COST-AWARE ACTIVE LEARNING OF OPTIMAL CONTROL POLICIES ON THE SWITCHING SET

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we investigate the discounted optimal control problem of control-affine systems under input constraints. The optimal feedback of such problems saturates and switches between input levels across surfaces in the state space. Imitating such a controller requires one optimal control solve per expert sample, so the samples must be placed where they matter. For the policy over a finite input grid, learned as a classifier, we show that these samples should lie on the switching surfaces. Their value there is set by the cost slope g, the rate at which the cost of the wrong input grows with the distance from the surface. We prove that the cost-sensitive loss of the learned policy is the g-weighted squared error of its switching surfaces and that the optimal sample density on them is proportional to $g^{a/(a+2)}$, with a the local rate of the learner. Margin sampling loses a constant factor, computable from the expert before learning, and an idealised learner improves the loss rate of uniform sampling from $N^{-2/n}$ to $N^{-2/(n-1)}$ in n state dimensions. The practical sampler estimates g from the already-returned action values, with no additional expert samples. On the inverted pendulum, two mountain-car configurations and a four-dimensional cart-pole, the sampler matches the accuracy of uniform sampling with 1.8 to more than four times fewer expert samples and its closed-loop cost gap with 1.4 to 4.7 times fewer. The factor computed beforehand predicts on which systems the cost weighting pays off.

1 INTRODUCTION

The optimal feedback of a constrained control problem is defined one state at a time, as the first input of a dynamic-programming recursion or a trajectory optimisation. Evaluating it online, thousands of times per second, is out of reach for most systems, so the feedback law is fitted offline to examples that the optimiser, the expert, provides. This is the setting of explicit model predictive control (Bemporad et al., 2002) in its learned form (Hertneck et al., 2018; Karg & Lucia, 2020) and of imitation learning from a planner (Ross et al., 2011). Every example costs one solve, and the budget of solves bounds the quality of the learned law. This paper asks where in the state space the examples should be requested.

Let us take an example of swing up a pendulum with a bounded torque. The optimal input is saturated over most of the state plane and reverses along curves, and with a finite set of input levels the law is piecewise constant, a classifier whose decision boundary is the switching set. Uniform sampling spends most of its queries where the law is constant and one label per region would do, while the error sits at the switching set. Margin sampling, the oldest idea of active learning (Tong & Koller, 2001; Settles, 2009), sends the queries there, in proportion to surface area. But the switching set is not uniform in value. Where the input steps to the neighbouring level a wrong decision costs almost nothing, where it reverses it costs a great deal, and margin sampling cannot tell the two apart. Cost-sensitive active learning treats such costs as an abstract cost vector attached to each query (Krishnamurthy et al., 2017; Njike & Siebert, 2023). The control problem gives them geometry. The switching set decomposes into surfaces, each with its own cost slope, and that decomposition turns the question where to query into an allocation problem across surfaces.

The value of a query on the switching set is set by the cost slope, the rate at which the cost of the wrong input grows with the distance from the surface. The budget should follow a power of this

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.

slope, not the surface area, and the slope can be estimated from the action values the expert has already returned.

We learn the policy over the input grid as a classifier by batch-mode active learning from an exact expert. Each iteration keeps the low-margin band of the current classifier, draws the next queries from it with a density proportional to a power of a local slope estimate regressed from the returned action values, and refits under the cost-sensitive loss of classification-based policy iteration (Lazaric et al., 2016). No importance weights are needed because the expert's labels are deterministic.

- **Theory** (Section 4). A geometric decomposition of the policy-learning problem into switching surfaces with heterogeneous misclassification costs and the surface-wise query allocation it yields. The cost-sensitive loss is the slope-weighted squared error of the surfaces, on the surfaces the input grid adds the slope factorises into the input jump and the gradient of the switching function and on the switching set of the continuous problem it is of order one, the optimal query density is proportional to $g^{a/(a+2)}$ in the cost slope g with a the learner's local rate, an idealised learner improves the rate from $N^{-2/n}$ to $N^{-2/(n-1)}$ in n state dimensions, margin sampling loses a factor computable from the expert before learning, and the same error bounds the closed-loop cost gap.
- **Method** (Section 3). A mesh-free band sampler weighted by the slope estimate, at no extra expert query, with the softmax input jump as its label-free proxy.
- **Evidence** (Section 5). On the pendulum, two mountain-car configurations and a four-dimensional cart-pole the sampler matches the accuracy of uniform sampling with 1.8 to more than 4 times fewer queries and its closed-loop cost gap with 1.4 to 4.7 times fewer, the factor computed beforehand predicts where the cost weighting pays, and occupation weighting, the on-policy state collection of DAgger-style methods, fails for a reason the theory names.

2 Problem Statement

The task is to learn the optimal feedback law of a constrained optimal control problem from an expert that is queried at states chosen by the learner, under a budget of N queries. We state the control problem, the input grid that turns its feedback law into a classifier and defines the expert, and the three errors by which the learned law is evaluated.

Constrained optimal control problem. We consider control-affine dynamics on a compact state space $\mathcal{X} \subset \mathbb{R}^n$ with a scalar input $u \in \mathcal{U} = [-u_{\max}, u_{\max}]$,

$$\dot{x} = f(x) + g_u(x)\, u, \tag{1}$$

with f and g_u Lipschitz continuous, and the discounted infinite-horizon cost of a policy $\pi : \mathcal{X} \to \mathcal{U}$,

$$J^\pi(x_0) = \int_0^\infty e^{-\rho t}\, \ell\big(x(t), \pi(x(t))\big)\, dt, \qquad \ell(x,u) = q(x) + w_u\, u^2, \tag{2}$$

with discount rate $\rho > 0$, input weight $w_u > 0$ and a continuous state penalty $q \geq 0$. The trajectories stay in $\mathcal{X}$ by construction of the systems considered, through periodic angles and reflecting walls (Appendix D). The value function $V^*(x) = \inf_\pi J^\pi(x)$ satisfies the Hamilton–Jacobi–Bellman equation $\rho V^* = \min_u\{\ell + \nabla V^{*\mathsf{T}}(f + g_u u)\}$ (Bardi & Capuzzo-Dolcetta, 1997; Fleming & Soner, 2006), whose minimiser is the saturated feedback $u^*(x) = \mathrm{sat}_{u_{\max}}(-\sigma(x)/(2w_u))$ with the switching function $\sigma = g_u^{\mathsf{T}} \nabla V^*$ of the maximum principle (Pontryagin et al., 1962; Liberzon, 2012). Under input constraints V^* is in general not differentiable everywhere and solves the equation only in the viscosity sense (Bardi & Capuzzo-Dolcetta, 1997). Its gradient jumps across surfaces in the state space, and the feedback is continuous wherever ∇V^* is and jumps with it, by up to the full input range, on the switching set of the continuous problem.

Input grid and expert. Fix a step $h > 0$, a one-step integrator $\Phi_h(x,u)$ of equation 1 with the input held constant, the exact discount $\gamma_h = e^{-\rho h}$, and the uniform input grid $\mathcal{A}_M = \{-u_{\max} + (j-1)\Delta u\}_{j=1}^M$ with $\Delta u = 2u_{\max}/(M-1)$. The semi-Lagrangian Bellman operator (Dolcetta & Falcone, 1989; Falcone & Ferretti, 2013)

$$(\mathcal{T}_{\mathcal{A}} V)(x) := \min_{u \in \mathcal{A}_M} \big\{\, h\, \ell(x,u) + \gamma_h\, V\big(\Phi_h(x,u)\big) \,\big\} \tag{3}$$

is the dynamic-programming principle over one step with the input held constant on the grid. It is a monotone γ_h-contraction with a unique fixed point $V^*_{h,\mathcal{A}}$ (Bertsekas, 2012), which converges to V^* as $h \downarrow 0$ and $\Delta u \downarrow 0$ (Dolcetta & Falcone, 1989; Chow & Tsitsiklis, 1991). We write V^* for this fixed point from here on, and

$$\begin{aligned} Q^*(x,u) &= h\,\ell(x,u) + \gamma_h V^*(\Phi_h(x,u)), \qquad A^*(x,u) = Q^*(x,u) - V^*(x) \geq 0, \\ \pi^*(x) &\in \arg\min_{u \in \mathcal{A}_M} Q^*(x,u) \end{aligned} \tag{4}$$

for the action values, the advantage and the optimal policy over the grid. The advantage $A^*(x,u)$ is the excess cost of input u over the optimal one at x. The policy π^* is piecewise constant with M levels. Its switching set contains the surfaces of the continuous problem, across which π^* jumps by several levels, and the surfaces between neighbouring levels that the input grid adds (Lemma 2).

On a state grid with multilinear interpolation the operator equation 3 is the classical semi-Lagrangian scheme and its fixed point is computable exactly, sweep by sweep. That table is our reference. The expert returns, for a queried state x, the row $Q^*(x,\cdot)$, and hence the label $\pi^*(x)$ and the advantage row $A^*(x,\cdot)$ with no further cost.

