Non-Parametric Stochastic Sequential Assignment With Random Arrival Times

Danial Dervovic[1] , **Parisa Hassanzadeh**[1] , **Samuel Assefa**[1] , **Prashant Reddy**[1]
[1]J.P. Morgan AI Research
{danial.dervovic, parisa.hassanzadeh, samuel.a.assefa, prashant.reddy}@jpmorgan.com

Abstract

We consider a problem wherein jobs arrive at random times and assume random values. Upon each job arrival, the decision-maker must decide immediately whether or not to accept the job and gain the value on offer as a reward, with the constraint that they may only accept at most n jobs over some reference time period. The decision-maker only has access to M independent realisations of the job arrival process. We propose an algorithm, Non-Parametric Sequential Allocation (NPSA), for solving this problem. Moreover, we prove that the expected reward returned by the NPSA algorithm converges in probability to optimality as M grows large. We demonstrate the effectiveness of the algorithm empirically on synthetic data and on public fraud-detection datasets, from where the motivation for this work is derived.

1 Introduction

In industrial settings it is often the case that a positive class assignment by a classifier results in an expensive manual intervention. A problem frequently arises whereby the number of these alerts exceeds the capacity of operators to manually investigate alerted examples [Beyer *et al.*, 2016]. A common scenario is one where each example is further endowed with an intrinsic value along with its class label, with all negative examples having zero value to the operator. Given their limited capacity, operators wish to maximise the cumulative value gained from expensive manual interventions. As an example, in financial fraud detection [Bolton and Hand, 2002], this is manifested as truly fraudulent transactions having value to the operator as (some function of) the monetary value of the transaction, and non-fraudulent transactions yielding zero value [Dal Pozzolo, 2015].

In this paper we systematically account for the constraint on intervention capacity and desire to maximise reward, in the setting where selections are made in real-time and we have access to a large backlog of training data. This problem structure is not limited to fraud, for example in cybersecurity [Vaněk *et al.*, 2012], automated content moderation [Consultants, 2019], compliance verification [Avenhaus *et al.*, 1996] and automated inspection in manufacturing [Morishita and Okumura, 1983] there is a need for filtering a stream of comparable examples that are too numerous for exhaustive manual inspection, with the imperative of maximising the value of inspected examples. We shall take an abstract view of *jobs* arriving, each having an intrinsic value.

To this end, we extend a problem first considered by [Albright, 1974], in which jobs arrive according to a random process and take on random nonnegative values. At each job arrival, the decision-maker must decide immediately whether or not to accept the job and gain the value on offer as a reward. They may only accept at most n jobs over some reference time period. In [Albright, 1974], this problem is solved optimally by way of a system of ordinary differential equations (ODE). Importantly, the job arrival process is assumed to be known and admits a closed-form mathematical expression. Solving the resulting system of ODEs analytically quickly becomes impractical, even for trivial job arrival processes. We propose an efficient algorithm, *Non-parametric Sequential Allocation Algorithm* (NPSA), which allows one merely to observe M realisations of the job arrival process and still recover the optimal solution, as defined by this solution of ODEs, with high probability. We empirically validate NPSA on both synthetic data and public fraud data, and rigorously prove its optimality.

This work plugs the gap in the literature where the following must be simultaneously accounted for: *i.* explicit constraints on the number of job acceptances; *ii.* maximising reward; *iii.* treating job arrivals as a continuous-time random process; and *iv.* learning the job value distribution and arrival process from data.

Related Work. The framework of *Cost-sensitive learning* [Elkan, 2001] seeks to minimise the misclassification cost between positive and negative examples, even on an example-by-example basis [Bahnsen *et al.*, 2014], but often the methods are tuned to a specific classification algorithm and do not admit specification of an explicit constraint on the number of positive labels. In [Shen and Kurshan, 2020], the authors formulate fraud-detection as an RL problem. They explicitly take into account the capacity of inspections and costs, but operate in discrete-time and provide no theoretical guarantees. Solving a *Constrained MDP* [Altman, 1999] optimises reward under long-term constraints that can be violated instantaneously but must be satisfied on average, such as in [Efroni *et al.*, 2020; Zheng and

Ratliff, 2020]. Works such as [Mannor and Tsitsiklis, 2006; Jenatton *et al.*, 2016] on *Constrained Online Learning* focus on a setting where the decision-maker interacts with an adversary and is simultaneously learning the adversary's behaviour and the best response, as measured by regret and variants thereof [Zhao *et al.*, 2020]. Constraints typically relate to quantities averaged over sample paths [Mannor and Shimkin, 2004], whereas in our problem we have a discrete finite resource that is exhausted. In this work we consider a non-adversarial environment that we learn before test-time from training data. Moreover, the setting we focus on here explicitly is continuous-time and finite horizon, contrasting with the constrained MDP and online learning literature which considers discrete-time with an often infinite horizon. Our problem aligns most closely with the framework of *Stochastic Sequential Assignment Problems (SSAP)* [Derman *et al.*, 1972; Khoshkhou, 2014], where it is assumed that distributions of job values and the arrival process are known and closed-form optimal policies derived analytically; the question of learning from data is ignored [Dupuis and Wang, 2002].

