

Learning to Sample under Energy Constraints

Suresh Rao, Anil Kumar, Prasad Rao
Institute of Aeronautical Engineering

Abstract

A sampling robot must trade scientific value against the energy and time needed to return its payload. Unknown terrain consumption complicates this tradeoff even when the possible outcomes of each action are known. We study finite-horizon sampling with stochastic traversal energy, uncertain acquisition success, and charging stations. A backward viability computation restricts actions to those that preserve a return on every supported trajectory. Within this action set, spatially pooled confidence intervals support pessimistic planning over energy and acquisition probabilities. We establish separate guarantees for returning safely and for the expected value of the returned sample. Experiments used a cratered terrain with narrow passages, three science regions, six charging stations, and three capacity–deadline combinations. We evaluated policies exactly over 24 independently generated training datasets per observation budget. Spatial transfer improved low-data decisions and value bounds, while a point-estimate planner often approached the oracle’s science value. An affordability-only planner exposed a different tradeoff: additional expected science came with a nonzero probability of failing to return. The study combines spatial information reuse with a substantive resource-constrained mission objective.

DOI: 10.18091/ijsts.v7i1.02

1 Introduction

Collecting a sample changes what a robot must protect. An empty robot can choose among several scientifically useful sites. After acquisition, the remaining task is to deliver the selected material before its energy or time expires. A planner must therefore compare the value of reaching a site with the value of returning from it. Uncertainty in terrain consumption affects both decisions.

Energy constraints couple each action to the feasibility of completing the mission. Consumption MDPs represent this dependence through a bounded resource and designated reload states [1]. Synthesis algorithms address qualitative resource-constrained objectives, with implementations available in FiMDP [2, 3]. Our mission also assigns value to the returned sample, requiring quantitative optimization among actions that preserve return feasibility.

Spatial side information connects observations with planning at related states [4–6]. Bounds on differences between transition probabilities link direct observations with evidence from related states [7]. Measurements on previously visited terrain can therefore inform decisions at locations with similar consumption or acquisition behavior. We use these bounds to estimate action probabilities within the set of actions that preserve a return.

We combine these ideas in a quantitative finite-horizon problem. The robot chooses a science region, attempts acquisition, and returns to a charging station. Motion has two possible energy costs, while acquisition can succeed or fail. The energy outcomes and transition support are known; their probabilities are learned. We compute the states from which a return remains possible under every supported outcome sequence. We then optimize sample value inside that set using confidence intervals built from spatial observations.

The experiments distinguish three quantities that are often conflated: actual science value, a statistical lower bound on that value, and the probability of returning. A well-calibrated model can improve the lower bound without materially changing the executed policy. A policy can also increase expected science by accepting a small chance of becoming stranded. Our comparison measures both effects directly.

1.1 Related Work

Budgeted informative planning selects robot paths that maximize information gathered under travel or sensing constraints [8, 9]. Approximation algorithms address discrete sensing locations, while sampling-based search also accommodates continuous motion spaces. Those objectives depend on measurements collected along a path. Our terminal reward instead depends on successfully delivering one acquired payload. Adaptive site selection treats scientific value and limited sampling opportunities as sequential decisions [10]. We use a fully observed physical state and a single returned sample, which permits exact policy evaluation. The uncertainty lies in stationary action probabilities and their finite-data estimates.

Safe exploration formulations account for whether an agent can recover after visiting an uncertain state [11]. Analogies between state–action pairs can support safe exploration and near-optimal learning when stochastic dynamics are unknown [12]. Our model assumes known outcome support and fixed training observations. We enforce return feasibility before estimating probabilities, which separates the safety calculation from the statistical quality of the selected policy.

Shielding restricts learning through an action-admissibility specification [13]. Predictive safety filters similarly assess a proposed action through the availability of a safe continuation [14]. Our finite-state restriction requires a return by a deadline, which is stronger than checking whether the next action is affordable. Robust dynamic programming provides the quantitative optimization over interval transition models [15, 16]. Confidence-based model learning supplies the statistical interpretation of those intervals [17, 18]. We use the standard distinction between expected-cost constraints and pathwise resource feasibility [19, 20].

The resulting algorithm is an explicit finite-horizon construction. Its contribution is the composition of spatial estimation, return viability, and payload-dependent reward, together with an evaluation of their separate effects. The capacity-dependent dynamic program has different complexity from the efficient qualitative algorithms in FiMDP. We do not benchmark against those algorithms on a different objective.

We restrict the cited literature to publications available by December 31, 2023. We conducted the computational experiments in September 2026.

2 Sampling Mission Model

2.1 State and Resource Model

Let $s \in S$ denote a traversable location, $b \in \{0, \dots, C\}$ the available energy, and h the remaining number of actions. The payload state $p \in \{0, 1, \dots, J\}$ is zero before acquisition and identifies the acquired science class afterward. A successful return with class p earns v_p , with $v_0 = 0$. The set $R \subseteq S$ contains charging stations. The augmented state is $x = (s, p, b)$.

