
Kolmogorov Approximation

Anonymous Author(s)

Affiliation

Address

email

1 Introduction

Many different approaches to approximation of probability distributions are studied in the literature [9, 12, 13]. The papers vary in the types random variables involved, how they are represented, and in the criteria used for evaluation of the quality of the approximations. This paper is on approximating discrete distributions represented as explicit probability mass functions with ones that are simpler to store and to manipulate. This is needed, for example, when a discrete distribution is given as a large data-set, obtained, e.g., by sampling, and we want to represent it approximately with a small table.

The main contribution of this paper is an efficient algorithm for computing the best possible approximation of a given random variable with a random variable whose complexity is not above a prescribed threshold, where the measures of the quality of the approximation and the complexity of the random variable are as specified in the following two paragraphs.

We measure the quality of an approximation by the distance between the original variable and the approximate one. Specifically, we use the Kolmogorov distance which is one of the most used in statistical practice and literature. Given two random variables X and X' whose cumulative distribution functions (cdfs) are F_X and $F_{X'}$, respectively, the Kolmogorov distance between X and X' is $d_K(X, X') = \sup_t |F_X(t) - F_{X'}(t)|$ (see, e.g., [8]). We say that X' is a good approximation of X if $d_K(X, X')$ is small.

The complexity of a random variable is measured by the size of its support, the number of values that it can take, $|\text{support}(X)| = |\{x: \Pr(X = x) \neq 0\}|$. When distributions are maintained as explicit tables, as done in many implementations of statistical software, the size of the support of a variable is proportional to the amount of memory needed to store it and to the complexity of the computations around it.

In summary, the exact notion of optimality of the approximation targeted in this paper is:

Definition 1. A random variable X' is an optimal m -approximation of a random variable X if $|\text{support}(X')| \leq m$ and there is no random variable X'' such that $|\text{support}(X'')| \leq m$ and $d_K(X, X'') < d_K(X, X')$.

The main contribution of the paper is a constructive proof of:

Theorem 2. Given a random variable X and a number m , there is an algorithm with memory and time complexity $O(|\text{support}(X)|^2 \cdot m)$ that computes an optimal m -approximation of X .

30 The rest of the paper is organized as follows. In Section 2 we describe how our work relates to other
 31 algorithms and problems studied in the literature. In Section 3 we detail the proposed algorithm,
 32 analyze its properties, and prove Theorem 2. In Section 4 we demonstrate how the proposed approach
 33 performs on the problem of estimating the probability of hitting deadlines is plans and compare it to
 34 alternatives approximation approaches from the literature. We also demonstrate the performance of
 35 our approximation algorithm on some randomly generated random variables. The paper is concluded
 36 with a discussion in Section 5.

37 2 Related Work

38 The problem studied in this paper is related to the theory of Sparse Approximation (aka Sparse
 39 Representation) that deals with sparse solutions for systems of linear equations, as follows.

Given a matrix $D \in \mathbb{R}^{n \times p}$ and a vector $x \in \mathbb{R}^n$, the most studied sparse representation problem is finding the sparsest possible representation $\alpha \in \mathbb{R}^p$ satisfying $x = D\alpha$:

$$\min_{\alpha \in \mathbb{R}^p} \|\alpha\|_0 \text{ subject to } x = D\alpha.$$

40 where $\|\alpha\|_0 = |\{i : \alpha_i \neq 0, i = 1, \dots, p\}|$ is the ℓ_0 pseudo-norm, counting the number of non-zero
 41 coordinates of α . This problem is known to be NP-Hard with a reduction to NP-complete subset
 42 selection problems.

