\right| &= \left|\mathbb{E}_{\bar{w}}(r_\theta - r_\star)(\widehat{Q}(s,\cdot) - \mathbb{E}_{\bar{w}}\widehat{Q}(s,\cdot))\right| \\ &\le \sqrt{\varepsilon_2 \operatorname{Var}_{\bar{w}}(\widehat{Q}(s,\cdot))} \le \frac{D}{2}\sqrt{\varepsilon_2},\end{aligned}$$

where the last inequality is Popoviciu's variance bound. This proves (214). Moreover,

$$\begin{aligned}\mathbb{E}_{\pi_\theta}Q(s,\cdot) - \mathbb{E}_{\bar{w}}Q(s,\cdot) &= \mathbb{E}_{\pi_\theta}\widehat{Q}(s,\cdot) - \mathbb{E}_{\bar{w}}\widehat{Q}(s,\cdot) - \mathbb{E}_{\bar{w}}(r_\theta - 1)\Delta_Q(s,\cdot) \\ &\ge \widehat{G}_\star - D\sqrt{\varepsilon_{\mathrm{KL}}/2} - B_\theta \mathbb{E}_{\bar{w}}|\Delta_Q(s,\cdot)|,\end{aligned}$$

which is (215).

For (216), let $\phi(x) = x\log x$. On $[L,U]$, $1/U \le \phi''(x) \le 1/L$. Taylor's theorem therefore bounds the pointwise Bregman divergence by

$$\frac{(x-y)^2}{2U} \le \phi(x) - \phi(y) - \phi'(y)(x-y) \le \frac{(x-y)^2}{2L}.$$

Set $x = r_\star$, $y = r_\theta$, integrate, and use $\mathbb{E}_{\bar{w}} x = \mathbb{E}_{\bar{w}} y = 1$; the integrated Bregman divergence equals $\mathrm{KL}(\pi_\star \| \pi_\theta)$.

Finally, when $q_s = Q^{\bar{w}}(s,\cdot)$, the policy-difference identity gives

$$V^{\pi_\theta} - V^{\bar{w}} = (I - \gamma P^{\pi_\theta})^{-1}\left(\mathbb{E}_{\pi_\theta} q_s - \mathbb{E}_{\bar{w}_s} q_s\right).$$

The local bound is at least $-[d_\theta]_+$, and the nonnegative Neumann resolvent has ℓ_∞ norm at most $(1-\gamma)^{-1}$. This proves (218).

Proof of Proposition 8. For equal-variance Gaussians, each KL direction is $\delta^2/2$. Their likelihood ratio is

$$r_\delta(a) = \exp(\delta a - \delta^2/2).$$

The Gaussian moment-generating function gives $\mathbb{E}_w r_\delta = 1$ and $\mathbb{E}_w r_\delta^2 = e^{\delta^2}$, proving the displayed $L_2(w)$ identity. If $\delta \neq 0$, $r_\delta(a)$ tends to zero in one tail and infinity in the other, proving the essential endpoint claim.

Proof of Lemma 1. The approximate calibration assumption gives

$$1 - d \leq \mathbb{E}_{\bar{w}} \widetilde{r} \leq 1 + d.$$

Divide the pointwise bounds $L \leq \widetilde{r} \leq U$ by these normalizer endpoints to obtain (220). For behavior-model transfer, the chain rule for Radon–Nikodym derivatives gives

$$\frac{d\pi}{dw} = \frac{d\pi}{d\bar{w}} \frac{d\bar{w}}{dw};$$

multiplying the corresponding endpoint bounds proves both stated intervals.