An action a has two known outcomes, indexed by $z \in \{0, 1\}$. Outcome z specifies the successor $f_z(s, p, a)$ and energy consumption $c_z(s, a)$. Its probability is $q(s, a)$ for $z = 1$ and $1 - q(s, a)$ for $z = 0$. The planner learns q while preserving the outcome correspondence. At a station, energy resets to capacity before the next action. For an outcome ending at physical location s' , the resource update is

$$F_z(s, b, a) = \begin{cases} C, & s' \in R, \\ \tilde{b} - c_z(s, a), & s' \notin R, \end{cases} \quad \tilde{b} = \begin{cases} C, & s \in R, \\ b, & s \notin R. \end{cases} \quad (1)$$

An action is unaffordable if any supported cost exceeds $\tilde{b}$.

For a motion action, both outcomes reach the same neighboring cell but consume different energy amounts. For an acquisition action, both outcomes consume the same energy; one acquires the site's payload and the other leaves the robot empty. The robot attempts acquisition only while empty. A station permits termination with the current payload. Termination away from a station fails the return requirement.

Outcome-dependent consumption can be represented in a conventional consumption MDP by inserting an outcome state with deterministic consumption. The physical decision consumes one mission step in our model; the auxiliary accounting state consumes none. This expansion connects the semantics to established consumption models while retaining a compact implementation [1, 3].

2.2 Mission Objective and Execution Assumptions

The objective is expected science delivered within H actions:

$$\max_{\pi} \mathbb{E}^{\pi}[v_{p_{\tau}}], \quad (2)$$

$\tau \leq H, \quad s_{\tau} \in R, \quad b_{\tau} \geq 0$ on every supported execution.

The terminal time τ is chosen by the policy at a station. Returning without a sample satisfies the resource requirement and earns zero science. This permits a conservative fallback when an attempted acquisition fails. The policy must still preserve that fallback after every admitted action.

Assumption 2.1 (Execution model). *The transition support, outcome-dependent consumption, station locations, and recharge behavior are known and valid during deployment. The robot observes location, payload, remaining energy, and remaining time exactly. Every action consumes one mission step, and termination is available only at a station.*

A physical system can violate this model through unmodeled housekeeping power, uncertain charging, localization error, or an unlisted motion outcome. Those effects require additional state or conservative support bounds. The results below apply to the finite model in Assumption 2.1.

3 Return Viability

3.1 Viability Computation

Let K_h contain physical-resource states from which the robot can return within h actions for every supported outcome sequence. The terminal set is

$$K_0 = \{(s, b) : s \in R, 0 \leq b \leq C\}. \quad (3)$$

Define the admissible action set

$$A_h(s, b) = \left\{ a : \tilde{b} \geq \max_z c_z(s, a), (s'_z, F_z(s, b, a)) \in K_{h-1} \forall z \right\}. \quad (4)$$

A station is viable through termination. Every other state belongs to K_h precisely when $A_h(s, b)$ contains a motion action. Sampling cannot create return viability because it leaves the physical state unchanged while spending time and energy. After constructing K_h , we admit sampling actions through the same successor test in (4).

The construction depends on the support of each action, so it uses no learned probability. A very unlikely high-consumption outcome has the same status as a frequent one in the viability computation. The probability model matters when the planner chooses among viable actions. This distinction prevents a favorable energy estimate from silently enlarging the safe action set. The station set serves as a terminal recovery condition, analogous to the terminal safe set used in predictive safety filtering [14]. Here, the remaining mission time decreases after every action, so the recursion enforces a specific return deadline.

Proposition 3.1 (Return guarantee). *Under Assumption 2.1, a policy starting in K_H and selecting only actions in (4) returns to a station within H steps on every supported execution. No admitted action exhausts the resource.*

Proof. At remaining horizon h , each admitted action is affordable and sends every supported outcome into K_{h-1} . Induction preserves this membership until termination or horizon zero. The terminal set K_0 contains only station states with nonnegative energy. Station termination also satisfies both requirements. $\square$

3.2 Affordability and Return Feasibility

A next-step energy check can admit an action that leaves the robot outside K_{h-1} . For example, consider a site one motion from a station. A return can require either one or three energy units. Sampling at that site with four units and an acquisition cost of two is affordable, but leaves insufficient energy for one return outcome. A planner using only affordability can accept this gamble if the expected sample value is high enough. The viability test rejects it regardless of the estimated probability of the expensive return.

The deadline requires every supported outcome sequence to reach a station within the remaining horizon. Almost-sure eventual return can permit trajectories of unbounded duration and therefore does not imply this finite-horizon guarantee. Consumption-MDP synthesis addresses broader qualitative objectives, including recurrent service [2, 3]. Our backward recursion enforces the bounded mission specification directly.

