

Learning Optimal Search Strategies

Stefan Ankirchner ^{*} Maximilian Philipp Thiel [†]

January 22, 2026

Abstract

We explore the question of how to learn an optimal search strategy within the example of a parking problem where parking opportunities arrive according to an unknown inhomogeneous Poisson process. The optimal policy is a threshold-type stopping rule characterized by an indifference position. We propose an algorithm that learns this threshold by estimating the integrated jump intensity rather than the intensity function itself. We show that our algorithm achieves a logarithmic regret growth, uniformly over a broad class of environments. Moreover, we prove a logarithmic minimax regret lower bound, establishing the growth optimality of the proposed approach.

2020 MSC : Primary: 60G40, 93E35; secondary: 62L05, 68Q32.

Keywords : optimal stopping, parking problem, reinforcement learning, regret.

Introduction

The parking problem is a classical example of a search problem. To describe a standard version of it, suppose that an agent is driving along a street, that she can not make a U-turn and that she can only see whether the next lot is free, but not which of the following ones. If the agent arrives at a free lot, then she has to decide whether to take it or not. Once a free lot is discarded, it is discarded forever. A decision rule can be modeled as a stopping time. The parking problem consists of finding the stopping time that minimizes the expected distance of the taken lot to some given target.

^{*}Stefan Ankirchner, Institute for Mathematics, University of Jena, Inselplatz 5, 07743 Jena, Germany. *Email*: s.ankirchner@uni-jena.de.

[†]Maximilian Thiel, Institute for Mathematics, University of Jena, Inselplatz 5, 07743 Jena, Germany. *Email*: maximilian.thiel@uni-jena.de.

To determine an optimal stopping time, one needs to know the distribution of the position of free parking lots. But what if the agent does not know the distribution?

The parking problem is also a paradigm of a problem that usually needs to be solved not once, but in many consecutive rounds, e.g. every morning when driving to your workplace. If the stopping agent does not know the distribution of the free lots, then she can learn it over time by observing in each round the positions of free lots up to the actual chosen one.

In the present article we address the question of what constitutes a good strategy for choosing a parking lot round after round. We do so within the framework of a continuous-time model, where free parking lots arrive according to an inhomogeneous Poisson process. Within this model the optimal stopping rule is of threshold type: there is a position b^* after which it is optimal to take the first free lot. The position b^* can be characterized as an indifference level: if the lot at b^* is free, then the agent is indifferent between taking it or taking the next free one.

We assume that the stopping agent does not know the jump intensity of the Poisson process. We propose a specific algorithm, called indifference level updating (ILU), that estimates the integrated jump intensity and determines a stopping rule in every round, based on the observations made so far.

In order to assess the quality of our algorithm we compute the growth rate of the regret, i.e. the accumulated difference of the expected distance to the target and the minimal expected distance when choosing the optimal stopping rule. Indeed, we show that the regret implied by the ILU algorithm grows logarithmically, uniformly for a large class of distributions.

A crucial property of the ILU algorithm is that it does not estimate the jump intensity function, but the *integrated* jump intensity. For the latter we have estimators with a mean square error (MSE) converging to zero at the rate $1/n$, where n is the number of independent process observations. We assume that the jump intensity is once continuously differentiable, which implies that the roundwise regret can be bounded against the MSE of the estimator. Hence a MSE in $\mathcal{O}(1/n)$ entails a logarithmic growth of the accumulated regret. We remark that any estimator of the intensity function itself, e.g. a kernel estimator, converges at a slower rate than $\mathcal{O}(1/n)$, and hence entails a regret growing faster than the logarithm.

To show that the ILU algorithm is good, we prove that the minimax regret grows logarithmically as well. This means that there is no algorithm that has, uniformly over all instances, a regret rate that grows slower than the logarithm. In this sense the ILU algorithm is a good method for choosing parking lots.

While we present our results in the context of the parking problem, the

underlying methods are not specific to this application. They apply more broadly to a class of timing and search problems with stochastic opportunity arrivals. To simplify the exposition, we develop our approach using the parking problem as a canonical example; it should be possible to carry out extensions to other settings with minor and natural adaptations.

Comparison with the literature

There is an extensive literature on the parking problem formulated as a stopping problem, see e.g. [12] for an early reference in discrete time. A version of the parking problem in continuous time has been considered in [14]. We refer to Chapter 2.5 in [8] for a recent overview on the variants of the parking problem that have been solved already.

The ILU algorithm described in the present article can be seen as an example of a model-based reinforcement learning algorithm. Reinforcement learning (RL) usually is understood to comprise algorithms, e.g. the q -learning algorithm, that make few assumptions on the model and the system distribution. RL algorithms are, therefore, universally applicable. However, the algorithms can be quite inefficient for some applications.

To obtain efficient algorithms for learning solutions to stochastic control problems it is natural to use as much information on the model as possible. The ILU algorithm presented in this article is such an algorithm: it makes use of the fact that optimal stopping rules are of threshold type, i.e. there exists a position after which it is optimal to take the first free lot. Moreover, the algorithm exploits the fact that free lots arrive according to an inhomogeneous Poisson process. A crucial ingredient of the algorithm is the estimation of the integrated jump intensity of the Poisson process.

There is quite some literature on model-based RL algorithms for models where the state dynamics are described as a Markov decision process. We refer to [11] for a survey. Examples of model-based RL algorithms for continuous-time stochastic control problems are still rather scarce. A first approach for linear-quadratic problems in continuous time is presented in [7]. The authors introduce a weighted least-squares algorithm for learning drift rates and hence the optimal control. The growth rate of the regret is not provided, and it seems that its determination has not yet been made. [2, 10, 15] consider linear-quadratic problems in an episodic finite time horizon setting, providing non-asymptotic regret bounds.

The article [13] considers propagator models and describe algorithms achieving sublinear regrets. [6, 3, 5, 4] examine singular and impulsive control problems in a non-parametric, ergodic setting, and obtain algorithms with

sublinear regret rates of power type. The article [1] introduces a steering algorithm to keep a process as close to some given path as possible, exploiting that the process is driven by a Brownian motion and that the drift rates are from an unknown bounded interval.

1 The parking problem in continuous time

In this section we summarize some results on the parking problem in continuous time.

Suppose that the parking spaces are located on the interval $[S, \infty)$, where $S \in (-\infty, 0)$. Assume that free parking lots arrive according to an inhomogeneous Poisson process with jump intensity $\lambda : [S, \infty) \rightarrow (0, \infty)$. More precisely, let $(N_t^\lambda)_{t \geq S}$ a right-continuous process, defined on some probability space (Ω, \mathcal{F}, P) , such that the increments are independent and $(N_t^\lambda - N_s^\lambda)$ is Poisson distributed with the parameter $\int_s^t \lambda(u) du$, for all $s \leq t$. We interpret the jump times of the process (N_t) as the positions of the free parking lots the agent can use while driving along the street from S to the right.

In this section we assume that the jump intensity function λ is known by the stopping agent. In the next section we will omit this assumption and assume that the stopping agent does not know λ , but can learn it by observing independent samples of the jump process (N_t) .

We suppose that the agent wants to park her car as close as possible to the target 0, in expectation. A policy for searching a parking lot can be described in terms of decision rule describing for each parking lot $t \in [S, \infty)$ whether to take it or not, in case it is free.

It is straightforward to show that it is enough to consider only decision rules of the following threshold type: there is a threshold $b \in [S, 0]$ after which the first free lot is accepted; and before b any free lot is discarded. Indeed, for any decision rule one can construct a threshold rule with an expected distance to the target that is not larger.

We model the decision rule with threshold $b \in [S, 0]$ as the stopping time $\tau_b := \inf\{t \geq b : N_t > N_b\}$. Observe that τ_b is the first jump time of N after b ; we interpret τ_b as the position of the first free parking space after b .

Note that the aim to park the car as close as possible to the target amounts in the problem of finding the threshold b for which $E|\tau_b|$ becomes minimal. We refer to b^* such that $E|\tau_{b^*}| = \min_{b \in [S, 0]} E[|\tau_b|]$ as an optimal threshold.

The next theorem provides a sufficient condition for b^* to be an optimal threshold.

Theorem 1.1. Let $b^* \in [S, 0)$ be such that

$$\int_{b^*}^0 e^{\int_y^0 \lambda(u) du} dy = \int_0^\infty e^{-\int_0^y \lambda(u) du} dy. \quad (1.1)$$

Then b^* is an optimal threshold.