Classifier and error measures. A policy classifier $\pi_\phi : \mathcal{X} \to \mathcal{A}_M$ is the argmax of M logits $\phi(x)$ computed by a network on a fixed embedding of the state, with softmax probabilities $p_u(x)$ over the levels. It is learned from N expert queries at states chosen by the learner and judged by three errors under the uniform distribution ν on $\mathcal{X}$. The zero-one error $e_{01}(\pi_\phi) = \nu(\{\pi_\phi \neq \pi^*\})$, the share of the state space on which the classifier disagrees with π^*, is the quantity of active-learning theory, and its complement is the accuracy of the classifier, its agreement with π^*. The cost-sensitive loss $L(\pi_\phi) = \mathbb{E}_{x\sim\nu}[A^*(x,\pi_\phi(x))]$ weighs each error by the excess cost of the wrong input. The closed-loop cost gap $R(\pi_\phi) = \mathbb{E}_{x_0\sim\nu}[J^{\pi_\phi}(x_0) - J^{\pi^*}(x_0)]$, with J^π from here on the discrete-time closed-loop cost $\sum_k \gamma_h^k\, h\,\ell(x_k,\pi(x_k))$ along $x_{k+1} = \Phi_h(x_k,\pi(x_k))$, is the loss that matters once the law is deployed.

3 Active Learning on the Switching Set

Switching set and margin. The switching set of a policy π over $\mathcal{A}_M$ is $\Sigma(\pi) = \bigcup_u \partial\{x : \pi(x) = u\} \cap \mathcal{X}$. For π^* these are the surfaces of Section 2. For the classifier it is the set on which its two largest logits tie, and the top-two logit margin $m_\phi(x)$ vanishes on it.

Batch-mode active learning. Algorithm 1 (Appendix D) learns π_ϕ in iterations. Each iteration draws B new query states from a density q built from the current classifier, obtains their advantage rows from the expert, and refits the classifier on all labels so far by minibatch gradient steps. The sampler itself never calls the expert, it uses the classifier and the rows already returned. The density is the mixture $q = (1-\alpha)\, q_{\mathrm{sw}} + \alpha\, \nu$ of a density q_{sw} on the low-margin band of the current classifier, defined next, and the uniform density, with a coverage share $\alpha \in (0,1)$, the defensive mixture of importance sampling (Hesterberg, 1995). The coverage share keeps labelling the cell interiors and finds surfaces the classifier does not yet represent.

Sampling on the switching set. The band is the lowest τ-quantile of the classifier's margin among uniform candidates, the batch version of margin-based active learning (Balcan et al., 2007). The sampler draws from the band with probability proportional to $\hat{g}(x)^\beta$, where $\hat{g}(x)$ is the gradient norm of a linear regression of the advantage difference $A^*(\cdot,b) - A^*(\cdot,a)$ of the classifier's two best inputs at x over the eight nearest labelled states. It uses the rows the expert has already returned, makes no further query, and estimates the cost slope of Lemma 2 by which Theorem 3 allocates. Three ablations replace the weight, *margin* ($\beta = 0$), *jump*, the expected input jump under the classifier's softmax $w(x) = \sum_u p_u(x)\,|u - \hat{u}(x)|$ with $\hat{u}(x)$ the argmax, which needs no labels, and *switchmap*, a histogram of the classifier's switches on a grid (Appendix D).

Cost-sensitive loss. Two losses on the pooled labels $y_i = \pi^*(x_i)$ are used,

$$\mathcal{L}_{\mathrm{ce}}(\phi) = \tfrac{1}{N}\sum_i -\log p_{y_i}(x_i), \qquad \mathcal{L}_{\mathrm{adv}}(\phi) = \tfrac{1}{N}\sum_i \sum_u p_u(x_i)\, A^*(x_i,u) + \eta\, \mathcal{L}_{\mathrm{ce}}(\phi), \tag{5}$$

cross-entropy on the label, and the cost-sensitive loss, the expected advantage of the classifier's input under its softmax, with A^* scaled to unit mean largest advantage over the pool and a small cross-entropy term ($\eta = 0.1$) that keeps the gradient alive at confident errors. The second is the classification loss of classification-based policy iteration (Lazaric et al., 2016) and of AggreVaTe (Ross & Bagnell, 2014), which estimate the advantage by rollouts where the expert here returns it exactly. The first needs the label alone, the second the advantage row, and both are reported.

No importance weights. With a deterministic label function the population minimiser of either loss is the same under every query density with full support, so the unweighted fit on the pooled labels is consistent and the weights of importance-weighted active learning (Beygelzimer et al., 2009) are not needed (Lemma 7 in Appendix A, the coverage share α enters only as the factor $1/\alpha$ between the losses under q and under ν). This is the structural difference to regression of the value function, whose fitted fixed point moves with the sampling density unless importance weights hold it in place.

4 Label Complexity and Allocation

Section 3 placed the queries on the band around the classifier's switching set and weighted them by an estimate of the cost slope. This section says why, for an idealised learner that locates each surface by bisection, in the tradition of the minimax analysis of active learning on boundary-fragment classes (Willett et al., 2005; Castro & Nowak, 2008). The rates in Theorem 3(i), (ii) and (iv) are assembled from that theory and from the nonparametric cost-sensitive rates of Njike & Siebert (2023), new are the decomposition of the loss over the surfaces with their slopes, the allocation across them, the gap of margin sampling and the closed-loop weighting. Proofs are in Appendix A.

Assumptions. Let λ be Lebesgue measure on $\mathcal{X}$ and $\Sigma = \bigcup_{i \le m} \Sigma_i$ the switching set of π^*.

- (A1) (*Boundary fragments*, Korostelev & Tsybakov, 1993; Mammen & Tsybakov, 1999.) For each i there are orthonormal coordinates $x = (z, y)$, $z \in \mathbb{R}^{n-1}$, a compact base $D_i \subset \mathbb{R}^{n-1}$ with $|D_i| > 0$ and an $f_i : D_i \to \mathbb{R}$, Lipschitz continuous with constant L, with $\Sigma_i = \{(z, f_i(z))\}$. The tubes $T_i = \{(z, y) : z \in D_i,\ |y - f_i(z)| < \delta_0\}$ are disjoint, and $\pi^* = a_i$ below and b_i above Σ_i in T_i.
- (A2) (*Linear cost growth*, the noise-free analogue of the margin condition of Tsybakov, 2004.) There are $0 < g_i^- \le g_i^+$ with $g_i^- |y - f_i(z)| \le A^*((z, y), \bar{\pi}) \le g_i^+ |y - f_i(z)|$ on T_i, where $\bar{\pi}$ is the wrong one of $\{a_i, b_i\}$, and $A^*(x, c) \ge \Delta_0 > 0$ on T_i for $c \notin \{a_i, b_i\}$.
- (A3) (*Exact expert.*) A query at x returns $\pi^*(x)$ and $A^*(x, \cdot)$.
- (A4) (*Learner class.*) $\hat{\pi} = a_i$ below and b_i above an estimate $\hat{f}_i$ on T_i, and $\hat{\pi} = \pi^*$ outside the tubes.

Set $e_i = \hat{f}_i - f_i$, with norms taken over D_i. Junctions of surfaces are $(n-2)$-dimensional and contribute at lower order to every statement below.

Proposition 1 (The cost-sensitive loss is a cost-weighted squared boundary error)**.** *Under (A1), (A2), (A4) and $\|e_i\|_\infty \le \delta_0$,*

$$\tfrac{1}{2} \sum_i g_i^- \, \|e_i\|_2^2 \;\le\; \int_{\mathcal{X}} A^*(x, \hat{\pi}(x))\, dx \;\le\; \tfrac{1}{2} \sum_i g_i^+ \, \|e_i\|_2^2, \qquad \lambda(\{\hat{\pi} \ne \pi^*\}) = \sum_i \|e_i\|_1.$$

The coordinate y is the vertical of the fragment, not the normal of the surface, and (A2) is stated in it. Its two constants bound the one-sided slopes of Lemma 2 from below and above over the tube. The zero-one error is the L^1 error of the surfaces and the cost-sensitive loss their cost-weighted L^2 error (Mammen & Tsybakov, 1999), hence the square in the rates.

Lemma 2 (Cost slope)**.** *In this lemma V^* denotes the continuous-time value function inserted into the one-step operator equation 3, $g_u \in C^1$, and the explicit Euler step is used, a Runge–Kutta step changes only the $O(h^2)$ terms. Let Σ_i be a surface of π^* between the inputs $a_i < b_i$ and $\sigma = g_u^{\mathsf{T}} \nabla V^*$ the switching function.*

(i) If V^* is C^2 on either side of Σ_i and C^1 across it, then near Σ_i

$$Q^*(x, b_i) - Q^*(x, a_i) = h\,(b_i - a_i)\big[w_u(a_i + b_i) + \gamma_h\, \sigma(x)\big] + O(h^2),$$

so Σ_i lies within an $O(h)$ neighbourhood of the level set $\{\sigma = -w_u(a_i + b_i)\}$ wherever $\nabla\sigma \neq 0$ on it, and the one-sided slopes are $g_i^{\pm} = h\,\gamma_h\,|b_i - a_i|\,|\nabla\sigma^{\pm}| + o(h)$, with $\nabla\sigma^{\pm}$ the limits of $\nabla\sigma$ from the two sides.

(ii) If instead ∇V^* jumps across Σ_i by $\kappa\, n$ with n the unit normal, as on the switching set of the continuous problem, then within the $O(h)$ band around Σ_i, $Q^*(x, b_i) - Q^*(x, a_i) = \gamma_h\,\kappa\, n^{\mathsf{T}}(x - x_\Sigma) + O(h)$ with x_Σ the foot of x on Σ_i, and the slope is $\gamma_h|\kappa| + O(h)$, of order one.

In both cases (A2) holds on a tube around Σ_i, with g_i^- and g_i^+ bounding the one-sided slopes, taken in the fragment coordinate y, from below and above over the tube.

