DIFFERENTIATING THROUGH MOVING RECOURSE: FEASIBLE POLICY OPTIMIZATION AND FINITE-SAMPLE CERTIFICATION FOR MULTISTAGE STOCHASTIC PROGRAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multistage stochastic programs model decisions under uncertainty where earlier decisions change later feasible sets. We propose a feasible pathwise gradient method (FeasPG) that can update the policy as new sample paths arrive. The policy proposes a decision and projects it onto the current feasible set, so every decision is feasible. We derive the full trajectory gradient, accounting for how earlier decisions change later constraint bounds. Under regularity conditions, it is unbiased at almost every parameter value. We also bound the error from treating the constraint bounds as fixed during differentiation. We analyze a proposal-distance penalty that can provide a training signal where projection makes cost locally flat. We give a decreasing-weight schedule that preserves the $O(K^{-1/2})$ stationarity rate for the expected cost over K training updates under standard stochastic-gradient assumptions. To assess the learned policy, we use held-out trajectories to construct a finite-sample upper confidence bound on its optimality gap. The certificates are also anytime-valid across checkpoints, so the bounds can guide when training stops. In reservoir-management experiments, FeasPG obtains lower mean costs than stochastic dual dynamic programming and reinforcement-learning baselines on model-generated paths. A separate reservoir study demonstrates informative certificates with no coverage failures in 200 repetitions per setting.

1 INTRODUCTION

Multistage stochastic programs (MSPs) model sequential decisions in applications such as power generation, reservoir operation, and inventory management. These decisions must satisfy operational constraints while accounting for uncertain future conditions. Effective solution methods must therefore find low-cost policies that remain feasible as uncertainty unfolds. In MSPs, the stages are coupled by *moving recourse constraints*, i.e., earlier decisions change an endogenous state, which determines the feasible sets of later stages. In hydrothermal scheduling, the classical application, releasing water satisfies demand cheaply but lowers the storage, which limits how much water the decision-maker can release at later stages.

Stochastic dual dynamic programming (SDDP) by Pereira & Pinto (1991) is a leading method for MSPs (Füllner & Rebennack, 2025). It approximates each stage's expected value function from below by affine functions, known as cuts, and uses these approximations in the stage problems that determine decisions. The standard formulation assumes convex stage problems and stagewise independent uncertainties. Extensions handle nonconvex stage problems (Zou et al., 2019; Philpott et al., 2020) or Markovian data (Silva et al., 2021; Park et al., 2022). When new data change an estimated uncertainty distribution, existing cuts may need to be updated. Sequential-sampling methods address this issue under stagewise independence, which can be violated in practice (Gangammanavar & Sen, 2021; Xu & Sen, 2023). We pursue an alternative: directly updating a parameterized feasible policy as new sample paths arrive via policy gradient updates.

Policy gradient methods are widely used in reinforcement learning to improve policies from sampled trajectories (Sutton et al., 2000; Schulman et al., 2017). They offer a natural way to make these

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.

policy updates in MSPs. The transition and cost functions are known, and uncertainty does not depend on decisions, so the same sample path can be replayed under different policies. Under suitable differentiability conditions, we can differentiate the replayed cost to update the policy (Glasserman, 1991). Costs and dynamics need not be convex, and uncertainty may be dependent across stages.

In the MSPs we study, each decision must satisfy linear constraints whose bounds depend on the current state. Our policy therefore maps the state to a candidate decision, the *proposal*, and projects it onto the current feasible set. Every decision is then feasible during training and at deployment, but the projection introduces two difficulties. First, earlier decisions change the state and hence the projection's constraint bounds. For a reservoir with current storage s, a proposed release $u > s$ is projected to $x = s$, so $\partial x/\partial u = 0$ while $\partial x/\partial s = 1$. Treating the bounds as fixed incorrectly sets the latter derivative to zero. Second, projection can map a range of proposals to the same decision, making cost locally flat in those directions (Lin et al., 2021). To help training in such regions, we add the squared distance between each proposal and its projection as a regularizer. To assess when a policy is sufficiently accurate, we construct optimality-gap certificates that can guide stopping across checkpoints.

Contributions.

1. We propose a feasible pathwise gradient (FeasPG) method for MSPs, which trains projected policies with the complete pathwise gradient, and establish the gradient's almost-everywhere unbiasedness under regularity conditions.
2. We give a sufficient condition under which the proposal-distance penalty provides a descent direction for the regularized training objective of a parameterized policy. We also show that a suitable decreasing-weight schedule retains the $O(K^{-1/2})$ stationarity rate for the original expected cost.
3. We construct a finite-sample upper confidence bound on the selected policy's optimality gap from held-out trajectories. The sample-size exponent in our error bound depends on the dimension of the features used to predict uncertainty, not on the number of state and decision variables. With fresh data at each checkpoint, the bounds are anytime-valid and can guide when to stop training.

Related work. Projection has been used to make MSP decision rules feasible (Guigues & Henrion, 2017), and differentiation through optimization layers provides their gradients (Amos & Kolter, 2017; Agrawal et al., 2019). PROF also penalizes the squared distance between a proposal and its projection (Chen et al., 2021). In MSP, Ghadimi et al. (2019) optimize parameterized policies through endogenous states, while Bertsimas & Villalobos Carballo (2023) learn decision rules based on uncertainty histories. While completing our experiments, we became aware of the concurrent work of Helm et al. (2026), who differentiate through projections for inventory policies and recover integer decisions. Our analysis bounds how errors from ignoring moving constraint bounds affect later stages. We also give a sufficient condition for descent of the regularized training objective for parameterized policies and analyze decreasing regularization weights. For certification, information relaxation provides the lower-bound construction (Brown et al., 2010; Balseiro & Brown, 2019). We add a finite-sample correction for the conditional means estimated from trajectories. The correction covers every state, decision, and history combination available to the perfect-information optimizer, not only those visited by the learned policy. Related work by Glanzer et al. (2025) uses conditional-mean bands for optimal stopping. Our construction instead certifies the optimality gap of a selected feasible MSP policy. Appendix A.4 gives the detailed comparisons.

2 Problem Setup and Method Overview

An MSP seeks sequential decisions that minimize expected total cost over T stages while satisfying the constraints of every stage. At stage t, we choose a decision x_t using the current state s_t and the uncertainty $\xi_1, \ldots, \xi_t$ observed so far. The state s_t depends on earlier decisions, whereas the distribution of the exogenous uncertainty is unaffected by them.

Let $h_t = (\xi_1, \ldots, \xi_t)$ be the history observed by the beginning of stage t. After x_t is chosen, ξ_{t+1} is revealed and the stage cost $c_t(s_t, x_t, \xi_{t+1})$ is incurred. The dynamics of the endogenous state s_t and the constraints on the decision x_t are known. Starting from $s_1 = s_1^0(\xi_1)$, the state evolves as

$$s_{t+1} = f_t(s_t, x_t, \xi_{t+1}). \tag{1}$$

To express the constraints, let $z_t = \varphi_t(h_t) \in \mathbb{R}^{d_t}$ be a fixed summary of the observed history; taking $z_t = h_t$ retains the full history. We require the decision x_t to satisfy the following linear constraints:

$$x_t \in \mathcal{X}_t(s_t, z_t), \qquad \mathcal{X}_t(s, z) = \{x : A_t x \le b_t(z) + B_t s\}. \tag{2}$$

Here, $B_t s$ captures moving recourse: earlier decisions change the endogenous state via (1), which affects the feasible set (2) at later stages. The maps c_t, f_t, b_t and fixed matrices A_t, B_t are known. Costs and dynamics may be nonconvex. We assume relatively complete recourse: for almost every history, the feasible set is nonempty at every state reachable by a feasible decision (Shapiro, 2011).

A policy $\pi = (\pi_1, \dots, \pi_T)$ is a sequence of maps $\pi_t : (s_t, h_t) \mapsto x_t \in \mathcal{X}_t(s_t, z_t)$; it is *nonanticipative*, i.e., x_t uses the observation (s_t, h_t) available up to stage t. Let Π be the class of such policies with integrable total cost. The MSP minimizes the expected total cost over Π:

$$V^\star = \inf_{\pi \in \Pi} J(\pi), \qquad J(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} c_t(s_t^\pi, x_t^\pi, \xi_{t+1})\right]. \tag{3}$$

Here, the superscript π denotes the states and decisions generated by policy π. We assume $V^\star$ is finite.

Rather than optimize over all policies in Π, we restrict to policies π_θ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$. At stage t, a proposal map $\mu_{\theta,t}$ outputs a candidate decision u_t, the *proposal*, which is projected onto the current feasible set:

$$u_t = \mu_{\theta,t}(s_t, z_t), \qquad x_t = \bar{P}_t(u_t, s_t; z_t) := \arg\min_{x \in \mathcal{X}_t(s_t, z_t)} \tfrac{1}{2}\|x - u_t\|^2. \tag{4}$$

Thus $\pi_{\theta,t}(s_t, h_t) = \bar{P}_t(\mu_{\theta,t}(s_t, z_t), s_t; z_t)$, and we call π_θ the deployed policy. The initial-state map s_1^0 and history summary φ_t do not depend on θ. Under relatively complete recourse, exact projection gives feasible, nonanticipative decisions at every training iterate (Proposition A.1). Training seeks to reduce their expected cost.

The data consist of independent, identically distributed *sample paths* $\xi = (\xi_1, \dots, \xi_{T+1})$ from an unknown distribution. Observations within each sample path may be dependent across stages. For ξ and a decision sequence $x = (x_1, \dots, x_T)$, let $C(x, \xi) = \sum_t c_t(s_t, x_t, \xi_{t+1})$ denote its realized cost, with s_t given by (1). For π_θ, we abbreviate x^{π_θ} as x^θ and let $C(\theta, \xi) = C(x^\theta, \xi)$ and $J(\theta) = J(\pi_\theta)$. Because uncertainty is exogenous and the costs and dynamics are known, we can replay the same sample path under different policies, recomputing their states, decisions, and costs $C(\theta, \xi)$.

To separate policy assessment from training, we split the sample paths into three independent sets: $\mathcal{D}_{\text{fit}}$ for training and selecting the policy (Section 3), $\mathcal{D}_{\text{cal}}$ for calibration, and $\mathcal{D}_{\text{ev}}$ for evaluation (Section 4). The full model and technical conventions are given in Appendices A.1–A.2.

3 Feasible Policy Optimization

Our method, FeasPG, replays sample paths under π_θ and differentiates the cost $C(\theta, \xi)$ through the proposals, projections, and state transitions. Because of moving recourse, an early decision changes both later proposals and their feasible sets.

3.1 Training by simulation and differentiation

The expected cost $J(\theta)$ is what we want to minimize. However, projection can leave that cost unchanged as an infeasible proposal moves, giving no cost-gradient signal in that direction. To move proposals toward their projected decisions, we add the regularization term

$$d_{\text{pd}}(\theta, \xi) = \tfrac{1}{2} \sum_{t=1}^{T} \|u_t - x_t\|^2. \tag{5}$$

Let $\mathcal{D}_{\text{pd}}(\theta) = \mathbb{E}[d_{\text{pd}}(\theta, \xi)]$. Then, the regularized training objective is $H_\lambda = J + \lambda \mathcal{D}_{\text{pd}}$, where setting $\lambda = 0$ reduces to optimizing J in (3). On a sample path, let $\widehat{g}_{\text{PW}} = \nabla_\theta C$ denote the pathwise gradient and $\widehat{g}_\lambda = \nabla_\theta(C + \lambda d_{\text{pd}})$ the regularized gradient, where these derivatives exist. Algorithm 1 computes them by backpropagating through the policy trajectory on the fixed sample path.

Algorithm 1: Feasible Pathwise Gradient (FeasPG). Given θ_0, batch size N_{bat}, learning rates η_k keeping updates in Θ, and regularization weights $\lambda_k \geq 0$, repeat for $k = 0, \ldots, K-1$:
1. Take the next N_{bat} previously unused training sample paths from $\mathcal{D}_{\mathrm{fit}}$.
2. On each path, initialize $s_1 = s_1^0(\xi_1)$ and replay the policy (4) with dynamics (1) at θ_k.
3. Differentiate $C + \lambda_k d_{\mathrm{pd}}$ through the replay, including each projection's dependence on its proposal and, through moving recourse, on the state, and update $\theta_{k+1} = \theta_k - \eta_k G_k$, where $G_k = N_{\mathrm{bat}}^{-1} \sum_{i=1}^{N_{\mathrm{bat}}} \widehat{g}_{\lambda_k}(\theta_k; \xi^{k,i})$.

Save parameter values as checkpoints, each defining a feasible policy, and select one, denoted $\widehat{\theta}$, using only information from $\mathcal{D}_{\mathrm{fit}}$.

The displayed updates use stochastic gradient descent (SGD) with fresh minibatches. In the policy-comparison experiments, we use Adam on reused training pools and select the checkpoint with the lowest mean cost on separate validation paths within $\mathcal{D}_{\mathrm{fit}}$ (Section 5). The stationarity result below instead analyzes a randomly selected SGD iterate.

3.2 The pathwise gradient

Changing the parameters affects a projected decision in two ways: it moves the proposal, and it changes the state that determines the feasible set. We first differentiate the projection with respect to these two inputs, then propagate their effects through the state recursion. Below, D denotes a Jacobian and ∇ the gradient of a scalar function.

Write $D_u \bar{P}_t$ and $D_s \bar{P}_t$ for the projection's Jacobians with respect to its proposal and state. We evaluate them where the projection is differentiable and nearby states also admit feasible decisions. Proposition B.1 gives their formulas and Appendix B.3 explains this condition.

The state, decision, and proposal sensitivities record how these quantities change with θ. Let $S_t = D_\theta s_t$, $X_t = D_\theta x_t$, and $U_t = D_\theta u_t = D_\theta \mu_{\theta,t} + D_s \mu_{\theta,t} S_t$, where $D_\theta \mu_{\theta,t}$ holds the state fixed. Since the initial state does not depend on θ, $S_1 = 0$. The sensitivities then satisfy

$$X_t = D_u \bar{P}_t U_t + \underbrace{D_s \bar{P}_t S_t}_{\text{moving recourse}}, \qquad S_{t+1} = D_s f_t S_t + D_x f_t X_t. \tag{6}$$

These sensitivities give the regularized gradient

$$\widehat{g}_\lambda = \sum_{t=1}^{T} \left[S_t^\top \nabla_s c_t + X_t^\top \nabla_x c_t + \lambda (U_t - X_t)^\top (u_t - x_t) \right]. \tag{7}$$

Setting $\lambda = 0$ gives the pathwise gradient of the operating cost, $\widehat{g}_{\mathrm{PW}} = \widehat{g}_0$. These formulas require differentiable proposal, transition, and cost maps along the policy trajectory. Algorithm 1 evaluates the same derivative by reverse-mode differentiation, without forming the full sensitivity matrices. The moving-recourse term $D_s \bar{P}_t S_t$ in (6) is the derivative of the projection through $B_t s$ in (2). Omitting this term treats the constraint bounds as constant during differentiation, while leaving the forward states, decisions, and costs unchanged. The resulting gradient error does not generally average out over larger batches; Corollary B.2 bounds how it propagates across stages.

The next result establishes when these sampled derivatives provide unbiased stochastic gradients for the expected training objective.

Theorem 1 (Unbiased pathwise gradients (informal))**.** *Fix $\lambda \geq 0$. Under the local Lipschitz and measurability conditions of Theorem B.1 and corollary B.1, the sampled loss $C + \lambda d_{\mathrm{pd}}$ is differentiable almost surely at Lebesgue-almost every fixed θ. At any such parameter satisfying the integrability and local domination conditions of these results,*

$$\mathbb{E}[\widehat{g}_\lambda(\theta; \xi)] = \nabla H_\lambda(\theta).$$

The recursion (7) *computes this derivative when the maps along the policy trajectory satisfy the differentiability conditions above.*

Projection is nonexpansive in its proposal input, but moving recourse and the dynamics can still amplify pathwise gradients. Theorem B.3 and corollary B.8 bound their second moments and quantify how sensitivity can accumulate over the horizon.

3.3 What proposal-distance regularization guarantees

Moving freely chosen proposals toward their projected decisions reduces the regularizer without changing any decision or cost. This holds throughout the trajectory because unchanged earlier decisions preserve later feasible sets (Theorem B.5). In practice, at each stage, the deployed policy π_θ uses the same θ to generate proposals at different (s_t, z_t). Updating those parameters changes these proposals together, so it need not move each proposal toward its projection. Corollary B.7 gives a condition under which a small parameter update approximates this inward movement. A fixed regularization weight may favor smaller proposal distances over lower operating cost. This motivates the decreasing-weight schedule analyzed below.

Stationarity with decreasing regularization (informal). Under standard smoothness and stochastic-gradient assumptions, together with a uniform bound $\|\nabla \mathcal{D}_{\mathrm{pd}}(\theta_k)\| \le G_D$ whenever $\lambda_k > 0$, consider Algorithm 1's SGD updates with fresh minibatches for K iterations. Choose the learning rate $\eta_k = c/\sqrt{K}$ and regularization weights $\lambda_k = \lambda_0(k+1)^{-1/4}$, where $c > 0$ is a sufficiently small fixed constant and $\lambda_0 \ge 0$ is the initial regularization weight. If $\hat{k}$ is selected uniformly from $\{0, \dots, K-1\}$, independently of training, then

$$\mathbb{E}\|\nabla J(\theta_{\hat{k}})\|^2 = O(K^{-1/2}). \tag{8}$$

Theorem B.2 and corollary B.3 give the complete conditions and explicit bound. This rate concerns the original cost J at a randomly selected iterate. Algorithm 2 separately certifies the checkpoint selected for deployment.

4 Certifying the selected policy

Evaluating the feasible policy $\widehat{\pi}$ selected by Algorithm 1 estimates its expected cost, but does not reveal how far it is from optimal. We construct an upper confidence bound on $J(\widehat{\pi}) - V^\star$ by comparing the policy with a penalized perfect-information benchmark and correcting for estimation error. The policy is fixed using $\mathcal{D}_{\mathrm{fit}}$, with independent data $\mathcal{D}_{\mathrm{cal}}$ and $\mathcal{D}_{\mathrm{ev}}$ used for calibration and evaluation, respectively. The certificate does not require training to have converged.

4.1 A lower benchmark from perfect information

Suppose an optimizer could observe the entire sample path ξ before choosing its decisions, while still obeying the constraints (2) and dynamics (1). Let $\mathcal{X}^{\mathrm{PI}}(\xi)$ be the set of all feasible decision sequences x given ξ. The expected optimal value of this perfect-information problem is a lower bound on $V^\star$, but its information advantage can make the bound loose. We seek a tighter benchmark by adding a penalty whose expected value is zero for every policy $\pi \in \Pi$.

The penalty compares current cost plus approximate remaining cost with its conditional mean. Fix the selected feasible policy $\widehat{\pi}$, continuation functions $v_{2:T+1}$ ($v_{T+1} = 0$), and certification design using $\mathcal{D}_{\mathrm{fit}}$. Here $v_t(s, h)$ approximates the optimal expected cost remaining from stage t, given state s and history h. Its accuracy affects certificate tightness, not validity, so even $v \equiv 0$ is allowed. For a single certification run, probabilities and expectations below condition on all training and selection information, including the randomness used. For every state–decision–history query $a = (s, x, h)$ available to the perfect-information optimizer and possible next observation w, define the stagewise response

$$H_t(a, w) = \underbrace{c_t(s, x, w)}_{\text{current cost}} + \underbrace{v_{t+1}(f_t(s, x, w), h \oplus w)}_{\text{approximate remaining cost}}, \tag{9}$$

where $h \oplus w$ appends the next observation to the history. For certification, we assume that the conditional distribution of ξ_{t+1} given h_t depends only on z_t, as in (A.1). For a fixed query a, its conditional mean is $Q_t(a, z) = \int H_t(a, w)\overline{K}_t(\mathrm{d}w \mid z)$, where $\overline{K}_t(\cdot \mid z)$ is the unknown conditional distribution of the next observation given history summary z. For a decision sequence $x \in \mathcal{X}^{\mathrm{PI}}(\xi)$, let $a_t = (s_t, x_t, h_t)$ denote its query at stage t, with s_t given by (1). The penalty adds the difference between this conditional mean and the realized response at each stage. The penalized benchmark is

$$Y_v(\xi) = \inf_{x \in \mathcal{X}^{\mathrm{PI}}(\xi)} \left\{ C(x, \xi) + \sum_t [Q_t(a_t, z_t) - H_t(a_t, \xi_{t+1})] \right\}. \tag{10}$$

Every implementable policy is feasible for this perfect-information problem. Such a policy chooses its query before the next observation, and Q_t is the corresponding conditional mean response. Its penalty therefore has expectation zero, so $\mathbb{E}Y_v \leq J(\pi)$ for every $\pi \in \Pi$, and taking the infimum gives $\mathbb{E}Y_v \leq V^\star$ (Brown et al., 2010). With exact optimal continuation values $v = V$, Proposition D.2 gives $\mathbb{E}Y_V = V^\star$ under the Bellman conditions in Appendix A.2. Including the current-stage cost c_t in the penalty can be essential for a tight bound: correcting only the continuation value can leave the benchmark arbitrarily loose, even when that continuation value is exact (Proposition D.1). When Q_t is estimated from data, however, the resulting penalty need not have mean zero, even if the estimate is fitted independently. The next subsection uses a uniform error band to preserve the lower bound with high confidence.

4.2 Learning the correction from sample paths

The benchmark in (10) requires the unknown conditional mean $Q_t(a, z)$. We estimate it from $\mathcal{D}_{\mathrm{cal}}$ and bound the estimation error uniformly over a set $\mathcal{A}_t$ containing all queries reachable under perfect information. Uniformity is essential: the optimizer can choose decisions, and hence the states they reach, after seeing the entire path. An error bound only at queries visited by the selected policy would not suffice.

To estimate $Q_t(a, z)$, group the calibration observations by their history summaries z_t. Before calibration, partition the domain of z_t into finitely many cells. Let $N_{\mathrm{cal}} = |\mathcal{D}_{\mathrm{cal}}|$, and let $N_{t,B}$ count the observations whose stage-t history summary lies in cell B. For a requested $z \in B$ with $N_{t,B} > 0$, average the responses at the requested query a:

$$\widehat{Q}_t(a, z) = \frac{1}{N_{t,B}} \sum_{i: z_t^i \in B} H_t(a, \xi_{t+1}^i). \tag{11}$$

Each summand uses the same query a but a different recorded observation ξ_{t+1}^i. With the model and fitted continuation function fixed, we can evaluate these responses without observing the queried state or decision in the calibration data. An empty cell uses the midpoint of a certified response range and retains a valid, possibly wide bound.

We accompany this estimate with a computable error radius $\beta_t(z)$. It accounts for sampling error, the bias from pooling nearby values of z, and extension of the bound to every query. The construction uses certified response ranges and bounds on how H_t and Q_t vary between nearby queries and history summaries. These bounds and all design choices are fixed before calibration. Appendix C.1 gives the formal assumptions and the explicit radius.

Theorem 2 (Uniform calibration)**.** *Fix a failure probability $\delta \in (0, 1)$. Under Assumption C.1, the radius in* (C.2) *satisfies, with probability at least $1 - \delta$,*

$$|\widehat{Q}_t(a, z) - Q_t(a, z)| \leq \beta_t(z)$$

simultaneously for every stage t, query $a \in \mathcal{A}_t$, and history summary $z \in \mathrm{supp}(z_t)$. Theorem C.1 gives the full statement and proof.

Rate of the calibration radius. Under the conditions of Corollary C.1, suitable resolutions yield $\beta_t(z) = O((\log N_{\mathrm{cal}}/N_{\mathrm{cal}})^{\tau/(2\tau+d_z)})$ uniformly in t, z, with high probability at fixed confidence levels. Here $d_z = \max_t d_t$ is the largest history-summary dimension and $\tau \in (0, 1]$ is the smallest stagewise Hölder exponent of the conditional mean in z. For fixed continuation functions, query complexity enters through a logarithmic covering-number term, rather than the rate exponent. Theorem C.2 gives a matching exponent for uniform one-sided corrections.

4.3 Computing the certificate

We compare the policy $\widehat{\pi}$ with a penalized perfect-information benchmark on each evaluation path, then correct for calibration and sampling errors.

Constructing a corrected gap. First set $Q_t^- = \widehat{Q}_t - \beta_t$ and replace Q_t by Q_t^- in (10). Let $M^-(x, \xi) = \sum_t [Q_t^-(a_t, z_t) - H_t(a_t, \xi_{t+1})]$ denote the resulting penalty and $Y^-(\xi)$ the resulting inner optimum. On the uniform calibration event, $0 \leq Q_t - Q_t^- \leq 2\beta_t$. The replacement can

therefore only lower the benchmark, preserving $\mathbb{E}Y^- \le V^\star$ for each fixed calibration dataset on this event, with expectation over a fresh path.

For each evaluation path, compute a certified global lower bound $\underline{Y}^-(\xi) \le Y^-(\xi)$ and evaluate the same penalized objective at the policy's decisions. Their difference is nonnegative because those decisions are feasible for the perfect-information problem. However, replacing Q_t by Q_t^- also lowers the policy's penalty. We add $2\sum_t \beta_t(z_t)$ to cover this possible downward bias, giving

$$G^{\mathrm{band}}(\xi) = \underbrace{C^{\widehat{\pi}}(\xi) + M^{-,\widehat{\pi}}(\xi)}_{\text{penalized policy cost}} - \underbrace{\underline{Y}^-(\xi)}_{\text{inner lower bound}} + \underbrace{2\sum_t \beta_t(z_t)}_{\text{calibration correction}} . \tag{12}$$

Here $C^{\widehat{\pi}}$ and $M^{-,\widehat{\pi}}$ are evaluated along the selected policy on the same path as the inner problem. Conditional on the training and calibration information, on the uniform calibration event, $\mathbb{E}G^{\mathrm{band}} \ge J(\widehat{\pi}) - V^\star$, where expectation also includes any independent solver randomness. Thus the mean corrected gap bounds the optimality gap, although an individual pathwise gap need not do so. The inner computation must supply a lower bound, for example from a dual-feasible LP solution. A feasible inner decision sequence supplies an upper bound and cannot serve this purpose.

From pathwise gaps to a confidence bound. Algorithm 2 adds an empirical Bernstein margin (Maurer & Pontil, 2009) to the sample mean of corrected gaps. Here G denotes G^{band} or the paired binary refinement below.

Theorem 3 (Selected-policy coverage)**.** *Let $\widehat{\pi}$ be an integrable feasible policy selected using $\mathcal{D}_{\mathrm{fit}}$ only. Under the calibration, sampling, solver, and range conditions in Theorem D.1, Algorithm 2 satisfies, conditional on that training and selection information,*

$$\mathbb{P}\{J(\widehat{\pi}) - V^\star \le U_{\alpha,\delta}\} \ge 1 - \delta - \alpha.$$

The probability is over the calibration and evaluation data and any solver randomness. The guarantee requires neither accurate continuation values nor successful policy optimization.

Corollaries D.1 and D.2 also bound how far the certificate $U_{\alpha,\delta}$ can exceed the true optimality gap. More evaluation paths reduce sampling uncertainty, but do not eliminate the difference between the lower bound and the true optimal value.

Paired binary refinement used in our experiment. For binary uncertainty independent across stages and of predecision information, we correct the policy–benchmark gap directly, rather than first constructing a lower benchmark as in the band method. This exploits cancellation of shared probability-estimation errors and can tighten the certificate. The empirical inner optimum need not be a lower bound on $V^\star$ in expectation conditional on the calibration data. The binary construction in Appendix D.2 and Corollary D.4 give the correction and its certified range. The evaluation steps of Algorithm 2 then apply unchanged.

Algorithm 2: certify the selected policy.
Input. The selected feasible policy $\widehat{\pi}$ and continuation approximation fixed using $\mathcal{D}_{\mathrm{fit}}$, independent blocks $\mathcal{D}_{\mathrm{cal}}, \mathcal{D}_{\mathrm{ev}}$, and calibration and evaluation failure budgets $\delta, \alpha \in (0,1)$ with $\delta + \alpha < 1$.

1. **Calibrate.** Fix the correction construction, solver settings, and range rule before calibration. Use $\mathcal{D}_{\mathrm{cal}}$ to construct G from (12) or the paired correction (D.13).
2. **Bound the range.** Obtain certified finite bounds $0 \le A_G \le G \le B_G$, including the prescribed solver and numerical errors. Fix these endpoints and $W_G = B_G - A_G$ before evaluation.
3. **Evaluate.** On each of the $n = N_{\mathrm{ev}} = |\mathcal{D}_{\mathrm{ev}}| \ge 2$ independent paths, simulate $\widehat{\pi}$, compute a certified global lower bound for the corresponding inner problem, and form the complete corrected gap G_i.
4. **Report.** Add the empirical Bernstein margin to the mean corrected gap: $U_{\alpha,\delta} = \overline{G} + R_{\mathrm{EB}}(\alpha)$. The margin in (D.15) uses the gaps' sample variance and certified range width W_G. If any required check fails, report $+\infty$ rather than discard that path. Appendix D.2 gives the range constructions and numerical safeguards.

Using the certificate to stop training. Choose a target optimality gap ϵ. At each check, fix the policy, sample sizes, and settings before running Algorithm 2 on fresh calibration and evaluation paths. Stop if its bound is at most ϵ; otherwise, policy optimization may continue with Algorithm 1. Preassign failure probabilities across checks with a sum at most the desired overall failure probability. Corollary D.3 bounds the probability of any incorrect report by this total, including one that triggers stopping. Complete each prescribed evaluation sample before deciding whether to stop.

5 EXPERIMENTS

We compare FeasPG with SDDP and reinforcement learning baselines on three reservoir-management benchmarks: 1) a nonlinear 12-reservoir, 48-stage synthetic system, 2) ONS with four aggregate reservoirs, and 3) CAMELS-BR with 12 semi-synthetic reservoirs (Chagas et al., 2020). The ONS model is fitted to historical inflow and load records, and the CAMELS-BR model to historical inflow records. All three benchmarks use model-generated test paths; ONS and CAMELS-BR also use separately reserved historical records. We assess certificate coverage and tightness in a separate, tractable reservoir model.

We use Adam (Kingma & Ba, 2015) to train on a fixed pool of N_{fit} sample paths, reused during training, and select hyperparameters and the checkpoint $\widehat{\theta}$ on a separate validation subset of $\mathcal{D}_{\mathrm{fit}}$; N_{fit} counts training paths only. For each benchmark, every method is evaluated on the same 4,096 test paths, and the process is repeated five times. Additional experiments (gradient checks, the effect of regularization, etc.) can be found in Appendix F.2.

5.1 HOW DO POLICIES COMPARE ON SYNTHETIC AND HISTORICAL DATA?

In the experiments, we vary N_{fit} on the synthetic system, and fix $N_{\mathrm{fit}} = 2500$ on ONS and CAMELS-BR. We compare the following methods.

- **FeasPG+MLP**, **FeasPG+Graph**, **FeasPG+DeepSets**: our method with different proposal networks. Each architecture comparison approximately matches parameter counts. Graph also changes the proposal parameterization and initialization (Appendix G).
- **Projected SHAC** (Xu et al., 2022): the same proposal network, simulator, and projection as FeasPG+MLP, but differentiates through only eight stages, with a learned critic beyond them.
- **PPO** (Schulman et al., 2017) and **SAC** (Haarnoja et al., 2018): model-free reinforcement learning methods, projected onto the feasible set in the same way.
- **Linearized SDDP** and **Finite-state SDDP**: SDDP references. The former trains on a linear approximation of the nonconvex synthetic system, with its decisions projected at deployment (Appendix H); the latter runs directly on ONS and CAMELS-BR (Appendix K).
- **Seasonal**: a fixed storage-feedback rule.

Figure 1 reports test cost on the synthetic system (benchmark 1) over a growing number of sample paths, N_{fit}. FeasPG+Graph has the lowest cost at every N_{fit}, followed by FeasPG+MLP. In an additional experiment (Appendix J), we observe that omitting the moving-recourse term of the gradient increases the mean test cost of FeasPG+MLP from 78.4k to 80.9k with the same regularization weight, which shows that the naive gradient may underperform.

On ONS and CAMELS-BR, whose data models are fitted to historical records, Table 1 reports test cost on model-generated paths at $N_{\mathrm{fit}} = 2500$. FeasPG+MLP has the lowest cost on ONS, while FeasPG+DeepSets does on CAMELS-BR. We also observe that the model-free baselines are not reliable overall: PPO performs worse than the simple Seasonal heuristic on both benchmarks, and SAC on CAMELS-BR, whereas every model-based method outperforms Seasonal. This suggests the benefit of exploiting the model for well-structured problems, where it is available. In Appendix K, we further test the same policies on the historical records themselves and observe that FeasPG+DeepSets again achieves the lowest cost on CAMELS-BR.

5.2 IS THE CERTIFICATE VALID AND INFORMATIVE?

In this experiment, we evaluate the coverage and tightness of the certificate on a three-reservoir, 18-stage model with binary uncertainty, chosen so that the inner problems of Algorithm 2 can be solved to verified global optimality and the true optimality gap can be bracketed. We fix a projected

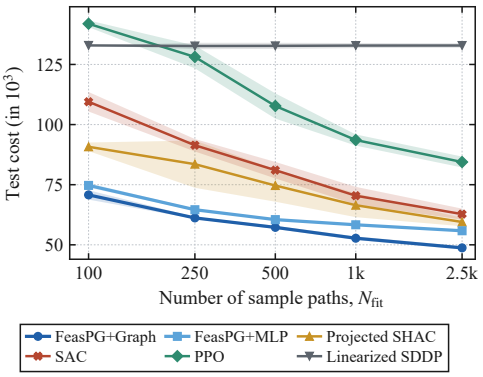

Figure 1: Test cost on the synthetic system over N_{fit}: lines represent means over five training runs (4,096 test paths for each) with ± 1 SD bands.

Method	ONS (10^9)	CAMELS-BR
Seasonal	214.4	222.3
FeasPG+MLP	$\mathbf{55.5 \pm 1.2}$	170.2 ± 0.3
FeasPG+DeepSets	–	$\mathbf{167.4 \pm 0.1}$
Projected SHAC	57.3 ± 1.1	200.7 ± 0.9
PPO	317.7 ± 49.7	490.2 ± 7.9
SAC	60.4 ± 1.6	606.8 ± 14.2
Finite-state SDDP	63.9 ± 0.8	174.8 ± 0.4

Table 1: Test cost on model-generated paths for ONS and CAMELS-BR at $N_{\text{fit}} = 2500$, mean $\pm$ SD over five training runs; bold marks the lowest cost.

affine policy $\widehat{\pi}$ and apply Algorithm 2 with independent calibration and evaluation sets, using either zero or fitted piecewise-linear (PWL) continuation functions v. Reference calculations with the true probabilities, which Algorithm 2 does not see, bracket $J(\widehat{\pi}) - V^\star$ in $[8.1, 10.7]$ (Appendix M).

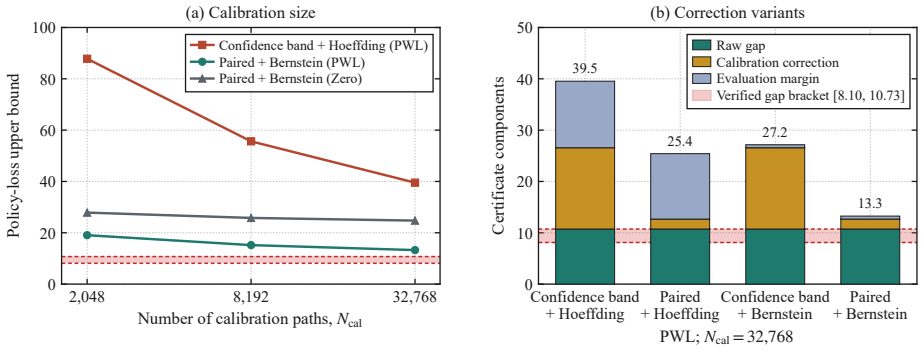

Figure 2: Certificates for a fixed policy on a three-reservoir, 18-stage model at nominal 95% coverage: mean and central 95% range over 200 repetitions per setting, $N_{\text{ev}} = 8{,}192$. Left: calibration size. Right: PWL correction variants at $N_{\text{cal}} = 32{,}768$. Paired/Confidence band are calibration corrections; Bernstein/Hoeffding are evaluation margins.

Figure 2 compares four certificates, combining confidence-band or paired calibration with Hoeffding or empirical Bernstein evaluation. Every certificate lies above the true gap in all 200 repetitions, although this does not establish exact nominal coverage. The results show that most of the gap between the certificate and the true value is due to the two corrections, and the paired correction with the Bernstein margin reduces it substantially. Appendix M details the range verification and the correction variants, and Appendix L gives a check with an exactly known optimum.

6 Conclusion

FeasPG learns feasible MSP policies while accounting for how earlier decisions change later constraints. Ignoring this dependence can introduce gradient errors that do not disappear with larger batches. A suitable decreasing regularization weight preserves SGD's stationarity rate for operating cost. The finite-sample certificate complements this guarantee by bounding the selected policy's optimality gap with high confidence. Future work could use our sensitivity bounds to shorten backward passes and our certificate error bounds to determine whether additional calibration or evaluation data would be most useful.

AI use statement. Generative AI tools assisted software development, experiment orchestration, result auditing, literature search, drafting, and editing. The authors reviewed all AI-assisted code, mathematics, artifacts, and text and take responsibility for the final content.

Ethics statement. The experiments use public hydrology and reservoir data. Appendix K reports provenance, processing, missing-data handling, and limitations. We make no operational-deployment claim.

Reproducibility statement. The appendices give full assumptions and proofs, experimental models, architectures, data splits, hyperparameters, solver and numerical checks, and complete results. The anonymized artifact will contain code, frozen manifests, processed-data provenance, and reproduction commands.

REFERENCES

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter. Differentiable convex optimization layers. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Matias Alvo, Daniel Russo, Yash Kanoria, and Minuk Lee. Deep reinforcement learning for inventory networks: Toward reliable policy optimization. *arXiv preprint arXiv:2306.11246*, 2025.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145, 2017.

Santiago R. Balseiro and David B. Brown. Approximations to stochastic dynamic programs via information relaxation duality. *Operations Research*, 67(2):577–597, 2019. doi: 10.1287/opre.2018.1782.

Dimitris Bertsimas and Kimberly Villalobos Carballo. Multistage stochastic optimization via kernels. *arXiv preprint arXiv:2303.06515*, 2023. doi: 10.48550/arXiv.2303.06515.

Dimitris Bertsimas, Shimrit Shtern, and Bradley Sturt. A data-driven approach to multistage stochastic linear optimization. *Management Science*, 69(1):51–74, 2023. doi: 10.1287/mnsc.2022.4352.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927, 2024. doi: 10.1287/opre.2021.0014.

David B. Brown, James E. Smith, and Peng Sun. Information relaxations and duality in stochastic dynamic programs. *Operations Research*, 58(4-part-1):785–801, 2010. doi: 10.1287/opre.1090.0796.

Vinícius B. P. Chagas, Pedro L. B. Chaffe, Nans Addor, Fernando M. Fan, Ayan S. Fleischmann, Rodrigo C. D. Paiva, and Vinícius A. Siqueira. CAMELS-BR: Hydrometeorological time series and landscape attributes for 897 catchments in brazil. *Earth System Science Data*, 12:2075–2096, 2020. doi: 10.5194/essd-12-2075-2020.

Bingqing Chen, Priya L. Donti, Kyri Baker, J. Zico Kolter, and Mario Bergés. Enforcing policy feasibility constraints through differentiable projection for energy optimization. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, pp. 199–210, 2021. doi: 10.1145/3447555.3464874.

Nan Chen, Xiang Ma, Yanchu Liu, and Wei Yu. Information relaxation and a duality-driven algorithm for stochastic dynamic programs. *Operations Research*, 72(6):2302–2320, 2024. doi: 10.1287/opre.2020.0464.

Xin Chen, Yifan Hu, and Minda Zhao. Landscape of policy optimization for finite horizon MDPs with general state and action. *arXiv preprint arXiv:2409.17138*, 2026. doi: 10.48550/arXiv.2409.17138. Revised March 2026.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818, 2014. doi: 10.1214/14-AOS1235.

Anukal Chiralaksanakul and David P. Morton. Assessing policy quality in multi-stage stochastic programming. Technical Report 2004-12, Stochastic Programming E-Print Series, 2004. URL https://doi.org/10.18452/8319.

Boris Defourny, Damien Ernst, and Louis Wehenkel. Scenario trees and policy selection for multistage stochastic programming using machine learning. *INFORMS Journal on Computing*, 25(3): 488–501, 2013. doi: 10.1287/ijoc.1120.0516.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1587–1596, 2018.

Christian Füllner and Steffen Rebennack. Stochastic dual dynamic programming and its variants: A review. *SIAM Review*, 67(3):415–539, 2025. doi: 10.1137/23M1575093.

Giulio Galvan, Marco Sciandrone, and Stefano Lucidi. A parameter-free unconstrained reformulation for nonsmooth problems with convex constraints. *Computational Optimization and Applications*, 80(1):33–53, 2021. doi: 10.1007/s10589-021-00296-1.