In these terms, using also the ℓ_∞ norm that represents the maximal coordinate and the ℓ_1 norm that represents the sum of the coordinates, our problem can be phrased as:

$$\min_{\alpha \in [0, \infty)^p} \|x - D\alpha\|_\infty \text{ subject to } \|\alpha\|_0 = m \text{ and } \|\alpha\|_1 = 1.$$

43 where D is the all-ones triangular matrix (the entry at row i and column j is one if $i \leq j$ and zero
 44 otherwise), x is related to X such that the i th coordinate of x is $F_X(x_i)$ where $\text{support}(X) =$
 45 $\{x_1 < x_2 < \dots < x_n\}$ and α is related to X' such that the i th coordinate of α is $f_{X'}(x_i)$. The
 46 functions F_X and $f_{X'}$ represent, respectively, the cumulative distribution function of X and the
 47 mass distribution function of X' . This, of course, means that the coordinates of x are assumed to be
 48 positive and monotonically increasing and that the last coordinate of x must be one. We demonstrate
 49 an application for this specific sparse representation problem and show that it can be solve in $O(n^2m)$
 50 time and memory.

51 3 An Algorithm for Optimal Approximation

52 We next describe in details the proof of theorem 2.

53 In the following we set X as a random variable with a finite support of size n , and we set $0 < m \leq n$.
 54 We need to find an m -optimal approximation random variable X' .

55 Our first step is to show that it is enough to limit our search to X' 's such that $\text{support}(X') \subseteq$
 56 $\text{support}(X)$.

57 **Lemma 3.** *There is an m -optimal-approximation X' of X such that $\text{support}(X') \subseteq \text{support}(X)$.*

58 *[DF: This proof is unclear to me, please clean.]* Assume for contradiction is a random variable X''
 59 with support size m such that $d_K(X, X'')$ is minimal but $\text{support}(X'') \not\subseteq \text{support}(X)$. We will
 60 show how to transform X'' support such that it will be contained in $\text{support}(X)$. Let v' be the first

61 $v' \in \text{support}(X'')$ and $v' \notin \text{support}(X)$. Let $v = \max\{i : i < v' \wedge i \in \text{support}(X)\}$. Every v'
62 we will replace with v and name the new random variable X' , we will show that $d_K(X, X'') =$
63 $d_K(X, X')$. First, note that: $F_{X''}(v') = F_{X'}(v)$, $F_X(v') = F_X(v)$. Second, $F_{X'}(v') - F_X(v') =$
64 $F_{X'}(v) - F_X(v)$. Therefore, $d_K(X, X'') = d_K(X, X')$ and X' is also an optimal approximation of
65 X . \square

66 Next, note that every random variable X'' with support of size at most m that is contained in
67 $\text{support}(X)$ be described by first setting the (at most m) elements of the support of X'' ; then for
68 every such option, determine X'' by setting probability values for the elements in the chosen support
69 of X' , and setting 0 for rest of the elements.

70 Since from Lemma 3 we can assume wlog that if X' is an m -optimal approximation variable for
71 X then $\text{support}(X') \subseteq \text{support}(X)$, our search to find such X' takes two steps. Denote the set of
72 random variables with support S by \mathbb{X}_S . In step 1, we find the m -optimal approximation random
73 variable among all random variables in \mathbb{X}_S , and denote the m -optimal distance for \mathbb{X}_S by $\varepsilon(X, S)$.
74 Next, in Step 2, among all the possible supports we find the support setting S of size $\leq m$ for which
75 $\varepsilon(X, S)$ is minimal: We describe an efficient way to do so.

76 3.1 Step 1

77 We first fix a set $S \subseteq \text{support}(X)$ of size at most m , and among all the random variables in \mathbb{X}_S
78 find one with a minimal distance from X . To that, set $S = \{x_1 < \dots < x_m\} \subseteq \text{support}(X)$. To
79 simplify the proofs set $x_0 = -\infty$, and $x_{m+1} = \infty$. Then $x_0 < x_1$ and $x_m < x_{m+1}$. In addition
80 recall that for every random variable X'' $F_{X''}(-\infty) = 0$ and $F_{X''}(\infty) = 1$. For the rest of this
81 section we assume S is fixed and therefore is not necessarily included in the notation.

82 Next, as the elements of S are also elements of $\text{support}(X)$, we can define the following weight
83 function that we use to find the m -optimal distance $\varepsilon(X, S)$.

