On Learning the Optimal Waiting Time

Tor Lattimore[1] and **András György**[1] and **Csaba Szepesvári**[1,2]

[1]Department of Computing Science, University of Alberta, Canada
[2]Microsoft Research, Redmond
tor.lattimore@gmail.com,{gyorgy,szepesva}@ualberta.ca

Abstract. Consider the problem of learning how long to wait for a bus before walking, experimenting each day and assuming that the bus arrival times are independent and identically distributed random variables with an unknown distribution. Similar uncertain optimal stopping problems arise when devising power-saving strategies, e.g., learning the optimal disk spin-down time for mobile computers, or speeding up certain types of satisficing search procedures by switching from a potentially fast search method that is unreliable, to one that is reliable, but slower. Formally, the problem can be described as a repeated game. In each round of the game an agent is waiting for an event to occur. If the event occurs while the agent is waiting, the agent suffers a loss that is the sum of the event's "arrival time" and some fixed loss. If the agent decides to give up waiting before the event occurs, he suffers a loss that is the sum of the waiting time and some other fixed loss. It is assumed that the arrival times are independent random quantities with the same distribution, which is unknown, while the agent knows the loss associated with each outcome. Two versions of the game are considered. In the full information case the agent observes the arrival times regardless of its actions, while in the partial information case the arrival time is observed only if it does not exceed the waiting time. After some general structural observations about the problem, we present a number of algorithms for both cases that learn the optimal weighting time with nearly matching minimax upper and lower bounds on their regret.

1 Introduction

Each day a student travels to school, either by bus or on foot, whichever is faster. The expected travel time for the bus is five minutes and is denoted by β while walking takes twenty minutes and is denoted by ω. Unfortunately, the bus is not always on time, so on each day t the student must decide how long he wants to wait for the bus, Y_t. The bus comes at random time X_t and if $X_t \leq Y_t$, then the student catches the bus. If $X_t > Y_t$, then they walk. The loss at time step t is the total travel time, which the student wants to keep as small as possible and is defined by

$$\ell_t(Y_t) \doteq \begin{cases} X_t + \beta, & \text{if } X_t \leq Y_t\,; \\ Y_t + \omega, & \text{otherwise}\,. \end{cases}$$

	$p(x)$	$\ell(y)$	y^*	$\ell(y^*)$
Power	$\frac{\mathbb{1}\{x\geq 1\}}{x^2}$	$\beta+1+\frac{\omega-\beta}{y}+\log\, y$	$\omega-\beta$	$\beta+2+\log(\omega-\beta)$
Exponential	$\lambda e^{-\lambda x}$	$\beta+\frac{1}{\lambda}+e^{-\lambda y}\left[(\omega-\beta)-\frac{1}{\lambda}\right]$	0, if $\omega-\beta<\frac{1}{\lambda}$; ∞, otherwise	ω, if $\omega-\beta<\frac{1}{\lambda}$; $\beta+\frac{1}{\lambda}$, otherwise

Table 1: Examples of losses and optimal waiting times for particular choices of arrival distributions. The first column gives the density of the arrival times.

We assume that $\beta < \omega$ and that the arrival times for the bus $(X_t)_t$ are identically and independently distributed according to some unknown distribution over the positive real line. We make no additional assumptions on the distribution of arrival times.[1] The expected loss of a fixed deterministic waiting time $y \geq 0$ is

$$\ell(y) = \mathbb{E}\left[\mathbb{1}\{X \leq y\}\,(X+\beta) + \mathbb{1}\{X > y\}\,(y+\omega)\right]\,,$$

where X is identically distributed to X_t.[2] An optimal waiting time is given by $y^* \doteq \arg\min_{y\geq 0} \ell(y)$, which we will show to exist (y^* may be infinite and the optimal waiting may not be unique, as we will demonstrate below).

To guide the reader's intuition, in Table 1 we tabulate the loss function, the optimal action, and the loss of the optimal action for two particular arrival time distributions. The examples show that the loss may be convex, or concave, it can be unbounded and the optimal action can also take on any value between 0 and infinity. These examples should not mislead the reader. Our methods do not need to know the form of the arrival time distributions, i.e., we consider the nonparametric setting.

Since the distribution of arrival times is unknown, the student cannot know when to stop waiting and must experiment to gain information. The regret at time step t is the difference between the actual travel time and the travel time under an optimal waiting time, $r_t \doteq \ell_t(Y_t) - \ell_t(y^*)$. Note that r_t may be negative, but has non-negative expectation, and that r_t does not depend on the choice of y^*. The cumulative regret until time step n is

$$R_n \doteq \sum_{t=1}^{n} \Big(\ell_t(Y_t) - \ell_t(y^*)\Big).$$

In the long run the student hopes to choose Y_t in such a way as to learn the optimal waiting time, in which case $\lim_{n\to\infty} \mathbb{E}[R_n]/n = 0$.

Two observation models will be considered. The first is a *full information setting* where X_t is always observed. This assumption is unnatural for the problem of waiting for a bus because the student would not usually observe the arrival time of the bus if they decided to walk. There are, however, waiting problems for which the full information setting is appropriate. An example is maximising hard-disk efficiency in mobile computing, previously considered by Krishnan,

[1] Note that the game is trivial if $\beta \geq \omega$, since in this case the student should always walk regardless of the expected arrival time of the bus.

[2] One can show that the expected loss is minimized by a fixed deterministic waiting time, i.e., there is no advantage to using a stopping rule. The simple reason is that when the bus arrives, due to our assumption that $\omega > \beta$ it is better to take the bus then to continue waiting and then eventually walk (since no more buses are coming).

Long and Vitter [12] where a hard-disk controller must decide after each interaction how long to wait before spinning down the disk to conserve energy. This is modelled by choosing $\beta = 0$ and ω to be some value that reflects the cost (in terms of time/energy/annoyance) of spinning up the disk. The goal of the controller is to minimise the sum of energy consumption and spin-up costs.

The second setting, called the *partial information setting*, is trickier, but often more natural, e.g., for the bus-stop problem, when the student in general will not observe X_t unless $X_t \leq Y_t$. More precisely, the student observes the pair (Z_t, δ_t) where $Z_t = \min\{X_t, Y_t\}$ and $\delta_t = \mathbb{1}\{X_t \leq Y_t\}$. So δ_t is 1 if the student travelled by bus and 0 otherwise, while Z_t is the time at which the travel starts. Another application of the partial information setting is the problem of combining algorithms to solve a number of instances of a satisficing search problems.[3] We assume that the agent has access to two algorithms for a given type of search problem. The first is potentially fast, but unreliable, while the second is typically slower, but has known guarantees on its performance. For each problem instance, the agent tries to use the potentially fast solver, switching to the more consistent algorithm if the first fails to deliver a solution within a certain amount of time. The task of the agent is to learn when to switch between solvers. Formally, the unreliable solver provides a solution to instance t at random time X_t, and the completion time of the slower method is deterministically ω for all instances. As for the hard-disk problem, $\beta = 0$. Comparisons between stochastic satisficing search algorithms have been made before (e.g., [16] and references there-in), but to our knowledge the sequential setting combined with the regret criterion are new.

