Improved guarantees for the a priori TSP

Jannis Blauth
Research Inst. for Discrete Mathematics, Hausdorff Center for Math., University of Bonn, Germany

Meike Neuwohner
Research Inst. for Discrete Mathematics, Hausdorff Center for Math., University of Bonn, Germany

Luise Puhlmann
Research Inst. for Discrete Mathematics, Hausdorff Center for Math., University of Bonn, Germany

Jens Vygen
Research Inst. for Discrete Mathematics, Hausdorff Center for Math., University of Bonn, Germany

Abstract

We revisit the A PRIORI TSP (with independent activation) and prove stronger approximation guarantees than were previously known. In the A PRIORI TSP, we are given a metric space (V, c) and an activation probability $p(v)$ for each customer $v \in V$. We ask for a TSP tour T for V that minimizes the expected length after cutting T short by skipping the inactive customers.

All known approximation algorithms select a nonempty subset S of the customers and construct a *master route solution*, consisting of a TSP tour for S and two edges connecting every customer $v \in V \setminus S$ to a nearest customer in S.

We address the following questions. If we randomly sample the subset S, what should be the sampling probabilities? How much worse than the optimum can the best master route solution be? The answers to these questions (we provide almost matching lower and upper bounds) lead to improved approximation guarantees: less than 3.1 with randomized sampling, and less than 5.9 with a deterministic polynomial-time algorithm.

2012 ACM Subject Classification Theory of computation → Approximation algorithms analysis

Keywords and phrases A priori TSP, random sampling, stochastic combinatorial optimization

1 Introduction

Many algorithms for stochastic discrete optimization problems sample a sub-instance, solve the resulting deterministic problem (often by some approximation algorithm), and extend this solution to the original instance [7, 10, 13, 14, 15, 25]. A nice and well-studied example is the A PRIORI TRAVELING SALESPERSON PROBLEM (A PRIORI TSP), which is the focus of this paper. What guarantee can we obtain by such an approach, even if we take an optimal sample? If we sample randomly, according to which distribution? What guarantee can we obtain by a deterministic polynomial-time algorithm? These are the questions addressed in this paper.

In the A PRIORI TSP (with independent activation), we are given a (semi-)metric space (V, c); the elements of V are called *customers*. Each customer v comes with an *activation probability* $0 < p(v) \leq 1$, so it will be active independently with probablity $p(v)$. However, we need to design a TSP tour T (visiting all of V) *before* knowing which customers will be active. *After* we know which customers are active we can cut the tour T short by skipping the inactive customers. The goal is to minimize the expected cost of the resulting tour (visiting the active customers).

Note that computing an optimum a priori tour is APX-hard as the metric TSP is APX-hard [23], which is the special case where all activation probabilities are 1. We study approximation algorithms. A *ρ-approximation algorithm* for the A PRIORI TSP is a polynomial-time algorithm that computes a tour of expected cost at most $\rho \cdot \mathrm{OPT}$ for any given instance, where OPT denotes the expected cost of an optimum a priori tour.

Shmoys and Talwar [25] devised a randomized 4-approximation algorithm and a deterministic 8-approximation algorithm. A randomized constant-factor approximation algorithm was discovered independently by Garg, Gupta, Leonardi and Sankowski [10]. The randomized Shmoys–Talwar algorithm easily improves to a 3.5-approximation by using the Christofides–Serdyukov algorithm instead of the double tree algorithm as a subroutine for TSP (as noted by [6]), and slightly better using the new Karlin–Klein–Oveis Gharan algorithm [19]. The deterministic algorithm was improved to a 6.5-approximation by van Zuylen [28]; a slight improvement of this guarantee follows from the recent deterministic version of the Karlin–Klein–Oveis Gharan algorithm [20].

All known approximation algorithms for the A PRIORI TSP are of the following type. Select a nonempty subset S of customers and find a TSP tour for S (the *master tour*). Connect each other customer $v \in V \setminus S$ with a pair of parallel edges to a nearest point $\mu(v)$ in the master tour. We call this a *master route solution*. Once we know the set of active customers, we pay for the entire master tour (pretending to visit also its inactive customers!) and pay $2c(\mu(v), v)$ for each active customer v outside S to cover the round trip visiting v from $\mu(v)$. See Figure 1 for an example. Of course, we could cut the resulting tour shorter (we visit some inactive customers, and we visit some customers several times), but we will not account for this possible gain (unless fewer than two customers are active).

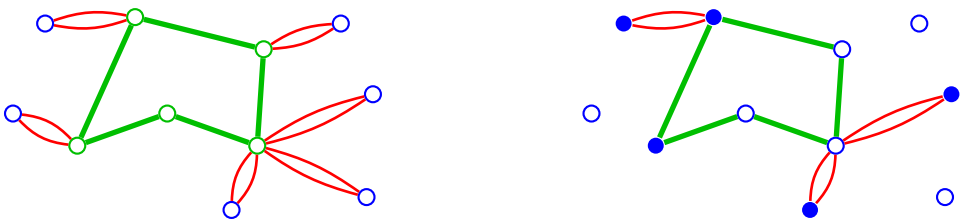


Figure 1 Left: A master route solution with a master tour (green, thick) and connections of the other customers to that master tour (red, curved). Right: After knowing which customers are active (filled), the master route solution reduces to a tour visiting all of the master tour and the other active customers.

1.1 Motivating questions

We start by reviewing the randomized algorithm by Shmoys and Talwar [25]. If fewer than two customers are active, any a priori tour can be cut short to a single point, resulting in cost zero. The algorithm by Shmoys and Talwar [25] selects each customer v independently into S with probability $p(v)$: exactly the activation probability. Assuming that the resulting set S is nonempty, there exists an associated master route solution with expected cost at most

$$\mathrm{MR}(S) \;:=\; \mathbb{E}_{A\sim p}\left[\mathbb{1}_{|A|\geq 2} \cdot \left(\mathrm{OPT}_{\mathrm{TSP}}(S,c) + 2\cdot\sum_{v\in A} c(v,S)\right)\right].$$

Here $\mathrm{OPT}_{\mathrm{TSP}}(S, c)$ denotes the length of an optimum TSP tour for S, and $c(v, S) = \min\{c(v, s) : s \in S\}$ denotes the distance between v and a nearest customer in S (which is zero if $v \in S$); moreover, $\mathbb{E}_{A\sim p}$ denotes the expectation when the set A of active customers is sampled with respect to the given activation probabilities. Later on, $\mathbb{P}_{A\sim p}$ is used analogously. We multiply with $\mathbb{1}_{|A|\geq 2}$ (which is 1 if $|A| \geq 2$ and 0 otherwise) because the cost of the solution is zero if fewer than two customers are active.

If there is a customer d with $p(d) = 1$ (a *depot*), then S is never empty and we can bound

$$\mathrm{MR}(S) \;\leq\; \mathrm{OPT}_{\mathrm{TSP}}(S, c) + 2 \cdot \mathbb{E}_{A\sim p}\left[\sum_{v\in A\setminus\{d\}} c(v, S \setminus \{v\})\right].$$

Note that the above upper bound also accounts for connecting active customers in S to the nearest other customer in S, which is not necessary but will allow the following. Taking the expectation over the random choice of S, an upper bound on the expected cost of that master route solution is

$$\mathbb{E}_{S\sim p}\left[\mathrm{MR}(S)\right] \;\leq\; \mathbb{E}_{S\sim p}\left[\mathrm{OPT}_{\mathrm{TSP}}(S, c)\right] + 2 \cdot \mathbb{E}_{S\sim p}\left[\sum_{v\in S\setminus\{d\}} c(v, S \setminus \{v\})\right]$$

as the probability distributions to choose S and A are identical and the vertices are sampled independently. Since $\sum_{v\in S\setminus\{d\}} c(v, S\setminus\{v\}) \leq \mathrm{OPT}_{\mathrm{TSP}}(S, c)$ for all S and $\mathbb{E}_{S\sim p}\left[\mathrm{OPT}_{\mathrm{TSP}}(S, c)\right] \leq \mathrm{OPT}$, where OPT again denotes the expected cost of an optimum a priori tour, this yields

$$\mathbb{E}_{S\sim p}\left[\mathrm{MR}(S)\right] \;\leq\; 3 \cdot \mathrm{OPT}. \tag{1}$$

The work of Shmoys and Talwar [25] implies that (1) also holds when there is no depot and when we take the conditional expectation under the condition that $|S| \geq 2$ (see also [28]). The Shmoys–Talwar algorithm cannot find an optimum TSP tour for S but uses the double tree algorithm with approximation guarantee 2. As noted by [6], one can as well use the Christofides–Serdyukov algorithm with approximation guarantee $\frac{3}{2}$, or in fact any α-approximation algorithm for TSP. Then the expected cost of the resulting master route solution is at most $(\alpha + 2) \cdot \mathrm{OPT}$.