On the grid surfaces the slope is the input jump, which the sampler reads off the policy, times the gradient of the switching function, Δu between neighbouring levels and $2u_{\max}$ across a reversal, a ratio of $M - 1$. On the switching set of the continuous problem the slope is of order one, larger still. Both regimes favour the jump weighting, and Table 2 shows that the jump tracks the slope measured on the references by finite differences across neighbouring nodes (Appendix D).

Theorem 3 (Rates and optimal allocation)**.** *Let $s = 2/(n-1)$, and let $c, c' > 0$ be constants that depend on n, L, δ_0 and the g_i but not on N.*

(i) Any learner from N labels at independent uniform states satisfies $\sup \mathbb{E}\, L(\hat\pi) \geq c\, N^{-2/n}$ over the boundary-fragment class of (A1) and (A2).

(ii) The bisection learner with budget N_i on Σ_i achieves $\|e_i\|_\infty \leq \varepsilon_i$ with

$$N_i = |D_i|\Big(\frac{L\sqrt{n-1}}{\varepsilon_i}\Big)^{n-1} k_i, \qquad k_i = \lceil \log_2(2\delta_0/\varepsilon_i)\rceil,$$

$$\textit{hence} \quad L(\hat\pi) \leq \sum_i c_i N_i^{-s}, \qquad c_i = \tfrac{1}{2} g_i^+ (L\sqrt{n-1})^2 k_i^s |D_i|^{1+s}.$$

(iii) With $k_i \equiv k$, the bound in (ii) is minimised over $\sum_i N_i = N$ by

$$N_i^\star = N\,\frac{|D_i|\,(g_i^+)^{1/(1+s)}}{|\Sigma|_g}, \qquad |\Sigma|_g := \sum_i |D_i|\,(g_i^+)^{1/(1+s)},$$

$$L^\star(N) = \tfrac{1}{2}(L\sqrt{n-1})^2\, k^{\,s}\, |\Sigma|_g^{1+s}\, N^{-s},$$

a label density per unit base area proportional to $g^{(n-1)/(n+1)}$.

(iv) Any active learner satisfies $\sup \mathbb{E}\, L(\hat\pi) \geq c' N^{-2/(n-1)}$ up to logarithms over the same class, so (iii) is rate-optimal over it.

The lower bounds are inherited from the boundary-fragment class, of which the switching sets considered are members by (A1), so rate-optimality is claimed over that class and not over the class of control problems. The allocation in (iii) is the Neyman allocation of stratified sampling (Neyman, 1934) and the equidistribution principle of adaptive meshes (de Boor, 1973), with the cost slope as the local error constant. To reach a loss ϵ the active learner needs $N \asymp |\Sigma|_g^{(1+s)/s}\,\epsilon^{-(n-1)/2}$ labels, the uniform learner $N \asymp \lambda(\mathcal{X})\,\epsilon^{-n/2}$, so a system with a small and expensive switching set gains a factor that grows like $\epsilon^{-1/2}$, a system whose switching set fills the box gains a constant.

Corollary 4 (Hölder gap of margin and linear-jump sampling)**.** *Let L_{mar} and L_{lin} be the bound of Theorem 3(ii) under the margin allocation $N_i \propto |D_i|$ and the linear-jump allocation $N_i \propto |D_i| g_i$, and $L^\star$ that of (iii). Then*

$$\frac{L_{\mathrm{mar}}}{L^\star} = \frac{\big(\sum_i g_i|D_i|\big)\big(\sum_i |D_i|\big)^s}{\big(\sum_i |D_i|\, g_i^{1/(1+s)}\big)^{1+s}} \geq 1, \qquad \frac{L_{\mathrm{lin}}}{L^\star} = \frac{\big(\sum_i g_i^{1-s}|D_i|\big)\big(\sum_i g_i|D_i|\big)^s}{\big(\sum_i |D_i|\, g_i^{1/(1+s)}\big)^{1+s}} \geq 1,$$

with equality if and only if g_i is constant over the components.

Margin sampling is optimal only when every surface costs the same. With $M = 21$ levels most of the switching set separates neighbouring levels, whose slope is twenty times smaller than at a reversal, so the gap is large wherever such surfaces dominate, and it can be computed from the reference before any learning run, which Section 5 does. When a component is starved of labels its error is capped by the tube width, and in that scarce regime the optimal exponent moves from $1/(1+s)$ toward one (Appendix B).

Corollary 5 (The exponent depends on the learner's local rate)**.** *For a learner whose local precision scales as $\varepsilon \propto q^{-1/a}$ in the label density q per unit base area, the allocation optimum is $q^\star \propto g^{a/(a+2)}$. Bisection has $a = n-1$ and the exponent $(n-1)/(n+1)$, a learner that fills a band of fixed width around the surface has $a = n$ and the exponent $n/(n+2)$, and its loss decays as $N^{-2/n}$ with the band-to-box volume ratio as the prefactor.*

The local rate of a plug-in or nearest-neighbour classifier under a margin condition is the standard one (Audibert & Tsybakov, 2007; Chaudhuri & Dasgupta, 2014). The practical sampler fills a band, so its exponent is $1/2$ in two dimensions and its gain over uniform labelling is a constant, the volume ratio, while bisection gains an order in the rate. Appendix E identifies the exponent for a local learner and finds $1/2$, approached from above as the budget grows.

Theorem 6 (Closed-loop cost gap)**.** *For any stationary policy π with bounded closed-loop cost, $J^\pi(x_0) - J^{\pi^*}(x_0) = \sum_{k\geq 0} \gamma_h^k A^*(x_k, \pi(x_k))$ along the closed loop of π (Kakade & Langford, 2002), hence with the discounted occupation measure d_ν^π of the closed loop from $x_0 \sim \nu$,*

$$R(\pi) = \frac{1}{1-\gamma_h} \int A^*(x, \pi(x))\, d_\nu^\pi(dx) \leq \frac{C_\pi}{1-\gamma_h}\, L(\pi), \qquad C_\pi = \left\| \frac{d\, d_\nu^\pi}{d\nu} \right\|_\infty,$$

the concentrability coefficient of approximate dynamic programming (Munos, 2007).

The bound explains why the cost-sensitive loss and the closed-loop cost gap can rank two learners differently. It is conditional on C_π, which can be large, so a smaller cost-sensitive loss need not translate into a proportionally smaller closed-loop cost gap, and the pendulum, where the closed-loop gain holds in the median and not in the mean, is that case. If the occupation density of the optimal closed loop is continuous on the tubes, the bound becomes an identity that weights the error of each surface by that density (Proposition 8), which suggests weighting the band by the closed-loop occupation. Section 5 tests that suggestion and refutes it, because the continuity assumption fails at the budgets of interest.

5 Experiments

Four configurations with an exact expert, the converged semi-Lagrangian table of Section 2 with $M = 21$ input levels. The *inverted pendulum* swing-up, whose switching set fills the state box, the *continuous mountain car* in two configurations, $w_u = 0.001$, which is close to bang-bang, and $w_u = 0.1$, whose optimal policy uses many intermediate levels, and the four-dimensional *cart-pole* swing-up with a reference on a coarse $21 \times 21 \times 25 \times 21$ grid. The classifier is a three-layer $\tanh$ network with 64 units, refitted after every batch of $B = 32$ queries, ten seeds, budgets up to 2016 queries (992 on the cart-pole), every sampler under the cost-sensitive loss unless stated. Metrics are the agreement with π^* and the cost-sensitive loss on the reference grid and the closed-loop cost gap R from 64 fixed uniform starts. Differences are tested at pre-specified endpoints with one-sided rank tests over seeds, corrected for the number of endpoints per system, and label complexities carry bootstrap intervals (dynamics, costs, hyperparameters and statistics in Appendix D).

Fewer expert queries for the same accuracy and cost gap. Figure 1 and Table 1 give the main result. From about 320 queries on, the switching-set samplers exceed uniform labelling in agreement on every configuration at every budget. Below 200 queries the band is placed by a still poor classifier and they are not better, the usual warm-up of active learning. The label-complexity ratio of Table 1 is the number of queries uniform labelling needs to reach the agreement, or the cost gap, that it attains at the end of the budget, divided by the number the sampler needs for the same level. The slope-estimate sampler reaches the agreement of uniform labelling with 1.8 to more than 4 times fewer queries and its closed-loop cost gap with 1.4 to 4.7 times fewer, and it improves both endpoints on every configuration (corrected $p \leq 0.004$ for the agreement at both endpoints, $p \leq 0.04$ for the cost gap at the final budget). The pendulum, whose switching set fills the box, gains a constant factor, the mountain cars, whose switching sets are small, gain a factor that grows with the target, as Theorem 3 predicts.

The gap statistic predicts where cost weighting pays. Table 2 gives, for each reference and before any learning run, the size of its switching set, the share of surfaces between neighbouring

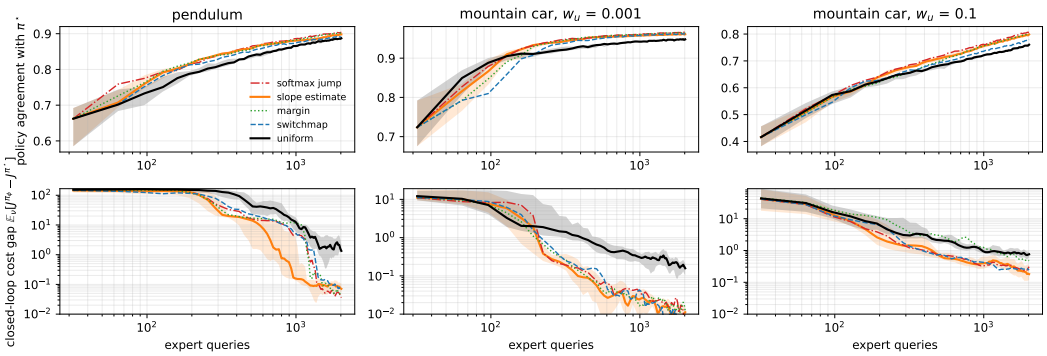

Figure 1: Agreement with π^* (top) and closed-loop cost gap R (bottom) against the number of expert queries on the three two-dimensional configurations, median over ten seeds, interquartile band for uniform labelling and the slope-estimate sampler.