Proof. First note that for all $b \in [S, 0]$ we have

$$E|\tau_b| = - \int_b^0 y \lambda(y) e^{-\int_b^y \lambda(u) du} dy + \int_0^\infty y \lambda(y) e^{-\int_b^y \lambda(u) du} dy.$$

Next observe that for b to be optimal it is sufficient that it satisfies the FOC $\frac{\partial E|\tau_b|}{\partial b} = 0$. We now show that (1.1) is equivalent to the FOC. Indeed,

$$\frac{\partial E|\tau_b|}{\partial b} = \lambda(b) \left[b - \int_b^0 y \lambda(y) e^{-\int_b^y \lambda(u) du} dy + \int_0^\infty y \lambda(y) e^{-\int_b^y \lambda(u) du} dy \right],$$

and hence $\frac{\partial E|\tau_b|}{\partial b} = 0$ is equivalent to

$$b = \int_b^0 y \lambda(y) e^{-\int_b^y \lambda(u) du} dy - \int_0^\infty y \lambda(y) e^{-\int_b^y \lambda(u) du} dy.$$

By using $\lambda(y) e^{-\int_b^y \lambda(u) du} = -\frac{\partial}{\partial y} e^{-\int_b^y \lambda(u) du}$ and applying integration by parts, the last equation can be rewritten as

$$\begin{aligned} b &= - \int_b^0 y \frac{\partial}{\partial y} e^{-\int_b^y \lambda(u) du} dy + \int_0^\infty y \frac{\partial}{\partial y} e^{-\int_b^y \lambda(u) du} dy \\ &= - \left[y e^{-\int_b^y \lambda(u) du} \right]_b^0 + \int_b^0 e^{-\int_b^y \lambda(u) du} dy + \left[y e^{-\int_b^y \lambda(u) du} \right]_0^\infty - \int_0^\infty e^{-\int_b^y \lambda(u) du} dy \\ &= -(0 - b) + \int_b^0 e^{-(\int_b^0 \lambda(u) du - \int_b^y \lambda(u) du)} dy + 0 - \int_0^\infty e^{-(\int_b^0 \lambda(u) du + \int_0^y \lambda(u) du)} dy \\ &= b + e^{-\int_b^0 \lambda(u) du} \left(\int_b^0 e^{\int_b^y \lambda(u) du} dy - \int_0^\infty e^{-\int_0^y \lambda(u) du} dy \right). \end{aligned}$$

This implies that the FOC is equivalent to

$$\int_b^0 e^{\int_b^y \lambda(u) du} dy = \int_0^\infty e^{-\int_0^y \lambda(u) du} dy.$$

□

In the following we rewrite the LHS of (1.1) in terms of the integrated jump intensity

$$\Lambda(y) := \int_y^0 \lambda(u) du, \quad y \in [S, 0].$$

Hence, if b^* satisfies the equation

$$\int_{b^*}^0 e^{\Lambda(y)} dy = \int_0^\infty e^{-\int_0^y \lambda(u) du} dy, \quad (1.2)$$

then b^* is an optimal threshold.

We finally remark that since λ is assumed to be positive, there exists at most one real $b^* \in [S, 0]$ satisfying equation (1.1).

Intuition behind the optimality criterion

Note that for all $t > 0$

$$P(\tau_0 > t) = P(N_t^\lambda - N_0^\lambda = 0) = e^{-\int_0^t \lambda(u) du},$$

and hence the RHS of (1.1) coincides with the expectation $E[\tau_{b^*} | \tau_{b^*} > 0] = E(\tau_0) = \int_0^\infty P(\tau_0 > t) dt$. In other words, the right-hand side of (1.1) corresponds exactly to the expected value of the first jump time after the parking space 0. The right-hand side, therefore, measures the expected costs after the driver has passed parking space 0.

The left-hand side of Equation (1.1) is equal to $\frac{|b^*| - E[\tau_{b^*} | 1_{\{\tau_{b^*} < 0\}}]}{P(\tau_{b^*} \geq 0)}$. Rearranging terms yields that (1.1) is equivalent to

$$|b^*| = E[\tau_{b^*}]. \quad (1.3)$$

Equation (1.3) characterizes b^* as the position at which the agent is indifferent between taking the lot, provided it is free, and continuing until the next free lot.

Optimality gap

We denote by

$$\Delta(b) := E|\tau_b| - E|\tau_{b^*}| \quad (1.4)$$

the optimality gap of choosing the stopping rule with threshold $b \in [S, 0]$ instead of the optimal threshold b^* . The next lemma collects some properties of the optimality gap.

Lemma 1.2. Δ is differentiable. Moreover, if λ is continuously differentiable, then Δ is twice continuously differentiable, and the second derivative is given by

$$\begin{aligned} \Delta''(b) &= \lambda'(b) \left[b - \int_b^0 y\lambda(y)e^{-\int_b^y \lambda(u)du} dy + \int_0^\infty y\lambda(y)e^{-\int_b^y \lambda(u)du} dy \right] \\ &\quad + \lambda(b) \left[1 + b\lambda(b) - \int_b^0 y\lambda(y)\lambda(b)e^{-\int_b^y \lambda(u)du} dy + \int_0^\infty y\lambda(y)\lambda(b)e^{-\int_b^y \lambda(u)du} dy \right]. \end{aligned} \quad (1.5)$$

Proof. The claims are straightforward to show. \square

We close this section by considering the special case where λ is constant.

Corollary 1.3. Suppose that λ is constant and greater than $\ln(2)/|S|$. Then $b^* = -\ln(2)/\lambda$ is an optimal threshold and

$$\Delta_\lambda(b) = \frac{1}{\lambda}(2e^{\lambda b} - 1) - b - \frac{\ln(2)}{\lambda}. \quad (1.6)$$

2 Learning optimal rules

Based on the Equation (1.1), the optimal stopping time τ_{b^*} can be determined with the knowledge of the intensity function λ . In the following we assume that the agent does not know the intensity function λ . We do assume, however, that the agent has to solve the stopping problem in many consecutive rounds and that in each round she observes the jump process up to the chosen stopping time. With the observations the agent can estimate the true intensity function λ . Hence, in round n the agent can approximate the optimal stopping rule by using the observations made in previous rounds $0, 1, \dots, n-1$.

Let $(N_t^0)_{t \in [S, \infty)}, (N_t^1)_{t \in [S, \infty)}, \dots$ be a sequence of stochastic processes on a measurable space (Ω, \mathcal{F}) . Assume that for every positive and measurable intensity function $\lambda : [S, \infty) \rightarrow (0, \infty)$ there exists a probability measure P_λ such that $(N_t^0)_{t \in [S, \infty)}, (N_t^1)_{t \in [S, \infty)}, \dots$ is an independent sequence of Poisson processes with jump intensity λ .

Similar as in the previous section, we denote by τ_b^i the first jump time of N^i after time $b \in [S, 0]$. Moreover, the optimality gap under P_λ is denoted by $\Delta_\lambda(b)$.

Definition 2.1 (Policy). A policy $(\pi_n)_{n \geq 0}$ is a sequence of random variables with values in $[S, 0]$ such that π_n is measurable with respect to \mathcal{F}_n , where

$\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n := \bigvee_{k=0}^{n-1} \sigma(N_t^{\lambda, k} : t \in [S, \tau_{\pi_k}^k])$ for every $n \geq 1$. (Recall that if $(\mathcal{A}_i)_{i \in I}$ is a family of σ -algebras indexed by a set I , then $\bigvee_{i \in I} \mathcal{A}_i$ denotes the smallest σ -algebra containing all \mathcal{A}_i .) We denote the set of policies by Π .

We interpret π_n as the threshold the agent chooses in round n for the stopping rule.

For a given positive and measurable intensity function λ , we define the regret of an arbitrary policy $(\pi_n)_{n \geq 0}$ in round $T \in \mathbb{N}$ as

$$\mathcal{R}_\lambda^\pi(T) := \sum_{n=0}^T E_\lambda[\Delta_\lambda(\pi_n)],$$

where Δ_λ is the optimality gap function defined.

Next we describe a specific algorithm that leads to a policy with a regret growing logarithmically as the number of rounds converge to infinity. We later prove that there is no policy that can achieve a regret growing slower than the logarithm. In this sense the algorithm is asymptotically optimal.

Algorithm (Indifference level updating (ILU)).

1. Initialize $\mathcal{I} = \{0\}$.
2. Choose the stopping time with threshold 0 in round 0.
3. In any round $n \geq 1$ do the following:
 - (a) Compute

$$\hat{\Gamma}(y) := \frac{\sum_{i \in \mathcal{I}} (N_0^i - N_y^i)}{|\mathcal{I}|}, \quad y \in [S, 0], \quad (2.1)$$

and

$$\hat{\varphi} := \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \tau_0^i \quad (2.2)$$

- (b) Determine \hat{b} such that

$$\int_{\hat{b}}^0 e^{\hat{\Gamma}(y)} dy = \hat{\varphi}. \quad (2.3)$$

If there is no solution $\hat{b} \in [S, 0]$, then set $\hat{b} = S$.

- (c) Choose the stopping time with threshold \hat{b} .
- (d) If the stopping time takes a value greater than zero, then $\mathcal{I} \leftarrow \mathcal{I} \cup \{n\}$.

Remark: Note that \mathcal{I} is the set of rounds where the algorithm stops after 0. We refer to \mathcal{I} as the set of rounds with full information. The quantity $N_0^j - N_y^j$ is the number of jumps on the interval $[y, 0]$ in round j , for $y \in [S, 0]$. The function $\hat{\Gamma}(y)$ is an estimator of the integrated jump intensity function $\Lambda(y)$, $y \in [S, 0]$. Moreover, $\hat{\varphi}$ is an estimator of $E_\lambda(\tau_0^i)$, the expected first jump time after 0.