This motivates the following questions:

(i) Is it optimal to sample S with exactly the activation probabilities (which is crucially used in the above analysis), or can we improve on the factor $\alpha + 2$ by sampling fewer or more?
(ii) How bad can the best master route solution be? We will call this the *master route ratio*: by the Shmoys–Talwar analysis, it is at most 3.
(iii) Can we obtain an approximation guarantee equal to the master route ratio by a master route solution based on random sampling, assuming that we can find optimum TSP tours? What is the best we can achieve with a $\frac{3}{2}$-approximation algorithm for TSP?
(iv) Can we obtain a better deterministic algorithm without a better TSP algorithm?

We give almost complete answers to all these questions.

1.2 Our results

The possibility that we sample the empty set or that no customer is active causes significant complications. The previous works [25] and [28] gave ad hoc proofs that *their* algorithms (which are also formulated with a depot) generalize to the non-depot case. We aim for a

general reduction, losing only an arbitrarily small constant: Fortunately, instances in which the expected number of active customers is small can be solved easily with an approximation factor $3+\varepsilon$ (for any $\varepsilon > 0$; similar to [7]; cf. Lemma 29), and hence much better than the known guarantees. For instances with a large expected number of active customers, one can assume without loss of generality (with an arbitrarily small loss) that there is a customer d that is always active, i.e., $p(d) = 1$ (Lemma 30). So we assume this henceforth and call d the depot. We summarize (and refer to Section 6 for the proof):

▶ **Theorem 1.** *Let $\varepsilon > 0$ and $\rho \geq 3$ be constants. If there exists a (randomized) polynomial-time ρ-approximation algorithm for instances (V, c, p) of the* A PRIORI TSP *that have a depot (i.e., a customer d with $p(d) = 1$), then there is a (randomized) polynomial-time $(\rho+\varepsilon)$-approximation algorithm for general instances of the* A PRIORI TSP*.*

The Shmoys-Talwar algorithm [25] includes a customer v into S with probability $p(v)$: the sampling probability is exactly the activation probability. Although this is natural and allows for the simple analysis in Section 1.1 (assuming a depot), we show that this is not optimal. Decreasing the probability of including a customer into the master tour improves the approximation guarantee. To be more precise, in Section 2 and Section 3, we analyze the following *sampling algorithm* for A PRIORI TSP instances with depot. Let $f\colon (0,1] \to [0,1]$ with $f(1) = 1$.

(i) Sample a subset $S \subseteq V$ by including every customer v independently with probability $f(p(v))$.
(ii) Call an α-approximation algorithm for (metric) TSP in order to compute a TSP tour for S, which serves as master tour.
(iii) Connect every customer outside S to the nearest customer in S by a pair of parallel edges.

For a given instance this algorithm has expected approximation ratio at most

$$\frac{1}{\mathrm{OPT}} \cdot \mathbb{E}_{S\sim f\circ p}\left[\alpha \cdot \mathrm{OPT}_{\mathrm{TSP}}(S,c) + 2 \cdot \sum_{v\in V} p(v)\cdot c(v,S)\right] \tag{2}$$

(where $\frac{0}{0} \coloneqq 1$). Shmoys and Talwar [25] used the identity function $f(p) = p$. It is easy to construct examples where sampling less or more is better. For example, if $c(v,w) = 1$ for all $v, w \in V$ with $v \neq w$ (and all activation probabilities except for the depot are tiny), it is best to include only the depot in the master tour: this yields an approximation ratio of 2 instead of 3. On the other hand, if $V = \{v_0, \dots, v_{n-1}\}$ and $c(v_i, v_j) = \min\{j-i, n+i-j\}$ for $i < j$ (i.e., (V,c) is the metric closure of a cycle), the more we sample, the better. However, even if we choose f depending on the instance, there is a limit on what we can achieve:

▶ **Theorem 2.** *No matter how f is chosen, even depending on the instance in an arbitrary way, the sampling algorithm has no better approximation ratio than*

- 2.655 *even if it computes an optimum TSP tour on the sampled customers;*
- 3.049 *assuming that we never compute a TSP tour on the sampled customers of cost less than* 1.4999 *times the cost of an optimum tour.*

See Section 3 for the proof. We do not have a matching upper bound, but we come close. For $\alpha = 1.5$ we prove (in Section 2):

▶ **Theorem 3.** *For $\alpha = 1.5$ and $f(p) = 1-(1-p)^\sigma$ with $\sigma = 0.663$, the sampling algorithm for* A PRIORI TSP *instances with depot has approximation guarantee less than 3.1.*

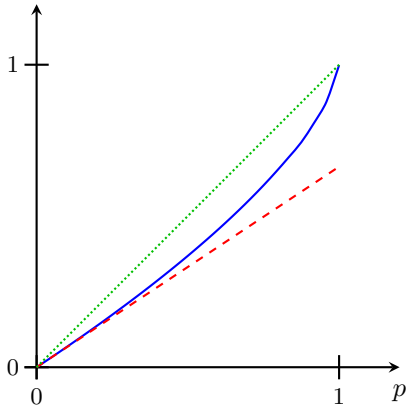


Figure 2 The function $p \mapsto 1 - (1-p)^\sigma$ with $\sigma = 0.663$ (blue, solid) defines the sampling probability in Theorem 3, which is always at most the identity function (green, dotted), and for small p approximately equal to $p \mapsto \sigma \cdot p$ (red, dashed).

Figure 2 shows this function f. Together with Theorem 1 this immediately implies one of our main results:

▶ **Corollary 4.** *There is a randomized* 3.1*-approximation algorithm for* A PRIORI TSP. ◀

We conjecture that the bounds in Theorem 2 are actually attained by the sampling algorithm with $f(p) = 1 - (1-p)^\sigma$, independent of the instance, where σ is a positive constant that depends on α only. See Comment 17 for details.

Having explored the limits of the random sampling approach, one might ask what is the limit of choosing an *optimal* master route solution. By van Zuylen's work [28], the answer to this question is the key to obtain a better deterministic approximation algorithm (see Section 8). Let us define:

▶ **Definition 5** (master route ratio). *The* master route ratio *is defined to be the supremum of*

$$\frac{\min\{\mathrm{MR}(S) : \emptyset \neq S \subseteq V\}}{\mathrm{OPT}}$$

taken over all A PRIORI TSP *instances (where* $\frac{0}{0} := 1$*).*

It is very easy to see that the master route ratio is at least 2 (for example, if $c(v,w) = 1$ for all $v, w \in V$ with $v \neq w$). By the Shmoys–Talwar analysis, it is at most 3. We will show in Section 4 (upper bound) and Section 5 (lower bound):

▶ **Theorem 6.** *The master route ratio for* A PRIORI TSP *instances with depot is at least* $\frac{1}{1-e^{-1/2}} > 2.541$ *and less than* 2.6*.*

We conjecture that the master route ratio is exactly $\frac{1}{1-e^{-1/2}}$.

As van Zuylen's [28] analysis reveals (cf. Theorem 35 in Section 8), her algorithm is a $(2+\alpha\rho)$-approximation algorithm if the master route ratio is ρ and we have an algorithm for TSP that guarantees to produce a tour of cost at most α times the value of the subtour relaxation. So our new upper bound on the master route ratio immediately implies a better guarantee (combining Theorems 1, 6, and 35 with $\alpha = \frac{3}{2}$ [27]):

▶ **Corollary 7.** *There is a deterministic 5.9-approximation algorithm for* A PRIORI TSP*.* ◀

1.3 Our techniques

The lower bounds (Theorem 2 and the lower bound in Theorem 6) are obtained by analyzing simple examples. The main technical difficulty is in proving the upper bounds.