84 **Definition 4.** For $0 \leq i < m$ let

$$w(x_i, x_{i+1}) = \begin{cases} P(x_i < X < x_{i+1}) & \text{if } i = 0 \text{ or } i = m; \\ P(x_i < X < x_{i+1})/2 & \text{otherwise.} \end{cases}$$

85 Note that when $i = 0$ (resp. $i = m + 1$) then $x_i = -\infty$ (resp. $x_i = \infty$).

86 Finally define:

$$\varepsilon(X, S) = \max_{i=0, \dots, m} w(x_i, x_{i+1}) \quad (1)$$

87 We first show that $\varepsilon(X, S)$ is a lower bound. That is, every random variable in \mathbb{X}_S has a distance at
88 least $\varepsilon(X, S)$. Then, we present a random variable $X' \in \mathbb{X}_S$ with distance $\varepsilon(X, S)$. It then follows
89 that such X' is an m -optimal approximation random variable among all random variables in \mathbb{X}_S .

90 The intuition behind choosing these specific weights and $\varepsilon(X, S)$ being a lower bound is as follows.
91 For every $1 \leq i \leq m$ let \hat{x}_i be the maximal element of $\text{support}(X)$ that is smaller than x_i . Then
92 since for every $X' \in \mathbb{X}_S$ the probability values of X' for the elements not in S are set to 0, we
93 have that $F_{X'}(\hat{x}_{i+1}) = F_{X'}(x_i)$. Therefore the distance between X' and X at points x_i and \hat{x}_{i+1} is
94 increased by $F_X(\hat{x}_{i+1}) - F_X(x_i) = P(x_i < X < x_{i+1})$.

95 Formally we have the following.

96 **Proposition 5.** For every random variable X' with $\text{support}(X') = S$ we have $d_k(X, X') \geq$
 97 $\varepsilon(X, S)$.

98 *Proof.* Let X' be a random variable with support S . Then by definition, for every $0 \leq i \leq m$,
 99 $d_k(X, X') \geq \max\{|F_X(x_i) - F_{X'}(x_i)|, |F_X(\hat{x}_{i+1}) - F_{X'}(\hat{x}_{i+1})|\}$. Note that $F_{X'}(x_i) = F_{X'}(\hat{x}_{i+1})$
 100 since the probability value for all the elements not in S is set to 0.

101 If $i = 0$, that is $x_i = -\infty$, we have that $F_X(x_i) = F_{X'}(x_i) = F_{X'}(\hat{x}_{i+1}) = 0$ and therefore
 102 $d_k(X, X') \geq |F_X(\hat{x}_{i+1}) - F_{X'}(\hat{x}_{i+1})| = |F_X(\hat{x}_{i+1}) - F_X(x_i)| = P(x_i < X < x_{i+1}) = w(x_i, x_{i+1})$.

103 If $i = m$, that is $x_{i+1} = \infty$, note that $F_X(\hat{x}_{i+1}) = F_{X'}(\hat{x}_{i+1}) = 1$. Therefore $F_{X'}(x_i) = 1$ as well.
 104 Therefore $d_k(X, X') \geq |F_X(\hat{x}_i) - F_{X'}(\hat{x}_{i+1}) - F_X(x_i)| = P(x_i < X < x_{i+1}) = w(x_i, x_{i+1})$.
 105 [[DF: fix]]

106 Otherwise for every $1 \leq i < m$, we use the fact that $\max\{|a|, |b|\} \geq |a - b|/2$ for every $a, b \in \mathfrak{R}$, to
 107 have $d_k(X, X') \geq \max\{|F_X(x_i) - F_{X'}(x_i)|, |F_X(\hat{x}_{i+1}) - F_{X'}(\hat{x}_{i+1})|\}$, and therefore $d_k(X, X') \geq$
 108 $1/2|F_X(x_i) - F_X(\hat{x}_{i+1}) + F_{X'}(\hat{x}_{i+1}) - F_{X'}(x_i)|$. Since it is given that $F_{X'}(\hat{x}_{i+1}) - F_{X'}(x_i) =$
 109 $P(x_i < X' < x_{i+1}) = 0$, we have that $d_k(X, X') \geq 1/2|F_X(x_i) - F_X(\hat{x}_{i+1})| = P(x_1 < X <$
 110 $x_2)/2 = w(x_i, x_{i+1})$.