2 Problem Setup

We follow a modified version of the problem setup in [Albright, 1974]. We assume a finite time horizon, from $t = 0$ to $t = T$, over which jobs arrive according to a nonhomogeneous Poisson process with continuous intensity function $\lambda(t)$. There are a fixed number of indistinguishable workers, n, that we wish to assign to the stream of incoming jobs. Each worker may only accept one job. Every job has a nonnegative value associated to it that is gained as a reward by the decision-maker if accepted. Any job that is not assigned immediately when it arrives can no longer be assigned. It is assumed the total expected number of jobs that arrive over the horizon $[0, T]$ is much larger than the number of available workers, that is, $n \ll \int_0^T \lambda(t)\,\mathrm{d}t$.

We take the job values to be i.i.d. nonnegative random variables drawn from a cumulative distribution F with finite mean $0 < \mu < \infty$ and density f. Moreover, we assume that the job value distribution is independent of the arrival process. The decision-maker's goal is to maximise the total expected reward accorded to the n workers over the time horizon $[0, T]$. We hereafter refer to Albright's problem as SeqAlloc (short for Sequential Allocation).

In the SeqAlloc model, it is assumed that $\lambda(t)$ and F are known to the decision-maker ahead of time, and an optimal *critical curve* $y_k(t)$ is derived for each of the n workers. When the k^{th} worker is active, if a job arrives at time t with value greater than $y_k(t)$ the job is accepted, at which point the $(k-1)^{\text{th}}$ worker is then active, until all n workers have been exhausted. These critical curves are addressed in more detail in Theorem 1.

We modify the SeqAlloc problem setting in the following way. The arrival intensity $\lambda(t)$ and F are *unknown* to the decision-maker ahead of time. Instead, they have access to M independent realisations of the job arrival process. Each realisation consists of a list of tuples (x_i, t_i), where x_i is the reward for accepting job i and t_i its arrival time. The goal for the decision-maker is the same as in the previous paragraph, that is, to derive critical curves for the n workers so as to maximise the expected cumulative reward at test time. In Section 3.1 we present an efficient algorithm for deriving these critical curves. We hereafter refer to the modified problem we address in this paper as Non-Parametric SeqAlloc, or SeqAlloc-NP for short.

3 Optimal Sequential Assignment

Following [DeGroot, 1970; Sakaguchi, 1977] we define a function that will take centre-stage in the sequel.

Definition 1 (Mean shortage function)**.** *For a nonnegative random variable X with pdf f and finite mean μ, the* mean shortage function *is given as $\phi(y) := \int_y^\infty (x-y) f(x)\,\mathrm{d}x$ for $y \geq 0$.*

The next result follows from [Albright, 1974, Theorem 2].

Theorem 1 (SeqAlloc critical curves)**.** *The (unique) optimal critical curves $y_n(t) \leq \ldots \leq y_1(t)$ solving the SeqAlloc problem satisfy the following system of ODEs (where $1 \leq k \leq n$):*

$$\frac{\mathrm{d}y_{k+1}(t)}{\mathrm{d}t} = -\lambda(t)\left(\phi(y_{k+1}(t)) - \phi(y_k(t))\right),$$
$$\phi(y_0(t)) = 0, \qquad y_k(T) = 0, \qquad t \in [0, T].$$

Indeed, solving this system of ODEs exactly is generally intractable, as we shall see in more detail in Section 4. Theorem 1 provides the optimal solution to the SeqAlloc problem.

3.1 Numerical Algorithm for SeqAlloc-NP

An algorithm to solve the non-parametric problem, SeqAlloc-NP, immediately suggests itself as shown in Algorithm 1: use the M independent realisations of the job arrival process to approximate the intensity $\lambda(t)$ and the mean shortage function $\phi(y)$, then use a numerical ODE solver with Theorem 1 to extract critical curves.

Algorithm 1: NPSA: solution to SeqAlloc-NP

Input: Number of workers n, ODE solver $\mathcal{D}$
Data: M realisations of job arrival process, $\mathcal{M}$
Output: Critical curves $\{\widetilde{y}_k(t)\}_{k=1}^n$

begin
  Estimate $\widetilde{\lambda}(t)$ and $\widetilde{\phi}(y)$ from $\mathcal{M}$
  $\widetilde{y}_0(t) \leftarrow \infty$, $Y \leftarrow \{\widetilde{y}_0(t)\}$
  for k **in** $(1, \ldots, n)$ **do**
    Solve via $\mathcal{D}$: $\widetilde{y}_k(T) = 0, t \in [0, T]$.
    $\frac{\mathrm{d}\widetilde{y}_k(t)}{\mathrm{d}t} = -\widetilde{\lambda}(t)\left(\widetilde{\phi}(\widetilde{y}_k(t)) - \widetilde{\phi}(\widetilde{y}_{k-1}(t))\right)$,
    $Y \leftarrow Y \cup \{\widetilde{y}_k(t)\}$
  return $Y \setminus \widetilde{y}_0(t)$

Algorithm 1 is a meta-algorithm in the sense that the estimators $\widetilde{\lambda}(t)$ and $\widetilde{\phi}$ must be defined for a full specification. These estimators must be accurate so as to give the correct solution and be efficient to evaluate, as the numerical ODE solver will call these functions many times. In Section 3.2

we define $\widetilde{\lambda}(t)$ and in Section 3.3 we define $\widetilde{\phi}$ appropriately. Taken together with Algorithm 1 this defines the *Nonparametric Sequential Allocation Algorithm*, which we designate by NPSA for the remainder of the paper.