To prove Theorem 3 and the upper bound in Theorem 6, we will show that it suffices to consider instances in which all customers (except the depot) have the same tiny activation probability. We call these instances *normalized*.

▶ **Definition 8.** *Let $\varepsilon > 0$. An instance (V, c, p) of* A PRIORI TSP *is called* ε-normalized *if the instance contains a depot $d \in V$ (with $p(d) = 1$), and $p(v) = \varepsilon$ for all $v \in V \setminus \{d\}$.*

Given an instance of A PRIORI TSP with a depot d, one can transform it to a normalized instance by replacing each customer $v \in V \setminus \{d\}$ by many copies, each with the same tiny activation probability, such that the probability that at least one of these copies is active is roughly $p(v)$. This way, the master route ratio and the approximation guarantee of the sampling algorithm can only get worse. More precisely, we show (in Section 7):

▶ **Lemma 9.** *Let $(\varepsilon_i)_{i\in\mathbb{N}} \in (0,1]^{\mathbb{N}}$ with $\lim_{i\to\infty} \varepsilon_i = 0$. Let $\mathcal{I}$ be the class of all ε-normalized instances with $\varepsilon = \varepsilon_i$ for some $i \in \mathbb{N}$. Then*

(i) The master route ratio is the same when restricting it to instances in $\mathcal{I}$ and when restricting it to all instances with depot.

(ii) Let $\sigma \in (0,1)$. Every upper bound on (2) for $f(p) = \sigma p \ \forall p \in (0,1)$ for all instances in $\mathcal{I}$ implies the same upper bound on (2) for $f(p) = 1 - (1-p)^\sigma$ for arbitrary instances with depot.

On a high level, our proofs of Theorem 3 and Theorem 6 are similar. In both cases we will design a linear program that encodes the metric c by variables and minimizes the expected cost of an optimum a priori tour subject to (a relaxation of) the constraint that the expected cost of the output of the sampling algorithm is at least 1 (for Theorem 3) or the expected cost of any master route solution is at least 1 (for Theorem 6), respectively. Then the reciprocals of the LP values yield the desired upper bounds.

However, this approach has to overcome several obstacles. First, it is not obvious how to encode the metric c by finitely many variables, given that we need to consider arbitrary instance sizes. We do this by fixing an optimum a priori tour T^* (a cyclic order of the customers) and carefully aggregating distances of customer pairs with the same number of hops in between on T^*. Of course we exploit the structure of normalized instances.

In the end, we will (almost) ignore variables that correspond to a very large number of hops (where it is very unlikely that none of the customers "in between" is active). These variables have negligible impact because the probability that these edges occur decreases exponentially with increasing number of hops on T^*, whereas the average length of these edges can only grow linearly due to the triangle inequality.

The next idea is to consider certain structured solutions only. Rather than connecting a customer v that is not in the master tour to the *nearest* customer $\mu(v)$ in the master tour, we consider only two possible members of the master tour: we traverse T^* from v in each of the two possible directions, and consider the first customer that we meet and that is contained in our master tour. None of these two may be a nearest one in the master tour, but we still obtain an upper bound. For bounding the master route ratio, we will in addition only consider master tours whose customers are equidistantly distributed on T^* (except for the depot).

In this way, we obtain an optimization problem for a fixed uniform activation probability p (i.e., for p-normalized instances). However, we must let $p \to 0$ according to Lemma 9

and hence need a description that is independent of p. This is another major obstacle. To overcome it, we use a second level of aggregation (buckets, rounding the number of hops to integer multiples of, say, $\frac{1}{100p}$). However, this causes several difficulties. In the case of the sampling algorithm, describing the expected cost of the output of the sampling algorithm in terms of the buckets is nontrivial. In case of the master route ratio, the same holds for master route solutions and actually requires a third level of aggregation (bucket intervals).

In the end, we obtain (in both cases) a single, relatively compact, linear program that yields an upper bound for all instance sizes and all activation probabilities from a sequence that converges to zero. We solve the dual LP numerically and just need to check feasibility to prove the desired upper bounds.

1.4 Further related work

The TSP has also been studied under the aspect of robust optimization, where the set of customers that need to be visited is known in advance, but the edge lengths are chosen probabilistically or even adversarially [9, 26]. The a priori optimization problem where the set of customers is chosen adversarially is known as universal TSP [10, 12, 24]. The probabilistic version that we consider was introduced by Jaillet [16] and Bertsimas [2]. Since then, various aspects of the problem have been investigated, including the asymptotic behavior of random instances [2, 3, 4, 16, 17], online variants [10], or exact algorithms [1]. Approximation algorithms have also been studied for general probability distributions [5, 12, 24].

Other problems that have been considered in an a priori setting include vehicle routing, traveling repairman, Steiner tree, and network design [6, 7, 8, 10, 13, 14, 15, 22]. However, none of these works managed to determine the approximation guarantee of their algorithms exactly.

Previous approaches to design a linear program that yields the approximation ratio of a certain algorithm for some optimization problem (e.g., [11, 18]) typically required an infinite family of linear programs and could not obtain a bound for general instances by just solving a single linear program.

2 Upper bound on the approximation ratio of random sampling

In this section we will prove Theorem 3. As mentioned earlier, we will design a single linear program such that the reciprocal of its optimum value is an upper bound on the approximation ratio of the sampling algorithm for a certain class of normalized instances. For this sake, let $\beta, b_0 > 0$ be constants that we will choose later. We will consider ε-normalized instances where ε is of the form $\varepsilon = \frac{\beta}{b}$ for some odd integer $b \geq b_0$. The meaning of these constants will become clear in Section 2.2. For such instances we will obtain an upper bound on the approximation ratio of the sampling algorithm, when sampling each customer with probability σp for $\sigma = 0.663$ (in addition to the depot). Combined with Lemma 9, this immediately yields the same upper bound on the approximation guarantee of the sampling algorithm that samples each customer v with probability $1-(1-p(v))^\sigma$ for arbitrary A PRIORI TSP instances with depot.

2.1 An optimization problem to bound the approximation ratio

In this section, we first describe an upper bound for all p-normalized instances (for a fixed uniform activation probability p) by a single optimization problem. We will consider the algorithm that samples each customer with probability σp. Let T^* be a fixed optimum a

priori tour, with customers appearing in the order $v_0, v_1, \ldots, v_{n-1}$; here v_0 denotes the depot. Let $v_i := v_0$ for $i < 0$ or $i > n-1$. For $k \in \mathbb{Z}_{\geq 1}$ we define

$$C_k \;:=\; p^2 \cdot \sum_{j \in \mathbb{Z}} c(v_j, v_{j+k}).$$

Observe that only finitely many summands are nonzero. See Figure 3 for an example. Since c is a metric, the numbers C_k are nonnegative and satisfy the triangle inequality, that is, for all $i, j \geq 1$

$$C_{i+j} \;\leq\; C_i + C_j. \tag{3}$$

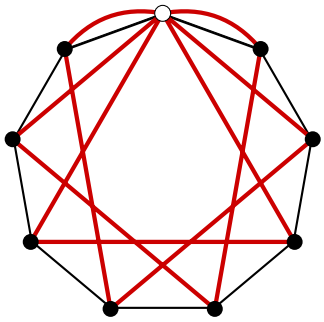


Figure 3 The depot v_0 is the white circle at the top; the tour T^* is drawn in black. Adding up the costs of the edges marked in red gives $\frac{C_3}{p^2}$.

Moreover, we can express the expected cost of T^* in terms of the C_i.