Estimating the common distribution of X_t (or other quantities depending on this distribution) in the partial information (or "censored") setting is heavily studied in the statistics literature [e.g. 6], but the focus tends to be on the natural medical applications where the censoring times are uncontrolled and independent of the arrival times. We know of no previous work on the decision problem studied here. Optimising the regret is more complex when only partial information is available because in this case the actions influence the observations.

The censored information problem is an instance of stochastic partial monitoring, first studied by Agrawal, Teneketzis and Anantharam [1]. In recent years there has been significant progress towards understanding partial monitoring with finitely many actions, both in the stochastic and adversarial settings [4, 9, 3], but the case where the number of actions is infinite/continuous the work has been more limited and specialised [11].

Summary of Results. The full information setting is analysed in Section 3 where we present two algorithms. The first is based on discretising the action space and applying the exponential weighting algorithm (EWA), while the second is an instance of the Follow-the-Leader (FTL) algorithm. We prove that EWA suffers a regret of at most $O(\log^{3/2}(n)\sqrt{n})$ while for FTL we were able to shave

[3] A search problem is satisficing if the searcher can stop once a satisfactory solution has been found, with SAT being a prototypical example.

off a small amount and bound the regret by $O(\log(n)\sqrt{n})$. We also establish a lower bound of $\Omega(\sqrt{n})$.

For the partial information case we also consider two algorithms (Section 4). The first is again based on a variant of the exponential weights algorithm, which cleverly controls the exploration of actions to deal with the partial information setting [2]. We establish that this algorithm enjoys a regret of $O(\log^2(n)\sqrt{n})$. Next we propose a novel optimistic algorithm that conservatively waits for the longest time that it cannot prove to be sub-optimal with high probability. We prove that this algorithm enjoys a regret of $O(\log^{3/2}(n)\sqrt{n})$. Thus, for both algorithms, the cost of partial information is surprisingly small and of order $O(\log^{1/2}(n))$. Some proofs have been omitted or sketched, but complete versions may be found in our report [13].

The theoretical findings are complemented by computer simulations in a variety of controlled scenarios (Section 5). Results are presented for the full information setting only (similar results were observed in the partial information setting). The most interesting finding here is that for the exponential distribution, both algorithms perform better than is predicted by theory, with at least FTL achieving $O(\log n)$ regret. The EWA algorithm behaves comparably to FTL, but only when the learning rate is tuned to be much larger than is theoretically justified.

Notation. At time step t define the empirical probability measure by $\mathbb{P}_t\{A\} \doteq \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}\{X_s \in A\}$ where A is any Borel-measurable subset of the real line. The cumulative distribution of the samples $X_1, \ldots, X_t$ is $F_t(x) = \mathbb{P}_t\{(-\infty, x]\}$. Expectations with respect to the empirical distribution $\mathbb{P}_t\{\cdot\}$ are denoted by $\mathbb{E}_t\{\cdot\}$. Further, by slightly abusing the notation for any measurable function $f : \mathbb{R} \to \mathbb{R}$, we define $\mathbb{E}_t[f(X)] = \int f(x) d\mathbb{P}_t(x)$ and for any Borel measurable subset A of the real line, $\mathbb{P}_t\{X \in A\} = \mathbb{E}_t[\mathbb{1}\{X \in A\}]$.

2 Structure of the Waiting Problem

Before the main theorems we present a crucial lemma that characterises the cumulative distribution of the arrival times in terms of the optimal action y^*. The result shows that the tail of X decays exponentially for times before y^*. As a consequence, if the optimal waiting time is large then the loss of choosing y much smaller than optimal cannot be too large. This latter fact should not be surprising. If it is optimal to wait for the bus for a very long time, then there must be a reasonable probability that it will arrive soon. This means that the bus is still likely to arrive if you wait for a shorter time. The critical case occurs when arrival times are exponentially distributed. As a result, it is not hard to see that to achieve a polynomially decreasing regret in n time steps, it is enough to consider waiting times below some $O(\log n)$ threshold.

Lemma 1. *Let $0 < \hat{y} \leq \tilde{y}$ such that $\inf_{y \in [0,\hat{y}]} \ell(y) \geq \ell(\tilde{y})$. Then, the following hold true for any $y \in [0, \hat{y}]$:*

1. $\mathbb{P}\{X > y\} \leq 2^{-\left\lfloor \frac{y}{2(\omega-\beta)} \right\rfloor}$.

2. *$\ell(y) - \ell(\hat{y}) \le (\omega - \beta) 2^{-\left\lfloor \frac{y}{2(\omega-\beta)} \right\rfloor}$.*

In particular, if $y^* \in [0, \infty]$ is optimal, then the above holds with $\tilde{y} = \hat{y} = y^*$.

The proof of the lemma utilizes the following bounds on loss differences, which will also be useful later and follows trivially from the definitions. The proof may be found in the technical report [13].

Lemma 2. *Let $y_2 \ge y_1$, then*

1. *$\ell(y_2) - \ell(y_1) = \mathbb{E}[\mathbb{1}\{y_1 < X \le y_2\}(X - y_1 + \beta - \omega) + \mathbb{1}\{X > y_2\}(y_2 - y_1)]$.*
2. *$\ell(y_2) - \ell(y_1) \ge (y_2 - y_1)(1 - F(y_2)) - (\omega - \beta)(F(y_2) - F(y_1))$.*
3. *$\ell(y_2) - \ell(y_1) \le (y_2 - y_1)(1 - F(y_1)) - (\omega - \beta)(F(y_2) - F(y_1))$.*

Proof (Lemma 1). Let $c \ge 0$ be some constant to be chosen later and $0 \le y \le \hat{y} - c$. Then we have

$$\begin{aligned}
0 &\overset{(a)}{\le} \ell(y) - \ell(\tilde{y}) \overset{(b)}{=} \mathbb{E}[(y + \omega - \beta - X)\mathbb{1}\{y < X \le \tilde{y}\} + (y - \tilde{y})\mathbb{1}\{X > \tilde{y}\}] \\
&\overset{(c)}{\le} \mathbb{E}[(y + \omega - \beta - X)\mathbb{1}\{y < X \le y + c\} + (\omega - \beta - c)\mathbb{1}\{X > y + c\}] \\
&\overset{(d)}{\le} (\omega - \beta)\mathbb{E}[\mathbb{1}\{y < X \le y + c\} + (\omega - \beta - c)\mathbb{1}\{X > y + c\}] \\
&\overset{(e)}{=} (\omega - \beta)\mathbb{P}\{y < X \le y + c\} + (\omega - \beta - c)\mathbb{P}\{X > y + c\} \qquad (1) \\
&\overset{(f)}{\le} (\omega - \beta)\mathbb{P}\{y < X\}\,, \qquad (2)
\end{aligned}$$

where (a) follows since $\ell(\tilde{y}) \le \ell(y)$ by assumption, (b) follows from Part 1 of Lemma 2, (c) follows by breaking $\mathbb{1}\{y + c < X \le \tilde{y}\}$ off from both indicators and since $y \le \hat{y} - c \le \tilde{y} - c$, while (d) is true by noting that $y + \omega - \beta - X \le \omega - \beta$ for $y \le X$. (e) and (f) are trivial. Choosing $c = 2(\omega - \beta) > 0$, (1) implies

$$\mathbb{P}\{y < X \le y + 2(\omega - \beta)\} \ge \mathbb{P}\{X > y + 2(\omega - \beta)\}\,.$$

