Kolmogorov Approximation

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Many different approaches to approximation of probability distributions are studied in the literature [4, 6, 7]. 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., [3]. We say taht 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, $|\operatorname{support}(X)| = |\{x\colon Pr(X = x) \neq 0\}|$. 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 $|\operatorname{support}(X')| \leq m$ and there is no random variable X'' such that $|\operatorname{support}(X'')| \leq m$ 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 exists an algorithm with memory and time complexity $O(|\operatorname{support}(X)|^2 \cdot m)$ that computes an optimal m-approximation of X.*

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

2 An Algorithm for Optimal Approximation

- We now start our story: Given X and m how can we find X'?
- We first show that it is enough to limit our search to X's such that $\mathrm{support}(X') \subseteq \mathrm{support}(X)$.

Lemma 3. *For any discrete random variable X and any $m \in \mathbb{N}$, there is an m-optimal-approximation X' of X such that* $\mathrm{support}(X') \subseteq \mathrm{support}(X)$.

Proof. Assume there is a random variable X'' with support size m such that $d_K(X, X'')$ is minimal but $\mathrm{support}(X'') \nsubseteq \mathrm{support}(X)$. We will show how to transform X'' support such that it will be contained in $\mathrm{support}(X)$. Let v' be the first $v' \in \mathrm{support}(X'')$ and $v' \notin \mathrm{support}(X)$. Let $v = \max\{i : i < v' \wedge i \in \mathrm{support}(X)\}$. Every v' we will replace with v and name the new random variable X', we will show that $d_K(X, X'') = d_K(X, X')$. First, note that: $F_{X''}(v') = F_{X'}(v)$, $F_X(v') = F_X(v)$. Second, $F_{X'}(v') - F_X(v') = F_{X'}(v) - F_X(v)$. Therefore, $d_K(X, X'') = d_K(X, X')$ and X' is also an optimal approximation of X. ☐

Observation 4. $max\{|a|, |b|\} \geq |a - b|/2$

- The next lemma states a lower bound on the distance $d_K(X, X')$ when a range of elements is excluded from the support of X'.

Lemma 5. *For $x_1, x_2 \in \mathrm{support}(X) \cup \{-\infty, \infty\}$ such that $x_1 < x_2$, if $P(x_1 < X' < x_2) = 0$ then $d_k(X, X') \geq P(x_1 < X < x_2)/2$.*

Proof. Let $\hat{x} = \max\{x \in \mathrm{support}(X) \cap \{-\infty, \infty\} \colon x < x_2\}$. By definition, $d_k(X, X') \geq \max\{|F_X(x_1) - F_{X'}(x_1)|, |F_X(\hat{x}) - F_{X'}(\hat{x})|\}$. From Observation 4, $d_k(X, X') \geq 1/2|F_X(x_1) - F_X(\hat{x}) + F_{X'}(\hat{x}) - F_{X'}(x_1)|$. Since it is given that $F_{X'}(\hat{x}) - F_{X'}(x_1) = P(x_1 < X' < x_2) = 0$, $d_k(X, X') \geq 1/2|F_X(x_1) - F_X(\hat{x})| = P(x_1 < X \leq \hat{x})/2 = P(x_1 < X < x_2)/2$. ☐

- The next lemma strengthen the lower bound.

Lemma 6. *For $x_1, x_2 \in \mathrm{support}(X) \cup \{-\infty, \infty\}$ such that $x_1 = -\infty$ or $x_2 = \infty$, if $P(x_1 < X' < x_2) = 0$ then $d_k(X, X') \geq P(x_1 < X < x_2)$.*