4 Spatial Probability Estimation

4.1 Spatial Similarity and Observation Pools

We transfer observations only between state–action pairs with matching outcome semantics. Let $\phi(s)$ denote spatial features and let $g(s, a)$ identify the admissible transfer class. We assume

$$|q(s, a) - q(u, a)| \leq L \|\phi(s) - \phi(u)\|_2 \quad \text{when } g(s, a) = g(u, a). \quad (5)$$

The class can distinguish geology and acquisition type. Consequently, a short Euclidean distance does not force pooling across a discontinuous material boundary. This explicit correspondence follows the information-reuse principle of spatial side-information methods [5, 7].

Each candidate pool $B_r(s, a)$ has a radius fixed before observing outcomes. It contains independent Bernoulli observations with total count N_r , empirical fraction $\hat{q}_r$, and sample-weighted distance $\bar{d}_r$ from the target. The observations can have different means. The supplied bound implies that their average mean differs from $q(s, a)$ by at most $L\bar{d}_r$.

Assumption 4.1 (Training data). *Training observations are independent conditional on the fixed terrain model. Counts and candidate pools are chosen before outcomes are observed. The probabilities remain stationary, and (5) holds with the supplied class labels and constant.*

4.2 Two-Sided Confidence Intervals

Let M be the number of valid state–action pairs and K the number of candidate radii. Relative-entropy confidence indices exploit the geometry of Bernoulli observations [21]. We use both endpoints and allocate error across the fixed candidate pools. For each pool, we invert binary relative entropy:

$$[\ell_r, u_r] = \left\{ v \in [0, 1] : N_r \text{kl}(\hat{q}_r \| v) \leq \log \frac{2MK}{\delta} \right\}. \quad (6)$$

For an empty pool we use $[0, 1]$. The spatial interval is the intersection

$$\underline{q}(s, a) = \max_r \max(0, \ell_r - L\bar{d}_r), \quad \bar{q}(s, a) = \min_r \min(1, u_r + L\bar{d}_r). \quad (7)$$

An empty intersection triggers a direct-data fallback in the implementation. Such an intersection cannot occur on the simultaneous coverage event below.

Lemma 4.1 (Simultaneous coverage). *Under Assumption 4.1, every true probability belongs to (7) with training-data probability at least $1 - \delta$.*

Proof. For independent Bernoulli variables with possibly different means, concavity of the log moment-generating function yields the usual Bernoulli exponential bound at their average mean. Each one-sided relative-entropy endpoint therefore has error at most $\delta/(2MK)$. The spatial bias bound transfers coverage of the pooled mean to the target mean. A union bound covers both endpoints, all targets, and all candidate pools. Their intersection retains the true target probability on that event [22]. $\square$

Two-sided intervals are necessary because the learned event has different consequences across actions. For motion, outcome one consumes more energy. For acquisition, outcome one obtains a payload. Even within one action type, recharge and remaining time can reverse the ordering of

continuation values. The planner evaluates both endpoints instead of assigning a fixed favorable direction to q .

5 Planning and Performance Bounds

Let $\underline{V}_h(s, p, b)$ be the lower value at remaining horizon h . A station permits immediate value v_p . Away from a station, horizon zero has no admissible mission completion. For each action in $A_h(s, b)$, define

$$W_z = \underline{V}_{h-1}(f_z(s, p, a), F_z(s, b, a)), \quad (8)$$

$$\underline{Q}_h(s, p, b, a) = \min_{q \in [\underline{q}(s, a), \bar{q}(s, a)]} [(1 - q)W_0 + qW_1]. \quad (9)$$

The minimum uses $\underline{q}$ when $W_1 \geq W_0$ and $\bar{q}$ otherwise. We maximize (9) over admitted actions and station termination. The stored maximizer defines a policy on viable augmented states. This is rectangular robust dynamic programming on a resource-augmented finite MDP [15, 16].

Theorem 5.1 (Science-value certificate). *Under Assumptions 2.1 and 4.1, with training-data probability at least $1 - \delta$, the stored policy satisfies*

$$\mathbb{E}^\pi[v_{p_\tau} \mid s, p, b, h] \geq \underline{V}_h(s, p, b) \quad (10)$$

for every viable augmented state and horizon. The return guarantee in Proposition 3.1 holds independently of this statistical event.

Proof. On the coverage event, the true probability is an admissible value in every minimization in (9). The claim holds for terminal station rewards. Assume it holds for the continuation policy at horizon $h - 1$. The adverse expectation cannot exceed the expectation under the true probability, and replacing continuation bounds with actual policy values cannot decrease that expectation. Induction proves the performance statement. The separate support induction proves return feasibility. $\square$

Overlapping spatial pools produce dependent estimates. The coverage argument permits that dependence through the union bound. The planner then uses a product of marginal intervals, allowing probability choices that may be incompatible with a single shared spatial model. This enlargement preserves validity while adding conservatism. It helps explain why an executed policy can perform well before its reported value approaches the oracle. Coupled-uncertainty MDPs address a related source of conservatism by restricting how departures from nominal dynamics can occur jointly [23]. Their deviation budgets differ from our spatial constraints, but illustrate how structural restrictions can rule out incompatible adverse models.