Harsha Gangammanavar and Suvrajeet Sen. Stochastic dynamic linear programming: A sequential sampling algorithm for multistage stochastic linear programming. *SIAM Journal on Optimization*, 31(3):2111–2140, 2021.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Saeed Ghadimi, Raymond T. Perkins, and Warren B. Powell. Reinforcement learning via parametric cost function approximation for multistage stochastic programming. *Optimization Online*, 2019. URL https://optimization-online.org/wp-content/uploads/2018/12/6999.pdf. Optimization Online preprint.

Martin Glanzer, Sebastian Maier, and Georg Ch. Pflug. Guaranteed bounds for optimal stopping problems using kernel-based non-asymptotic uniform confidence bands. *European Journal of Operational Research*, 327(1):162–173, 2025. doi: 10.1016/j.ejor.2025.05.028.

Paul Glasserman. *Gradient Estimation via Perturbation Analysis*. Kluwer Academic Publishers, Boston, 1991. URL https://link.springer.com/book/9780792390954.

Sébastien Gros, Mario Zanon, and Alberto Bemporad. Safe reinforcement learning via projection on a safe set: How to achieve optimality? *IFAC-PapersOnLine*, 53(2):8076–8081, 2020. doi: 10.1016/j.ifacol.2020.12.2276.

Vincent Guigues and René Henrion. Joint dynamic probabilistic constraints with projected linear decision rules. *Optimization Methods and Software*, 32(5):1006–1032, 2017. doi: 10.1080/10556788.2016.1233972.

Vincent Guigues, Alexander Shapiro, and Yi Cheng. Duality and sensitivity analysis of multistage linear stochastic programs. *European Journal of Operational Research*, 308(2):752–767, 2023. doi: 10.1016/j.ejor.2022.11.051.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1861–1870, 2018.

Patrick P. Helm, Jan-Niklas Doerr, Joren Gijsbrechts, and Stefan Minner. Hard constraints, smooth gradients: Learning feasible inventory policies via differentiable projection. *arXiv preprint arXiv:2608.02343*, 2026. doi: 10.48550/arXiv.2608.02343.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012. doi: 10.1214/ECP.v17-2079.

Heinrich Jiang. Non-asymptotic uniform rates of consistency for k-NN regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3999–4006, 2019. doi: 10.1609/aaai.v33i01.33013999.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020. URL https://www.jmlr.org/papers/v21/19-827.html.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: Sharp partial identification under smoothness. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23734–23757. PMLR, 2024. URL https://proceedings.mlr.press/v235/khan24b.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Jyun-Li Lin, Wei Hung, Shang-Hsuan Yang, Ping-Chun Hsieh, and Xi Liu. Escaping from zero gradient: Revisiting action-constrained reinforcement learning via Frank–Wolfe policy optimization. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 397–407, 2021.

Mark G. Low. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554, 1997. doi: 10.1214/aos/1030741084.

Dhruv Madeka, Kari Torkkola, Carson Eisenach, Anna Luo, Dean P. Foster, and Sham M. Kakade. Deep inventory management. *arXiv preprint arXiv:2210.03137*, 2022.

Wai-Kei Mak, David P. Morton, and R. Kevin Wood. Monte carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24(1–2):47–56, 1999. doi: 10.1016/S0167-6377(98)00054-6.

Hannah Markgraf, Shambhuraj Sawant, Hanna Krasowski, Lukas Schäfer, Sébastien Gros, and Matthias Althoff. Safe reinforcement learning using action projection: Safeguard the policy or the environment? *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=DDrGSEYxGU.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Vladimir I. Norkin. The exact projective penalty method for constrained optimization. *Journal of Global Optimization*, 89(2):259–276, 2024. doi: 10.1007/s10898-023-01350-4.

Hyuk Park, Zhuangzhuang Jia, and Grani A. Hanasusanto. Data-driven stochastic dual dynamic programming: Performance guarantees and regularization schemes. *Optimization Online*, 2022. Revised May 2025. Available from Optimization Online.

Mario V. F. Pereira and Leontina M. V. G. Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52:359–375, 1991. doi: 10.1007/BF01582895.

Tu-Hoa Pham, Giovanni De Magistris, and Ryuki Tachibana. OptLayer: Practical constrained optimization for deep reinforcement learning in the real world. In *2018 IEEE International Conference on Robotics and Automation*, pp. 6236–6243, 2018. doi: 10.1109/ICRA.2018.8460547.

A. B. Philpott, F. Wahid, and J. F. Bonnans. MIDAS: A mixed integer dynamic approximation scheme. *Mathematical Programming*, 181(1):19–50, 2020.

L. C. G. Rogers. Pathwise stochastic optimal control. *SIAM Journal on Control and Optimization*, 46(3):1116–1132, 2007. doi: 10.1137/050642885.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Alexander Shapiro. Inference of statistical bounds for multistage stochastic programming problems. *Mathematical Methods of Operations Research*, 58(1):57–68, 2003. doi: 10.1007/s001860300280.

Alexander Shapiro. Analysis of stochastic dual dynamic programming method. *European Journal of Operational Research*, 209(1):63–72, 2011. doi: 10.1016/j.ejor.2010.08.007.

Thuener Silva, Davi Valladão, and Tito Homem-de Mello. A data-driven approach for a class of stochastic dynamic optimization problems. *Computational Optimization and Applications*, 80(3):687–729, 2021. doi: 10.1007/s10589-021-00320-4.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. doi: 10.1214/aos/1176345969.

Hyung Ju Suh, Max Simchowitz, Kaiqing Zhang, and Russ Tedrake. Do differentiable simulators give better policy gradients? In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20668–20696, 2022.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pp. 1057–1063, 2000.

Gonçalo Terça and David Wozabal. Envelope theorems for multistage linear stochastic optimization. *Operations Research*, 69(5):1608–1629, 2021. doi: 10.1287/opre.2020.2038.

Petter Tøndel, Tor Arne Johansen, and Alberto Bemporad. An algorithm for multi-parametric quadratic programming and explicit MPC solutions. *Automatica*, 39(3):489–497, 2003. doi: 10.1016/S0005-1098(02)00250-9.

Jiajun Xu and Suvrajeet Sen. Compromise policy for multi-stage stochastic linear programming: Variance and bias reduction. *Computers & Operations Research*, 153:106132, 2023.

Jie Xu, Viktor Makoviychuk, Yashraj Narang, Fabio Ramos, Wojciech Matusik, Animesh Garg, and Miles Macklin. Accelerated policy learning with parallel differentiable simulation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ZSKRQMvttc.

Yunhao Yan and Henry Lam. Data-driven solutions and uncertainty quantification for multistage stochastic optimization. In *Proceedings of the 2024 Winter Simulation Conference*, pp. 3300–3311, 2024. doi: 10.1109/WSC63780.2024.10838792.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Shenao Zhang, Boyi Liu, Zhaoran Wang, and Tuo Zhao. Model-based reparameterization policy gradient methods: Theory and practical algorithms. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Helin Zhu, Fan Ye, and Enlu Zhou. Solving the dual problems of dynamic programs via regression. *IEEE Transactions on Automatic Control*, 63(5):1340–1355, 2018. doi: 10.1109/TAC.2017.2747405.

Jikai Zou, Shabbir Ahmed, and Xu Andy Sun. Stochastic dual dynamic integer programming. *Mathematical Programming*, 175(1–2):461–502, 2019.

GUIDE TO THE APPENDICES

Appendix A gives the full model and further related literature. Appendix B states and proves the learning results, including projection sensitivity, decreasing-weight convergence, and proposal-distance geometry. Appendix C develops the uniform bands and their statistical boundaries, while Appendix D gives the full certification, paired-correction, and sequential-stopping results. Appendix E works through an inventory instance. Appendices F and G give complete comparisons and implementations. Appendices H–K give the training benchmarks and diagnostics, while Appendices L and M give the certificate studies.

A MODEL, ASSUMPTIONS, AND FURTHER RELATED LITERATURE

A.1 MULTISTAGE MODEL AND PROJECTED POLICIES

The endogenous state changes with past decisions, whereas the distribution of the exogenous observations does not depend on the policy. Training differentiates through the state transitions. Certification reuses the observations to evaluate alternative decisions.

Notation. For a vector-valued map g, $D_s g$ denotes its Jacobian with respect to s, while for a scalar-valued map c, $\nabla_s c$ denotes its column gradient. I_m is the $m \times m$ identity, and $\|\cdot\|$ denotes the Euclidean norm of a vector and the operator norm of a matrix.

Data process and information. Let $\xi = (\xi_1, \ldots, \xi_{T+1})$ denote a sample path of the exogenous data, such as an inflow and load sequence, where ξ_t takes values in Ξ_t. The information available at stage t is $\mathcal{F}_t = \sigma(\xi_{1:t})$. We adopt the standard exogenous-sampling model of stochastic programming: the data consist of independent sample paths from an unknown distribution that does not depend on the decisions or policy parameters (Shapiro, 2011; Füllner & Rebennack, 2025). This independence concerns different sample paths, not the observations within a path, which may be dependent across stages.

Timing. At stage t, the decision maker observes the endogenous state $s_t \in \mathbb{R}^{n_t}$ and a history summary $z_t = \varphi_t(\xi_{1:t}) \in \mathcal{Z}_t$, and chooses $x_t \in \mathbb{R}^{m_t}$. The next uncertainty realization ξ_{t+1} is then observed and the state advances. Consequently, the decision and its feasible set may depend on information available through stage t, but not on ξ_{t+1}.

The history summary allows the policy to use past observations when the current observation alone is insufficient under serial dependence (Füllner & Rebennack, 2025). Unlike s_t, this summary remains fixed when the policy changes on a given sample path. We take φ_t and the initial-state map $s_1 = s_1^0(\xi_1)$ to be measurable and independent of θ. Policy optimization does not require the summary to predict the next observation. Certification additionally uses the predictive sufficiency condition specified next.

Predictively sufficient history summary. For the certification construction in Appendix D.1, we use a history summary that retains the information needed to predict the next-stage uncertainty (Füllner & Rebennack, 2025). Specifically, we assume that the conditional distribution of ξ_{t+1} given the observed history depends on that history only through z_t. Write $\overline{K}_t(\cdot \mid z)$ for this conditional distribution. Thus, for every measurable set A,

$$\mathbb{P}\{\xi_{t+1} \in A \mid \mathcal{F}_t\} = \overline{K}_t(A \mid z_t) \quad \text{almost surely.} \tag{A.1}$$

The map φ_t is known, but the conditional distribution is not. The full history $z_t = \xi_{1:t}$ always satisfies this condition. When a lower-dimensional sufficient summary is available, it reduces the dimension of the conditional-mean estimation problem. For a Markov process, one may take $z_t = \xi_t$. We write $h_t = \xi_{1:t}$ for the full observed history and use h for a possible history at the relevant stage. Thus $z_t = \varphi_t(h_t)$ is its chosen summary. We write $\mathsf{Z}_t = \operatorname{supp}(z_t)$ for the support of z_t and let $\mathcal{Z}_t \supseteq \mathsf{Z}_t$ denote the modeled domain on which the policy and, when required, the certificate are defined.

Policy. Let $\Theta \subseteq \mathbb{R}^{d_\theta}$ be the parameter domain. The policy π_θ is generated by a sequence of parameterized proposal maps $\mu_{\theta,t} : \mathbb{R}^{n_t} \times \mathcal{Z}_t \to \mathbb{R}^{m_t}$. These maps may be affine or implemented by feedforward neural networks (multilayer perceptrons, or MLPs), graph neural networks, or architectures that preserve permutation symmetries. In each case, the output $u_t^\theta = \mu_{\theta,t}(s_t^\theta, z_t)$ is an unconstrained *proposal*, which we project onto the current feasible set to obtain the implemented (or deployed) decision. Superscripts record the induced policy trajectory: s_t^θ and x_t^θ are the state and decision generated by π_θ.

Feasible set. We adopt the standard fixed-recourse representation used in multistage linear models: the constraint normals are fixed, while the observed history and endogenous state move the right-hand side (Shapiro, 2011). Thus, the decision at stage t must lie in the polyhedron

$$\mathcal{X}_t(s, z) = \{x \in \mathbb{R}^{m_t} : A_t x \le q_t(s, z)\}, \qquad q_t(s, z) = b_t(z) + B_t s, \tag{A.2}$$

where $A_t \in \mathbb{R}^{k_t \times m_t}$ and $B_t \in \mathbb{R}^{k_t \times n_t}$ are known matrices and $b_t : \mathcal{Z}_t \to \mathbb{R}^{k_t}$ is a known measurable map. The constraint coefficients may have either sign, and the matrix may contain dependent or redundant rows, including those introduced by writing balance equalities as paired inequalities. The bounds change with the state through $B_t s$ and with observed data through $b_t(z)$. In a reservoir model, for example, $B_t s$ captures storage-dependent release limits, while $b_t(z)$ may account for inflows already observed at decision time. The representation makes the state dependence of the projection explicit without requiring convex costs or dynamics.

As is standard in MSP, we impose relatively complete recourse: $\mathcal{X}_t(s, z)$ is nonempty at every state reachable by a feasible decision prefix, for almost every realization of the data process (Shapiro, 2011). The policy can therefore implement the Euclidean projection of its proposal at every stage,

$$P_t(u, q) := \arg\min_{A_t x \le q} \tfrac{1}{2}\|x - u\|^2, \qquad x_t^\theta = P_t\big(u_t^\theta, q_t(s_t^\theta, z_t)\big), \tag{A.3}$$

a strongly convex quadratic program with a unique solution. The same rule is used in training and in deployment.

Dynamics and objective. Once x_t is implemented and ξ_{t+1} is revealed, the state advances and a stage cost is incurred:

$$s_{t+1}^\theta = f_t(s_t^\theta, x_t^\theta, \xi_{t+1}), \qquad J(\theta) := \mathbb{E}\left[\sum_{t=1}^{T} c_t(s_t^\theta, x_t^\theta, \xi_{t+1})\right], \tag{A.4}$$

where $f_t : \mathbb{R}^{n_t} \times \mathbb{R}^{m_t} \times \Xi_{t+1} \to \mathbb{R}^{n_{t+1}}$ is the transition function and $c_t : \mathbb{R}^{n_t} \times \mathbb{R}^{m_t} \times \Xi_{t+1} \to \mathbb{R}$ the stage cost. Both maps are known. For the pathwise sensitivity analysis, we assume local Lipschitz continuity in (s, x) for each fixed uncertainty realization on the relevant trajectory domains, as is customary in this analysis (Glasserman, 1991). This regularity allows nonlinear and nonconvex costs and dynamics. All pathwise derivatives below hold the exogenous observations and history summaries fixed. A terminal cost $\psi(s_{T+1})$ is absorbed into the last stage, with $c_T(s_T, x_T, \xi_{T+1})$ understood to include $\psi(f_T(s_T, x_T, \xi_{T+1}))$, and all derivatives below use this composed c_T. We write $J(\theta)$ as shorthand for $J(\pi_\theta)$, the expected cost of the policy determined by θ.

The multistage program. Let Π contain nonanticipative policies whose induced decisions are feasible almost surely, whose states follow the stated dynamics, and whose total costs are integrable. The multistage stochastic program is

$$V^\star := \inf_{\pi \in \Pi} J(\pi), \qquad J(\pi) = \mathbb{E}\Big[\sum_{t=1}^{T} c_t(s_t, x_t, \xi_{t+1})\Big], \tag{A.5}$$

assumed finite. We write $C(x, \xi) = \sum_{t=1}^{T} c_t(s_t, x_t, \xi_{t+1})$ for the realized cost of a decision sequence on path ξ, with the associated states generated by (A.4).

Proposition A.1 (Nonanticipativity and feasibility)**.** *Suppose the model and policy maps are measurable, relatively complete recourse holds, and* (A.3) *is solved exactly. Then, for every θ, the decision x_t^θ is $\mathcal{F}_t$-measurable and satisfies $A_t x_t^\theta \le q_t(s_t^\theta, z_t)$ for $t = 1, \dots, T$, for almost every realization of the data process. Hence every projected policy with integrable cost is a nonanticipative feasible policy for the multistage problem, so that $\pi_\theta \in \Pi$ and $V^\star \le J(\theta)$.*

Every iterate therefore defines a feasible policy, whether or not training has converged. Training seeks to reduce its expected cost. From Appendix B.1 onward, we omit the superscript θ when describing the trajectory of a fixed policy.

Data for training and certification. The available data consist of N independent complete sample paths $\{\xi^i_{1:T+1}\}_{i=1}^N$. Neither the conditional distribution nor an exact conditional simulator is supplied. Our information about the distribution comes only from these sample paths. The model functions of the MSP remain known. Following standard sample-splitting practice in out-of-sample solution assessment (Mak et al., 1999; Shapiro, 2003), we partition the paths into a training and selection block $\mathcal{D}_{\mathrm{fit}}$, a calibration block $\mathcal{D}_{\mathrm{cal}}$ for constructing the certificate correction, and an evaluation block $\mathcal{D}_{\mathrm{ev}}$ for computing the reported bound. The training and selection phase also fixes the continuation approximation, query domains, and certification design. With these choices held fixed, the calibration data estimate the conditional means and bound their estimation errors. The evaluation data are used only for the final confidence bound. Let $\mathcal{G}_{\mathrm{fit}}$ collect all information fixed after training and selection—formally, the sigma-field generated by the training and selection data, selected policy, fixed continuation approximation and certification choices, and training randomness. Conditional on $\mathcal{G}_{\mathrm{fit}}$, the calibration and evaluation blocks remain independent samples from the sample-path distribution, allowing certification to assess the selected policy without reusing its training data.

A.2 Formal setting and scope

The main text introduces the model conditions where they are used. This appendix gives the technical conditions and specifies which results require them. The additional conditions for rates, approximation bounds, and certificate width are not required for the basic feasibility and coverage results.

Dimensions and information. The parameter domain is $\Theta \subseteq \mathbb{R}^{d_\theta}$; $s_t \in \mathbb{R}^{n_t}$, $x_t, u_t \in \mathbb{R}^{m_t}$, $A_t \in \mathbb{R}^{k_t \times m_t}$, $B_t \in \mathbb{R}^{k_t \times n_t}$, and $q_t \in \mathbb{R}^{k_t}$. The uncertainty spaces are standard Borel spaces equipped with fixed compatible Polish topologies, $\mathcal{F}_t = \sigma(\xi_{1:t})$, $s_1 = s_1^0(\xi_1)$, and $z_t = \varphi_t(\xi_{1:t})$. The causal maps s_1^0 and φ_t are Borel measurable and independent of θ. The distribution of the complete exogenous sample path does not depend on the decisions or on θ. Standard Borel spaces admit a measurable version of the conditional distribution in (A.1), and $\overline{K}_t$ denotes one fixed such version. For the conditional-mean construction, conditional on $\mathcal{G}_{\mathrm{fit}}$, each $(\mathcal{A}_t, d_{\mathcal{A},t})$ is a separable Borel metric space, H_t is jointly Borel measurable in the query and the uncertainty realization, and Q_t is a jointly measurable version of the conditional mean. These standard conditions ensure that the displayed conditional expectations, suprema, and confidence events are well defined. Let Π be the class of integrable, nonanticipative feasible policies and define

$$J(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} c_t(s_t, x_t, \xi_{t+1})\right], \qquad V^\star = \inf_{\pi \in \Pi} J(\pi).$$

For the projected policy, $J(\pi_\theta) = J(\theta)$ as defined in (A.4).

Feasibility and measurability. The maps $b_t, f_t, c_t, \mu_{\theta,t}$ are Borel measurable, and the projection map P_t is measurable wherever its feasible polyhedron is nonempty. For almost every history and every state reached by a feasible decision prefix along that history, relatively complete recourse requires $\mathcal{X}_t(s_t, z_t) \neq \varnothing$. The mathematical policy uses the exact minimizer in (A.3).

Bellman values and exactness. For the exactness statement in Appendix D.1.1, we impose the standard measurability and selection conditions of finite-horizon dynamic programming and assume that the Bellman values in (D.6) are finite (Shapiro, 2011; Brown et al., 2010). In particular, the value functions are measurable, and measurable minimizing sequences suffice when the Bellman infima are not attained. Under these conditions and the required integrability, $\mathbb{E}[V_1(s_1, h_1)] = V^\star$. These dynamic-programming conditions justify exactness with $v = V$. The lower-bound argument for a general continuation approximation instead uses integrability and a nonpositive expected penalty under every feasible nonanticipative policy.

Differentiability and domination. For Theorem B.1, Θ is open and, for almost every sample path, the composite $C(\cdot, \xi)$ is locally Lipschitz. Its nondifferentiability set is jointly measurable in the parameter and the sample path. At a fixed parameter satisfying the theorem's hypotheses, the sample-path difference quotients on a fixed neighborhood are bounded in absolute value by a Lipschitz constant that may depend on the sample path and has finite expectation. To identify the derivative with (B.4)–(B.6), the encountered proposals are differentiable in (θ, s) and the transition and cost maps in (s, x) at their trajectory arguments, with uncertainty realizations and history summaries held fixed. In addition, each encountered $\bar{P}_t(\cdot, \cdot; z_t)$ is differentiable at (u_t, s_t) in an open feasible neighborhood as in Proposition B.1, almost surely. This additional neighborhood condition does not follow from relatively complete recourse only on reachable states, and ambient almost-everywhere differentiability does not ensure differentiability at the sampled arguments. The uniform derivative bounds needed for the horizon theorem are precisely (B.15).

Stochastic-gradient specialization. The experiments use Adam for training on reused path pools, whereas the SGD analysis of Algorithm 1 uses fresh minibatches. Theorem B.2 analyzes its displayed SGD update under the stated conditional-mean and variance conditions. Fresh independent minibatches of sample paths and exact sample derivatives satisfying the interchange conditions yield the required conditional mean. Corollary B.4 records the zero-weight specialization. Either $\Theta = \mathbb{R}^{d_\theta}$, or the parameterization and step size keep every update segment inside the open convex regularity domain. Given the pre-update sigma-field $\mathcal{G}_k^{\mathrm{tr}}$, paths within a batch are conditionally independent draws from the sample-path distribution, with arbitrary serial dependence allowed inside each path. Repeated reuse of one fixed sample instead corresponds to empirical-risk optimization.

A.3 Proof of Proposition A.1

Proceed by induction over the decision stages. The initial state $s_1 = s_1^0(\xi_1)$ and z_1 are $\mathcal{F}_1$-measurable. Suppose the previous projected decisions form a feasible prefix and s_t is $\mathcal{F}_t$-measurable. Relatively complete recourse makes the current projection set nonempty. Measurability of the proposal, right-hand side, and projection then makes x_t^θ $\mathcal{F}_t$-measurable, while the definition of the projection gives

$$A_t x_t^\theta \le q_t(s_t^\theta, z_t).$$

The transition map makes s_{t+1}^θ $\mathcal{F}_{t+1}$-measurable and extends the feasible prefix. The induction holds almost surely through stage T. If the total cost is integrable, the induced policy lies in Π, so its expected cost upper-bounds $V^\star$.

A.4 Related Literature

Parameterized policies and feasible learning. Reusable policies for MSPs have been optimized over several policy classes. Ghadimi et al. (2019) tune right-hand-side parameters of a deterministic lookahead policy evaluated in a stochastic base model, propagating parameter sensitivities through the endogenous state and the state-dependent right-hand side of the lookahead linear program. They study both gradient-based stochastic approximation and smoothed search based only on function values, with a finite-time stationarity bound for the smoothed objective. Bertsimas & Villalobos Carballo (2023) instead parameterize stage decisions directly by side information and exogenous histories in a reproducing-kernel Hilbert space (RKHS) and optimize the resulting decision rules by functional stochastic gradient descent, establishing asymptotic optimality under their convexity and approximation assumptions. An alternative is to learn a policy from optimized decisions: Defourny et al. (2013) fit nonlinear policy-function approximations to scenario-tree decisions and restore feasibility at deployment through a state-dependent optimization problem. Their construction is supervised policy extraction followed by out-of-sample policy selection, whereas we optimize the cost of trajectories generated by the projected policy directly and develop a finite-sample optimality-gap certificate relative to the unrestricted MSP optimum.

Guigues & Henrion (2017) also enforce feasibility on each realized path, projecting decision rules parameterized by exogenous histories onto hard constraints. Their projection sets can depend recursively on earlier projected decisions. In their principal tractable specification, the proposals are linear in exogenous history under Gaussian or truncated-Gaussian models. For each fixed parameter, our state-feedback policy likewise defines a rule based on the observed history. We study how

sensitivities grow along the resulting trajectory and quantify the error from omitting the derivative of the constraint bounds.

Differentiable constrained policies and pathwise gradients. Implicit differentiation through optimization layers is established (Amos & Kolter, 2017; Agrawal et al., 2019). Constraint-enforcing layers have also been used in control and energy policies (Pham et al., 2018; Chen et al., 2021), with PROF augmenting the cost objective by the squared distance between the proposal and its projection and motivating this regularization term through one-step gradient calculations (Chen et al., 2021). Building on this regularizer and established projection derivatives, we analyze its behavior when projected decisions change later recourse sets and the policy shares parameters across stages or histories. We also study stochastic approximation to the original MSP objective with decreasing regularization weights. Static projective-penalty methods establish analogous exactness and stationarity properties for constrained optimization (Galvan et al., 2021; Norkin, 2024). Our geometric result extends this perspective to sequential, state-dependent recourse sets. Safe and action-constrained reinforcement learning also studies projection corrections and projection-induced loss of gradient information (Gros et al., 2020; Lin et al., 2021; Markgraf et al., 2026).

Differentiable simulators propagate derivatives through known dynamics and show how sensitivities can grow or decay over long horizons (Suh et al., 2022; Zhang et al., 2023). In inventory control, Madeka et al. (2022) and Alvo et al. (2025) learn policies by differentiating simulated costs. Alvo et al. also use feasibility layers tailored to inventory-allocation constraints. Related multistage sensitivity analyses differentiate optimized value functions or model data (Terça & Wozabal, 2021; Guigues et al., 2023).

The concurrent work of Helm et al. (2026) is discussed in the introduction. Here we record the technical relation. Their Lemma 1 gives the target and right-hand-side Jacobians of a weighted projection under a locally fixed, linearly independent active set, and their Corollary 1 shows that the projection is nonexpansive in its target in the projection's weighted norm. Proposition B.1 gives both Jacobians at differentiability points using the full active set, including dependent rows and the paired inequalities that represent balance equalities. Their integer recovery requires that componentwise reductions of a feasible decision remain feasible, which nonnegative resource coefficients ensure, whereas continuous projection needs no such condition. Beyond the Jacobians, our multistage results have no counterpart there: the horizon bounds of Theorem B.3 and Corollary B.2, the proposal-distance analysis and the stationarity bound of Theorem B.2, and the certificate of Appendix D.1. These results complement global policy-gradient guarantees obtained under problem-specific landscape conditions for other finite-horizon objectives (Bhandari & Russo, 2024; Chen et al., 2026).

Information relaxation and policy assessment. Pathwise dual formulations originate with Rogers (2007) and Brown et al. (2010), while Balseiro & Brown (2019) develop heuristic performance bounds from approximate value functions. For estimating these bounds, particularly relevant are Zhu et al. (2018), who fit information-relaxation penalties by regression while preserving the lower-bound property, and Chen et al. (2024), who develop a duality-driven algorithm with regression Monte Carlo and confidence-interval estimates. Simulation-based solution-quality assessment also has a long history in stochastic programming (Mak et al., 1999; Shapiro, 2003; Chiralaksanakul & Morton, 2004).

We build on classical information-relaxation penalties with conditionally mean-zero increments, evaluating the policy and relaxation on the same path to obtain finite-sample optimality-gap certificates from recorded uncertainty realizations. Arbitrary policy–penalty pairing is permitted by Balseiro & Brown (2019). Here, the policy and continuation approximation used to construct the penalty may be chosen separately, using only training data, and certification requires neither a known transition distribution nor an exact conditional simulator. To account for the unknown distribution, conditional-mean bands bound estimation error uniformly over the perfect-information optimizer's queries and give a conservative correction. For independent binary uncertainty, we instead correct the policy–benchmark gap directly, allowing shared estimation errors to cancel. Both constructions also account for inner-solve and evaluation errors. Separately, our tightness analysis gives a smooth convex example in which centering the realized cost is necessary for tightness even with an exact continuation value, and shows that a continuation approximation can still give an exact lower bound when its centered error is dominated by each alternative decision's one-step suboptimality in the Bellman recursion.

Uniform estimation of conditional means. To control these conditional means uniformly without imposing a parametric model, we draw on nonparametric estimation results. For an s-Hölder-smooth function of a d-dimensional feature, the best possible worst-case uniform-error rate is $(\log N/N)^{s/(2s+d)}$ (Stone, 1982). Finite-sample uniform rates for local estimators are developed, for example, by Jiang (2019). Confidence bands with uniform coverage over a function class, and the limits of adaptation across smoothness classes are studied by Low (1997); Chernozhukov et al. (2014). Our construction is deliberately fixed in advance: the smoothness bound, partition, query cover, and tuning parameters are chosen before calibration rather than from the calibration sample.

Within control, Glanzer et al. (2025) provide a particularly close band-based result. Their Corollary 3.2 constructs a nonasymptotic uniform RKHS band for a regression function, and their Theorem 4.1 propagates one-sided bands through a backward recursion to obtain confidence bounds on the optimal value of a finite-horizon problem with binary stop–continue decisions. We instead control the conditional mean uniformly over both the predictively sufficient history summary and the state, decision, and history arguments queried by a multistage information relaxation with vector-valued decisions and hard recourse. Reusing recorded uncertainty realizations avoids requiring the historical data-collection policy to cover the state–decision space. The covering number of the query domain enters the estimation bound logarithmically. On its coverage event, the resulting one-sided band defines a penalty that preserves the information-relaxation lower bound, which we use to certify the optimality gap of any feasible policy selected from training data.

Data-driven multistage optimization and overlap. Recent multistage methods construct policies from sample paths through robust data-driven models, estimated Markov transitions, or RKHS decision rules (Bertsimas et al., 2023; Park et al., 2022; Bertsimas & Villalobos Carballo, 2023). For data-driven solution-quality assessment, the closest work is Yan & Lam (2024), who introduce a multisample version of sample average approximation (SAA) and a multisample single-replication procedure. Under stagewise independence and a parameterized policy class, they establish consistency and asymptotically valid coverage for the gap between a candidate and the best policy in that class.

Our certificate permits dependence across stages through a predictively sufficient history summary and gives finite-sample coverage relative to the unrestricted MSP optimum. It uses the observed exogenous uncertainty and the known model functions, together with certified regularity bounds and certified lower bounds for the perfect-information inner problems. The approaches therefore require different model information and provide different optimality-gap guarantees.

In offline reinforcement learning, outcomes are observed under a behavior policy, so evaluating another policy typically requires overlap: the relevant histories and action neighborhoods must be adequately represented under the data-collection distribution (Kallus & Uehara, 2020). Smoothness may still bound the range of possible values when overlap fails (Khan et al., 2024). Our boundary result applies the same observational-equivalence argument to an MSP optimality gap: the selected policy's cost may be identifiable while $V^\star$ depends on a decision not represented in the observed data. Observed exogenous uncertainty and known model functions avoid this limitation by allowing each recorded realization to be reevaluated at every state–decision query. When the uncertainty is not directly observed, the model functions are unknown, or decisions affect the successor distribution, overlap or extrapolation is again needed.

B Policy Optimization: Full Results and Proofs

B.1 Feasible Policy Optimization

We replay sample paths under the projected policy and differentiate the resulting costs to update its parameters. Projection enforces the stage constraints throughout training, but those constraints must also enter the derivative calculation: an early decision changes both later proposals and the sets onto which they are projected. We first state the training procedure in Algorithm 1, then derive the local projection derivatives and combine them into the complete pathwise gradient.

B.1.1 OPTIMIZING A FEASIBLE POLICY

For one sample path and a parameter at which the realized cost is differentiable, write

$$C(\theta,\xi) := C(x^\theta,\xi) = \sum_{t=1}^{T} c_t(s_t^\theta, x_t^\theta, \xi_{t+1}), \qquad \widehat{g}_{\mathrm{PW}}(\theta;\xi) = \nabla_\theta C(\theta,\xi),$$

so that $J(\theta) = \mathbb{E}[C(\theta,\xi)]$. On a fixed path, θ enters the cost through the proposals and the resulting projections and transitions. When the maps along the trajectory are differentiable, as specified below, reverse-mode automatic differentiation computes $\widehat{g}_{\mathrm{PW}}$ by applying the chain rule backward from total cost to the parameters (Glasserman, 1991). No derivative of the distribution of the exogenous sample path ξ is needed, because that distribution is independent of θ. For any fixed parameter satisfying Theorem B.1, the sampled total cost is nondifferentiable only on a null event, on which we set the sampled gradient to zero.

We can also penalize the distance between each proposal and its deployed decision by adding a proposal-distance regularizer with weight $\lambda \geq 0$:

$$d_{\mathrm{pd}}(\theta,\xi) := \frac{1}{2}\sum_{t=1}^{T} \|u_t^\theta - x_t^\theta\|^2, \qquad \mathcal{D}_{\mathrm{pd}}(\theta) := \mathbb{E}[d_{\mathrm{pd}}(\theta,\xi)], \qquad H_\lambda(\theta) := J(\theta) + \lambda \mathcal{D}_{\mathrm{pd}}(\theta). \tag{B.1}$$

On a realized trajectory, d_{pd} is zero exactly when every proposal is already feasible and therefore equals its deployed decision. Its gradient enters the parameter update, while the deployed policy retains the same projection rule. Setting $\lambda_k = 0$ at every iteration recovers pathwise-gradient optimization of the expected cost J, without the proposal-distance regularizer.

Algorithm 1 (expanded): feasible pathwise policy optimization. Given $\theta_0 \in \Theta$, batch size N_{bat}, step sizes η_k chosen so that the displayed updates remain in Θ, and optional proposal-distance weights $\lambda_k \geq 0$, for $k = 0, \ldots, K-1$:

1. *Sample.* Take the next N_{bat} previously unused full sample paths $\{\xi^{k,i}\}_{i=1}^{N_{\mathrm{bat}}}$ from $\mathcal{D}_{\mathrm{fit}}$.

2. *Replay.* For each path $\xi = \xi^{k,i}$, initialize $s_1 = s_1^0(\xi_1)$ and, for $t = 1, \ldots, T$, form $z_t = \varphi_t(\xi_{1:t})$ and compute

$$u_t = \mu_{\theta_k,t}(s_t, z_t), \qquad x_t = P_t(u_t, b_t(z_t) + B_t s_t),$$
$$s_{t+1} = f_t(s_t, x_t, \xi_{t+1}).$$

   Accumulate $C(\theta_k, \xi^{k,i})$ and $d_{\mathrm{pd}}(\theta_k, \xi^{k,i})$. Each projection is solved exactly as in (A.3).

3. *Differentiate and update.* Compute

$$\widehat{g}_{\lambda_k}(\theta_k;\xi^{k,i}) := \nabla_\theta \big[C(\theta,\xi^{k,i}) + \lambda_k d_{\mathrm{pd}}(\theta,\xi^{k,i})\big]\big|_{\theta=\theta_k}$$

   through the proposals, projection solves, and transitions, including the effects of the state on both proposals and constraint bounds in (B.4). Use the projection Jacobians in Appendix B.1.2 at the differentiability points specified there. Appendix B.1.3 gives the conditions for unbiasedness. Set

$$\theta_{k+1} = \theta_k - \frac{\eta_k}{N_{\mathrm{bat}}} \sum_{i=1}^{N_{\mathrm{bat}}} \widehat{g}_{\lambda_k}(\theta_k;\xi^{k,i}).$$

Save the iterates as *checkpoints* and select one using only information from the training and selection data, including any internal validation split. The selected checkpoint $\widehat{\theta}$ determines the *selected policy* $\widehat{\pi} = \pi_{\widehat{\theta}}$.

For later reference, denote the batch-averaged gradient in Algorithm 1 by

$$G_k := \frac{1}{N_{\mathrm{bat}}} \sum_{i=1}^{N_{\mathrm{bat}}} \widehat{g}_{\lambda_k}(\theta_k; \xi^{k,i}), \qquad \theta_{k+1} = \theta_k - \eta_k G_k. \tag{B.2}$$

Over the full run, Algorithm 1 uses $N_{\mathrm{bat}}K$ full sample paths and $N_{\mathrm{bat}}KT$ projection solves. Its backward pass computes the gradient specified by the sensitivity recursion below, using reverse mode to avoid forming the full sensitivity matrices. Every saved checkpoint remains feasible by Proposition A.1, regardless of the regularization weights. These weights do affect the optimization target: a fixed weight targets H_λ, whereas suitably decreasing weights can retain stationarity guarantees for J. Theorem B.2 in Appendix B.2 covers both decreasing and zero weights. The regularization term can provide an update direction when changes in a proposal leave its projected decision unchanged, but it is not required for the stationarity guarantee.

B.1.2 Projection Jacobian

The projection depends on the policy parameter through its proposal and incoming state. We therefore differentiate it with respect to these two inputs, after substituting the state-dependent bounds:

$$\bar{P}_t(u, s; z) := P_t\big(u, b_t(z) + B_t s\big).$$

Thus $\bar{P}_t$ expresses the projection in its proposal and state inputs while holding the observed history summary fixed. Fix z and suppress the stage index, writing $A \in \mathbb{R}^{k\times m}$. At $x = \bar{P}(u, s; z)$, the *active set* $\mathcal{I} := \{i : (Ax)_i = b_i(z) + (Bs)_i\}$ collects the constraints that hold with equality. Let $A_{\mathcal{I}}$ and $B_{\mathcal{I}}$ denote the submatrices formed by these active rows.

Proposition B.1 (Projection Jacobian). *For fixed z, the feasible state domain $\mathsf{D}_z := \{s : \exists x,\ Ax \le b(z)+Bs\}$ is a polyhedron, and $\bar{P}(\cdot,\cdot; z)$ is continuous and piecewise affine on $\mathbb{R}^m \times \mathsf{D}_z$. It is locally Lipschitz and differentiable almost everywhere on the interior of that domain. At any differentiability point (u, s) with $s \in \operatorname{int} \mathsf{D}_z$, its Jacobians are*

$$D_u\bar{P} = I_m - A_{\mathcal{I}}^{+} A_{\mathcal{I}}, \qquad D_s\bar{P} = A_{\mathcal{I}}^{+} B_{\mathcal{I}}, \tag{B.3}$$

where $A_{\mathcal{I}}^{+}$ is the Moore–Penrose pseudoinverse. In particular, $D_u\bar{P}$ is the orthogonal projector onto $\ker A_{\mathcal{I}}$, so $\|D_u\bar{P}\| \le 1$. For an empty active set, the formulas mean $D_u\bar{P} = I_m$ and $D_s\bar{P} = 0$.

The interior condition means that nearby states also admit feasible decisions, but it does not require the decision set itself to have an interior, so balance equalities are allowed. At such a differentiability point, the first Jacobian retains proposal changes that preserve the active equalities, while the second describes how the decision responds to changes in the state-dependent bounds. Only active rows enter these formulas, and the pseudoinverse accommodates linear dependence among them. Working directly with $\bar{P}$ also respects the linked changes in the bounds: a derivative with respect to independently varying every component of q may fail to exist even when the derivative with respect to s exists.

The piecewise-affine structure is standard for strictly convex parametric quadratic programs (Tøndel et al., 2003). Appendix B.3 proves the formula and illustrates the distinction between dependent active rows and genuine corners of the projection map. Because a pseudoinverse does not remove these corners, the trajectory recursion requires differentiability at the sampled inputs, not merely almost-everywhere differentiability over all possible inputs.

B.1.3 Policy-trajectory sensitivities and unbiasedness

Let $S_t := D_\theta s_t \in \mathbb{R}^{n_t \times d_\theta}$ and $X_t := D_\theta x_t \in \mathbb{R}^{m_t \times d_\theta}$ be the sensitivities of the state and of the deployed decision to the parameter along a sampled policy trajectory. Since s_1 and z_t do not depend on θ, $S_1 = 0$ and $D_\theta z_t = 0$. Along a sampled trajectory where the proposal, transition, and cost maps are differentiable and each state-dependent projection $\bar{P}_t(\cdot,\cdot; z_t)$ is differentiable at (u_t, s_t) in

an open feasible neighborhood as in Proposition B.1, the chain rule through (A.3) and (A.4) gives

$$X_t = D_u\bar{P}_t\, D_\theta\mu_{\theta,t} + \Big(\underbrace{D_u\bar{P}_t\, D_s\mu_{\theta,t}}_{\text{through the proposal}} + \underbrace{D_s\bar{P}_t}_{\text{through the bounds}} \Big) S_t, \tag{B.4}$$

$$S_{t+1} = D_s f_t\, S_t + D_x f_t\, X_t, \tag{B.5}$$

$$\widehat{g}_{\mathrm{PW}} = \sum_{t=1}^{T} \left(S_t^\top \nabla_s c_t + X_t^\top \nabla_x c_t \right), \tag{B.6}$$

with $D_u\bar{P}_t, D_s\bar{P}_t$ evaluated at $(u_t, s_t; z_t)$ and $D_\theta\mu_{\theta,t}, D_s\mu_{\theta,t}$ at (s_t, z_t). Together, these equations compute $\nabla_\theta C(\theta, \xi)$ recursively along the policy trajectory on one sample path. The first term in (B.4) captures the direct change in the proposal, with $D_\theta\mu_{\theta,t}$ holding the state and history summary fixed. The bracket accounts for how state changes caused by earlier decisions affect both the proposal and the constraint bounds. Equation (B.5) passes both effects to the next stage. The moving-recourse derivative is explicitly $D_s\bar{P}_t = A^+_{t,\mathcal{I}_t} B_{t,\mathcal{I}_t}$, where $\mathcal{I}_t$ is the stage-t active set. Computing it directly retains the linked changes in the bounds $q_t(s_t, z_t) = b_t(z_t) + B_t s_t$.