Table 1: Endpoints at the final budget, ten seeds, cost-sensitive loss. Mean agreement and closed-loop cost gap R, and label-complexity ratios $N^\star_{\text{uniform}}/N^\star_{\text{sampler}}$ (median over seeds, 95 percent bootstrap interval) at the level uniform labelling reaches at the end of the budget, or the lower end of the interval where uniform labelling reaches that level in fewer than half the seeds.

system	sampler	agreement	cost gap R	ratio agreement	ratio cost gap
pendulum	uniform	0.89	4.9		
1856 queries	slope estimate	0.90	0.14	1.8 [1.5, 2.0]	1.4 [0.9, 2.3]
	jump	0.90	3.5	2.0 [1.6, 2.4]	1.0 [0.7, 2.0]
	margin	0.90	3.3		
mountain car	uniform	0.95	0.19		
$w_u = 0.001$	slope estimate	0.96	0.02	> 4.0	4.7 [2.3, 5.4]
1856 queries	jump	0.97	0.04	> 3.7	4.2 [2.0, 6.3]
	margin	0.96	0.02		
mountain car	uniform	0.75	0.89		
$w_u = 0.1$	slope estimate	0.80	0.39	1.9 [1.6, 2.1]	2.9 [1.7, 3.3]
1856 queries	jump	0.80	0.26	2.1 [1.7, 2.4]	2.5 [1.3, 3.0]
	margin	0.79	0.88		
cart-pole (4D)	uniform	0.78	9.96		
992 queries	slope estimate	0.80	3.93	1.9 [1.4, 2.3]	2.1 [1.2, 2.8]
	jump ($\beta = 1$)	0.81	3.89	1.9 [1.3, 2.2]	1.6 [1.0, 2.1]
	margin	0.80	4.96		

levels, and the Hölder gap of Corollary 4 in its capped form. The mountain car with $w_u = 0.1$ has a switching set of which 89 percent separates neighbouring levels and a gap of 12. On it margin sampling does not beat uniform labelling in the closed-loop cost gap at any budget (Figure 1, Table 1) and loses to it in the cost-sensitive loss, exactly the failure Corollary 4 describes, while the two cost-weighted samplers win all three metrics. On the pendulum the gap is 4.6 and margin and jump weighting are within noise of one another on the cost-sensitive loss. On the near bang-bang car the samplers coincide although the uncapped gap is large, its capped version at the budgets used is close to one (Appendix B).

The gain transfers to four dimensions. On the cart-pole the classifier, the sampler and the accounting are unchanged, and since the coarse reference's own closed loop does not settle from the fixed starts within 300 steps, the cost gap is measured against the expert's closed loop. The gap statistic of the reference is 1.5 (Table 2), which predicts a constant-factor gain from the switching

Table 2: Switching-set statistics of the four references, computed before any learning run. Share of neighbouring node pairs on which π^* switches, share of those between neighbouring levels, correlation of cost slope and input jump across the crossings, and the Hölder gap of margin sampling at saturation.

system	crossings	adjacent levels	corr($\log g$, $\log \lvert\Delta u\rvert$)	gap of margin
pendulum	10.5%	20%	0.62	4.6
mountain car $w_u = 0.001$	4.8%	20%	0.71	13.9
mountain car $w_u = 0.1$	17.7%	89%	0.65	11.8
cart-pole (4D)	14.0%	13%	0.66	1.5

set and little from cost weighting, and both hold (Table 1). The slope-estimate sampler reaches the agreement of uniform labelling with 1.9 and its cost gap with 2.1 times fewer queries (corrected $p \leq 0.013$ at 992 queries), and margin, jump and slope weighting lie within 0.03 of one another on the cost-sensitive loss.

The slope estimate beats its softmax proxy. The band is weighted by the true slope from the table (extra expert rows, not counted), by the slope estimate from the labelled rows, and by the softmax jump, at $\beta \in \{1/2, 1\}$ with ten seeds. The rank correlation between the weight and the true slope on the band lies between -0.07 and 0.32 for the softmax jump and between 0.28 and 0.64 for the estimate over the four systems, so the softmax jump is not an estimate of the slope on most of them. At $\beta = 1/2$ and 992 queries the estimate halves the cost-sensitive loss of the proxy on every two-dimensional configuration, matches or improves the closed-loop cost gap and keeps the agreement within 0.004 (Appendix D). The true slope lowers the cost-sensitive loss further and the agreement with it, the trade-off of Proposition 1, and $\beta = 1$ over-concentrates for both. The slope estimate at $\beta = 1/2$, the band-filling exponent of Corollary 5, is therefore the sampler of this paper and the softmax jump its ablation.

The jump exponent trades accuracy against cost. Figure 2 (Appendix E) sweeps the exponent β of the softmax-jump sampler over $\{0, 1/3, 1/2, 1, 2\}$, $\beta = 0$ being margin sampling. On the wide-band mountain car every $\beta > 0$ beats margin sampling, the cost-sensitive loss keeps improving up to $\beta = 2$ while the agreement peaks at $\beta \in [1/3, 1]$, the trade-off of Proposition 1, since a large exponent starves the cheap surfaces whose errors count in the zero-one loss and cost nothing in the cost-sensitive loss. On the near bang-bang car every $\beta > 0$ improves on margin at the largest budget, on the pendulum nothing separates, as the gaps predict. Appendix E identifies the theorem's exponent with the true slope and a local learner and finds one half.

The cost-sensitive loss is a second gain, and only the slope estimate needs the row. Under uniform labelling the cost-sensitive loss lowers the pendulum's closed-loop cost gap at 992 queries from 70 to 12 at no extra expert cost, which is why every comparison is made under the same loss. With cross-entropy, which needs the label alone, the margin, jump and switchmap samplers still win agreement on all three two-dimensional configurations and the closed-loop cost gap on both mountain cars, so switching-set sampling does not depend on the expert returning the row, only the slope estimate does.

Robust to coverage share, band quantile and capacity. The agreement gain is unchanged for $\alpha \in [0.03, 0.6]$ and $\tau \in [0.05, 0.4]$, the closed-loop gain needs $\alpha \geq 0.1$ and $\tau \geq 0.1$, and a 128-unit classifier keeps the ratios on the pendulum and grows them on the wide-band car (Appendix D).

Occupation weighting fails, as Theorem 6 predicts. Theorem 6 suggests, under a continuity assumption on the occupation (Proposition 8), weighting the band by the closed-loop occupation of the current policy, the on-policy state collection of DAgger and its uncertainty-gated variants restricted to the band. Weighting the band by a histogram of the states the current classifier's closed loop visits from 64 uniform starts, with exponents $\{0, 1/3, 1\}$, worsens every metric at 512 queries and above on both systems tested (closed-loop cost gap at 992 queries 11.4 against 5.7 without it on the pendulum, 3.4 against 0.6 on the wide-band car). The occupation of the current classifier

concentrates where that policy goes and starves the outer switching set that trajectories cross briefly but must cross correctly, so the continuity assumption fails. Theorem 6 explains why the two losses rank learners differently, it is not a recipe, and the sampler weights by ν.

6 Related Work

Active learning. Margin and uncertainty sampling query near the decision boundary of the current classifier or a level set (Tong & Koller, 2001; Settles, 2009; Balcan et al., 2007; Gotovos et al., 2013), and Willett et al. (2005); Castro & Nowak (2008) give minimax rates for boundary-fragment classes with the bisection learner. Kpotufe et al. (2022) show that such gains require a unique Bayes classifier, which deterministic expert labels provide. Cost-sensitive active learning weighs queries by misclassification cost (Margineantu, 2005; Krishnamurthy et al., 2017), with nonparametric rates governed by the mass near the decision boundary (Njike & Siebert, 2023). These results bound the label complexity of a single boundary with a homogeneous cost. Our decomposition into surfaces with heterogeneous slopes turns the query problem into an allocation across them, with the slope as the local error constant.

Active imitation learning. Judah et al. (2014) reduce imitation learning to i.i.d. active learning and query near the expert's decision boundary. DAgger and its gated variants (Ross et al., 2011; Chernova & Veloso, 2009; Zhang & Cho, 2017; Menda et al., 2019; Hoque et al., 2021) collect states along the learner's own trajectories and gate the expert query by confidence or a learned risk, and AggreVaTe (Ross & Bagnell, 2014) trains the policy by cost-sensitive classification against the expert's cost-to-go. These methods place queries along trajectories or by uncertainty, not by the geometry and cost of the decision boundary, and Section 5 refutes occupation weighting, a caution for trajectory-based gating while the policy is still poor.

Classification-based policy iteration and explicit MPC. Lagoudakis & Parr (2003) represent the policy as a classifier, Lazaric et al. (2016) analyse the cost-sensitive loss, and Dimitrakakis & Lagoudakis (2008) allocate rollouts to states whose best action is uncertain, all under a fixed state distribution. Learned explicit MPC samples the state space randomly (Hertneck et al., 2018) or from closed-loop data (Karg & Lucia, 2020). We supply the sampling rule these methods leave open.