Therefore, for any $y \ge 0$ such that $y + 2(\omega - \beta) \le \hat{y}$,

$$\mathbb{P}\{X \le y + 2(\omega - \beta) | X > y\} \ge \frac{1}{2} \qquad (3)$$

and if $2k(\omega - \beta) \le \hat{y}$, then

$$\begin{aligned}
&\mathbb{P}\{X > 2k(\omega - \beta)\} \overset{(a)}{=} \prod_{i=1}^{k} \mathbb{P}\{X > 2i(\omega - \beta) | X > 2(i-1)(\omega - \beta)\} \\
&\overset{(b)}{=} \prod_{i=1}^{k} (1 - \mathbb{P}\{X \le 2i(\omega - \beta) | X > 2(i-1)(\omega - \beta)\}) \overset{(c)}{\le} 2^{-k}\,, \qquad (4)
\end{aligned}$$

where (a) follows from the chain rule for probability, (b) is just $\mathbb{P}\{A|B\} = 1 - \mathbb{P}\{A^c|B\}$ for events A and B and (c) follows by substituting (3), which is permitted thanks to $2k(\omega - \beta) \le \hat{y}$. The above inequality immediately implies Part 1 (for $y \le \hat{y} < 2k(\omega - \beta)$ the result holds trivially) and, combined with (2) for $c = 0$, it also yields Part 2. □

That an optimal waiting time is guaranteed to exist follows from Lemma 1 by a tedious case-based analysis. See the technical report for the proof [13].

Theorem 1. *For any arrival time distribution there exists a* $y^* \in [0, \infty]$ *such that* $\ell(y^*) = \inf_{y \in [0,\infty]} \ell(y)$*.*

Part 2 of Lemma 1 also shows that to guarantee an ε-optimal action, it suffices to consider the waiting times in an interval of length $O(\log(1/\varepsilon))$ starting at zero:

Corollary 1. *Let* $\varepsilon > 0$ *and* $\bar{y}(\varepsilon) \doteq 2(\omega - \beta) \max \left\{ 1 + \log_2(\frac{\omega-\beta}{\varepsilon}),\, 0 \right\}$*. Then* $\inf_{y \in [0, \bar{y}(\varepsilon)]} \ell(y) - \ell(y^*) < \varepsilon$*.*

Proof. The result follows immediately from Part 2 of Lemma 1.

3 Full Information Setting

We consider the case when X_t is always observed in round t. Our first algorithm discretises the set of actions and then applies the exponential weighting algorithm [e.g., 5]. The key observation is that by Corollary 1, to guarantee an ε-optimal action, it suffices to play in the interval of length $O(\log(1/\varepsilon))$. Since the exponential weights algorithm assumes a finite action set, we need to discretise the action space. The following elementary observation, which follows directly from Part 1 of Lemma 2 shows that to achieve an ε-error, it suffices to discretise the interval with an accuracy of ε.

Proposition 1. *For any* $y_2 \geq y_1 \geq 0$*,* $\ell(y_2) - \ell(y_1) \leq y_2 - y_1$*.*

The exponential weights algorithm enjoys a regret smaller than $R\sqrt{n \log(K)/2}$, where n is the number of rounds, K is the number of actions, and R is the range of losses [5, §4.2, Thm 2.2]. So we see that this method suffers a regret of at least $O(\sqrt{n})$. This suggests choosing $\varepsilon = (\omega - \beta)/\sqrt{n}$ and using the action set $A = \{k\varepsilon : 0 \leq k \leq \bar{y}(\varepsilon)/\varepsilon, k \in \mathbb{N}\}$, leading to Algorithm 1, where for tuning the learning rate η we use that the range of the loss function is $m + \omega$ when the largest waiting time is $m = \max A$. The running time of the algorithm is $O(|A|)$ per time step, which in this case is $O(\sqrt{n} \log(n))$.

Algorithm 1 EWA for Optimal Waiting

```
Input: $\omega, \beta$, and $n$
$\varepsilon \leftarrow (\omega - \beta)/\sqrt{n}$, $A \leftarrow \{k\varepsilon : 0 \leq k \leq \bar{y}(\varepsilon)/\varepsilon, k \in \mathbb{N}\}$, $R \leftarrow \bar{y}(\varepsilon) + \omega$
$\eta \leftarrow \sqrt{8 \log(|A|)/n}/R$ and $w_1(y) \leftarrow 1$ for all $y \in A$
for $t = 1, \ldots, n$ do
    $W_t \leftarrow \sum_{y \in A} w_t(y)$ and $p_t(y) \leftarrow w_t(y)/W_t$ for each $y$
    Sample waiting time $Y_t$ from distribution $p_t$ on $A$ and observe $X_t$
    for $y \in A$ do        // Update the weights
            $\ell_t(y) \leftarrow \mathbb{1}\{X_t \leq y\}(X_t + \beta) + \mathbb{1}\{X_t > y\}(y + \omega)$
            $w_{t+1}(y) \leftarrow w_t(y) \exp(-\eta \ell_t(y))$
    end for
end for
```

Theorem 2 (EWA Regret). *Let $n > 0$ and R_n be the regret of Algorithm 1 when used for n rounds. Then $\mathbb{E}[R_n] \in O((\omega - \beta)\log^{3/2}(n)\sqrt{n})$.*

Proof. Let ε, A and R be as in the pseudo-code of the algorithm. As noted beforehand, the expected regret[4] of EWA against the best action in A is $R\sqrt{n/2\log K}$, where $K = |A| \leq \lceil \bar{y}(\varepsilon)/\varepsilon \rceil = \lceil \bar{y}((\omega-\beta)/\sqrt{n})\sqrt{n}/(\omega-\beta) \rceil$. By Proposition 1, $\min_{y \in A} \ell(y) - \inf_{y \in [0,\bar{y}(\varepsilon)]} \ell(y) \leq \varepsilon$ and by Corollary 1, $\inf_{y \in [0,\bar{y}(\varepsilon)]} \ell(y) - \ell(y^*) \leq \varepsilon$. Hence, $\mathbb{E}[R_n] \leq R\sqrt{n/2\log(\bar{y}((\omega-\beta)/\sqrt{n})\sqrt{n}/(\omega-\beta)+1)} + 2(\omega-\beta)\sqrt{n} \in O((\omega-\beta)\log^{3/2}(n)\sqrt{n})$, where we used $R = \bar{y}((\omega-\beta)/\sqrt{n}) + \omega$ and that $\bar{y}((\omega-\beta)/\sqrt{n}) \in O((\omega-\beta)\log(n))$.