▶ **Proposition 10.** *The expected cost of T^* is exactly*

$$\sum_{i=1}^{\infty} (1-p)^{i-1} \cdot C_i. \tag{4}$$

Proof. Let $1 \leq i \leq n-2$ and $1 \leq j \leq n-i-1$. Then v_j and v_{j+i} are consecutive active customers with probability $p^2 \cdot (1-p)^{i-1}$; note that the cost of the edge $\{v_j, v_{j+i}\}$ is counted with exactly the same coefficient in (4). Moreover, for $1 \leq j \leq n-1$, v_j is the first active customer after the depot with probability $p \cdot (1-p)^{j-1}$, and the cost of the edge $\{v_0, v_j\}$ is counted $\sum_{i=j}^{\infty} p^2 \cdot (1-p)^{i-1} = p \cdot (1-p)^{j-1}$ times in (4). By symmetry, the terms also match for the last active customer before the depot. ◀

We now consider the master route solution resulting from sampling each customer with probability σp (in addition to the depot). Let α again denote the approximation guarantee of the TSP algorithm that we use. We will now show that the expected cost of this master route solution is at most

$$\sigma^2 \sum_{k=1}^{\infty} (1-\sigma p)^{k-1} \cdot \left(\alpha \cdot C_k + 2p \cdot \sum_{i=1}^{k-1} \min\left\{C_i,\, C_{k-i}\right\} \right). \tag{5}$$

By the same argumentation as in the proof of Proposition 10, the master tour has expected cost at most

$$\alpha \cdot \mathbb{E}_{S \sim q}[c(T^*[S])] \;=\; \alpha \cdot \sigma^2 \cdot \sum_{k=1}^{\infty} (1-\sigma p)^{k-1} \cdot C_k,$$

where $q(v) = \sigma p$ for all $v \in V \setminus \{d\}$ and $q(d) = 1$.

Next we bound the expected cost of connecting the active customers to the master tour. Instead of connecting v to the nearest customer in the master tour, we consider only two options: the first sampled customer that we meet when traversing T^* from v in either direction. Note that sampling v_0 with probability 1 is equivalent to sampling each v_j with $j \leq 0$ and $j \geq n$ with probability σp. Now, for $j \in \mathbb{Z}$ and $k \geq 2$, the probability that v_j and v_{j+k} are sampled, but none of the intermediate customers is, equals $(\sigma p)^2 \cdot (1-\sigma p)^{k-1}$. In this case, the total expected cost of connecting the intermediate active customers can be bounded by $2p \cdot \sum_{i=1}^{k-1} \min\{c(v_j, v_{j+i}), c(v_{j+i}, v_{j+k})\}$. Thus we can bound the expected cost of connecting all active customers to the master tour by

$$\begin{aligned}
&\sum_{k=2}^{\infty} \sigma^2 p^2 \cdot (1-\sigma p)^{k-1} \cdot \sum_{j \in \mathbb{Z}} 2p \cdot \sum_{i=1}^{k-1} \cdot \min\left\{c(v_j, v_{j+i}),\, c(v_{j+i}, v_{j+k})\right\} \\
&\leq\ 2\sigma^2 p^3 \cdot \sum_{k=1}^{\infty} (1-\sigma p)^{k-1} \cdot \sum_{i=1}^{k-1} \min\left\{\sum_{j \in \mathbb{Z}} c(v_j, v_{j+i}),\, \sum_{j \in \mathbb{Z}} c(v_{j+i}, v_{j+k})\right\} \\
&=\ 2\sigma^2 p^3 \cdot \sum_{k=1}^{\infty} (1-\sigma p)^{k-1} \cdot \sum_{i=1}^{k-1} \min\left\{\sum_{j \in \mathbb{Z}} c(v_j, v_{j+i}),\, \sum_{j \in \mathbb{Z}} c(v_j, v_{j+(k-i)})\right\} \\
&=\ 2\sigma^2 p \cdot \sum_{k=1}^{\infty} (1-\sigma p)^{k-1} \cdot \sum_{i=1}^{k-1} \min\left\{C_i,\, C_{k-i}\right\}.
\end{aligned}$$

We conclude that the ratio of (5) to (4) is an upper bound on the approximation guarantee of the sampling algorithm for that instance. Note that the number of customers appears neither in (4) nor in (5). In other words, minimizing (4) subject to the constraints that (5) is equal to 1 and the C_i are nonnegative and satisfy the triangle inequality (3) yields the reciprocal of an upper bound on the approximation guarantee of the sampling algorithm on all p-normalized instances. We arrive at the following optimization problem:

$$\min \sum_{i=1}^{\infty} (1-p)^{i-1} \cdot C_i \qquad \text{(Sampling-OP)}$$

$$\text{subject to} \qquad C_i \geq 0 \qquad \text{for } i \in \mathbb{N} \quad (6)$$

$$C_i + C_j \geq C_{i+j} \qquad \text{for } i, j \in \mathbb{N} \quad (7)$$

$$\sum_{k=1}^{\infty} (1-\sigma p)^{k-1} \cdot \left(\alpha \cdot C_k + 2p \cdot \sum_{i=1}^{k-1} \min\left\{C_i,\, C_{k-i}\right\}\right) \geq \sigma^{-2}. \quad (8)$$

Note that in (8) we only require that (5) is at least 1 instead of exactly 1. This does not change the infimum because we can always scale all the C_i's. We have proved:

▶ **Lemma 11.** *Let $0 < p < 1$. The reciprocal of the value of* (Sampling-OP) *is an upper bound on the approximation guarantee for the sampling algorithm with $f(p) = \sigma p$ for all p-normalized instances.* ◀

2.2 Obtaining a single linear program

Note that we have an infinite set of optimization problems (one for each choice of p), and, in view of Lemma 9, we have to consider the limit for $p \to 0$.

In the following, we require that p is of the form $p = \frac{\beta}{b}$ for some odd integer $b \geq b_0$. Note that $p \to 0$ as $b \to \infty$. In order to obtain a single optimization problem for all such values of p, we put subsequent C_i's into buckets of size b. More precisely, we define buckets

$$B_i \;:=\; \sum_{j=\max\{1,ib-\frac{b-1}{2}\}}^{ib+\frac{b-1}{2}} C_j \tag{9}$$

for $i \geq 0$. In the following, we show that we can use the constraints in (Sampling-OP) to generate (slightly relaxed) constraints that only depend on these buckets. First, we note that the buckets are chosen such that they still satisfy the triangle inequality.

▶ **Proposition 12.** *For all $i, j \geq 1$,*

$$B_{i+j} \;\leq\; B_i + B_j.$$

Proof. Indeed, using (7), as illustrated in Figure 4,

$$\begin{aligned} B_{i+j} \;&=\; \sum_{k=-\frac{b-1}{2}}^{\frac{b-1}{2}} C_{(i+j)b+k} \;=\; \sum_{k=0}^{\frac{b-1}{2}} C_{(i+j)b-\frac{b-1}{2}+2k} + \sum_{k=1}^{\frac{b-1}{2}} C_{(i+j)b-\frac{b-1}{2}+2k-1} \\ &\leq\; \sum_{k=0}^{\frac{b-1}{2}} \left(C_{ib-\frac{b-1}{2}+k} + C_{jb+k}\right) + \sum_{k=1}^{\frac{b-1}{2}} \left(C_{ib+k} + C_{jb-\frac{b-1}{2}+k-1}\right) \\ &=\; \sum_{k=-\frac{b-1}{2}}^{\frac{b-1}{2}} C_{ib+k} + \sum_{k=-\frac{b-1}{2}}^{\frac{b-1}{2}} C_{jb+k} \;=\; B_i + B_j. \end{aligned}$$

◀

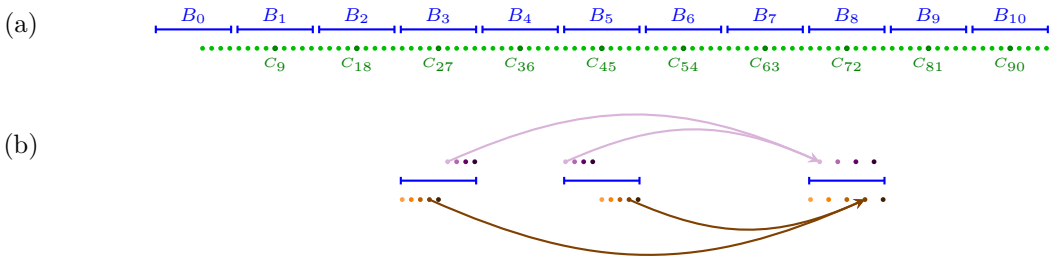


Figure 4 (a): The green dots stand for $C_1, C_2, \ldots$, and the centers of the buckets (C_{ib} for $i \geq 1$) are highlighted. Here the bucket size is $b = 9$, and the blue intervals show the buckets $B_0, B_1, B_2, \ldots$ (b): Combining the triangle inequalities for the C_i's leads to triangle inequalities for the B_i's; here shown for $B_i = B_3$ and $B_j = B_5$: We add up all triangle inequalities for C_k from B_3 and C_ℓ from B_5 where C_k and C_ℓ have the same color; illustrated with $C_{26} + C_{48} \leq C_{74}$ and $C_{28} + C_{41} \leq C_{69}$.