111 We saw that $d_k(X, X') \geq w(x_i, x_{i+1})$ for every $0 \leq i \leq m$. Therefore by definition of $\varepsilon(X, S)$,
 112 proof follows. \square

113 [[DF: here I stopped]]

114 Let X' be defined by $f_{X'}(x_i) = w(x_{i-1}, x_i) + w(x_i, x_{i+1}) + f_X(x_i)$ for $i = 1, \dots, m$ and $f_{X'}(x) =$
 115 0 for $x \notin S$.

116 **Lemma 6.** For $i > 1$, if $F_{X'}(x_i) - F_X(x_i) = w(x_i, x_{i+1})$ then $F_{X'}(x_{i+1}) - F_X(x_{i+1}) =$
 117 $w(x_{i+1}, x_{i+2})$.

Proof.

$$F_{X'}(x_{i+1}) - F_X(x_{i+1}) = \quad (2)$$

$$\begin{aligned} &= f_{X'}(x_{i+1}) - f_X(x_{i+1}) - P(X < x_{i+1}) + P(X' < x_{i+1}) \\ &= f_{X'}(x_{i+1}) - f_X(x_{i+1}) - F_X(x_i) - P(x_i < X < x_{i+1}) + F_{X'}(x_i) \\ &= f_{X'}(x_{i+1}) - f_X(x_{i+1}) - F_X(x_i) - 2w(x_i, x_{i+1}) + F_{X'}(x_i) \end{aligned} \quad (3)$$

$$= f_{X'}(x_{i+1}) - f_X(x_{i+1}) - 2w(x_i, x_{i+1}) + w(x_i, x_{i+1}) \quad (4)$$

$$= w(x_i, x_{i+1}) + w(x_{i+1}, x_{i+2}) - 2w(x_i, x_{i+1}) + w(x_i, x_{i+1}) \quad (5)$$

$$= w(x_{i+1}, x_{i+2})$$

118 By Definition 4 the probability $P(x_{i-1} < X < x_i) = 2w(x_{i-1}, x_i)$ as in Equation (3). Equation (4)
 119 is deduced by the induction hypothesis and Equation (5) where $f_{X'}(x_i) - f_X(x_i) = w(x_{i-1}, x_i) +$
 120 $w(x_i, x_{i+1})$ is true by construction, see Definition?? \square

121 **Lemma 7.** Base case: $i = 1, F_{X'}(x_1) - F_X(x_1) = w(x_1, x_2)$.

Proof.

$$\begin{aligned}
F_{X'}(x_1) - F_X(x_1) &= \\
&= f_{X'}(x_1) - f_X(x_1) - w(x_0, x_1) \\
&= w(x_0, x_1) + w(x_1, x_2) - w(x_0, x_1) \\
&= w(x_1, x_2)
\end{aligned}$$

122

□

123 **Proposition 8.** *There exists X' such that $\text{support}(X') = S$ and $d_k(X, X') = \varepsilon(X, S)$.*

124 3.2 Step 2

125 Chakravarty, Orlin, and Rothblum [2] proposed a polynomial-time method that, given a certain
 126 objective functions (additive), finds an optimal consecutive partition. Their method involves the
 127 construction of a graph such that the (consecutive) set partitioning problem is reduced to the problem
 128 of finding the shortest path in that graph.