3.2 Estimation of Non-Homogeneous Poisson Processes

In this section we discuss estimation of the non-homogeneous Poisson process P with rate function $\lambda(t) > 0$ for all $t \in [0, T]$. We make the assumption that we have M i.i.d. observed realisations of this process. In this case, we adopt the well known technique of [Law and Kelton, 1991] specialised by [Henderson, 2003]. Briefly, the rate function estimator is taken to be piecewise constant, with breakpoints spaced equally according to some fixed width δ.

Denote by $\widetilde{\lambda}^{(M)}(t)$ the estimator of $\lambda(t)$ by M independent realisations of P. Let the subinterval width used by the estimator be $\delta_M > 0$. We denote by $C_i(a, b)$ the number of jobs arriving in the interval $[a, b)$ in the i^{th} independent realisation of P. For $t \geq 0$, let $\ell(t) := \lfloor t/\delta_M \rfloor \cdot \delta_M$ so that $t \in [\ell(t), \ell(t) + \delta_M]$. Our estimator is the number of arrivals recorded within a given subinterval, averaged over independent realisations of P and normalised by the binwidth δ_M, that is,

$$\widetilde{\lambda}^{(M)}(t) = \frac{1}{M\delta_M} \sum_{i=1}^{M} C_i(\ell(t), \ell(t) + \delta_M). \tag{1}$$

From [Henderson, 2003, Remark 2] we have the following result.

Theorem 2 (Arrival rate estimator convergence)**.** *Suppose that $\delta_M = O(M^{-a})$ for any $a \in (0, 1)$ and fix $t \in [0, T)$. Then, $\widetilde{\lambda}^{(M)}(t) \to \lambda(t)$ almost surely as $M \to \infty$.*

For the NPSA algorithm we use Eq. (1) with $\delta_M = T \cdot M^{-\frac{1}{3}}$ as the estimator for the intensity $\lambda(t)$. There are $\lceil T/\delta_M \rceil$ ordered subintervals, so the time complexity of evaluating $\widetilde{\lambda}(t)$ is $O(\log(T/\delta_M))$ and the space complexity is $O(T/\delta_M)$, owing respectively to searching for the correct subinterval $[\ell(t), \ell(t) + \delta_M]$ via binary search and storing the binned counts C_i . The initial computation of the C_i incurs a time cost of $O(MN_{\text{max}})$, where N_{max} denotes the maximum number of jobs over the M realisations.

3.3 Mean Shortage Function Estimator

The following result (with proof in the Supplementary Material) leads us to the $\widetilde{\phi}$ estimator for NPSA.

Lemma 1. *The mean shortage function of Definition 1 can be written as $\phi(y) = \int_y^\infty (1 - F(x))\mathrm{d}x$, where F is the cdf of the random variable X.*

Lemma 1 suggests the following estimator: perform the integral in Lemma 1, replacing the cdf F with the *empirical cdf* for the job value r.v. X computed with the samples $(x_1, \ldots, x_N)$, $F_N(x) := \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{x_i \leq x}$, where $\mathbf{1}_\omega$ is the indicator variable for an event ω. Since the empirical cdf is piecewise constant, the integral is given by the sum of areas of $O(N)$ rectangles. Concretely, we cache the integral values ϕ_i evaluated at each data sample x_i and linearly interpolate for intermediate $y \in [x_i, x_{i+1})$ at evaluation time. Indeed, after initial one-time preprocessing, this estimate $\widetilde{\phi}_N(y)$ has a runtime complexity of $O(\log N)$ per function call (arising from a binary search of the precomputed values) and space complexity $O(N)$, where N is the number of data samples used for estimation. Pseudocode for these computations is given in the Supplementary Material.

We have shown that the NPSA mean-shortage function estimator is computationally efficient. It now remains to show that it is accurate, that is, statistically consistent.

Theorem 3. *Let X be a nonnegative random variable with associated mean shortage function ϕ. Then, the estimate of the mean shortage function converges in probability to the true value, that is,*

$$\lim_{N\to\infty} \mathbb{P}\left[\sup_{y\geq 0} \left|\widetilde{\phi}_N(y) - \phi(y)\right| > \epsilon\right] = 0$$

for any $\epsilon > 0$, where the estimate computed by the estimator using N independent samples of X is denoted by $\widetilde{\phi}_N(y)$.

Proof Sketch. It can be shown that an upper-bound on $|\widetilde{\phi}_N(y) - \phi(y)|$ is induced by an upper-bound on $|F_N(x) - F(x)|$. The Dvoretzky–Kiefer–Wolfowitz inequality [Dvoretzky *et al.*, 1956; Massart, 1990] furnished with this bound yields the result. □

3.4 NPSA Performance Bounds

We have shown that the individual components of the NPSA algorithm, namely the intensity $\widetilde{\lambda}(t)$ and mean shortage $\widetilde{\phi}(y)$ estimators, are computationally efficient and statistically consistent. However, our main interest is in the output of the overall NPSA algorithm, that is, will following the derived threshold curves at test time yield an expected reward that is optimal with high probability? The answer to this question is affirmative under the assumptions of the SeqAlloc-NP problem setup as described in Section 2.