Next we aim for an upper bound on the left-hand side of (8) that only depends on the buckets. First we show:

▶ **Lemma 13.**

$$\sum_{k=1}^{\infty} (1-\sigma p)^{k-1} \cdot \sum_{i=1}^{k-1} \min\left\{C_i,\, C_{k-i}\right\} \;\leq\; b \cdot \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} e^{-(i+j-1)\cdot\sigma b p} \cdot \min\{B_i, B_j\}.$$

Proof. For $i \in \mathbb{Z}_{\geq 0}$, let $I_i = \{\max\{1, ib - \frac{b-1}{2}\}, \dots, ib + \frac{b-1}{2}\}$ be the set of indices in the i-th bucket. Then

$$\begin{aligned}
&\sum_{k=1}^{\infty}(1-\sigma p)^{k-1} \cdot \sum_{i=1}^{k-1} \min\left\{C_i,\, C_{k-i}\right\} \\
&= \sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty}(1-\sigma p)^{k+\ell-1} \cdot \min\left\{C_k,\, C_\ell\right\} \\
&\leq \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(1-\sigma p)^{\max\{0,i+j-1\}\cdot b} \cdot \sum_{k\in I_i}\sum_{\ell\in I_j} \min\{C_k, C_\ell\} \\
&\leq \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} e^{-(i+j-1)\cdot\sigma b p} \cdot \sum_{k\in I_i}\sum_{\ell\in I_j} \min\{C_k, C_\ell\}.
\end{aligned}$$

In the last inequality we used $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$. Now, for $i, j \in \mathbb{Z}_{\geq 0}$, consider the complete bipartite graph H where one bipartition consists of $|I_j|$ copies of every element of I_i, and the other bipartition consists of $|I_i|$ copies of every element of I_j. Then

$$\sum_{(k,\ell)\in E(H)} \min\{C_k, C_\ell\} \;=\; |I_i| \cdot |I_j| \cdot \sum_{k\in I_i}\sum_{\ell\in I_j} \min\{C_k, C_\ell\}.$$

We can partition $E(H)$ into $t := |I_i| \cdot |I_j|$ perfect matchings $M_1, \dots, M_t$. Then

$$\begin{aligned}
\sum_{(k,\ell)\in E(H)} \min\{C_k, C_\ell\} &= \sum_{s=1}^{t} \sum_{(k,\ell)\in M_s} \min\{C_k, C_\ell\} \\
&\leq \sum_{s=1}^{t} \min\left\{ \sum_{(k,\ell)\in M_s} C_k,\; \sum_{(k,\ell)\in M_s} C_\ell \right\} \\
&= \sum_{s=1}^{t} \min\left\{ |I_j| \cdot \sum_{k\in I_i} C_k,\; |I_i| \cdot \sum_{\ell\in I_j} C_\ell \right\} \\
&= |I_i| \cdot |I_j| \cdot \min\left\{|I_j| \cdot B_i,\, |I_i| \cdot B_j\right\}.
\end{aligned}$$

Note that the second equality follows from the fact that $V(H)$ contains $|I_j|$ copies of each element in I_i and vice versa. Moreover, summing over the endpoints of the edges in a perfect matching in M is the same as summing over $V(H)$. Division by $|I_i| \cdot |I_j|$ yields

$$\sum_{k\in I_i}\sum_{\ell\in I_j} \min\{C_k, C_\ell\} \;\leq\; \min\left\{|I_j| \cdot B_i, |I_i| \cdot B_j\right\} \;\leq\; b \cdot \min\{B_i, B_j\}. \qquad \blacktriangleleft$$

Using Lemma 13 and $\beta = bp$, the left-hand side of (8) can be upper bounded by

$$\begin{aligned}
&\alpha \cdot \sum_{k=1}^{\infty}(1-\sigma p)^{k-1} \cdot C_k \;+\; 2bp \cdot \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} e^{-(i+j-1)\cdot\sigma b p} \cdot \min\{B_i, B_j\} \\
&\leq\; \alpha \cdot \sum_{k=0}^{\infty}(1-\sigma p)^{\max\{0, kb - \frac{b-1}{2} - 1\}} \cdot B_k \;+\; 2bp \cdot \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} e^{-(i+j-1)\cdot\sigma b p} \cdot \min\{B_i, B_j\} \\
&\leq\; \alpha \cdot \sum_{k=0}^{\infty} e^{-(k-1)\cdot\sigma\beta} \cdot B_k \;+\; 2\beta \cdot \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} e^{-(i+j-1)\cdot\sigma\beta} \cdot \min\{B_i, B_j\}. \qquad (10)
\end{aligned}$$

The last inequality follows from $(1-\sigma p)^{kb-\frac{b-1}{2}-1} \leq (1-\sigma p)^{kb-b}$ and $1+x \leq e^x$ for all $x \in \mathbb{R}$. Note that we still sum over infinitely many variables. Hence, in order to get a finite linear program, we aim for an upper bound on (10) that only depends on the buckets B_i with $i \leq N$ for some integer N that we will choose later. For this, we use the triangle inequality (Proposition 12) to bound the terms depending on buckets B_i with $i > N$ by some term depending on $B_1, \dots, B_N$ only. For large N this will result in a negligible error as the coefficients in (10) decrease exponentially. In Section 2.4 we prove the following bound on the error term:

▶ **Lemma 14.** *Let* $\delta_1 \coloneqq \frac{4\beta}{e^{N\sigma\beta}(e^{\sigma\beta}-1)}$ *and* $\delta_2 \coloneqq \left(\alpha + \frac{2\beta}{e^{N\sigma\beta}(e^{\sigma\beta}-1)}\right) \cdot \frac{e^{-N\sigma\beta}}{(1-e^{-\sigma\beta})^2} \cdot (1+N-e^{-\sigma\beta}N)$. *Then*

$$\alpha \cdot \sum_{k=N+1}^{\infty} e^{-(k-1)\cdot\sigma\beta} \cdot B_k \;+\; 2\beta \cdot \sum_{i,j\in\mathbb{Z}_{\geq 0}:\max\{i,j\}>N} e^{-(i+j-1)\cdot\sigma\beta} \cdot \min\{B_i, B_j\}$$
$$\leq\; \delta_1 \cdot \sum_{k=0}^{N} e^{-(k-1)\cdot\sigma\beta} \cdot B_i + \delta_2 \cdot B_1.$$

Therefore, we get a lower bound on (Sampling-OP) by minimizing $\sum_{i=1}^{\infty}(1-p)^{i-1} \cdot C_i$ subject to (9) and $B_i \geq 0$ for $i \geq 0$, $B_{i+j} \leq B_i + B_j$ for $i, j \geq 1$ with $i + j \leq N$, and

$$(\alpha+\delta_1) \cdot \sum_{k=0}^{N} e^{-(k-1)\cdot\sigma\beta} \cdot B_k + 4\beta \cdot \sum_{j=0}^{N}\sum_{i=0}^{j} e^{-(i+j-1)\cdot\sigma\beta} \cdot \min\{B_i, B_j\} + \delta_2 \cdot B_1 \geq \sigma^{-2}. \quad (11)$$

Note that the objective still contains infinitely many variables and depends on p. The first problem can easily be resolved by bounding

$$\sum_{i=1}^{\infty}(1-p)^{i-1} \cdot C_i \;\geq\; \sum_{i=0}^{\infty}(1-p)^{bi+\frac{b-1}{2}-1} \cdot B_i \;\geq\; \sum_{i=0}^{N}(1-p)^{(i+\frac{1}{2})b} \cdot B_i. \quad (12)$$

It remains to get rid of the dependence on b and p (recall that $p = \frac{\beta}{b}$). To this end, we exploit that $\lim_{b\to\infty}(1-\frac{\beta}{b})^{(i+\frac{1}{2})b} = e^{-(i+\frac{1}{2})\beta}$ for all $i = 0, \dots, N$, and that by Lemma 9, we can choose b_0 arbitrarily large. This will allow us to conclude that we can replace the objective by $\sum_{i=0}^{N} e^{-(i+\frac{1}{2})\beta} \cdot B_i$ and still obtain an upper bound (see the proof of Lemma 15 for the technical details). Putting everything together, we arrive at the following LP.