Dynamic programming. Munos & Moore (2002) refine grid cells where the policy disagrees, the switching-set idea for a mesh, and the action gap (Singh & Yee, 1994; Farahmand, 2011) controls how value error becomes policy error, of which Proposition 1 is the classification counterpart. Adaptive collocation for value functions (Lu et al., 2021; Wu et al., 2023; Nakamura-Zimmerer et al., 2021; Moore & Atkeson, 1993; Schaul et al., 2016) is the value-side analogue (Appendix C).

7 Conclusion

The switching set of a constrained optimal control problem is where the labels of its explicit feedback law are worth an expert query, and the cost of a wrong decision on that set is not uniform. Weighting the queries by the cost slope, estimated from the action values the expert has already returned, follows from a label-complexity analysis whose constants can be computed from the expert before learning, and which predicts, on four benchmark configurations, where the weighting pays, where margin sampling suffices and where nothing separates. *Limitations.* The theorem is for an idealised bisection learner and the method is its batch instantiation, a bisection along the classifier's normals is not implemented. The evidence is three two-dimensional configurations and one four-dimensional system, all with a table as the expert, and the allocation exponent is identified for a local learner with an expert-informed allocation, not for the network with the practical sampler, which prefers a flatter allocation. Above two dimensions the band's volume fraction shrinks, candidate generation grows even though it never calls the expert, and the classifier's capacity floor appears earlier. The closed-loop statement is conditional on a concentrability constant that can be large. An optimisation-based expert returns the label with one solve but the advantage row, which the slope estimate needs, only with more, and it adds label noise that the analysis does not cover.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, in the form of parts of the experiment code, and to assist in the writing of proofs. We have not used generative AI tools to generate synthetic data sets, to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide critical ingredients for proving mathematical claims, to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, or to interpret results, and translation, dataset cleaning and reformatting, and qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability. We have reviewed all AI-assisted work. The authors reviewed, tested and executed all code, so that every number in the paper comes from runs executed by the authors, checked every step of every proof against the statements of the main text, and read every edited passage for meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The assumptions (A1)–(A4) and complete proofs of every statement of Section 4 are given in Appendix A, the capped allocation of the scarce regime in Appendix B. Appendix D specifies every system with its dynamics, stage cost, discount, integrator, state box and reference grid, the semi-Lagrangian reference and the expert, the state embedding, classifier architecture, optimiser and training schedule, every sampler with its hyperparameters and defaults, the seeds, the ablation grids, the pre-specified endpoints with their targets and statistics, and the computation of the switching-set statistics of Table 2. Appendix E describes the exponent identification. Section 5 states the setup shared by all runs, the systems, classifier, batch size, seeds, budgets, loss and metrics. Together these specify each experiment completely, so that every table and figure of the paper can be reproduced from the information given.

REFERENCES

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001217.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In Nader H. Bshouty and Claudio Gentile (eds.), *Learning Theory (COLT)*, pp. 35–50. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-72927-3.

Martino Bardi and Italo Capuzzo-Dolcetta. *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*. Birkhäuser, Boston, 1997. doi: 10.1007/978-0-8176-4755-1.

Alberto Bemporad, Manfred Morari, Vivek Dua, and Efstratios N. Pistikopoulos. The explicit linear quadratic regulator for constrained systems. *Automatica*, 38(1):3–20, 2002. doi: https://doi.org/10.1016/S0005-1098(01)00174-1.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. II: Approximate Dynamic Programming*. Athena Scientific, 4 edition, 2012. ISBN 978-1-88652-908-3.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 49–56, 2009. ISBN 9781605585161. doi: 10.1145/1553374.1553381.

Rui M. Castro and Robert D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008. doi: 10.1109/TIT.2008.920189.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, pp. 3437–3445. MIT Press, 2014.

Sonia Chernova and Manuela Veloso. Interactive policy learning through confidence-based autonomy. *Journal of Artificial Intelligence Research*, 34:1–25, 2009. ISSN 1076-9757. doi: 10.1613/jair.2584.

C.-S. Chow and J.N. Tsitsiklis. An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE Transactions on Automatic Control*, 36(8):898–914, 1991. doi: 10.1109/9.133184.

Carl de Boor. Good approximation by splines with variable knots. In *Spline Functions and Approximation Theory*, volume 21 of *International Series of Numerical Mathematics*, pp. 57–72. Birkhäuser, 1973. ISBN 978-3-0348-5979-0. doi: 10.1007/978-3-0348-5979-0_3.

Christos Dimitrakakis and Michail G. Lagoudakis. Rollout sampling approximate policy iteration. *Machine Learning*, 72(3):157—-171, 2008. ISSN 1573-0565. doi: 10.1007/s10994-008-5069-3.

I. Capuzzo Dolcetta and M. Falcone. Discrete dynamic programming and viscosity solutions of the bellman equation. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 6:161–183, 1989. ISSN 0294-1449. doi: https://doi.org/10.1016/S0294-1449(17)30020-3.

Maurizio Falcone and Roberto Ferretti. *Semi-Lagrangian Approximation Schemes for Linear and Hamilton–Jacobi Equations*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013. doi: 10.1137/1.9781611973051.

Amir-massoud Farahmand. Action-gap phenomenon in reinforcement learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Wendell H. Fleming and Halil Mete Soner. *Controlled Markov Processes and Viscosity Solutions*. Springer, 2nd edition, 2006. doi: 10.1007/0-387-31071-1.

Geoffrey J. Gordon. Stable function approximation in dynamic programming. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, pp. 261–268, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603778.

Alkis Gotovos, Nathalie Casati, Gregory Hitz, and Andreas Krause. Active learning for level set estimation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pp. 1344—1350. AAAI Press, 2013. ISBN 9781577356332.

Michael Hertneck, Johannes Köhler, Sebastian Trimpe, and Frank Allgöwer. Learning an approximate model predictive controller with guarantees. *IEEE Control Systems Letters*, 2(3):543–548, 2018. doi: 10.1109/LCSYS.2018.2843682.

Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995. doi: https://doi.org/10.2307/1269620.

Ryan Hoque, Ashwin Balakrishna, Ellen Novoseller, Albert Wilcox, Daniel S. Brown, and Ken Goldberg. ThriftyDAgger: Budget-aware novelty and risk gating for interactive imitation learning. In *Conference on Robot Learning (CoRL)*, 2021.

Kshitij Judah, Alan P. Fern, Thomas G. Dietterich, and Prasad Tadepalli. Active imitation learning: Formal and practical reductions to I.I.D. learning. *Journal of Machine Learning Research*, 15: 4105–4143, 2014.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274. Morgan Kaufmann Publishers Inc., 2002. ISBN 1558608737.

Benjamin Karg and Sergio Lucia. Efficient representation and approximation of model predictive control laws via deep learning. *IEEE Transactions on Cybernetics*, 50(9):3866–3878, 2020. ISSN 2168-2275. doi: 10.1109/tcyb.2020.2999556.

Alexander P. Korostelev and Alexandre B. Tsybakov. *Minimax Theory of Image Reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer, 1993. doi: https://doi.org/10.1007/978-1-4612-2712-0.

Samory Kpotufe, Gan Yuan, and Yunfan Zhao. Nuances in margin conditions determine gains in active learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8112–8126. PMLR, 2022.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1915–1924, 2017.

Michail G. Lagoudakis and Ronald Parr. Reinforcement learning as classification: Leveraging modern classifiers. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pp. 424–431, 2003.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Analysis of classification-based policy iteration algorithms. *Journal of Machine Learning Research*, 17(19):1–30, 2016.

Daniel Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton University Press, 2012.

Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. DeepXDE: A deep learning library for solving differential equations. *SIAM Review*, 63(1):208–228, 2021. doi: 10.1137/19M1274067.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999. doi: 10.1214/aos/1017939240.

Dragos D. Margineantu. Active cost-sensitive learning. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 1622–1623, 2005.

Kunal Menda, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. Ensembledagger: A bayesian approach to safe imitation learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5041–5048, 2019. doi: 10.1109/IROS40897.2019.8968287.

Andrew W. Moore and Christopher G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13(1):103–130, October 1993. ISSN 1573-0565. doi: 10.1007/BF00993104.

Rémi Munos. Performance bounds in L_p-norm for approximate value iteration. *SIAM Journal on Control and Optimization*, 46(2):541–561, 2007. doi: 10.1137/040614384.

Rémi Munos and Andrew Moore. Variable resolution discretization in optimal control. *Machine Learning*, 49:291–323, 2002. doi: https://doi.org/10.1023/A:1017992615625.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008.

Tenavi Nakamura-Zimmerer, Qi Gong, and Wei Kang. Adaptive deep learning for high-dimensional hamilton–jacobi–bellman equations. *SIAM Journal on Scientific Computing*, 43(2): A1221–A1247, 2021. ISSN 1095-7197. doi: 10.1137/19m1288802.

Jerzy Neyman. On the two different aspects of the representative method : The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–606, 07 1934. ISSN 0952–8385. doi: 10.1111/j.2397-2335.1934.tb04184.x.

Boris Ndjia Njike and Xavier Siebert. Nonparametric active learning for cost-sensitive classification. *arXiv preprint arXiv:2310.00511*, 2023.