129 The KolmogorovApprox algorithm (Algorithm 1) starts by constructing a directed weighted graph
 130 G similar to the method of Chakravarty, Orlin, and Rothblum [2]. The nodes V consist of the
 131 support of X together with an extra two nodes, $-\infty$ and ∞ for technical reasons, whereas the
 132 edges E connect every pair of nodes in one direction (lines 1-2). The weight w of each edge
 133 $e = (x, y) \in E$ is determined by one of two cases as in Definition 4. The first is where nodes
 134 x or y are the source or target nodes respectively. In this case, the weight is the probability of X
 135 to get a value between x and y , non inclusive, i.e., $w(e) = \Pr(x < X < y)$. The second case
 136 is where x and y are not a source or target nodes, here the weight is the probability of X to get a
 137 value between x and y , non inclusive, divided by two i.e., $w(e) = \Pr(x < X < y)/2$. The values
 138 taken are non inclusive, since we are interested only in the error value. The source node of the
 139 shortest path problem at hand corresponds to the $-\infty$ node added to G in the construction phase,
 140 and the target node is the extra node ∞ . The set of all solution paths in G , i.e., those starting at
 141 $-\infty$ and ending in ∞ with at most m edges, is called $\text{paths}(G, -\infty, \infty)$. The goal is to find the
 142 path l in $\text{paths}(G, -\infty, \infty)$ with the lightest bottleneck (line 3). This can be achieved by using the
 143 *Bellman – Ford* algorithm with two tweaks. The first is to iterate the graph G in order to find only
 144 paths with length of at most m edges. The second is to find the lightest bottleneck as opposed to
 145 the traditional objective of finding the shortest path. This is performed by modifying the manner of
 146 “relaxation” to $\text{bottleneck}(x) = \min[\max(\text{bottleneck}(v), w(e))]$, done also in [14]. Consequently,
 147 we find the lightest maximal edge in a path of length $\leq m$, which represents the minimal error,
 148 $\varepsilon(X, S)$, defined in Definition ?? where the nodes in path l represent the elements in set S . The
 149 approximated random variable X' is then derived from the resulting path l (lines 4-5). Every node
 150 $x \in l$ represent a value in the new calculated random variable X' , we then iterate the path l to find
 151 the probability of the event $f_{X'}(x)$ as described in Definition ?. For every edge $(x_i, x_j) \in l$ we
 152 determine: if (x_i, x_j) is the first edge in the path l (i.e. $x_i = -\infty$), then node x_j gets the full weight
 153 $w(x_i, x_j)$ and it's own weight in X such that $f_{X'}(x_j) = f_X(x_j) + w(x_i, x_j)$. If (x_i, x_j) is not the
 154 first nor the last edge in path l then we divide it's weight between nodes x_i and x_j in addition to their
 155 own original weight in X and the probability that already accumulated. If (x_i, x_j) is the last edge
 156 in the path l (i.e. $x_j = \infty$) then node x_i gets the full weight $w(x_i, x_j)$ in addition to what was already
 157 accumulated such that $f_{X'}(x_j) = f_{X'}(x_i) + w(x_i, x_j)$.

158 **Theorem 9.** *KolmogorovApprox(X, m) is an m -optimal-approximation of X .*

Algorithm 1: KolmogorovApprox(X, m)

```
1  $S = \text{support}(X) \cup \{\infty, -\infty\}$ 
2  $G = (V, E) = (S, \{(x, y) : x < y\})$ 
3  $(x_0, \dots, x_{m+1}) = l = \text{argmin}_{l \in \text{paths}(G, -\infty, \infty), |l| \leq m} \max\{w(e) : e \in l\}$ 
4 for  $0 < i < m + 1$  do
5    $f_{X'}(x_i) = w(x_{i-1}, x_i) + w(x_i, x_{i+1}) + f_X(x_i)$ 
6 return  $X'$ 
```

159 **Theorem 10.** *The KolmogorovApprox(X, m) algorithm runs in time $O(mn^2)$, using $O(n^2)$ mem-*
160 *ory where $n = |\text{support}(X)|$.*

161 *Proof.* Constructing the graph G takes $O(n^2)$. The number of edges is $O(E) \approx O(n^2)$ and for every
162 edge the weight is at most the sum of all probabilities between the source node $-\infty$ and the target
163 node ∞ , which can be done efficiently by aggregating the weights of already calculated edges. The
164 construction is also the only stage that requires memory allocation, specifically $O(E + V) = O(n^2)$.
165 Finding the shortest path takes $O(m(E + V)) \approx O(mn^2)$. Since G is DAG (directed acyclic graph)
166 finding shortest path takes $O(E + V)$. We only need to find paths of length $\leq m$, which takes
167 $O(m(E + V))$. Deriving the new random variable X' from the computed path l takes $O(mn)$. For
168 every node in l (at most m nodes), calculating the probability $P(s < X < \infty)$ takes at most n .
169 To conclude, the worst case run-time complexity is $O(n^2 + mn^2 + mn) = O(mn^2)$ and memory
170 complexity is $O(E + V) = O(n^2)$. \square