Under the full information stochastic setting the FTL algorithm, which at each round chooses the waiting time that minimises the empirical loss so far, is also expected to do well. The next theorem shows that FTL does indeed improve slightly on EWA.

Theorem 3 (FTL Regret). *Let Y_t be defined by $Y_1 \doteq 0$ and, for all $t \geq 2$,*

$$Y_t \in \arg\min_{y} \sum_{s=1}^{t-1} \left(\mathbb{1}\{X_s \leq y\}(X_s + \beta) + \mathbb{1}\{X_s > y\}(y + \omega)\right).$$

Then, $\mathbb{E}[R_n] \leq (\omega - \beta)(11.6\sqrt{n}\log n - 11\sqrt{n} + \log n + 12)$.

Remark 4 *It is easy to see that for any $t \geq 1$, $Y_t = X_s$ for some $1 \leq s \leq t-1$, hence Y_t can be computed in $O(t)$ time. Note that Y_t is not unique.*

Proof. The empirical loss of wait-time y at time step t is

$$\hat{\ell}_t(y) \doteq \frac{1}{t}\sum_{s=1}^{t} \left(\mathbb{1}\{X_s \leq y\}(X_s + \beta) + \mathbb{1}\{X_s > y\}(y + \omega)\right).$$

The expected regret at time step t may be decomposed. Let $(s_t)_t$ be a sequence of constants to be chosen later. Then,

$$\begin{aligned}
&\mathbb{E}[r_t|Y_t] = \ell(Y_t) - \ell(y^*) = \ell(Y_t) - \ell(s_t) + \ell(s_t) - \ell(y^*) \\
&= \ell(Y_t) - \hat{\ell}_{t-1}(Y_t) + \hat{\ell}_{t-1}(Y_t) - \ell(s_t) + \hat{\ell}_{t-1}(s_t) - \hat{\ell}_{t-1}(s_t) + \ell(s_t) - \ell(y^*) \\
&\overset{(a)}{\leq} \left|\ell(Y_t) - \hat{\ell}_{t-1}(Y_t)\right| + \left|\ell(s_t) - \hat{\ell}_{t-1}(s_t)\right| + \ell(s_t) - \ell(y^*),
\end{aligned}$$

where in (a) we used the fact that $\hat{\ell}_{t-1}(Y_t) \leq \hat{\ell}_{t-1}(s_t)$. Now,

$$\begin{aligned}
|\ell(y) - \hat{\ell}_t(y)| &\overset{(a)}{=} |(\mathbb{E} - \mathbb{E}_t)[\mathbb{1}\{X \leq y\}(X + \beta) + \mathbb{1}\{X > y\}(\omega + y)]| \\
&\overset{(b)}{=} |(\mathbb{E} - \mathbb{E}_t)[\mathbb{1}\{X \leq y\}(X - y) + \mathbb{1}\{X > y\}(\omega - \beta)]|
\end{aligned}$$

[4] Bounds for adversarial algorithms like EWA are typically proven for the regret without the expectation, but in the stochastic case this distinction is not important with bounds on the expected regret following from a straight-forward application of standard concentration inequalities.

$$\overset{(c)}{\le} y\left|F(y) - F_t(y)\right| + (\omega - \beta)\left|F(y) - F_t(y)\right| ,$$

where (a) is simply the definition of the losses and $(\mathbb{E} - \mathbb{E}_t)$, (b) by rearranging and using the fact that $(\mathbb{E} - \mathbb{E}_t)\alpha = 0$ for any constant α, (c) by $|X - y| \le y$ which holds for $0 \le X \le y$ and the definition of the cumulative distribution. Combined with [7, Thm. 3.3], which states that $\mathbb{E}[\sup_x |F_t(x) - F(x)|] \le 1/\sqrt{t}$, the last inequality gives

$$\mathbb{E}\left[\sup_{y \le s}\left|\ell(y) - \hat{\ell}_t(y)\right|\right] \le \frac{s + \omega - \beta}{\sqrt{t}} . \tag{5}$$

Next we show that $Y_{t+1} \in O(\log t)$ for any $t \ge 1$. Since Y_{t+1} is the optimal waiting time for the empirical distribution of the arrival times, we can apply Part 1 of Lemma 1 to obtain

$$\begin{aligned}\mathbb{P}_t\left\{X \ge Y_{t+1}\right\} = \inf_{\varepsilon > 0} \mathbb{P}_t\left\{X > Y_{t+1} - \varepsilon\right\} &\le \inf_{\varepsilon > 0} 2^{-\left\lfloor \frac{Y_{t+1} - \varepsilon}{2(\omega - \beta)} \right\rfloor} \\ &\le \inf_{\varepsilon > 0} 2^{1 - \frac{Y_{t+1} - \varepsilon}{2(\omega - \beta)}} = 2^{1 - \frac{Y_{t+1}}{2(\omega - \beta)}} .\end{aligned}$$

Therefore, if $Y_{t+1} > m_{t+1} \doteq 2(\omega - \beta)(1 + \log_2 t)$, then $\mathbb{P}_t\left\{X \ge Y_{t+1}\right\} < 1/t$. On the other hand, $\mathbb{P}_t\left\{X \ge Y_{t+1}\right\} \ge 1/t$ since $Y_{t+1} \in \{X_1, \ldots, X_t\}$. Thus, $Y_{t+1} \le m_{t+1}$. Choose $s_t = \min\{y^*, m_t\}$. Then, by (5),

$$\begin{aligned}\mathbb{E}\left[r_{t+1}\right] &\le \mathbb{E}\left[\left|\ell(Y_{t+1}) - \hat{\ell}_t(Y_{t+1})\right| + \left|\ell(s_{t+1}) - \hat{\ell}_t(s_{t+1})\right|\right] + \ell(s_{t+1}) - \ell(y^*) \\ &\le \mathbb{E}\left[\sup_{y \le m_{t+1}}\left|\ell(y) - \hat{\ell}_t(y)\right| + \left|\ell(s_{t+1}) - \hat{\ell}_t(s_{t+1})\right|\right] + \ell(s_{t+1}) - \ell(y^*) \\ &\le \left(m_{t+1} + s_{t+1} + 2(\omega - \beta)\right)\frac{1}{\sqrt{t}} + \ell(s_{t+1}) - \ell(y^*) \\ &\le \left(m_{t+1} + s_{t+1} + 2(\omega - \beta)\right)\frac{1}{\sqrt{t}} + \frac{\omega - \beta}{t} ,\end{aligned}$$