$$\min \sum_{i=0}^{N} e^{-(i+\frac{1}{2})\beta} \cdot B_i \quad \text{(Sampling-LP)}$$

$$\text{subject to} \quad (\alpha+\delta_1) \cdot \sum_{k=0}^{N} e^{-(k-1)\cdot\sigma\beta} \cdot B_k + \delta_2 \cdot B_1 + 4\beta \cdot \sum_{j=0}^{N}\sum_{i=0}^{j} e^{-(i+j-1)\cdot\sigma\beta} \cdot M_{i,j} \;\geq\; \sigma^{-2} \quad (13)$$

$$B_i + B_j \;\geq\; B_{i+j} \qquad \text{for } 1 \leq i \leq j \leq N \text{ with } i+j \leq N \quad (14)$$

$$B_i \;\geq\; M_{i,j} \qquad \text{for } 0 \leq i \leq j \leq N \quad (15)$$

$$B_j \;\geq\; M_{i,j} \qquad \text{for } 0 \leq i \leq j \leq N \quad (16)$$

$$B, M \;\geq\; 0. \quad (17)$$

Recall that δ_1 and δ_2 were defined in Lemma 14. We conclude:

▶ **Lemma 15.** *Let* N *be an integer and* $\beta > 0$. *The reciprocal of the optimum value of* (Sampling-LP) *is an upper bound on the approximation guarantee of the sampling algorithm for* $f(p) = 1-(1-p)^\sigma$ *(using an* α*-approximation algorithm for TSP), for all* A PRIORI *TSP instances with depot.*

Proof. We compare the value of (Sampling-LP) to the value of (Sampling-OP). We showed above that for any feasible solution C to (Sampling-OP) we obtain a feasible solution (B, M) to (Sampling-LP) via (9) and $M_{i,j} = \min\{B_i, B_j\}$.

Fix $\delta > 0$. Then there exists $b_0 \in \mathbb{N}$ such that for all odd integers $b \geq b_0$

$$\sum_{i=0}^{N} e^{-(i+\frac{1}{2})\beta} \cdot B_i \;\leq\; (1+\delta) \cdot \sum_{i=0}^{N} \left(1 - \tfrac{\beta}{b}\right)^{(i+\frac{1}{2})b} \cdot B_i \;\overset{(12)}{\leq}\; (1+\delta) \cdot \sum_{i=0}^{\infty} \left(1 - \tfrac{\beta}{b}\right)^{i-1} \cdot C_i.$$

Thus the value of (Sampling-LP) is at most $(1+\delta)$ times the value of (Sampling-OP) with $p = \frac{\beta}{b}$ for all odd integers $b \geq b_0$. Hence, by Lemma 11, $(1+\delta)$ times the reciprocal of the optimum value of (Sampling-LP) is an upper bound on (2) for all $\frac{\beta}{b}$-normalized instances for all odd integers $b \geq b_0$. By Lemma 9, the same bound then holds for all instances with depot. Since this bound holds for all $\delta > 0$, it also holds for $\delta = 0$. ◀

2.3 The dual LP

In order to obtain a lower bound on the optimum value of (Sampling-LP), we provide a *feasible solution* to the *dual* linear program. For the dual LP, we introduce variables $x_{i,j}$ for the inequalities of type (14), variables $v_{i,j}$ and $w_{i,j}$ for the inequalities of type (15) and (16), respectively, and a variable y for inequality (13). Using these variables, the dual LP looks as follows:

$$\max\ \sigma^{-2} \cdot y \qquad \text{(Dual-Sampling-LP)}$$

$$\text{subject to} \qquad 4\beta \cdot e^{-(i+j-1)\cdot\sigma\beta} \cdot y \;\leq\; v_{i,j} + w_{i,j} \qquad \text{for } 0 \leq i \leq j \leq N \quad (18)$$

$$(\alpha+\delta_1) \cdot e^{-(k-1)\cdot\sigma\beta} \cdot y + \sum_{j=k}^{N} v_{k,j} + \sum_{j=0}^{k} w_{j,k} + \mathbb{1}_{k=1} \cdot \delta_2 \cdot y$$

$$+\mathbb{1}_{k>0} \cdot \left(\sum_{i=1}^{\min\{k,N-k\}} x_{i,k} + \sum_{i=k}^{N-k} x_{k,i} - \sum_{\substack{1 \leq i \leq j \leq N,\\ i+j=k}} x_{i,j} \right) \;\leq\; e^{-(k+\frac{1}{2})\beta} \qquad \text{for } 0 \leq k \leq N \quad (19)$$

$$x, y, v, w \;\geq\; 0. \qquad (20)$$

▶ **Corollary 16.** *Let N be an integer and $\beta > 0$. For any feasible solution (x, y, v, w) to* (Dual-Sampling-LP)*, σ^2/y is an upper bound on the approximation ratio of the sampling algorithm with $f(p) = 1 - (1-p)^\sigma$ restricted to* A PRIORI *TSP instances with depot.* ◀

We have computed a dual solution using Gurobi 10.0.1 with $\alpha = 1.5$, $\beta = \frac{1}{100}$, $N = 2500$, and $\sigma = 0.663$, yielding an upper bound of 3.094 and thus proving Theorem 3. The dual solution and a Python script that verifies that this is a feasible solution to (Dual-Sampling-LP) can be found at `https://doi.org/10.60507/FK2/JCUIRI`. For $\alpha = 1$, we get an upper bound of 2.694.

▶ **Comment 17.** Solving (Sampling-LP) with the same values for α, β, N, and σ yields an A PRIORI TSP instance of the same shape as the example provided in Section 3. Hence we conjecture that the upper bound given by (Dual-Sampling-LP) converges to the lower bound given in Theorem 2 for $\beta \to 0$ and $N\beta \to \infty$.

2.4 Bounding the error term (Proof of Lemma 14)

We first prove the following auxiliary lemma:

▶ **Lemma 18.** *Let $n \in \mathbb{N}$ and $q \in (0,1)$. Then*

$$\sum_{k=n+1}^{\infty} k \cdot q^{k-1} \;=\; \frac{q^n}{(1-q)^2} \cdot (1+n-qn). \tag{21}$$

Proof. By induction on n. For $n = 0$, the statement is equivalent to the well-known formula

$$\sum_{k=1}^{\infty} k \cdot (1-q) \cdot q^{k-1} \;=\; \frac{1}{1-q}$$

for the expected value of a geometrically distributed random variable. Next, assume that (21) holds for some $n \in \mathbb{N}$. Then

$$\begin{aligned}\sum_{k=n+2}^{\infty} k \cdot q^{k-1} \;&=\; \sum_{k=n+1}^{\infty} k \cdot q^{k-1} - (n+1) \cdot q^n \overset{(21)}{=} \frac{q^n}{(1-q)^2} \cdot (1+n-qn) - (n+1) \cdot q^n \\ &=\; \frac{q^n}{(1-q)^2} \cdot (1+n-qn-(n+1)\cdot(1-q)^2) \;=\; \frac{q^{n+1}}{(1-q)^2} \cdot (n+2-q(n+1)),\end{aligned}$$

which is (21) for $n+1$. ◀

Now we are ready to prove Lemma 14:

Proof of Lemma 14. We compute

$$\begin{aligned}&2\beta \cdot \sum_{i,j \in \mathbb{Z}_{\geq 0} : \max\{i,j\} > N} e^{-(i+j-1)\cdot\sigma\beta} \cdot \min\{B_i, B_j\} \\ &\leq\; 4\beta \cdot \sum_{i=0}^{N} \sum_{j=N+1}^{\infty} e^{-(i+j-1)\cdot\sigma\beta} \cdot B_i \;+\; 2\beta \cdot \sum_{i=N+1}^{\infty} \sum_{j=N+1}^{\infty} e^{-(i+j-1)\cdot\sigma\beta} \cdot B_i \\ &=\; 4\beta \cdot \sum_{i=0}^{N} e^{-(i-1)\cdot\sigma\beta} \cdot B_i \cdot \sum_{j=N+1}^{\infty} e^{-j\sigma\beta} \;+\; 2\beta \cdot \sum_{i=N+1}^{\infty} e^{-(i-1)\cdot\sigma\beta} \cdot B_i \cdot \sum_{j=N+1}^{\infty} e^{-j\sigma\beta} \\ &=\; \delta_1 \cdot \sum_{i=0}^{N} e^{-(i-1)\cdot\sigma\beta} \cdot B_i \;+\; \frac{2\beta}{e^{N\sigma\beta}(e^{\sigma\beta}-1)} \cdot \sum_{i=N+1}^{\infty} e^{-(i-1)\cdot\sigma\beta} \cdot B_i\end{aligned}$$