Since $\hat{\Gamma}$ and $\hat{\varphi}$ make only use of observations of the rounds in \mathcal{I} , the sequence of thresholds chosen by the ILU algorithm is a policy in the sense of Definition 2.1.

3 Main results

We start by defining an environment class of smooth intensity functions λ .

Definition 3.1 (Environment class $\mathcal{M}(L)$, $L \in (1, \infty)$).

We say $\lambda \in \mathcal{M}(L)$ if and only if

1. λ is continuously differentiable;
2. λ is bounded from below by $\ln(2)/|S| + 1/L$;
3. λ and λ' are bounded by L on $[S, \infty)$, i.e. $\lambda(u) \leq L$ and $|\lambda'(u)| \leq L$, for all $u \in [S, \infty)$.

Notice that property 2 guarantees the following.

Lemma 3.2. *For any $\lambda \in \mathcal{M}(L)$ there exists a unique $b^* \in (S, 0]$ satisfying (1.1); moreover this b^* satisfies $b^* - S \geq |S| - \left(\frac{\ln(2)}{|S|} + \frac{1}{L}\right)^{-1} \ln(2) > 0$.*

Proof. The smallest possible value for b^* , denoted by b_{min}^* , is achieved by $\lambda \in \mathcal{M}(L)$ that is constant equal to the minimal value. Hence b_{min}^* satisfies

$$\int_{b_{min}^*}^0 e^{-(\frac{\ln(2)}{|S|} + \frac{1}{L})u} du = \left(\frac{\ln(2)}{|S|} + \frac{1}{L}\right)^{-1}. \quad (3.1)$$

A straightforward computation shows that

$$b_{min}^* = -\left(\frac{\ln(2)}{|S|} + \frac{1}{L}\right)^{-1} \ln(2) > -\frac{|S|}{\ln(2)} \ln(2) = S.$$

□

In the following we denote by $\mathcal{R}_\lambda^{ILU}(T)$ the regret in round $T \in \mathbb{N}$ entailed by the ILU algorithm under the probability measure P_λ .

Theorem 3.3 (Upper bound of ILU on class $\mathcal{M}(L)$). *There exists a constant $C \in \mathbb{R}$, depending only on S and L , such that for all $T \in \mathbb{N}$, we have*

$$\sup_{\lambda \in \mathcal{M}(L)} \mathcal{R}_\lambda^{ILU}(T) \leq C \ln(T + 1).$$

The proof is provided in Section 5.

Theorem 3.4 (Lower bound on class $\mathcal{M}(L)$). *Suppose that L is larger than $\ln(2)/|S| + 1/L$, and hence $\mathcal{M}(L)$ is non-empty. Then there exists $c \in (0, \infty)$ such that for all $T \in \mathbb{N}$*

$$\inf_{\text{policy } \pi} \sup_{\lambda \in \mathcal{M}(L)} \mathcal{R}_\lambda^\pi(T) \geq c \ln(T).$$

The proof is provided in Section 6.

Theorem 3.3 and Theorem 3.4 imply that the ILU algorithm has an asymptotically optimal growth rate.

4 MSE of the estimated threshold

We use the setting of Section 2. To simplify notation we omit the superscript zero and write $\tau_b = \tau_b^0$. Note that τ_b has the same distribution as τ_b^i for any $i \geq 1$, under any P_λ .

In the following let $\lambda \in \mathcal{M}(L)$. To simplify notation we omit λ in P_λ , E_λ , Var_λ etc.

Lemma 4.1. *We have $\text{Var}(\tau_0) < \infty$.*

Proof. Since $\lambda \in \mathcal{M}(L)$, we have $\lambda(u) \geq C$ for all $u \in [S, \infty)$, where $C :=$

$\ln(2)/|S| + \frac{1}{L}$. Moreover,

$$\begin{aligned}
E(\tau_0^2) &= \int_0^\infty P(\tau_0^2 > t) dt = \int_0^\infty P(\tau_0 > \sqrt{t}) dt \\
&= 2 \int_0^\infty P(\tau_0 > u) u du \\
&= 2 \int_0^\infty e^{-\int_0^u \lambda(y) dy} u du \\
&\leq 2 \int_0^\infty e^{-Cu} u du \\
&= 2 \left(\left[-\frac{1}{C} e^{-Cu} u \right]_0^\infty + \frac{1}{C} \int_0^\infty e^{-Cu} \right) \\
&= \frac{2}{C^2} < \infty.
\end{aligned}$$

It is obvious that $|E(\tau_0)| < \infty$. Therefore $\text{Var}(\tau_0) < \infty$. \square

We first analyze the ILU algorithm under the additional assumption of full observation. We use this assumption to make it easier to work with the estimators.

Definition 4.2 (Full Information Policy). A policy $(\pi_n^{full})_{n \geq 0}$ is a sequence of random variables with values in $[S, 0]$ such that π_n is measurable with respect to \mathcal{F}_n , where $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n := \bigvee_{k=0}^{n-1} \sigma(N_t^{\lambda, k} : t \in [S, \tau_0^k])$ for every $n \geq 1$. We denote the set of policies by Π^{full} .

In the following, we consider the estimators for $n \geq 1$

$$\hat{\Lambda}_n(y) := \frac{\sum_{i=0}^{n-1} (N_0^i - N_y^i)}{n},$$

and

$$\hat{\tau}_{0,n} := \frac{1}{n} \sum_{i=0}^{n-1} \tau_0^i,$$

respectively.

Note that $(\hat{\Lambda}_n)$ and $(\hat{\tau}_{0,n})$ are exactly the counterparts to the estimators $\hat{\Gamma}$ and $\hat{\varphi}$ from the ILU Algorithm, under the additional assumption that full information is always available.

Definition 4.3. Let $\hat{b}_0 := 0$. For every $n \in \mathbb{N}$ let \hat{b}_n be the real in $[S, 0]$ that satisfies

$$\int_{\hat{b}_n}^0 e^{\hat{\Lambda}_n(y)} dy = \hat{\tau}_{0,n}. \quad (4.1)$$

If there is no $\hat{b}_n \in [S, 0]$ satisfying Equation (4.1), we set $\hat{b}_n = S$.

Note that $(\hat{b}_n) \in \Pi^{full}$, but not necessarily a policy in the sense of Definition 2.1.

Lemma 4.4. *For all $y \in [S, 0]$ we have that $MSE(\hat{\Lambda}_n(y)) = \frac{1}{n}\Lambda(y)$ and $\hat{\Lambda}_n(y)$ is an unbiased estimator.*

Proof. Straightforward. □

Lemma 4.5. *It holds that the $MSE(\hat{\tau}_{0,n}) = \frac{1}{n} \text{Var}(\tau_0)$ and $\hat{\tau}_{0,n}$ is an unbiased estimator.*

Proof. Straightforward. □

Lemma 4.6. *For all $n \in \mathbb{N}$ we have*

$$E\left(\sup_{x \in [S, 0]} (\hat{\Lambda}_n(x) - \Lambda(x))^2\right) \leq 4\Lambda(S) \frac{1}{n} \leq 4L|S| \frac{1}{n} \quad (4.2)$$

Proof. First, notice that $M(y) := \hat{\Lambda}_n(-y) - \Lambda(-y)$, $y \in [0, |S|]$, is a time-continuous martingal w.r.t. the natural filtration $(\mathcal{F}_y)_{y \in [0, |S|]}$, where $\mathcal{F}_y = \sigma(N_0^i - N_t^i : -y \leq t \leq 0, i = 0, \dots, n-1)$. Note that $(M(y))$ is a square integrable martingale with left-continuous paths. Therefore, Doob's L^2 -maximal inequality for is applicable and we get

$$\begin{aligned} E\left(\sup_{x \in [S, 0]} (\hat{\Lambda}_n(x) - \Lambda(x))^2\right) &= E\left[\sup_{y \in [0, |S|]} M(y)^2\right] \\ &\leq 4E[M(|S|)^2] = E((\hat{\Lambda}_n(S) - \Lambda(S))^2). \end{aligned}$$

With Lemma 4.4 we get

$$E\left(\sup_{x \in [S, 0]} (\hat{\Lambda}_n(x) - \Lambda(x))^2\right) \leq 4\Lambda(S) \frac{1}{n},$$

and with the simple estimate $\Lambda(S) \leq L|S|$ we have (4.2). □

Proposition 4.7. Let $M := \min\{m \in \mathbb{N}_0 : S + m\frac{1}{2L} \geq 0\}$. Then, for all $n \in \mathbb{N}$, we have

$$\begin{aligned} E\left(\left(\hat{b}_n - b^*\right)^2\right) &\leq \left(S^2 \frac{4\Lambda(S) + \text{Var}(\tau_0)}{\varepsilon^2}\right) \frac{1}{n} + \\ &\quad + \left(S^2 4\Lambda(S)(M+1) + 2\text{Var}(\tau_0) + 8|S| \int_S^0 e^{2\Lambda(y)} \Lambda(y) dy\right) \frac{1}{n}, \end{aligned} \quad (4.3)$$

where $0 < \varepsilon < \min\{E(\tau_0), 1, \frac{b^* - S}{3+2E(\tau_0)}\}$.