where in the last step we used Part 2 of Lemma 1 to bound $\ell(s_{t+1}) - \ell(y^*)$. Summing over t ultimately leads to

$$\begin{aligned}\mathbb{E}\left[R_n\right] = \mathbb{E}\left[r_1\right] + \sum_{t=2}^{n} \mathbb{E}\left[r_t\right] &\le (\omega - \beta)\left(1 + \sum_{t=1}^{n-1}\left[\frac{6 + 4\frac{\log t}{\log 2}}{\sqrt{t}} + \frac{1}{t}\right]\right) \\ &\le (\omega - \beta)(11.6\sqrt{n}\log n - 11\sqrt{n} + \log n + 12) .\end{aligned}$$

as required. □

If the arrival time X_t is exponentially distributed, then the regret of the FTL algorithm may be shown to be at most poly-logarithmic. Experimental results suggest that the true regret is actually logarithmic in n, but so far the proof eludes us.

Theorem 5. *Assume that X_t is exponentially distributed with parameter λ such that $1/\lambda < \omega - \beta$. Then, for the algorithm of Theorem 3, we have $\mathbb{E}\left[R_n\right] \in O(\log^2 n)$.*

3.1 Lower Bound

The general upper bounds given in the previous section cannot be greatly improved in the worst-case. Note that the following theorem is proven for the easier full information setting, so translates immediately to give an identical lower bound in the partial information setting.

Theorem 6. *There exists a universal constant $c > 0$ such that for each algorithm and fixed n there exists a distribution such that $\mathbb{E}R_n \geq c(\omega - \beta)\sqrt{n}$.*

Proof. For $p \in [0,1]$ let $\mathbb{P}_p$ be a measure defined such that $\mathbb{P}_p(X = 1/2) = p$ and $\mathbb{P}_p(X = \infty) = 1 - p$. Let us denote the expected loss under measure $\mathbb{P}_p$ by ℓ_p. A simple calculation shows that

$$\ell_p(y) = \begin{cases} y + \omega, & \text{if } y < \frac{1}{2}; \\ p(\frac{1}{2} + \beta) + (1-p)(y + \omega), & \text{otherwise.} \end{cases}$$

Thus, ℓ_p is piecewise linear, with two increasing segments. The two local minima of ℓ_p are at 0 and 1/2 with values $\ell_p(0) = \omega$ and $\ell_p(1/2) = 1/2 + \omega - p(\omega - \beta)$. For simplicity, we set $\omega = 1$, $\beta = 0$, the full result can be obtained by scaling. Thus, $\ell_p(0) = 1$, $\ell_p(1/2) = 3/2 - p$ and the optimal waiting time y^*_p is 0 for $p < 1/2$ and 1/2 for $p > 1/2$. If $p = 1/2$, then 0 and 1 are both optimal. It is also clear that for the "rounding function" ρ defined by $\rho(y) = \frac{1}{2}\mathbb{1}\{y \geq 1/2\}$, then for any $y \geq 0$ it holds that $\ell_p(\rho(y)) \leq \ell_p(y)$: By "rounding down" the waiting time y to either 0 or 1/2, one can only win in terms of the expected loss. Based on $\mathbb{P}_p$, we construct three environments and will use a fairly standard technique based on the relative entropy that shows that the regret will be large in at least in one of the environments. The three environments are given by the measures $\mathbb{P}_{1/2}$, $\mathbb{P}_{1/2+\varepsilon}$ and $\mathbb{P}_{1/2-\varepsilon}$ for some $\varepsilon \in [0, 1/2)$ to be chosen later. Note that $|\ell_{1/2+\sigma\varepsilon}(0) - \ell_{1/2+\sigma\varepsilon}(1/2)| = \sigma\varepsilon$. Fix $n > 0$. Now, take any algorithm A and let Y_t be the choice made by A in round $1 \leq t \leq n$. Let R_σ be the expected regret of A during the first n rounds when used on the measure $\mathbb{P}_{1/2+\sigma\varepsilon}$, $\sigma \in \{-1, 0, +1\}$. Denoting by $\mathbb{E}_\sigma$ the expectation under $\mathbb{P}_{1/2+\sigma\varepsilon}$, we thus have $R_\sigma = \mathbb{E}_\sigma[\sum_{t=1}^n \ell_{1/2+\sigma\varepsilon}(Y_t) - \ell_{1/2+\sigma\varepsilon}(y^*_{1/2+\sigma\varepsilon})]$. Let $\hat{Y}_t = \rho(Y_t)$ be the "rounded" decision and let $N(y) = \sum_{t=1}^n \mathbb{1}\left\{\hat{Y}_t = y\right\}$, $y \in \{0, 1/2\}$. Then,

$$R_\sigma \geq \mathbb{E}_\sigma\left[\sum_{t=1}^n \ell_{1/2+\sigma\varepsilon}(\hat{Y}_t) - \ell_{1/2+\sigma\varepsilon}(y^*_{1/2+\sigma\varepsilon})\right]$$

and thus

$$R_1 \geq \varepsilon\mathbb{E}_1[N(0)], \qquad R_{-1} \geq \varepsilon\mathbb{E}_{-1}[N(1/2)]\,. \tag{6}$$

Now, a standard argument shows that

$$\mathbb{E}_0[N(0)] - \mathbb{E}_1[N(0)] \leq n\sqrt{\frac{n}{2}D(\mathbb{P}_0||\mathbb{P}_1)} \leq 2n\varepsilon\sqrt{\frac{n}{2}},$$

$$\mathbb{E}_0[N(1/2)] - \mathbb{E}_{-1}[N(1/2)] \leq n\sqrt{\frac{n}{2}D(\mathbb{P}_0||\mathbb{P}_{-1})} \leq 2n\varepsilon\sqrt{\frac{n}{2}}\,,$$

where $D(\mathbb{P}_0||\mathbb{P}_{-1})$ denotes the relative entropy between $\mathbb{P}_0$ and $\mathbb{P}_{-1}$. Summing up these two inequalities and using (6), $n-(R_1/\varepsilon+R_{-1}/\varepsilon) \leq 4n\varepsilon\sqrt{n/2}$. Setting $\varepsilon = c/\sqrt{n}$ and reordering gives $\sqrt{n}c(1-2\sqrt{2}c) \leq R_1+R_{-1}$. Choose $c = 1/(4\sqrt{2})$ and note that $2\max(R_1, R_{-1}) \geq R_1 + R_{-1}$ to finish the proof. □