Proof. Let $\hat{x} = \max\{x \in \mathrm{support}(X) \cap \{-\infty, \infty\} \colon x < x_2\}$. By definition $d_k(X, X') \geq \max\{|F_X(x_1) - F_{X'}(x_1)|, |F_X(\hat{x}) - F_{X'}(\hat{x})|\}$. If $x_1 = -\infty$ then $d_k(X, X') \geq \{|F_X(\hat{x}) - F_{X'}(\hat{x})|\}$ since $F_X(-\infty) = F_{X'}(-\infty) = 0$. Furthermore, $F_{X'}(\hat{x}) = P(x_1 < X' < x_2) = 0$. Therefore $d_k(X, X') \geq F_X(\hat{x}) = P(x_1 < X \leq \hat{x}) = P(x_1 < X < x_2)$. If $x_2 = \infty$ then $d_k(X, X') \geq \{|F_X(x_1) - F_{X'}(x_1)|\}$ since $F_X(\hat{x}) = F_{X'}(\hat{x}) = F_X(\infty) = F_{X'}(\infty) = 1$. Furthermore, $F_{X'}(x_1) = 1$ since it is given that $P(x_1 < X' < x_2) = 0$. Therefore we get that $d_k(X, X') \geq |F_X(x_1) - 1| = |1 - F_X(\hat{x}) - | = P(x_1 < X \leq \hat{x}) = P(x_1 < X < x_2)$. ☐

Definition 7. *For $x_1, x_2 \in \mathrm{support}(X) \cup \{-\infty, \infty\}$ let*

$$w(x_1, x_2) = \begin{cases} P(x_1 < X < x_2) & \text{if } x_1 = -\infty \text{ or } x_2 = \infty; \\ P(x_1 < X < x_2)/2 & \text{otherwise.} \end{cases}$$

Definition 8. *For $S = \{x_1 < \cdots < x_m\} \subseteq \mathrm{support}(X)$, $x_0 = -\infty$, and $x_{m+1} = \infty$, let*

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

- From here on, until the end of the section, S is fixed.

Proposition 9. *There is no X' such that* $\mathrm{support}(X') = S$ *and* $d_k(X, X') < \varepsilon(X, S)$.

Proof. Let i be the index that maximizes $w(x_i, x_{i+1})$. If $0 < i < n-1$ then $d_k(X, X') \geq w(x_i, x_{i+1})$ by Lemma 5. If $i = 0$ or $i = n+1$ the same follows from Lemma 6. ☐

Definition 10. *Let X' to by $f_{X'}(x_i) = w(x_{i-1}, x_i) + w(x_i, x_{i+1}) + f_X(x_i)$ for $i = 1, \ldots, m$ and $f_{X'}(x) = 0$ for $x \notin S$.*

Lemma 11. *For $i > 1$, if $F_{X'}(x_i) - F_X(x_i) = w(x_i, x_{i+1})$ then $F_{X'}(x_{i+1}) - F_X(x_{i+1}) = w(x_{i+1}, x_{i+2})$.*

Proof.

$$
\begin{aligned}
&F_{X'}(x_{i+1}) - F_X(x_{i+1}) = \qquad (1)\\
&= f_{X'}(x_{i+1}) - f_X(x_{i+1}) - P(X < x_{i+1}) + P(X' < x_{i+1})\\
&= f_{X'}(x_{i+1}) - f_X(x_{i+1}) - F_X(x_i) - P(x_i < X < x_{i+1}) + F_{X'}(x_i)\\
&= f_{X'}(x_{i+1}) - f_X(x_{i+1}) - F_X(x_i) - 2w(x_i, x_{i+1}) + F_{X'}(x_i) \qquad (2)\\
&= f_{X'}(x_{i+1}) - f_X(x_{i+1}) - 2w(x_i, x_{i+1}) + w(x_i, x_{i+1}) \qquad (3)\\
&= w(x_i, x_{i+1}) + w(x_{i+1}, x_{i+2}) - 2w(x_i, x_{i+1}) + w(x_i, x_{i+1}) \qquad (4)\\
&= w(x_{i+1}, x_{i+2})
\end{aligned}
$$

By Definition 7 the probability $P(x_{i-1} < X < x_i) = 2w(x_{i-1}, x_i)$ as in Equation (2). Equation (3) is deduced by the induction hypothesis and Equation (4) where $f_{X'}(x_i) - f_X(x_i) = w(x_{i-1}, x_i) + w(x_i, x_{i+1})$ is true by construction, see Definition10. ☐

Lemma 12. *Base case: $i = 1$, $F_{X'}(x_1) - F_X(x_1) = w(x_1, x_2)$.*

Proof.