171 4 A case study and experimental results

172 The case study examined in our experiments is the problem of task trees with deadlines [4, 3].
173 Hierarchical planning is a well-established field in AI [5, 6, 7], and is still relevant nowadays [1, 15].
174 A hierarchical plan is a method for representing problems of automated planning in which the
175 dependency among tasks can be given in the form of networks, here we focus on hierarchical plans
176 represented by task trees. The leaves in a task tree are *primitive* actions (or tasks), and the internal
177 nodes are either *sequence* or *parallel* actions. The plans we deal with are of stochastic nature, where
178 the duration of a primitive action is given by a random variable.

179 A sequence node denotes a series of tasks that should be performed consecutively, whereas a parallel
180 node denotes a set of tasks that begin at the same time. A *valid* plan is one that is fulfilled before some
181 given *deadline*, i.e., its *makespan* is less than or equal to the deadline. The objective in this context
182 is to compute the probability that a given plan is valid, or more formally computing $P(X < T)$,
183 where X is a random variable representing the makespan of the plan and T is the deadline. As said
184 above, resource consumption (task duration) is uncertain, and described as probability distributions
185 in the leaf nodes. We assume that the distributions are independent but *not* necessarily identically
186 distributed and that the random variables are discrete and have a finite support.

187 The problem of finding the probability that a task tree satisfies a deadline is known to be NP-hard. In
188 fact, even the problem of summing a set of random variables is NP-hard [10]. This is an example of
189 an explicitly given random variable that we need to estimate deadline meeting probabilities for.

190 In the first experiment we focus on is the problem of task trees with deadlines, and consider three
191 types of task trees. The first type includes logistic problems of transporting packages by trucks and
192 airplanes (from IPC2 <http://ipc.icaps-conference.org/>). Hierarchical plans of those logistic problems

193 were generated by the JSHOP2 planner [11] (see example problem, Figure 1, one parallel node with
 194 all descendant task nodes being in sequence). The second type consists of task trees used as execution
 195 plans for the ROBIL team entry in the DARPA robotics challenge (DRC simulation phase), and the
 196 third type is of linear plans (sequential task trees). The primitive tasks in all the trees are modeled as
 197 discrete random variables with support of size M obtained by discretization of uniform distributions
 over various intervals. The number of tasks in a tree is denoted by N .

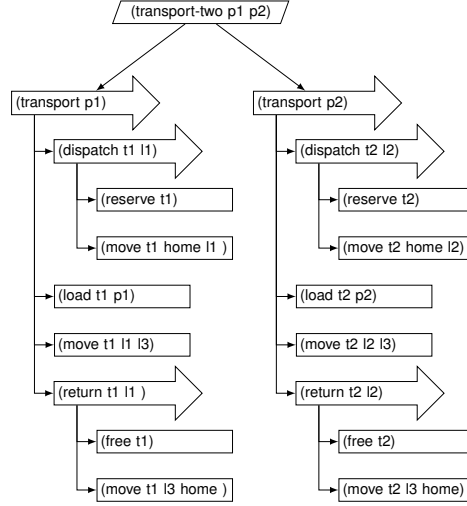


Figure 1: A plan generated by the JSHOP2 algorithm. Arrow shapes represent sequence nodes, parallelograms represent parallel nodes, and rectangles represent primitive nodes.