Holding the bounds fixed when differentiating would instead remove $D_s\bar{P}_t S_t$ from the decision derivative without changing the simulated decisions or costs. Because the resulting state sensitivity is carried to later stages, the gradient error includes both the current omission and the accumulated effects of earlier omissions. Corollary B.2 in Appendix B.6 bounds these two components. Increasing the minibatch size does not generally remove this error, which arises from the derivative calculation rather than sampling noise.

The same sensitivities also give the derivative of the proposal-distance term. Let

$$U_t := D_\theta u_t = D_\theta\mu_{\theta,t} + D_s\mu_{\theta,t} S_t$$

denote the proposal's full parameter derivative, including its dependence on the state. At these differentiability points, we use U_t and the full parameter derivative X_t of the deployed decision from (B.4). The latter includes the response to the state-dependent constraint bounds. Together they give the regularized pathwise gradient

$$\widehat{g}_\lambda(\theta; \xi) = \underbrace{\widehat{g}_{\mathrm{PW}}(\theta; \xi)}_{\text{pathwise cost gradient}} + \underbrace{\lambda \sum_{t=1}^{T} (U_t - X_t)^\top (u_t - x_t)}_{\text{weighted proposal-distance gradient}}. \tag{B.7}$$

Equation (B.7) is the sampled gradient used in Algorithm 1, with $\widehat{g}_0(\theta; \xi) = \widehat{g}_{\mathrm{PW}}(\theta; \xi)$. Its dependence on the moving bounds enters through the term $D_s\bar{P}_t S_t$ in X_t. To make that dependence explicit, note that projection optimality makes the residual $u_t - x_t$ orthogonal to directions that preserve the active equalities, and $D_u\bar{P}_t$ projects onto those directions. Thus $(D_u\bar{P}_t)^\top (u_t - x_t) = 0$ and (B.7) can be written as

$$\widehat{g}_\lambda = \widehat{g}_{\mathrm{PW}} + \lambda \sum_{t=1}^{T} \left[U_t^\top (u_t - x_t) - S_t^\top (D_s\bar{P}_t)^\top (u_t - x_t) \right]. \tag{B.8}$$

The first term inside the brackets is the derivative obtained by treating the deployed decision as constant in the proposal-distance term, while the second corrects for the moving constraint bounds. This correction vanishes whenever $D_s\bar{P}_t = 0$, in particular when $B_t = 0$ and the bounds are state-independent, but is generally present when the state moves those bounds.

To establish unbiasedness of the sampled gradients, we prove the expectation identities for J and then for H_λ under standard pathwise-differentiation conditions (Glasserman, 1991). For almost every sample path, local Lipschitz continuity in θ ensures differentiability of the sampled cost at Lebesgue-almost every parameter. An integrable pathwise Lipschitz bound then permits differentiation under the expectation. The resulting identities describe the objective at a fixed parameter. The conditions needed along a sequence of training iterates are addressed in Appendix B.2.

Theorem B.1 (Almost-everywhere differentiability and unbiasedness)**.** *Suppose that, for almost every ξ, $C(\cdot, \xi)$ is locally Lipschitz on an open parameter domain and that the set of nondifferentiability pairs (θ, ξ) is jointly measurable. Then, for Lebesgue-almost every fixed θ, $C(\cdot, \xi)$ is*

differentiable at θ almost surely. Fix any such θ. If $\mathbb{E}|C(\theta,\xi)| < \infty$ and, on a common deterministic neighborhood of θ,

$$|C(\theta_1,\xi) - C(\theta_2,\xi)| \le \Lambda_C(\xi)\|\theta_1 - \theta_2\|, \qquad \mathbb{E}[\Lambda_C(\xi)] < \infty,$$

then J is differentiable at θ and

$$\nabla J(\theta) = \mathbb{E}[\widehat{g}_{\mathrm{PW}}(\theta;\xi)]. \tag{B.9}$$

The projection need not be globally smooth for this expectation identity to hold under the stated assumptions. In particular, different constraints may be active on different trajectories or at different parameter values.

Corollary B.1 (Unbiased regularized pathwise gradient)**.** *Fix $\lambda \ge 0$. Suppose the assumptions of Theorem B.1 hold with the regularized sampled objective $C(\theta,\xi) + \lambda d_{\mathrm{pd}}(\theta,\xi)$ in place of $C(\theta,\xi)$. Define $\widehat{g}_\lambda$ as the derivative of this regularized sampled objective wherever that derivative exists, and set it to zero on the almost-sure null event of nondifferentiability. Then, for Lebesgue-almost every parameter satisfying these hypotheses,*

$$\nabla H_\lambda(\theta) = \mathbb{E}[\widehat{g}_\lambda(\theta;\xi)].$$

If, in addition, the proposal, transition, and cost maps are differentiable and, almost surely at that parameter, every encountered $\bar{P}_t(\cdot,\cdot;z_t)$ is differentiable at (u_t,s_t) in an open feasible neighborhood as in Proposition B.1, then this sample derivative can be computed from (B.7)*, using the sensitivities in* (B.4)–(B.5)*.*

The corollary separates two requirements: the derivative of the complete sampled loss must have the correct expectation, and the intermediate maps must be differentiable for the recursion to compute it. The first does not imply the second, since sampled projection inputs can concentrate at points where the projection is nondifferentiable. Moreover, the almost-everywhere identity at fixed parameters does not establish smoothness of J or conditional unbiasedness at data-dependent iterates. Appendix B.2 states these training conditions explicitly.

For nonsmooth objectives, Appendix B.10 describes a Gaussian parameter-smoothing alternative with its own differentiation conditions. It averages costs over nearby parameters, whereas the proposal-distance term guides proposals toward their deployed decisions. The experiments do not use parameter smoothing.

B.2 Guarantees for Policy Optimization

The pathwise gradient supplies an update direction, and we now examine what those updates achieve. Under the smoothness and sampling conditions below, SGD with zero or suitably decreasing regularization weights approaches stationarity, as measured by the expected squared norm of ∇J. We first bound this quantity in terms of the training budget and then explain how the constants depend on the multistage model and horizon. Stationarity concerns the parameterized policy class, whereas certification in Appendix D.1 addresses the optimality gap relative to the unrestricted MSP optimum.

B.2.1 Convergence of the policy optimization algorithm

Relative to a descent step for J, the update contains both sampling noise and the gradient of the proposal-distance regularizer. Decreasing the weight lets the regularizer influence proposals early in training while its contribution to later updates vanishes. To quantify progress toward stationarity, we use the standard SGD assumptions of a smooth lower-bounded objective and controlled conditional mean and variance (Ghadimi & Lan, 2013): smoothness bounds the cost change after each update, while the conditional mean and variance bounds control the regularizer gradient and sampling noise. Unlike the almost-everywhere identity in Theorem B.1, these conditions are imposed at the actual training iterates. Write $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{G}_k^{\mathrm{tr}}]$ for conditional expectation given the pre-update training information. Here $\mathcal{G}_k^{\mathrm{tr}}$ records the iterates, previously used minibatches, and training randomness before iteration k, excluding the current fresh minibatch.

Theorem B.2 (Stationarity with zero or vanishing regularization)**.** *For the updates* (B.2)*, let θ_0 be deterministic. Suppose J is bounded below by a finite constant $J_{\inf}$ and has an L_J-Lipschitz gradient on an open convex domain containing every iterate and every line segment between consecutive*

iterates. Let $\{\eta_k\}_{k\geq 0}$ and $\{\lambda_k\}_{k\geq 0}$ be nonnegative deterministic sequences fixed before training, with $\eta_k \leq 1/L_J$. Suppose, for some finite $G_D, \sigma \geq 0$,

$$\mathbb{E}_k G_k = \nabla J(\theta_k) + b_k, \qquad \|b_k\| \leq \lambda_k G_D, \qquad \mathbb{E}_k\|G_k - \mathbb{E}_k G_k\|^2 \leq \frac{\sigma^2}{N_{\mathrm{bat}}} \quad \textit{almost surely}. \tag{B.10}$$

For $S_K = \sum_{k=0}^{K-1} \eta_k > 0$, choose $\hat{k}$ independently of training with $\mathbb{P}\{\hat{k} = k\} = \eta_k / S_K$. Then

$$\begin{aligned}\mathbb{E}\|\nabla J(\theta_{\hat{k}})\|^2 &= \frac{1}{S_K}\sum_{k=0}^{K-1} \eta_k \mathbb{E}\|\nabla J(\theta_k)\|^2 \\ &\leq \underbrace{\frac{2[J(\theta_0) - J_{\inf}]}{S_K}}_{\textit{initial cost gap}} + \underbrace{\frac{G_D^2}{S_K}\sum_{k=0}^{K-1}\eta_k\lambda_k^2}_{\textit{regularization}} + \underbrace{\frac{L_J\sigma^2}{N_{\mathrm{bat}}S_K}\sum_{k=0}^{K-1}\eta_k^2}_{\textit{sampling noise}}.\end{aligned} \tag{B.11}$$

In particular, this expected squared gradient tends to zero if $S_K \to \infty$, $\sum_{k=0}^{K-1}\eta_k\lambda_k^2/S_K \to 0$, and $\sum_{k=0}^{K-1}\eta_k^2/S_K \to 0$.

For Algorithm 1 with fresh independent minibatches of exogenous paths from $\mathcal{D}_{\mathrm{fit}}$, exact sampled derivatives, and valid differentiation under the expectation, $b_k = \lambda_k \nabla \mathcal{D}_{\mathrm{pd}}(\theta_k)$ at positive-weight iterates, while $b_k = 0$ at zero-weight iterates, so G_D is needed only where the regularization term is used. If $S_K \to \infty$ and $\lambda_k \to 0$, the initial-cost-gap and regularization contributions in (B.11) vanish. Convergence to stationarity then requires the stepsizes and batch size to make the sampling-noise contribution vanish as well. Appendix B.7 gives the proof.

For an explicit rate, fix a training budget K and choose

$$\eta_k = \frac{c}{\sqrt{K}}, \qquad \lambda_k = \lambda_0(k+1)^{-1/4}, \qquad 0 < c \leq 1/L_J, \quad \lambda_0 \geq 0.$$

The stepsize is constant within the run, while the regularization weight decreases. With $\Delta_0 = J(\theta_0) - J_{\inf}$, Corollary B.3 gives

$$\mathbb{E}\|\nabla J(\theta_{\hat{k}})\|^2 \leq \frac{2\Delta_0/c + 2G_D^2\lambda_0^2 + L_J\sigma^2 c/N_{\mathrm{bat}}}{\sqrt{K}}, \tag{B.12}$$

where $\hat{k}$ is now uniform on $\{0, \ldots, K-1\}$. Regularization therefore need not change the $O(K^{-1/2})$ rate or the stationarity target J. Setting $\lambda_0 = 0$ recovers unregularized policy optimization without assumptions on $\mathcal{D}_{\mathrm{pd}}$. Its detailed specialization is in Appendix B.7.

For fixed batch size, schedule parameters, and problem constants, (B.12) gives a sufficient budget of $N_{\mathrm{fit}} = N_{\mathrm{bat}}K = O(\epsilon^{-2})$ full sample paths to achieve $\mathbb{E}\|\nabla J(\theta_{\hat{k}})\|^2 \leq \epsilon$. The budget controls the expected squared gradient at a randomly selected iterate under Algorithm 1's fresh-minibatch SGD scheme. The experimental implementation uses Adam on fixed path pools and validation-based checkpoint selection, as detailed in Section 5.

An important feature of this training bound is that policy optimization uses pathwise gradients directly, without estimating conditional expectations. Consequently, with the policy class, batch size, schedule parameters, and constants in (B.12) fixed, the uncertainty dimension does not change the $O(\epsilon^{-2})$ sample-size exponent, although it may affect policy size and sensitivity constants. The kernel-based data-driven SDDP method of Park et al. (2022), in contrast, estimates conditional expectations at a rate that slows as the uncertainty dimension grows. The comparison concerns different accuracy criteria: their guarantee assesses a learned first-stage decision followed by optimal recourse, while ours controls the gradient of the represented policy's cost. It therefore does not establish a performance ordering, nor does the training rate carry over to our conditional-mean certificate, whose estimation rate remains dimension dependent (Corollary C.1).

B.2.2 How the planning horizon affects the gradient

The constants G_D and σ depend on how parameter perturbations propagate through the MSP. A change in the current decision alters the next state, which in turn changes later proposals and feasible

sets. Repeating this process can amplify the pathwise gradient as the horizon grows. We quantify this effect first for the pathwise cost gradient $\widehat{g}_{\mathrm{PW}}$ and then for the gradient of the regularizer. Since the analysis concerns a trajectory at a given parameter, it does not depend on the regularization weights used to obtain that parameter.

To bound the accumulated effect, consider how much the state sensitivity from earlier stages can grow in a single transition. Let L_f^s, L_f^x, L_μ^s, and L_{MR} bound, respectively, $\|D_s f_t\|$, $\|D_x f_t\|$, $\|D_s \mu_{\theta,t}\|$, and $\|D_s \bar{P}_t\|$ uniformly over the parameter and sample-path pairs under consideration. Then

$$\rho = L_f^s + L_f^x (L_\mu^s + L_{\mathrm{MR}}) \tag{B.13}$$

bounds the propagation of an existing state sensitivity: the first term comes directly through the dynamics, and the second through the current decision. The moving-bound response contributes through L_{MR}. Appendix B.5 records the remaining derivative bounds and the explicit gradient bound Γ_T.

Theorem B.3 (Horizon-dependent bound on the gradient norm)**.** *Suppose the proposal, transition, and cost maps are differentiable and the state-dependent projection maps $\bar{P}_t(\cdot, \cdot; z_t)$ are differentiable at the encountered (u_t, s_t) in open feasible neighborhoods as in Proposition B.1, with the uniform bounds in* (B.15)*. Then $\|\widehat{g}_{\mathrm{PW}}\| \le \Gamma_T$, where Γ_T is given in* (B.16)*. If these conditions hold almost surely, then $\mathbb{E}\|\widehat{g}_{\mathrm{PW}}\|^2 \le \Gamma_T^2$. Holding all constants in* (B.15) *fixed as T varies,*

$$\Gamma_T^2 = \begin{cases} O(d_\theta T^2), & 0 \le \rho < 1, \\ O(d_\theta T^4), & \rho = 1, \\ O(d_\theta \rho^{2T}), & \rho > 1. \end{cases} \tag{B.14}$$

When $\rho < 1$, the bound on an earlier state perturbation decays as it passes through later stages. The total gradient can nevertheless grow because new parameter effects enter at each stage and the objective sums T cost derivatives, resulting in the quadratic squared-gradient bound. At $\rho = 1$, inherited effects can accumulate, giving a quartic bound, while for $\rho > 1$, repeated amplification can instead produce exponential growth. These regimes describe worst-case bounds, not the gradients on typical sampled paths.

Projection onto a fixed set does not amplify differences between proposals, yet the full trajectory also transmits perturbations through the dynamics, the proposal's state dependence, and the moving-bound response $D_s \bar{P}_t$. These additional effects explain why projection alone does not prevent amplification. The moving-bound response can be large when the active-row matrix has small positive singular values and the bounds move in the corresponding directions, as the pseudoinverse formula in Proposition B.1 shows.

For the regularized gradient, a uniform bound on proposal distances gives

$$\|\widehat{g}_{\lambda_k}\| \le \Gamma_T + \lambda_k \Delta_T,$$

where Δ_T bounds the sampled gradient of the proposal-distance regularizer (Corollary B.8). When the corollary's sensitivity and distance bounds hold throughout the update domain, and differentiation under the expectation is valid there, fresh independent minibatches allow $G_D = \Delta_T$ and $\sigma^2 = (\Gamma_T + \lambda_0 \Delta_T)^2$ for the decreasing-weight schedule in Appendix B.2.1, where $\lambda_k \le \lambda_0$. Bounds on both gradients enter the training rate, and L_J may also depend on T.

Together, these bounds explain how the multistage sensitivities affect the stationarity guarantee. They do not, however, determine the selected policy's optimality gap relative to the unrestricted MSP optimum. Assessing that gap requires the separate lower benchmark developed in Appendix D.1, whose validity does not depend on policy optimization having reached a stationary point.

B.3 Derivation of the projection Jacobian

We first establish the piecewise-affine structure and then identify the derivative wherever it exists. Fix z and write $b = b(z)$. The feasible state domain D_z is a polyhedron because it is the projection of the polyhedron $\{(s, x) : Ax \le b + Bs\}$ onto the state coordinates. For each linearly independent row subset J of A, define

$$\lambda_J(u, s) = (A_J A_J^\top)^{-1}(A_J u - b_J - B_J s), \qquad F_J(u, s) = u - A_J^\top \lambda_J(u, s).$$

For $J = \varnothing$, set $F_J(u, s) = u$ and omit the multiplier condition below. The projection optimality conditions imply $\bar{P}(u, s; z) = F_J(u, s)$ on the closed polyhedral region

$$\mathsf{C}_J = \{(u, s) : \lambda_J(u, s) \geq 0,\ AF_J(u, s) \leq b + Bs\}.$$

These finitely many regions cover $\mathbb{R}^m \times \mathsf{D}_z$: the projection residual is a nonnegative combination of active constraint normals, and such a combination can be represented using a linearly independent subset of those normals. Uniqueness of the projection makes the affine formulas agree on overlaps, so this finite closed cover establishes continuity and the piecewise-affine representation. Subdividing a feasible line segment at the region boundaries also gives a Lipschitz bound equal to the maximum norm of the finitely many linear parts $[I_m - A_J^+ A_J,\ A_J^+ B_J]$. Rademacher's theorem then gives almost-everywhere differentiability on the interior.

Now fix a differentiability point (u, s) in an open feasible neighborhood. Let $H = A_{\mathcal{I}}$ contain all rows active at that point and let $N = I_m - H^+ H$. Each active slack $b_i + (Bs)_i - (A\bar{P}(u, s; z))_i$ is nonnegative nearby and zero at the point. Its derivative must therefore vanish, giving

$$HD_u\bar{P} = 0, \qquad HD_s\bar{P} = B_{\mathcal{I}}.$$

Continuity keeps initially inactive constraints inactive in a sufficiently small neighborhood. Every nearby projection residual consequently belongs to the span of the rows of H, even if some initially active constraints become inactive. Hence $N(u - \bar{P}(u, s; z)) = 0$ throughout that neighborhood, and differentiation gives

$$ND_u\bar{P} = N, \qquad ND_s\bar{P} = 0.$$

Combining these identities yields

$$D_u\bar{P} = N, \qquad D_s\bar{P} = (I_m - N)D_s\bar{P} = H^+ HD_s\bar{P} = H^+ B_{\mathcal{I}},$$

which proves (B.3). The matrix N is the orthogonal projector onto $\ker H$. Projection optimality also gives $N^\top(u - x) = 0$, as used in (B.8). For $\mathcal{I} = \varnothing$, these arguments use $N = I_m$ and give $D_s\bar{P} = 0$.

Dependent rows versus nondifferentiability. Consider the balance equality $x = s$, encoded as $x \leq s$ and $-x \leq -s$. Both rows are active for every proposal, yet $\bar{P}(u, s) = s$ is differentiable everywhere. Here $A_{\mathcal{I}} = B_{\mathcal{I}} = [1, -1]^\top$, and the pseudoinverse $A_{\mathcal{I}}^+ = \frac{1}{2}[1, -1]$ gives $D_u\bar{P} = 0$ and $D_s\bar{P} = 1$. Every state admits a feasible decision, although its feasible decision set is a singleton. In contrast, for the constraints $x \leq s$ and $x \leq -s$, the projection of $u = 1$ is $\bar{P}(1, s) = -|s|$ near zero. Its state derivative does not exist at zero. The pseudoinverse is now $\frac{1}{2}[1, 1]$, whose product with $B_{\mathcal{I}} = [1, -1]^\top$ is zero but does not produce a derivative. Thus, the formula covers dependent active rows at differentiability points, not arbitrary active-set boundaries. The boundary of the feasible-state domain D_z is a separate issue: nearby states need not all admit feasible decisions. At such states, or when D_z is lower dimensional, derivatives along feasible state directions require a separate formulation and are not covered by the proposition. This restriction concerns the state domain, not active constraints at the projected decision, which the formula already includes.

Observed recourse coefficients. The local derivative formulas also apply when A_t and B_t are known functions of the observed history summary z_t, since these coefficients remain fixed when a sample path is differentiated with respect to θ. The same formulas apply to $\bar{P}_t(u, s; z_t)$ with these observed coefficients substituted in. Differentiability at the sampled inputs and the uniform sensitivity bounds must still hold. Dependence on the unobserved successor uncertainty realization is excluded by the decision timing.

Numerical scope. An exact minimizer is covered by the proposition. A solver-certified feasible but nonoptimal output remains implementable, but it defines a different policy and its derivative is not covered here. A positive primal-feasibility tolerance gives approximate rather than exact feasibility. At a differentiability point covered by the proposition, the pseudoinverse can be computed by a rank-revealing factorization. At a nondifferentiable point, neither this formula nor an arbitrary generalized derivative is justified as a classical derivative. Null-probability exceptional paths can be assigned any finite gradient without changing an expectation. Positive-probability nondifferentiability is not covered by this result. The smoothing alternative below requires the perturbed backward pass to return a classical derivative almost surely.

B.4 Proof of Theorem B.1

Fix a sample path in the probability-one set on which $\theta \mapsto C(\theta, \xi)$ is locally Lipschitz. Rademacher's theorem gives differentiability for Lebesgue-almost every parameter on this path. Joint measurability of the nondifferentiability set and Fubini's theorem reverse the quantifiers: there is a full-Lebesgue-measure set Θ_{reg} such that, at every fixed $\theta \in \Theta_{\mathrm{reg}}$, the pathwise derivative exists almost surely. When all encountered proposal, transition, and cost maps are differentiable and each encountered $\bar{P}_t(\cdot, \cdot; z_t)$ is differentiable at (u_t, s_t) in an open feasible neighborhood as in Proposition B.1, the classical chain rule is exactly (B.4)–(B.6).

Now fix $\theta \in \Theta_{\mathrm{reg}}$ satisfying the theorem's integrability and local-domination conditions. Differentiability and the local Lipschitz bound imply $\|\widehat{g}_{\mathrm{PW}}(\theta; \xi)\| \le \Lambda_C(\xi)$ almost surely. For sufficiently small h,

$$R(h, \xi) := C(\theta + h, \xi) - C(\theta, \xi) - \langle \widehat{g}_{\mathrm{PW}}(\theta; \xi), h \rangle$$

satisfies $R(h, \xi)/\|h\| \to 0$ almost surely and $|R(h, \xi)|/\|h\| \le 2\Lambda_C(\xi)$. Dominated convergence, applied along every sequence $h \to 0$, gives

$$J(\theta + h) - J(\theta) = \langle \mathbb{E}[\widehat{g}_{\mathrm{PW}}(\theta; \xi)], h \rangle + o(\|h\|).$$

This proves differentiability and (B.9). The law of ξ contributes no score term because it is independent of θ.

B.5 Uniform bounds and proof of Theorem B.3

To obtain an explicit deterministic bound with horizon dependence, assume that the following standard sensitivity bounds hold uniformly on the relevant path domains at differentiability points:

$$\begin{aligned} &f_t : \|D_s f_t\| \le L_f^s,\ \|D_x f_t\| \le L_f^x, && \mu_{\theta,t} : \|D_s \mu_{\theta,t}\| \le L_\mu^s,\ \|D_\theta \mu_{\theta,t}\| \le L_\mu^\theta \sqrt{d_\theta}, \\ &\bar{P}_t : \|D_u \bar{P}_t\| \le 1,\ \|D_s \bar{P}_t\| \le L_{\mathrm{MR}}, && c_t : \|\nabla_s c_t\| \le L_c^s,\ \|\nabla_x c_t\| \le L_c^x. \end{aligned} \tag{B.15}$$

For the fixed matrices A_t, B_t, a finite L_{MR} can be chosen as the maximum of $\|A_{t,\mathcal{I}}^+ B_{t,\mathcal{I}}\|$ over stages and row subsets, with the empty-subset value set to zero. Proposition B.1 then bounds the state derivative at every differentiability point it covers. With ρ defined in (B.13), set

$$\chi = L_f^x L_\mu^\theta, \qquad \mathcal{H}_j(\rho) = \sum_{\ell=0}^{j-1} \rho^\ell, \qquad \mathcal{H}_0(\rho) = 0.$$

The geometric sum records how the bound on inherited state sensitivity accumulates across stages. The resulting explicit cost-gradient bound is

$$\|\widehat{g}_{\mathrm{PW}}\| \le \sqrt{d_\theta} \left[T L_c^x L_\mu^\theta + (L_c^s + L_c^x (L_\mu^s + L_{\mathrm{MR}})) \chi \sum_{t=1}^{T} \mathcal{H}_{t-1}(\rho) \right] =: \Gamma_T. \tag{B.16}$$

The first term inside the brackets bounds the direct effect of each proposal on its stage cost, while the second accounts for effects transmitted through states generated by earlier decisions.

Proof. Recall that $S_t = D_\theta s_t$ and $X_t = D_\theta x_t$. Since $D_\theta z_t = 0$, the chain rule and (B.15) imply

$$\|X_t\| \le L_\mu^\theta \sqrt{d_\theta} + (L_\mu^s + L_{\mathrm{MR}}) \|S_t\|, \qquad \|S_{t+1}\| \le L_f^s \|S_t\| + L_f^x \|X_t\| \le \rho \|S_t\| + \chi \sqrt{d_\theta}.$$

Because $S_1 = 0$, induction gives

$$\|S_t\| \le \chi \sqrt{d_\theta} \mathcal{H}_{t-1}(\rho).$$

The parameter derivative of the stage-t cost is therefore bounded by

$$\begin{aligned} \left\| \nabla_\theta [c_t(s_t^\theta, x_t^\theta, \xi_{t+1})] \right\| &\le L_c^s \|S_t\| + L_c^x \|X_t\| \\ &\le \sqrt{d_\theta} \left[L_c^x L_\mu^\theta + (L_c^s + L_c^x (L_\mu^s + L_{\mathrm{MR}})) \chi \mathcal{H}_{t-1}(\rho) \right]. \end{aligned}$$

Summing over t proves (B.16). If $\rho < 1$, $\mathcal{H}_j(\rho)$ is uniformly bounded; if $\rho = 1$, it equals j; and if $\rho > 1$, it is $O(\rho^j)$. Squaring the resulting bounds proves the three regimes in (B.14).

□

B.6 Error from omitting moving recourse

To measure the effect of the moving constraint bounds, compute a policy trajectory and compare its complete gradient with an incomplete calculation, denoted $\widehat{g}_{\mathrm{noMR}}$. In this calculation, delete $D_s \bar{P}_t S_t$ from the decision derivative at every stage, then use the modified sensitivities at later stages. All Jacobians are evaluated at the original states and decisions, so the forward decisions and costs are unchanged. The next result bounds the resulting gradient error.

Corollary B.2 (Error from omitting moving recourse)**.** *Under Theorem B.3, let $\rho_0 = L_f^s + L_f^x L_\mu^s$, and, with an empty sum interpreted as zero, define*

$$\mathcal{E}_t := L_f^x L_{\mathrm{MR}} \chi \sqrt{d_\theta} \sum_{j=1}^{t-1} \rho_0^{t-1-j} \mathcal{H}_{j-1}(\rho).$$

For each parameter and sample path satisfying the preceding differentiability conditions,

$$\|\widehat{g}_{\mathrm{PW}} - \widehat{g}_{\mathrm{noMR}}\| \le \sum_{t=1}^{T} \Big[(L_c^s + L_c^x L_\mu^s)\mathcal{E}_t + L_c^x L_{\mathrm{MR}} \chi \sqrt{d_\theta} \mathcal{H}_{t-1}(\rho) \Big]. \tag{B.17}$$

In particular, the error is zero when $L_{\mathrm{MR}} = 0$. For $L_{\mathrm{MR}} > 0$, the bound accounts for propagation of earlier omissions, but it does not assert that realized error is monotone in T.

The bound separates the current omission from its later consequences: $\mathcal{E}_t$ bounds the state-sensitivity error inherited from earlier stages, while the second term in (B.17) bounds the newly omitted response at stage t. Thus correct forward decisions and costs do not guarantee a correct training gradient. This is a derivative error, not sampling noise, so increasing the minibatch size does not generally remove it.

Proof. Let $\widetilde{S}_t, \widetilde{X}_t$ denote the sensitivities computed after deleting $D_s \bar{P}_t S_t$, without changing the forward simulation. With $e_t = \|S_t - \widetilde{S}_t\|$ and $e_1 = 0$,

$$\begin{aligned} \|X_t - \widetilde{X}_t\| &\le L_\mu^s e_t + L_{\mathrm{MR}} \|S_t\|, \\ e_{t+1} &\le L_f^s e_t + L_f^x \|X_t - \widetilde{X}_t\| \le \rho_0 e_t + L_f^x L_{\mathrm{MR}} \|S_t\|. \end{aligned}$$

The state-sensitivity bound from Theorem B.3 and induction give

$$e_t \le L_f^x L_{\mathrm{MR}} \chi \sqrt{d_\theta} \sum_{j=1}^{t-1} \rho_0^{t-1-j} \mathcal{H}_{j-1}(\rho) = \mathcal{E}_t.$$

Finally, subtracting the two versions of (B.6) and applying the cost-gradient bounds gives

$$\|\widehat{g}_{\mathrm{PW}} - \widehat{g}_{\mathrm{noMR}}\| \le \sum_{t=1}^{T} \left[(L_c^s + L_c^x L_\mu^s) e_t + L_c^x L_{\mathrm{MR}} \|S_t\| \right].$$

Substituting the bounds on e_t and $\|S_t\|$ proves (B.17). This modified backward calculation need not give the gradient of any scalar objective. The bound concerns the calculation obtained by deleting the specified derivative term.

□

B.7 Proof and specializations of the stationarity bound

We first prove Theorem B.2, then derive the decreasing-weight rate and the unregularized specialization.

Proof. Write $g_k = \nabla J(\theta_k)$. Smoothness and (B.10) give

$$\mathbb{E}_k[J(\theta_{k+1})] \le J(\theta_k) - \eta_k g_k^\top (g_k + b_k) + \frac{L_J \eta_k^2}{2} \|g_k + b_k\|^2 + \frac{L_J \eta_k^2 \sigma^2}{2 N_{\mathrm{bat}}}.$$

The identity

$$-\eta_k g_k^\top (g_k + b_k) + \frac{L_J \eta_k^2}{2}\|g_k + b_k\|^2 = -\frac{\eta_k}{2}\|g_k\|^2 + \frac{\eta_k}{2}\|b_k\|^2 - \frac{\eta_k(1 - L_J\eta_k)}{2}\|g_k + b_k\|^2$$

and $\eta_k \le 1/L_J$ therefore imply

$$\mathbb{E}_k[J(\theta_{k+1})] \le J(\theta_k) - \frac{\eta_k}{2}\|g_k\|^2 + \frac{\eta_k \lambda_k^2 G_D^2}{2} + \frac{L_J \eta_k^2 \sigma^2}{2N_{\mathrm{bat}}}.$$

Taking expectations, summing over k, and using $J(\theta_K) \ge J_{\inf}$ proves the bound. The distribution of $\hat{k}$ identifies the weighted average. □

B.7.1 A DECREASING-WEIGHT SCHEDULE

The schedule used in Appendix B.2 balances the regularization and sampling terms at order $K^{-1/2}$.

Corollary B.3 (Finite-budget rate with decreasing regularization)**.** *Under Theorem B.2, fix a budget $K \ge 1$ and set*

$$\eta_k = \frac{c}{\sqrt{K}}, \qquad \lambda_k = \lambda_0 (k+1)^{-1/4}, \qquad 0 < c \le 1/L_J, \quad \lambda_0 \ge 0.$$

With $\Delta_0 = J(\theta_0) - J_{\inf}$ and $\hat{k}$ uniform on $\{0, \dots, K-1\}$, independently of training, the bound (B.12) *holds.*

Proof. Here $S_K = c\sqrt{K}$ and $\sum_{k<K} \eta_k^2 = c^2$. Since $\sum_{k=0}^{K-1}(k+1)^{-1/2} \le 2\sqrt{K}$,

$$\sum_{k=0}^{K-1} \eta_k \lambda_k^2 = \frac{c\lambda_0^2}{\sqrt{K}} \sum_{k=0}^{K-1} (k+1)^{-1/2} \le 2c\lambda_0^2.$$

Substitute these bounds into (B.11). □

The stepsize is constant within each run, while the regularization weight decreases at every update. If the training budget is not fixed in advance, one may instead use $\eta_k = c(k+1)^{-1/2}$ with the same λ_k. Then (B.11) gives $O((1 + \log K)/\sqrt{K})$ for an iterate selected with probabilities η_k / S_K, rather than uniformly. These are stationarity guarantees for J, but they do not assert that regularization improves the cost. A fixed positive weight instead permits a stationarity analysis of H_λ under its own smoothness, lower-bound, and unbiased-gradient assumptions.

B.7.2 UNREGULARIZED POLICY OPTIMIZATION

The next specialization connects the stochastic-gradient conditions to the fresh sample paths and pathwise derivatives used by Algorithm 1. Its assumptions do not involve the proposal-distance term.

Corollary B.4 (Stationarity of unregularized policy optimization)**.** *Under the objective and update-domain assumptions of Theorem B.2, run Algorithm 1 from deterministic θ_0 with $\lambda_k = 0$ for every k. Suppose Theorem B.1 holds at every iterate and, conditionally on the pre-update training information $\mathcal{G}_k^{\mathrm{tr}}$, the N_{bat} full sample paths are independent and identically distributed (i.i.d.) draws from the sample-path distribution. Assume additionally that the backward pass in Algorithm 1 returns the exact sampled derivative almost surely at every iterate. This holds, for example, when the proposal, transition, and cost maps are differentiable and every encountered $\bar{P}_t(\cdot, \cdot; z_t)$ is differentiable at (u_t, s_t) in an open feasible neighborhood as in Proposition B.1, almost surely at every iterate. Finally, for a fresh exogenous path ξ independent of $\mathcal{G}_k^{\mathrm{tr}}$, suppose*

$$\mathbb{E}\big[\|\widehat{g}_{\mathrm{PW}}(\theta_k; \xi) - \nabla J(\theta_k)\|^2 \,\big|\, \mathcal{G}_k^{\mathrm{tr}}\big] \le \nu_T^2.$$

For a constant $0 < \eta \le 1/L_J$ and $\hat{k}$ uniform on $\{0, \dots, K-1\}$, independently of training,

$$\mathbb{E}\|\nabla J(\theta_{\hat{k}})\|^2 \le \frac{2[J(\theta_0) - J_{\inf}]}{\eta K} + \frac{L_J \eta \nu_T^2}{N_{\mathrm{bat}}}. \tag{B.18}$$

Under Theorem B.3, one may take $\nu_T^2 \le \Gamma_T^2$.

Proof. For $\lambda_k = 0$, the minibatch estimator is

$$G_k = \frac{1}{N_{\mathrm{bat}}} \sum_{i=1}^{N_{\mathrm{bat}}} \widehat{g}_{\mathrm{PW}}(\theta_k; \xi^{k,i}).$$

Conditional freshness, the assumed applicability of Theorem B.1 at each iterate, and within-batch independence give

$$\mathbb{E}_k G_k = \nabla J(\theta_k), \qquad \mathbb{E}_k \|G_k - \nabla J(\theta_k)\|^2 \le \frac{\nu_T^2}{N_{\mathrm{bat}}}.$$

Apply Theorem B.2 with $b_k = 0$, $G_D = 0$, $\sigma = \nu_T$, and $\eta_k = \eta$. Then $S_K = \eta K$ and $\hat{k}$ is uniform, giving (B.18). Under the uniform premises of Theorem B.3, the conditional variance is at most the conditional second moment, which is bounded by Γ_T^2. □

B.7.3 Training samples versus parameter updates

Corollary B.5. *Under Corollary B.4 and Theorem B.3, choose the variance bound ν_T so that $\nu_T \le \Gamma_T$. Let $\Delta_0 = J(\theta_0) - J_{\mathrm{inf}}$, $N_{\mathrm{fit}} = N_{\mathrm{bat}} K$, and suppose $\Delta_0 \nu_T > 0$. With*

$$\eta = \min\left\{ \frac{1}{L_J}, \sqrt{\frac{2\Delta_0 N_{\mathrm{bat}}}{L_J \nu_T^2 K}} \right\},$$

one has

$$\begin{aligned} \mathbb{E}\|\nabla J(\theta_{\hat{k}})\|^2 &\le \frac{2L_J\Delta_0}{K} + 2\nu_T \sqrt{\frac{2L_J\Delta_0}{N_{\mathrm{fit}}}} \\ &\le \frac{2L_J\Delta_0}{K} + 2\Gamma_T \sqrt{\frac{2L_J\Delta_0}{N_{\mathrm{fit}}}}. \end{aligned} \tag{B.19}$$

The first term depends on the number of parameter updates, whereas the second depends on the total number of sample paths. At a fixed sample-path budget, larger minibatches leave fewer updates and therefore enlarge the first term in this bound.

Proof. Let $\eta_\star = \sqrt{2\Delta_0 N_{\mathrm{bat}}/(L_J \nu_T^2 K)}$. Since $\eta = \min\{1/L_J, \eta_\star\}$,

$$\frac{1}{\eta} \le L_J + \frac{1}{\eta_\star}, \qquad \eta \le \eta_\star.$$

Substitution in (B.18) yields

$$\mathbb{E}\|\nabla J(\theta_{\hat{k}})\|^2 \le \frac{2L_J\Delta_0}{K} + \frac{2\Delta_0}{K\eta_\star} + \frac{L_J \eta_\star \nu_T^2}{N_{\mathrm{bat}}}.$$

The last two terms are equal and sum to $2\nu_T\sqrt{2L_J\Delta_0/(N_{\mathrm{bat}}K)}$. Using $N_{\mathrm{fit}} = N_{\mathrm{bat}}K$ and the horizon premise $\nu_T \le \Gamma_T$ proves (B.19). □

B.8 Best-in-class convergence under a PL condition

The following result imposes a Polyak–Łojasiewicz (PL) condition on the parameterized objective J and upgrades stationarity to objective-value convergence within the chosen policy class for unregularized policy optimization. Convex stage costs alone do not imply this condition because projection can make the objective locally flat. Appendix B.8.1 verifies PL for a shared base-stock policy whose expected cost can be nonconvex and whose projection constraints can bind or become inactive. Recall that $J_\Theta^\star = \inf_{\theta \in \Theta_{\mathrm{int}}} J(\theta)$ is the best expected cost among represented policies with integrable total cost, and $\Delta_\Theta = J_\Theta^\star - V^\star$ is the policy-class approximation gap.

Corollary B.6 (Best-in-class convergence under PL)**.** *For the unregularized updates under the regularity, noise, and update-domain assumptions of Corollary B.4, replace the constant stepsize by the diminishing sequence below, suppose the displayed parameter updates remain in Θ, and assume that, for some $\mu > 0$,*

$$\tfrac{1}{2}\|\nabla J(\theta)\|^2 \ge \mu[J(\theta) - J_\Theta^\star], \qquad \theta \in \Theta. \tag{B.20}$$

With

$$\eta_k = \frac{2}{\mu(k+k_0)}, \qquad k_0 \geq \max\{2, 2L_J/\mu\},$$

the terminal checkpoint obeys

$$\mathbb{E}[J(\theta_K) - J^\star_\Theta] \leq \frac{A_{\mathrm{PL}}}{K+k_0}, \qquad A_{\mathrm{PL}} = \max\left\{k_0[J(\theta_0) - J^\star_\Theta], \frac{2L_J\nu_T^2}{N_{\mathrm{bat}}\mu^2}\right\}. \tag{B.21}$$

Therefore

$$\mathbb{E}[J(\theta_K) - V^\star] \leq \frac{A_{\mathrm{PL}}}{K+k_0} + \Delta_\Theta. \tag{B.22}$$

The PL inequality is imposed directly on J and need not follow from convex stage costs.