We will need some results on approximation of ODEs. Following the presentation of [Brauer, 1963], consider the initial value problem

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(t, x), \tag{2}$$

where x and f are d-dimensional vectors and $0 \leq t < \infty$. Assume that $f(t, x)$ is continuous for $0 \leq t < \infty$, $\|x\| < \infty$ and $\|\cdot\|$ is a norm. Recall that a continuous function $x(t)$ is an *ϵ-approximation* to (2) for some $\epsilon \geq 0$ on an interval if it is differentiable on an interval I apart for a finite set of points S, and $\|\frac{\mathrm{d}x(t)}{\mathrm{d}t} - f(t, x(t))\| \leq \epsilon$ on $I \setminus S$. The function $f(t, x)$ satisfies a *Lipschitz condition* with constant L_f on a region $D \subset \mathbb{R} \times \mathbb{R}^d$ if $\|f(t, x) - f(t, x')\| \leq L_f\|x - x'\|$ whenever $(t, x), (t, x') \in D$. We will require the following lemma from [Brauer, 1963].

Lemma 2. *Suppose that $x(t)$ is a solution to the initial value problem* (2) *and $x'(t)$ is an ϵ-approximate solution to* (2)*. Then*

$$\|x(t) - x'(t)\| \leq \|x(0) - x'(0)\|e^{L_f t} + \tfrac{\epsilon}{L_f}(e^{L_f t} - 1),$$

where L_f is the Lipschitz-constant of $f(t,x)$.

Now consider two instantiations of the problem setup with differing parameters, which we call *scenarios*: one in which the job values are nonnegative r.v.s X with mean μ, cdf F and mean shortage function ϕ; in the other, the job values are nonnegative r.v.s X' with mean μ', cdf F' and mean shortage function ϕ'. We stipulate that X and X' have the same support and admit the (bounded) densities f and f' respectively. In the first scenario the jobs arrive with intensity function $\lambda(t) > 0$ and in the second they arrive with intensity $\lambda'(t) > 0$. In both scenarios there are n workers. We are to use the preceding results to show that the difference between threshold curves $|y_k(t) - y'_k(t)|$ computed between these two scenarios via NPSA can be bounded by a function of the scenario parameters.

We further stipulate that the scenarios do not differ by too great a degree, that is, $(1-\delta_\lambda)\lambda(t) \le \lambda'(t) \le (1+\delta_\lambda)\lambda(t)$ for all $t \in [0,T]$ and $(1-\delta_\phi)\phi(y) \le \phi'(y) \le (1+\delta_\phi)\phi(y)$ for all $y \in [0,\infty)$, where $0 < \delta_\lambda, \delta_\phi < 1$. Moreover, define $\lambda_{\max} = \max_{t\in[0,T]}\{\lambda(t)\}$. We are led to the following result.

Lemma 3. *For any $k \in \{1,\dots,n\}$, $y'_k(t)$ is an ϵ-approximator for $y_k(t)$ when $\epsilon > 2\mu\lambda_{max}(\delta_\phi + \delta_\lambda)$.*

Proof Sketch. Upper bound the left-hand-side of

$$\left|\lambda'(t)\big(\phi'(y'_{k+1}) - \phi'(y'_k)\big) - \lambda(t)\big(\phi(y'_{k+1}) - \phi(y'_k)\big)\right| < \epsilon,$$

where the ODE description of y_k and y'_k is employed from Theorem 1 and use the definition of an ϵ-approximator. □

We are now able to compute a general bound on the difference between threshold curves derived from slightly differing scenarios.

Lemma 4. *For any $k \in \{1,\dots,n\}$*

$$|y_k(t) - y'_k(t)| \le (\delta_\lambda + \delta_\phi)\mu\Big(e^{2\lambda_{max}(T-t)} - 1\Big).$$

Proof Sketch. Compute the Lipschitz constant $2\lambda_{\max}$ for the ODE system of Theorem 1, then use Lemmas 2 and 3. □

Having bounded the difference between threshold curves in two differing scenarios, it remains to translate this difference into a difference in reward. Define

$$H(y) := \textstyle\int_y^\infty x f(x)\mathrm{d}x; \qquad \overline{F}(y) := 1 - F(y), \quad (3)$$

so that $\phi(y) = H(y) - y\overline{F}(y)$ by Lemma 1. We also have from [Albright, 1974, Theorem 2] that for a set of (not necessarily optimal) threshold curves $\{\widetilde{y}_k(t)\}_{k=1}^n$, the expected reward to be gained by replaying the thresholds from a time $t \in [0,T]$ is given by

$$E_k(t;\widetilde{y}_k,\dots,\widetilde{y}_1) = \int_t^T \Big[H(\widetilde{y}_k(\tau)) + \overline{F}(\widetilde{y}_k(\tau))\cdot E_{k-1}(\tau;\widetilde{y}_{k-1},\dots,\widetilde{y}_1)\Big] \times \Big[\lambda(\tau)\exp\big[-\textstyle\int_t^\tau \lambda(\sigma)\overline{F}(\widetilde{y}_k(\sigma))\mathrm{d}\sigma\big]\Big]\mathrm{d}\tau. \quad (4)$$

We also have from [Albright, 1974, Theorem 1] that

$$E_n(t;\widetilde{y}_n,\dots,\widetilde{y}_1) = \textstyle\sum_{k=1}^n \widetilde{y}_k(t) \quad (5)$$

We wish to lower-bound the expected total reward at test time using the thresholds $\{y'_k\}_{k=1}^n$, when the job arrival process has value distribution F and intensity $\lambda(t)$.