Proof. Fix $n \in \mathbb{N}$ and suppose first that we are on the event $B_n := \{|S| > |\hat{b}_n|\}$.

Note that \hat{b}_n is determined in such a way that

$$\int_{\hat{b}_n}^0 e^{\hat{\Lambda}_n(y)} dy = \hat{\tau}_{0,n}.$$

Furthermore, the optimal b^* and the unknown, true intensity function λ satisfy

$$\int_{b^*}^0 e^{\Lambda(y)} dy = E(\tau_0).$$

Therefore

$$\int_{b^*}^0 e^{\Lambda(y)} dy - \int_{\hat{b}_n}^0 e^{\hat{\Lambda}_n(y)} dy = E(\tau_0) - \hat{\tau}_{0,n}.$$

This is equivalent to

$$\int_{b^*}^0 e^{\Lambda(y)} dy - \int_{\hat{b}_n}^0 e^{\Lambda(y)} dy + \int_{\hat{b}_n}^0 \left(e^{\Lambda(y)} - e^{\hat{\Lambda}_n(y)} \right) dy = E(\tau_0) - \hat{\tau}_{0,n}. \quad (4.4)$$

1st case: Let $\hat{b}_n \geq b^*$. It follows from (4.4):

$$\begin{aligned} \int_{b^*}^{\hat{b}_n} e^{\Lambda(y)} dy &= E(\tau_0) - \hat{\tau}_{0,n} - \int_{\hat{b}_n}^0 \left(e^{\Lambda(y)} - e^{\hat{\Lambda}_n(y)} \right) dy \\ &= E(\tau_0) - \hat{\tau}_{0,n} + \int_{\hat{b}_n}^0 e^{\Lambda(y)} \left(e^{\hat{\Lambda}_n(y) - \Lambda(y)} - 1 \right) dy. \end{aligned} \quad (4.5)$$

In the following, we estimate the left-hand side of equation (4.5) from below and the right-hand side from above. We define for this

$$\begin{aligned} A &:= \int_{b^*}^{\hat{b}_n} e^{\Lambda(y)} dy, \\ B &:= E(\tau_0) - \hat{\tau}_{0,n} + \int_{\hat{b}_n}^0 e^{\Lambda(y)} \left(e^{\hat{\Lambda}_n(y) - \Lambda(y)} - 1 \right) dy. \end{aligned}$$

Note that

$$\begin{aligned} A &\geq (\hat{b}_n - b^*) \min_{y \in [S, 0]} e^{\Lambda(y)} \geq \hat{b}_n - b^*; \\ B &\leq |E(\tau_0) - \hat{\tau}_{0,n}| + \int_S^0 e^{\Lambda(y)} \left| e^{\hat{\Lambda}_n(y) - \Lambda(y)} - 1 \right| dy. \end{aligned}$$

Observe that for all $x \in [-1, 1]$ we have

$$|e^x - 1| \leq 2|x|. \quad (4.6)$$

Define $C_n := \left\{ \left| \hat{\Lambda}_n(y) - \Lambda(y) \right| \leq 1, \forall y \in [S, 0] \right\}$.
Then on C_n we have

$$\left| e^{\hat{\Lambda}_n(y) - \Lambda(y)} - 1 \right| \leq 2 \left| \hat{\Lambda}_n(y) - \Lambda(y) \right|.$$

In particular, using the estimates of A and B, we get

$$\hat{b}_n - b^* \leq |E(\tau_0) - \hat{\tau}_{0,n}| + 2 \int_S^0 e^{\Lambda(y)} \left| \hat{\Lambda}_n(y) - \Lambda(y) \right| dy.$$

2nd case: Let $\hat{b}_n < b^*$. Then (4.4) implies

$$- \int_{\hat{b}_n}^{b^*} e^{\Lambda(y)} dy = E(\tau_0) - \hat{\tau}_{0,n} + \int_{\hat{b}_n}^0 e^{\Lambda(y)} \left(e^{\hat{\Lambda}_n(y) - \Lambda(y)} - 1 \right) dy,$$

which is equivalent to

$$\int_{\hat{b}_n}^{b^*} e^{\Lambda(y)} dy = (\hat{\tau}_{0,n} - E(\tau_0)) + \int_{\hat{b}_n}^0 e^{\Lambda(y)} \left(- \left(e^{\hat{\Lambda}_n(y) - \Lambda(y)} - 1 \right) \right) dy. \quad (4.7)$$

In the following, we estimate the left-hand side of equation (4.7) from below and the right-hand side from above. We define for this

$$A := \int_{\hat{b}_n}^{b^*} e^{\Lambda(y)} dy,$$

$$B := (\hat{\tau}_{0,n} - E(\tau_0)) + \int_{\hat{b}_n}^0 e^{\Lambda(y)} \left(- \left(e^{\hat{\Lambda}_n(y) - \Lambda(y)} - 1 \right) \right) dy.$$

Then

$$A \geq (b^* - \hat{b}_n) \min_{y \in [S, 0]} e^{\Lambda(y)} \geq b^* - \hat{b}_n;$$

$$B \leq |E(\tau_0) - \hat{\tau}_{0,n}| + \int_S^0 e^{\Lambda(y)} \left| e^{\hat{\Lambda}_n(y) - \Lambda(y)} - 1 \right| dy.$$

Let C_n be as in the first case. Then on C_n we have

$$\left| e^{\hat{\Lambda}_n(y) - \Lambda(y)} - 1 \right| \leq 2 \left| \hat{\Lambda}_n(y) - \Lambda(y) \right|.$$

Therefore applies

$$b^* - \hat{b}_n \leq |E(\tau_0) - \hat{\tau}_{0,n}| + 2 \int_S^0 e^{\Lambda(y)} \left| \hat{\Lambda}_n(y) - \Lambda(y) \right| dy.$$

To sum up, in any case, if B_n and C_n occur, we have

$$\left| \hat{b}_n - b^* \right| \leq |E(\tau_0) - \hat{\tau}_{0,n}| + 2 \int_S^0 e^{\Lambda(y)} \left| \hat{\Lambda}_n(y) - \Lambda(y) \right| dy. \quad (4.8)$$

Thus, using the estimate $(e + f)^2 \leq 2(e^2 + f^2)$, $e, f \in \mathbb{R}$

$$\begin{aligned} E \left(\left| \hat{b}_n - b^* \right|^2 \mathbf{1}_{B_n} \mathbf{1}_{C_n} \right) &\leq 2 \left\{ E \left[(E(\tau_0) - \hat{\tau}_{0,n})^2 \mathbf{1}_{B_n} \mathbf{1}_{C_n} \right] + \right. & (4.9) \\ &\quad \left. + 4E \left[\mathbf{1}_{B_n} \mathbf{1}_{C_n} \left(\int_S^0 e^{\Lambda(y)} \left| \hat{\Lambda}_n(y) - \Lambda(y) \right| dy \right)^2 \right] \right\} \\ &\leq 2 \left\{ E \left[(E(\tau_0) - \hat{\tau}_{0,n})^2 \right] + \right. \\ &\quad \left. + 4E \left[\left(\int_S^0 e^{\Lambda(y)} \left| \hat{\Lambda}_n(y) - \Lambda(y) \right| dy \right)^2 \right] \right\}. \end{aligned}$$

Using Jensen's inequality for integrals we further obtain

$$\begin{aligned} &E \left(\left| \hat{b}_n - b^* \right|^2 \mathbf{1}_{B_n} \mathbf{1}_{C_n} \right) \\ &\leq 2 \left\{ E \left[(E(\tau_0) - \hat{\tau}_{0,n})^2 \right] + 4E \left[|S| \left(\int_S^0 \left(e^{\Lambda(y)} \left| \hat{\Lambda}_n(y) - \Lambda(y) \right| \right)^2 dy \right) \right] \right\} \\ &= 2 \left\{ \text{MSE}(\hat{\tau}_{0,n}) + 4|S| \int_S^0 e^{2\Lambda(y)} \text{MSE}(\hat{\Lambda}_n(y)) dy \right\}. \end{aligned}$$

With Lemma 4.4 and 4.5 we further obtain that

$$E \left(\left| \hat{b}_n - b^* \right|^2 \mathbf{1}_{B_n} \mathbf{1}_{C_n} \right) \leq \frac{1}{n} \left(2\text{Var}(\tau_0) + 8|S| \int_S^0 e^{2\Lambda(y)} \Lambda(y) dy \right). \quad (4.10)$$

In summary, we have

$$\begin{aligned} &E((\hat{b}_n - b^*)^2) \\ &= E((\hat{b}_n - b^*)^2 \mathbf{1}_{B_n^c}) + E((\hat{b}_n - b^*)^2 \mathbf{1}_{B_n} \mathbf{1}_{C_n^c}) + E((\hat{b}_n - b^*)^2 \mathbf{1}_{B_n} \mathbf{1}_{C_n}) \\ &\leq S^2 P(B_n^c) + S^2 P(C_n^c) + \frac{1}{n} \left(2\text{Var}(\tau_0) + 8|S| \int_S^0 e^{2\Lambda(y)} \Lambda(y) dy \right). \quad (4.11) \end{aligned}$$

Due to Lemma 4.5, 4.6 and Lemma 4.8 below we have

$$\begin{aligned} P(B_n^c) &\leq \frac{E(\|\hat{\Lambda}_n - \Lambda\|^2) + \text{Var}(\hat{\tau}_{0,n})}{\varepsilon^2} \\ &\leq \frac{4\Lambda(S) + \text{Var}(\tau_0)}{\varepsilon^2} \frac{1}{n} \end{aligned} \quad (4.12)$$

where $0 < \varepsilon < \min\{E(\tau_0), 1, \frac{b^* - S}{3 + 2E(\tau_0)}\}$.