Unlike the general stationarity bound, this result controls the expected cost of the terminal iterate. Under the PL condition, within-class optimization error vanishes asymptotically, but the policy-class approximation error Δ_Θ remains.

Proof of Corollary B.6. The preceding descent calculation with a varying stepsize and (B.20) gives

$$e_{k+1} \leq (1-\mu\eta_k)e_k + \frac{L_J\nu_T^2}{2N_{\mathrm{bat}}}\eta_k^2, \qquad e_k = \mathbb{E}[J(\theta_k) - J^\star_\Theta].$$

Let $s = k + k_0$ and $q = 2L_J\nu_T^2/(N_{\mathrm{bat}}\mu^2)$. With $\eta_k = 2/(\mu s)$,

$$e_{k+1} \leq (1 - 2/s)e_k + q/s^2.$$

If $e_k \leq A_{\mathrm{PL}}/s$, then $A_{\mathrm{PL}} \geq q$ implies

$$e_{k+1} \leq \frac{A_{\mathrm{PL}}(s-2) + q}{s^2} \leq \frac{A_{\mathrm{PL}}}{s+1}.$$

The definition of A_{PL} establishes the induction at $k = 0$, proving (B.21). Adding $J^\star_\Theta - V^\star = \Delta_\Theta$ proves (B.22). □

B.8.1 An inventory example satisfying PL

The following inventory example shows that PL does not require a convex policy objective or a fixed projection active set. It uses a single base-stock target shared across periods. Base-stock policy optimization is studied by Bhandari & Russo (2024) and Chen et al. (2026). The elementary specialization below makes the condition explicit without requiring an affine deployed trajectory.

Consider $T \geq 2$ periods with independent, identically distributed nonnegative demands D_t, where D_t is revealed after decision t and corresponds to ξ_{t+1} in the general model. Replenishment is immediate and uncapacitated, unmet demand is backlogged, and the objective includes holding and backlog costs but no ordering cost. Let s_t be net inventory before ordering and x_t inventory after ordering. A single target $b \in I = [a, \bar{b}]$ is shared across periods:

$$x_t = \mathrm{Proj}_{[s_t,\infty)}(b) = \max\{s_t, b\}, \qquad s_{t+1} = x_t - D_t, \qquad s_1 = s_0 > \bar{b}. \tag{B.23}$$

The order quantity is $x_t - s_t \geq 0$: the policy replenishes up to the target when inventory is low and otherwise keeps the existing stock. Thus the initial stock exceeds the target, but later demand can make replenishment necessary. For holding and backlog rates $h, p > 0$, the unregularized objective is

$$J(b) = \mathbb{E}\sum_{t=1}^{T}\left[h(s_{t+1})^+ + p(-s_{t+1})^+\right].$$

Here $\theta = b$ and $\lambda_k = 0$, so no proposal-distance term is used.

Suppose demand has finite mean and an absolutely continuous distribution function F_D, with density $f_D \geq f_{\min} > 0$ almost everywhere on I. Assume that the newsvendor target $b^\star$, defined by $F_D(b^\star) = p/(h+p)$, lies in the interior of I, and that

$$p_0 := \mathbb{P}\{D_1 > s_0 - a\} > 0.$$

The density condition gives curvature to the expected one-period cost. The last condition ensures a positive probability that demand reduces inventory below every target in I, so the target affects a later decision.

To see how these conditions imply PL, set $A_t = s_0 - \sum_{r<t} D_r$ and $\ell(y) = \mathbb{E}[h(y - D_1)^+ + p(D_1 - y)^+]$. Nonnegative demand and (B.23) give $x_t = \max\{A_t, b\}$ by induction. Because A_t is independent of D_t, differentiation yields

$$J'(b) = w_T(b)\ell'(b), \qquad w_T(b) = \sum_{t=2}^{T} \mathbb{P}\{A_t < b\}, \qquad \ell'(b) = (h+p)F_D(b) - p. \tag{B.24}$$

The factor $w_T(b)$ is the expected number of periods in which replenishment occurs: it measures how often the target affects the deployed decision. The first period contributes nothing to J' because $x_1 = s_0$. For later periods, $\mathbb{P}\{A_t = b\} = 0$ at each fixed b, and bounded one-period slopes justify differentiation under the expectation. Since $A_t \le s_0 - D_1$ for $t \ge 2$, we have $(T-1)p_0 \le w_T(b) \le T - 1$. Also, ℓ is strongly convex on I with modulus $m = (h+p)f_{\min}$. Hence J' has the sign of $b - b^\star$, so $J_I^\star := \min_{b\in I} J(b) = J(b^\star)$. Integrating (B.24) and using strong convexity of ℓ gives

$$J(b) - J(b^\star) \le (T-1)\big[\ell(b) - \ell(b^\star)\big] \le \frac{T-1}{2m}|\ell'(b)|^2.$$

Together with $|J'(b)| \ge (T-1)p_0|\ell'(b)|$, this proves

$$\tfrac{1}{2}|J'(b)|^2 \ge \underbrace{(T-1)p_0^2(h+p)f_{\min}}_{\mu}\big[J(b) - J_I^\star\big], \qquad b \in I. \tag{B.25}$$

Thus a positive probability of replenishment, together with curvature of the expected holding and backlog cost, prevents a suboptimal plateau on the target interval.

A nonconvex two-period instance. Take $T = 2$, $s_0 = 2$, $h = p = 1$, uniform demands on $[0, 2]$, and $I = [1/4, 3/2]$. Then $b^\star = 1$, $x_1 = 2$, and $x_2 = \max\{b, 2 - D_1\}$. The second-period constraint $x_2 \ge 2 - D_1$ is active with probability $1 - b/2$ and inactive with probability $b/2$, both positive throughout I. Direct integration gives

$$J'(b) = \frac{b(b-1)}{2}, \qquad J(b) - J(1) = \frac{(b-1)^2(2b+1)}{12}.$$

For $b \ne 1$, their ratio satisfies

$$\frac{|J'(b)|^2}{2[J(b) - J(1)]} = \frac{3b^2}{2(2b+1)} \ge \frac{1}{16}, \qquad b \in I.$$

At $b = 1$ the PL inequality holds trivially, so $\mu = 1/16$ is valid on the whole interval. Nevertheless, $J''(b) = b - 1/2 < 0$ on $(1/4, 1/2)$. This instance therefore satisfies PL without convexity of the policy objective or a fixed projection active set.

The target interval is important: in this two-period instance, every $b \le 0$ gives the same suboptimal policy and a zero gradient. The example establishes PL for the direct base-stock parameter on I, not for arbitrary neural parameterizations of that target. Applying Corollary B.6 still requires its smoothness, sampling, and update-domain assumptions. In particular, the stochastic updates must remain in I. Adding a projection of the parameter onto I changes the update and requires a corresponding projected-SGD analysis.

B.9 A POLICY-CLASS APPROXIMATION GUARANTEE

We next ask how accurately the projected policy class can approximate an optimal policy. The argument compares both policies on the same uncertainty path and requires the optimal policy's feedback rule to remain feasible at states reached by the represented policy. Recall $\Theta_{\text{int}} = \{\theta \in \Theta : \mathbb{E}|C(\theta, \xi)| < \infty\}$, $J_\Theta^\star = \inf_{\theta\in\Theta_{\text{int}}} J(\theta)$, and $\Delta_\Theta = J_\Theta^\star - V^\star$. Write $h_t = \xi_{1:t}$, and let $\mathcal{R}_t$ contain every state–history pair reachable by either the projected class or an optimal policy $\pi^\star$ attaining $V^\star$, along a history in the support of the data distribution. Suppose $\pi^\star$ has a feedback representation $\pi_t^\star(s, h)$ on $\mathcal{R}_t$, is L_π-Lipschitz in s, and remains feasible throughout this comparison domain:

$$\pi_t^\star(s, h) \in \mathcal{X}_t(s, \varphi_t(h)), \qquad (s, h) \in \mathcal{R}_t.$$

This condition extends feasibility beyond the policy's own trajectories to the comparison domain $\mathcal{R}_t$ and is precisely what permits the projection argument to compare the represented and optimal policies away from the optimal trajectory. Assume also the global pathwise Lipschitz bounds

$$\|f_t(s,x,w) - f_t(\bar{s},\bar{x},w)\| \le L_f^s\|s-\bar{s}\| + L_f^x\|x-\bar{x}\|,$$
$$|c_t(s,x,w) - c_t(\bar{s},\bar{x},w)| \le L_c^s\|s-\bar{s}\| + L_c^x\|x-\bar{x}\|,$$

and suppose some $\theta_{\mathrm{app}} \in \Theta_{\mathrm{int}}$ satisfies

$$\sup_{t,(s,h)\in\mathcal{R}_t} \|\mu_{\theta_{\mathrm{app}},t}(s,\varphi_t(h)) - \pi_t^\star(s,h)\| \le \varepsilon_\pi. \tag{B.26}$$

Theorem B.4 (Approximation by projected policies)**.** *Under the preceding conditions, let* $\kappa = L_f^s + L_f^x L_\pi$. *Then*

$$0 \le \Delta_\Theta \le K_T^\pi \varepsilon_\pi, \qquad K_T^\pi = TL_c^x + L_f^x(L_c^s + L_c^x L_\pi)\sum_{t=1}^{T}\mathcal{H}_{t-1}(\kappa). \tag{B.27}$$

The constant is $O(T)$, $O(T^2)$, *or geometric in* T *according as* $\kappa < 1$, $\kappa = 1$, *or* $\kappa > 1$.

This is a guarantee about what the policy class can represent, not what training will find. It bounds the best expected cost in the class rather than the cost of an arbitrary trained checkpoint.

Proof. Couple $\pi_{\theta_{\mathrm{app}}}$ and $\pi^\star$ using the same exogenous sample path ξ, and write $\delta_t = \|s_t^{\theta_{\mathrm{app}}} - s_t^\star\|$ and $e_t = \|x_t^{\theta_{\mathrm{app}}} - x_t^\star\|$. Feasibility throughout the comparison domain makes $\pi_t^\star(s_t^{\theta_{\mathrm{app}}}, h_t)$ a fixed point of the projection used by the learned policy. Projection onto a fixed feasible set does not increase distances. Combining this fact with (B.26) and Lipschitz continuity of $\pi^\star$ gives

$$e_t \le \varepsilon_\pi + L_\pi\delta_t, \qquad \delta_{t+1} \le \kappa\delta_t + L_f^x\varepsilon_\pi.$$

Since $\delta_1 = 0$, $\delta_t \le L_f^x\varepsilon_\pi\mathcal{H}_{t-1}(\kappa)$. Therefore the pathwise cost difference is at most

$$\sum_{t=1}^{T}(L_c^s\delta_t + L_c^x e_t) \le K_T^\pi\varepsilon_\pi.$$

Taking expectations and using the optimality of $\pi^\star$ proves the claim. The horizon regimes follow from the geometric sum. □

B.10 Parameter smoothing at nondifferentiable projections

If a fixed parameter reaches a point where the projection is nondifferentiable with positive probability, its sample-path derivative need not exist almost surely. When $\Theta = \mathbb{R}^{d_\theta}$, an optional alternative is the smoothed target with smoothing scale $\delta > 0$,

$$J_\delta(\theta) = \mathbb{E}_\varepsilon[J(\theta + \delta\varepsilon)], \qquad \varepsilon \sim \mathcal{N}(0, I), \quad \varepsilon \perp \xi.$$

The set Θ_{reg} from the proof of Theorem B.1 contains parameters at which the sampled cost is differentiable almost surely and has full Lebesgue measure. Because the perturbation has a density, $\theta + \delta\varepsilon \in \Theta_{\mathrm{reg}}$ almost surely. Under an integrable domination condition, differentiating the perturbed policy trajectory gives an unbiased gradient of J_δ, provided the perturbed sampled loss is differentiable and the backward pass returns its derivative almost surely. Smoothing does not validate an arbitrary derivative at a nondifferentiable state-dependent projection. If J is globally L_0-Lipschitz, then

$$|J_\delta(\theta) - J(\theta)| \le L_0\delta\,\mathbb{E}\|\varepsilon\| \le L_0\delta\sqrt{d_\theta}.$$

Every perturbed policy trajectory remains feasible by Proposition A.1. Smoothing changes the training target, not the deployed feasibility guarantee.

B.11 PROJECTION PLATEAUS AND PROPOSAL-DISTANCE GEOMETRY

Projection explains one reason why stationarity need not imply optimality. An open, unbounded set of proposals can project to the same suboptimal vertex. The deployed objective is then constant on that entire set, even when the cost in decision space is strongly convex. For example, let the feasible set be $[0, 1]$ and the deployed cost be $(x - 1)^2$. Every proposal $u < 0$ projects to the suboptimal decision $x = 0$, so the unregularized proposal gradient is zero throughout that open, unbounded region. The proposal-distance term supplies a negative-gradient direction toward the boundary, although its iterates need not cross the boundary in finite time. The next proposition formalizes this distinction between a zero cost gradient and a nonzero proposal-distance gradient.

Proposition B.2 (Projection-induced flat regions and proposal-distance gradients)**.** *Let $\mathcal{X} \subset \mathbb{R}^m$ be a compact full-dimensional polytope, let ϕ be continuously differentiable on an open neighborhood of $\mathcal{X}$, let $P_{\mathcal{X}}$ be Euclidean projection, and let $N_{\mathcal{X}}(v)$ be the normal cone at a vertex v. Then*

$$P_{\mathcal{X}}(u) = v \quad \textit{for every } u \in v + \operatorname{int} N_{\mathcal{X}}(v).$$

If v is suboptimal for ϕ on $\mathcal{X}$, then $\phi(P_{\mathcal{X}}(u))$ is constant and suboptimal on this open, unbounded proposal region. Hence no positive global Polyak–Łojasiewicz constant exists on the full proposal space, even when ϕ is strongly convex.

For $\lambda > 0$, define

$$F_\lambda(u) := \phi(P_{\mathcal{X}}(u)) + \frac{\lambda}{2}\|u - P_{\mathcal{X}}(u)\|^2.$$

Then $\min_{u\in\mathbb{R}^m} F_\lambda(u) = \min_{x\in\mathcal{X}} \phi(x)$. At every $u \notin \mathcal{X}$ at which $P_{\mathcal{X}}$ is differentiable, with $Q = DP_{\mathcal{X}}(u)$,

$$\nabla F_\lambda(u) = Q\nabla\phi(P_{\mathcal{X}}(u)) + \lambda\{u - P_{\mathcal{X}}(u)\} \neq 0, \tag{B.28}$$

and the two displayed terms are orthogonal.

Proof. For a closed convex set, $P_{\mathcal{X}}(u) = v$ if and only if $u - v \in N_{\mathcal{X}}(v)$. At a vertex of a full-dimensional polytope, this normal cone is full dimensional, has nonempty interior, and is unbounded. The projected objective is therefore constant on the asserted region. If its value there exceeds the optimum, a positive global Polyak–Łojasiewicz inequality is impossible because the proposal gradient vanishes throughout that region.

For the regularized objective, $F_\lambda(u) \geq \phi(P_{\mathcal{X}}(u)) \geq \min_{\mathcal{X}} \phi$, while any minimizer $x^\star \in \mathcal{X}$ satisfies $P_{\mathcal{X}}(x^\star) = x^\star$ and $F_\lambda(x^\star) = \phi(x^\star)$. At a differentiability point of a polyhedral Euclidean projection, Q is the orthogonal projector onto the tangent space of the active face. The residual $u - P_{\mathcal{X}}(u)$ lies in the orthogonal normal space. The chain rule and the standard identity

$$\nabla_u \left\{ \frac{1}{2}\|u - P_{\mathcal{X}}(u)\|^2 \right\} = u - P_{\mathcal{X}}(u)$$

give (B.28) and its orthogonal decomposition. If the gradient vanished, each orthogonal term would vanish, which is impossible when $u \notin \mathcal{X}$. □

B.11.1 SEQUENTIAL PROJECTIONS

The same property extends to a multistage policy: moving each proposal toward its projected decision leaves earlier decisions, and therefore later states and feasible sets, unchanged. More generally, every point on the ray from the deployed process x through the proposal process u deploys the same x. For $\lambda > 0$, moving toward x strictly decreases the regularized objective whenever proposal distance is positive, while leaving cost unchanged. To state this property before imposing a shared parameterization, let $\mathbb{L}^2_{\mathcal{F}}$ be the space of adapted (nonanticipative) processes $v = (v_t)_{t=1}^T$ satisfying $\|v\|^2_{\mathbb{L}^2_{\mathcal{F}}} = \mathbb{E}\sum_{t=1}^T \|v_t\|^2 < \infty$. Let $\mathbb{U}_2 \subseteq \mathbb{L}^2_{\mathcal{F}}$ be the set of all adapted proposal processes for which the sequential projection recursion is well defined, its deployed process belongs to $\mathbb{L}^2_{\mathcal{F}}$, and the realized cost is integrable. Write $\mathcal{R}(u) = x$ for the deployed decision process obtained by applying these proposals to the stagewise projections and state transitions. Thus $\mathcal{R}$ maps a complete proposal process to its feasible decision process, rather than projecting each stage onto a set fixed in advance. Define $\Pi_2 := \{\pi \in \Pi : x^\pi \in \mathbb{L}^2_{\mathcal{F}}\}$.

Theorem B.5 (Causal radial invariance and descent)**.** *Under relatively complete recourse and the measurability conditions of Proposition A.1, every $u \in \mathbb{U}_2$, with $x = \mathcal{R}(u)$, satisfies*

$$\mathcal{R}(x) = x, \qquad \mathcal{R}(x + \alpha(u - x)) = x \quad \textit{for every } \alpha \geq 0 \quad \textit{almost surely}. \tag{B.29}$$

For $\lambda > 0$, define the unrestricted process objective

$$\overline{H}_\lambda(u) := \mathbb{E}\left[C(\mathcal{R}(u), \xi) + \frac{\lambda}{2} \sum_{t=1}^{T} \|u_t - [\mathcal{R}(u)]_t\|^2 \right].$$

Its right directional derivative toward x is

$$\overline{H}'_\lambda(u; x - u) = -\lambda\, \mathbb{E} \sum_{t=1}^{T} \|u_t - x_t\|^2. \tag{B.30}$$

Thus every nonfixed proposal has a strict descent direction. Moreover,

$$\inf_{u \in \mathbb{U}_2} \overline{H}_\lambda(u) = \inf_{\pi \in \Pi_2} J(\pi).$$

In particular, this value is $V^\star$ whenever restricting to Π_2 does not change the MSP optimum, as in models with uniformly bounded recourse decisions.

The regularized objective can therefore decrease strictly while every deployed decision, and hence the cost, remains unchanged. The regularizer removes unnecessary proposal distance along this direction, but it does not make that direction a policy-improvement step. The theorem also shows that the proposal-distance term preserves the unrestricted optimal value under the stated conditions. This conclusion requires neither a scenario tree nor differentiability of the projection map. Its proof applies the projection normal-cone condition stage by stage: keeping earlier decisions unchanged keeps later states and feasible sets unchanged, even when the bounds depend on the state.

Proof of Theorem B.5. Fix a sample path outside the null set on which the sequential recursion is well defined, and write $x = \mathcal{R}(u)$. Euclidean projection optimality at stage t gives

$$u_t - x_t \in N_{\mathcal{X}_t(s_t, z_t)}(x_t).$$

Because a normal cone is a cone, x_t is also the projection of $x_t + \alpha(u_t - x_t)$ onto the same feasible set for every $\alpha \geq 0$. At the first stage the feasible set is unchanged. Inductively, if all earlier deployed decisions remain x, then their generated states remain unchanged, so the current moving right-hand side and feasible set also remain unchanged. This proves the second identity in (B.29), while the first follows by using the feasible process x itself as the proposal. These identities hold pathwise whenever the original recursion is well defined. The moment conditions are used below for the expected-objective conclusions. The process $x + \alpha(u - x)$ is adapted and square-integrable. Its recursion is well defined by the same induction, its deployed process is x, and its realized cost is therefore integrable, so it belongs to $\mathbb{U}_2$.

For $0 \leq \tau \leq 1$, take $\alpha = 1 - \tau$ to obtain

$$\overline{H}_\lambda(u + \tau(x - u)) = \mathbb{E}[C(x, \xi)] + \frac{\lambda}{2}(1 - \tau)^2 \mathbb{E} \sum_{t=1}^{T} \|u_t - x_t\|^2.$$

The right derivative at zero is (B.30). Finally, every $u \in \mathbb{U}_2$ is mapped to the square-integrable decision process of a policy in Π_2, so $\overline{H}_\lambda(u) \geq \inf_{\pi \in \Pi_2} J(\pi)$. Conversely, each policy $\pi \in \Pi_2$ can use its induced process x^π as its own proposal, giving $\mathcal{R}(x^\pi) = x^\pi$ with zero residual. Taking infima proves the last claim. □

Theorem B.5 concerns the unrestricted adapted proposal process. A shared policy may be unable to express its radial direction. To state a sufficient condition, let u_θ and $x_\theta = \mathcal{R}(u_\theta)$ denote the closed-loop proposal and deployed processes, including their dependence on preceding states. A shared policy must use one parameter perturbation across all histories. The condition below asks for a perturbation that moves every history-specific proposal, to first order, toward its own deployed decision. The Hadamard qualifier used below requires this directional response to remain stable when the step size and direction are perturbed simultaneously.

Corollary B.7 (Shared-policy radial visibility)**.** *Fix $\lambda > 0$. Suppose $\theta \mapsto u_\theta \in \mathbb{U}_2$ is Hadamard directionally differentiable as a map into $\mathbb{L}^2_{\mathcal{F}}$, $\mathcal{R} : \mathbb{U}_2 \to \mathbb{L}^2_{\mathcal{F}}$ is locally Lipschitz at u_θ relative to the $\mathbb{L}^2_{\mathcal{F}}$ norm, and the deployed-cost functional $x \mapsto \mathbb{E}[C(x, \xi)]$ is locally Lipschitz at x_θ. If $\mathcal{D}_{\mathrm{pd}}(\theta) > 0$ and there is a parameter direction e such that $\theta + \tau e \in \Theta$ for every sufficiently small $\tau > 0$ and*

$$u'_\theta(e) = x_\theta - u_\theta \quad \text{in } \mathbb{L}^2_{\mathcal{F}}, \tag{B.31}$$

then $x'_\theta(e) = 0$ and

$$H'_\lambda(\theta; e) = -\lambda\, \mathbb{E} \sum_{t=1}^{T} \|u_t^\theta - x_t^\theta\|^2. \tag{B.32}$$

Indeed, $u_{\theta+\tau e} = u_\theta + \tau(x_\theta - u_\theta) + o(\tau)$. Local Lipschitz continuity of $\mathcal{R}$ and (B.29) then give $x_{\theta+\tau e} = x_\theta + o(\tau)$. Local Lipschitz continuity of the deployed-cost functional makes its directional derivative zero, and differentiating the squared process norm yields (B.32). Hence a parameter satisfying (B.31) cannot be directionally stationary for H_λ unless its proposals already equal their projected decisions almost surely. We call (B.31) *radial visibility*. It is sufficient, not necessary. Parameter sharing can violate it, and residual contributions from different histories can cancel. Thus a fixed λ can change the best policy in the represented class even though Theorem B.5 preserves the optimum in unrestricted proposal space.

Remark B.1 (Why shared-policy visibility is needed). Consider two equally likely observed histories, feasible sets $\mathcal{X}_+ = [1, 2]$ and $\mathcal{X}_- = [-2, -1]$, and the shared scalar proposal $u_\theta = \theta$. Let the respective costs be $c_+(x) = (x - 2)^2$ and $c_-(x) = 0$. For $|\theta| < 1$,

$$x_+^\theta = 1, \qquad x_-^\theta = -1, \qquad J(\theta) = \frac{1}{2}, \qquad \mathcal{D}_{\mathrm{pd}}(\theta) = \frac{1}{2}(\theta^2 + 1).$$

Hence $H'_\lambda(\theta) = \lambda\theta$, and $\theta = 0$ is stationary despite a positive proposal residual and a represented policy $\theta = 2$ with strictly lower expected cost J. The two history-specific radial directions have opposite signs and cancel through the shared parameter. This example also shows why Theorem B.5 cannot be transferred unconditionally to a low-dimensional policy class.

B.11.2 Horizon bound for the regularizer gradient

Corollary B.8 (Horizon bound for the regularized gradient)**.** *Under Theorem B.3, suppose additionally that $\|u_t - x_t\| \le \bar{d}$ along the relevant trajectories. Define*

$$\Delta_T := \bar{d}\sqrt{d_\theta}\left[T L_\mu^\theta + (L_\mu^s + L_{\mathrm{MR}})\chi \sum_{t=1}^{T} \mathcal{H}_{t-1}(\rho)\right]. \tag{B.33}$$

Then, for every $\lambda \ge 0$,

$$\|\widehat{g}_\lambda\| \le \Gamma_T + \lambda\Delta_T.$$

Consequently, for fixed λ, the corresponding second moment is at most $(\Gamma_T + \lambda\Delta_T)^2$ whenever the assumptions hold almost surely. Moreover, whenever differentiation and expectation can be interchanged for d_{pd}, $\|\nabla\mathcal{D}_{\mathrm{pd}}(\theta)\| \le \Delta_T$.

For fixed sensitivity bounds, keeping proposals close to their projected decisions controls the additional gradient introduced by regularization. The factor Δ_T also accounts for the propagation of these derivatives through later stages, so the residual distance alone does not determine their size.

Proof of Corollary B.8. The proof of Theorem B.3 gives

$$\|S_t\| \le \chi\sqrt{d_\theta}\mathcal{H}_{t-1}(\rho).$$

Because $D_u\bar{P}_t$ is an orthogonal projector, $\|I_{m_t} - D_u\bar{P}_t\| \le 1$. Combining the definitions of U_t and X_t yields

$$U_t - X_t = (I_{m_t} - D_u\bar{P}_t)U_t - D_s\bar{P}_t S_t,$$

and therefore

$$\|U_t - X_t\| \le \sqrt{d_\theta}\left[L_\mu^\theta + (L_\mu^s + L_{\mathrm{MR}})\chi\mathcal{H}_{t-1}(\rho)\right].$$

Multiplying by $\|u_t - x_t\| \le \bar{d}$, summing over t, and using (B.7) proves the pathwise bound. Jensen's inequality and the stated derivative–expectation interchange give $\|\nabla\mathcal{D}_{\mathrm{pd}}\| \le \Delta_T$. □

C CALIBRATION: FULL RESULTS, PROOFS, AND STATISTICAL BOUNDARIES

C.1 LEARNING CONDITIONAL MEANS FROM SAMPLE PATHS

We give the complete construction of the estimator $\widehat{Q}_t$ and radius β_t in Theorem 2, followed by the coverage and rate results. Recall the response H_t in (9) and its conditional mean Q_t. The set $\mathcal{A}_t$ contains every query reachable under perfect information. Throughout this subsection, we condition on the training information $\mathcal{G}_{\mathrm{fit}}$, so the continuation approximation, query sets, maps defining the history summaries, and tuning parameters are fixed before using $\mathcal{D}_{\mathrm{cal}}$.

Regularity assumptions. The construction controls three sources of error. A bounded response range controls sampling error, regularity in the history summary controls the bias from pooling nearby observations, and regularity in the query extends a bound from finitely many representative queries to the whole domain. These are standard ingredients of uniform nonparametric regression (Stone, 1982; Jiang, 2019; Low, 1997; Chernozhukov et al., 2014). Here the conditions must hold over all of $\mathcal{A}_t$, because the optimizer may choose its query after observing the path. They concern the fixed response map and its conditional mean, not the accuracy of the continuation approximation.

Assumption C.1 (Uniform band regularity)**.** *The standard measurability conditions recorded in Appendix A hold. The domains, partitions, query covers, numerical bounds, and regularity constants below are deterministic or $\mathcal{G}_{\mathrm{fit}}$-measurable and are fixed before calibration. The bounds are certified on the stated domains, and conservative model-based values are admissible. For every stage t:*

(a) We use a known compact domain $\mathcal{Z}_t \subseteq \mathbb{R}^{d_t}$ for the history summary. It contains the support $\mathsf{Z}_t = \mathrm{supp}(z_t)$ and has finite partitions $\mathcal{P}_{t,h}$ into measurable cells of diameter at most h. Here $h > 0$ is a scalar cell-resolution parameter, distinct from the history argument in a query.

(b) The query set $\mathcal{A}_t$, with distance $d_{\mathcal{A},t}$, has finite covering numbers:

$$\mathcal{N}(\mathcal{A}_t, d_{\mathcal{A},t}, \epsilon) < \infty, \qquad 0 < \epsilon \leq 1.$$

Here $\mathcal{N}(\mathcal{A}_t, d_{\mathcal{A},t}, \epsilon)$ is the covering number: the minimum number of points needed so that every query lies within distance ϵ of one of them. For vector-valued queries, the distance may simply be $d_{\mathcal{A},t}(a, a') = \|a - a'\|$.

(c) For fixed scalar response bounds $\ell_t^H \leq u_t^H$, let $W_t = u_t^H - \ell_t^H$. The bounds

$$H_t(a, w) \in [\ell_t^H, u_t^H], \qquad |H_t(a, w) - H_t(\bar{a}, w)| \leq L_{a,t} d_{\mathcal{A},t}(a, \bar{a})$$

hold simultaneously for all $a, \bar{a} \in \mathcal{A}_t$, almost surely in $w = \xi_{t+1}$. They are not restricted to queries historically paired with that realization.

(d) The conditional mean satisfies

$$|Q_t(a, z) - Q_t(a, \bar{z})| \leq L_{z,t} \|z - \bar{z}\|^{\tau_t}$$

for some Hölder exponent $\tau_t \in (0, 1]$, all $a \in \mathcal{A}_t$, and all $z, \bar{z} \in \mathsf{Z}_t$. This is Lipschitz continuity when $\tau_t = 1$.

Estimator and error radius. The partition of $\mathcal{Z}_t$ groups observations according to z_t, whereas the query cover controls variation in the state, decision, and history arguments retained by H_t. Predictive sufficiency of z_t does not prevent the fitted continuation approximation from depending on the history. We therefore use separate tuning parameters for cell diameter h and query-cover resolution ϵ, both fixed before calibration. Let $\mathcal{A}_{t,\epsilon}$ be a fixed query cover, also called an ϵ-net, with a point within distance ϵ of every query. Fix its finite cardinality or a certified upper bound before calibration. Index the $N_{\mathrm{cal}} = |\mathcal{D}_{\mathrm{cal}}|$ calibration sample paths by i. For a cell $B \in \mathcal{P}_{t,h}$, define the number of calibration samples falling into that cell as

$$N_{t,B} = \sum_{i=1}^{N_{\mathrm{cal}}} \mathbf{1}\{z_t^i \in B\}.$$

For $z \in B$, the conditional-mean estimate averages responses evaluated at the common query a using the uncertainty realizations from that cell:

$$\widehat{Q}_t(a,z) = \begin{cases} \dfrac{1}{N_{t,B}} \displaystyle\sum_{i:z_t^i \in B} H_t(a, \xi_{t+1}^i), & N_{t,B} > 0, \\ (\ell_t^H + u_t^H)/2, & N_{t,B} = 0. \end{cases} \tag{C.1}$$

Thus each observation's history summary determines its cell, and its recorded uncertainty realization is evaluated at the requested query a. No observed state or decision needs to match that query. When a cell is empty, the known response range supplies its midpoint and radius $W_t/2$.

For the simultaneous sampling bound, count the combinations of stage, cell, and representative query:

$$R(h,\epsilon) = \sum_{t=1}^{T} |\mathcal{P}_{t,h}|\,|\mathcal{A}_{t,\epsilon}|, \qquad \Lambda = \log\left(\frac{2R(h,\epsilon)}{\delta}\right).$$

Only the size of the query cover enters that bound, so the implementation need not enumerate its representatives when a certified cardinality bound is available. Replacing $|\mathcal{A}_{t,\epsilon}|$ by this upper bound preserves validity. For $z \in B$, define the error radius

$$\beta_t(z) = \begin{cases} \min\left\{ W_t,\ L_{z,t} h^{\tau_t} + 2L_{a,t}\epsilon + W_t \sqrt{\dfrac{\Lambda}{2N_{t,B}}} \right\}, & N_{t,B} > 0, \\ W_t/2, & N_{t,B} = 0. \end{cases} \tag{C.2}$$

For a nonempty cell, the three summands in (C.2) control the bias from pooling nearby history summaries, extension from the representative queries, and sampling error, respectively. The factor two accounts for moving both the empirical and population means from a representative query to the requested query. The cap W_t is valid because both means lie in $[\ell_t^H, u_t^H]$.

Coverage and shrinking width.

Theorem C.1 (Uniform band from reused uncertainty realizations)**.** *Under the data model of Appendix A.1 and Assumption C.1, with conditional probability at least $1-\delta$ given $\mathcal{G}_{\mathrm{fit}}$,*

$$\left|\widehat{Q}_t(a,z) - Q_t(a,z)\right| \le \beta_t(z) \tag{C.3}$$

simultaneously for every stage t, query $a \in \mathcal{A}_t$, and history summary $z \in \mathsf{Z}_t$.

Because the band holds simultaneously, it remains valid at a query selected by the inner optimizer after seeing the entire path. Its radius uses the observed cell counts, so validity does not require a lower bound on cell probabilities. For an explicit uniform rate, the next corollary adds growth bounds on the partitions and query covers, together with a lower bound on cell probabilities, as in uniform nonparametric regression (Stone, 1982; Jiang, 2019).

Corollary C.1 (Uniform rate: conditioning dimension versus query-cover size)**.** *Conditional on $\mathcal{G}_{\mathrm{fit}}$, suppose Assumption C.1 holds uniformly in N_{cal}. In addition, suppose there are fixed constants $C_{Z,t}, C_{A,t}, \kappa_t > 0$ and exponents $p_t \ge 0$ such that*

$$|\mathcal{P}_{t,h}| \le C_{Z,t} h^{-d_t}, \qquad |\mathcal{A}_{t,\epsilon}| \le (C_{A,t}/\epsilon)^{p_t},$$
$$\mathbb{P}\{z_t \in B\} \ge \kappa_t h^{d_t}$$

for every stage t, every cell $B \in \mathcal{P}_{t,h}$ with $B \cap \mathsf{Z}_t \neq \varnothing$, and all sufficiently small $h, \epsilon > 0$, uniformly in N_{cal}. Let $\tau = \min_t \tau_t$ and $d_z = \max_t d_t$. For a failure probability $\delta_{\mathrm{occ}} \in (0,1)$, as $N_{\mathrm{cal}} \to \infty$ assume

$$\log(N_{\mathrm{cal}}/\delta_{\mathrm{occ}}) = o(N_{\mathrm{cal}}), \qquad \log(1/\delta) = O(\log(N_{\mathrm{cal}}/\delta_{\mathrm{occ}})).$$

With

$$h \asymp \left(\frac{\log(N_{\mathrm{cal}}/\delta_{\mathrm{occ}})}{N_{\mathrm{cal}}}\right)^{1/(2\tau+d_z)}, \qquad \epsilon \asymp h^{\tau},$$

the maximum band radius satisfies, with conditional probability at least $1-\delta_{\mathrm{occ}}$ given $\mathcal{G}_{\mathrm{fit}}$,

$$\max_t \sup_{z \in \mathsf{Z}_t} \beta_t(z) = O\left[\left(\frac{\log(N_{\mathrm{cal}}/\delta_{\mathrm{occ}})}{N_{\mathrm{cal}}}\right)^{\tau/(2\tau+d_z)}\right], \tag{C.4}$$

with constants depending on the stagewise smoothness, cell-probability, range, and covering-number parameters. In particular, p_t affects only the logarithmic covering-number term, not the exponent in (C.4).

Reusing each recorded uncertainty realization at many state–decision queries removes the need for data collected near every such query. The history-summary dimension d_z therefore determines the rate exponent, while the query-cover exponents p_t enter through a logarithmic term. This rate concerns estimation of the conditional means Q_t induced by the fixed continuation approximation, not estimation of v itself. Any admissible h and ϵ fixed before calibration give the stated finite-sample coverage, while the displayed choices balance the terms in the radius to obtain its shrinking-width rate. The two failure probabilities have different roles: δ controls failure of the simultaneous band, whereas δ_{occ} controls the small cell counts that could prevent the stated rate. Both conclusions hold with total failure probability at most $\bar{\delta}$ when $\delta = \delta_{\mathrm{occ}} = \bar{\delta}/2$.

The remaining dependence on history-summary dimension is unavoidable over general Hölder classes. Theorem C.2 in Appendix C.3 gives a matching exponent for the worst-case expected shortfall radius of uniformly valid one-sided centering rules, placing a limit on this nonparametric calibration task while leaving room for certificates that exploit additional structure.

C.2 Band proofs and an optional construction

We first prove the uniform band and rate results from Appendix C.1, then give an alternative construction for separable responses. Throughout, we condition on the training information $\mathcal{G}_{\mathrm{fit}}$.

C.2.1 Proofs of uniform coverage and its rate

Proof of Theorem C.1. Fix a stage t and condition on the history summaries in the calibration sample, $(z_t^1, \dots, z_t^{N_{\mathrm{cal}}})$. Within each nonempty cell and at each point in the finite representative-query set, the calibration residuals

$$H_t(a, \xi_{t+1}^i) - Q_t(a, z_t^i)$$

are conditionally independent, mean zero, and have range width at most W_t. Hoeffding's inequality therefore bounds the conditional failure probability for each stage–cell–net-point triple by $\delta/R(h,\epsilon)$. Averaging over these stage-t summaries and taking a union bound over all triples gives the sampling term in (C.2), without requiring joint conditioning on history summaries from different stages. Hölder continuity contributes $L_{z,t}h^{\tau_t}$, while moving between a query and its nearest net point changes both empirical and population means by at most $L_{a,t}\epsilon$. For an empty cell, the midpoint of the known response range has error at most $W_t/2$. This proves (C.3). □

Proof of Corollary C.1. For each cell intersecting Z_t, a multiplicative Chernoff bound and the corollary's cell-probability condition give

$$\mathbb{P}\big\{N_{t,B} < \tfrac{1}{2}N_{\mathrm{cal}}\kappa_t h^{d_t}\big\} \le \exp\left(-\frac{N_{\mathrm{cal}}\kappa_t h^{d_t}}{8}\right).$$

A union bound over the stage–cell pairs makes $N_{t,B} \ge \frac{1}{2}N_{\mathrm{cal}}\kappa_t h^{d_t}$ simultaneous with probability at least $1-\delta_{\mathrm{occ}}$ for all sufficiently large N_{cal} under the displayed choice of h. Substitute this count into (C.2). The partition of $\mathcal{Z}_t$ contributes $d_t \log(1/h)$ and the query cover contributes $p_t \log(C_{A,t}/\epsilon)$ to Λ. Choosing $\epsilon \asymp h^{\tau}$ balances the query-extension and pooling-bias terms. The specified h then balances h^{τ} against the sampling term $\sqrt{\log(N_{\mathrm{cal}}/\delta_{\mathrm{occ}})/(N_{\mathrm{cal}}h^{d_z})}$, yielding (C.4). □

C.2.2 Optional construction for separable responses

The query cover can be avoided when the response is a finite sum of products of query-dependent coefficients and functions of uncertainty. In that case, we estimate the conditional means of those functions and combine their error bounds, rather than extending a bound from representative queries. Suppose the fixed response has the representation

$$H_t(a, w) = \sum_{\ell=1}^{R_t} a_{t\ell}(a)\psi_{t\ell}(w). \tag{C.5}$$

Here $a_{t\ell}(a)$ depends only on the state–decision–history query, while $\psi_{t\ell}(w)$ depends only on the next uncertainty realization. Only the conditional means

$$\mu_{t\ell}(z) = \mathbb{E}[\psi_{t\ell}(\xi_{t+1}) \mid z_t = z]$$

must then be estimated. Define the corresponding conditional-mean estimator by

$$\widehat{Q}_t(a, z) = \sum_{\ell=1}^{R_t} a_{t\ell}(a)\widehat{\mu}_{t\ell}(z).$$

If, on an event of conditional probability at least $1 - \delta$ given $\mathcal{G}_{\text{fit}}$, $|\widehat{\mu}_{t\ell}(z) - \mu_{t\ell}(z)| \le b_{t\ell}(z)$ simultaneously for every stage t, $\ell = 1, \dots, R_t$, and $z \in \mathsf{Z}_t$, then

$$|\widehat{Q}_t(a, z) - Q_t(a, z)| \le \sum_{\ell=1}^{R_t} |a_{t\ell}(a)| b_{t\ell}(z). \tag{C.6}$$

To obtain a radius that depends only on the history summary, define

$$\beta_t^{\text{sep}}(z) = \sup_{a \in \mathcal{A}_t} \sum_{\ell=1}^{R_t} |a_{t\ell}(a)| b_{t\ell}(z). \tag{C.7}$$

When this supremum is finite and can be computed or certified from above, β_t^{sep} replaces β_t in Algorithm 2: subtract it from $\widehat{Q}_t$ and add $2\sum_t \beta_t^{\text{sep}}(z_t)$ in the corrected gap (12). This gives the same type of certificate without constructing a query cover.