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

☐

Proposition 13. *There exists X' such that* $\mathrm{support}(X') = S$ *and* $d_k(X, X') = \varepsilon(X, S)$.

Chakravarty, Orlin, and Rothblum [1] proposed a polynomial-time method that, given certain objective functions (additive), finds an optimal consecutive partition. Their method involves the construction of a graph such that the (consecutive) set partitioning problem is reduced to the problem of finding the shortest path in that graph.

The $\mathrm{KolmogorovApprox}$ algorithm (Algorithm 2) starts by constructing a directed weighted graph G similar to the method of Chakravarty, Orlin, and Rothblum [1]. The nodes V consist of the support of X together with an extra two nodes, $-\infty$ and ∞ for technical reasons, whereas the edges E connect every pair of nodes in one direction (lines 1-2). The weight w of each edge $e = (i, j) \in E$ is determined by one of two cases as in Definition 7. The first is where i or j are the source or target nodes respectively. In this case the weight is the probability of X to get a value between i

and j, non inclusive, i.e., $w(e) = Pr(i < X < j)$ (lines 4-5). The second case is where i or j are not a source or target nodes, here the weight is the probability of X to get a value between i and j, non inclusive, divided by two i.e., $w(e) = Pr(i < X < j)/2$ (lines 6-7). The values taken are non inclusive, since we are interested only in the error value. The source node of the shortest path problem at hand corresponds to the $-\infty$ node added to G in the construction phase, and the target node is the extra node ∞. The set of all solution paths in G, i.e., those starting at $-\infty$ and ending in ∞ with at most m edges, is called $paths(G, -\infty, \infty)$. The goal is to find the path l in $paths(G, -\infty, \infty)$ with the lightest bottleneck (lines 8-9). This can be achieved by using the $Bellman - Ford$ algorithm with two tweaks. The first is to iterate the graph G in order to find only paths with length of at most m edges. The second is to find the lightest bottleneck as opposed to the traditional objective of finding the shortest path. This is performed by modifying the manner of "relaxation" to $bottleneck(x) = min[max(bottleneck(v), w(e))]$, done also in [8]. Consequently, we find the lightest maximal edge in a path of length $\leq m$, which represents the minimal error, $\varepsilon(X, S)$, defined in Definition 8 where the nodes in path l represent the elements in set S. The approximated random variable X' is then derived from the resulting path l (lines 10-17). Every node $n \in l$ represent a value in the new calculated random variable X', we than iterate the path l to fine the probability of the event $f_{X'}(n)$ as described in Definition 10. For every edge $(i, j) \in l$ we determine: if (i, j) is the first edge in the path l (i.e. $i == -\infty$), then node j gets the full weight $w(i, j)$ and it's own weight in X such that $f_{X'}(j) = f_X(j) + w(i, j)$ (lines 11-12). If (i, j) in not the first nor the last edge in path l then we divide it's weight between nodes i and j in addition to their own original weight in X and the probability that already accumulated (lines 16-17). If (i, j) is the last edge in the path l (i.e. $i == \infty$) then node i gets the full weight $w(i, j)$ in addition to what was already accumulated such that $f_{X'}(j) = f_{X'}(j) + w(i, j)$ (lines 13-14).

Algorithm 1: $\mathrm{KolmogorovApprox}(X, m)$

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

Theorem 14. $\mathrm{KolmogorovApprox}(X, m) = X'$ *where* X' *is an* m*-optimal-approximation.*

Theorem 15. *The* $\mathrm{KolmogorovApprox}(X, m)$ *algorithm runs in time* $O(mn^2)$*, using* $O(n^2)$ *memory where* $n = |\,\mathrm{support}(X)|$.