We first make the assumptions that the functions H and $\overline{F}$ do not differ too greatly between the critical curves derived for the two scenarios. Concretely, there exist $\epsilon_{\overline{F}}, \delta_{\overline{F}}, \delta_H \in (0,1)$ such that $e^{\delta_H}H(y_k(t)) \ge H(y'_k(t)) \ge e^{-\delta_H}H(y_k(t))$, $e^{\delta_{\overline{F}}}\overline{F}(y_k(t)) \ge \overline{F}(y'_k(t)) \ge e^{-\delta_{\overline{F}}}\overline{F}(y_k(t))$ and $\big|\overline{F}(y'_k(t)) - \overline{F}(y_k(t))\big| \le \epsilon_{\overline{F}}$ for all $k \in \{1,\dots,n\}$ and $t \in [0,T]$. Furthermore, we define the mean arrival rate $\overline{\lambda} := \frac{1}{T}\int_0^T \lambda(t)\mathrm{d}t$. We are then able to prove the following lower bound on the total reward under incorrectly specified critical curves.

Lemma 5. *Let $\delta = \max\{\delta_H, \delta_{\overline{F}}\}$. Then*

$$E_n(t;y'_k,\dots,y'_1) \ge e^{-n(\delta+\overline{\lambda}\epsilon_{\overline{F}}T)}E_n(t;y_n,\dots,y_1).$$

Proof Sketch. Use induction on n and Eq. (4). □

We have the ingredients to prove the main result. There is a technical difficulty to be overcome, whereby we need to push additive errors from Lemma 4 through to the multiplicative errors required by Lemma 5. This is possible when the functions H, $\overline{F}$ and ϕ are all Lipschitz continuous and have positive lower-bound on $\bigcup_{k=1}^n \mathrm{Range}\,(y_k(t)) \cup \mathrm{Range}\,(y'_k(t))$, facts which are established rigorously in the Supplementary Material.

As a shorthand, we denote the expected reward gained by using the optimal critical curves by $r^\star := E_n(0;y_n,\dots,y_1)$. Moreover, let the critical curves computed by NPSA from M job arrival process realisations be $\{\widetilde{y}_k^{(M)}\}_{k=1}^n$ and the associated expected reward under the true data distribution be the random variable $R^{(M)} := E_n(0;\widetilde{y}_n^{(M)},\dots,\widetilde{y}_1^{(M)})$.

Theorem 4. *Fix an arbitrary $\epsilon \in (0,1)$. Then,*

$$\lim_{M\to\infty}\mathbb{P}\left[\frac{R^{(M)}}{r^\star} \ge 1-\epsilon\right] = 1.$$

Proof Sketch. Using Lemma 5 one can lower bound the probability by

$$1 - \mathbb{P}\big[\delta_H > \tfrac{2\epsilon}{n}\big] - \mathbb{P}\big[\delta_{\overline{F}} > \tfrac{2\epsilon}{n}\big] - \mathbb{P}\Big[\epsilon_{\overline{F}} > \tfrac{2\epsilon}{n\overline{\lambda}T}\Big]. \quad (6)$$

Lemma 4 and Theorem 3 can be used to show that the third term of (6) vanishes as $M \to \infty$. The first and second term also vanish by pushing through the additive error from Theorem 3 and Lemma 4 into multiplicative errors, then verifying these quantities are small enough when M is sufficiently large. □

Theorem 4 demonstrates that the NPSA algorithm solves the SeqAlloc-NP problem optimally when the number of realisations of the job arrival process, M, is sufficiently large.

4 Experiments

We now empirically validate the efficacy of the NPSA algorithm. Three experiments are conducted: *i.* observing the convergence of NPSA to optimality; *ii.* assessing the impact on NPSA performance when the job value distribution F and

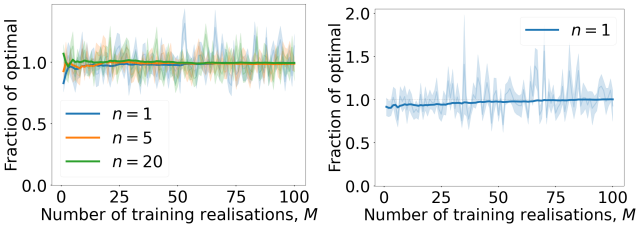

Figure 1: Convergence experiments for job values that are exponentially distributed (left) with mean $\mu = 5$ and job values that are Lomax-distributed (right) with shape $\alpha = 3.5$ and scale $\xi = 5$. The time horizon $T = 2\pi$ and the job arrival rate is $\lambda = 1$. The horizontal dashed line at $y = 1$ indicates optimal reward. The rolling (Cesàro) average is drawn with thick lines to highlight convergence.

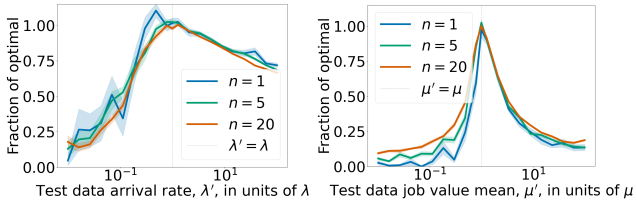

Figure 2: Robustness experiments for NPSA. Expected reward is evaluated on the processes obtained by independently varying λ' and μ' at test time. A y-value of 1 corresponds to the best possible expected reward. The curvature of the plots at $x = 10^0 = 1$ shows how robust the algorithm is with respect to changes in arrival intensity (left) and value mean (right), with small curvature indicating robustness and large curvature showing the opposite.

the arrival intensity $\lambda(t)$ of the data-generating process differ between training and test time; and finally *iii.* applying NPSA to public fraud data and evaluating its effectiveness in detection of the most valuable fraudulent transactions.