The viability computation costs $O(H|S|C|A|)$ for two outcomes. Quantitative planning costs $O(H(J + 1)|S|C|A|)$ and stores $O(H(J + 1)|S|C|)$ policy entries. The implementation retains only two value layers during backward recursion. Its dependence on numerical capacity is pseudopolynomial when capacity is encoded in binary. A large continuous battery state would require abstraction or approximation with a separate error analysis.

Algorithm 1. Spatial learning with a guaranteed return

Require: Outcome support and costs, stations R , capacity C , horizon H , observations D

- 1: $K_0 \leftarrow \{(s, b) : s \in R\}$
- 2: **for** $h = 1, \dots, H$ **do**
- 3: Compute K_h and A_h by supported-successor tests
- 4: **end for**
- 5: Construct class-restricted intervals using (7)
- 6: Initialize station rewards and terminal feasibility
- 7: **for** $h = 1, \dots, H$ **do**
- 8: **for** each viable (s, p, b) **do**
- 9: Evaluate both interval endpoints for every $a \in A_h(s, b)$
- 10: Store the maximizing action, including termination when $s \in R$
- 11: **end for**
- 12: **end for**
- 13: Execute with the observed payload, energy, and remaining horizon

6 Experimental Design**6.1 Terrain and Sampling Model**

We generated a 23×19 terrain containing four crater interiors and two ridges with restricted passages. After removing obstacles, the graph contained 360 locations and 1298 valid motion or acquisition actions. The robot began at (1,2). Six stations supported repeated energy replenishment. Science regions centered near (5,3), (11,12), and (20,12) offered returned values 1, 1.6, and 2.4. An acquisition attempt consumed three energy units and one time step. Figure 1 shows the resulting choice between nearby, moderate-value, and distant sites.

A known binary geology class was defined by $g(x, y) = \mathbf{1}\{x + 0.55y > 17\}$. Motion consumed $1 + g$ or $3 + g$ energy units. For cardinal action a , the probability of the expensive outcome was

$$q_{\text{move}}(x, y, a) = 0.12 + 0.004x + 0.003y + 0.13g(x, y) + 0.035\mathbf{1}\{a \in \{\text{east}, \text{north}\}\}. \quad (11)$$

Acquisition succeeded with probability

$$q_{\text{sample}}(x, y) = 0.38 + 0.005x + 0.006y - 0.05g(x, y). \quad (12)$$

Both within-class gradient norms are below the supplied $L = 0.009$ per cell. These analytic fields provide a known truth for evaluating the guarantees. They are synthetic terrain and energy models, with no claim of calibration to a rover platform.

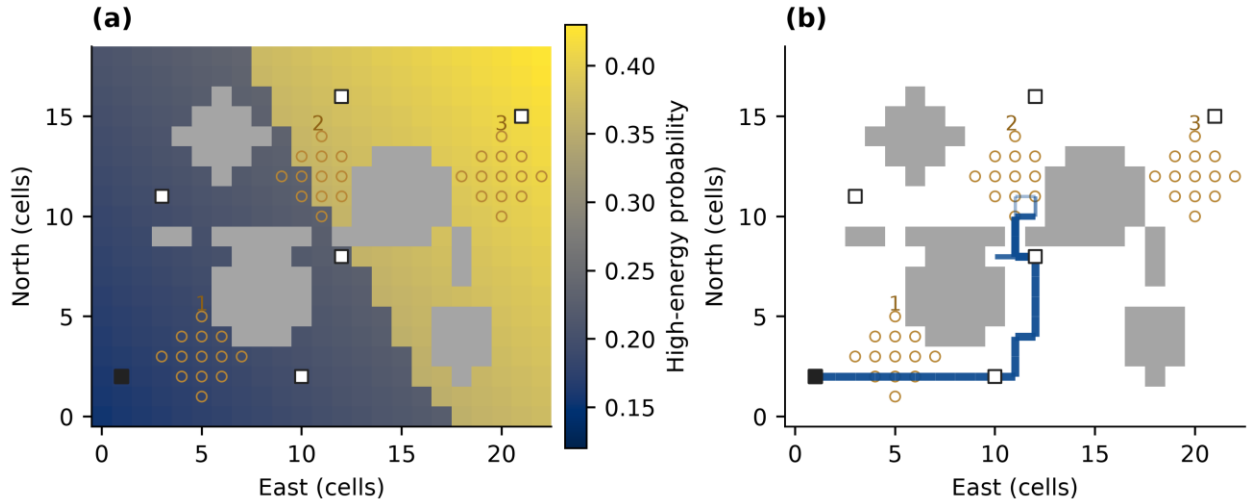


Figure 1. Sampling terrain. (a) Probability of the expensive eastward motion outcome; grey cells are obstacles, outlined squares are charging stations, and the filled square is the start. Numbered science regions have values 1, 1.6, and 2.4. (b) Expected motion counts under one spatial-confidence policy; line width and opacity increase with expected use. Counts were computed by forward probability propagation. The policy used 64 observations per action, capacity 28, and horizon 50.