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

Algorithm 2: $\mathrm{KolmogorovApprox}(X, m)$

$S = \mathrm{support}(X) \cup \{\infty, -\infty\}$
$G = (V, E) = (S, \{e = (x, y) \in S^2 \colon x < y\})$
foreach $e = (x, y) \in E$ **do**
  if $i = \infty$ *OR* $j = -\infty$ **then**
    $w(e) = Pr(i < X < j)$
  else
    $w(e) = Pr(i < X < j)/2$
/* The following can be obtained, e.g., using the Bellman-Ford algorithm */
$l^* = \mathrm{argmin}_{l \in paths(G, -\infty, \infty, |l| \leq m} \max\{w(e) \colon e \in l\}$
foreach $e = (i, j) \in l^*$ **do**
  if $i = -\infty$ **then**
    $f_{X'}(j) = f_X(j) + Pr(i \leq X < j)$
  else if $j == \infty$ **then**
    $f_{X'}(i) = f_{X'}(i) + Pr(i \leq X < j)$
  else
    $f_{X'}(i) = f_{X'}(i) + Pr(i \leq X < j)/2$
    $f_{X'}(j) = f_X(j) + Pr(i \leq X < j)/2$
return X'

3 Experiments and Results

In the first experiment we focus on the problem of task trees with deadlines, and consider three types of task trees. The first type includes logistic problems of transporting packages by trucks and airplanes (from IPC2 http://ipc.icaps-conference.org/). Hierarchical plans of those logistic problems were generated by the JSHOP2 planner [5] (see example problem, Figure 1). The second type consists of task trees used as execution plans for the ROBIL team entry in the DARPA robotics challenge (DRC simulation phase), and the third type is of linear plans (sequential task trees). The primitive tasks in all the trees are modeled as 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.

We implemented the approximation algorithm for solving the deadline problem with four different methods of approximation. The first two are for achieving a one-sided Kolmogorov approximation – the $\mathrm{OptTrim}$ [**?**] and the Trim [2] operators, and the third is a simple sampling scheme. We used those methods as a comparison to the Kolmogorov approximation with the suggested $\mathrm{KolmogorovApprox}$ algorithm. The parameter m of $\mathrm{OptTrim}$ and $\mathrm{KolmogorovApprox}$ corresponds to the inverse of ε given to the Trim operator. Note that in order to obtain some error ε, one must take into consideration the size of the task tree N, therefore, $m/N = 1/(\varepsilon \cdot N)$. We ran also an exact computation as a reference to the approximated one in order to calculate the error. The experiments conducted with the following operators and their parameters: $\mathrm{KolmogorovApprox}$ operator with $m = 10 \cdot N$, the $\mathrm{OptTrim}$ 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 $s = 10^6$.

Table 1 shows the results of the main experiment. The quality of the solutions provided by using the $\mathrm{OptTrim}$ 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 $\mathrm{OptTrim}$. Nevertheless, the approximation algorithm comes

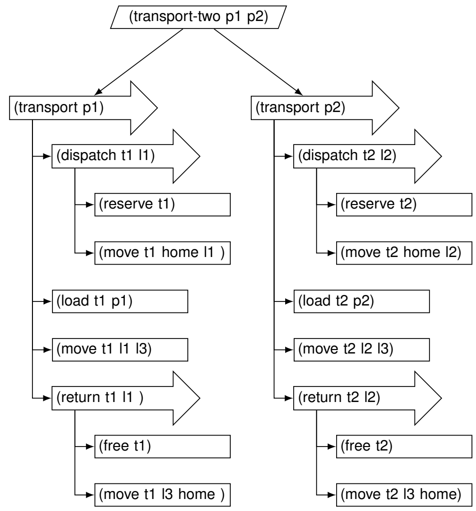

Figure 1: A plan generated by the JSHOP2 algorithm. Arrow shapes represent sequence nodes, parallelograms represent parallel nodes, and rectangles represent primitive nodes.

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.

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 $\mathrm{KolmogorovApprox}$, $\mathrm{OptTrim}$, 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.