Convergence to Optimality. We require a job value distribution F and arrival intensity $\lambda(t)$ such that we can derive the optimal reward *exactly*. We fit $\widetilde{\phi}$ and $\widetilde{\lambda}$ using M simulated realisations of the job arrival process. The reward observed from using NPSA-derived thresholds on further simulations is then compared to the known optimal reward as M grows.

Part of the motivation for the development of NPSA stems from the intractability of exactly solving the system of ODEs necessitated by Theorem 1 for the optimal critical curves. This strictly limits the F and $\lambda(t)$ that we can use for this experiment. Thus, we restrict the job arrival process to be homogeneous, that is, $\lambda(t) = \lambda$ for all $t \in [0, T]$. We consider two job-value distributions, *i.* exponential, that is,

$$F(z) = 1 - e^{-\frac{z}{\mu}}, \quad \phi(z) = \mu e^{-\frac{z}{\mu}},$$

where μ is the mean job value; and *ii.* Lomax, that is,

$$F(z) = 1 - (1 + \tfrac{z}{\xi})^{-\alpha}, \quad \phi(z) = \tfrac{\xi^\alpha z + \xi^{\alpha+1}}{(\alpha-1)(\xi+z)^\alpha},$$

where $\alpha > 0$ is the shape parameter and $\xi > 0$ is the scale. The exponential distribution is the "simplest" distribution in the maximum entropy sense for a nonnegative r.v. with known mean. Lomax-distributed r.v.s are related to exponential r.v.s by exponentiation and a shift and are heavy-tailed.

Using the SageMath [The Sage Developers, 2020] interface to Maxima [Maxima, 2014], we are able to symbolically solve for the optimal thresholds (using Theorem 1) when $n \leq 20$ for exponentially distributed job values and $n = 1$ for Lomax-distributed job values. The optimal reward $r^\star$ is computed using the identity $r^\star = \sum_{k=1}^{n} y_k(0)$ from (5). We then simulate the job arrival process for $M \in \{1, \ldots, 100\}$ independent realisations. For each M, NPSA critical curves are derived using the M realisations. Then, using the same data-generating process $M' = 50$ independent realisations are played out, recording the cumulative reward obtained. The empirical mean reward over the M' simulations is computed along with its standard error and is normalised relative to $r^\star$, for each M.

The result is plotted in Figure 1 for both exponentially and Lomax-distributed jobs. We observe that in both cases convergence is rapid in M. Convergence is quicker in the exponential case, which we attribute to the lighter tails than in the Lomax case, where outsize job values are more often observed that may skew the empirical estimation of ϕ. In the exponential case we observe that convergence is quicker as n increases, which we attribute to noise from individual workers' rewards being washed out by their summation. We further note that we have observed these qualitative features to be robust to variation of the experimental parameters.

Data Distribution Shift. In this experiment, jobs arrive over time horizon $T = 2\pi$ according to a homogeneous Poisson process with fixed intensity $\lambda = 500$ and have values that are exponentially distributed with mean $\mu = 200$. We simulate $M = 30$ realisations of the job arrival process and derive critical curves via NPSA. We then compute modifiers δ_j for $j \in \{1, \ldots, 20\}$, where the δ_j are logarithmically spaced in the interval $[10^{-2}, 10^2]$. The modifiers δ_j give rise to $\lambda'_j = \delta_j \cdot \lambda$ and $\mu'_j = \delta_j \cdot \mu$. We fix a $j \in \{1, \ldots, 20\}$. Holding μ (resp. λ) constant, we then generate $M' = 20$ realisations of the job arrival process with arrival rate λ'_j (resp. mean job value μ'_j) during which we accept jobs according to the thresholds derived by NPSA for μ, λ. The mean and standard error of the reward over the M' realisations is recorded and normalised by the optimal reward for the true data generating process at test time, $r'^\star$.

The result is shown in Figure 2. Note first that the reward is very robust to variations in arrival rate. Indeed, using thresholds that have been derived for an arrival process where the rate differs by an order of magnitude (either an increase or decrease) incurs a relatively small penalty in reward (up to 60%), especially when the jobs at test time arrive more frequently than during training time. The reward is less robust with respect to variations in the mean of the job value, wherein a difference by an order of magnitude corresponds to $\approx 80\%$ loss of reward when $n = 20$. Nevertheless, for more modest deviations from the true μ value the reward is robust.