198

199 We implemented the approximation algorithm for solving the deadline problem with four different
 200 methods of approximation. The first two are for achieving a one-sided Kolmogorov approximation –
 201 the OptTrim [3] and the Trim [4] operators, and the third is a simple sampling scheme. We used those
 202 methods as a comparison to the Kolmogorov approximation with the suggested KolmogorovApprox
 203 algorithm. The parameter m of OptTrim and KolmogorovApprox corresponds to the inverse of ε
 204 given to the Trim operator. Note that in order to obtain some error ε , one must take into consideration
 205 the size of the task tree N , therefore, $m/N = 1/(\varepsilon \cdot N)$. We ran also an exact computation as
 206 a reference to the approximated one in order to calculate the error. The experiments conducted
 207 with the following operators and their parameters: KolmogorovApprox operator with $m = 10 \cdot N$,
 208 the OptTrim operator with $m = 10 \cdot N$, the Trim as operator with $\varepsilon = 0.1/N$, and two simple
 simulations, with a different samples number $s = 10^4$ and $s = 10^6$.

Task Tree	M	KolmogorovApprox	OptTrim	Trim	Sampling	
		$m/N=10$	$m/N=10$	$\varepsilon \cdot N=0.1$	$s=10^4$	$s=10^6$
Logistics ($N=34$)	2	0	0	0.0019	0.007	0.0009
	4	0	0.0046	0.0068	0.0057	0.0005
Logistics ($N=45$)	2	0.0002	0.0005	0.002	0.015	0.001
	4	0	0.003	0.004	0.008	0.0006
DRC-Drive ($N=47$)	2	0	0.004	0.009	0.0072	0.0009
	4	0	0.008	0.019	0.0075	0.0011
Sequential ($N=10$)	2	0.009	0.015	0.024	0	0
	4	0.001	0.024	0.04	0.008	0.0016
	10	0	0.028	0.06	0.0117	0.001

Table 1: Comparison of estimated errors with respect to the reference exact computation on various task trees.

209

Table 1 shows the results of the main experiment. The quality of the solutions provided by using the OptTrim operator are better (lower errors) than those provided by the Trim operator, following the optimality guarantees, but is interesting to see that the quality gaps happen in practice in each of the examined task trees. However, in some of the task trees the sampling method produced better results than the approximation algorithm with OptTrim. Nevertheless, the approximation algorithm comes with an inherent advantage of providing an exact quality guarantees, as opposed to the probabilistic guarantees provided by sampling.

In order to better understand the quality gaps in practice between KolmogorovApprox, OptTrim, and Trim, we investigate their relative errors when applied on single random variables with support size $n = 100$, and different support sizes of the resulting random variable approximation (m). In each instance of this experiment, a random variable is randomly generated by choosing the probabilities of each element in the support from a uniform distribution and then normalizing these probabilities so that they sum to one.

Figure 2 present the error produced by the above methods. The depicted results are averages over several instances (50 instances) of random variables. The curves in the figure show the average error of OptTrim and Trim operators with comparison to the average error of the optimal approximation provided by KolmogorovApprox as a function of m .

According to the depicted results it is evident that increasing the support size of the approximation m reduces the error, as expected, in all three methods. However, errors produced by the KolmogorovApprox are significantly smaller, safe to say, a half of the error produced by OptTrim and Trim, it is clear both in the table (the relative error is mostly above 1) and in the graph.

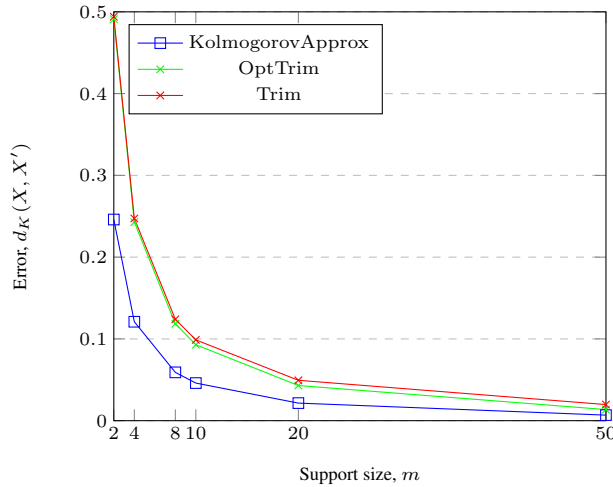


Figure 2: Error comparison between KolmogorovApprox, OptTrim, and Trim, on randomly generated random variables as function of m .