Furthermore, with the Estimate (4.14) from Lemma 4.10 below, we have

$$P(C_n^c) \leq (4\Lambda(S)(M + 1)) \frac{1}{n}. \quad (4.13)$$

By combining the estimates (4.12), (4.13) and (4.11) we get (4.7). \square

Lemma 4.8. *Let $B_n := \{|S| > |\hat{b}_n|\}$, $n \in \mathbb{N}$. Then we have*

$$P(B_n^c) \leq \frac{E(\|\hat{\Lambda}_n - \Lambda\|_\infty^2) + \text{Var}(\hat{\tau}_{0,n})}{\varepsilon^2},$$

where $0 < \varepsilon < \min\{E(\tau_0), 1, \frac{b^* - S}{3 + 2E(\tau_0)}\}$.

Proof. We want to determine $\varepsilon := \varepsilon(\lambda) > 0$ such that $\|\hat{\Lambda}_n - \Lambda\|_\infty < \varepsilon$ and $|E(\tau_0) - \hat{\tau}_{0,n}| < \varepsilon$ implies $|\hat{b}_n| < |S|$. Let $\delta := b^* - S$.

We show via contradiction that $\varepsilon < \min\{1, \frac{\delta}{3 + 2E(\tau_0)}\}$ implies $|\hat{b}_n| < |S|$.

Assume $|\hat{b}_n| = |S|$. We have

$$\begin{aligned} \int_{\hat{b}_n}^0 e^{\hat{\Lambda}_n(u)} du &\geq \int_{\hat{b}_n}^0 e^{\Lambda(u) - \varepsilon} du = \int_S^0 e^{\Lambda(u) - \varepsilon} du = e^{-\varepsilon} \int_{b^* - \delta}^0 e^{\Lambda(u)} du \\ &= e^{-\varepsilon} \left(\int_{b^*}^0 e^{\Lambda(u)} du + \int_{b^* - \delta}^{b^*} e^{\Lambda(u)} du \right) = e^{-\varepsilon} \left(E(\tau_0) + \int_{b^* - \delta}^{b^*} e^{\Lambda(u)} du \right) \\ &\geq e^{-\varepsilon} \left(E(\tau_0) + \delta \min_{u \in [S, 0]} e^{\Lambda(u)} \right) = e^{-\varepsilon} (E(\tau_0) + \delta). \end{aligned}$$

Moreover

$$E(\tau_0) + \varepsilon \geq \hat{\tau}_{0,n} \geq \int_{\hat{b}_n}^0 e^{\hat{\Lambda}_n(u)} du \geq e^{-\varepsilon} (E(\tau_0) + \delta).$$

Therefore we have a contradiction if $e^\varepsilon(E(\tau_0) + \varepsilon) < (E(\tau_0) + \delta)$. Note that for $0 \leq \varepsilon \leq 1$ it holds that $e^\varepsilon \leq 1 + \varepsilon + \varepsilon^2$. It follows for $\varepsilon \in [0, 1]$

$$\begin{aligned} e^\varepsilon(E(\tau_0) + \varepsilon) &\leq (1 + \varepsilon + \varepsilon^2)(E(\tau_0) + \varepsilon) \leq (1 + 2\varepsilon)(E(\tau_0) + \varepsilon) \\ &\leq E(\tau_0) + \varepsilon + 2\varepsilon E(\tau_0) + 2\varepsilon^2 \\ &\leq E(\tau_0) + (3 + 2E(\tau_0))\varepsilon. \end{aligned}$$

With $\varepsilon < \min\{1, \frac{\delta}{3+2E(\tau_0)}\}$ we have $e^\varepsilon(E(\tau_0) + \varepsilon) \leq E(\tau_0) + (3 + 2E(\tau_0))\varepsilon < E(\tau_0) + \delta$ and therefore a contradiction.

Note that $E(\tau_0) - \varepsilon \geq 0$ and therefore we have in the end $\varepsilon < \min\{E(\tau_0), 1, \frac{\delta}{3+2E(\tau_0)}\}$. Finally,

$$P(B_n^c) \leq P(\|\hat{\Lambda}_n - \Lambda\| > \varepsilon) + P(|E(\tau_0) - \hat{\tau}_{0,n}| > \varepsilon) \leq \frac{E(\|\hat{\Lambda}_n - \Lambda\|_\infty^2) + \text{Var}(\hat{\tau}_{0,n})}{\varepsilon^2}$$

□

In the following let $\tilde{C}_n(y, r) := \left\{ \left| \hat{\Lambda}_n(y) - \Lambda(y) \right| \leq r \right\}$, $r \in \mathbb{R}$, $n \in \mathbb{N}$. In particular,

$$C_n = \bigcap_{y \in [S, 0]} \tilde{C}_n(y, 1),$$

where the event C_n is defined as in the proof of Proposition 4.7.

Lemma 4.9. *Let $y \in [S, 0]$ and define $k := \frac{1}{2L}$. Then for all $t \in [y, y + k \wedge 0]$ and $\omega \in \tilde{C}_n(y, \frac{1}{2}) \cap \tilde{C}_n(y + k, \frac{1}{2})$:*

$$\left| \hat{\Lambda}_n(t) - \Lambda(t) \right| \leq 1.$$

Proof. W.l.o.g. let $y + k < 0$. If $y + k \geq 0$, then replace $y + k$ by 0 in the following considerations and the estimates are still valid.

Let $\omega \in \tilde{C}_n(y, \frac{1}{2}) \cap \tilde{C}_n(y + k, \frac{1}{2})$. It follows that $\left| \hat{\Lambda}_n(y, \omega) - \Lambda(y) \right| \leq \frac{1}{2}$ and $\left| \hat{\Lambda}_n(y + k, \omega) - \Lambda(y + k) \right| \leq \frac{1}{2}$. Moreover, it follows directly from the definition that $\hat{\Lambda}_n(y_1, \omega) \geq \hat{\Lambda}_n(y_2, \omega)$ if $y_1 \leq y_2$ and $y_1, y_2 \in [S, 0]$. Therefore, the following applies for any $t \in [y, y + k]$:

$$\begin{aligned} \hat{\Lambda}_n(t, \omega) - \Lambda(t) &\leq \hat{\Lambda}_n(y, \omega) + \Lambda(y) - \Lambda(y) - \Lambda(t) \\ &\leq \frac{1}{2} + L(t - y) \\ &\leq \frac{1}{2} + Lk = 1. \end{aligned}$$

Furthermore

$$\begin{aligned}\Lambda(t) - \hat{\Lambda}_n(t, \omega) &\leq \Lambda(t) - \Lambda(y+k) + \Lambda(y+k) - \hat{\Lambda}_n(y+k, \omega) \\ &\leq Lk + \frac{1}{2} = 1.\end{aligned}$$

□

Lemma 4.10. *Let $M := \min\{m \in \mathbb{N}_0 : S + m\frac{1}{2L} \geq 0\}$. Then*

$$P(C_n^c) \leq 4\Lambda(S)(M+1)\frac{1}{n}, \quad (4.14)$$

where $C_n = \left\{ \left| \hat{\Lambda}_n(y) - \Lambda(y) \right| \leq 1, \forall y \in [S, 0] \right\}$.