6.2 Mission Parameters and Training Data

We tested three capacity–deadline pairs: tight (22,38), nominal (28,50), and extended (34,64). Changing both quantities alters the reachable science regions and the number of acquisition attempts available before a safe return. Each valid state–action pair received 4, 16, 64, or 256 independent observations. We generated 24 training datasets at each budget, using common data for all planners within a dataset. Candidate radii were 0, 2, 4, 7, and 12 cells, with $\delta = 0.05$. Pools matched action and geology; acquisition pools also matched science class.

Direct used simultaneous intervals from the target action alone. *Spatial* used the bias-corrected pooled intervals. *Kernel* used the pooled empirical probability from the radius with the smallest corrected interval width. It optimized a point model under the same return restriction. *No bias* removed the spatial bias allowance while retaining the concentration calculation. *Oracle* used the generating probabilities and the same return restriction. These comparisons isolate information reuse, statistical conservatism, and probability-model error.

Budget only used the same point estimates as *Kernel*, but required only next-action affordability for every supported cost. It optimized returned science and thus had an incentive to return, while permitting trajectories that could become stranded. This is a quantitative risk-taking comparison with the same reward, data, and resource accounting. Its difference from *Kernel* is the return-viability restriction.

6.3 Policy Evaluation and Statistical Analysis

We computed actual science value and return probability by exact backward evaluation of each stored policy under the generating probabilities. Independent forward probability propagation checked representative policies. We also audited all supported successors of stored shielded actions. No Monte Carlo rollout error enters the reported policy values. The uncertainty bands

describe variation across independent training datasets and use two-sided Student-*t* intervals for mean performance.

We measured dynamic-program execution time after compilation. Timing excludes probability estimation, JIT compilation, and the viability computation shared by planners in a mission regime. It is an implementation diagnostic on the execution host, rather than a hardware-independent comparison with published planners. The saved results retain individual seeds, estimates, and policy snapshots.

7 Results

7.1 Science Value and Confidence Bounds

Spatial pooling improved science value when individual action observations were scarce. In the nominal mission with four observations per action, Direct achieved 1.10 expected science, while Spatial achieved 1.40. Their corresponding lower bounds were 0.23 and 1.07. The difference reflects a practical decision: uncertain long trips can make the nearby science region preferable under a pessimistic model. Pooling reduces probability uncertainty enough to justify more valuable missions.

At 64 observations, Spatial achieved 1.60 expected science and a lower bound of 1.59 in the nominal mission. Kernel achieved 1.60 under the same return restriction. Its point-model values are predictions and have no simultaneous lower-confidence interpretation. Figure 2 shows that good executed performance can coexist with a substantial certificate gap.

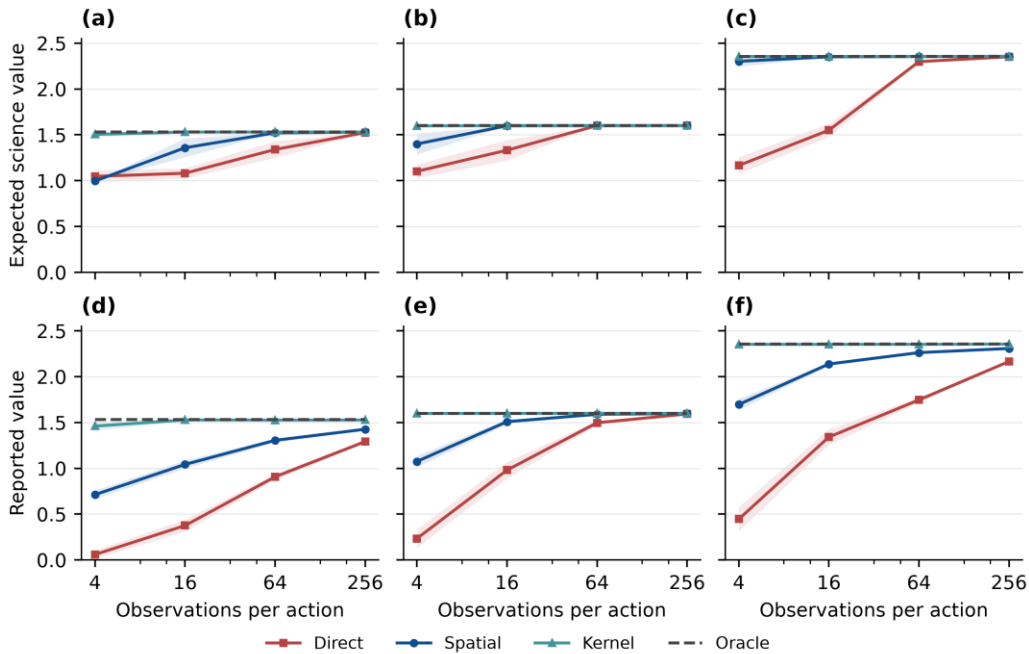


Figure 2. Learning curves across mission regimes. Columns show the tight, nominal, and extended missions. Panels (a)–(c) report expected returned science under the true model; panels (d)–(f) report the planner’s value. Direct and Spatial values are lower bounds on the stated coverage event; Kernel values are point-model predictions. Shading gives 95% Student-*t* intervals over 24 training datasets. Oracle uses the true probabilities with the return constraint.