Evaluation on Public Fraud Data. We augment the SeqAlloc-NP problem setup in Section 2 with the following. Each job (transaction) is endowed with a feature $x \in \mathcal{X}$ and a true class label $y \in \{0, 1\}$. The decision-maker has access to x when a job arrives, but not the true label y. They also have access to a *discriminator*, $D : \mathcal{X} \to [0, 1]$ that represents a

Dataset	M_{train}	M_{test}	$N^{\text{tot}}_{\text{daily}}$	$N^{\text{fraud}}_{\text{daily}}$	$v^{\text{fraud}}_{\text{daily}}$	`clf` F_1-score
`cc-fraud`	2	1	94,935	122	17,403	0.9987
`ieee-fraud`	114	69	$2{,}853 \pm 54$	103 ± 4	$16{,}077 \pm 738$	0.8821

Table 1: Dataset properties for Figure 3 experiments. Each dataset has M_{train} realisations of training data and M_{test} realisations for testing. There are $N^{\text{tot}}_{\text{daily}}$ transactions per day in the test data, out of which $N^{\text{fraud}}_{\text{daily}}$ are fraudulent, with a total monetary value of $v^{\text{fraud}}_{\text{daily}}$. We indicate the F_1-score of the `clf` classifier on the training set.

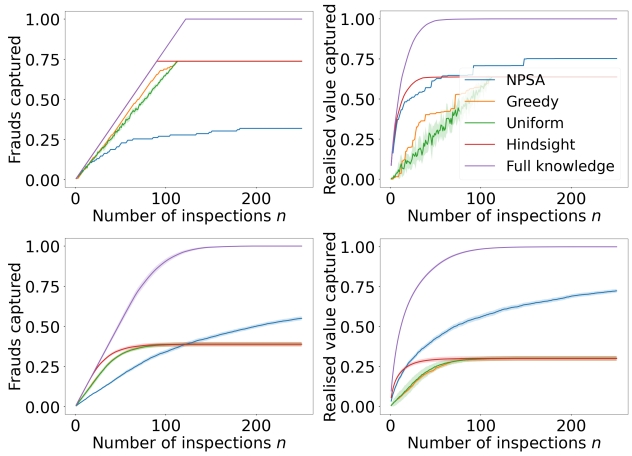

Figure 3: Fraud detection results for `cc-fraud` (top) and `ieee-fraud` (bottom) datasets. The left plots show the fraction of daily fraudulent transactions captured and the right show the fraction of fraudulent monetary value captured.

subjective assessment of probability of a job with side information x being a member of the positive class, $\mathbb{P}[y = 1 \mid x]$. The *adjusted value* of a job $V(x, v)$, where v is the job value, is given by the *expected utility*, $V(x, v) = D(x) \cdot v$, where we stipulate that a job being a member of the positive class yields utility v, being a member of the negative class yields zero utility and the decision-maker is risk-neutral. The decision-maker now seeks to maximise total expected utility.

The reference time frame T is set to one day and the individual realisations are split into M_{test} test and M_{train} training realisations, ensuring all test realisations occur chronologically after all training realisations. A classifier `clf` is trained on the transactions in the training set using the F_1-score as a loss function. This yields the discriminator $D(\,\cdot\,) \equiv$ `clf.predict_proba(·)` where `clf` has a scikit-learn [Pedregosa *et al.*, 2011]-type interface. For each transaction with side information x and value v, we compute the adjusted value $V(x, v)$. The mean-shortage function $\widetilde{\phi}$ for the adjusted job value distribution is learned on this data via the scheme described in Section 3.3. Given $\widetilde{\lambda}(t)$ and $\widetilde{\phi}$, we derive critical curves via NPSA for $n \in \{1, \ldots, 250\}$, which are replayed on the M_{test} test realisations. Full details of the `clf` training and data preparation procedure are given in the Supplementary Material.

We are interested in two quantities: *i.* the total monetary value of inspected transactions that are truly fraudulent, which we call *realised value* and *ii.* how many are truly fraudulent, or *captured frauds*. We compare these quantities obtained from the NPSA algorithm with those obtained from a number of baselines, in order of increasing capability: *i. Greedy.* Choose the first n transactions `clf` marks as having positive class; *ii. Uniform.* From all the transactions `clf` marks as positive class, choose n transactions uniformly at random; *iii. Hindsight.* From all the transactions `clf` marks as positive class, choose the n transactions with highest monetary value; *iv. Full knowledge.* From all the transactions with $y = 1$, choose the n transactions with highest monetary value. Note that *iv.* is included to serve as an absolute upper-bound on performance. We use two public fraud detection datasets, which we denote `cc-fraud` [Dal Pozzolo *et al.*, 2015] and `ieee-fraud` [IEEE-CIS, 2019]. The relevant dataset properties are given in Table 1.

The results are shown in Figure 3. First observe that NPSA shows favourable results even when trained on two realisations ($M_{\text{train}} = 2$ for the `cc-fraud` dataset), outperforming even the *Hindsight* baseline for $n \geq 60$ in terms of captured realised value. On the `ieee-fraud` dataset with $M_{\text{train}} = 114$, NPSA is outperformed only by *Full Knowledge* after $n \geq 15$. In terms of the number of captured frauds, the intuition that NPSA is waiting to inspect only the most valuable transactions to select is validated, evidenced by the the NPSA curve in these plots lying below the baseline curves, contrasted with the high realised value.

5 Conclusion

In this work we introduce the SeqAlloc-NP problem and its efficient, provably optimal solution via the NPSA algorithm. Given M independent realisations of a job arrival process, we are able to optimally select the n most valuable jobs in real-time assuming the incoming data follows the same arrival process. This algorithm is robust to variations in the data-generating process at test-time and has been applied to the financial fraud detection problem, when the value of each transaction is evaluated in a risk-neutral manner.

Future work will go down several paths: including investigating risk-hungry and risk-averse decision-makers; studying adversarial job arrival processes; addressing the effect of jobs taking up a finite time; and specialising to different application domains.