4 Partial Information

We now consider the more challenging case where X_t is not observed if $Y_t < X_t$ and so the waiting time directly influences the amount of information gained at each time step. Just like in the previous section, our first algorithm is based on a discretisation idea. As before, we first notice that it is enough to consider stopping times in an interval of length $O((\omega-\beta)\log(n))$ and also that a discretisation accuracy of $\varepsilon = (\omega-\beta)/\sqrt{n}$ will suffice to get a $\tilde{O}(\sqrt{n})$ regret, which is conjectured to hold. In this case, however, an appropriately modified version of the exponential weights algorithm is needed which works with estimated losses and adds exploration to facilitate the estimation of losses. In fact, as it turns out, after discretisation, our problem falls into the framework of prediction with expert advice with side-observations, where after the learner chooses an action $Y_t \in A$ it observes the losses for a subset $S(Y_t) \subset A$ of actions. In our case, $S(Y_t) = \{y \in A : y \leq Y_t\}$, which means that waiting for a longer time leads to more information than waiting for a shorter time. This framework was first studied by Mannor and Shamir [14]. Here, we will use the Exp3-DOM algorithm of Alon et. al. as this algorithm improves upon the results of Mannor and Shamir for our setting [2]. The general idea of Exp3-DOM is to restrict exploration to actions in a dominating set D, which is a subset of actions such that $\cup_{a\in D} S(a) = A$. In particular, exploration is restricted to a minimal dominating set. In our case, the minimal dominating set contains a single element, $y_{\max} = \max A$. This results in Algorithm 2. If the learning rate η is chosen carefully, then Theorem 7 of [2] shows that the algorithm suffers a $\tilde{O}(\sqrt{n})$ regret. Recall the definition of $\bar{y}(\varepsilon)$

Theorem 7. *Pick $n > 0$ and let R_n be the regret of Algorithm 1 when used for n rounds. Then, $\mathbb{E}[R_n] \in O((\omega-\beta)\log^2(n)\sqrt{n})$.*

Proof. Let ε, A and R be as in the pseudo-code of the algorithm. Using that in our case the the observation sets $S(a)$ are fixed, Theorem 7, Eq. (2) of [2] gives that the regret of Exp3-DOM against the best waiting time in A is $O(K\log(K) + \log(K)(1/\eta + \eta R^2 \sum_{t=1}^{n}(1+Q_t)))$, where $K = |A|$ is the number of actions and $Q_t = \sum_{y\in A} p_t(y)/q_t(y)$ and where we used that the dominant set in our case has a single element. Now, Lemma 13 of [2] gives that $Q_t = \alpha\ln(K/\eta)$, where α is the so-called independence number of the graph (A, E) underlying the observation system: $(a_1, a_2) \in E$ if $a_1 \in S(a_2)$ or $a_2 \in S(a_1)$. In our case, the graph is a clique and hence its independence number is $\alpha = 1$. Choosing $\eta = 1/(R\sqrt{n})$ thus gives that the regret of Exp3-DOM against the best waiting time in A is $O(R\log(K)\sqrt{n} + K\log(K))$. By Proposition 1, $\min_{y\in A}\ell(y) - \inf_{y\in[0,\bar{y}(\varepsilon)]}\ell(y) \leq \varepsilon$, while by Corollary 1, $\inf_{y\in[0,\bar{y}(\varepsilon)]}\ell(y) - \ell(y^*) \leq$

Algorithm 2 Exp3-Dom

```
Input: $\omega, \beta$, and $n$
// Recall definition of $\bar{y}(\varepsilon)$ given in Corollary 1
$\varepsilon \leftarrow (\omega - \beta)/\sqrt{n}$, $A \leftarrow \{k\varepsilon : 0 \leq k \leq \bar{y}(\varepsilon)/\varepsilon, k \in \mathbb{N}\}$, $R \leftarrow \bar{y}(\varepsilon) + \omega$, $\eta \leftarrow 1/(R\sqrt{n})$
$w_1(y) \leftarrow 1$ for all $y \in A$
for $t = 1, \ldots, n$ do
    $W_t \leftarrow \sum_{y \in A} w_t(y)$ and $p_t(y) \leftarrow \eta w_t(y)/W_t + (1 - \eta)\mathbb{1}\{y = \max A\}$ for each $y$
    Sample waiting time $Y_t$ from distribution $p_t$ on $A$ and observe $Z_t, \delta_t$
    $w_{t+1}(y) \leftarrow w_t(y)$ for all $y \in A$
    for $y \in A \cap [0, Y_t]$ do        // Update the weights
        $\ell_t(y) \leftarrow \mathbb{1}\{Z_t \leq y\}(Z_t + \beta) + \mathbb{1}\{Z_t > y\}(y + \omega)$
        $q_t(y) \leftarrow \sum_{y' \in A : y' \geq y} p_t(y')$ and $\tilde{\ell}_t(y) \leftarrow \ell_t(y)/q_t(y)$
        $w_{t+1}(y) \leftarrow w_t(y)\exp(-\eta\tilde{\ell}_t(y))$
    end for
end for
```

ε. Hence, $\mathbb{E}[R_n] \in O(R\log(K)\sqrt{n} + K\log(K) + (\omega - \beta)\sqrt{n})$. Now, using the definition of $\bar{y}$, $\bar{y}((\omega-\beta)/\sqrt{n}) \in O((\omega - \beta)\log(n))$. Thus, $K = |A| \in O(\lceil \bar{y}(\varepsilon)/\varepsilon \rceil) = O(\bar{y}((\omega - \beta)/\sqrt{n})\sqrt{n}/(\omega - \beta)) = O(\log(n)\sqrt{n})$ and $R = \bar{y}((\omega - \beta)/\sqrt{n}) + \omega \in O((\omega - \beta)\log(n))$. Plugging these into the previous bound, we get $\mathbb{E}[R_n] \in O((\omega - \beta)\log^2(n)\sqrt{n})$.

Note that since the partial information setting is strictly more difficult than the full information setting, our previous lower bound shows that the regret cannot be better than $\Omega(\sqrt{n})$. However, as in the full information setting, we can expect to improve upon the performance of Exp3-DOM by using an algorithm that exploits the fact that the environment is stochastic. In particular, as is common in sequential learning algorithms we make use of an optimistic strategy, which will wait for the bus as long as reasonably possible. The algorithm maintains an estimate of the cumulative distribution and chooses a non-increasing sequence of waiting times starting from a carefully chosen upper bound. The waiting times decrease at a data-dependent rate that is chosen to ensure some nearly-optimal waiting time is always smaller than the action chosen. This results in Algorithm 3.

The following theorem bounds the expected regret of Algorithm 3. The bound is worse by a factor of $O(\sqrt{\log n})$ than that obtained in the full information setting described in Section 3, but improves the bound announced in Theorem 7.

Theorem 8. *The regret is bounded by* $\mathbb{E}[R_n] \leq (\omega - \beta)(42 + 7\log^{3/2}(n)\sqrt{n})$.

Lemma 3. *Define* $\bar{y} = \min\{y^*, y_{\max}\}$*, where* $y_{\max}$ *is given in Algorithm 3. Then* $\ell(\bar{y}) - \ell(y^*) \leq (\omega - \beta)/\sqrt{n}$.