Proof. We have $M = \min\{m \in \mathbb{N}_0 : S + m\frac{1}{2L} \geq 0\}$. Define $y_m := S + m\frac{1}{2L}$ with $m \in \{0, 1, \dots, M-1\}$. Furthermore let $y_M := 0$. According to Lemma 4.9 it holds for all $t \in [S, 0]$ and $\omega \in \tilde{C}_n(y_0, \frac{1}{2}) \cap \tilde{C}_n(y_1, \frac{1}{2}) \cap \dots \cap \tilde{C}_n(y_M, \frac{1}{2})$:

$$\left| \hat{\Lambda}_n(t) - \Lambda(t) \right| \leq 1.$$

In particular, the following applies:

$$\begin{aligned}\tilde{C}_n\left(y_0, \frac{1}{2}\right) \cap \dots \cap \tilde{C}_n\left(y_M, \frac{1}{2}\right) &\subseteq \left\{ \left| \hat{\Lambda}_n(y) - \Lambda(y) \right| \leq 1, \forall y \in [S, 0] \right\} \\ &= C_n.\end{aligned}$$

Therefore, using $C_n^c \subseteq \left(\tilde{C}_n(y_0, \frac{1}{2}) \cap \tilde{C}_n(y_1, \frac{1}{2}) \cap \dots \cap \tilde{C}_n(y_M, \frac{1}{2}) \right)^c$ and the subadditivity of probability measures, we get

$$\begin{aligned}P(C_n^c) &\leq P\left(\left(\tilde{C}_n\left(y_0, \frac{1}{2}\right) \cap \tilde{C}_n\left(y_1, \frac{1}{2}\right) \cap \dots \cap \tilde{C}_n\left(y_M, \frac{1}{2}\right)\right)^c\right) \\ &\leq \sum_{i=0}^M P\left(\tilde{C}_n\left(y_i, \frac{1}{2}\right)^c\right) \\ &= \sum_{i=0}^M P\left(\left\{ \left| \hat{\Lambda}_n(y_i) - \Lambda(y_i) \right| > \frac{1}{2} \right\}\right) \\ &\leq \sum_{i=0}^M \frac{1}{(1/2)^2} \text{Var}(\hat{\Lambda}_n(y_i)) \\ &= \sum_{i=0}^M 4\Lambda(y_i)\frac{1}{n} \\ &\leq 4\Lambda(S)(M+1)\frac{1}{n}.\end{aligned}$$

□

5 Proof of Theorem 3.3

Let (π_n) be the sequence of thresholds chosen by the ILU algorithm. Recall that (π_n) is a policy in the sense of Definition 2.1. We need to show that there exists a constant C such that for all $\lambda \in \mathcal{M}(L)$ and $T \in \mathbb{N}$ we have $\mathcal{R}_\lambda^\pi(T) \leq C \ln(T + 1)$.

We estimate the expected optimality gaps $E_\lambda[\Delta_\lambda(\pi_n)]$ by using the optimality gaps entailed by the sequence (\hat{b}_n) . To this end we define for $n \geq 0$

$$r_\lambda(n) := E_\lambda[\Delta_\lambda(\hat{b}_n)].$$

Next we define a sequence $\sigma(j)$, $j \geq 0$, recursively as follows. We set $\sigma(1) = 0$ and for all $j \geq 1$

$$\sigma(j + 1) = \min\{n > \sigma(j) : \tau_{\pi_n}^n > 0\}.$$

Notice that $\sigma(j)$ represents the round where for the j -th time we have full information. Moreover, $E_\lambda(\sigma(j + 1) - \sigma(j))$ is exactly the expected waiting time until a round with full observation. The expected value is bounded from above by $E_\lambda(\sigma(j + 1) - \sigma(j)) \leq e^{L|S|}$. Finally, let \mathcal{I}_n denote the random set of rounds with full observation during the first n rounds of search. We have $\mathcal{I}_0 = \emptyset$ and $\mathcal{I}_n = \{\sigma(j) : \sigma(j) \leq n - 1\}$ for $n \geq 1$. Moreover, note that $|\mathcal{I}_n| = \max\{k : \sigma(k) \leq n - 1\}$ and $\sigma(|\mathcal{I}_n|) < n \leq \sigma(|\mathcal{I}_n| + 1)$.

The crucial observation now is that for all $n \geq 0$ we have

$$E_\lambda[\Delta_\lambda(\pi_n)] = E_\lambda[r_\lambda(|\mathcal{I}_n|)]. \quad (5.1)$$

Note that the equation for $n = 0$ is trivial, since $E_\lambda[\Delta_\lambda(\pi_0)] = E_\lambda[\Delta_\lambda(0)] = E_\lambda[\Delta_\lambda(\hat{b}_0)] = r_\lambda(0)$. In order to prove the equation for $n \geq 1$, we can assume that (\hat{b}_n) is independent of (π_n) and (\mathcal{I}_n) (define (\hat{b}_n) on an independent copy of the probability space). Then on the event $\{|\mathcal{I}_n| = i\}$ the threshold π_n has the same distribution as \hat{b}_i . Hence $E_\lambda[\Delta_\lambda(\pi_n)1_{\{|\mathcal{I}_n|=i\}}] = E_\lambda[\Delta_\lambda(\hat{b}_i)1_{\{|\mathcal{I}_n|=i\}}] = E_\lambda[\Delta_\lambda(\hat{b}_i)]P_\lambda(|\mathcal{I}_n| = i) = r_\lambda(i)P_\lambda(|\mathcal{I}_n| = i)$. Consequently,

$$E_\lambda[\Delta_\lambda(\pi_n)] = \sum_{i=0}^{n-1} r_\lambda(i)P_\lambda(|\mathcal{I}_n| = i) = E_\lambda[r_\lambda(|\mathcal{I}_n|)].$$

From (5.1) we get

$$\mathcal{R}_\lambda^\pi(T) = \sum_{n=0}^T E_\lambda[r_\lambda(|\mathcal{I}_n|)].$$

Now we use that λ belongs to the environment class $\mathcal{M}(L)$. This implies that λ is continuously differentiable and hence Δ_λ twice continuously differentiable (see Lemma 1.2). Therefore, using $\Delta_\lambda(b^*) = \Delta'_\lambda(b^*) = 0$, Taylor's theorem implies

$$\Delta_\lambda(b) = \frac{1}{2}\Delta''_\lambda(\theta)(b - b^*)^2 \quad (5.2)$$

for some θ between b and b^* .

Moreover, we define $c := \sup_{\lambda \in \mathcal{M}(L)} \sup_{b \in [S, 0]} \Delta''_\lambda(b)$. Since $\lambda \in \mathcal{M}(L)$ one can derive from Equation (1.5) that c depends on S and L only, and that $c < \infty$. Hence, for all λ and $b \in [S, 0]$ we have $\Delta_\lambda(b) \leq \frac{1}{2}c(b - b^*)^2$. Thus,

$$\sup_{\lambda \in \mathcal{M}(L)} r_\lambda(n) \leq \frac{1}{2}cE_\lambda[(\hat{b}_n - b^*)^2].$$

For notational convenience, let $L_{low} := \frac{\ln(2)}{|S|} + \frac{1}{L}$. Note that L_{low} is exactly the lower bound from property 2 of the environment class $\mathcal{M}(L)$ and only depends on the environment parameters S and L . The following estimates are valid for all $\lambda \in \mathcal{M}(L)$

1. $\Lambda(S) \leq |S|L$;
2. $\min\{m \in \mathbb{N}_0 : S + m\frac{1}{2L} \geq 0\} \leq \lceil 2|S|L \rceil \leq 2|S|L + 1$;
3. $E(\tau_0) \geq \frac{1}{L}$;
4. $\text{Var}(\tau_0) \leq 2L_{low}^2$.

The last estimate follows from the proof of Lemma 4.1. Moreover, using Lemma 3.2 we have

$$\frac{b^* - S}{3 + 2E(\tau_0)} \geq \frac{|S| - L_{low}^{-1} \ln(2)}{3 + 2L_{low}^{-1}}.$$

Therefore, Proposition 4.7 implies that $\sup_{\lambda \in \mathcal{M}(L)} r_\lambda(n) \leq \frac{1}{2}cD\frac{1}{n}$, where

$$D = \left[\left(S^2 \frac{4L|S| + 2L_{low}^2}{\left(\min\left\{ \frac{1}{L}, \frac{|S| - L_{low}^{-1} \ln(2)}{3 + 2L_{low}^{-1}} \right\} \right)^2} \right) + (4L_{low}^2 + 8|S|^3 L e^{2|S|L} + |S|^3 4L(2|S|L + 2)) \right] \frac{1}{n}.$$

Therefore, for all $\lambda \in \mathcal{M}(L)$ we have

$$\begin{aligned}
\mathcal{R}_\lambda^\pi(T) &= E_\lambda \left[\sum_{n=0}^T r_\lambda(|\mathcal{I}_n|) \right] \leq E_\lambda(\tau_0^0) + E_\lambda \left[\sum_{n=1}^T r_\lambda(|\mathcal{I}_n|) \right] \\
&\leq L_{low}^{-1} + E_\lambda \left[\sum_{j=1}^{|\mathcal{I}_T|} (\sigma(j+1) - \sigma(j)) r_\lambda(j) \right] \\
&\leq L_{low}^{-1} + \sum_{j=1}^T E_\lambda[\sigma(j+1) - \sigma(j)] r_\lambda(j) \\
&\leq L_{low}^{-1} + e^{L|S|} \sum_{j=1}^T r_\lambda(j) \\
&\leq L_{low}^{-1} + e^{L|S|} \frac{1}{2} c D \ln(T+1) \\
&\leq L_{low}^{-1} \frac{\ln(T+1)}{\ln(2)} + e^{L|S|} \frac{1}{2} c D \ln(T+1) \\
&= \left(\frac{L_{low}^{-1}}{\ln(2)} + e^{L|S|} \frac{1}{2} c D \right) \ln(T+1).
\end{aligned}$$

Note that the RHS of the previous inequality does not depend on λ . Thus we have shown the theorem.