Tables 2 and Figure 2 present the error produced by the above methods. The depicted results in the table are averages over several instances of random variables for each entry (50 instances). The columns in the table show the average percentage of the relative error of the $\mathrm{OptTrim}$ and Trim operators with respect to the error of the optimal approximation provided by $\mathrm{KolmogorovApprox}$; the relative error of each instance is calculated by $(\mathrm{OptTrim} \,/\, \mathrm{KolmogorovApprox}) - 1$, $(\mathrm{Trim} \,/\, \mathrm{KolmogorovApprox}) - 1$, respectively. The figure shows the average error of each method, whereas each curve represent a different method 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

m	Relative error Kolmogorov Vs. OptTrim	Relative error Kolmogorov Vs. Trim
2	1.0054	0.994
4	1.0373	1.000
8	1.096	1.002
10	1.1221	0.9946
20	1.2986	1.001
50	1.888	0.994

Table 2: Relative error $\mathrm{KolmogorovApprox}$ Vs. $\mathrm{OptTrim}$ and $\mathrm{KolmogorovApprox}$ Vs. Trim on randomly generated random variables with original support size $n = 100$.

$\mathrm{KolmogorovApprox}$ are significantly smaller, safe to say, a half of the error produced by $\mathrm{OptTrim}$ and Trim, it is clear both in the table (the relative error is mostly above 1) and in the graph.

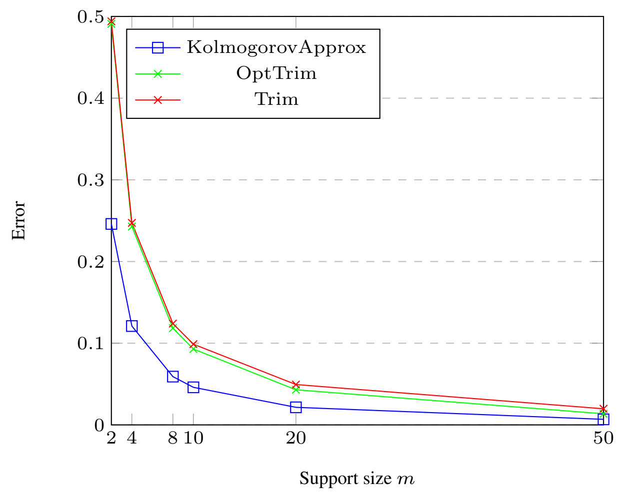

Figure 2: Error comparison between $\mathrm{KolmogorovApprox}$, $\mathrm{OptTrim}$, and Trim, on randomly generated random variables as function of m.

The above experiments display the quality of approximation provided by the $\mathrm{KolmogorovApprox}$ algorithm, as proven before to be optimal approximation under the Kolmogorv metric. One may wonder the need of such an algorithm where the use of linear programing in an easy valid option described and discussed in previews works [6]. In order to address this issue we executed an experiment to compare the run-time between $\mathrm{KolmogorovApprox}$ algorithm and a simple linear programing algorithm. The LP algorithm implemented in Mathematics as follows:.... The run-time comparison results were very clear and persuasive, for a random variable with support size $n = 10$ and $m = 5$, the LP algorithm run-time was 850 sec, where the $\mathrm{KolmogorovApprox}$ algorithm run-time was ≈ 0 sec. Furthermore, for a random variable with support size $n = 100$ and $m = 5$, the $\mathrm{KolmogorovApprox}$ algorithm run-time was 0.14 sec and the LP algorithm took significantly much longer, therefore, due to time limitations of the LP algorithm we did not proceed to examine it any farther. Since it is not trivial to deduce LP algorithm run-time we concluded by the conducted experiment that in this case the LP algorithm might not be as efficient as $\mathrm{KolmogorovApprox}$ algorithm were its run-time is proven to be polynomial 15.

References

[1] 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.

[2] L. Cohen, S. E. Shimony, and G. Weiss. Estimating the probability of meeting a deadline in hierarchical plans. In *IJCAI*, pages 1551–1557, 2015.

[3] J. D. Gibbons and S. Chakraborti. Nonparametric statistical inference. In *International encyclopedia of statistical science*, pages 977–979. Springer, 2011.

[4] A. C. Miller and T. R. Rice. Discrete approximations of probability distributions. *Management Science*, 29(3):352–362, 1983.

[5] 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.

[6] K. Pavlikov and S. Uryasev. CVaR distance between univariate probability distributions and approximation problems. Technical Report 2015-6, University of Florida, 2016.

[7] 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.

[8] E. Shufan, H. Ilani, and T. Grinshpoun. A two-campus transport problem. In *MISTA*, pages 173–184, 2011.