Proof. Apply Part 2 of Lemma 1 and the definition of $\bar{y}$. □

The following lemma shows that $Y_t \geq \bar{y}$ for all $1 \leq t \leq n$ with high probability. This means that if $y^* \geq y_{\max}$, then with high probability the algorithm will

Algorithm 3 Optimistic Waiting

```
Input: $\omega$, $\beta$, and $n$
$\alpha \leftarrow 3/2$ and $y_{\max} \leftarrow 2(\omega-\beta)(1+\log_2\sqrt{n})$ and $Y_1 \leftarrow y_{\max}$
for $t = 1, \dots, n$ do
    Observe $Z_t$, $\delta_t$
    Compute the empirical distribution: $G_t(y) \doteq \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}\{Z_s \le y\}$
    Compute waiting time for next day:
        $\varepsilon_t \leftarrow \sqrt{\frac{\log 2n^\alpha}{2t}}$
        $Y_{t+1} \leftarrow \max\Big\{0 \le y \le Y_t : G_t(y) - G_t(y') + 2\varepsilon_t \ge \frac{y-y'}{\omega-\beta}(1 - G_t(y) - \varepsilon_t) - \frac{1}{\sqrt{n}},\ 0 \le y' \le y\Big\}$
end for
```

always choose $Y_t = y_{\max}$ and suffer no more than $(\omega-\beta)/\sqrt{n}$ regret per time step. On the other hand, if $y^* < y_{\max}$, then the algorithm will choose $Y_t \ge y^*$, which guarantees that it is continually learning information about the loss of the optimal action.

Lemma 4. *For $1 \le t \le n$, we have that $\mathbb{P}\{A_n^c\} \le n^{1-\alpha}$, where event A_t is defined by $A_t = \bigcap_{s\le t}\left\{Y_s \ge \bar{y} \text{ and } \sup_{y\le Y_{s-1}} |G_{s-1}(y) - F(y)| \le \varepsilon_{t-1}\right\}$.*

Proof. Define event $B = \bigcup_{t\le n}\left\{\sup_{x\le Y_{t-1}} |G_{t-1}(x) - F(x)| \ge \varepsilon_{t-1}\right\}$ and recall that $F_t(x) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}\{X_s \le x\}$, which is unknown to the learner. The Dvoretzky–Kiefer–Wolfowitz–Massart theorem [8, 15] gives that

$$\mathbb{P}\left\{\sup_x |F_t(x) - F(x)| \ge \varepsilon_t\right\} \le 2\exp\left(-2\varepsilon_t^2 t\right) = n^{-\alpha}.$$

Therefore, by the union bound, with probability at least $1 - n^{1-\alpha}$ it holds that $|F_t(x) - F(x)| \le \varepsilon_t$ for all $t \le n$ and $x \in \mathbb{R}$. By the definition of the observations $(Z_s)_s$, $G_t(y) = F_t(y)$ for all $y \le \min_{1\le s\le t} Y_s$. Further, since by construction $(Y_t)_t$ is non-increasing, $\min_{1\le s\le t} Y_s = Y_t$ and so $G_t(y) = F_t(y)$ for all $y \le Y_t$. Therefore $\mathbb{P}\{B\} \le n^{1-\alpha}$. We now show that if B does not occur then A_t holds for $1 \le t \le n$. We prove this by induction on t. That B^c implies A_1 is trivial. Now, assume that B^c implies that A_t holds for some $1 \le t < n$. On B^c we have

$$\sup_{y\le Y_t} |G_t(y) - F(y)| \le \varepsilon_t. \tag{7}$$

Thus, it suffices to show that on B^c, $Y_{t+1} \ge \bar{y}$ also holds. By the induction hypothesis, $Y_t \ge \bar{y}$. Combining this with (7) we get

$$\sup_{y\le \bar{y}} |G_t(y) - F(y)| \le \varepsilon_t. \tag{8}$$

Now let $y' \leq \bar{y} \leq y^*$. Then

$$\begin{aligned}
0 &\overset{(a)}{\geq} \ell_t(y^*) - \ell_t(y') \overset{(b)}{\geq} \ell_t(\bar{y}) - \ell_t(y') - \frac{\omega - \beta}{\sqrt{n}} \\
&\overset{(c)}{\geq} (\bar{y} - y')(1 - F(\bar{y})) - (\omega - \beta)(F(\bar{y}) - F(y')) - \frac{\omega - \beta}{\sqrt{n}} \\
&\overset{(d)}{\geq} (\bar{y} - y')(1 - G_t(\bar{y}) - \varepsilon_t) - (\omega - \beta)(G_t(\bar{y}) - G_t(y') + 2\varepsilon_t) - \frac{\omega - \beta}{\sqrt{n}},
\end{aligned}$$

where (a) follows since y^* is the optimal waiting time, (b) by Lemma 3, (c) by Part 2 of Lemma 2, and (d) holds by (8). Rearranging we obtain

$$G_t(\bar{y}) - G_t(y') + 2\varepsilon_t \geq \frac{\bar{y} - y'}{\omega - \beta}(1 - G_t(\bar{y}) - \varepsilon_t) - \frac{1}{\sqrt{n}},$$

which implies, by the definition of Y_{t+1}, that $Y_{t+1} \geq \bar{y}$. Therefore A_{t+1} holds and so B^c implies that A_n holds. Therefore $\mathbb{P}\{A_n\} \geq \mathbb{P}\{B^c\} \geq 1 - n^{1-\alpha}$. □

Proof (of Theorem 8). The proof follows almost immediately from Lemmas 2 to 4. Assume that A_n holds. Then

$$\begin{aligned}
\ell(Y_t) - \ell(y^*) &\overset{(a)}{\leq} \ell(Y_t) - \ell(\bar{y}) + \frac{\omega - \beta}{\sqrt{n}} \\
&\overset{(b)}{\leq} (Y_t - \bar{y})(1 - F(\bar{y})) - (\omega - \beta)(F(Y_t) - F(\bar{y})) \\
&\overset{(c)}{\leq} (Y_t - \bar{y})\left(1 - G_{t-1}(\bar{y}) + \varepsilon_{t-1}\right) - (\omega - \beta)\left(G_{t-1}(Y_t) - G_{t-1}(\bar{y}) - 2\varepsilon_{t-1}\right) \\
&\overset{(d)}{\leq} 2\varepsilon_{t-1}(Y_t - \bar{y}) + 4\varepsilon_{t-1}(\omega - \beta) + \frac{\omega - \beta}{\sqrt{n}}, \qquad (9)
\end{aligned}$$

where (a) follows from Lemma 3, (b) by Part 2 of Lemma 2 and the fact that $Y_t \geq \bar{y}$, (c) follows from the definition of A_n while (d) follows from the definition of Y_t. Therefore, on A_n,