Implementation requires three checks. First, compute the supremum in (C.7), or certify a finite upper bound. Second, clip $\widehat{Q}_t$ pointwise to $[\ell_t^H, u_t^H]$. Since $Q_t \in [\ell_t^H, u_t^H]$, clipping cannot increase error and makes Proposition D.3 applicable. Third, fix any basis selected from the training and policy selection block before calibration. The representation in (C.5) must then be exact. An approximate factorization requires a certified uniform residual added to β_t^{sep}.

C.3 Statistical Boundaries: Results and Proofs

The certificate uses known model functions to reuse observations across endogenous states and decisions, but this benefit has two limits. First, reusing uncertainty realizations cannot generally remove the calibration rate's dependence on the history-summary dimension. Second, when decisions change the uncertainty distribution, observations under one decision need not identify outcomes under another. The next two results formalize these limits.

C.3.1 Unavoidable dependence on history-summary dimension

Theorem C.2 below shows that the history-summary dimension cannot generally be removed from the calibration rate. In a Hölder model of smoothness s on a d-dimensional history summary, every one-sided centering rule (q_N^-, r_N) that reports a lower estimate q_N^- and a valid allowance r_N for its shortfall, and is uniformly valid with probability $1 - \delta$, must have worst-case expected shortfall radius $\sup_q \mathbb{E}_q \|r_N\|_\infty \ge c\,(\log N/N)^{s/(2s+d)}$ for large N, where the supremum is over the stated Hölder class. Our histogram band construction attains this exponent (Corollary C.1), up to constants and the stated finite-cover factors. The lower bound concerns uniformly valid one-sided centering over this nonparametric class, rather than every calibration choice in Algorithm 2.

Theorem C.2 (Minimax shortfall of uniformly valid lower-centering rules)**.** *Fix $d \ge 1$, $s \in (0, 1]$, $\delta < 1/4$, and a radius L large enough that $q_0 \equiv 1/2$ lies in the interior of the Hölder ball below. Consider the favorable submodel*

$$z \sim \text{Unif}([0,1]^d), \qquad W \mid z = z_0 \sim \text{Bernoulli}(q(z_0)), \quad z_0 \in [0,1]^d,$$

where q ranges over a Hölder ball $\mathcal{H}_d^s(L)$ and takes values in $[1/4, 3/4]$. For every pair (q_N^-, r_N) that is jointly measurable as a function of the N observations and z, has Borel sample paths and $r_N \ge 0$, and for which both the event and the sup norm in the displays below are measurable, if

$$\inf_{q \in \mathcal{H}_d^s(L)} \mathbb{P}_q\big\{0 \le q(z) - q_N^-(z) \le r_N(z) \text{ for all } z \in [0,1]^d\big\} \ge 1 - \delta,$$

then there is a constant $c > 0$ such that, for all sufficiently large N,

$$\sup_{q \in \mathcal{H}_d^s(L)} \mathbb{E}_q \left\| r_N \right\|_\infty \geq c \left(\frac{\log N}{N} \right)^{s/(2s+d)}. \tag{C.8}$$

Thus the shortfall radius required by a uniformly valid one-sided centering correction cannot shrink faster on this subclass.

This limitation already appears with binary responses and uniform sampling of history summaries, before any difficult recourse optimization is introduced. The comparison with Corollary C.1 can be made in the same expected metric. Fix the band failure probability δ and take $\delta_{\mathrm{occ}} = N_{\mathrm{cal}}^{-2}$. Because $\sup_z \beta_t(z) \leq W_t$ deterministically, integrating the high-probability rate in (C.4), specialized to $\tau = s$, gives

$$\mathbb{E}\left[\max_t \sup_{z \in \mathsf{Z}_t} \beta_t(z) \,\middle|\, \mathcal{G}_{\mathrm{fit}} \right] = O\left[\left(\frac{\log N_{\mathrm{cal}}}{N_{\mathrm{cal}}} \right)^{s/(2s+d_z)} \right].$$

In the theorem's regression submodel, our construction uses $q_N^- = \widehat{q}_N - \beta_N$ and $r_N = 2\beta_N$. It therefore attains the lower-bound exponent, up to constants and the stated finite-cover factors.

The result concerns the shortfall control needed for an informative one-sided correction. It does not rule out a valid lower bound obtained from an arbitrarily conservative lower-centering rule, and certificates that exploit additional parametric structure or a known conditional distribution fall outside its scope. Without restrictions on the conditional means, observations need not determine their values throughout the support of the history summary. At a point with zero probability mass, the conditional distribution is not identified by the joint data distribution. Even for almost-everywhere validity, a unit discrepancy can be hidden on a region of probability $O(1/N)$. Hence no band valid at every point in this support can shrink uniformly over all measurable conditional means.

Proof of Theorem C.2. The pair (q_N^-, r_N) defines the uniformly valid random band $[q_N^-, q_N^- + r_N]$ of width r_N. We apply the usual packing argument to this band. Let ϕ be a nonnegative Hölder-s bump supported on the unit ball, with $\phi(0) = 1$. Place $M \asymp h^{-d}$ disjoint translates in $[0,1]^d$ and define

$$q_0(z) = \frac{1}{2}, \qquad q_j(z) = \frac{1}{2} + a\phi\left(\frac{z - z_j}{h} \right), \quad j = 1, \ldots, M,$$

where $a = c_0 h^s$. For sufficiently small c_0, every q_j belongs to the stated Hölder ball and takes values in $[1/4, 3/4]$. The alternatives have pairwise sup-norm separation a, whereas the Bernoulli likelihoods satisfy

$$D_{\mathrm{KL}}(P_{q_j}^N \,\|\, P_{q_0}^N) \leq CNa^2h^d.$$

Choose $h \asymp (\log N/N)^{1/(2s+d)}$ and then choose c_0 so that the last display is at most a sufficiently small multiple of $\log M \asymp \log N$.

If $\|r_N\|_\infty < a$, the induced band cannot contain two members of the packing because they are separated by a in sup norm. It therefore induces a selector of the data-generating alternative whenever coverage holds. With the preceding constant chosen small enough and N sufficiently large, Fano's inequality gives an average error probability of at least $1/2$ for every selector. Subtracting the coverage failure probability $\delta < 1/4$, at least one alternative must produce $\|r_N\|_\infty \geq a$ with constant probability. Multiplying by a proves (C.8).

Finally, embed the experiment in a two-stage model with no nontrivial constraints and one-step response $H(z, W) = W$. Its conditional centering target is exactly $q(z)$, so the lower bound applies to the part of the certificate based on the uniform centering band. □

C.3.2 Why endogenous uncertainty requires decision coverage

When the distribution of the next observation depends on the decision, outcomes observed under one decision need not identify the cost of another. Such identification requires data covering the alternative decision—often called *overlap*—or additional model structure. The next result shows why reusing observations alone cannot resolve this ambiguity.

Theorem C.3 (Impossibility without decision coverage under endogenous uncertainty)**.** *Let $A \in \{0,1\}$ denote the decision. Consider observational trajectories generated by the data-collection policy π_0, which always chooses zero. There are two models, P_0 and P_1, with identical distributions for the observed trajectories such that*

$$J_{P_0}(\pi_0) = J_{P_1}(\pi_0) = 1, \qquad V^\star_{P_0} = 0, \qquad V^\star_{P_1} = 1.$$

Thus the optimality gap of the data-collection policy is one under P_0 and zero under P_1. Any data-dependent upper confidence bound U_N satisfying

$$\inf_{P \in \{P_0, P_1\}} \mathbb{P}_P\{J_P(\pi_0) - V^\star_P \le U_N\} \ge 1 - \alpha$$

must satisfy

$$\mathbb{P}_{P_1}\{U_N \ge 1\} \ge 1 - \alpha$$

for every N, even though the true optimality gap under P_1 is zero. No uniformly valid certificate can shrink without data covering both decisions or additional structural assumptions.

Even exact knowledge of the data-collection policy's cost does not identify its distance from optimality: an unobserved decision may be better. The two models generate the same observed trajectories for every sample size, so additional data from that policy cannot distinguish them. Validity under P_0 therefore forces the same conservative bound under P_1, where the policy is already optimal. In the model of Appendix A.1, observed exogenous uncertainty and known model functions permit evaluation of alternative decisions under the same uncertainty distribution, removing this particular ambiguity.

Proof of Theorem C.3. Under both models, choosing zero gives terminal uncertainty realization $W = 0$. Choosing one gives $W = 0$ under P_0 and $W = 1$ under P_1. Use the known cost

$$c(a, w) = \mathbf{1}\{a = 0\} + w\mathbf{1}\{a = 1\}.$$

The data-collection policy π_0 always chooses zero, so the distributions of the observed trajectories are identical and $J_{P_0}(\pi_0) = J_{P_1}(\pi_0) = 1$. Choosing one costs zero under P_0 and one under P_1, which gives $V^\star_{P_0} = 0$ and $V^\star_{P_1} = 1$. Uniform validity under P_0 requires $\mathbb{P}_{P_0}\{U_N \ge 1\} \ge 1 - \alpha$. Observational equivalence transfers the same probability statement to P_1, where the true optimality gap is zero. □

D Certification: Full Results and Proofs

D.1 Certifying the Selected Policy from Sample Paths

Let $\widehat{\pi}$ be a feasible policy selected using $\mathcal{D}_{\text{fit}}$. For Algorithm 1, $\widehat{\pi} = \pi_{\widehat{\theta}}$ at the selected checkpoint $\widehat{\theta}$. We hold the selected policy fixed throughout calibration and evaluation. Its decision rule does not change, but its decisions still respond to the observed state and history. To assess this policy without knowing the optimal MSP cost, we construct a lower bound $L \le V^\star$. Comparing the policy's expected cost with this benchmark gives

$$0 \le J(\widehat{\pi}) - V^\star \le J(\widehat{\pi}) - L.$$

This comparison leads to a computable upper confidence bound $U_{\alpha,\delta}$ on the optimality gap, with coverage at least $1 - \alpha - \delta$ conditional on the training information $\mathcal{G}_{\text{fit}}$ under the assumptions below. The failure probabilities δ and α account for calibration and evaluation, respectively. Fixing the policy and certification choices before using $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{ev}}$ allows us to assess the selected policy without requiring it to be stationary or optimal.

The construction has three steps: form an information-relaxation benchmark, estimate the conditional means in its penalty from recorded sample paths, and evaluate the corrected gap on fresh sample paths. The final report accounts for estimation and inner-optimization errors as well as evaluation sampling. We learn about the uncertainty distribution from sample paths. The construction also requires known model functions, known bounds on the functions being estimated, and certified global lower bounds for the pathwise inner problems. Accurate continuation approximations can tighten the benchmark, but are not required for coverage.

D.1.1 CONSTRUCTING A LOWER BENCHMARK

Optimizing separately on each recorded sample path with perfect foresight of all future uncertainty realizations, while retaining the physical dynamics and recourse constraints, gives an optimistic benchmark. This *information relaxation* can be tightened by an *information-relaxation penalty*: a charge for information unavailable to an implementable policy. If the charge has mean zero under every feasible, nonanticipative policy, adding it leaves each such policy's expected cost unchanged while the perfect-information minimum remains no larger. A penalty with nonpositive conditional mean also preserves a lower bound, since adding it cannot increase any such policy's expected cost. Unlike the proposal-distance term in Algorithm 1, this penalty is used for certification and does not enter the policy-optimization objective.

One such mean-zero penalty subtracts the realized current-stage cost plus approximate remaining cost from its conditional mean. The difference averages to zero for an implementable decision because that decision is made before the next uncertainty realization is observed. It can nevertheless penalize a perfect-information decision that uses the future observation. The main statistical difficulty is that the conditional mean is unknown: a fitted regression need not preserve the mean-zero property.

Fix a *continuation approximation* $v = (v_2, \dots, v_{T+1})$ using $\mathcal{D}_{\text{fit}}$. The role of $v_{t+1}(s, h)$ is to approximate the optimal expected cost from stage $t+1$ through T, starting from state s and history h and using feasible, nonanticipative decisions. We set $v_{T+1} = 0$ because no costs remain after the horizon. Algorithm 1 does not require these functions, but they are used here to construct the certificate.

The perfect-information optimizer can encounter states and decisions outside the selected policy's trajectories. Each state–decision–history tuple at which we evaluate a conditional mean is called a *query*. Before calibration, we fix a measurable set $\mathcal{A}_t$ containing every stage-t query feasible for that optimizer along any sample path in the data distribution's support. Known state and decision bounds can supply a conservative box or polyhedral superset without solving the relaxed problem.

For a query $a = (s, x, h) \in \mathcal{A}_t$ and next-stage uncertainty realization w, define

$$H_t^v(a, w) := \underbrace{c_t(s, x, w)}_{\text{realized current-stage cost}} + \underbrace{v_{t+1}\big(f_t(s, x, w), h \oplus w\big)}_{\text{approximate expected remaining cost}}, \tag{D.1}$$

$$Q_t^v(a, z) := \underbrace{\int H_t^v(a, w) \overline{K}_t(\mathrm{d}w \mid z)}_{\text{conditional mean of } H_t^v}, \tag{D.2}$$

where $h \oplus w$ appends the realization w to the history. The continuation approximation is defined at all states and histories appearing in (D.1). Once v is fixed, $H_t^v(a, w)$ is evaluated directly from the known cost and transition functions and the fitted continuation approximation, with no separate regression for H_t^v needed. In contrast, $Q_t^v(a, z)$ averages this response over the unknown conditional distribution of the next uncertainty realization. Calibration uses $\mathcal{D}_{\text{cal}}$ to estimate Q_t^v and bound its estimation error. We display z separately because calibration groups observations with nearby history summaries while evaluating their recorded realizations at a common query a. In the inner problem, $z = \varphi_t(h)$, and we abbreviate $Q_t^v(a) = Q_t^v(a, \varphi_t(h))$.

The penalty below subtracts the entire response H_t^v from its conditional mean, because both the current-stage cost and approximate remaining cost may still be random when the decision is chosen. If the stage cost is already known when the decision is made, it cancels from the centered expression. The resulting stagewise penalty

$$m_t^v(a, w) = Q_t^v(a) - H_t^v(a, w) \tag{D.3}$$

has conditional mean zero whenever the query is chosen before the next uncertainty realization is observed. A perfect-information optimizer can instead choose the query after observing that realization. The penalty adds back the difference whenever the realized response $H_t^v(a, w)$ falls below its conditional mean $Q_t^v(a)$.

Let $\mathcal{X}^{\text{PI}}(\xi)$ contain all decision sequences satisfying the physical dynamics and recourse constraints on path ξ, with nonanticipativity relaxed. Here "PI" stands for perfect information. For a decision

sequence x and path ξ, let $a_t = (s_t, x_t, h_t)$ be the resulting stage-t query, with $h_t = \xi_{1:t}$ and states generated by (A.4). Write $M^v(x,\xi) = \sum_{t=1}^{T} m_t^v(a_t, \xi_{t+1})$ and set

$$Y_v(\xi) = \inf_{x \in \mathcal{X}^{\mathrm{PI}}(\xi)} \{C(x,\xi) + M^v(x,\xi)\}, \qquad L(v) = \mathbb{E}[Y_v]. \tag{D.4}$$

Here $Y_v(\xi)$ is the optimal value for one fixed path, while $L(v)$ averages these values over the exogenous path distribution. The functions H_t^v and Q_t^v define the penalty, but they are not themselves lower bounds on the MSP optimum. The lower benchmark is $L(v)$. To establish this property, fix any feasible nonanticipative policy π. Its decisions remain available to the perfect-information optimizer, and its expected penalty is zero by conditional centering. The expected perfect-information value therefore satisfies $L(v) \le J(\pi)$. Taking the infimum over all such policies gives the left inequality in

$$L(v) \le V^\star \le J(\pi). \tag{D.5}$$

This is the classical weak-duality relation for information relaxation (Brown et al., 2010). Equivalently, every feasible policy satisfies

$$0 \le J(\pi) - V^\star \le J(\pi) - L(v).$$

To quantify tightness, we compare v_t with the exact optimal cost-to-go functions V_t. These are theoretical reference functions: Algorithm 2 does not require solving the Bellman recursion that defines them. With $V_{T+1} = 0$, the recursion is

$$V_t(s,h) = \inf_{x \in \mathcal{X}_t(s,\varphi_t(h))} \int [c_t(s,x,w) + V_{t+1}(f_t(s,x,w), h \oplus w)]\, \overline{K}_t(\mathrm{d}w \mid \varphi_t(h)). \tag{D.6}$$

With the exact continuation values $v = V$, the classical information-relaxation bound is exact (Brown et al., 2010):

$$Y_V(\xi) = V_1(s_1, h_1) \quad \text{almost surely}, \qquad L(V) = V^\star.$$

Appendix A.2 records the technical conditions for this identity.

Why the realized cost is centered. When the current-stage cost is unknown at decision time, a perfect-information optimizer can exploit its realization. Centering only the continuation value leaves this source of information advantage uncorrected. The resulting lower bound remains valid, but Proposition D.1 shows that it can be arbitrarily weak even with exact continuation values in a smooth convex two-stage model. Centering both terms gives the exact bound in the same example.

The role of the continuation approximation. With exact conditional centering and the stated integrability conditions, lower-bound validity does not require an accurate continuation approximation. In particular, $v \equiv 0$ is a valid choice: it still centers the current-stage cost through the penalty $m_t^0 = Q_t^0 - c_t$, rather than removing the penalty altogether. The conditional expected current cost must still be estimated when it is unknown. A fitted continuation approximation can improve tightness, but approximation error does not necessarily weaken the bound. Constant shifts cancel under centering, and Proposition D.2 gives a broader sufficient condition for exactness despite approximation error.

Accounting for conditional-mean estimation error. Using independent data to fit a conditional mean does not ensure that the estimated penalty has nonpositive conditional mean. Even unbiasedness across datasets does not ensure this property after the fitted estimate is fixed. The band construction in Appendix C.1 addresses this through calibration: conditional on the training information $\mathcal{G}_{\mathrm{fit}}$, its lower confidence bounds lie below the true conditional means simultaneously at every stage and query in $\mathcal{A}_t$, with probability at least $1-\delta$ over the calibration data. On this event, the corrected penalty has nonpositive conditional mean under every feasible nonanticipative policy, preserving the lower-bound property in (D.5). The expected optimum of the corrected perfect-information problem is therefore a lower bound on $V^\star$ on this calibration event. Appendix C.1 constructs these confidence bounds, and Appendix D.2 accounts separately for evaluation sampling error and inner-solver error in the reported optimality-gap bound.

D.2 COMPUTING AND INTERPRETING THE CERTIFICATE

After bounding the conditional-mean estimation error using $\mathcal{D}_{\mathrm{cal}}$, we use the reserved evaluation block $\mathcal{D}_{\mathrm{ev}}$ to construct the upper confidence bound $U_{\alpha,\delta}$ on the selected policy's optimality gap. The two data blocks average different quantities: calibration averages responses to estimate conditional means at each query, whereas evaluation averages corrected policy–benchmark gaps across complete paths.

For each sample path in $\mathcal{D}_{\mathrm{ev}}$, the calibration data and all uncertainty realizations on that path are fixed. The resulting perfect-information problem is therefore deterministic, with the decisions still to be optimized. We compute a certified lower value for its objective, compare it with the selected policy's penalized cost on the same path, and apply the calibration correction. We then average these corrected gaps over $\mathcal{D}_{\mathrm{ev}}$ and add a confidence margin for evaluation sampling error. Thus the expectation in the lower-bound argument is not evaluated analytically. The inner computation requires a certified lower bound, not merely a feasible inner solution.

We give two calibration constructions with a common reporting rule. The uniform band corrects the information-relaxation penalty itself so that, on the simultaneous calibration event, the expected perfect-information value remains a lower bound on $V^\star$. This construction allows serially dependent, decision-independent uncertainty. For independent binary uncertainty, the construction used in the certificate experiments instead corrects the complete paired gap. Correcting the gap directly can retain useful cancellations even when the uncorrected empirical inner objective does not give a valid lower benchmark. Either construction can be evaluated using a fixed Hoeffding or empirical Bernstein rule under its stated assumptions. Readers using the general band can proceed from the conditional-mean band construction directly to the evaluation rules. The intervening binary construction is an alternative, not an additional step required by the band method.

Let $\mathcal{G}_{\mathrm{cal}} = \sigma(\mathcal{G}_{\mathrm{fit}}, \mathcal{D}_{\mathrm{cal}})$ collect all training information and calibration data. Each calibration construction gives a statistic G from one fresh evaluation sample path and an event $\mathcal{E}_{\mathrm{cal}} \in \mathcal{G}_{\mathrm{cal}}$ satisfying

$$\mathbb{P}(\mathcal{E}_{\mathrm{cal}} \mid \mathcal{G}_{\mathrm{fit}}) \geq 1 - \delta, \qquad \mathbb{E}[G \mid \mathcal{G}_{\mathrm{cal}}] \geq J(\widehat{\pi}) - V^\star \quad \text{on } \mathcal{E}_{\mathrm{cal}}. \tag{D.7}$$

The first inequality gives the confidence level of the calibration event. On that event, the second says that the mean corrected gap bounds the optimality gap from above. This expectation averages over a fresh sample path and any independent solver seed, with the calibrated construction held fixed. To bound the sampling error in the evaluation average, we require certified finite bounds $0 \leq A_G \leq G \leq B_G$. Their difference, $W_G = B_G - A_G$, enters the confidence margin added to that average. The range-construction rule is fixed before calibration. Its endpoints may depend on calibration data but must be fixed before evaluation. In bounded MSP models, physical cost and value ranges can be combined with a prescribed upper bound on the inner-solver error to obtain such a range, as shown explicitly in Proposition D.3.

Conditional-mean band construction. Let $\mathcal{E}_{\mathrm{band}}$ be the event on which (C.3), or the separable moment band, holds simultaneously over all queries in $\mathcal{A}_t$. With its decision-independent radius $\beta_t(z)$, define

$$Q_t^-(a,z) = \underbrace{\widehat{Q}_t(a,z)}_{\text{estimated conditional mean}} - \underbrace{\beta_t(z)}_{\text{error bound}}, \qquad r_t(z) = 2\beta_t(z). \tag{D.8}$$

On this event, $0 \leq Q_t - Q_t^- \leq r_t$: the factor two covers both sides of the original estimation band. Replacing the exact mean in (D.3) by its lower confidence bound gives $m_t^-(a,w) = Q_t^-(a,\varphi_t(h)) - H_t(a,w)$ for $a = (s,x,h)$. Sum these increments along the path to obtain M^-, and set

$$Y^-(\xi) = \inf_{x \in \mathcal{X}^{\mathrm{PI}}(\xi)} \{C(x,\xi) + M^-(x,\xi)\}. \tag{D.9}$$

On $\mathcal{E}_{\mathrm{band}}$, the conditional mean of m_t^- is nonpositive under every feasible nonanticipative policy. Adding the penalty therefore cannot increase any such policy's expected cost, and perfect-information minimization can only lower it further. The same comparison-policy argument as in (D.5) gives $\mathbb{E}[Y^- \mid \mathcal{G}_{\mathrm{cal}}] \leq V^\star$. A *certified lower value* is a computable lower bound on the true inner optimum. For any such $\underline{Y}^- \leq Y^-$, use

$$G^{\mathrm{band}}(\xi) = \underbrace{C^{\widehat{\pi}} + M^{-,\widehat{\pi}}}_{\text{penalized policy cost}} - \underbrace{\underline{Y}^-}_{\text{inner lower bound}} + \underbrace{\sum_t r_t(z_t)}_{\text{calibration correction}}. \tag{D.10}$$

Here $C^{\widehat{\pi}}$ and $M^{-,\widehat{\pi}}$ are the cost and penalty evaluated along $\widehat{\pi}$ on that same path. For the band construction and worked examples, write $G_i^{\mathrm{cert}} = G^{\mathrm{band}}(\xi^i)$ and $\underline{Y}_i = \underline{Y}^-(\xi^i)$. Feasibility of the policy for the relaxed problem makes this statistic nonnegative, while the added correction compensates for the reduction in the policy's penalty caused by replacing Q_t with Q_t^-. With $e_t = Q_t - Q_t^- \in [0, r_t]$ evaluated along that policy, the statistic's conditional mean is $J(\widehat{\pi}) - \mathbb{E}[\underline{Y}^- \mid \mathcal{G}_{\mathrm{cal}}] + \mathbb{E}[\sum_t (r_t - e_t) \mid \mathcal{G}_{\mathrm{cal}}]$. Thus (D.7) holds with $\mathcal{E}_{\mathrm{cal}} = \mathcal{E}_{\mathrm{band}}$. To complete the construction, Proposition D.3 supplies $A_G = 0$ and a finite B_G from response and cost bounds and a fixed upper bound on the inner-solver error, though a sharper certified range may also be used.

Direct gap correction for independent binary uncertainty. We next construct a corrected gap G^{bin} that can yield a tighter certificate than the band construction in (D.10). The policy and perfect-information objectives use the same estimated probabilities, so their estimation errors can partially cancel when we correct their difference directly. The construction takes the form $G^{\mathrm{bin}} = D + \kappa$, where D is the computed policy–benchmark gap and κ corrects for probability-estimation error. We call this a *paired* correction because both objectives use the same path and probability estimates.

Here binary refers to the uncertainty, not the decisions. Write $w_t = \xi_{t+1} \in \{0, 1\}$ for the uncertainty observed after the decision, and suppose these Bernoulli variables are independent across stages and of the predecision information $\mathcal{F}_t$, including initial exogenous information. Let $p_t = \mathbb{P}(w_t = 1 \mid \mathcal{F}_t)$ be the unknown, deterministic stagewise probability. The calibration sample paths give the estimate $\widehat{p}_t = N_{\mathrm{cal}}^{-1} \sum_{j=1}^{N_{\mathrm{cal}}} w_t^j$. For the fixed responses H_t in (D.1), the difference between the two outcomes and the estimated conditional mean are

$$\Delta_t(a) = H_t(a, 1) - H_t(a, 0), \qquad \widehat{Q}_t(a) = H_t(a, 0) + \widehat{p}_t \Delta_t(a).$$

Thus calibration estimates one probability per stage. This construction needs neither a history-summary partition nor a query cover. Instead, fix certified bounds $m_t \le \Delta_t(a) \le M_t$ over the entire query domain $\mathcal{A}_t$. The probability estimates satisfy $\|\widehat{p} - p\|_2 \le r_2$ with conditional probability at least $1 - \delta$ given the training information $\mathcal{G}_{\mathrm{fit}}$, where

$$r_2 = \sqrt{\frac{T + 2\sqrt{T \log(1/\delta)} + 2\log(1/\delta)}{4 N_{\mathrm{cal}}}}. \tag{D.11}$$

This radius bounds the combined estimation error across stages (Hsu et al., 2012, Theorem 2.1).

Unlike (D.10), we use $\widehat{Q}_t$ directly in the penalty rather than subtracting a band radius. The empirical penalized objective and its policy–benchmark gap are

$$\begin{aligned} F_{\widehat{p}}(x, \xi) &= C(x, \xi) + \sum_t \{\widehat{Q}_t(a_t) - H_t(a_t, w_t)\}, \\ D(\xi) &= F_{\widehat{p}}(\widehat{\pi}, \xi) - \underline{Y}_{\widehat{p}}(\xi), \end{aligned} \tag{D.12}$$

where $\underline{Y}_{\widehat{p}}$ is a certified lower bound on $\inf_{x \in \mathcal{X}^{\mathrm{PI}}(\xi)} F_{\widehat{p}}(x, \xi)$. Here $F_{\widehat{p}}(\widehat{\pi}, \xi)$ evaluates that objective at the selected policy's decisions, so the policy cost and inner solve use the same estimated penalty. Write $a_t^{\widehat{\pi}} = (s_t^{\widehat{\pi}}, x_t^{\widehat{\pi}}, h_t)$ for the query generated by the selected policy on that path. For the empirical gap D, the corrected statistic and its path-dependent calibration correction are

$$\begin{aligned} G^{\mathrm{bin}}(\xi) &= \underbrace{D(\xi)}_{\text{computed policy–benchmark gap}} + \underbrace{\kappa(\xi)}_{\text{calibration correction}}, \\ \kappa(\xi) &= r_2 \left[\sum_t \max\left\{ \Delta_t(a_t^{\widehat{\pi}}) - m_t,\ M_t - \Delta_t(a_t^{\widehat{\pi}}) \right\}^2 \right]^{1/2}. \end{aligned} \tag{D.13}$$

At each stage, the maximum in κ bounds how much the policy's response difference can differ from that of any feasible perfect-information alternative. The correction therefore depends on variation across queries, not on the absolute size of the responses. If the response difference is constant across queries at every stage, probability-estimation errors cancel from every policy–alternative objective difference, and tight bounds then give $\kappa = 0$.

With conditional probability at least $1 - \delta$ over calibration, G^{bin} satisfies (D.7). Validity is established for the corrected gap as a whole, so the empirical inner objective need not itself give a lower

bound on $V^\star$ in expectation. Corollary D.4 gives the proof and shows how a certified bound on D, including prescribed solver and arithmetic errors, determines the range $[A_G, B_G]$. Once this range is fixed, the evaluation step is the same as for the band construction. The independence assumption is specific to this binary construction and does not restrict the general band construction.

Evaluation rules. For each of the $n = N_{\mathrm{ev}} = |\mathcal{D}_{\mathrm{ev}}| \geq 2$ sample paths in $\mathcal{D}_{\mathrm{ev}}$, compute G_i using the band statistic in (D.10) or the binary statistic in (D.13), according to the construction fixed before calibration. Let $\overline{G} = n^{-1}\sum_i G_i$ and $S_G^2 = (n-1)^{-1}\sum_i (G_i - \overline{G})^2$ denote the sample mean and sample variance of the corrected gaps, respectively. Conditional on $\mathcal{G}_{\mathrm{cal}}$, $\overline{G}$ estimates the population mean $\mathbb{E}[G \mid \mathcal{G}_{\mathrm{cal}}]$ in (D.7). Because the sample mean can fall below this expectation, the report adds a confidence margin rather than using $\overline{G}$ alone. For a fixed evaluation failure probability $a \in (0,1)$, the two reporting radii are

$$R_{\mathrm{H}}(a) = W_G \sqrt{\frac{\log(2/a)}{2n}}, \tag{D.14}$$

$$R_{\mathrm{EB}}(a) = \sqrt{\frac{2S_G^2 \log(2/a)}{n}} + \frac{7W_G \log(2/a)}{3(n-1)}. \tag{D.15}$$

Hoeffding uses only the certified range, whereas empirical Bernstein also uses the sample variance of the complete corrected gap (Maurer & Pontil, 2009, Theorem 4). In the binary construction, a path-dependent κ therefore contributes both its own variance and covariance with D. A constant calibration shift, by contrast, changes neither the empirical variance nor the evaluation radius, although it shifts the reported upper bound. Conditional on calibration, both concentration inequalities require independent evaluation sample paths and a fixed finite range. The band construction permits dependence within each sample path, whereas the binary construction requires stagewise independence. As in Algorithm 2, the choice between Hoeffding and empirical Bernstein is fixed before calibration, and W_G must be certified rather than replaced by the observed sample range.

Algorithm 2 (expanded): certifying the selected policy. *Input.* A feasible policy $\widehat{\pi}$ determined using $\mathcal{D}_{\mathrm{fit}}$ only, the continuation approximation $\widehat{v} = v$ fixed during training and policy selection, history-summary and query domains, and known model-based bounds. Before calibration, choose the band or binary paired construction and an evaluation rule $\mathsf{m} \in \{\mathrm{H}, \mathrm{EB}\}$, with error budgets (δ, α). Fix the range-construction rule, solver settings, and an a priori bound on the difference between the true inner optimum and its certified lower value, valid for every path. Reserve the independent calibration and evaluation blocks $\mathcal{D}_{\mathrm{cal}}$ and $\mathcal{D}_{\mathrm{ev}}$, with $N_{\mathrm{cal}} \geq 1$ and $n = N_{\mathrm{ev}} \geq 2$.

1. **Calibrate.** For the band construction, form $(\widehat{Q}_t, \beta_t)$ by (C.1)–(C.2) or (C.6)–(C.7), and clip the separable estimate to $[\ell_t^H, u_t^H]$. For the binary construction, form $\widehat{p}$, r_2, and the correction rule in (D.13). Fix the resulting rule for computing G.

2. **Fix the range.** Apply the range-construction rule fixed before calibration, including the prescribed bounds on solver and arithmetic errors, to obtain finite $0 \leq A_G \leq G \leq B_G$ and $W_G = B_G - A_G$. Fix all endpoints before examining evaluation data.

3. **Evaluate.** Replay each sample path in $\mathcal{D}_{\mathrm{ev}}$ under $\widehat{\pi}$. Compute a certified lower bound on the inner optimum for the band construction in (D.9) or the binary construction in (D.12), for example from an LP dual solution or global branch-and-bound. Verify that the inner-solver error is no larger than the fixed bound. Form G_i by (D.10) or (D.13). If certification fails on any path, issue no finite certificate and do not discard that path. A feasible inner solution gives an upper bound on the inner optimum and cannot replace the required lower bound.

4. **Report.** Compute $\overline{G}$ and, for empirical Bernstein, the sample variance of the full G_i. Report
$$U_{\alpha,\delta} = \underbrace{\overline{G}}_{\text{mean corrected gap}} + \underbrace{R_{\mathsf{m}}(\alpha)}_{\text{evaluation margin}}.$$

When finite, $U_{\alpha,\delta}$ is the upper endpoint of the confidence interval $[0, U_{\alpha,\delta}]$, whose coverage is established by Theorem D.1. The displayed formula uses exact arithmetic, but a certified upward-rounded

implementation is also valid. The resulting numerical error is included in the certificate-width bound of Corollary D.1. Defining $U_{\alpha,\delta} = +\infty$ whenever no finite certificate is issued ensures that the coverage claim does not condition on successful reporting. With this convention, the calibration mean inequality and fixed finite range yield the following guarantee. All distributional statements below condition on the training information $\mathcal{G}_{\mathrm{fit}}$.

Theorem D.1 (End-to-end finite-sample validity)**.** *Let the training information $\mathcal{G}_{\mathrm{fit}}$ contain the data and randomness used to train and select an integrable, nonanticipative feasible policy $\widehat{\pi} \in \Pi$, fit $\widehat{v}$, and fix the calibration construction, domains, range-construction rule, and evaluation method in Algorithm 2. The policy may be any checkpoint selected using the training and policy selection block, including an output of Algorithm 1.*

Conditional on $\mathcal{G}_{\mathrm{fit}}$, let $\mathcal{D}_{\mathrm{cal}}$ consist of independent paths from the exogenous data distribution. Conditional on $\mathcal{G}_{\mathrm{cal}}$, let the paths in $\mathcal{D}_{\mathrm{ev}}$ be i.i.d. from that same distribution. Use the same measurable pathwise solver and reporting map on every path. If randomized, its path-specific seeds are i.i.d. and independent of $\mathcal{G}_{\mathrm{cal}}$ and all sample paths. Suppose the chosen calibration construction supplies a measurable statistic G, defined for every fresh sample path and solver seed, that satisfies (D.7)*. On $\mathcal{E}_{\mathrm{cal}}$, require $0 \le A_G \le G \le B_G < \infty$ almost surely under this joint conditional law, with calibration-measurable endpoints fixed before evaluation and including all prescribed numerical and solver errors. Whenever the implementation returns a finite report, its inner lower values are certified and its numerical checks meet the prescribed bounds. If a solver fails to return such a lower value or any prescribed numerical or range check fails, Algorithm 2 reports $+\infty$, so coverage is not conditioned on successful computation. Then either fixed evaluation rule in Algorithm 2 satisfies*

$$\mathbb{P}\{J(\widehat{\pi}) - V^\star \le U_{\alpha,\delta} \,|\, \mathcal{G}_{\mathrm{fit}}\} \ge 1 - \delta - \alpha. \tag{D.16}$$

For both stated calibration constructions, validity requires neither accurate continuation values nor global optimality of the selected policy. The band construction also needs no lower bound on cell probabilities for coverage.

Because coverage is conditional on the training information $\mathcal{G}_{\mathrm{fit}}$, searching over candidate policies using the training and policy selection block does not consume the calibration or evaluation error budget. The guarantee applies to the selected policy regardless of whether policy optimization reached a stationary or optimal point. A useful numerical bound, however, also requires control of the certificate's excess over the true optimality gap, as the next result quantifies.

Corollary D.1 (When the certificate is informative)**.** *Under the conditions of Theorem D.1, suppose the Bellman conditions in Appendix A.2 hold, so that $Y_V(\xi) = V_1(s_1, h_1)$ almost surely and $\mathbb{E}[V_1(s_1, h_1)] = V^\star$. On $\mathcal{E}_{\mathrm{cal}}$, assume the computation succeeds almost surely under the same joint conditional law, every certified inner lower value is within $\varepsilon_{\mathrm{IP}}$ of the true optimum of the inner problem it bounds, and numerical errors have fixed nonnegative bounds $\varepsilon_{\mathrm{path}}$ for the uniform upward path-statistic error and $\varepsilon_{\mathrm{report}}$ for the additional upward error in the final mean/radius computation. Let $\varepsilon_{\mathrm{path}} + \varepsilon_{\mathrm{report}} \le \varepsilon_{\mathrm{num}}$ and write $\varepsilon_{\mathrm{comp}} = \varepsilon_{\mathrm{IP}} + \varepsilon_{\mathrm{num}}$. If the path statistics and final report are evaluated in exact arithmetic, then $\varepsilon_{\mathrm{num}} = 0$, even when the inner-solver error $\varepsilon_{\mathrm{IP}}$ is positive. Suppose*

$$\|\widehat{v}_t - V_t\|_\infty \le \varepsilon_{v,t} \quad (t = 2, \dots, T+1), \qquad \varepsilon_{v,T+1} = 0 \tag{D.17}$$

where, for each $t = 1, \dots, T$, the bound for $\widehat{v}_{t+1}$ holds on the successor-state–history domain generated by queries in $\mathcal{A}_t$. Define

$$\mathcal{C}_{\mathrm{cal}} = \begin{cases} 4\sum_t \beta_t^{\mathrm{av}}, & \textit{band construction}, \\ 2\mathbb{E}[\kappa \mid \mathcal{G}_{\mathrm{cal}}], & \textit{binary paired construction}, \end{cases} \qquad \beta_t^{\mathrm{av}} = \mathbb{E}[\beta_t(z_t^{\mathrm{ev}}) \mid \mathcal{G}_{\mathrm{cal}}]. \tag{D.18}$$

Here z_t^{ev} is the stage-t history summary on a fresh evaluation sample path. Both expectations average over such paths conditional on $\mathcal{G}_{\mathrm{cal}}$. These expectations enter the theoretical width bound but are not required inputs to Algorithm 2. For Hoeffding set $(\mathcal{T}_{\mathrm{ev}}, \epsilon_{\mathrm{ev}}) = (2R_{\mathrm{H}}(\alpha), \alpha)$. For empirical Bernstein choose any fixed $0 < \eta < 1 - \alpha$ for the width analysis and set

$$(\mathcal{T}_{\mathrm{ev}}, \epsilon_{\mathrm{ev}}) = \big(R_{\mathrm{EB}}(\alpha) + R_{\mathrm{EB}}(\eta),\, \alpha + \eta\big). \tag{D.19}$$

The quantity $\mathcal{T}_{\mathrm{ev}}$ bounds the evaluation contribution to certificate width, while ϵ_{ev} is its joint coverage-and-width failure budget. Then, with probability at least $1 - \delta - \epsilon_{\mathrm{ev}}$ conditional on the training information $\mathcal{G}_{\mathrm{fit}}$,

$$\begin{aligned} 0 \le J(\widehat{\pi}) - V^\star &\le U_{\alpha,\delta} \\ &\le J(\widehat{\pi}) - V^\star + 2\sum_t \varepsilon_{v,t+1} + \mathcal{C}_{\mathrm{cal}} + \varepsilon_{\mathrm{comp}} + \mathcal{T}_{\mathrm{ev}}. \end{aligned} \tag{D.20}$$

For the band construction, replacing β_t^{av} by $\bar{\beta}_t^{\mathrm{rate}} = \sup_{z\in\mathsf{Z}_t}\beta_t(z)$ gives a uniform width bound. Substituting the histogram rate (C.4) *additionally uses the growth and cell-probability conditions in Corollary C.1. The separate probability δ_{occ} that some relevant cell receives too few observations must be included when that rate is substituted into a joint statement. Proposition D.3 uses the computable supremum $\bar{\beta}_t^{\mathrm{env}} = \sup_{z\in\mathcal{Z}_t}\beta_t(z)$ over the entire modeled domain of z_t to construct the certified range.*

The bound explains how far the certificate can exceed the true optimality gap. The contributions come from continuation-approximation error, calibration uncertainty, inner-solver and numerical error, and evaluation sampling error. Enlarging the evaluation sample addresses only this sampling error, so even a precise estimate of the corrected gap can remain conservative when the continuation approximation or inner lower bounds are weak. The uniform value-error term is sufficient but need not be sharp: the refinement in Proposition D.2 of Appendix D.3 accounts for cancellation under centering and compares the remaining value errors with the extra cost of suboptimal decisions.