References

[Albright, 1974] C. S. Albright. Optimal sequential assignments with random arrival times. *Management Science*, 21(1):60–67, 2020/10/21/ 1974.

[Altman, 1999] E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

[Avenhaus *et al.*, 1996] R. Avenhaus, M. J. Canty, and F. Calogero. *Compliance Quantified: An Introduction to Data Verification*. Cambridge University Press, 1996.

[Bahnsen *et al.*, 2014] A. C. Bahnsen, D. Aouada, and B. Ottersten. Example-dependent cost-sensitive logistic regression for credit scoring. In *Proc. ICMLA*, pages 263–269, 2014.

[Beyer *et al.*, 2016] B. Beyer, C. Jones, J. Petoff, and N. R. Murphy. *Site Reliability Engineering: How Google Runs Production Systems*. O'Reilly Media, Inc., 1st edition, 2016.

[Bolton and Hand, 2002] R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statist. Sci.*, 17(3):235–255, 08 2002.

[Brauer, 1963] F. Brauer. Bounds for solutions of ordinary differential equations. *Proc. American Mathematical Society*, 14(1):36–43, 1963.

[Consultants, 2019] Cambridge Consultants. Use of AI in Content Moderation. *Produced on behalf of Ofcom*, 2019.

[Dal Pozzolo *et al.*, 2015] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi. Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 159–166, 2015.

[Dal Pozzolo, 2015] A. Dal Pozzolo. *Adaptive machine learning for credit card fraud detection*. PhD thesis, 2015.

[DeGroot, 1970] M. H. DeGroot. *Optimal statistical decisions*. McGraw-Hill, New York, NY, 1970.

[Derman *et al.*, 1972] C. Derman, G. J. Lieberman, and S. M. Ross. A sequential stochastic assignment problem. *Management Science*, 18(7):349–355, 1972.

[Dupuis and Wang, 2002] Paul Dupuis and Hui Wang. Optimal stopping with random intervention times. *Advances in Applied Probability*, 34(1):141–157, 2002.

[Dvoretzky *et al.*, 1956] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27(3):642–669, 09 1956.

[Efroni *et al.*, 2020] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *CoRR*, abs/2003.02189, 2020.

[Elkan, 2001] C. Elkan. The foundations of cost-sensitive learning. In *Proc. IJCAI*, page 973–978, 2001.

[Henderson, 2003] S. G. Henderson. Estimation for nonhomogeneous poisson processes from aggregated data. *Operations Research Letters*, 31(5):375 – 382, 2003.

[IEEE-CIS, 2019] IEEE Computational Intelligence Society. IEEE-CIS. IEEE-CIS Fraud Detection, 2019. https://www.kaggle.com/c/ieee-fraud-detection/datasets.

[Jenatton *et al.*, 2016] R. Jenatton, J. Huang, and C. Archambeau. Adaptive algorithms for online convex optimization with long-term constraints. In *Proc. ICML*, volume 48, pages 402–411, 2016.

[Khoshkhou, 2014] G.B. Khoshkhou. *Stochastic sequential assignment problem*. PhD thesis, University of Illinois at Urbana-Champaign, 2014.

[Law and Kelton, 1991] A. M. Law and D. W. Kelton. *Simulation modeling and analysis*. McGraw-Hill, 2nd edition, 1991.

[Mannor and Shimkin, 2004] Shie Mannor and Nahum Shimkin. A geometric approach to multi-criterion reinforcement learning. *JMLR*, 5:325–360, 2004.

[Mannor and Tsitsiklis, 2006] S. Mannor and J. N. Tsitsiklis. Online learning with constraints. In *Learning Theory*, pages 529–543. Springer Berlin Heidelberg, 2006.

[Massart, 1990] P. Massart. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *Ann. Probab.*, 18(3):1269–1283, 07 1990.

[Maxima, 2014] Maxima. *Maxima, a Computer Algebra System. Version 5.34.1*, 2014.

[Morishita and Okumura, 1983] I. Morishita and M. Okumura. Automated visual inspection systems for industrial applications. *Measurement*, 1(2):59 – 67, 1983.

[Pedregosa *et al.*, 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.

[Sakaguchi, 1977] M. Sakaguchi. A sequential allocation game for targets with varying values. *Journal of the Operations Research Society of Japan*, 20(3):182–193, 1977.

[Shen and Kurshan, 2020] H. Shen and E. Kurshan. Deep Q-Network-based Adaptive Alert Threshold Selection Policy for Payment Fraud Systems in Retail Banking. In *Proc. ICAIF*, 2020.

[The Sage Developers, 2020] The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 9.2)*, 2020.

[Vaněk *et al.*, 2012] O. Vaněk, Z. Yin, M. Jain, B. Bošanský, M. Tambe, and M. Pěchouček. Game-theoretic resource allocation for malicious packet detection in computer networks. In *Proc. AAMAS*, page 905–912, 2012.

[Zhao *et al.*, 2020] Peng Zhao, Guanghui Wang, Lijun Zhang, and Zhi-Hua Zhou. Bandit convex optimization in non-stationary environments. In *Proc. AISTATS*, volume 108, pages 1508–1518, 2020.

[Zheng and Ratliff, 2020] Liyuan Zheng and Lillian J. Ratliff. Constrained upper confidence reinforcement learning. *CoRR*, abs/2001.09377, 2020.