Bounding $B_i \leq i \cdot B_1$ for $i > N$ by using to the triangle inequality (Proposition 12), we obtain

$$\begin{aligned}&\alpha \cdot \sum_{k=N+1}^{\infty} e^{-(k-1)\cdot\sigma\beta} \cdot B_k \;+\; 2\beta \cdot \sum_{i,j \in \mathbb{Z}_{\geq 0} : \max\{i,j\} > N} e^{-(i+j-1)\cdot\sigma\beta} \cdot \min\{B_i, B_j\} \\ &\leq\; \delta_1 \cdot \sum_{i=0}^{N} e^{-(i-1)\cdot\sigma\beta} \cdot B_i + \left(\alpha + \frac{2\beta}{e^{N\sigma\beta}(e^{\sigma\beta}-1)}\right) \cdot \sum_{k=N+1}^{\infty} e^{-(k-1)\cdot\sigma\beta} \cdot B_k \\ &\leq\; \delta_1 \cdot \sum_{i=0}^{N} e^{-(i-1)\cdot\sigma\beta} \cdot B_i + \left(\alpha + \frac{2\beta}{e^{N\sigma\beta}(e^{\sigma\beta}-1)}\right) \cdot \sum_{k=N+1}^{\infty} e^{-(k-1)\cdot\sigma\beta} \cdot k \cdot B_1 \\ &=\; \delta_1 \cdot \sum_{i=0}^{N} e^{-(i-1)\cdot\sigma\beta} \cdot B_i + \delta_2 \cdot B_1,\end{aligned}$$

where we used Lemma 18 in the final equality with $n = N$ and $q = e^{-\sigma\beta}$. ◀

3 Lower bound on the approximation ratio of the sampling algorithm

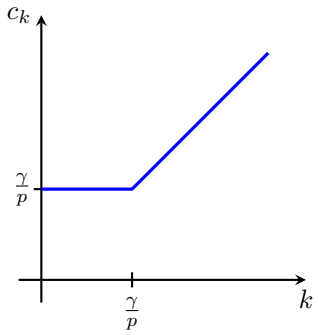
In this section we prove Theorem 2. We will provide a family of instances for which the sampling algorithm that we described in Section 1.2 has no better approximation ratio than 2.655, no matter how we choose f, and even when assuming that we can compute optimal TSP tours on the sampled customers. Using the currently best approximation guarantee for metric TSP leads to a ratio of more than 3.049 (again, even if we try all f).

In the proof we exploit that each instance in the family that we describe has uniform activation probability, i.e., $p(v) = p$ for each $v \in V$, where p is a small positive number. Hence, it suffices to consider functions f with $f(1) = 1$ and $f(p) = \sigma p$ for some $\sigma \in \left[0, \frac{1}{p}\right]$.

Let $\gamma \in [1, 2]$ be a parameter chosen later, depending only on the approximation ratio α for TSP. Let $0 < \varepsilon \ll 1$. We will choose $p > 0$ (very small) and $n \in \mathbb{N}$ (very large), depending on γ and ε only (cf. Lemma 19). We then consider a p-normalized instance of the A PRIORI TSP with $V = \{v_0, \cdots, v_{n-1}\}$ with v_0 being the depot. Then for $0 \leq i < j \leq n-1$ with $k = \min\{j - i, i + n - j\}$ define distances $c(v_i, v_j) = \frac{c_k}{n}$, where (cf. Figure 5)

$$c_k := \begin{cases} \frac{\gamma}{p} & \text{if } k \leq \frac{\gamma}{p} \\ k & \text{otherwise.} \end{cases}$$

Note that they form a metric. We will show the claimed lower bounds for this set of instances.



α	γ	best choice for σ	lower bound on approx. ratio of the sampling algorithm
1	1.623	≈ 0.623	> 2.655
$\frac{4}{3}$	1.383	≈ 0.651	> 2.914
1.4999	1.291	≈ 0.663	> 3.049

Figure 5 Left: The distance of v_i and v_j for $0 \leq i < j \leq n-1$ is $\frac{c_k}{n}$, depending on $k = \min\{j - i, i + n - j\}$. The figure shows the dependence of c_k on k. Right: The table shows, for a given approximation ratio α that our black box TSP approximation algorithm achieves, how we should choose γ such that the sampling algorithm performs worst possible even if we choose σ best possible.

We choose p and n such that the following properties hold that we will use later:

▶ **Lemma 19.** *For every $\gamma \in [1, 2]$ and every $\varepsilon \in (0, \frac{1}{4})$, there are $p > 0$ and $n \in \mathbb{N}$ such that $\frac{\gamma}{p}$ is integral and*

(i) $p \leq (1 - 4\varepsilon) \cdot \varepsilon^2$,
(ii) $(1 - \varepsilon p) \cdot (1 - xp)^{\frac{y}{p}} \geq e^{-xy} - \varepsilon$ *for all* $x \in \left[0, \frac{1}{p}\right]$ *and* $y \in \left[1, \frac{2}{\varepsilon}\right]$,
(iii) $n \geq \frac{4}{\varepsilon p} + \frac{3}{\varepsilon}$,
(iv) $(1 + x) \cdot e^{-p\varepsilon^2 x} \leq \varepsilon - \frac{1}{n}$ *for all* $x \geq \varepsilon^2 pn - 3$.

Proof. Fix $\gamma \in [1, 2]$ and $\varepsilon \in (0, \frac{1}{4})$. As $\lim_{t \to 0}(1 - t)^{-\frac{2\ln(\varepsilon)}{\varepsilon}} = 1$, pick $\delta \in (0, 1)$ with

$$1 - \frac{\varepsilon}{2} \leq (1 - \delta)^{-\frac{2\ln(\varepsilon)}{\varepsilon}}. \tag{22}$$

Moreover, as $\lim_{t\to 0}(1-t)^{\frac{1}{t}} = e^{-1}$, we can pick $p \in (0, \frac{1}{2})$ such that (i) holds, $\frac{\gamma}{p}$ is integral and such that

$$\forall t \in (0, -\ln(\varepsilon)\cdot p]: \ (1-t)^{\frac{1}{t}} \ \geq \ (1-\delta)\cdot e^{-1}.$$

In particular,

$$\forall x \in (0, -\ln(\varepsilon)]: \ (1-xp)^{\frac{1}{xp}} \ \geq \ (1-\delta)\cdot e^{-1}. \tag{23}$$

We show that (ii) holds. First of all, if $x > -\ln(\varepsilon)$, then the right-hand-side of (ii) is negative, whereas the left-hand-side is nonnegative, so we are done in this case. Next, assume that $x \leq -\ln(\varepsilon)$ holds. If $x = 0$, then (ii) is equivalent to $1-\varepsilon p \geq 1-\varepsilon$. For $x > 0$, we calculate

$$\begin{aligned}
(1-\varepsilon p)\cdot(1-xp)^{\frac{y}{p}} \ &= \ (1-xp)^{\frac{y}{p}} - \varepsilon p\cdot(1-xp)^{\frac{y}{p}} && \mid p \in \left(0, \frac{1}{2}\right) \\
&\geq \ \left((1-xp)^{\frac{1}{xp}}\right)^{xy} - \frac{\varepsilon}{2} && \mid (23) \\
&\geq \ (1-\delta)^{xy}\cdot e^{-xy} - \frac{\varepsilon}{2} && \mid x \leq -\ln(\varepsilon) \\
&\geq \ (1-\delta)^{-\frac{2\ln(\varepsilon)}{\varepsilon}}\cdot e^{-xy} - \frac{\varepsilon}{2} && \mid (22) \\
&\geq \ \left(1-\frac{\varepsilon}{2}\right)\cdot e^{-xy} - \frac{\varepsilon}{2} \\
&\geq \ e^{-xy} - \varepsilon.
\end{aligned}$$

Finally, as $\lim_{x\to\infty}(1+x)\cdot e^{-p\varepsilon^2 x} = 0$, we may choose n subject to (iii) and (iv). ◀

To prove Theorem 2 we use the following auxiliary lemma multiple times:

▶ **Lemma 20.** *For $\beta \in \left(0, \frac{1}{p}\right]$ and $k \in \mathbb{N}_{\geq \frac{\gamma}{p}}$ we have*

$$\sum_{i=1}^{k}(\beta p)^2\cdot(1-\beta p)^{i-1}\cdot c_i \ = \ \beta\gamma + (1-\beta p)^{\frac{\gamma}{p}} - (1+\beta pk)(1-\beta p)^k, \tag{24}$$

where $0^0 \coloneqq 1$.