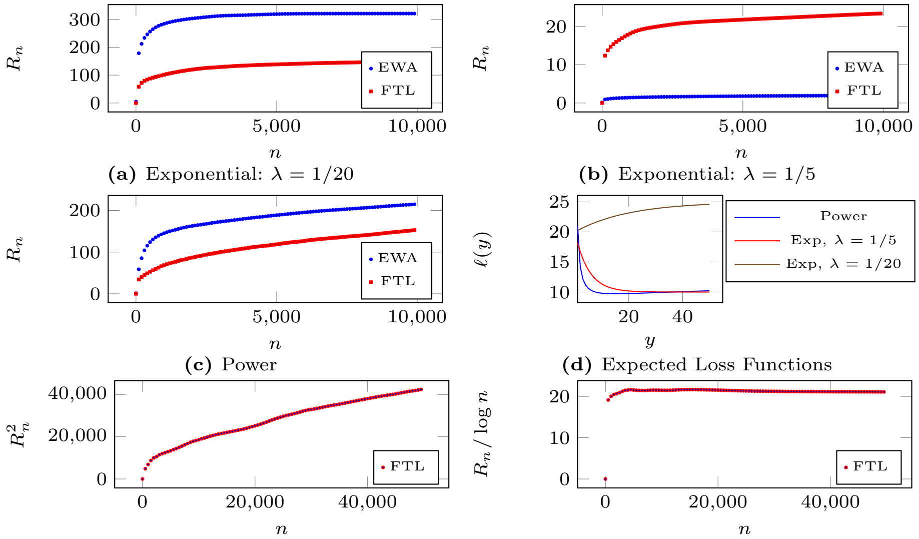
$$\begin{aligned}
&\sum_{t=1}^n \ell(Y_t) - \ell(y^*) \overset{(a)}{\leq} \ell(Y_1) - \ell(y^*) + \sum_{t=2}^n \left(2\varepsilon_{t-1}(Y_t - \bar{y}) + 4\varepsilon_{t-1}(\omega - \beta) + \frac{\omega - \beta}{\sqrt{n}}\right) \\
&\overset{(b)}{\leq} y_{\max} + \sum_{t=2}^n \left(2\varepsilon_{t-1}(y_{\max} + 2(\omega - \beta)) + \frac{\omega - \beta}{\sqrt{n}}\right) \overset{(c)}{\leq} (\omega - \beta)(40 + 5\log^{\frac{3}{2}}(n)\sqrt{n}),
\end{aligned}$$

where (a) follows from (9), (b) follows by naively bounding $Y_t - \bar{y} \leq y_{\max}$, while (c) follows arduously from the definition of ε_t and $y_{\max}$. In case A_n does not hold, the regret may be as much as $y_{\max}$ per day, but $\mathbb{P}\{A_n^c\} \leq n^{1-\alpha} = 1/\sqrt{n}$. Combining with the previous display completes the result. □

5 Experiments

We performed three experiments comparing EWA with FTL in the full-information case with $\omega = 20$ and $\beta = 5$. We used two exponential distributions with $\lambda = 1/20$ and $1/5$ respectively, as well as a power law distribution (see

Table 1). The horizon was set to $n = 10,000$ and the learning rate of exponential weighting was tuned to be a factor of 100 larger than the theoretical optimum, which was observed to give a good performance across all three problems. The FTL algorithm generally out-performs the exponential weighting algorithm, but not by an enormous margin. If the theoretically optimal learning rate is used then the performance of exponential weighting deteriorates significantly. Figures (e) and (f) suggest that FTL suffers $\sqrt{n}$ regret on the power-law distribution, but logarithmic regret for exponentially distributed arrival times with parameter $\lambda = 1/20$. Each data point is the average of 20 independent trials. Code is available at http://downloads.tor-lattimore.com/projects/optimal_waiting.



(a) Exponential: $\lambda = 1/20$

(b) Exponential: $\lambda = 1/5$

(c) Power

(d) Expected Loss Functions

(e) Power Law, squared regret

(f) Exponential, $\lambda = 1/20$

6 Conclusions

We introduced the problem of learning an optimal waiting time with two variants. In both cases, we presented two general algorithms relying on no assumptions that were shown to enjoy near-optimal worst-case regret. Interesting future work is to further analyse the problem-dependent regret bounds of FTL and other algorithms in both full and partial information settings beyond exponentially distributed arrivals. One approach for less conservative algorithms may be to use the Kaplan-Meier estimator rather than the standard empirical distribution, but the mathematical theory behind this estimator is not yet well-developed for this setting where the censoring times are known and not i.i.d. One exception is by Ganchev et. al., but unfortunately their confidence bound depends on the scale and is not suitable for obtaining optimal regret bounds in our problem [10]. Another challenge is to improve the running time of the algorithms to $O(1)$ per time step. While our results are the first in this setting, we expect various extensions to related problems, such as when one can choose between multiple options with random durations.

Acknowledgements This work was supported by the Alberta Innovates Technology Futures and NSERC.

Bibliography

[1] Rajeev Agrawal, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space. *IEEE Transaction on Automatic Control*, 34:258–267, 1989.

[2] Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems*, pages 1610–1618, 2013.

[3] Gábor Bartók. A near-optimal algorithm for finite partial-monitoring games against adversarial opponents. In *COLT*, pages 696–710, 2013.

[4] Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *COLT 2011*, pages 133–154, 2011.

[5] Nicolo Cesa-Bianchi. *Prediction, learning, and games.* Cambridge University Press, 2006.

[6] A Clifford Cohen. *Truncated and censored samples: theory and applications.* CRC press, 1991.

[7] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation.* Springer, 2001.

[8] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27:642–669, 1956.

[9] Dean P. Foster and Alexander Rakhlin. No internal regret via neighborhood watch. *Journal of Machine Learning Research - Proceedings Track (AISTATS)*, 22:382–390, 2012.

[10] Kuzman Ganchev, Yuriy Nevmyvaka, Michael Kearns, and Jennifer Wortman Vaughan. Censored exploration and the dark pool problem. *Communications of the ACM*, 53(5):99–107, 2010.

[11] Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 594–605. IEEE, 2003.

[12] P. Krishnan, Philip M. Long, and Jeffrey Scott Vitter. Adaptive disk spindown via optimal rent-to-buy in probabilistic environments. *Algorithmica*, 23(1):31–56, 1999.

[13] Tor Lattimore, András György, and Csaba Szepesvári. On learning the optimal waiting time. http://downloads.tor-lattimore.com/projects/optimal_waiting/, 2014.

[14] Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *NIPS*, pages 684–692, 2011.

[15] Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18:1269–1283, 1990.

[16] Celso C Ribeiro, Isabel Rosseti, and Reinaldo Vallejos. Exploiting run time distributions to compare sequential and parallel stochastic local search algorithms. *Journal of Global Optimization*, 54(2):405–429, 2012.