6 Lower bound: Proof of Theorem 3.4

In the following we prove Theorem 3.4. We need to show that the minimax regret

$$\inf_{\pi \in \Pi} \sup_{\lambda \in \mathcal{M}(L)} \mathcal{R}_\lambda^\pi(T)$$

grows at least logarithmically. The crux of the argument is that it is enough to derive a logarithmic lower bound for $\mathcal{H}(L) := \{\lambda \in \mathcal{M}(L) : \lambda \text{ constant}\}$, the subclass of constant intensity functions. Indeed, note that

$$\inf_{\pi \in \Pi} \sup_{\lambda \in \mathcal{M}(L)} \mathcal{R}_\lambda^\pi(T) \geq \inf_{\pi \in \Pi} \sup_{\lambda \in \mathcal{H}(L)} \mathcal{R}_\lambda^\pi(T).$$

Therefore, a logarithmic lower bound $\mathcal{H}(L)$ implies a lower bound for the whole environment class $\mathcal{M}(L)$.

Note that if the unknown intensity function is from the class $\mathcal{H}(L)$, then the stopping agent observes a homogeneous Poisson processes in each round.

The estimation of the intensity function is thus simply a one parameter estimation, namely the estimation of $\lambda \in [a, b]$, where $a := \frac{\ln(2)}{|S|} + \frac{1}{L}$ and $b := L$.

In addition, we assume that we have full information in every round, i.e. that we can observe the homogeneous Poisson process on $[S, 0]$ in every round. Since the class of policies with full information is larger than the class of policies with partial information, we have

$$\inf_{\pi \in \Pi} \sup_{\lambda \in \mathcal{H}(L)} \mathcal{R}_\lambda^\pi(T) \geq \inf_{\pi \in \Pi^{full}} \sup_{\lambda \in \mathcal{H}(L)} \mathcal{R}_\lambda^\pi(T). \quad (6.1)$$

Moreover, to simplify the following analysis we introduce the class of so-called cut-off full information policies

$$\Pi^{cut} := \left\{ \pi^{cut} = (\pi_n^{cut})_{n \geq 1} \mid \exists \pi^{full} \in \Pi^{full} \text{ s.t. } \forall n \geq 0 : \pi_n^{cut} = \min\{\pi_n^{full}, -\frac{\ln(2)}{L}\} \right\}.$$

The idea is that for $\lambda \in \mathcal{H}(L)$ we know that $b^* \in (S, -\frac{\ln(2)}{L}]$ by Corollary 1.3. Therefore, cutting off every policy at $-\frac{\ln(2)}{L}$ can not worsen the regret. Let $\pi \in \Pi^{full}$ and π^{cut} the corresponding cut-off policy. Fix $n \in \mathbb{N}_0$.

case 1: We have $\pi_n < -\frac{\ln(2)}{L}$. Therefore

$$\Delta(\pi_n) = E|\tau_{\pi_n}| - E|\tau_{b^*}| = E|\tau_{\pi_n^{cut}}| - E|\tau_{b^*}| = \Delta(\pi_n^{cut})$$

case 2: We have $\pi_n \in [L^*, 0]$, where $L^* := -\frac{\ln(2)}{L}$. By Theorem 1.3 we know the cost function for $\lambda \in \mathcal{H}(L)$ is $\Delta_\lambda(b) = \frac{1}{\lambda}(2e^{\lambda b} - 1) - b - \frac{\ln(2)}{\lambda}$ for $b \in [S, 0]$. We want to show that $\Delta_\lambda(x) \geq \Delta_\lambda(L^*)$ for all $x \in [L^*, 0]$. We have $\Delta'_\lambda(x) = 2e^{\lambda x} - 1$. Recall that $\lambda \in [\ln(2)/|S| + \frac{1}{L}, L]$ and $x \in [L^*, 0]$. It follows directly $\lambda x \geq -\ln(2)$ and therefore $\Delta'_\lambda(x) \geq 0$ for all $x \in [L^*, 0]$.

It follows

$$\Delta_\lambda(\pi_n) \geq \Delta_\lambda(\pi_n^{cut}).$$

In summary, the regret rate does not get worse by substituting policies by cut-off policies and hence we get

$$\inf_{\pi \in \Pi} \sup_{\lambda \in \mathcal{H}(L)} \mathcal{R}_\lambda^\pi(T) \geq \inf_{\pi \in \Pi^{full}} \sup_{\lambda \in \mathcal{H}(L)} \mathcal{R}_\lambda^\pi(T) = \inf_{\pi \in \Pi^{cut}} \sup_{\lambda \in \mathcal{H}(L)} \mathcal{R}_\lambda^\pi(T).$$

Therefore a lower bound for cut-off policies is valid for the policies from Definition 2.1 as well.

When estimating a constant jump intensity, observing $(N_t^j)_{t \in [S, 0]}$ is equivalent, in the sense of sufficiency, to observing $N_0^j \sim Poi(\lambda|S|)$. Therefore, in the following we can assume that any estimator of the parameter λ makes solely use of the values $N_0^0, N_0^1, N_0^2, \dots$. Note that the latter sequence is *i.i.d.*

and Poisson distributed with parameter $\lambda|S|$. W.l.o.g. we choose $|S| = 1$. Otherwise just replace λ by $\tilde{\lambda} := |S|\lambda \in [|S|a, |S|b]$ in the following considerations.

Recall from Lemma 1.3 that for any constant λ we have $E_\lambda(|\tau_{b^*}|) = \frac{\ln(2)}{\lambda}$. There is a one-to-one correspondence between the optimal threshold b^* and the given constant intensity λ . Indeed, we have $b^*(\lambda) = -\frac{\ln(2)}{\lambda}$ and $\lambda(b^*) = -\frac{\ln(2)}{b^*}$. We use this correspondence to reduce the problem of finding a minimizing cut-off policy to a problem of determining an estimator for λ with minimal mean square error.

Proposition 6.1. There exists a constant $c \in (0, \infty)$ such that for all $x \in [S, 0]$ and $\lambda \in \mathcal{H}(L)$ we have

$$\Delta_\lambda(x) \geq c(x - b^*(\lambda))^2.$$

Proof. We make a Taylor expansion of $\Delta_\lambda(x)$ around the optimal threshold $b^*(\lambda)$. Note that all necessary smoothness conditions are fulfilled and therefore we get

$$\Delta_\lambda(x) = \Delta_\lambda(b^*(\lambda)) + \Delta'_\lambda(b^*(\lambda))(x - b^*(\lambda)) + \Delta''_\lambda(\nu)(x - b^*(\lambda)),$$

where $\nu \in [\min\{x, b^*(\lambda)\}, \max\{x, b^*(\lambda)\}]$. By definition and the first order optimality criteria we know $\Delta_\lambda(b^*(\lambda)) = \Delta'_\lambda(b^*(\lambda)) = 0$. Moreover, for $\lambda \in \mathcal{H}(L)$ we know that $\Delta_\lambda(x) = \frac{1}{\lambda}(2e^{\lambda x} - 1) - x - \frac{\ln(2)}{\lambda}$. It follows $\Delta''_\lambda(x) = 2\lambda e^{\lambda x} > 0$ for all x . Therefore, we have

$$\Delta_\lambda(x) \geq c(x - b^*(\lambda))^2,$$

where $c := \inf_{\lambda \in [a, b]} \inf_{x \in [S, 0]} \Delta''_\lambda(x) > 0$. □

Proposition 6.1 implies that for all $(\pi_n) \in \Pi^{cut}$ we have

$$\begin{aligned} E_\lambda(\Delta_\lambda(\pi_n)) &\geq cE_\lambda[(\pi_n - b^*(\lambda))^2] = cE_\lambda \left[\left(b^*\left(-\frac{\ln(2)}{\pi_n}\right) - b^*(\lambda) \right)^2 \right] \\ &\geq \tilde{c}E_\lambda \left[\left(-\frac{\ln(2)}{\pi_n} - \lambda \right)^2 \right], \end{aligned} \tag{6.2}$$

where $\tilde{c} := c \min_{\lambda \in [a, b]} (b^*)'(\lambda) = c \frac{\ln(2)}{L^2}$.