This distinction between coverage and width also affects the probability budget. Coverage uses $\delta + \alpha$, whereas the empirical Bernstein width statement controls upward evaluation error as well, at an additional failure probability η without changing the reported certificate. To obtain a joint coverage-and-width statement with total evaluation budget a, one can choose $\alpha = \eta = a/2$ before calibration.

To connect this guarantee for a selected policy with policy optimization, restrict the parameter class to its integrable members,

$$\Theta_{\mathrm{int}} = \{\theta\in\Theta : \mathbb{E}|C(\theta,\xi)| < \infty\}, \qquad J_\Theta^\star = \inf_{\theta\in\Theta_{\mathrm{int}}} J(\theta), \qquad \Delta_\Theta = J_\Theta^\star - V^\star,$$

and assume $\Theta_{\mathrm{int}} \neq \varnothing$. Here $C(\theta,\xi)$ is the total path cost of π_θ. The quantity Δ_Θ measures the approximation error from restricting the policy class, while $J(\widehat{\theta}) - J_\Theta^\star$ measures the selected policy's excess cost over the class infimum.

Corollary D.2 (Selected-policy optimality gap and certificate-width decomposition)**.** *Let $\widehat{\theta}\in\Theta$ be any checkpoint selected using the training and policy selection block only, with integrable total cost, and fix the certification choices listed in Algorithm 2 before calibration. Under Proposition A.1 and theorem D.1, the selected policy has coverage at least $1-\delta-\alpha$. This statement applies to either calibration construction and either fixed evaluation rule, without assuming that the checkpoint minimizes either the training cost or the expected cost globally.*

For width, impose the conditions of Corollary D.1, and suppose the value-approximation event

$$\mathcal{E}_v = \{\|\widehat{v}_t - V_t\|_\infty \le \varepsilon_{v,t},\ t = 2,\dots,T+1;\ \varepsilon_{v,T+1} = 0\} \in \mathcal{G}_{\mathrm{fit}}$$

has probability at least $1-\gamma$. Assume validity for almost every realization of the training information and the width hypotheses on $\mathcal{E}_v$. Under the joint law of the data and randomness underlying $\mathcal{G}_{\mathrm{fit}}$, the calibration and evaluation data, and the solver randomness, with probability at least $1-\gamma-\delta-\epsilon_{\mathrm{ev}}$,

$$\begin{aligned} 0 \le J(\widehat{\theta}) - V^\star &\le U_{\alpha,\delta} \\ &\le \underbrace{J(\widehat{\theta}) - J_\Theta^\star}_{\textit{training and selection}} + \underbrace{\Delta_\Theta}_{\textit{policy-class approximation}} + \underbrace{2\sum_t \varepsilon_{v,t+1}}_{\textit{value approximation}} \\ &\quad + \underbrace{\mathcal{C}_{\mathrm{cal}}}_{\textit{calibration}} + \underbrace{\varepsilon_{\mathrm{comp}}}_{\textit{inner and numerical error}} + \underbrace{\mathcal{T}_{\mathrm{ev}}}_{\textit{evaluation}}. \end{aligned} \tag{D.21}$$

The training-and-selection and policy-class approximation terms sum to the selected policy's optimality gap. The value-approximation, calibration, computation, and evaluation terms bound the certificate's excess over that gap. Algorithm 2 computes its report without knowing the true optimality gap, the policy-class approximation error, or the continuation-approximation errors. The optional conditions in Appendix B connect the first two terms to policy optimization: the Polyak–Łojasiewicz condition yields best-in-class convergence for unregularized policy optimization, and uniform approximation of an optimal feasible policy controls the policy-class approximation error. Neither condition is needed for certification. In particular, the random iterate selected in Theorem B.2, if subsequently fixed before calibration, retains its expected stationarity bound while separately receiving the conditional coverage guarantee. A checkpoint chosen by validation within the training and policy selection block also receives coverage, though it need not inherit the stationarity bound.

Certificate-guided stopping. The certificate can guide whether to stop policy optimization or continue training. To certify successive checkpoints, allocate per-report failure probabilities whose sum does not exceed a prescribed total and reserve fresh calibration and evaluation blocks for each attempted report. For a desired optimality-gap tolerance $\epsilon > 0$, stop at the first checkpoint whose reported bound is at most ϵ. Corollary D.3 in Appendix D.3.2 ensures that the probability of ever stopping with an optimality gap greater than ϵ is at most the total allocated failure probability. Earlier reports may inform further training and later certification choices, but the decision to attempt each report, its policy, block sizes, and certification settings must be fixed before opening its calibration block. The guarantee does not permit arbitrary stopping within a block. A bound above the tolerance means that the desired accuracy has not yet been certified, but it does not by itself establish that further policy optimization is needed.

D.3 Additional Certificate Results and Proofs

This subsection proves the cost-timing separation and refined tightness result, establishes the certificate theorem and its corollaries, extends coverage to repeated checkpoints, and derives a computable range bound.

Proposition D.1 (Continuation-only centering can be arbitrarily loose)**.** *Let* $R_t^v(a,w) = v_{t+1}(f_t(s,x,w), h \oplus w)$ *and define*

$$m_t^{v,\mathrm{cont}}(a,w) = \int R_t^v(a,\bar{w})\overline{K}_t(\mathrm{d}\bar{w} \mid \varphi_t(h)) - R_t^v(a,w).$$

Define $M^{v,\mathrm{cont}}$, Y_v^{cont}, *and* $L^{\mathrm{cont}}(v)$ *as in* (D.4)*, with this increment replacing* m_t^v*. Whenever these quantities are measurable and integrable so that the information-relaxation comparison is well defined,* $L^{\mathrm{cont}}(v) \le V^\star$*. For every* $\epsilon \in (0,1)$*, moreover, a two-stage MSP with compact polyhedral recourse, a nonconstant exact continuation value, and nonnegative smooth convex costs bounded by four has*

$$V^\star = 1, \qquad L^{\mathrm{cont}}(V) = \epsilon.$$

Moreover, on every sample path in the support of the data distribution,

$$Y_V^{\mathrm{cont}}(\xi) = \epsilon, \qquad Y_V(\xi) = 1.$$

Proof of Proposition D.1. For any feasible nonanticipative policy π, conditional centering gives $\mathbb{E}[m_t^{v,\mathrm{cont}}(a_t^\pi, \xi_{t+1}) \mid \mathcal{F}_t] = 0$. Its decisions are also feasible for the perfect-information problem, so $L^{\mathrm{cont}}(v) \le \mathbb{E}[C^\pi + M^{v,\mathrm{cont},\pi}] = J(\pi)$. Taking the infimum over π proves weak duality.

For the separation, fix $\epsilon \in (0,1)$, let ξ_1 be deterministic, and let $Z := \xi_2$ and $W := \xi_3$ be independent Bernoulli$(1/2)$ random variables. Stage one has a single feasible decision, $c_1 = 0$, and $s_2 = Z$. At stage two, the policy observes Z and chooses $x \in [0,1]$ before observing W, incurring

$$c_2(s,x,w) = 2\epsilon s + 4(1-\epsilon)(x-w)^2.$$

Both decision sets are compact polyhedra. On the relevant domain $s, x \in [0,1]$ and $w \in \{0,1\}$, this cost is smooth and convex, and lies between zero and $2\epsilon + 4(1-\epsilon) \le 4$.

The optimal causal value. Independence of W and Z gives $\mathbb{E}[(x-W)^2 \mid Z] = (x-\frac{1}{2})^2 + \frac{1}{4}$ for a decision made before W. Since $V_3 = 0$, the Bellman calculation is

$$Q_2^V(s,x) = 2\epsilon s + (1-\epsilon) + 4(1-\epsilon)(x-\tfrac{1}{2})^2,$$
$$V_2(s) = \min_{x\in[0,1]} Q_2^V(s,x) = 2\epsilon s + 1 - \epsilon.$$

The minimum is attained at $x = 1/2$. Thus V_2 is nonconstant and $V^\star = \mathbb{E}[V_2(Z)] = 1$.

Centering only the continuation value. Using these exact values gives

$$m_1^{V,\mathrm{cont}} = \mathbb{E}[V_2(Z)] - V_2(Z) = \epsilon(1-2Z), \qquad m_2^{V,\mathrm{cont}} = 0,$$

where the second equality follows from $V_3 = 0$. Adding the penalties to the realized cost cancels the $2\epsilon Z$ term:

$$C + M^{V,\mathrm{cont}} = \epsilon + 4(1-\epsilon)(x-W)^2.$$

The perfect-information optimizer can choose $x = W$, so $Y_V^{\mathrm{cont}}(\xi) = \epsilon$ on every path and $L^{\mathrm{cont}}(V) = \epsilon$.

Including the current cost in centering. Since $c_1 = 0$, the first penalty is unchanged. At stage two the full penalty is $m_2^V = Q_2^V(Z, x) - c_2(Z, x, W)$, which replaces the realized stage-two cost by its conditional mean. Therefore

$$C + M^V = \epsilon(1 - 2Z) + Q_2^V(Z, x) = 1 + 4(1 - \epsilon)(x - \tfrac{1}{2})^2.$$

This expression is minimized at $x = 1/2$, giving $Y_V(\xi) = 1$ on every path. Centering the current cost removes the perfect-information advantage of matching the decision to the shock W. □

D.3.1 A REFINED PENALTY-TIGHTNESS BOUND

A uniform value-error bound treats errors at different successor states separately. Centering can cancel part of this error, and the extra cost of a suboptimal decision can offset what remains. The following result accounts for both effects.

Fix the versions of $\overline{K}_t$, Q_t^V, and V_t used in (D.6). Let $\mathcal{R}_t^{\mathrm{PI}}$ consist of the stage-t queries generated by a history in the support of the data distribution and a feasible perfect-information decision prefix through stage t, so it is contained in the conservative band domain $\mathcal{A}_t$. Set $\Delta_t^v = v_t - V_t$ for $t = 2, \ldots, T + 1$, with $\Delta_{T+1}^v = 0$, and define

$$\begin{aligned} A_t^V(a) &:= Q_t^V(a) - V_t(s, h) \geq 0, \\ \iota_t^v(a, w) &:= \Delta_{t+1}^v(f_t(s, x, w), h \oplus w) - \int \Delta_{t+1}^v(f_t(s, x, \bar{w}), h \oplus \bar{w})\, \overline{K}_t(\mathrm{d}\bar{w} \mid \varphi_t(h)), \\ \zeta_t(v) &:= \sup_{a \in \mathcal{R}_t^{\mathrm{PI}}} \sup_{w \in \operatorname{supp} \overline{K}_t(\cdot \mid \varphi_t(h))} [\, \iota_t^v(a, w) - A_t^V(a) \,]_+. \end{aligned} \tag{D.22}$$

Here A_t^V is the suboptimality of the candidate decision in the Bellman recursion, and ι_t^v is the centered fluctuation in continuation-value error across successor uncertainty realizations. The quantity $\zeta_t(v)$ records the largest amount by which the centered value error exceeds the decision's suboptimality. Only this excess can lower the penalized objective below the exact Bellman benchmark. The pointwise suprema make the bound simultaneous for every perfect-information candidate on almost every exogenous sample path.

Proposition D.2 (Refined penalty tightness)**.** *Suppose V_1, Y_v, and the residuals below are measurable and integrable, and $\zeta_t(v) < \infty$ for every stage. On almost every exogenous sample path, every pathwise feasible perfect-information decision sequence satisfies*

$$\begin{aligned} C(x, \xi) + M^v(x, \xi) &= V_1(s_1, h_1) + \sum_{t=1}^{T} \{A_t^V(a_t) - \iota_t^v(a_t, \xi_{t+1})\}, \\ 0 \leq V^\star - L(v) &\leq \sum_{t=1}^{T} \zeta_t(v), \qquad \zeta_t(v) \leq 2 \left\| v_{t+1} - V_{t+1} \right\|_\infty, \quad t = 1, \ldots, T. \end{aligned} \tag{D.23}$$

Here each sup norm is over successor state–history pairs generated by queries in $\mathcal{A}_t$ and successor uncertainty realizations in the support of their conditional distributions. If every $\zeta_t(v)$ is zero, then $Y_v = V_1$ almost surely and $L(v) = V^\star$. This conclusion does not require $v = V$. These are unobservable benchmark quantities used only to analyze tightness, not inputs to the certificate.

Proof of Proposition D.2. Work on the probability-one event on which every successor uncertainty realization belongs to the fixed support of its conditional distribution. Fix any pathwise feasible perfect-information decision sequence on this event. From (D.1)–(D.2), $m_t^v - m_t^V = -\iota_t^v$. For the Bellman penalty,

$$c_t + m_t^V = V_t(s_t, h_t) + A_t^V(a_t) - V_{t+1}(s_{t+1}, h_{t+1}).$$

Summing and using $V_{T+1} = 0$ proves the residual identity in (D.23). On this event, every residual term is at least $-\zeta_t(v)$ by (D.22), so $Y_v \geq V_1 - \sum_t \zeta_t(v)$ pathwise. Take expectations and combine $\mathbb{E}[V_1] = V^\star$ with weak duality to obtain the middle inequality in (D.23).

Finally, $|\iota_t^v| \leq 2 \left\| v_{t+1} - V_{t+1} \right\|_\infty$ and $A_t^V \geq 0$ on every feasible query. Thus $[\iota_t^v - A_t^V]_+ \leq |\iota_t^v|$, proving the last inequality. If every defect is zero, then $Y_v \geq V_1$ pathwise, and weak duality and integrability force equality almost surely and therefore $L(v) = V^\star$. □

Proof of Theorem D.1. Condition on a calibration realization in $\mathcal{E}_{\mathrm{cal}}$ and let $\mu_G = \mathbb{E}[G \mid \mathcal{G}_{\mathrm{cal}}]$. The calibration contract (D.7) gives $\mu_G \geq J(\widehat{\pi}) - V^\star$. The endpoints A_G, B_G are fixed under this conditioning, and the path statistics are i.i.d. under the joint distribution of sample paths and solver seeds. If $W_G = 0$, the statistic is constant almost surely and both radii vanish. Otherwise, Hoeffding gives

$$\mathbb{P}\{|\overline{G} - \mu_G| > R_{\mathrm{H}}(\alpha) \mid \mathcal{G}_{\mathrm{cal}}\} \leq \alpha.$$

For empirical Bernstein, apply Maurer & Pontil (2009, Theorem 4) to $(G - A_G)/W_G$ to obtain

$$\mathbb{P}\{\mu_G > \overline{G} + R_{\mathrm{EB}}(\alpha) \mid \mathcal{G}_{\mathrm{cal}}\} \leq \alpha.$$

Integrating over calibration proves (D.16) for either fixed reporting rule. Replacing an output by $+\infty$ on an operational failure preserves this upper-coverage statement. No conditioning on successful reporting is used.

For completeness, the band construction's mean inequality follows directly from the simultaneous band. Put $e_t = Q_t - Q_t^- \in [0, r_t]$ along the policy. Under every adapted policy the penalty increments satisfy $\mathbb{E}[m_t^- \mid \mathcal{F}_t \vee \mathcal{G}_{\mathrm{cal}}] = -e_t \leq 0$, so $L^- := \mathbb{E}[Y^- \mid \mathcal{G}_{\mathrm{cal}}] \leq V^\star$. Feasibility of the policy for the inner problem gives $G^{\mathrm{band}} \geq 0$. If the path statistic is evaluated without numerical error, while retaining the certified inner lower bound $\underline{Y}^-$, then

$$\mu_G = J(\widehat{\pi}) - L^- + \mathbb{E}\left[Y^- - \underline{Y}^- + \sum_t (r_t - e_t) \,\middle|\, \mathcal{G}_{\mathrm{cal}} \right].$$

Every term inside the expectation is nonnegative. The binary construction's mean inequality and range are proved in Corollary D.4. That argument establishes calibration before invoking the present concentration result. □

Proof of Corollary D.1. The exact-centered information relaxation generated by $\widehat{v}$ has value $L(\widehat{v})$ satisfying

$$0 \leq V^\star - L(\widehat{v}) \leq 2 \sum_t \varepsilon_{v,t+1}$$

under the additional Bellman and approximation hypotheses. For the band construction, on a fixed exogenous path $r_t(z_t)$ does not depend on the candidate control. Thus

$$Y^- \geq Y_{\widehat{v}} - \sum_t r_t(z_t), \qquad V^\star - L^- \leq 2 \sum_t \varepsilon_{v,t+1} + 2 \sum_t \beta_t^{\mathrm{av}}.$$

The identity for μ_G in the proof of Theorem D.1, together with $r_t - e_t \leq r_t$, the inner-solver error bound $\varepsilon_{\mathrm{IP}}$, and the upward numerical error bound $\varepsilon_{\mathrm{path}}$ for the path statistic, yields

$$0 \leq \mu_G - [J(\widehat{\pi}) - V^\star] \leq 2 \sum_t \varepsilon_{v,t+1} + 4 \sum_t \beta_t^{\mathrm{av}} + \varepsilon_{\mathrm{IP}} + \varepsilon_{\mathrm{path}}. \tag{D.24}$$

For the binary construction, fix an exogenous path ξ. Let F_p denote the objective $F_{\widehat{p}}$ in (D.12) with $\widehat{p}$ replaced by the true probability vector p. Suppressing the path argument in both objectives, define

$$D_{\widehat{p}}^{\mathrm{exact}} = F_{\widehat{p}}(\widehat{\pi}) - \inf_{x \in \mathcal{X}^{\mathrm{PI}}(\xi)} F_{\widehat{p}}(x), \qquad D_p = F_p(\widehat{\pi}) - \inf_{x \in \mathcal{X}^{\mathrm{PI}}(\xi)} F_p(x).$$

These gaps use estimated and true probabilities, respectively, but both use exact inner infima rather than solver lower bounds. On the calibration event $\{\|\widehat{p} - p\|_2 \leq r_2\}$, for every feasible relaxed x,

$$|\{F_{\widehat{p}}(x) - F_{\widehat{p}}(\widehat{\pi})\} - \{F_p(x) - F_p(\widehat{\pi})\}| \leq \kappa.$$

Taking infima of these relative objectives gives $|D_{\widehat{p}}^{\mathrm{exact}} - D_p| \leq \kappa$ without requiring attainment. If $h = \inf_x F_{\widehat{p}} - \underline{Y}_{\widehat{p}}$, then $0 \leq h \leq \varepsilon_{\mathrm{IP}}$ and the computed statistic obeys

$$D_p \leq G \leq D_p + 2\kappa + \varepsilon_{\mathrm{IP}} + \varepsilon_{\mathrm{path}}.$$

Since $\mathbb{E}[D_p \mid \mathcal{G}_{\mathrm{cal}}] = J(\widehat{\pi}) - L(\widehat{v})$, its mean width satisfies

$$0 \leq \mu_G - [J(\widehat{\pi}) - V^\star] \leq 2 \sum_t \varepsilon_{v,t+1} + 2\mathbb{E}[\kappa \mid \mathcal{G}_{\mathrm{cal}}] + \varepsilon_{\mathrm{IP}} + \varepsilon_{\mathrm{path}}. \tag{D.25}$$

These are the two values of $\mathcal{C}_{\mathrm{cal}}$ in (D.18).

The preceding bounds control the excess of the population mean μ_G over the optimality gap. We now bound how far the sample-based report can exceed μ_G. Fix a calibration realization in $\mathcal{E}_{\mathrm{cal}}$, so the evaluation statistics are i.i.d. in the fixed interval $[A_G, B_G]$. Let $U^{\mathrm{ex}}_{\alpha,\delta}$ denote the result of evaluating the sample-mean and confidence-radius formula in exact arithmetic on these statistics. The statistics themselves may already include upward numerical error; its contribution $\varepsilon_{\mathrm{path}}$ was included in (D.24) and (D.25). If $W_G = 0$, every statistic equals μ_G almost surely, both radii are zero, and $U^{\mathrm{ex}}_{\alpha,\delta} = \mu_G$. Below assume $W_G > 0$.

Hoeffding evaluation. With conditional probability at least $1-\alpha$, the two-sided bound gives $|\overline{G} - \mu_G| \le R_{\mathrm{H}}(\alpha)$. On this one event, the report both covers the mean and has bounded excess:

$$\mu_G \le U^{\mathrm{ex}}_{\alpha,\delta} = \overline{G} + R_{\mathrm{H}}(\alpha) \le \mu_G + 2R_{\mathrm{H}}(\alpha).$$

Thus the evaluation contribution is $\mathcal{T}_{\mathrm{ev}} = 2R_{\mathrm{H}}(\alpha)$, with failure budget $\epsilon_{\mathrm{ev}} = \alpha$.

Empirical Bernstein evaluation. The coverage argument controls underestimation of the mean: $\mu_G - \overline{G} \le R_{\mathrm{EB}}(\alpha)$ with conditional probability at least $1-\alpha$. To bound certificate width, we also need to control the opposite deviation, $\overline{G} - \mu_G$. For the fixed width-analysis budget η, apply the same one-sided inequality to the reflected and normalized observations

$$Z_i = \frac{B_G - G_i}{W_G} \in [0,1].$$

Their mean deviation and sample variance satisfy

$$\mathbb{E}[Z_i \mid \mathcal{G}_{\mathrm{cal}}] - \overline{Z} = \frac{\overline{G} - \mu_G}{W_G}, \qquad S_Z^2 = \frac{S_G^2}{W_G^2}.$$

Consequently, applying empirical Bernstein to the Z_i and multiplying by W_G gives $\overline{G} - \mu_G \le R_{\mathrm{EB}}(\eta)$ with conditional probability at least $1-\eta$. A union bound makes both deviations hold with conditional probability at least $1-\alpha-\eta$; independence of the two events is not needed. On their intersection,

$$\mu_G \le U^{\mathrm{ex}}_{\alpha,\delta} = \overline{G} + R_{\mathrm{EB}}(\alpha) \le \mu_G + R_{\mathrm{EB}}(\alpha) + R_{\mathrm{EB}}(\eta).$$

This proves the choices of $\mathcal{T}_{\mathrm{ev}}$ and ϵ_{ev} in (D.19). The additional budget η is used only to analyze width; the reported certificate still uses α.

Numerical errors and the joint guarantee. The implemented report rounds upward, with

$$U^{\mathrm{ex}}_{\alpha,\delta} \le U_{\alpha,\delta} \le U^{\mathrm{ex}}_{\alpha,\delta} + \varepsilon_{\mathrm{report}}.$$

Combine the appropriate evaluation bound with (D.24) or (D.25). This yields

$$U_{\alpha,\delta} \le J(\widehat{\pi}) - V^\star + 2\sum_t \varepsilon_{v,t+1} + \mathcal{C}_{\mathrm{cal}} + \mathcal{T}_{\mathrm{ev}} + \varepsilon_{\mathrm{IP}} + \varepsilon_{\mathrm{path}} + \varepsilon_{\mathrm{report}}.$$

The last two terms sum to at most $\varepsilon_{\mathrm{num}}$, so numerical error is counted once. The lower inequality in each evaluation event also gives $J(\widehat{\pi}) - V^\star \le \mu_G \le U_{\alpha,\delta}$. Adding the calibration failure probability δ proves (D.20) with total failure budget $\delta + \epsilon_{\mathrm{ev}}$.

The finite-width conclusion uses the extra assumption that computation succeeds almost surely on $\mathcal{E}_{\mathrm{cal}}$. Under this assumption, the algorithm reports a finite value almost surely on the event used above. Reporting $+\infty$ on failure preserves coverage, but by itself cannot give a finite-width guarantee.

Refined tightness and calibration rates. To obtain the refined value-error bound, replace the initial estimate $V^\star - L(\widehat{v}) \le 2\sum_t \varepsilon_{v,t+1}$ by $V^\star - L(\widehat{v}) \le \sum_t \zeta_t(\widehat{v})$ from Proposition D.2; the remaining steps are unchanged. For the band construction, averaging a nonnegative radius over the history-summary distribution gives $\beta_t^{\mathrm{av}} \le \bar{\beta}_t^{\mathrm{rate}}$. Under the additional assumptions of Corollary C.1, the histogram rate holds on the event that every relevant cell receives enough observations. Its failure probability δ_{occ} must be added to the joint budget, giving $\delta + \epsilon_{\mathrm{ev}} + \delta_{\mathrm{occ}}$ when that rate is substituted. □

Proof of Corollary D.2. Proposition A.1 applies to every parameter and hence every random checkpoint selected using the training and policy selection block. Conditional on the training information $\mathcal{G}_{\mathrm{fit}}$, the policy and all certification choices are fixed, so Theorem D.1 gives coverage with failure probability at most $\delta + \alpha$ for the band or binary paired statistic, with the Hoeffding or empirical Bernstein rule fixed before calibration. On $\mathcal{E}_v$, decompose

$$J(\widehat{\theta}) - V^\star = [J(\widehat{\theta}) - J^\star_\Theta] + \Delta_\Theta$$

and substitute into (D.20). By Corollary D.1, conditional failure of this width statement is at most $\delta + \epsilon_{\mathrm{ev}}$ for each such realization of the training information. Integrating and adding $\mathbb{P}(\mathcal{E}_v^c) \le \gamma$ proves (D.21). Under the additional hypotheses of Theorem B.4, its bound $\Delta_\Theta \le K_T^\pi \varepsilon_\pi$ can be substituted into the same decomposition. □

D.3.2 Anytime-valid certification across checkpoints

The preceding guarantee covers one policy fixed before calibration. During further training, we may certify several checkpoints and stop after observing their reported bounds. Allocating failure probabilities across fresh calibration and evaluation blocks so that their sum is bounded gives simultaneous validity for all attempted reports. This is anytime validity over the checkpoint index: coverage is preserved when the reported history determines when to stop, in the sense of the confidence-sequence literature (Howard et al., 2021).

Let $\mathcal{F}_{k-1}^{\mathrm{data}}$ denote the information available before calibrating checkpoint k. Using data made available for training and policy selection after earlier reports, train the policy and fix it together with the continuation approximation, calibration construction, evaluation rule, and solver settings. Reserve fresh blocks for calibration and evaluation. After reporting, add both blocks to the training and policy selection pool for later checkpoints.

Corollary D.3 (Anytime-valid bounds across checkpoints)**.** *Let $(\alpha_k, \delta_k)_{k\ge1}$ be deterministic positive sequences with*

$$\sum_{k\ge1}(\alpha_k + \delta_k) \le \bar{\alpha}.$$

Before opening the kth calibration block, suppose the decision to attempt that checkpoint, its calibration and evaluation block sizes, the checkpoint itself, and all certification settings are $\mathcal{F}_{k-1}^{\mathrm{data}}$-measurable. Conditional on $\mathcal{F}_{k-1}^{\mathrm{data}}$, suppose every attempted checkpoint satisfies all conditions for validity in Theorem D.1 for its chosen construction and evaluation rule, including fresh reserved calibration and evaluation blocks, with error probabilities (α_k, δ_k). Then, with probability at least $1 - \bar{\alpha}$, every attempted checkpoint bound is valid simultaneously. In particular, the bound remains valid at any data-dependent checkpoint selected using the reported history.

The guarantee applies simultaneously to reported checkpoints, not to every sample size. The prespecified allocation schedule assigns the total failure probability across reports. For example, $\alpha_k = \delta_k = \bar{\alpha}/[2k(k+1)]$ satisfies the required sum. Whether to attempt a report must be decided before inspecting its calibration or evaluation block. Stopping adaptively within a current block would instead require calibration and evaluation guarantees that remain valid after arbitrary within-block stopping.

In particular, let U_k be the bound reported for $\widehat{\pi}_k$, setting $U_k = +\infty$ if that checkpoint is skipped or produces no report. Stop at $K_\epsilon = \inf\{k : U_k \le \epsilon\}$, with $\inf \varnothing = \infty$. Then

$$\mathbb{P}\{K_\epsilon < \infty,\ J(\widehat{\pi}_{K_\epsilon}) - V^\star > \epsilon\} \le \bar{\alpha}.$$

The decision to stop may therefore depend on all completed reports. This controls false declarations of ϵ-optimality but does not guarantee that the stopping criterion will eventually be met.

Proof of Corollary D.3. Apply the conditional coverage statement at checkpoint k and integrate over $\mathcal{F}_{k-1}^{\mathrm{data}}$. A countable union bound shows that the probability of any attempted-checkpoint failure is at most $\sum_k(\alpha_k + \delta_k) \le \bar{\alpha}$. On the complementary event all attempted-checkpoint bounds hold, including the one selected by any stopping rule based on the reported history. □

D.3.3 A COMPUTABLE RANGE BOUND FOR THE BAND STATISTIC

Proposition D.3 (Range bound for the band-based statistic)**.** *Suppose, for every stage t, that*

$$\begin{aligned} |c_t(s,x,w)| \le C_t, \quad H_t(a,w) \in [\ell_t^H, u_t^H] \quad & \forall\, a=(s,x,h) \in \mathcal{A}_t,\ w \in \operatorname{supp} \overline{K}_t(\cdot \mid \varphi_t(h)), \\ \widehat{Q}_t(a,z) \in [\ell_t^H, u_t^H] \quad & \forall\, (a,z) \in \mathcal{A}_t \times \mathcal{Z}_t, \\ \beta_t(z) \le \bar{\beta}_t^{\mathrm{env}} \quad & \forall\, z \in \mathcal{Z}_t. \end{aligned}$$

Let $W_t = u_t^H - \ell_t^H$. For each evaluation path ξ^i, write $Y_i^- = Y^-(\xi^i)$ and $\underline{Y}_i = \underline{Y}^-(\xi^i)$, and assume $0 \le Y_i^- - \underline{Y}_i \le \varepsilon_{\mathrm{IP}}$. Then the calibration-measurable bound

$$B_G = 2\sum_{t=1}^{T} \big(C_t + W_t + \bar{\beta}_t^{\mathrm{env}}\big) + \varepsilon_{\mathrm{IP}} \tag{D.26}$$

satisfies $0 \le G_i^{\mathrm{cert}} \le B_G$.

Because $\mathcal{A}_t$ contains every query available to the perfect-information optimizer, the displayed domain covers both the selected policy and every candidate in the inner problem on almost every path in the support of the data distribution. To use this bound in Algorithm 2, the inner-solver error cap must hold almost surely for a fresh sample path and solver seed, not merely on the observed evaluation sample. Any additional uniform upward error in computing the statistic must also be added to B_G before evaluation.

Proof. On a fixed sample path ξ, write $\underline{Y} = \underline{Y}^-(\xi)$ and leave the path arguments implicit in $C^{\widehat{\pi}}$, $M^{-,\widehat{\pi}}$, and Y^-. Define the unshifted penalized objective

$$\widetilde{F}(x,\xi) = C(x,\xi) + \sum_{t=1}^{T} \{\widehat{Q}_t(a_t,z_t) - H_t(a_t,\xi_{t+1})\}.$$

Because $\widehat{Q}_t$ and H_t lie in the same interval,

$$|\widehat{Q}_t - H_t| \le W_t, \qquad |\widetilde{F}| \le A_0 := \sum_{t=1}^{T} (C_t + W_t).$$

On this fixed path, the radius $\beta_t(z_t)$ does not depend on the decisions. Hence

$$C + M^- = \widetilde{F} - \sum_{t=1}^{T} \beta_t(z_t), \qquad Y^- = \inf_{x \in \mathcal{X}^{\mathrm{PI}}(\xi)} \widetilde{F}(x,\xi) - \sum_{t=1}^{T} \beta_t(z_t).$$

The common shift cancels when the selected policy's objective is compared with Y^-. The bound on the inner-solver error therefore gives

$$C^{\widehat{\pi}} + M^{-,\widehat{\pi}} - \underline{Y} \le 2A_0 + \varepsilon_{\mathrm{IP}}.$$

Finally, $\sum_t r_t(z_t) = 2\sum_t \beta_t(z_t) \le 2\sum_t \bar{\beta}_t^{\mathrm{env}}$, which proves (D.26). Nonnegativity follows because the selected policy is feasible for the perfect-information problem, $\underline{Y} \le Y^-$, and $r_t \ge 0$. □

D.4 INDEPENDENT-BINARY PAIRED CORRECTION

We now verify the direct gap correction from Appendix D.2 and derive its range for empirical Bernstein evaluation.

Corollary D.4 (Independent-binary instance of Algorithm 2)**.** *Use the independent-binary calibration construction of Appendix D.2, with the policy, continuation approximation, and complete query-domain contrast bounds fixed before using fresh independent calibration sample paths. Let D, κ, and r_2 be defined by* (D.11)–(D.13)*. Set*

$$K = r_2 \Big(\sum_t (M_t - m_t)^2 \Big)^{1/2}.$$

Assume integrable costs and inner values and a certified range bound $0 \le D \le B_{\mathrm{raw}}$, fixed before evaluation and including the prescribed bounds on solver and arithmetic errors. Use fresh evaluation paths and the measurable solver/seed conditions of Theorem D.1. Then the binary construction satisfies (D.7) *and admits $A_G = K/2$, $B_G = B_{\mathrm{raw}} + K$. In particular, Algorithm 2 with empirical Bernstein evaluation reports*

$$U^{\mathrm{bin}}_{\alpha,\delta} = \overline{G} + \sqrt{\frac{2S_G^2 \log(2/\alpha)}{n}} + \frac{7(B_{\mathrm{raw}} + K/2)\log(2/\alpha)}{3(n-1)}, \qquad n = N_{\mathrm{ev}} \ge 2, \tag{D.27}$$

with conditional coverage at least $1 - \delta - \alpha$ given the training information $\mathcal{G}_{\mathrm{fit}}$. The corresponding width statement is (D.20) *with $\mathcal{C}_{\mathrm{cal}} = 2\mathbb{E}[\kappa \mid \mathcal{G}_{\mathrm{cal}}]$ and the empirical Bernstein tail allocation in* (D.19)*, under the additional hypotheses of Corollary D.1.*

Proof. Condition first on the training information $\mathcal{G}_{\mathrm{fit}}$, so the policy, continuation approximation, and contrast bounds are fixed.

1. Control the probability-estimation error. Write $e = \widehat{p} - p$. Each coordinate is an average of N_{cal} independent centered Bernoulli observations. Hoeffding's lemma and independence across stages give, for every $u \in \mathbb{R}^T$,

$$\mathbb{E}\exp(u^\top e) = \prod_{t=1}^{T} \mathbb{E}\exp(u_t e_t) \le \exp\left(\frac{\|u\|_2^2}{8N_{\mathrm{cal}}}\right).$$

Thus the subgaussian variance proxy is $1/(4N_{\mathrm{cal}})$. Applying Hsu et al. (2012, Theorem 2.1) with the identity matrix gives the radius r_2 in (D.11). In particular, the calibration event $\mathcal{E}_{\mathrm{bin}} := \{\|e\|_2 \le r_2\}$ has conditional probability at least $1 - \delta$.

2. Compare the true and estimated gaps on one path. Fix a calibration realization in $\mathcal{E}_{\mathrm{bin}}$ and an evaluation path ξ. All infima and suprema in this step are over the same feasible set $\mathcal{X}^{\mathrm{PI}}(\xi)$; we suppress ξ in the objective arguments. Let F_p be $F_{\widehat{p}}$ with the estimated probabilities replaced by the true probabilities, and define $D_p = F_p(\widehat{\pi}) - \inf_x F_p(x)$. For a candidate sequence x, write $\Delta(x) = (\Delta_t(a_t))_{t=1}^T$. The binary conditional mean is linear in its probability, so

$$F_{\widehat{p}}(x) = F_p(x) + e^\top \Delta(x).$$

The certified inner lower value satisfies $\underline{Y}_{\widehat{p}} \le \inf_x F_{\widehat{p}}(x) \le F_{\widehat{p}}(x)$ for every candidate x. Consequently, $F_{\widehat{p}}(\widehat{\pi}) - F_{\widehat{p}}(x) \le D$. Substituting the preceding objective identity now gives, for each x,

$$\begin{aligned} F_p(\widehat{\pi}) - F_p(x) &= F_{\widehat{p}}(\widehat{\pi}) - F_{\widehat{p}}(x) + e^\top\{\Delta(x) - \Delta(\widehat{\pi})\} \\ &\le D + e^\top\{\Delta(x) - \Delta(\widehat{\pi})\}. \end{aligned}$$

Taking the supremum over x on both sides yields

$$D_p - D \le \sup_x e^\top\{\Delta(x) - \Delta(\widehat{\pi})\}.$$

This argument uses only a certified lower value; it does not require an exact inner solve or attainment of either infimum.

3. Bound the remaining perturbation by κ. For each stage, put

$$b_t(\xi) = \max\{\Delta_t(a_t^{\widehat{\pi}}) - m_t, M_t - \Delta_t(a_t^{\widehat{\pi}})\}.$$

Both the policy query and every candidate query have contrasts in $[m_t, M_t]$. Hence $|\Delta_t(a_t) - \Delta_t(a_t^{\widehat{\pi}})| \le b_t(\xi)$. Cauchy–Schwarz and the calibration event therefore give

$$\begin{aligned} \left|e^\top\{\Delta(x) - \Delta(\widehat{\pi})\}\right| &\le \|e\|_2 \, \|\Delta(x) - \Delta(\widehat{\pi})\|_2 \\ &\le r_2\Big(\sum_t b_t(\xi)^2\Big)^{1/2} = \kappa(\xi). \end{aligned}$$

The contrast bounds cover every relaxed query, so the same calibration event controls all candidates on every relevant path. Combining this bound with step 2 proves $D_p \le D + \kappa = G$.

4. Pass from the pathwise comparison to the optimality gap. Hold the calibration data fixed and take expectations over a fresh path and any independent solver seed. Exact centering gives $\mathbb{E}[F_p(\widehat{\pi}, \xi) \mid \mathcal{G}_{\mathrm{cal}}] = J(\widehat{\pi})$, while weak duality gives $\mathbb{E}[\inf_x F_p(x, \xi) \mid \mathcal{G}_{\mathrm{cal}}] \le V^\star$. Thus, on $\mathcal{E}_{\mathrm{bin}}$,

$$\mathbb{E}[G \mid \mathcal{G}_{\mathrm{cal}}] \ge \mathbb{E}[D_p \mid \mathcal{G}_{\mathrm{cal}}] \ge J(\widehat{\pi}) - V^\star.$$

Together with step 1, this proves the calibration contract. It does not assert that the empirical inner value alone is a lower benchmark for $V^\star$.

5. Establish the range and apply the evaluation bound. A point in $[m_t, M_t]$ has maximum distance to its two endpoints between $(M_t - m_t)/2$ and $M_t - m_t$. Applying this fact to the policy contrast in b_t yields

$$K/2 \le \kappa \le K.$$

Since $0 \le D \le B_{\mathrm{raw}}$, the corrected statistic has endpoints $A_G = K/2$ and $B_G = B_{\mathrm{raw}} + K$. Its interval width is therefore $W_G = B_{\mathrm{raw}} + K/2$. Conditional on calibration, these endpoints are fixed and the fresh evaluation statistics are i.i.d. under the stated solver conditions. The empirical Bernstein rule in Theorem D.1 gives (D.27), using the sample variance of the complete statistic $G = D + \kappa$. Combining the evaluation failure probability α with the calibration failure probability δ proves the stated coverage.