Table 1. Results at 64 observations per action, averaged over 24 datasets. Lower bounds are reported only for methods whose intervals include the spatial bias allowance or use direct observations. Return is an exact probability averaged over datasets. Time is median quantitative-planning time, excluding compilation and shared viability construction.

Mission	Planner	Science	Lower bound	Return (%)	Time (ms)
Tight	Direct	1.34	0.91	100.00	9
	Spatial	1.52	1.30	100.00	11
	Kernel	1.53	—	100.00	10
	No bias	1.53	—	100.00	10
	Budget only	1.58	—	98.46	20
	Oracle	1.53	—	100.00	—
Nominal	Direct	1.60	1.50	100.00	18
	Spatial	1.60	1.59	100.00	24
	Kernel	1.60	—	100.00	22
	No bias	1.60	—	100.00	23
	Budget only	1.67	—	82.40	35
	Oracle	1.60	—	100.00	—
Extended	Direct	2.30	1.75	100.00	31
	Spatial	2.35	2.26	100.00	44
	Kernel	2.36	—	100.00	39
	No bias	2.36	—	100.00	40
	Budget only	2.39	—	99.77	56
	Oracle	2.36	—	100.00	—

7.2 Effect of the Return Constraint

All planners using the viability restriction returned with probability one under the generating model. Budget only increased science in some regimes by accepting paths that could end away from a station. In the nominal mission at 64 observations, it achieved 1.67 expected science with a 17.60% failure-to-return probability. The same point estimates with the full return constraint achieved 1.60 expected science. This paired comparison measures the cost of the stronger specification.

The difference persists when the probability estimates are accurate because it comes from the admissible policy class. Increasing data can correct a model error, but it cannot make a risky supported trajectory satisfy a pathwise return requirement. Figure 3 separates that structural tradeoff from the statistical gap between actual science and its lower bound.

7.3 Capacity and Deadline Sensitivity

A separate oracle sweep evaluated seven capacities from 14 to 38 and seven deadlines from 24 to 72 actions. The 49 combinations isolated physical mission difficulty from probability-estimation error. Every evaluated policy satisfied the return condition. At capacity 14, all tested deadlines yielded zero returned science. Increasing time alone could not supply the energy needed for a viable acquisition trip.

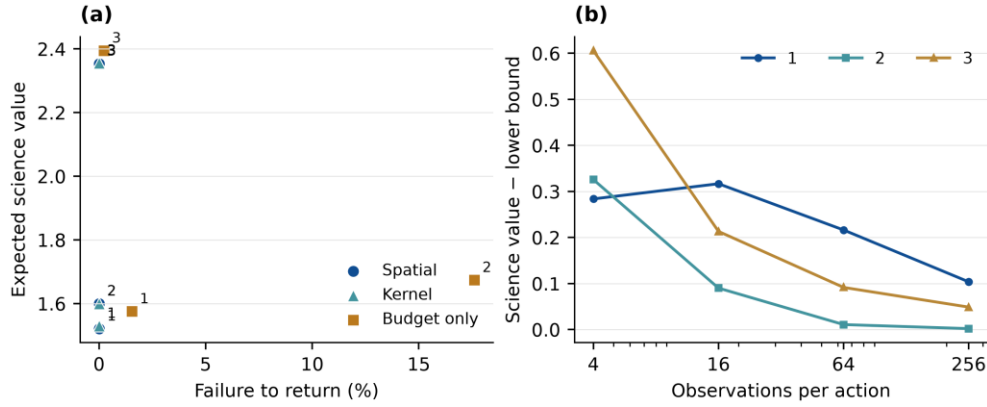


Figure 3. Safety and statistical conservatism. (a) Expected science versus failure-to-return probability at 64 observations per action. Labels 1, 2, and 3 identify the tight, nominal, and extended missions. Kernel and Budget only share the same point estimates. (b) Mean difference between actual science and the spatial lower bound, with the same mission labels. Values were evaluated exactly and then averaged across training datasets.

Larger capacities exposed successive science opportunities in Figure 4(a). Within an accessible region, additional time supported more acquisition attempts and improved expected return value. The viable state–energy fraction depended on capacity but was identical across the tested deadlines at each capacity. Thus, the shortest tested deadline already allowed basic recovery from the same states; longer deadlines mainly improved the sampling objective. This separates the constraint’s physical feasibility from the quantitative use of remaining time.