We also examined how our algorithm compares to linear programming as described and discussed, for example, in [12]. We ran an experiment to compare the run-time between the KolmogorovApprox algorithm with the run-time of a state-of-art implementation of linear programming. We used the “Minimize” function of Wolfram Mathematica and fed it with the equations $\min_{\alpha \in \mathbb{R}^n} \|x - \alpha\|_\infty$ subject to $\|\alpha\|_0 \leq m$ and $\|\alpha\|_1 = 1$. The run-time comparison results were clear and persuasive, for a random variable with support size $n = 10$ and $m = 5$, the LP algorithm run-time was 850 seconds, where the KolmogorovApprox algorithm run-time was less than a tenth of a second. For $n = 100$ and $m = 5$, the KolmogorovApprox algorithm run-time was 0.14 seconds and the LP

algorithm took more than a day. Due to these timing results of the LP algorithm we did not proceed to examine it any further. Since it is not trivial to formally analyze the run-time of the LP algorithm, we conclude by the reported experiment that in this case the LP algorithm might not be as efficient as KolmogorovApprox algorithm whose complexity is proven to be polynomial in Theorem 10.

5 Discussion

References

- [1] R. Alford, V. Shivashankar, M. Roberts, J. Frank, and D. W. Aha. Hierarchical planning: Relating task and goal decomposition with task sharing. In *IJCAI*, pages 3022–3029, 2016.
- [2] A. Chakravarty, J. Orlin, and U. Rothblum. A partitioning problem with additive objective with an application to optimal inventory groupings for joint replenishment. *Operations Research*, 30(5):1018–1022, 1982.
- [3] L. Cohen, T. Grinshpoun, and G. Weiss. Optimal approximation of random variables for estimating the probability of meeting a plan deadline. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.
- [4] L. Cohen, S. E. Shimony, and G. Weiss. Estimating the probability of meeting a deadline in hierarchical plans. In *IJCAI*, pages 1551–1557, 2015.
- [5] T. Dean, R. J. Firby, and D. Miller. Hierarchical planning involving deadlines, travel time, and resources. *Computational Intelligence*, 4(3):381–398, 1988.
- [6] K. Erol, J. Hendler, and D. S. Nau. HTN planning: Complexity and expressivity. In *AAAI*, volume 94, pages 1123–1128, 1994.
- [7] K. Erol, J. Hendler, and D. S. Nau. Complexity results for HTN planning. *Annals of Mathematics and Artificial Intelligence*, 18(1):69–93, 1996.
- [8] J. D. Gibbons and S. Chakraborti. Nonparametric statistical inference. In *International encyclopedia of statistical science*, pages 977–979. Springer, 2011.
- [9] A. C. Miller and T. R. Rice. Discrete approximations of probability distributions. *Management Science*, 29(3):352–362, 1983.
- [10] R. Möhring. Scheduling under uncertainty: Bounding the makespan distribution. *Computational Discrete Mathematics*, pages 79–97, 2001.
- [11] D. S. Nau, T.-C. Au, O. Ilghami, U. Kuter, J. W. Murdock, D. Wu, and F. Yaman. SHOP2: An HTN planning system. *Journal of Artificial Intelligence Research*, 20:379–404, 2003.
- [12] K. Pavlikov and S. Uryasev. CVaR distance between univariate probability distributions and approximation problems. Technical Report 2015-6, University of Florida, 2016.
- [13] A. N. Pettitt and M. A. Stephens. The kolmogorov-smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics*, 19(2):205–210, 1977.
- [14] E. Shufan, H. Ilani, and T. Grinshpoun. A two-campus transport problem. In *MISTA*, pages 173–184, 2011.
- [15] Z. Xiao, A. Herzig, L. Perrussel, H. Wan, and X. Su. Hierarchical task network planning with task insertion and state constraints. In *IJCAI*, pages 4463–4469, 2017.