Lev S. Pontryagin, Vladimir G. Boltyanskii, Revaz V. Gamkrelidze, and Evgenii F. Mishchenko. *The Mathematical Theory of Optimal Processes*. Interscience Publishers, John Wiley & Sons Inc., 1962.

Stéphane Ross and J. Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 627–635. PMLR, 2011.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2016.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Satinder P. Singh and Richard C. Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994. doi: https://doi.org/10.1023/A:1022693225949.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2001. doi: 10.1162/153244302760185243.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi: 10.1109/9.580874.

Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. doi: 10.1214/aos/1079120131.

Rebecca Willett, Robert Nowak, and Rui Castro. Faster rates in regression via active learning. In Y. Weiss, B. Schölkopf, and J. Platt (eds.), *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

Chenxi Wu, Min Zhu, Qinyang Tan, Yadhu Kartha, and Lu Lu. A comprehensive study of non-adaptive and residual-based adaptive sampling for physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 403:115671, 2023. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2022.115671.

Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end simulated driving. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 2891–2897. AAAI Press, 2017.

A PROOFS

Lemma 7 (Density-free consistency)**.** *Let $q \geq \alpha\nu$ on $\mathcal{X}$ and $L_q(\pi) = \mathbb{E}_{x\sim q}[A^*(x, \pi(x))]$. Then for every policy π, $L_\nu(\pi) \leq \alpha^{-1} L_q(\pi)$, and $L_q(\pi) = 0$ if and only if $\pi(x) \in \arg\min_u Q^*(x, u)$ for ν-almost every x. The same holds for the zero-one loss.*

Proof of Lemma 7. $A^* \geq 0$ and $\nu \leq q/\alpha$ give the inequality by monotonicity of the integral. $L_q(\pi) = 0$ forces $A^*(x, \pi(x)) = 0$ for q-almost every x, and q and ν are mutually absolutely continuous. The zero-one case is identical. □

Proof of Proposition 1. Only the wrong input is chosen, and only between the true and the estimated surface, so $\int_{\mathcal{X}} A^*(x, \hat{\pi}(x))\, dx = \sum_i \int_{D_i} \int_{f_i(z)}^{\hat{f}_i(z)} A^*((z, y), \bar{\pi})\, dy\, dz$, and by (ii) the inner integral lies between $g_i^- e_i(z)^2/2$ and $g_i^+ e_i(z)^2/2$. If y is not the normal coordinate, vertical and normal distances differ by at most $(1 + L^2)^{1/2}$, which is absorbed into $g^\pm$. The zero-one identity is the same computation without the integrand. □

Proof of Lemma 2. (i). Write $v_u = f + g_u u$ for the Euler velocity. For x near Σ_i with both successors $x + hv_{a_i}$ and $x + hv_{b_i}$ on the same side, V^* is C^2 there and $V^*(x + hv_u) = V^*(x) + h\nabla V^*(x)^\mathsf{T} v_u + O(h^2)$, so the value terms of Q^* differ by $h\gamma_h(b_i - a_i)\, g_u^\mathsf{T} \nabla V^*(x) + O(h^2) = h\gamma_h(b_i - a_i)\sigma(x) + O(h^2)$ and the stage costs by $hw_u(b_i^2 - a_i^2) = hw_u(b_i - a_i)(a_i + b_i)$, which does not depend on x. Since V^* is C^1 across Σ_i, σ is continuous and the same expansion holds with remainder $O(h^2)$ when the successors straddle Σ_i, so the display holds near Σ_i and its zero set lies within $O(h)$ of the level set of σ wherever $\nabla\sigma \neq 0$. For the slope, $\nabla_x[V^*(x+hv_{b_i}) - V^*(x+hv_{a_i})] = (I + hDv_{b_i})^\mathsf{T} \nabla V^*(x+hv_{b_i}) - (I + hDv_{a_i})^\mathsf{T} \nabla V^*(x+hv_{a_i}) =$

$h(b_i - a_i)\big[\nabla^2 V^* g_u + (Dg_u)^\mathsf{T} \nabla V^*\big] + o(h) = h(b_i - a_i)\nabla\sigma + o(h)$, one-sided where $\nabla^2 V^*$ is, the Df terms cancel, and the remainder is $O(h^2)$ if $\nabla^2 V^*$ and Dg_u are Lipschitz. (ii). Let $\nabla^\pm V^*$ be the one-sided gradients, $\nabla^+ V^* - \nabla^- V^* = \kappa n$ because V^* is continuous across Σ_i. In the band where $x + hv_{b_i}$ lies on the $+$ side and $x + hv_{a_i}$ on the $-$ side, $V^*(x + hv_{b_i}) - V^*(x + hv_{a_i}) = (\nabla^+ V^*)^\mathsf{T}(x + hv_{b_i} - x_\Sigma) - (\nabla^- V^*)^\mathsf{T}(x + hv_{a_i} - x_\Sigma) + O(h^2) = \kappa\, n^\mathsf{T}(x - x_\Sigma) + O(h)$, whose gradient in x is $\kappa n + O(h)$, and the stage-cost difference is $O(h)$. A Runge–Kutta step replaces hv_u by $hv_u + O(h^2)$ and changes only the remainders. □

Proof of Theorem 3. (i) and (iv). Willett et al. (2005); Castro & Nowak (2008) prove minimax lower bounds on the excess zero-one risk for the boundary-fragment class with Hölder exponent one and noise exponent $\kappa = 1$, the noise-free endpoint, of order $N^{-1/n}$ for passive and $N^{-1/(n-1)}$ for active learning. Cauchy–Schwarz on each base, $\|e_i\|_1 \le |D_i|^{1/2}\|e_i\|_2$, and Proposition 1 give $L(\hat\pi) \ge \frac{1}{2}\min_i g_i^-\, e_{01}(\hat\pi)^2/|D|$, and Jensen moves the square inside the expectation.

(ii). This is the cell-wise bisection of Willett et al. (2005) without noise. Partition D_i into cubes of side $\varepsilon_i/(L\sqrt{n-1})$, $|D_i|(L\sqrt{n-1}/\varepsilon_i)^{n-1}$ of them. At each cube centre z_c bisect in y over the tube interval of length $2\delta_0$ with k_i queries, which locates $f_i(z_c)$ to within $\varepsilon_i/2$, and set $\hat f_i$ to that value on the cube. Then $|e_i(z)| \le |f_i(z) - f_i(z_c)| + \varepsilon_i/2 \le \varepsilon_i/2 + \varepsilon_i/2 = \varepsilon_i$. Insert into Proposition 1.

(iii). Minimise $\sum_i c_i N_i^{-s}$ subject to $\sum_i N_i = N$, the Neyman allocation problem (Neyman, 1934) with exponent s. The objective is convex in N_i, stationarity $sc_iN_i^{-s-1} = \mu$ gives $N_i \propto c_i^{1/(1+s)} \propto |D_i|(g_i^+)^{1/(1+s)}$, and substituting back gives $\sum_i c_i N_i^{-s} = (\sum_i c_i^{1/(1+s)})^{1+s} N^{-s}$. □

Proof of Corollary 4. Insert the two allocations into (ii) of Theorem 3 and divide by $L^\star$. The inequalities are Hölder's inequality with exponents $1+s$ and $(1+s)/s$ applied to $|D_i|^{1/(1+s)} g_i^{1/(1+s)} \cdot |D_i|^{s/(1+s)}$, with equality if and only if g_i is constant. □

Proof of Corollary 5. The local precision of a nearest-neighbour rule scales as the local spacing of the labels (Chaudhuri & Dasgupta, 2014). With $\varepsilon(z) \propto q(z)^{-1/a}$ the loss of Proposition 1 is $\int_\Sigma g\, q^{-2/a}$, and minimising it subject to $\int_\Sigma q = N$ gives the pointwise stationarity condition $g\, q^{-2/a-1} = \text{const}$, hence $q \propto g^{a/(a+2)}$. A learner that fills a band of width 2δ with N labels has spacing $(2\delta|\Sigma|/N)^{1/n}$ and the resulting loss is of order $g|\Sigma|(2\delta|\Sigma|/N)^{2/n}$, against $g|\Sigma|(\lambda(\mathcal{X})/N)^{2/n}$ for uniform labels. □

Proposition 8 (Occupation weighting)**.** *If $g_i^- = g_i^+ =: g_i$, $d_\nu^{\pi^*}$ has a density $d^\star$ continuous on the tubes and $d_\nu^{\hat\pi} \to d_\nu^{\pi^*}$ uniformly on the tubes as $\max_i \|e_i\|_\infty \to 0$, then*

$$R(\hat\pi) = \frac{1 + o(1)}{2(1 - \gamma_h)} \sum_i g_i \int_{D_i} d^\star(z, f_i(z))\, e_i(z)^2\, dz, \qquad q^\star \propto (g\, d^\star)^{1/(1+s)},$$

the latter the pointwise allocation optimum.