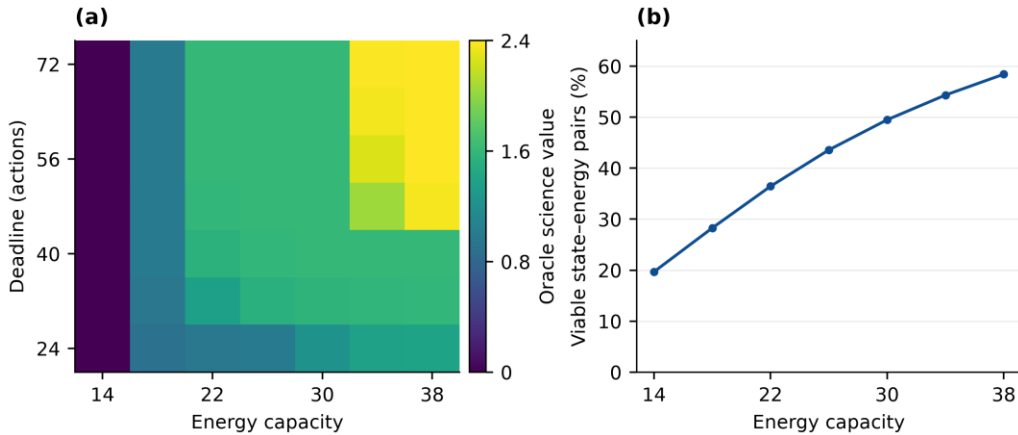


Figure 4. Capacity–deadline sweep under the true probability model. (a) Optimal expected returned science for 49 evaluated parameter pairs; each colored cell corresponds to one exact dynamic-program solution. (b) Fraction of physical-state and energy-level pairs in the return-viable set. This fraction was identical across the tested deadlines at each capacity. These deterministic model evaluations have no training or rollout confidence interval.

7.4 Experimental Limitations

The smooth within-class model favors methods supplied with the correct class labels and a valid spatial bound. Unknown geology, a wrong outcome correspondence, or drift during the mission can invalidate that information. The No bias ablation measures removal of one correction within this controlled setting; it does not establish calibration on arbitrary terrain.

The terrain contains competing routes and nontrivial return constraints, but it remains one synthetic map. The three mission regimes test resource and deadline sensitivity on that map. Physical validation would require measured consumption distributions, uncertain station availability, and energy used during sensing and waiting. A study across independent terrain families would be needed before making claims about general navigation performance.

8 Discussion

The experiment suggests a division of computational effort. The return set can be computed once from conservative support and energy bounds, then reused as probability estimates improve. Learning can focus on distinctions that change scientific decisions within that set. This architecture permits a high-value point-model policy when empirical performance is the priority, or a pessimistic policy when a value certificate is required. Both can preserve the same return specification.

Spatial information is most useful when observations are scarce and the bound on between-location variation is informative. Its benefit decreases when direct observations already determine the relevant action rankings. The remaining certificate gap then reflects both finite-data uncertainty and rectangular relaxation of the shared spatial model. A tighter planner could optimize over a coupled confidence set, at greater computational cost.

The explicit capacity expansion is appropriate for auditing small finite models. Larger missions should exploit consumption structure, abstract the battery state conservatively, or use rolling horizons with verified terminal return sets. Predictive safety filtering offers a related way to retain a feasible backup continuation during repeated finite-horizon planning [14]. The qualitative synthesis results and FiMDP implementation offer relevant building blocks for those extensions [2, 3]. Any approximation would need to preserve the supported-successor test before inheriting the return guarantee.

9 Conclusion

We studied learning for a sampling mission that must finish at a charging station. Backward viability preserved a return under every supported energy and acquisition outcome. Spatial confidence intervals improved low-data planning and supplied lower bounds on returned science. Exact evaluation separated these benefits from the additional science available to a planner that accepted a chance of becoming stranded. The remaining challenges concern uncertain support, continuous energy models, and validation with measured terrain consumption.

References

- [1] František Blahoudek, Tomáš Brázdil, Petr Novotný, Melkior Ornik, Pranay Thangeda, and Ufuk Topcu. Qualitative Controller Synthesis for Consumption Markov Decision Processes. In *Computer Aided Verification*, pages 421–447, 2020. doi: [10.1007/978-3-030-53291-8_22](https://doi.org/10.1007/978-3-030-53291-8_22).