6. Account for certificate width. Under the additional width hypotheses, let $D^{\mathrm{exact}}_{\widehat{p}} = F_{\widehat{p}}(\widehat{\pi}) - \inf_x F_{\widehat{p}}(x)$. Step 3 bounds the difference between every true and estimated policy–candidate gap in absolute value. Taking suprema of those gaps in both directions gives $|D^{\mathrm{exact}}_{\widehat{p}} - D_p| \le \kappa$. Replacing the exact inner infimum by its certified lower value adds between zero and $\varepsilon_{\mathrm{IP}}$ to the gap. After including the prescribed upward path-statistic error, this gives

$$D_p \le G \le D_p + 2\kappa + \varepsilon_{\mathrm{IP}} + \varepsilon_{\mathrm{path}}.$$

Thus calibration contributes at most $2\mathbb{E}[\kappa \mid \mathcal{G}_{\mathrm{cal}}]$ to the mean excess. The continuation error, evaluation tails, and final rounding are then handled as in Corollary D.1, proving the stated width specialization. □

E A Worked Capacitated-Inventory Example of the Certificate

This example shows how the abstract certificate inputs arise from familiar inventory-model ingredients: physical capacity bounds and a Hölder regularity bound for the conditional distribution of next-period demand. Let the scalar demand process satisfy $w_t \in [0, \bar{d}]$, where w_t denotes ξ_t in the general model. Let $z_t = w_t$ be the history summary, and assume it is predictively sufficient:

$$\mathbb{P}\{w_{t+1} \in A \mid \mathcal{F}_t\} = \overline{K}_t(A \mid z_t)$$

for the unknown conditional distribution $\overline{K}_t(\cdot \mid z_t)$ from (A.1). The known range of the history summary is $\mathcal{Z}_t = [0, \bar{d}]$. Let s_t denote inventory minus backlog before ordering. Assume $k_t^{\mathrm{ord}}, k_t^{\mathrm{hold}}, k_t^{\mathrm{back}} \ge 0$. With order capacity $\bar{u}_t > 0$ and storage capacity $\bar{S} \ge 0$, fix a known $\underline{s}_1 < \bar{S}$ such that $\underline{s}_1 \le s_1 \le \bar{S}$ almost surely, and set

$$\mathcal{X}_t(s_t) = \{x : 0 \le x \le \bar{u}_t,\ s_t + x \le \bar{S}\}, \qquad s_{t+1} = s_t + x_t - w_{t+1},$$

and use the piecewise-linear cost

$$c_t(s, x, w) = k_t^{\mathrm{ord}} x + k_t^{\mathrm{hold}}(s + x - w)_+ + k_t^{\mathrm{back}}(w - s - x)_+.$$

The decision set has the moving right-hand side $x \le \bar{S} - s$. Relatively complete recourse is immediate on the reachable domain because $x = 0$ is feasible whenever $s \le \bar{S}$, and the scalar feasibility layer is

$$\bar{P}_t(u, s) = \min\{\bar{u}_t, \bar{S} - s, \max\{0, u\}\}.$$

Relaxation query set and its cover. Define $\underline{s}_{t+1} = \underline{s}_t - \bar{d}$. Every feasible perfect-information or nonanticipative trajectory then satisfies $\underline{s}_t \le s_t \le \bar{S}$. A conservative relaxation query set is

$$\mathcal{A}_t = \left\{(s, x) : \underline{s}_t \le s \le \bar{S}, \quad 0 \le x \le \bar{u}_t, \quad s + x \le \bar{S}\right\}.$$

Although the general query includes history, the Markov response below depends on the query only through (s, x), while z_t is displayed separately as the history summary used for conditioning. No history-space cover is therefore needed. With $D_{s,t} = \bar{S} - \underline{s}_t > 0$ and

$$d_{\mathcal{A},t}((s,x),(\bar{s},\bar{x})) = \max\left\{\frac{|s-\bar{s}|}{D_{s,t}}, \frac{|x-\bar{x}|}{\bar{u}_t}\right\},$$

choosing one query from each nonempty cell of a rectangular grid yields an ϵ-net satisfying

$$|\mathcal{A}_{t,\epsilon}| \le (3/\epsilon)^2, \qquad 0 < \epsilon \le 1.$$

Thus the query-cover growth bound in Corollary C.1 holds with $p_t = 2$, independently of the history length.

Continuation approximation and band constants. Use an affine continuation approximation fitted on $\mathcal{D}_{\text{fit}}$ and depending on history only through the current demand z:

$$\widehat{v}_{t+1}(s,z) = a_{t+1}s + g_{t+1}z + d_{t+1}, \qquad \widehat{v}_{T+1} = 0,$$

and fix its coefficients before calibration. The response in (D.1) then reduces to

$$H_t((s,x),w) = c_t(s,x,w) + \widehat{v}_{t+1}(s+x-w,w).$$

Constraining the continuation coefficients to a fixed compact box guarantees finite constants for the range bound. Define bounds on the magnitude of the continuation approximation and the current-stage cost by

$$V_{t+1}^{\max} := |a_{t+1}| \max\{|\underline{s}_{t+1}|, \bar{S}\} + |g_{t+1}|\bar{d} + |d_{t+1}|,$$
$$C_t := k_t^{\text{ord}}\bar{u}_t + k_t^{\text{hold}}\bar{S} + k_t^{\text{back}}(-\underline{s}_{t+1})_+.$$

On every certified query and every demand $w \in [0, \bar{d}]$,

$$-V_{t+1}^{\max} \le H_t \le C_t + V_{t+1}^{\max}.$$

Hence one may declare

$$\ell_t^H = -V_{t+1}^{\max}, \qquad u_t^H = C_t + V_{t+1}^{\max}, \qquad W_t = C_t + 2V_{t+1}^{\max}.$$

Tighter response bounds over the known polyhedral domain can be computed by linear programs for the finitely many linear pieces.

A valid query-Lipschitz constant for the scaled metric is

$$L_{a,t} = (k_t^{\text{hold}} + k_t^{\text{back}} + |a_{t+1}|)D_{s,t} + (k_t^{\text{ord}} + k_t^{\text{hold}} + k_t^{\text{back}} + |a_{t+1}|)\bar{u}_t.$$

The response is Lipschitz in successor demand with constant

$$L_{w,t} = k_t^{\text{hold}} + k_t^{\text{back}} + |a_{t+1}| + |g_{t+1}|.$$

To control how the conditional mean changes with current demand, suppose that the next-demand distributions satisfy

$$W_1\big(\overline{K}_t(\cdot \mid z), \overline{K}_t(\cdot \mid \bar{z})\big) \le K_t^{\text{W}}|z-\bar{z}|^{\tau_t}$$

where $\tau_t \in (0,1]$ and W_1 denotes the 1-Wasserstein distance between probability distributions. Thus nearby current demands induce nearby next-demand distributions. Since the response is Lipschitz in next demand, this condition implies Assumption C.1(d) with $L_{z,t} = L_{w,t}K_t^{\text{W}}$. The constant K_t^{W} is a valid certified upper bound supplied by the model assumptions and fixed before calibration. It may be derived analytically or supplied conservatively. Partition the history-summary domain $[0, \bar{d}]$ into uniform bins. Here $d_t = 1$. If the history summary has full support on this interval and a marginal density bounded below, then the cell-probability condition in Corollary C.1 holds. This density condition is sufficient for the stated uniform rate, but is not needed to compute the band or guarantee its coverage.

Reuse of observed demand and the perfect-information LP. For $z \in B$ with $N_{t,B} > 0$, index the demand sample paths in $\mathcal{D}_{\mathrm{cal}}$ by i. Writing $\widehat{Q}_t(s,x,z)$ for $\widehat{Q}_t((s,x),z)$, the estimator (C.1) becomes

$$\widehat{Q}_t(s,x,z) = \frac{1}{N_{t,B}} \sum_{i:z_t^i \in B} \left[c_t(s,x,w_{t+1}^i) + \widehat{v}_{t+1}(s+x-w_{t+1}^i, w_{t+1}^i) \right].$$

Thus every observed demand is reevaluated at every feasible inventory–order pair. On a fixed demand sample path from $\mathcal{D}_{\mathrm{ev}}$,

$$C + M^- = \sum_{t=1}^{T} \left[\widehat{Q}_t(s_t,x_t,z_t) - \beta_t(z_t) - \widehat{v}_{t+1}(s_{t+1}, w_{t+1}) \right].$$

Minimizing this objective over the physical dynamics and recourse constraints, as in (D.9), is an LP. For each calibration observation used at stage t, introduce epigraph variables satisfying

$$r_{ti}^+ \geq s_t + x_t - w_{t+1}^i, \quad r_{ti}^+ \geq 0, \qquad r_{ti}^- \geq w_{t+1}^i - s_t - x_t, \quad r_{ti}^- \geq 0,$$

and replace the positive-part terms in $\widehat{Q}_t$ by $k_t^{\mathrm{hold}} r_{ti}^+ + k_t^{\mathrm{back}} r_{ti}^-$. The physical state equations use the evaluation-path demands, while empty calibration cells contribute a constant midpoint and preserve linearity.

A dual-feasible LP objective supplies the certified lower bound $\underline{Y}$. If the solver also returns a primal feasible point and terminates only when

$$\text{primal objective} - \text{dual lower bound} \leq \varepsilon_{\mathrm{IP}},$$

then $0 \leq Y^- - \underline{Y} \leq \varepsilon_{\mathrm{IP}}$. With $\bar{\beta}_t^{\mathrm{env}} = \sup_{z \in [0,\bar{d}]} \beta_t(z)$, Proposition D.3 gives a computable range bound once the stated certified inputs and the inner-solver error bound are fixed:

$$B_G = 2 \sum_{t=1}^{T} \left(C_t + W_t + \bar{\beta}_t^{\mathrm{env}} \right) + \varepsilon_{\mathrm{IP}}.$$

Under the displayed Wasserstein–Hölder condition and the cell-probability condition in Corollary C.1, the one-dimensional history summary gives the calibration rate $\widetilde{O}(N_{\mathrm{cal}}^{-\tau/(2\tau+1)})$ for $\tau = \min_t \tau_t$, while the two-dimensional inventory–order query contributes through the logarithm of its covering number.

Implementation. For this model, Algorithm 1 trains a proposal $\mu_{\theta,t}(s,z)$ using exact clipping. After selecting the policy, we fit the affine continuation approximation and choose the band design using $\mathcal{D}_{\mathrm{fit}}$. We hold both fixed while forming the histogram estimates and radii from demand observations in $\mathcal{D}_{\mathrm{cal}}$. On each sample path in $\mathcal{D}_{\mathrm{ev}}$, we simulate the selected policy and solve the displayed LP. Its dual lower bound and verified solver gap supply the inner quantities in (D.10), with solver error at most $\varepsilon_{\mathrm{IP}}$. Combining them with the policy quantities from the same sample path gives G_i^{cert}. The expanded Algorithm 2 in Appendix D.2 then uses Hoeffding evaluation to report $U_{\alpha,\delta}$. The inputs are the stated predictive sufficiency, known support, and model-side regularity bound. The procedure estimates only the conditional means required by the penalty, rather than a full transition distribution, and constructs no scenario tree.

Conservative history-summary and query domains, response ranges, and regularity bounds preserve validity, with their looseness reflected in the reported width. A piecewise-linear or ReLU continuation approximation is also admissible when the resulting perfect-information problem is represented as a MILP and solved with certified global branch-and-bound bounds.

F Experimental Overview and Complete Main Comparisons

F.1 Experimental setup

Benchmarks. We evaluate policy training on a synthetic nonlinear hydrothermal system with 12 reservoirs and 48 stages, and on two case studies using historical hydrology records: ONS,

with four aggregate reservoirs, and CAMELS-BR (Chagas et al., 2020), with 12 catchments and semi-synthetic reservoirs. These experiments assess gradient accuracy, training performance under matched sample-path budgets, and the benefits of structured policies. To assess certification, we use a separate three-reservoir, 18-stage model with verified optimality-gap bounds and examine the coverage and tightness of the certificates for a fixed policy. This separation lets us evaluate policy optimization on the hydrology models while testing certification where the required global inner bounds and finite ranges can be verified. Appendix E provides a worked inventory certificate as a further illustration.

Baselines. FeasPG (Algorithm 1) replays sample paths under the deployed policy π_θ and computes the complete pathwise gradient $\widehat{g}_{\mathrm{PW}}$, or $\widehat{g}_\lambda$ when using regularization. We implement the proposal map $\mu_{\theta,t}$ with an MLP or one of two structured alternatives: Graph uses reservoir-network connections, while DeepSets (Zaheer et al., 2017) pools information across reservoirs using shared parameters. The closest model-based baseline is Projected Short-Horizon Actor-Critic (Projected SHAC) (Xu et al., 2022): it uses the same model and projection but replaces the full-horizon pathwise gradient with short-window gradients and a learned value approximation. Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Soft Actor-Critic (SAC) (Haarnoja et al., 2018) provide projected model-free references, with all learned methods sharing the projection. For the hydrology comparisons, we also include empirical SDDP (Pereira & Pinto, 1991; Füllner & Rebennack, 2025) as a planning benchmark and Seasonal as a fixed, nonlearned storage-feedback rule. SDDP is not used as an optimality certificate in these comparisons.

Data and evaluation protocol. The synthetic benchmarks use simulated exogenous sample paths. The hydrology case studies use models fitted to historical records, with separate held-out historical evaluation (Appendix K). Training and validation are separate subsets of $\mathcal{D}_{\mathrm{fit}}$; test paths are held out from both. Certification uses independent sets $\mathcal{D}_{\mathrm{cal}}$ and $\mathcal{D}_{\mathrm{ev}}$. For policy optimization, the benchmarks use FeasPG's estimator with Adam on fixed training pools, whereas the stationarity theorem analyzes the fresh-minibatch SGD update and a randomly selected iterate. In the policy comparisons, validation selects hyperparameters and $\widehat{\theta}$ by the lowest mean cost $C(\theta, \xi)$ among evaluated checkpoints, separately within each training budget and seed. This fixes the deployed policy $\widehat{\pi} = \pi_{\widehat{\theta}}$ before evaluation on common test paths. Validation paths are not used for gradient updates. Comparisons are made within each benchmark, with N_{fit} counting distinct training sample paths, excluding validation. The appendices report detailed settings and computational costs.

F.2 Is the gradient correct, and can the regularizer escape a plateau?

F.2.1 Gradient mechanism

An early release changes later storage, affecting both later proposals and feasible release bounds. Algorithm 1 differentiates through both effects. We check whether its gradient predicts the resulting change in cost.

On 32 fixed training paths, we slightly perturb the parameters in a chosen direction and compare central finite differences with automatic differentiation. We then remove selected derivative terms while keeping the simulated decisions and costs identical. This isolates errors in the gradient calculation. Appendix J gives the diagnostic details.

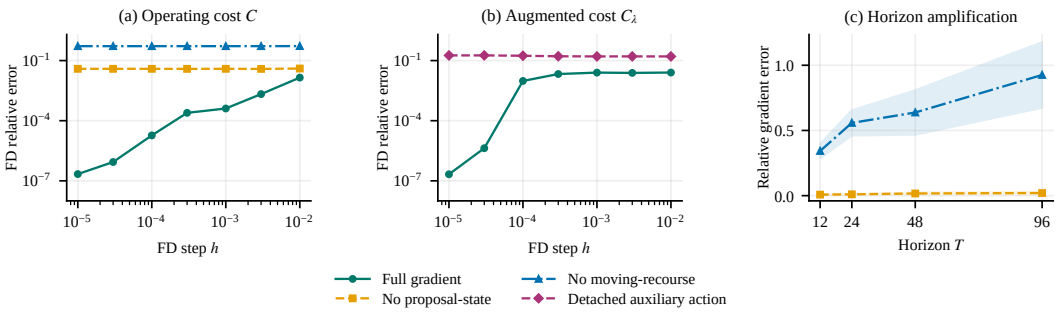

Figure 3: Checking the gradient calculation. (a) Agreement with finite differences for cost C, averaged over five seeds. (b) The same check for C_λ at $\lambda = 1$, using three fixed random directions at one update-30 checkpoint from the startup experiment in Appendix F.2.2. "Detached auxiliary action" treats x_t as constant only when differentiating d_{pd}, while derivatives of the cost and state dynamics remain complete. (c) Relative difference from the complete cost gradient as the number of stages grows, with mean and sample SD over five seeds.

Figure 3(a) shows close agreement only for the complete pathwise cost gradient. At the prespecified step $h = 10^{-5}$, its mean normalized error is 2.17×10^{-7}. The error rises to 3.88×10^{-2} when we ignore how later proposals respond to state changes ("No proposal-state"), and to 5.24×10^{-1} when we ignore how feasible bounds change with the state ("No moving-recourse").

Panel (b) checks the regularized pathwise gradient $\widehat{g}_\lambda$. Here $C_\lambda(\theta, \xi) = C(\theta, \xi) + \lambda d_{\mathrm{pd}}(\theta, \xi)$ abbreviates the sample-path loss in Algorithm 1. Treating the projected decision as constant within d_{pd} gives an error of 0.181, compared with 2.12×10^{-7} for the complete derivative at the same step size. Thus the dependence of feasible bounds on the state also matters for the gradient of the regularizer, as described by (B.8).

Panel (c) shows how omitting these effects changes the complete cost gradient over longer horizons. Without the moving-recourse derivative, the relative error increases from 0.343 at 12 stages to 0.926 at 96 stages. These checks assess derivative accuracy but do not guarantee lower cost after a fixed training budget. Appendix J reports separate training comparisons with selected derivative terms omitted throughout the policy replay.

In a matched-MLP training comparison, mean test cost is 78.423k with complete gradients and 80.872k without moving-recourse derivatives, using the full-gradient variant's selected regularization strength for both (Appendix J).

F.2.2 Effect of proposal-distance regularization

Even a correctly computed cost gradient can be zero. If a proposal exceeds available water, projection limits the actual release. Small changes to the proposal may then leave both the release and its cost unchanged. The proposal-distance term can provide an update direction: $C_\lambda = C + \lambda d_{\mathrm{pd}}$, with $d_{\mathrm{pd}} = \frac{1}{2}\sum_{t=1}^{T} \|u_t - x_t\|^2$. Here u_t is the proposed decision and x_t is the decision applied after projection. Their distance can change even while the applied decision stays fixed.

We construct this situation in the controlled benchmark's 12-reservoir, 48-stage nonlinear model with its MLP initialization. We set initial storage to 1% of capacity and inflows to 2% of nominal, leaving the other settings unchanged. A bound covering the full inflow support confirms that every initial proposal exceeds available water and the complete cost gradient is zero. This constructed stress test does not represent a historical drought.

We compare no proposal-distance term ($\lambda = 0$), a fixed weight ($\lambda = 1$), and a decreasing weight ($\lambda_k = (k+1)^{-1/2}$), with identical training budgets and checkpoint selection by validation cost.

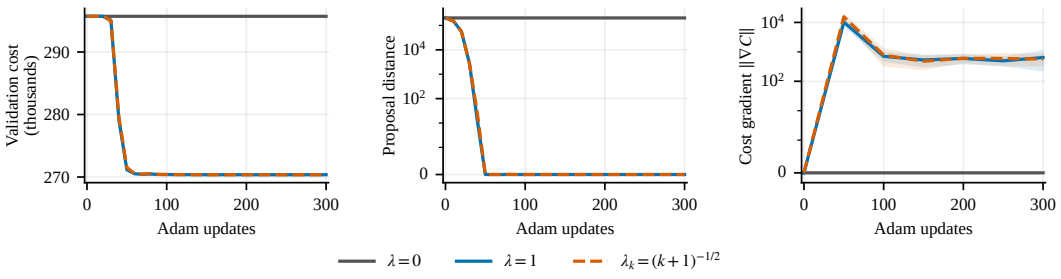

Figure 4: Training from the constructed zero-gradient initialization (12 reservoirs, 48 stages, 1% initial storage, 2% nominal inflow). Left: validation cost. Middle: distance between proposed and projected decisions. Right: cost gradient norm on 32 fixed training paths, including moving recourse. Curves show means over five seeds, with shading for sample SD. The figure displays the first 300 of 500 Adam updates, using 1,000 training paths per candidate.

In Figure 4, training without the proposal-distance term leaves decisions and costs unchanged, with exactly zero cost gradients in all five seeds. Either positive weight first reduces proposal distance (middle), after which decisions change, the cost gradient becomes nonzero (right), and validation cost falls (left). On 4,096 independent test paths, mean cost falls from 295,453 to 270,121 with either positive weight, an 8.57% reduction. The gain mainly reflects water conservation and a reduction in the low-storage component of cost, but it does not establish near-optimal dispatch.

The benefit depends on the setting. With the same prescribed weights, reductions are 15.99% and 16.16% at 5% nominal inflow, but only 0.50% and 0.91% in the original nominal case. Appendix I reports these comparisons, a no-release reference, and further mechanism checks.

For subsequent FeasPG policy-comparison experiments, we use $\lambda_k = a\lambda_{\text{ref}}/\sqrt{k+1}$, where λ_{ref} is computed from training data at initialization. Validation cost jointly selects $a \in \{0.1, 1, 10\}$, the learning rate, and the checkpoint. Appendix I gives the scale rule and tuning costs.

F.3 Sample-path efficiency

In the controlled benchmark, releases change future storage, ramp bounds, and delayed downstream inflows, so state changes affect both later proposals and their feasible sets. We use matched sample-path budgets to compare the policy-optimization methods: FeasPG+MLP and Projected SHAC share a policy network, simulator, and projection, but SHAC truncates the derivative every eight stages and uses a learned value approximation for the remaining horizon. A separate comparison between FeasPG+Graph and a matched MLP examines structured policy configurations. Validation selects each checkpoint, and all methods are assessed on the same 4,096 test paths.

Across the tested budgets, FeasPG+MLP's mean test cost is 6.13%–22.68% below that of Projected SHAC (Figure 1). Because truncation and value approximation change jointly, this comparison at matched training budgets does not isolate the causal value of full-horizon differentiation. FeasPG+Graph further reduces mean test cost by 5.28%–12.72% relative to the matched MLP and has the lowest mean test cost among the learned methods. This second comparison concerns parameter-matched policy configurations rather than architecture alone, because Graph also uses a different proposal parameterization and initialization (Appendix G). Deployment latency is of the same millisecond order for Graph and the matched MLP. Detailed cost, update, and resource comparisons appear in Appendix H.

Because this benchmark is nonconvex, the SDDP comparator solves a fixed first-order planning approximation with two joint outcomes per stage and 50 iterations, then evaluates its selected policy in the original simulator. A diagnostic finds approximately twofold overprediction of hydro generation, while 13 of 25 runs select the initial pre-cut checkpoint (Appendix H). Figure 1 therefore evaluates this specified approximate planner, not SDDP in general, and does not separate physical approximation, scenario compression, and finite training.

F.4 HYDROLOGY CASE STUDIES FITTED TO HISTORICAL RECORDS

The hydrology cases extend the comparison to aggregate energy storage and a portfolio of non-nested catchments. ONS uses FeasPG+MLP, while CAMELS-BR compares DeepSets with a parameter-matched MLP. Performance is reported on each model's own objective scale.

Exogenous data models estimated from the historical training intervals provide 500 validation paths, and five training seeds share 4,096 test paths. The reported 95% paired-bootstrap intervals resample both training seeds and test sample paths to account for both sources of uncertainty (Appendix K). Historical sequences reserved for evaluation are opened only after the exogenous data model, policy architecture, hyperparameters, and checkpoint have been fixed.

At $N_{\text{fit}} = 2500$, ONS FeasPG+MLP has 13.14% lower mean cost than SDDP (95% paired-bootstrap interval for the relative effect, $[-14.65, -11.61]\%$). Its relative effect versus Projected SHAC is -3.15% ($[-5.16, -1.25]\%$). CAMELS-BR DeepSets has 1.67% lower mean cost than its matched MLP ($[-1.86, -1.45]\%$), and its sample mean is lower at all five displayed budgets. On the single ONS historical evaluation, Seasonal costs 4.596×10^{12}, compared with $(5.161 \pm 0.098) \times 10^{12}$ for FeasPG+MLP. The historical and model-generated evaluations can therefore give different rankings.

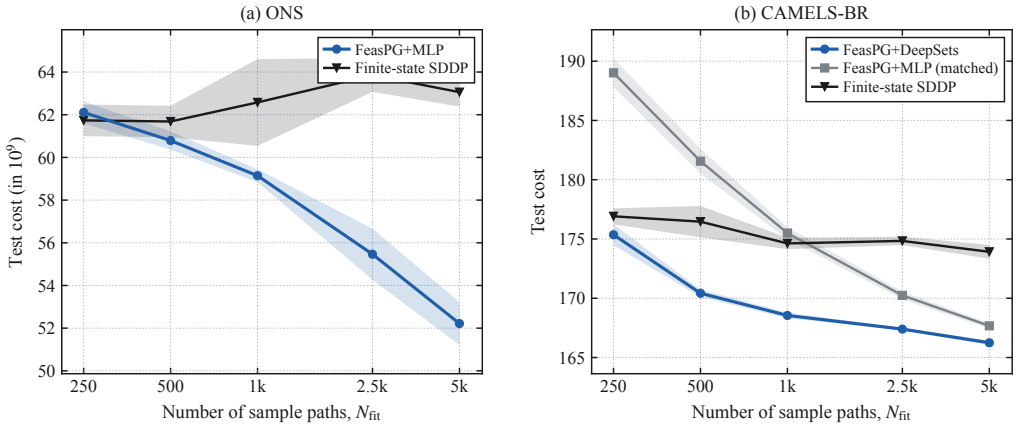

Figure 5: Objective on model-generated test paths versus training sample paths, mean ± sample SD over five seeds for ONS and CAMELS-BR. Each panel uses its case-specific raw objective scale. Blue is case-specific FeasPG, gray its matched MLP where distinct, and black finite-state empirical SDDP. SDDP uses the same five budgets, seeds, and test paths. The displayed budgets are 250, 500, 1,000, 2,500, and 5,000, with the 5,000-path points added as a subsequent extension of the original nested training pool. Table 1 reports the 2,500-path comparison, and Appendix K records execution provenance.

Together, the cases illustrate both the value of permutation-equivariant policy structure and the sensitivity of method rankings to the evaluation distribution. Appendix K gives the full paired contrasts, held-out historical results, protocol timing, and numerical audits.

F.5 SAMPLE-PATH CERTIFICATE ASSESSMENT

To assess certification separately from policy improvement, we hold a policy fixed and compare upper confidence bounds on its optimality gap $J(\widehat{\pi}) - V^\star$. The test model has three heterogeneous reservoirs over 18 stages, with continuous releases serving shared demand and finite storage capacities. A common binary random variable determines inflows and load and is observed after each decision, independently across stages. This structure permits the Paired + Bernstein instance of Algorithm 2: calibration estimates one probability per stage, and empirical Bernstein controls evaluation error.

We train a projected affine policy on 2,048 sample paths and take $\widehat{\pi}$ to be the prespecified update-400 policy. We reserve 1,024 validation paths but do not use them to select this checkpoint. The Zero

and additive piecewise-linear (PWL) continuation approximations use only training data. Separate reference computations give $J(\widehat{\pi}) \simeq 245.5919$ and $J(\widehat{\pi}) - V^\star \in [8.104, 10.730]$. True probabilities are used only for these reference computations, not for constructing the certificate. Appendix M specifies the model, exact continuous inner solver, and prespecified protocol.

The comparison changes how Algorithm 2 controls calibration and evaluation errors while leaving the policy and continuation fixed. Paired calibration uses the correction κ to retain cancellations between the policy and inner objectives. For evaluation, two deterministic dynamic programs bound the empirical penalized objective over all feasible decisions and binary uncertainty paths, yielding a fixed range without enumerating full paths. Empirical Bernstein then uses the variance of the complete corrected statistic, including κ.

With $\delta = \alpha = 0.025$, each reported upper confidence bound exceeds the verified optimality-gap upper endpoint in all 200 repetitions for each configuration. At $N_{\mathrm{cal}} = 32{,}768$ and $N_{\mathrm{ev}} = 8{,}192$, replacing Confidence band + Hoeffding by Paired + Bernstein reduces the mean PWL bound from 39.5354 to 13.2535. The calibration correction falls from 15.8324 to 1.9326 and the evaluation correction from 12.9709 to 0.5888, with the same policy, continuation, sample budgets, and exact inner problems. The mean reported certificate equals 5.40% of policy cost, compared with a verified relative optimality gap of 3.30%–4.37%. Each reported certificate is an upper confidence bound on the optimality gap, not an estimate of that gap. These results illustrate the gains from exploiting shared calibration errors and variance-sensitive evaluation. Table 14 separates the correction effects, and comparisons with other fixed checkpoints are in the experimental supplement. Appendix L gives an additional Confidence band + Hoeffding check with an exactly known optimum.

G Policy Classes and Implementation

This appendix specifies the common feasibility layer and the policy architectures used in the experiments.

Common projected-policy interface. Every learned method uses the same state-dependent feasibility layer: given the proposal $u_t = \mu_{\theta,t}(s_t, z_t)$, the deployed decision is

$$x_t = \bar{P}_t(u_t, s_t; z_t) = \arg\min_x \tfrac{1}{2}\|x - u_t\|_2^2 \quad \text{s.t.} \quad A_t x \le b_t(z_t) + B_t s_t.$$

In the nonlinear controlled benchmark, the set combines availability, release, ramping, and one weighted shared-release constraint. In the hydrology case-study models it combines nonnegativity, release capacity, available water, and, where applicable, a shared release limit. For box constraints with a single weighted shared limit, the implementation uses clipping and a fixed-iteration scalar bisection. This numerical routine approximates the exact map $\bar{P}_t$ used in the theory. The feasibility audits report its realized constraint residuals rather than treating finite-precision output as an exact projection solution. PPO and SAC (Schulman et al., 2017; Haarnoja et al., 2018) and Projected SHAC (Xu et al., 2022) are evaluated through this same map, which holds feasibility enforcement fixed across comparisons. The appendix-only cost–latency suite additionally includes TD3 (Fujimoto et al., 2018).

Features and proposal initialization. Continuous physical inputs are divided by their specified storage, release, or routing-queue capacities. The controlled benchmark appends four periodic time features. Hydrology case-study policies append the training-only hydrology features defined for each case. The proposal parameterizations are suite-specific. In the controlled benchmark, the flat MLP and the Projected SHAC policy network output $a_i = \tanh f_{\theta,i} \in [-1, 1]$, which is mapped around the stagewise reference release by

$$u_{t,i} = \mu_{\theta,t,i}(s_t, z_t) = \begin{cases} r^{\mathrm{ref}}_{t,i}(a_i + 1), & a_i < 0, \\ r^{\mathrm{ref}}_{t,i} + a_i(c_i - r^{\mathrm{ref}}_{t,i}), & a_i \ge 0, \end{cases}$$

where c_i is release capacity. Its zero output layer therefore initializes at $r^{\mathrm{ref}}_{t,i}$. The controlled Graph instead uses $r^{\mathrm{ref}}_{t,i} + \kappa_i(s_{t,i} - \bar{s}_{t,i})$ plus a zero-initialized residual scaled by $0.20\, R^{-1} \sum_i c_i$. Here R is the number of reservoirs, $\bar{s}_{t,i}$ is the reference storage, and κ_i is the storage-feedback gain. In the hydrology case-study suite, FeasPG+MLP and FeasPG+DeepSets use

$$u_{t,i} = \mu_{\theta,t,i}(s_t, z_t) = r^{\mathrm{target}}_{p(t),i} + 0.20\big(s_{t,i} - \tfrac{1}{2}(\underline{g}_{p(t),i} + \overline{g}_{p(t),i})\big) + 0.25 c_i \tanh f_{\theta,i},$$

with a zero-initialized final output. Here $p(t)$ indexes the seasonal period, r^{target} is the target release, and $\underline{g}, \overline{g}$ are the lower and upper storage guides. PPO and SAC use the common feasibility interface but not this initialization. Objective scaling is used only inside some optimizers, and all reported objectives are recomputed in the unscaled common simulator.

Proposal classes. Let o_t denote the normalized flat observation. The stagewise affine and all-pairs quadratic proposals are

$$\mu^{\mathrm{aff}}_{\theta,t}(s_t, z_t) = W^{\mathrm{aff}}_t[1; o_t], \qquad \mu^{\mathrm{poly}}_{\theta,t}(s_t, z_t) = W^{\mathrm{poly}}_t[1; o_t; \mathrm{vech}(o_t o_t^\top)].$$

Here vech lists the distinct entries of the symmetric matrix $o_t o_t^\top$. The flat MLP applies two hidden layers to o_t and outputs one proposal coordinate per reservoir. In the controlled benchmark, the Graph policy encodes reservoir i's local feature vector v_i as $h_i^0 = \phi(v_i)$ and updates it for three layers with shared weights using its own, normalized upstream, normalized downstream, and global messages. For $\ell = 0, 1, 2$,

$$h_i^{\ell+1} = \tanh\left(W_0^\ell h_i^\ell + W_\uparrow^\ell \sum_j \bar{A}^\uparrow_{ij} h_j^\ell + W_\downarrow^\ell \sum_j \bar{A}^\downarrow_{ij} h_j^\ell + W_{\mathrm{g}}^\ell \bar{h}^\ell + b^\ell \right),$$

where $\bar{h}^\ell = R^{-1} \sum_i h_i^\ell$. Thus the controlled Graph can coordinate allocation under the shared-release constraint. For the non-network CAMELS-BR portfolio, DeepSets instead uses $h_i = \phi(v_i)$, $\bar{h} = R^{-1} \sum_i h_i$, and $\delta_i = \psi([h_i, \bar{h}])$, which is permutation equivariant in basin labels. ONS uses the MLP, while CAMELS-BR uses DeepSets with a parameter-matched flat MLP comparator.

Matching and optimization. In the controlled data-budget suite the width of the two-layer flat MLP is chosen to minimize its parameter-count difference from the width-64 graph policy: 54,619 versus 54,721 deployed parameters. Projected SHAC uses the same 54,619-parameter deployed policy network—the *actor* in actor–critic terminology—and model access, but backpropagates in eight-stage windows, detaches the simulated state at each window boundary, and adds a learned cost-to-go model, or *critic*, there. Its training state contains 159,369 parameters across the actor and the online and target critic copies. Only the 54,619-parameter actor is deployed, and the target critic is updated by exponential averaging rather than directly optimized. PPO and SAC use separate policy sizes. The separate 11-method comparison uses an 82,444-parameter MLP. Direct policies and Projected SHAC are trained with Adam, gradient clipping at 10, and weight decay 10^{-6}. Learning rates and checkpoints are chosen by validation cost.

Scope. In the controlled suite, parameter matching covers the deployed Graph and MLP proposals and the Projected-SHAC actor. The CAMELS-BR DeepSets policy has a parameter-matched MLP comparator, but its Projected-SHAC actor is not parameter matched. Critic state and training compute are accounted for separately. PPO, SAC, and the 11-method suite use their prespecified architectures. The stationarity theorem concerns fresh-minibatch SGD with a randomly selected iterate, whereas the reported FeasPG implementations use Adam and validation-selected checkpoints.

H Controlled Nonlinear Benchmark

This appendix documents the controlled benchmark, its linearized SDDP comparator, and the supplementary implementation audits.

SDDP extension of the matched-budget comparison. The linearized SDDP curve in Figure 1 adds 25 runs: five training seeds at each of 100, 250, 500, 1,000, and 2,500 training paths. It reuses the exact nested training pool and the 500 validation and 4,096 test paths of that figure, verified by data checksums. Each training prefix is compressed to two joint inflow–demand outcomes per stage. Training performs 50 SDDP iterations, and validation selects among the initial policy, evaluated before any backward cut is added, and iterations 10, 20, 30, 40, and 50. The selected policy is evaluated in the original nonlinear simulator with the shared feasibility projection. This supplementary extension was conducted after the primary learned-policy comparison. The linearization supplies a planning comparator, not a lower bound on the nonlinear optimum. The latency panel below retains the original five learned methods.

Approximation diagnostic and interpretation. The planning model linearizes generation and evaporation at each stage's prescribed storage/release pair, retaining that expansion as realized states and decisions change. The true simulator instead evaluates the nonlinear head and two narrow efficiency bands whose centers move with storage. The planning model also permits discretionary spill, whereas the simulator spills only overflow. The contribution of this additional difference has not been quantified. Increasing N_{fit} changes the two fitted joint outcomes per stage but leaves this physical approximation and the 50-iteration training budget unchanged.

A two-seed replay at $N_{\text{fit}} = 2500$ found that the affine planning model predicted roughly twice the hydro generation obtained in the nonlinear simulator. This supports model mismatch as one explanation for the observed cost gap, but does not disentangle physical approximation, scenario compression, and finite SDDP training. Across all 25 SDDP runs, 13 select the initial pre-cut checkpoint, including all five runs at $N_{\text{fit}} = 100$.

Supplementary baseline comparison. The decision-switching benchmark has $R = 12$ reservoirs, $T = 48$ stages, four latent uncertainty factors, and a two-stage routing delay. The policy acts before the current inflow and demand are realized. The fixed full-comparison split contains 1,000 training, 500 validation, and 4,096 common test sample paths. Five optimization seeds share the same sample-path sets. The candidate and checkpoint are selected only by validation cost, and the common test set is read once after selection. Each candidate sees 1,000 distinct training paths.

Table 2: Complete controlled-benchmark comparison at $N_{\text{fit}} = 1000$. Cost is mean $\pm$ sample SD over five training seeds, with one run for each fixed method. Timing records each implementation's CPU execution. FeasPG and baseline timings were collected separately and do not support a controlled speed comparison. Lower cost is better.

Method	Test cost	Policy params	Tuning (s)	Batch-1 (ms)
FeasPG+Graph	54.512 ± 0.573k	54,721	1131.51	0.978
FeasPG+MLP	59.423 ± 0.153k	82,444	607.63	0.854
FeasPG+polynomial	61.122 ± 0.150k	705,600	629.06	0.806
FeasPG+affine	63.303 ± 0.175k	28,224	499.22	0.796
SAC	79.078 ± 1.359k	85,528	1042.16	0.984
TD3	81.049 ± 3.602k	82,444	683.83	0.932
PPO	105.871 ± 3.875k	82,456	84.20	0.944
Two-scenario linearized SDDP	131.588 ± 1.614k	$0^{\dagger}$	613.96	3.066
Constant cost-function approximation (CFA)	134.212 ± 0.013k	1	452.83	1.356
Deterministic lookahead	136.059k	0	0	1.376
Myopic	144.091k	0	0	0.758

Here "Policy params" counts trainable scalars in the deployed proposal or policy, excluding RL critics and other training-only state. The dagger marks SDDP, whose selected policies contain 960, 0, 0, 1,920, and 1,440 cuts across the five seeds (mean 864).

The graph policy's mean cost is 8.27% below the unmatched flat MLP and 31.07% below SAC. Its five paired Graph-minus-MLP differences on common test paths are -5.089, -4.071, -4.814, -4.669, -5.916k. All 47 reported runs have zero recorded adjusted-decision-bound, shared-capacity, and storage violations. Myopic proposes the stage's reference release at every reservoir and relies on the projection for feasibility. Deterministic lookahead and CFA use a planned-state 12-stage LP approximation, while the reported two-scenario linearized SDDP compresses the training data to two joint outcomes per stage and linearizes the nonlinear system. Two of five SDDP seeds select this initial pre-cut checkpoint. This full-comparison suite contains projected PPO/SAC/TD3 but not Projected SHAC.

In this suite, N_{fit} matches distinct training sample paths while optimizer work, experience-replay work, validation opportunities, and compute differ. Policy parameters count deployed policies only, and the simplified LP references provide no optimality bound for the nonlinear control problem.

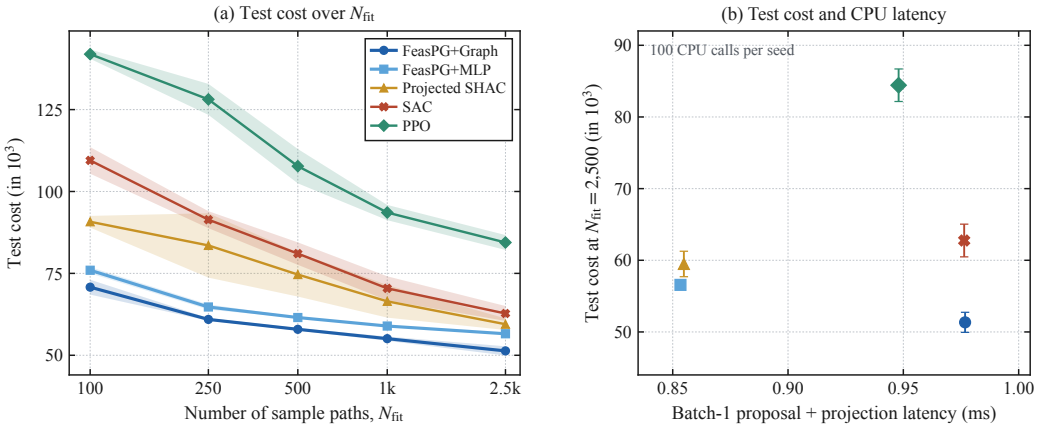

Figure 6: Matched-suite sample-path budget and deployed-decision timing for the learned methods. The left panel shows the learned-method curves from Figure 1, and the right panel uses the same $N_{\text{fit}} = 2500$ selected checkpoints. Cost coordinates are means over five training seeds with one-sample-SD vertical bars. Latency coordinates are the median across five within-seed medians, each measured from 100 CPU calls at the horizon midpoint to policy proposal plus the shared projection. The timed state comes from a replay of each policy, and the RL adapter includes its Torch–NumPy interface overhead. The panel reports deployed proposal-plus- projection latency separately from training resources. Timings were collected separately for FeasPG and baselines, so differences are descriptive.

Supplementary training audit. For each direct-policy candidate, 500 Adam updates with batch size $N_{\text{bat}} = 2$ consume each of the 1,000 training paths once and validation is checked every ten updates. PPO, SAC, and TD3 complete 1,000 training episodes per candidate, but PPO performs ten epochs per rollout and the off-policy methods reuse an experience-replay buffer. CFA uses symmetric perturbations and therefore 2,000 path evaluations. Learning-rate grids, all selected checkpoints, per-seed costs, paired pathwise intervals, training curves, cost components, throughput, and resource counters are included in the artifact. FeasPG evaluates three regularization strengths at each learning rate: six candidates per seed in this supplementary suite, and nine in the primary budget suite. Initialization adds 32 per-path cost and regularizer-gradient evaluations on the same training pool.