Proof of Theorem 6 and Proposition 8. $J^\pi(x) = h\ell(x, \pi(x)) + \gamma_h J^\pi(x')$ and $Q^*(x, \pi(x)) = h\ell(x, \pi(x)) + \gamma_h V^*(x')$ with $x' = \Phi_h(x, \pi(x))$, so $J^\pi(x) - V^*(x) = A^*(x, \pi(x)) + \gamma_h[J^\pi(x') - V^*(x')]$, and iterating along the closed loop gives the identity, the remainder vanishing because both costs are bounded. Averaging over $x_0 \sim \nu$ and writing $d_\nu^\pi = (1 - \gamma_h)\sum_k \gamma_h^k \nu P_\pi^k$ gives the integral form, and the bound follows from the definition of C_π. Under the continuity assumption the occupation of $\hat\pi$ on the tubes is $d^\star(1 + o(1))$ and the computation of Proposition 1 with the weight $d^\star$ gives the identity of the proposition. The allocation is the pointwise version of Theorem 3(iii) with g replaced by $gd^\star$. □

B The Capped Allocation and the Scarce Regime

The bound of Theorem 3(ii) lets ε_i grow without limit when a component is starved of labels, but the learner never errs by more than the tube half-width. With $\varepsilon_i = \delta_0 \min\{1, (q_{\text{sat}}/q_i)^{1/(n-1)}\}$, where q_i is the label density on Σ_i and q_{sat} the density at which bisection starts to beat the tube, the optimal

allocation depends on the resolution $r = N/(|\Sigma|q_{\mathrm{sat}})$, the number of labels per saturation cell. For $r \ll 1$ most components stay at the cap and the optimum spends the labels on the most expensive components, exponent toward one. For $r \gg 1$ every component is resolved and the exponent is $1/(1+s)$. Evaluated on the three reference tables the optimal exponent is 1.05, 0.90 and 0.75 at $r = 0.3$, about 0.5 at $r = 1$, and 0.35 from $r = 3$ on, for the pendulum, the near bang-bang and the wide-band mountain car. Between exponents $1/3$ and 1 the capped loss differs by at most 40 percent at $r = 1$, so the exponent is identifiable from data only once the surfaces are resolved. The gap of margin sampling saturates at 4.6, 13.9 and 11.8, the gap of linear-jump sampling grows without bound in r because it starves the cheap components entirely.

C Why Not the Value Function

The switching set is also where a smooth network fitted to the value function by semi-Lagrangian fitted value iteration is structurally worst, since the Bellman targets have a gradient discontinuity there. It is natural to ask whether collocation on the switching set lowers the value error at scarce budgets. In about nine hundred runs on the same three configurations with five seeds, fewer than in the main study and therefore supporting evidence rather than a general verdict, including histogram and mesh-free switching-set densities, trajectory-based action triggers, residual-based densities, local two-level function classes fitted by exact weighted least squares and by averaging, and drift mitigation by target networks and learning-rate schedules, no density lowered the value error below uniform collocation at any budget tried, and every apparent gain traced to a drift of the uniform baseline that the mitigation removed. The reason is structural. The fixed point of a fitted value iteration depends on the projection (Gordon, 1995; Tsitsiklis & Van Roy, 1997; Munos & Szepesvári, 2008), which depends on the sampling density unless importance weights hold it in place, and with the weights in place the density acts only through the estimation variance, whose reduction at the switching set is paid for by a curvature floor of any smooth class there. The policy classifier has neither problem, its target function is the same under every density (Lemma 7) and its class can represent a discontinuity. The closed-loop improvements that switching-set collocation did produce in the value-function experiments sat in the policy, which is what led to the present formulation.

D Experimental Details

Systems. Pendulum, state $(\theta, \omega) \in [-\pi, \pi) \times [-8, 8]$ with periodic angle, dynamics $\dot\theta = \omega$, $\dot\omega = -15\sin(\theta+\pi) + 3u$ (gravitational acceleration 10, mass and length 1), stage cost $(1-\cos\theta) + 0.1\,\omega^2 + 0.001\,u^2$, $u_{\max} = 2$, fourth-order Runge–Kutta step $h = 0.05$, $\gamma = 0.99$ per step, eight fixed starts for the settling statistics and 64 uniform starts for the closed-loop cost gap over 200 steps. Mountain car, state $(x, v) \in [-1.2, 0.6] \times [-0.07, 0.07]$, standard continuous dynamics with elastic walls and no termination, stage cost $(x - \pi/6)^2 + 100v^2 + w_u u^2$ per unit step, $\gamma = 0.995$, $u_{\max} = 1$, closed-loop cost gap over 300 steps. Cart-pole, state $(p, \dot p, \theta, \omega) \in [-2.4, 2.4] \times [-6, 6] \times [-\pi, \pi) \times [-10, 10]$ with periodic angle measured from the upright, force $|u| \le 10$, elastic walls at $|p| = 2.4$, stage cost $[(1-\cos\theta) + 0.1p^2 + 0.01\dot p^2 + 0.1\omega^2 + 10^{-4}u^2]/h$, $h = 0.05$, $\gamma = 0.99$ per step, reference on $21 \times 21 \times 25 \times 21$ nodes with 21 inputs converged to $\delta < 10^{-13}$, closed-loop cost gap over 300 steps.

References. Semi-Lagrangian value iteration on a 200×201 grid with multilinear interpolation, $M = 21$, converged to machine precision in float64. The expert at an arbitrary state interpolates the table at the successor states of all 21 inputs.

Learner and sampler. Classifier on the state embedding that maps periodic angles to $(\cos, \sin)$ and scales the box axes affinely to $[-1, 1]$, with three hidden layers of 64 (128 in the capacity ablation) $\tanh$ units, Adam with learning rate 2×10^{-3}, 400 steps per iteration on minibatches of 256 pooled labels, warm start across iterations. Batches of $B = 32$ queries after 32 uniform ones. Band sampler with 4096 uniform candidates per iteration (16384 on the cart-pole), band quantile $\tau = 0.1$, coverage share $\alpha = 0.1$, exponent $\beta = 1/2$ for the slope estimate over the eight nearest labelled states and $\beta = 1$ for the softmax jump unless swept. Switchmap on a 50×50 histogram with jitter inside the cell. Occupation histogram from 64 closed-loop rollouts of the current classifier on the same grid.

Algorithm 1 Active learning of the policy classifier on the switching set

Input: expert $x \mapsto A^*(x, \cdot)$, budget N, batch B, parameters α, τ, β
Output: policy classifier π_ϕ
$\mathcal{D} \leftarrow \{(x_i, A^*(x_i, \cdot))\}_{i \le B}$ with $x_i \sim \nu$, $\quad \phi \leftarrow \mathrm{fit}(\mathcal{D})$
while $|\mathcal{D}| < N$ **do**
  $\mathcal{C} \sim \nu$, $\quad \mathcal{B}_\phi \leftarrow \{c \in \mathcal{C} : \ m_\phi(c) \le \tau\text{-quantile of } m_\phi \text{ on } \mathcal{C}\}$
  $X_{\mathrm{sw}} \sim q_{\mathrm{sw}}^{(1-\alpha)B}$ with $q_{\mathrm{sw}}(c) \propto \hat{g}(c)^\beta \, \mathbf{1}[c \in \mathcal{B}_\phi]$, $\quad X_\nu \sim \nu^{\alpha B}$
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x, A^*(x, \cdot)) : \ x \in X_{\mathrm{sw}} \cup X_\nu\}$
  $\phi \leftarrow \mathrm{fit}(\mathcal{D})$ by minibatch gradient steps on the loss equation 5
end while

Seeds. Ten seeds (42 to 51) for the main, exponent, slope, quantile and cart-pole sweeps, five for the coverage and capacity ablations.

Ablations. The agreement gain of the switching-set sampler is unchanged for $\alpha \in \{0.03, 0.1, 0.3, 0.6\}$ ($p \le 0.03$ against uniform on the pendulum and the wide-band mountain car), the closed-loop gain needs some coverage ($\alpha = 0.03$ loses it on the pendulum) and disappears at $\alpha = 0.6$. The band quantile $\tau \in \{0.05, 0.1, 0.2, 0.4\}$ leaves the agreement gain unchanged ($p \le 0.01$ at every τ on both systems, ten seeds), the closed-loop gain is insensitive for $\tau \ge 0.1$ and the thinnest band loses it on the pendulum (29 against 12 at 992 queries), a band placed by a still imperfect classifier must be wide enough to contain the true surface. With a 128-unit classifier the ratios persist on the pendulum and grow on the wide-band mountain car (uniform labelling does not reach the agreement 0.78 or the cost gap 0.5 within the budget, the switching-set sampler does at 1241 and 902 queries), consistent with the estimation-error regime of Theorem 3 once the approximation floor moves up.

Rates. On the cost-sensitive loss the log-log slopes of the mean learning curves over the middle of the budget range are -0.8 to -1.4 for uniform labelling and -1.2 to -1.9 for the switching-set samplers on the three two-dimensional configurations, the ratio of Theorem 3 at absolute values below the idealised ones. The slopes of the zero-one error are -0.2 to -0.3 for every learner, the approximation floor of the classifier rather than a sampling effect.

Statistics. Endpoints are the agreement and the closed-loop cost gap at 512 and 1856 queries (992 on the cart-pole) and the per-seed number of queries to reach the agreement targets 0.86, 0.95, 0.75, 0.77 and the cost-gap targets 10, 0.1, 0.5, 8 (pendulum, near bang-bang car, wide-band car, cart-pole), the levels uniform labelling reaches at the end of the budget, computed by log-linear interpolation of each seed's own curve. Differences carry 95 percent bootstrap confidence intervals over the ten seeds (10^4 resamples), the ratio of label complexities is the ratio of medians with its bootstrap interval, and one-sided Mann–Whitney p values are Holm-corrected over the endpoints of each system.