- [2] František Blahoudek, Murat Cubuktepe, Petr Novotný, Melkior Ornik, Pranay Thangeda, and Ufuk Topcu. Fuel in Markov Decision Processes (FiMDP): A Practical Approach to Consumption. In *Formal Methods*, pages 640–656, 2021. doi: [10.1007/978-3-030-90870-6_34](https://doi.org/10.1007/978-3-030-90870-6_34).
- [3] František Blahoudek, Petr Novotný, Melkior Ornik, Pranay Thangeda, and Ufuk Topcu. Efficient Strategy Synthesis for MDPs With Resource Constraints. *IEEE Transactions on Automatic Control*, 68(8):4586–4601, 2023. doi: [10.1109/tac.2022.3209612](https://doi.org/10.1109/tac.2022.3209612).
- [4] Melkior Ornik, Jie Fu, Niklas T. Lauffer, W. K. Perera, Mohammed Alshiekh, Masahiro Ono, and Ufuk Topcu. Expedited Learning in MDPs with Side Information. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 1941–1948, 2018. doi: [10.1109/cdc.2018.8619134](https://doi.org/10.1109/cdc.2018.8619134).
- [5] Pranay Thangeda and Melkior Ornik. Safety-Guaranteed, Accelerated Learning in MDPs with Local Side Information. In *2020 American Control Conference (ACC)*, pages 1099–1104, 2020. doi: [10.23919/acc45564.2020.9147372](https://doi.org/10.23919/acc45564.2020.9147372).
- [6] Pranay Thangeda. Efficient learning and planning using spatial side information. Master’s thesis, University of Illinois at Urbana-Champaign, 2020. M.S. thesis.
- [7] Pranay Thangeda, Melkior Ornik, and Ufuk Topcu. Expedited Online Learning With Spatial Side Information. *IEEE Transactions on Automatic Control*, 68(3):1479–1491, 2023. doi: [10.1109/tac.2022.3153278](https://doi.org/10.1109/tac.2022.3153278).
- [8] A. Singh, A. Krause, C. Guestrin, and W. J. Kaiser. Efficient Informative Sensing using Multiple Robots. *Journal of Artificial Intelligence Research*, 34:707–755, 2009. doi: [10.1613/jair.2674](https://doi.org/10.1613/jair.2674).
- [9] Geoffrey A. Hollinger and Gaurav S. Sukhatme. Sampling-based robotic information gathering algorithms. *The International Journal of Robotics Research*, 33(9):1271–1287, 2014. doi: [10.1177/0278364914533443](https://doi.org/10.1177/0278364914533443).
- [10] Pranay Thangeda and Melkior Ornik. Adaptive Sampling Site Selection for Robotic Exploration in Unknown Environments. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4120–4125, 2022. doi: [10.1109/iros47612.2022.9982113](https://doi.org/10.1109/iros47612.2022.9982113).
- [11] Teodor Mihai Moldovan and Pieter Abbeel. Safe Exploration in Markov Decision Processes. In *Proceedings of the 29th International Conference on Machine Learning*, 2012. URL <https://arxiv.org/abs/1205.4810>.
- [12] Melrose Roderick, Vaishnavh Nagarajan, and J. Zico Kolter. Provably Safe PAC-MDP Exploration Using Analogies. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1216–1224. PMLR, 2021. URL <https://proceedings.mlr.press/v130/roderick21a.html>.
- [13] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe Reinforcement Learning via Shielding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: [10.1609/aaai.v32i1.11797](https://doi.org/10.1609/aaai.v32i1.11797).
- [14] Kim Peter Wabersich and Melanie N. Zeilinger. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica*, 129:109597, 2021. doi: [10.1016/j.automatica.2021.109597](https://doi.org/10.1016/j.automatica.2021.109597).

- [15] Arnab Nilim and Laurent El Ghaoui. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research*, 53(5):780–798, 2005. doi: [10.1287/opre.1050.0216](https://doi.org/10.1287/opre.1050.0216).
- [16] Garud N. Iyengar. Robust Dynamic Programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. doi: [10.1287/moor.1040.0129](https://doi.org/10.1287/moor.1040.0129).
- [17] Alexander L. Strehl and Michael L. Littman. An analysis of model-based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008. doi: [10.1016/j.jcss.2007.08.009](https://doi.org/10.1016/j.jcss.2007.08.009).
- [18] Michael Kearns and Satinder Singh. Near-Optimal Reinforcement Learning in Polynomial Time. *Machine Learning*, 49(2-3):209–232, 2002. doi: [10.1023/a:1017984413808](https://doi.org/10.1023/a:1017984413808).
- [19] Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999. doi: [10.1201/9781315140223](https://doi.org/10.1201/9781315140223). First edition.
- [20] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994. doi: [10.1002/9780470316887](https://doi.org/10.1002/9780470316887).
- [21] Aurélien Garivier and Olivier Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19, pages 359–376. PMLR, 2011. URL <https://proceedings.mlr.press/v19/garivier11a.html>.
- [22] Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: [10.1080/01621459.1963.10500830](https://doi.org/10.1080/01621459.1963.10500830).
- [23] Shie Mannor, Ofir Mebel, and Huan Xu. Lightning Does Not Strike Twice: Robust MDPs with Coupled Uncertainty. In *Proceedings of the 29th International Conference on Machine Learning*, 2012. URL <https://arxiv.org/abs/1206.4643>.