We can interpret $-\frac{\ln(2)}{\pi_n}$ as the corresponding estimator for the parameter λ defined by the policy $(\pi_n) \in \Pi^{cut}$. It follows directly by (6.2) that

$$\mathcal{R}_\lambda^\pi(T) = \sum_{n=1}^T E_\lambda(\Delta_\lambda(\pi_n)) \geq \tilde{c} \sum_{n=1}^T E_\lambda \left[\left(-\frac{\ln(2)}{\pi_n} - \lambda \right)^2 \right]$$

and therefore

$$\inf_{\text{policies } \pi} \sup_{\lambda \in \mathcal{H}(L)} \mathcal{R}_\lambda^\pi(T) \geq \inf_{\pi \in \Pi^{\text{cut}}} \sup_{\lambda \in \mathcal{H}(L)} \mathcal{R}_\lambda^\pi(T) \geq \inf_{\text{estimator } \hat{\lambda}} \sup_{\lambda \in [a,b]} \tilde{c} \sum_{n=1}^T E_\lambda [(\hat{\lambda}_n - \lambda)^2].$$

In summary, we have reduced the problem of deriving a lower bound for the minimax regret on the environment class $\mathcal{M}(L)$ to a problem of deriving a lower bound for the minimax risk when estimating $\lambda \in \mathcal{H}(L)$ using the observations $N_0^0, N_0^1, N_0^2, \dots$

Next observe that the minimax risk can be bounded from below with the Bayes risk with prior density q

$$\begin{aligned} \inf_{\text{estimator } \hat{\lambda}} \sup_{\lambda \in [a,b]} E_\lambda \left[\sum_{n=1}^T (\hat{\lambda}_n - \lambda)^2 \right] &\geq \inf_{\text{estimator } \hat{\lambda}} \int_{[a,b]} E_\lambda \left[\sum_{n=1}^T (\hat{\lambda}_n - \lambda)^2 \right] q(\lambda) d\lambda \\ &= \inf_{\text{estimator } \hat{\lambda}} \sum_{n=1}^T \int_{[a,b]} E_\lambda \left[(\hat{\lambda}_n - \lambda)^2 \right] q(\lambda) d\lambda. \end{aligned} \tag{6.3}$$

Recall that the the density of the $Beta(3, 3)$ -distribution on $[0, 1]$ is given by $q_{[0,1]}(x) = 30x^2(1-x)^2$. It is straightforward to show that $q_{[a,b]}(x) = \frac{1}{b-a} q_{[0,1]}(\frac{x-a}{b-a}) = \frac{30}{(b-a)^5} (x-a)^2 (b-x)^2$ is density function on $[a, b]$. In the following we choose the prior with density $q = q_{[a,b]}$.

Now, we apply the so-called van Trees inequality to the RHS of (6.3). This inequality can be found, for example, in Chapter 2 in [9]. For the choice of the specific prior with density $q = q_{[a,b]}$ all conditions for the application of the van Trees inequality are satisfied, especially the condition $q(a) = q(b) = 0$, and hence we get

$$\inf_{\text{estimator } \hat{\lambda}} \sum_{n=1}^T \int_{[a,b]} E_\lambda \left[(\hat{\lambda}_n - \lambda)^2 \right] q(\lambda) d\lambda \geq \sum_{n=1}^T \frac{1}{\mathbb{I}_{q_{[a,b]}} + \int_{[a,b]} \mathbb{I}_{p^{(n)}}(\lambda) q(\lambda) d\lambda},$$

where $\mathbb{I}_{q_{[a,b]}} := \int \left[\frac{\partial}{\partial \lambda} \ln(q(\lambda)) \right]^2 q(\lambda) d\lambda$ and $\mathbb{I}_{p^{(n)}}(\lambda)$ denotes the Fisher information of the n -fold product of the Poisson distribution. Recall that $\mathbb{I}_{p^{(n)}} = n \mathbb{I}_{p^{(1)}} = \frac{n}{\lambda}$.

To derive the value of $\mathbb{I}_{q_{[a,b]}}$ we first calculate $\mathbb{I}_{q_{[0,1]}}$. Note that $\mathbb{I}_{q_{[0,1]}} = \int \frac{q'(\lambda)^2}{q(\lambda)} d\lambda$ and hence

$$\mathbb{I}_{q_{[0,1]}} = \int_0^1 \frac{60^2 x^2 (1-x)^2 (1-2x)^2}{30x^2 (1-x)^2} dx = 120 \int_0^1 (1-2x)^2 dx = 40.$$

Next observe that $q'_{[a,b]}(x) = \frac{1}{b-a}q'_{[0,1]}(\frac{x-a}{b-a})\frac{1}{b-a}$ and thus

$$\begin{aligned}\mathbb{I}_{q_{[a,b]}} &= \int_a^b \frac{\frac{1}{(b-a)^4}(q'_{[0,1]}(\frac{x-a}{b-a}))^2}{\frac{1}{b-a}q_{[0,1]}(\frac{x-a}{b-a})} dx \\ &= \int_0^1 \frac{1}{(b-a)^3} \frac{q'_{[0,1]}(u)}{q_{[0,1]}(u)} (b-a) du \\ &= \frac{40}{(b-a)^2},\end{aligned}$$

where we substitute $u = \frac{x-a}{b-a}$. To sum up, we get

$$\begin{aligned}\sum_{n=1}^T \frac{1}{\mathbb{I}_{q_{[a,b]}} + \int_{[a,b]} \mathbb{I}_{p^{(n)},n}(\lambda)q(\lambda)d\lambda} &= \sum_{n=1}^T \frac{1}{\frac{40}{(b-a)^2} + n \int_{[a,b]} \frac{1}{\lambda} q(\lambda)d\lambda} \\ &\geq \sum_{n=1}^T \frac{1}{\frac{40}{(b-a)^2} + n \frac{1}{a} \int_{[a,b]} q(\lambda)d\lambda} \\ &\geq C' \ln(T),\end{aligned}$$

where $C' = \frac{1}{\frac{40}{(b-a)^2} + \frac{1}{a}}$. Therefore

$$\tilde{c} \inf_{\text{estimator}} \sup_{\hat{\lambda} \in [a,b]} E_{\lambda} \left[\sum_{n=1}^T (\hat{\lambda}_n - \lambda)^2 \right] \geq \tilde{c} C' \ln(T) = C \ln(T),$$

where $C = \tilde{c} C'$. In summary,

$$\inf_{\pi \in \Pi} \sup_{\lambda \in \mathcal{M}(L)} \mathcal{R}_{\lambda}^{\pi}(T) \geq \inf_{\text{estimator}} \sup_{\hat{\lambda} \in [a,b]} \tilde{c} E_{\lambda} \left[\sum_{n=1}^T (\hat{\lambda}_n - \lambda)^2 \right] \geq C \ln(T).$$

Thus Theorem 3.4 is proved.

Acknowledgements

We thank Nabil Kazi-Tani and Michael Neumann for very fruitful discussions. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 547236699.

References

- [1] S. Ankirchner, S. Christensen, J. Kallsen, P. L. Borne, and S. Perko. Learning to steer with Brownian noise. *arXiv preprint arXiv:2410.03221*, 2024.
- [2] M. Basei, X. Guo, A. Hu, and Y. Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Journal of Machine Learning Research*, 23(178):1–34, 2022.
- [3] S. Christensen, N. Dexheimer, and C. Strauch. Data-driven optimal stopping: A pure exploration analysis. *arXiv preprint arXiv:2312.05880*, 2023.
- [4] S. Christensen and C. Strauch. Nonparametric learning for impulse control problems—Exploration vs. exploitation. *The Annals of Applied Probability*, 33(2):1569 – 1587, 2023.
- [5] S. Christensen, C. Strauch, and L. Trottner. Learning to reflect: a unifying approach for data-driven stochastic control strategies. *Bernoulli*, 30(3):2074–2101, 2024.
- [6] S. Christensen, A. H. Thomsen, and L. Trottner. Data-driven rules for multidimensional reflection problems. *to appear in SIAM / ASA Journal on Uncertainty Quantification*, *arXiv preprint arXiv:2311.06639*, 2024.
- [7] T. E. Duncan, L. Guo, and B. Pasik-Duncan. Adaptive continuous-time linear quadratic gaussian control. *IEEE Transactions on automatic control*, 44(9):1653–1662, 2002.
- [8] T. S. Ferguson. Optimal stopping and applications, 2006. *URL <https://www.math.ucla.edu/~tom/Stopping/Contents.html>*, 2018.
- [9] R. D. Gill and B. Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. 1995.
- [10] X. Guo, A. Hu, and Y. Zhang. Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls, 2021.
- [11] T. M. Moerland, J. Broekens, A. Plaat, and C. M. Jonker. Model-based reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 16(1):1–118, 2023.

- [12] J. MacQueen and R. G. Miller, Jr. Optimal persistence policies. *Operations Res.*, 8:362–380, 1960.
- [13] E. Neuman and Y. Zhang. Statistical learning with sublinear regret of propagator models, 2023.
- [14] M. Sakaguchi and M. Tamaki. On the optimal parking problem in which spaces appear randomly. 1982.
- [15] L. Szpruch, T. Treetanthiploet, and Y. Zhang. Exploration-exploitation trade-off for continuous-time episodic reinforcement learning with linear-convex models. *arXiv preprint arXiv:2112.10264*, 2021.