Slope sources. At $\beta = 1/2$ and 992 queries the cost-sensitive loss is 0.17, 0.004 and 0.016 for the softmax jump and 0.07, 0.002 and 0.008 for the slope estimate on the pendulum, the near bang-bang and the wide-band car ($p = 0.09$, 0.00, 0.01), the closed-loop cost gap 7.4, 0.028 and 0.76 against 4.4, 0.027 and 0.57, and 8.0 against 0.72 at 512 queries on the wide-band car ($p < 0.01$), the agreement within 0.004. Rank correlations of the weight with the true slope on the band are -0.07, 0.03, 0.32 and -0.06 for the softmax jump and 0.39, 0.61, 0.64 and 0.28 for the estimate (pendulum, near bang-bang car, wide-band car, cart-pole). The true slope from the table uses three extra expert rows per candidate, not counted. Cart-pole endpoints at 992 queries, slope estimate against uniform, corrected $p = 0.002$ for the agreement and 0.013 for the cost gap.

Switching-set statistics. For every pair of neighbouring grid nodes with different optimal inputs, the cost slope is the finite difference of $Q_b^* - Q_a^*$ across the pair divided by the grid spacing, the base measure is the cell area of the other axes, and the gap of Corollary 4 is evaluated with one component per crossing. The capped version of Appendix B uses the same crossings.

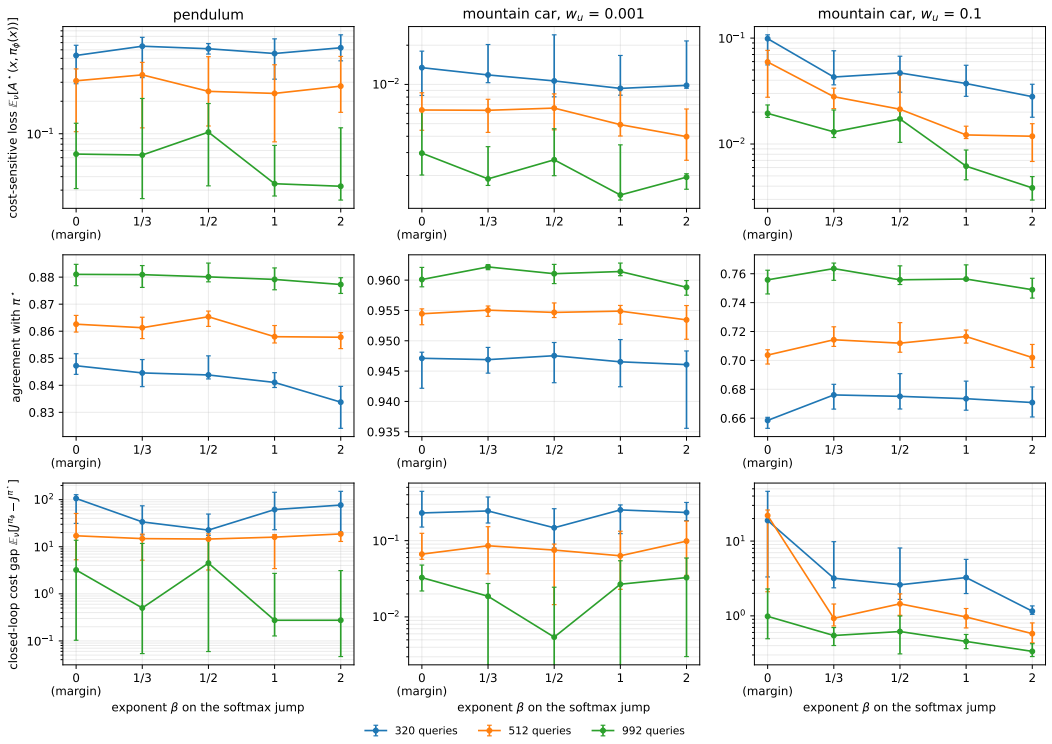

Figure 2: Sweep of the jump exponent β of the band sampler with the softmax proxy ($\beta = 0$ is margin sampling), cost-sensitive loss, ten seeds, median and interquartile range at 320, 512 and 992 queries. Rows are the cost-sensitive loss, the agreement and the closed-loop cost gap, columns the three configurations. On the wide-band mountain car (right) the cost-sensitive loss improves monotonically in β while the agreement peaks at moderate β.

E THE ALLOCATION EXPONENT

The sweep of Figure 2 cannot identify the allocation exponent, because the softmax jump is a proxy for the cost slope, the surfaces are unresolved at 10^3 queries, and the classifier has an approximation floor. The following experiment removes the three confounders and tests Corollary 5 directly. Labels are placed in the switching band of the reference, one box of two by one cells around every crossing of π^* between neighbouring grid nodes, with a crossing chosen with probability proportional to $|D|\, g^\beta$ and g the cost slope of that crossing from the table, plus a coverage share $\alpha = 0.1$ of uniform labels. The fit is one-shot, so the only difference between arms is β. Two learners, a one-nearest-neighbour classifier in the state embedding, which is local and has no approximation floor, and a 256-unit classifier with the cost-sensitive loss and 3000 steps with a cosine schedule. Budgets 10^3, $4 \cdot 10^3$, $1.6 \cdot 10^4$ and $6.4 \cdot 10^4$ labels, $\beta \in \{0, 1/4, 1/2, 3/4, 1\}$, both mountain cars, ten seeds, the cost-sensitive loss evaluated on 10^5 fixed uniform states through the same interpolated expert, so that the evaluation has no grid quantisation.

Figure 3 gives the result. Every β between $1/4$ and $3/4$ beats margin sampling ($\beta = 0$) on the cost-sensitive loss at every budget for both learners ($p < 0.01$, one-sided Mann–Whitney over seeds), by factors between 1.5 and 15. For the nearest-neighbour learner the grid argmin is $\beta = 1/2$ at every budget from $4 \cdot 10^3$ labels on for both configurations, and the vertex of a quadratic fit of the log loss in β falls with the budget, from 0.86 to 0.62 on the wide-band configuration and from 0.59 to 0.53 on the near bang-bang one. Corollary 5 predicts $1/2$ for a band-filling learner in two dimensions and Appendix B predicts the drift from above at scarce budgets. Evaluated with the band-filling rate on the two references, the capped optimum is 0.95 and 1.10 at $r = 0.3$ labels per saturation cell and 0.50 from $r = 3$ on, the gap of margin sampling saturates at 6.9 and 5.5, and the measured gaps at $1.6 \cdot 10^4$ and $6.4 \cdot 10^4$ labels are 15 and 3.3 (wide band) and 2.7 and 1.5 (near bang-bang). The

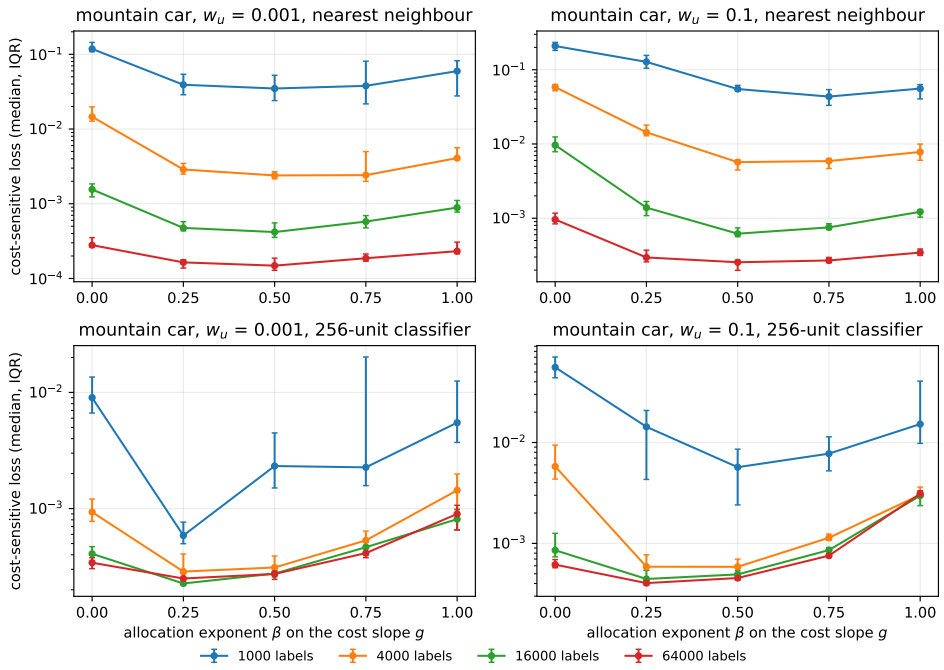

Figure 3: Cost-sensitive loss against the allocation exponent β for an expert-informed band allocation with density $\propto g^{\beta}$, median and interquartile range over ten seeds, one line per budget. Top row the nearest-neighbour learner, bottom row the 256-unit classifier, columns the two mountain-car configurations. For the local learner the minimum sits at $\beta = 1/2$ from $4 \cdot 10^3$ labels on and the vertex of a quadratic fit moves from 0.86 to 0.62 (wide band) and from 0.59 to 0.53 (near bang-bang) as the budget grows, the prediction of Corollary 5 for a band-filling learner with its scarce-regime drift. For the global classifier the optimum is flatter, between $1/4$ and $1/2$, falling to a vertex of 0.33.

agreement decreases monotonically in β at every budget above 10^3, the trade-off of Proposition 1 once more, since a large exponent starves the cheap surfaces whose errors count in the zero-one loss. The global classifier prefers a smaller exponent than the local learner, between $1/4$ and $1/2$ with a vertex falling to 0.33 at the largest budget, and $\beta = 1$ is worse than margin sampling for it from $1.6 \cdot 10^4$ labels on. A network shares capacity across surfaces, so starving a component also degrades its neighbours, which flattens the optimal allocation. The theory covers the local learner, and the local learner behaves as it says.