Matched-suite RL tuning. For every budget and training seed, validation selects both checkpoint and actor learning rate. Projected SHAC uses $\{10^{-4}, 3\times10^{-4}, 6\times10^{-4}\}$, while PPO and SAC use $\{10^{-4}, 3\times10^{-4}, 10^{-3}\}$. All other prespecified settings are fixed: PPO and SAC have two width-256 hidden layers and $\gamma = 1$, while SHAC uses the matched actor described in Appendix G. PPO uses 10-episode rollouts, 10 epochs, and batch size 256. SAC uses a 100,000-step buffer, 1,000 learning-start steps, batch size 256, and one gradient update per step. Across the 25 controlled budget–seed cells, the largest rate is selected for 24 SHAC, 14 PPO, and 25 SAC cells. At $N_{\text{fit}} = 2500$, the counts are 5/5, 3/5, and 5/5. In ONS and CAMELS-BR, endpoint selection varies by system, with frequent lower-bound selections on CAMELS-BR and upper-bound selections on ONS. The complete hydrology-case selections are reported with their artifacts.

Motivated by these upper-endpoint counts, a one-sided supplementary check, conducted after the primary comparison at $N_{\text{fit}} = 2500$, adds 10^{-3} for SHAC and 3×10^{-3} for PPO/SAC.

Table 3: Supplementary upper-rate sensitivity on the controlled benchmark. Upper-selected counts cover all 25 budget–seed cells and the five $N_{\text{fit}} = 2500$ cells. Expanded selection uses the same validation set. Costs are test mean ± training-seed SD in thousands.

Method	Prespecified grid	Upper selected all / 2.5k	Added rate	Added selected at 2.5k	Primary / enlarged-grid test cost
Projected SHAC	$\{1, 3, 6\} \times 10^{-4}$	24/25 / 5/5	10^{-3}	2/5	$59.48 \pm 1.78 / 58.91 \pm 1.31$
PPO	$\{1, 3, 10\} \times 10^{-4}$	14/25 / 3/5	3×10^{-3}	0/5	$84.43 \pm 2.27 / 84.43 \pm 2.27$
SAC	$\{1, 3, 10\} \times 10^{-4}$	25/25 / 5/5	3×10^{-3}	3/5	$62.76 \pm 2.28 / 61.67 \pm 1.41$

The enlarged-grid baseline test means are 58.91k for Projected SHAC, 61.67k for SAC, and 84.43k for PPO. For comparison, the FeasPG means are 48.74k for Graph and 55.84k for the matched MLP.

Scope. This baseline sensitivity analysis selects learning rate and checkpoint. The primary comparison uses the prespecified grids. The one-sided check targets the controlled $N_{\text{fit}} = 2500$ endpoint, and the hydrology-case grids remain fixed. Results therefore compare implementations selected over the declared candidates, not exhaustive tuning of each algorithm.

I Proposal-Distance Regularization: Mechanism and Protocol

Constructed startup stress test. The experiment in Appendix F.2.2 retains the controlled nonlinear model, cost coefficients, original MLP initialization, and all physical constraints. Initial storage is set to 1% of capacity, and the inflow mean, factor loadings, and idiosyncratic noise are scaled jointly by $q \in \{0.01, 0.02, 0.05, 0.10\}$. Demand and price distributions are unchanged. The nominal control retains the original 64% initial storage and unscaled inflow. All five settings and all five training seeds are reported. The extreme scarcity cases are mechanism tests rather than claims about the frequency of historical droughts.

The training-only screening used 32 paths. Before observing training outcomes, the $q = 0.02$ case was designated as primary because it is the less severe of the two cases for which an all-support plateau bound succeeds. A recursive upper bound on storage under all-water release, including the two-stage routing queue, gives minimum proposal–storage, ramp–storage, and shared-capacity margins of 3.871, 13.982, and 374.786, respectively. The positive margins prove inductively that every initial decision equals available storage and is locally independent of policy parameters. The cost gradient therefore vanishes even with complete moving recourse. The bound also succeeds at $q = 0.01$ but is inconclusive at $q = 0.05$ and $q = 0.10$. Observed results in the latter cases are empirical.

Each candidate uses 1,000 paths from the original nested training pool, 500 Adam updates with $N_{\text{bat}} = 2$ previously unused paths per update, learning rate 6×10^{-4}, weight decay 10^{-6}, and gradient clipping at 10. Prescribed weights are 0, 1, and $(k + 1)^{-1/2}$, where k starts at zero. On a new 500-path validation set, we compare checkpoints saved at ten-update intervals, including initialization, and select the one with the lowest cost. We fix this selection before reporting costs on a new 4,096-path test set. The validation and test seeds are 35202 and 35303. The five training seeds are 0–4.

Table 4: Startup stress-test cost reduction relative to the trained $\lambda = 0$ control (%). Values compare five-seed mean test costs. All positive-weight configurations improve in all five seeds.

Setting	$\lambda = 1$	$\lambda_k = (k + 1)^{-1/2}$
1% initial storage, 1% inflow	5.44	5.44
1% initial storage, 2% inflow	8.57	8.57
1% initial storage, 5% inflow	15.99	16.16
1% initial storage, 10% inflow	17.30	17.29
Original nominal case	0.50	0.91

What improves. For $q = 0.02$, all unregularized training gradients are exactly zero, and decisions and sample-path costs stay unchanged. Adam weight decay does not alter the zero output layer, so it

does not resolve this plateau. Positive weights first reduce proposal distance while leaving decisions unchanged. An independent replay of seed 0 with $\lambda = 1$ finds the first decision change and first nonzero cost gradient after update 26. In that run, validation thermal cost increases from 178,472 to 183,343, while the low-storage cost component falls from 92,586 to 65,612 and terminal cost falls from 24,646 to 21,412. The primary gain is water conservation. A post-training no-release reference has validation cost 270,507, compared with 270,366 for the learned policy. At 5% inflow, these values are 239,655 and 236,383. These descriptive references were not used for selection and do not certify optimality.

Derivative and optimizer checks. An independent implementation of the regularizer-gradient formula reproduces the first 30 updates exactly. At update 30, a double-precision directional finite-difference check uses 32 training paths, three fixed random directions, and the original seven step sizes. At $h = 10^{-5}$, the mean normalized error is 2.12×10^{-7} for the complete derivative and 1.03 when moving recourse is omitted from the entire policy replay, with identical forward values. Appendix J isolates the omission within the regularizer. All step sizes and directions are retained in the artifact. In the scalar example $\mathcal{X} = [0, 1]$, $c(x) = (x - 1)^2$, $u_0 = -1$, Adam with learning rate 0.1 crosses into positive decisions after 12 updates for $\lambda = 1$ and after 14 for the decreasing weight, while the unregularized decision remains zero. In contrast, plain SGD at that learning rate satisfies $u_{k+1} = (1 - 0.1\lambda_k)u_k < 0$ for the tested positive schedules and need not cross at any finite iteration. The mechanism therefore does not imply a universal finite-step escape guarantee.

Weight selection. The budget, scaling, hydrology, derivative-ablation, and supplementary comparisons use the following training-only scale rule. On the first 32 training sample paths ξ^j, let $g_j = \|\widehat{g}_{\mathrm{PW}}(\theta_0; \xi^j)\|$ and $h_j = \|\nabla_\theta d_{\mathrm{pd}}(\theta_0, \xi^j)\|$. Write $C_j = C(\theta_0, \xi^j)$. We set

$$\lambda_{\mathrm{ref}} = \frac{\mathrm{median}_j\, g_j}{\mathrm{median}_j\, h_j + 10^{-12}}.$$

If either median is zero, we instead use the prescribed positive fallback

$$\lambda_{\mathrm{ref}} = \frac{\max\{32^{-1} \sum_j |C_j|, 10^{-12}\}}{\max\{\frac{1}{2} T \sum_i c_i^2, 10^{-12}\}},$$

where c_i denotes release capacity. The fallback is activated in 0 of 220 benchmark initializations. Cost and its gradients use the training-only objective normalization in the hydrology optimizers, while reported costs are recomputed in the raw simulator. The gradient includes each projected decision's dependence on the proposal and on the state-dependent constraint bounds, except in the explicitly identified derivative ablations.

For each training budget and seed, the declared positive grid is $\lambda_k = a\lambda_{\mathrm{ref}}/\sqrt{k+1}$ with $a \in \{0.1, 1, 10\}$ and $k = 0, \ldots, K-1$. We use validation cost to jointly select the strength, learning rate, and checkpoint, including initialization. Exact ties favor the smaller strength and then the earlier learning-rate candidate. Scale estimation uses only that budget's training prefix and never borrows paths from a larger budget. The test and historical sets do not select the weight. The grid is fixed before the benchmark test reports and is not expanded in response to them.

Nominal controls. At $N_{\mathrm{fit}} = 1000$, $\lambda = 0$ controls use the same device, initialization, paths, optimizer, and learning-rate grid as the positive runs. The positive grid has three times as many candidates, so this comparison assesses the complete selection protocol and is not matched for tuning compute.

Table 5: Nominal $N_{\mathrm{fit}} = 1000$ test controls. Costs are mean $\pm$ training-seed SD. Synthetic costs are in thousands, ONS in billions, and CAMELS in native units. Positive reduction favors regularization.

Setting	$\lambda = 0$	Selected positive	Reduction (%)
Synthetic Graph	55.072 ± 0.659	52.737 ± 0.766	4.24
Synthetic MLP	58.927 ± 0.104	58.326 ± 0.191	1.02
ONS MLP	59.534 ± 0.718	59.144 ± 0.303	0.65
CAMELS MLP	175.392 ± 0.501	175.499 ± 0.512	-0.06
CAMELS DeepSets	168.440 ± 0.162	168.550 ± 0.233	-0.07

Table 6: Validation costs for the same selected controls, with scales and seed summary matching Table 5.

Setting	$\lambda = 0$	Selected positive
Synthetic Graph	55.017 ± 0.633	52.693 ± 0.835
Synthetic MLP	58.897 ± 0.109	58.303 ± 0.188
ONS MLP	58.959 ± 0.670	58.534 ± 0.501
CAMELS MLP	170.979 ± 0.504	171.079 ± 0.509
CAMELS DeepSets	163.863 ± 0.159	163.994 ± 0.218

The nominal controls show different effects across settings: synthetic Graph and MLP improve, ONS improves more modestly, and CAMELS has almost unchanged, slightly higher test means. These results support the corrective mechanism without establishing a universal improvement in held-out cost.

The effect also depends on the training budget. Relative to the unregularized controlled runs, Graph's sample mean is 0.38% higher at 250 paths, whereas it is 5.07% lower at 2,500 paths. The artifact records all 43 budget, scaling, and full-comparison contrasts against unregularized FeasPG. Comparisons to the unregularized hydrology training runs also differ in training device.

Auxiliary metrics and computational cost. Table 7 reports the proposal distance and fraction of decision coordinates changed by projection (tolerance 10^{-6}) on test paths. Gradient norms are evaluated on the same first 32 training paths at the selected checkpoint, before clipping, for the original training objective scale. The total gradient uses the checkpoint's decreasing weight. Consequently, gradient magnitudes across differently scaled systems are not directly comparable. All selected policies have zero recorded maximum decision and storage feasibility residuals.

Table 7: Nominal control diagnostics, five-seed means. $\overline{d}_{\mathrm{pd}}$ is mean test proposal distance, and G_C and G_λ are training-gradient norms. Projected is the percentage of decision coordinates changed by projection. Time sums all candidate training runs and scale estimation per seed, in seconds.

Setting, weight	$\overline{d}_{\mathrm{pd}}$	G_C	G_λ	Projected (%)	Time (s)
Synthetic Graph, 0	1.40×10^4	2.00×10^4	2.00×10^4	80.2	507
Synthetic Graph, positive	3.70×10^2	3.89×10^4	4.12×10^4	47.0	1515
Synthetic MLP, 0	1.42×10^3	5.01×10^4	5.01×10^4	39.9	249
Synthetic MLP, positive	3.82×10^2	5.77×10^4	5.89×10^4	37.3	805
ONS MLP, 0	1.15×10^{15}	3.84×10^1	3.84×10^1	100.0	456
ONS MLP, positive	9.38×10^{14}	4.22×10^1	4.32×10^1	99.2	1339
CAMELS MLP, 0	9.31×10^4	1.19×10^1	1.19×10^1	50.4	61
CAMELS MLP, positive	8.75×10^4	1.15×10^1	1.15×10^1	50.2	181
CAMELS DeepSets, 0	1.42×10^5	1.10×10^1	1.10×10^1	60.7	105
CAMELS DeepSets, positive	1.35×10^5	1.23×10^1	1.23×10^1	59.7	307

The primary and derivative-ablation grids have nine candidates per seed, while the supplementary full comparison has six. Candidates share the training pool and initialization scale estimate, so tuning does not increase distinct training data. The artifact records selected weights, checkpoints, validation/test costs, gradient norms, feasibility residuals, and runtimes. Tuning time includes scale estimation, optimizer updates, and periodic validation but excludes initialization validation and final held-out evaluation. FeasPG uses one CPU thread per worker, while the hydrology-case baselines use CUDA, so these timings do not provide a controlled speed comparison.

Hydrology-case derivative checks. We check an independent expression for the regularizer after 30 updates, using 32 training paths, three fixed random directions, and the original seven finite-difference step sizes. At $h = 10^{-5}$, the maximum normalized complete-gradient error is 4.75×10^{-10} for ONS MLP, 4.60×10^{-10} for CAMELS MLP, and 1.16×10^{-9} for CAMELS DeepSets. Removing moving recourse increases the ONS mean error to 0.358. The same omission leaves the two CAMELS derivatives unchanged at these particular checkpoints because that constraint dependence is inactive there, but this does not establish that the correction can generally be discarded.

J Additional Mechanism and Scaling Results

This appendix reports derivative diagnostics and supporting controlled-benchmark robustness, scaling, and numerical equivariance results.

Diagnostic metrics. The horizontal axis h in the finite-difference panels is the parameter perturbation size, not the training learning rate. Fix the parameters θ, a direction v, and the exogenous sample paths. Let $F(\theta)$ denote the average cost in panel (a), or the average regularized loss in panel (b), on these paths. We perturb the parameters in both directions and replay the sample paths under each perturbed policy, including proposals, projections, and state transitions:

$$d_{\mathrm{FD}}(h) = \frac{F(\theta + hv) - F(\theta - hv)}{2h}, \qquad d_{\mathrm{AD}} = \nabla_\theta F(\theta)^\top v.$$

Both perturbed evaluations use the same exogenous sample paths. The complete AD gradient should agree with this central finite difference at differentiability points as h decreases, until floating-point errors become significant. For the ablations, we replace the AD gradient by the derivative with the specified terms omitted, while retaining the same forward loss and FD reference.

Figure 3 compares automatic-differentiation (AD) and central finite-difference (FD) directional derivatives using $|d_{\mathrm{AD}} - d_{\mathrm{FD}}| / \max\{1, |d_{\mathrm{AD}}|, |d_{\mathrm{FD}}|\}$. Panel (a) uses 32 fixed training paths for the five-seed cost-gradient check, with reported errors at the prespecified step $h = 10^{-5}$. Panel (b) uses the seed-0, update-30 checkpoint from the startup experiment in Appendix I, with $\lambda = 1$, the first 32 training paths, three fixed normalized Gaussian parameter directions, and double precision. Both checks retain all seven step sizes $\{10^{-2}, 3\times10^{-3}, 10^{-3}, 3\times10^{-4}, 10^{-4}, 3\times10^{-5}, 10^{-5}\}$. The complete regularized derivatives reproduce the existing diagnostic. To isolate the derivative through the projected decision in the regularizer, we evaluate

$$C + \frac{\lambda}{2} \sum_t \|u_t - \mathrm{stopgrad}(x_t)\|^2$$

on the same full policy replay. Here $\mathrm{stopgrad}$ preserves its input value but assigns it zero derivative. States, proposals, and the cost retain their original computation graphs. The forward objective is identical to C_λ, and the detached gradient minus the complete gradient equals $\lambda \sum_t X_t^\top (u_t - x_t)$ by (B.7). At $h = 10^{-5}$, the mean normalized directional errors are 2.12×10^{-7} and 0.181, and their maxima over the three directions are 5.66×10^{-7} and 0.251, respectively. The relative full-vector difference is 0.0428.

The horizon panel (c) reports $\|\widehat{g}_{\mathrm{abl}} - \widehat{g}_{\mathrm{PW}}\| / \max\{\|\widehat{g}_{\mathrm{PW}}\|, 10^{-12}\}$, where "abl" omits the derivative through either the proposal's state input or the moving constraint bounds, and "PW" is the complete pathwise gradient.

Detaching the decision can reverse the derivative's sign. Consider two stages with $u_1 = \theta$, $x_1 = P_{[0,2]}(u_1)$, $s_2 = x_1$, and $u_2 = 1 + s_2/2$. For $0 < \theta < 2$, project the second proposal onto $[0, s_2]$, giving $x_2 = s_2 = \theta$ and

$$d_{\mathrm{pd}}(\theta) = \tfrac{1}{2}(1 - \theta/2)^2, \qquad d'_{\mathrm{pd}}(\theta) = -\tfrac{1}{2}(1 - \theta/2).$$

Detaching x_2 only in this residual instead returns $+\frac{1}{2}(1-\theta/2)$. At $\theta = 1$, the complete and detached derivatives are -0.25 and $+0.25$, respectively. A central difference at $h = 10^{-6}$ agrees with the complete derivative to within 10^{-9}. This example isolates the effect of the state-dependent bound in the regularization term.

Finite-budget comparison with gradient terms omitted. Figure 3 tests derivative correctness for C before optimization. We also train the same projected MLP with C_λ while omitting one derivative contribution throughout the policy replay. This separate mechanism experiment uses the original 0.10 ramp fraction, compared with 0.50 in the primary budget benchmark, and its own fixed data split and proposal initialization. The forward computation, policy class, projection, sample-path budget, and validation rule are held fixed. For each seed, the full-gradient variant selects the regularization strength. The two derivative-ablation variants then use that same strength and select their learning rate and checkpoint on validation data. At $N_{\mathrm{fit}} = 1000$, test costs are 78.423 ± 0.061k

for the complete gradient, 78.734 ± 0.575k without proposal–state feedback, and 80.872 ± 0.748k without moving recourse. The omission applies to both terms of C_λ.

Scope. Finite differences identify the derivative of the stated objective, while finite-budget Adam outcomes additionally reflect nonconvex optimization, implicit bias, and checkpoint selection.

Complete sample-path budget grid. Table 8 gives the learned-method values in Figure 1, together with the fixed myopic reference. All learned methods use the same sets of distinct training sample paths and common test paths. Resource accounting is in Appendix H.

Table 8: Controlled cost on model-generated test paths in thousands (lower is better), mean $\pm$ sample SD over five training seeds. Myopic is fixed.

N_{fit}	FeasPG+Graph	FeasPG+MLP (matched)	Proj. SHAC	SAC	PPO	Myopic
100	70.76 ± 1.96	74.70 ± 0.94	90.76 ± 1.79	109.48 ± 4.06	141.93 ± 1.42	144.25
250	61.18 ± 0.39	64.59 ± 0.44	83.53 ± 9.83	91.40 ± 2.59	128.12 ± 4.73	144.25
500	57.25 ± 0.67	60.48 ± 0.18	74.65 ± 6.72	81.04 ± 3.43	107.72 ± 5.24	144.25
1000	52.74 ± 0.77	58.33 ± 0.19	66.47 ± 4.97	70.43 ± 3.68	93.57 ± 2.33	144.25
2500	48.74 ± 0.73	55.84 ± 0.17	59.48 ± 1.78	62.76 ± 2.28	84.43 ± 2.27	144.25

Empirical scaling. We trained FeasPG+Graph and its parameter-matched MLP separately in seven reservoir–horizon configurations. Table 9 reports results on fresh sample paths from the fitted data model. Graph has a lower mean cost in 7 of seven settings, with relative reductions ranging from 3.65% to 16.23%. The deployed parameter counts remain approximately constant ($54{,}721$ for Graph and $54{,}615$–$54{,}907$ for MLP). Median batch-1 proposal-plus-projection latency ranges from 0.964 to 1.073 ms for Graph and 0.851 to 0.909 ms for MLP in the serial CPU audit.

Table 9: Seven-configuration scaling audit. Costs are in thousands, mean $\pm$ sample SD over five training seeds. Myopic is fixed. "Reduction" is the reduction in Graph mean cost relative to matched MLP.

Configuration	FeasPG+Graph	FeasPG+MLP (matched)	Myopic	Reduction (%)
$R = 6,\ T = 48$	20.28 ± 0.80	24.21 ± 0.07	56.88	16.23
$R = 12,\ T = 48$	52.87 ± 0.70	58.29 ± 0.16	144.04	9.29
$R = 24,\ T = 48$	127.02 ± 2.86	138.07 ± 2.20	316.54	8.01
$R = 48,\ T = 48$	282.24 ± 4.96	298.49 ± 1.58	722.88	5.45
$R = 12,\ T = 12$	11.30 ± 0.22	11.73 ± 0.08	27.28	3.65
$R = 12,\ T = 24$	22.96 ± 0.72	25.09 ± 0.19	64.66	8.52
$R = 12,\ T = 96$	114.35 ± 0.58	123.79 ± 0.34	309.91	7.62

Numerical equivariance audit. Across five trained policies from the $R = 12, T = 48$ scaling setting and five random node permutations (64 common paths per audit), Graph's mean decision equivariance error is 2.16×10^{-6}, with a largest decision error of 5.99×10^{-4} and largest path-cost invariance error of 1.95×10^{-2}. The matched flat MLP is not permutation equivariant: its corresponding maxima are 48.04 and 9.43×10^{4}. These are numerical audits of the implemented architecture.

K Hydrology Case Studies: Data and Additional Results

This appendix documents data provenance and sample splits, reports model-generated and held-out results, and records the numerical feasibility audit.

K.1 Data provenance, models, and splits

The two reported settings use public measurements, case-specific objectives, and deliberately different control models. ONS open data are CC BY 4.0, and CAMELS-BR v1.2 is the frozen Zenodo release under CC BY 4.0 (Chagas et al., 2020). Download metadata, source checksums, processed-data checksums, unit conversions, missingness rules, and checksummed held-out evaluation arrays are retained in the artifact.

ONS uses official natural-energy inflow (ENA) and load records for four electrical-subsystem aggregates in an equivalent-energy hydrothermal model with piecewise-linear thermal dispatch. The data model is fitted on 2000–2014, with 2015–2018 as an unused temporal buffer. Historical evaluation uses one 48-month chronology from 2022–2025.

CAMELS-BR uses 12 non-nested catchments with standardized semi-synthetic reservoirs. The data model is fitted on water years 1981–2005. Water years 2006–2014 form an unused temporal buffer, and historical evaluation uses five overlapping 48-month windows beginning in 2016–2020. The dataset contains no verified dam network or plant physics: capacities, release limits, and the objective are experimental specifications.

Neither temporal buffer is used to fit a data model, train a policy, or select hyperparameters or checkpoints. Historical evaluation observations are opened only after checkpoint selection and are distinct from the generated policy-validation paths described below.

For each case, a periodic autoregression is fitted using only the training interval. Calendar-conditioned residual vectors are jointly resampled, and ONS also contains an annual periodic-autoregressive (PAR) component. Its equivalent-energy hydrothermal model follows the classical multistage structure (Pereira & Pinto, 1991). The master training pool has nested prefixes $N_{\mathrm{fit}} \in \{100, 250, 500, 1000, 2500\}$. Validation, model-generated test, calibration, and evaluation contain 500, 4,096, 500, and 2,048 paths, respectively, with common random numbers within a benchmark. The five training seeds are fixed. Before choosing a decision, the policy observes opening storage, previous release, calendar, routing state, and the causal hydrology history summary. The current uncertainty is realized after the decision. Historical observations cannot select a hyperparameter or checkpoint.

K.2 Results on model-generated test paths and sample-path budgets

Table 10 reports 95% conditional value at risk, $\mathrm{CVaR}_{0.95}$, which summarizes the worst 5% cost tail and is computed within each policy from 4,096 path costs and then averaged across seeds. Table 11 reports the paired relative effects used in the main comparison. Scales are case specific, with cost means and training-seed SDs in Table 1.

Table 10: Model-generated test CVaR_{95} at $N_{\mathrm{fit}} = 2500$, averaged over five training seeds. ONS is in billions, while CAMELS-BR retains its raw objective units. Bold marks the lowest value in each column.

Method	ONS	CAMELS-BR 12
Seasonal	588.790	429.903
FeasPG+MLP	168.251	344.845
FeasPG+DeepSets	–	342.481
Projected SHAC	174.118	390.450
PPO	708.426	706.508
SAC	**159.030**	828.247
Finite-state SDDP	244.192	**335.881**

The tail and mean rankings need not agree: on ONS the lowest sample mean belongs to FeasPG+MLP and the lowest CVaR to SAC, while on CAMELS-BR the corresponding methods are FeasPG+DeepSets and Finite-state SDDP.

For candidate A and reference B, define the relative effect $\delta(A, B) = 100(\bar{C}_A/\bar{C}_B - 1)$, where negative values favor A. Each 95% crossed paired-bootstrap interval uses 10,000 replicates that independently resample the five training-seed rows and 4,096 common-path columns, with the same indices for both methods. The seed-level sensitivity forms a t_4 interval from the five paired path-mean cost differences and expresses it as a percentage of the observed grand reference mean.

These are unadjusted effect intervals, not a familywise declaration of binary significance. For ONS FeasPG+MLP versus Projected SHAC, the crossed-bootstrap interval excludes zero, whereas the five-seed t_4 sensitivity interval crosses zero.

The SDDP reference uses the same convex model, seeds, nested training pools, and common test paths as the learned methods, with five empirical joint scenarios per stage and hydrology regime and 80 forward/backward iterations. Validation selects among iterations 20/40/60/80. Its standalone

Table 11: Paired effects on model-generated test paths at $N_{\text{fit}} = 2500$. "Boot." is the crossed paired-bootstrap percentile interval, and "seed t_4" is the five-seed sensitivity interval.

Contrast (A/B)	$\delta(A, B)$ (%)	Boot. 95% CI	Seed t_4 95% CI
ONS: FeasPG+MLP / Projected SHAC	-3.15	$[-5.16, -1.25]$	$[-6.32, 0.02]$
ONS: FeasPG+MLP / finite-state SDDP	-13.14	$[-14.65, -11.61]$	$[-14.92, -11.36]$
CAMELS-BR: DeepSets / matched MLP	-1.67	$[-1.86, -1.45]$	$[-1.99, -1.35]$
CAMELS-BR: DeepSets / finite-state SDDP	-4.25	$[-4.55, -3.96]$	$[-4.49, -4.02]$

training runner is not included in the present artifact, so it is a planning reference rather than a fully re-executable baseline. Matching N_{fit} does not match solver work or compute, and these results provide no optimality bound.

Figure 5 includes a subsequent $N_{\text{fit}} = 5000$ extension using the original 2,500 training paths plus 2,500 fresh paths from the same fixed data model. Evaluation arrays and baseline selection rules are unchanged. Detailed budget values, extension results, and execution records are retained in the accompanying experimental supplement.

K.3 Held-out historical evaluation and physical audit

Table 12: Held-out historical cost at $N_{\text{fit}} = 2500$, mean $\pm$ sample SD over five training seeds. ONS is in trillions, while CAMELS-BR uses native raw units. Seasonal is fixed.

Method	ONS (1)	CAMELS-BR (5)
Seasonal	**4.596**	256.257
FeasPG+MLP	5.161 ± 0.098	245.081 ± 2.944
FeasPG+DeepSets	–	$\mathbf{215.875 \pm 1.596}$
Projected SHAC	5.111 ± 0.102	320.625 ± 3.422
PPO	5.262 ± 0.156	588.704 ± 16.959
SAC	5.616 ± 0.095	595.760 ± 87.875
Finite-state SDDP	5.495 ± 0.034	290.062 ± 8.761

The lowest historical sample means belong to Seasonal on ONS and FeasPG+DeepSets on CAMELS-BR. This evaluation uses the same selected checkpoints as the model-generated test.

The held-out historical evaluation contains one ONS chronology and five overlapping CAMELS windows. These shared histories provide descriptive assessments under a changed evaluation distribution, not independent population replications or a pooled cross-case test.

Across both reported cases, the maximum recorded post-projection decision and physical-storage residuals are zero. Shortage, spill, deficit, and guide-band quantities are objective terms or soft deviations rather than hard infeasibility.

L Single-Reservoir Certificate Check

This check assesses confidence-band calibration and Hoeffding evaluation against an exactly known optimality gap. There is one reservoir, $T = 6$, $s_1 = 3$, and independent binary random variables $w_t \in \{0, 1\}$, observed after each decision, with probabilities $(7, 12, 9, 13, 8, 11)/20$. Here w_t denotes the postdecision observation ξ_{t+1} in the general model. Decisions satisfy $0 \le x_t \le \min\{s_t, 3\}$, with

$$s_{t+1} = s_t - x_t + 2w_t, \qquad c_t = a_t(3 - w_t - x_t)_+ + \frac{s_{t+1}}{8} + \mathbf{1}\{t = 6\}\, 4(3 - s_{t+1})_+,$$

where $(a_1, \ldots, a_6) = (1, 4, 2, 6, 3, 8)$. Policy optimization and certification receive sample paths and known physics, not the true probabilities.

Policy and exact reference. An affine proposal with exact projection is trained on 2,048 sample paths and selected using 1,024 validation paths, all within $\mathcal{D}_{\text{fit}}$. The selected policy $\widehat{\pi}$ is fixed before calibration. Zero, affine, and convex piecewise-linear (PWL) continuation functions use only training data. Exact rational Bellman recursion agrees with a continuous scenario-tree LP. The optimal

expected cost is $V^\star = 27.67381525$, the policy's expected cost is $J(\widehat{\pi}) = 27.795488057$, and the optimality gap is 0.121672807. These reference computations are excluded from certification.

Calibration and inner solution. The binary response is $H_t(a, w) = H_t(a, 0) + w\{H_t(a, 1) - H_t(a, 0)\}$. With empirical wet frequency $\widehat{p}_t$ estimated from $\mathcal{D}_{\mathrm{cal}}$, the uniform radius is

$$\beta_t = R_t \min\left\{1, \sqrt{\frac{\log(2T/\delta)}{2N_{\mathrm{cal}}}}\right\}, \qquad R_t = \sup_{a \in \mathcal{A}_t} |H_t(a, 1) - H_t(a, 0)|.$$

The contrast extrema occur at integer PWL-cell vertices and are computed exactly. We use $Q_t^- = \widehat{Q}_t - \beta_t$ and $r_t = 2\beta_t$. Within each affine cell, reservoir balances form a network matrix with integral bounds. Thus dynamic programming over integer vertices solves the continuous inner problem exactly. Before evaluation, the range rule takes the smaller of the exact paired-gap supremum over all 64 possible paths and the sum of stagewise ranges at fixed uncertainty realizations, then adds $2\sum_t \beta_t$ and bounds on outward-rounding error. It uses no true probabilities or evaluation samples.

Results. We use $\delta = \alpha = 0.025$ and 200 repetitions with fresh, independent sets $\mathcal{D}_{\mathrm{cal}}$ and $\mathcal{D}_{\mathrm{ev}}$, conditional on the training information. Methods share samples within each repetition. Table 13 reports the largest budget. Its paired-gap column averages the penalized policy cost minus the inner lower bound in (D.10), before adding the calibration correction $2\sum_t \beta_t$. The evaluation column is the Hoeffding margin $R_{\mathrm{H}}(\alpha)$ from (D.14). These three components sum to the reported bound $U = U_{\alpha,\delta}$.

Table 13: Single-reservoir certificates at $N_{\mathrm{cal}} = N_{\mathrm{ev}} = 131{,}072$, averaged over 200 repetitions. True optimality gap: 0.121673. Range-only U uses the sum of stagewise ranges.

Continuation	Paired gap	Calibration	Evaluation	U	Range-only U
Zero	1.2100	0.2960	0.0297	1.5357	1.7953
Affine	1.2100	0.5849	0.0308	1.8257	2.0853
PWL	0.1220	0.8630	0.0049	0.9898	1.3741

Each row covers the true optimality gap in 200/200 repetitions (per-row 95% exact binomial interval $[98.17, 100]\%$). Affine and Zero have the same reported mean paired gap, but Affine has a larger calibration correction and hence a larger certificate. PWL sharply reduces the paired gap, yet calibration dominates its reported bound, which remains about eight times the optimality gap. This finite-support check concerns one fixed policy and does not establish coverage after repeated retraining or scalability to nonlinear networks. Full training settings, sample-size sweeps, and solver audits are in the experimental supplement.

M Certificate under Binary Uncertainty and Coupled-Reservoir Study

This appendix documents the coupled-reservoir certificate experiment reported in Section 5.2. We use the empirical penalized objective $F_{\widehat{p}}$ and raw gap D from (D.12), with the path-dependent correction κ from (D.13). The range construction below implements Corollary D.4.

Structural range bound and implementation. For the coupled-hydro implementation, with deterministic initial state and state-based continuations, choose any v_1 determined using training data only (which does not enter the penalty) and define $A_t(s, x) = \widehat{Q}_t(s, x) - v_t(s)$. Telescoping yields $F_{\widehat{p}}(x, \xi) = v_1(s_1) + \sum_t A_t(s_t, x_t)$. A valid exact-solver range bound is the sum of the query-domain ranges of A_t, or the difference between the global maximum and minimum of $\sum_t A_t$ over feasible decisions and all binary sample paths. For the coupled hydro model and its fitted PWL functions, two deterministic dynamic programs optimize these extrema over both release and the next binary uncertainty realization, without enumerating complete paths. The integral affine-cell argument below justifies the continuous extrema.

Add the prescribed bounds on solver and arithmetic errors before evaluation. Let B_{raw} bound the raw gap D, and let K be the upper bound on κ defined in Corollary D.4. The implementation uses $K_{\mathrm{floor}} \le K/2$ and $K \le K_{\mathrm{cap}}$, so the reported-statistic interval has width $B_{\mathrm{raw}} + K_{\mathrm{cap}} - K_{\mathrm{floor}}$. Its outward-rounded mean and variance refer to the complete reported statistic, including κ. Constant centering shifts can be removed when forming a range, but the random calibration correction cannot be omitted from the variance.

The independent-stage assumption is specific to this calibration rule. Serial dependence requires valid conditional-moment calibration.

A coordinate-band control. For comparison, let m_t and M_t bound the response contrast $\Delta_t(a) = H_t(a,1) - H_t(a,0)$ over all queries. Set $b_t = (m_t + M_t)/2$ and replace the response by $\widetilde{H}_t(a,w) = H_t(a,w) - b_t w$. Its empirical centered term differs by the decision-independent quantity $-(\widehat{p}_t - w_t)b_t$, which cancels between the policy and inner objective on the same path. Applying the coordinate Hoeffding band to this response gives the constant correction $\rho \sum_t (M_t - m_t)$, with $\rho = \sqrt{\log(2T/\delta)/(2N_{\mathrm{cal}})}$, instead of $2\rho \sum_t \sup_a |\Delta_t(a)|$. The coordinate probability bands and their union bound do not themselves require independence across stages. The marginal-moment centering used here still uses the independent-stage model, while serial dependence requires valid conditional-moment calibration. Its evaluation statistic has a constant shift, so empirical Bernstein uses the raw-gap variance and raw-gap range. The five reporting rules are prespecified, and we never choose the smallest reported certificate after evaluation.

Coupled model. For reservoir $r = 1, \dots, R$, capacity is $r+3$, the release limit is $r+1$, and initial storage and terminal target are 3. For $R = 3$ and $T = 18$,

$$0 \le x_{t,r} \le \min\{s_{t,r}, r+1\}, \qquad s_{t+1,r} = \min\{r+3, s_{t,r} - x_{t,r} + r w_t\}.$$

The common binary observation $w_t \in \{0,1\}$, denoting ξ_{t+1} in the general model, is revealed after the decision. It determines shared demand $d_t = R(3 - w_t)$ and stage cost

$$c_t = a_t\Big(d_t - \sum_r x_{t,r}\Big)_+ + \sum_r \frac{r-1}{8} x_{t,r} + \mathbf{1}\{t = T\}\, 4 \sum_r (3 - s_{t+1,r})_+.$$

Prices repeat $(1, 4, 2, 6, 3, 8)$. Storing water is free and the transition enforces overflow, with no voluntary spill decision. Shared thermal backup couples the heterogeneous reservoirs. The binary variables are independent across stages, with wet-state probabilities repeating $(7, 12, 9, 13, 8, 11)/20$. Policy optimization and certification receive sample paths and known physics, not these probabilities. There is one common binary random variable per stage, rather than separate independent inflow factors for each reservoir, and the history summary z_t is constant.

Fixed policy and continuation functions. The policy projects an affine proposal based on the storage vector. It is trained on 2,048 sample paths using 400 full-horizon Adam updates, learning rate 0.02, $\lambda = 0$, and batch size $N_{\mathrm{bat}} = 64$. The output $\widehat{\pi}$ is the prespecified update-400 policy, and 1,024 separate validation paths do not select it. Earlier-checkpoint comparisons are in the experimental supplement.

Only the training paths supply stagewise frequencies. A coupled integer-grid Bellman recursion supplies training targets, followed by nonnegative least squares for $v_t(s) = b_t + \sum_{r,k} a_{t,r,k}(k - s_r)_+$ with $a_{t,r,k} \ge 0$. Knot values are stored as multiples of $1/65536$, and $v_{T+1} = 0$. The additive approximation does not represent every cross-reservoir value interaction. Integer-grid training targets are not asserted to equal the unrestricted continuous stochastic optimum.

Continuous inner extrema and arithmetic. For a fixed path in the support, write $y_{t,r} = s_{t,r} - x_{t,r}$. The penalized objective is PWL in the aggregate release $\sum_r x_{t,r}$ and separable in post-release storages, with integer breakpoints. Within each affine cell, a reservoir balance has form $y_{t,r} + x_{t,r} - a y_{t-1,r} = b$, where $a \in \{0,1\}$ records whether the previous overflow cap is active and b is integral. Together with the aggregate-release equation, columns have the signed incidence structure of a network, and all interval bounds are integral. Thus both continuous minimum and maximum occur at integral cell vertices. This justifies the inner DP and the residual-maximization DP used to bound the statistic. This deterministic path argument is not extended to the nonanticipative stochastic optimum.

For $R = 3$, the solver uses 210 storage vertices and 5,400 feasible state/decision vertices. Rational scaling makes objective comparisons exact, and outward arithmetic bounds cover the policy and reported statistics. A failed 10^{-6} cap stops reporting. Independent continuous PWL solvers agree within 2.28×10^{-13} on audited path problems. Implementation and arithmetic audits are retained in the experimental supplement.

Reference assessment. True probabilities are used only by a separate assessor. Exact rational expectation of the information-relaxation lower value gives a lower bound on $V^\star$, and the optimal feasible Markov policy with integer releases supplies an upper bound. For the main model, $V^\star \in [234.8622, 237.4875]$. All 2^{18} paths in the finite support are used only for this assessment and the fixed policies' expected costs. The policy's expected cost is $J(\widehat{\pi}) \simeq 245.5919$, and the optimality gap $J(\widehat{\pi}) - V^\star$ in the continuous-decision model lies in $[8.104, 10.730]$. We do not replace this interval with an exact gap value.

Calibration, evaluation, and comparisons. Reporting rules are fixed before 200 fresh repetitions, each with independent calibration and evaluation sets $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{ev}}$. Calibration sizes are $N_{\text{cal}} \in \{2{,}048, 8{,}192, 32{,}768\}$, while evaluation uses $N_{\text{ev}} = 8{,}192$ sample paths. We report $U = U_{\alpha,\delta}$ with $\delta = \alpha = 0.025$. Calibration objects and range bounds are fixed before evaluation. Methods and nested calibration prefixes share samples within a repetition, and repetitions are independent conditional on the training information. For every reported setting, the certificate exceeds the verified optimality-gap upper endpoint in 200/200 repetitions. The per-setting exact 95% binomial lower bound is 98.17%, conservatively applying to actual coverage because the true optimality gap is below that endpoint. This is a marginal statement, and outputs across settings are paired. No failed run is dropped.

Table 14: Certificate components for the fixed update-400 policy and PWL continuation, averaged over 200 repetitions with $N_{\text{cal}} = 32{,}768$ and $N_{\text{ev}} = 8{,}192$. Verified optimality-gap interval: $[8.104, 10.730]$. The common evaluation mean of the raw gap D in (D.12) is 10.732.

Calibration + evaluation	Calibration	Evaluation	U	$U/J(\widehat{\pi})$
Confidence band + Hoeffding	15.832	12.971	39.535	16.10%
Paired + Hoeffding	1.933	12.751	25.416	10.35%
Confidence band + Bernstein	15.832	0.589	27.153	11.06%
Paired + Bernstein	1.933	0.589	13.253	5.40%
Centered band + Bernstein	5.914	0.589	17.235	7.02%

Switching to Bernstein evaluation changes both the deterministic range and the concentration bound. The centered-coordinate control removes a component that cancels from paired gaps, while retaining a coordinatewise calibration band. The main binary calibration choice exploits stage independence through the joint Euclidean confidence region and evaluates the variance of the complete corrected gap. All variants use the same policies, continuations, sample budgets, and exact inner optima. The reduction therefore concerns the certificate construction.

Remaining limitations include additive value approximation, the assumption of independent binary uncertainty, and the gap between the reported upper bound and the verified optimality-gap interval.