Proof. The case $\beta p = 1$ is straightforward. Next, assume $\beta \in (0, \frac{1}{p})$. For $k = \frac{\gamma}{p}$, we have

$$\begin{aligned}
&\sum_{i=1}^{\frac{\gamma}{p}}(\beta p)^2\cdot(1-\beta p)^{i-1}\cdot c_i \ = \ \sum_{i=1}^{\frac{\gamma}{p}}(\beta p)^2\cdot(1-\beta p)^{i-1}\cdot\frac{\gamma}{p} \ = \ (\beta\gamma)\cdot\sum_{i=1}^{\frac{\gamma}{p}}(\beta p)\cdot(1-\beta p)^{i-1} \\
&= \ (\beta\gamma)\cdot\left(1-(1-\beta p)^{\frac{\gamma}{p}}\right) \ = \ \beta\gamma + (1-\beta p)^{\frac{\gamma}{p}} - \left(1+\beta p\cdot\frac{\gamma}{p}\right)(1-\beta p)^{\frac{\gamma}{p}}.
\end{aligned}$$

For the case $k \geq \frac{\gamma}{p} + 1$, we use Lemma 18 to compute

$$\begin{aligned}
&\sum_{i=1}^{k}(\beta p)^2\cdot(1-\beta p)^{i-1}\cdot c_i \\
&= \ \sum_{i=1}^{\frac{\gamma}{p}}(\beta p)^2\cdot(1-\beta p)^{i-1}\cdot c_i + (\beta p)^2\cdot\left(\sum_{i=\frac{\gamma}{p}+1}^{\infty} i\cdot(1-\beta p)^{i-1} - \sum_{i=k+1}^{\infty} i\cdot(1-\beta p)^{i-1}\right) \\
&= \ (\beta\gamma)\cdot\left(1-(1-\beta p)^{\frac{\gamma}{p}}\right) + (1-\beta p)^{\frac{\gamma}{p}}\cdot\left(1+\frac{\gamma}{p}\beta p\right) - (1-\beta p)^k\cdot(1+k\beta p) \\
&= \ \beta\gamma + (1-\beta p)^{\frac{\gamma}{p}} - (1+\beta pk)(1-\beta p)^k.
\end{aligned}$$

◀

Now we proceed to the main part of the proof of Theorem 2. First, we aim for an upper bound on the expected cost of an optimum a priori tour:

▶ **Lemma 21.** *The optimum a priori tour for the considered instance has expected cost at most* $(1+\varepsilon)\cdot(\gamma+e^{-\gamma})$.

Proof. Consider the a priori tour T^* that visits $v_0,\dots,v_{n-1}$ in this order. This a priori tour has expected cost

$$\begin{aligned}
&\sum_{i=1}^{n-2}\sum_{j=1}^{n-1-i} p^2\cdot(1-p)^{i-1}\cdot c(v_j,v_{j+i}) + \sum_{i=1}^{n-1} p\cdot(1-p)^{i-1}\cdot\big(c(v_0,v_i)+c(v_{n-i},v_0)\big)\\
&\le \sum_{i=1}^{n-2} p^2\cdot(1-p)^{i-1}\cdot c_i + 2\sum_{i=1}^{n-1} p\cdot(1-p)^{i-1}\cdot\frac{c_i}{n}\\
&\le (1+\varepsilon)\cdot\sum_{i=1}^{n-1} p^2\cdot(1-p)^{i-1}\cdot c_i\\
&\le (1+\varepsilon)\cdot(\gamma+(1-p)^{\frac{\gamma}{p}})\\
&\le (1+\varepsilon)\cdot(\gamma+e^{-\gamma}),
\end{aligned}$$

where we used $\frac{2}{n}\le\varepsilon p$ (which follows from Lemma 19 (iii)) in the second inequality and Lemma 20 for $\beta=1$ and $k=n-1$ in the third inequality. ◀

Second, we aim for a lower bound on the expected cost of the master route solution: We define $q:V\to[0,1]$ by setting $q(v):=\sigma p$ for $v\in V\setminus\{v_0\}$ and $q(v_0):=1$.

▶ **Lemma 22.** *The expected cost of the master route solution when sampling each customer* $v\in V$ *with probability* $q(v)$ *is at least*

$$\min\left\{\frac{e^{-2}}{\varepsilon}-1,\ (1-\varepsilon)^3\cdot(1-4\varepsilon)\cdot\left(\alpha(\sigma\gamma+e^{-\sigma\gamma})+2\gamma+\frac{1}{\sigma}e^{-2\sigma\gamma}\right)\right\}.$$

Proof. We distinguish several cases, depending on how large σ is.
Case 1: $\sigma\le\varepsilon$.

In this case, we consider the connection cost only. For each v_i with $\lceil\frac{1}{\varepsilon p}\rceil\le i\le n-\lceil\frac{1}{\varepsilon p}\rceil$, the probability that no sampled vertex is fewer than $\lceil\frac{1}{\varepsilon p}\rceil$ hops on T^* away from v_i is at least

$$(1-\sigma p)^{2\lceil\frac{1}{\varepsilon p}\rceil-1}\ \ge\ (1-\varepsilon p)^{1+\frac{2}{\varepsilon p}}\ \ge\ e^{-2}-\varepsilon,$$

where we used Lemma 19 (ii) with $x=\varepsilon$ and $y=\frac{2}{\varepsilon}$. In this event, if v_i is active, we have to pay connection cost at least $2\cdot\frac{1}{n}\lceil\frac{1}{\varepsilon p}\rceil$. Hence the total connection cost is at least

$$(e^{-2}-\varepsilon)\cdot\left(n+1-2\left\lceil\frac{1}{\varepsilon p}\right\rceil\right)\cdot\frac{2p}{n}\left\lceil\frac{1}{\varepsilon p}\right\rceil\ \ge\ (e^{-2}-\varepsilon)\cdot\left(n-1-\frac{2}{\varepsilon p}\right)\cdot\frac{2}{\varepsilon n}\ \ge\ \frac{e^{-2}}{\varepsilon}-1,$$

where we used that $n\ge\frac{4}{\varepsilon p}+2$ by Lemma 19 (iii) in the last inequality.
Case 2: $\sigma>\varepsilon$. Let S denote the set of sampled customers, and, if $|S|\ge2$, let $e_{\max}[S]$ be a longest edge of the tour $T^*[S]$ that we get from T^* by skipping the customers that were not sampled. Note that S and $e_{\max}[S]$ are random variables that depend on the sampling.

We want to compute a lower bound on the expected cost of the master tour. It is not always true that $T^*[S]$ is an optimum TSP tour for S, but almost. Namely, if $|S|=1$, the statement is true, and otherwise, we claim that $c(T^*[S])-c(e_{\max}[S])$ is a lower bound on the cost of any TSP tour for S. This follows from

▷ Claim 23. For every $\{v_0\} \subsetneq S \subseteq V$, $T^*[S] \setminus \{e_{\max}[S]\}$ is a min-cost spanning tree in (S, c).

Proof. To prove this, we may assume (by the cyclic symmetry of c) that $S = \{v_i : i \in I\}$ and $e_{\max}[S] = \{v_{\min I}, v_{\max I}\}$; then let $i, \ell \in I$ with $i < \ell$, and let $\{j, k\}$ be any edge on the path from i to ℓ in the tree $T^*[S] \setminus \{e_{\max}[S]\}$ (i.e., $i \le j < k \le \ell$ and $j, k \in I$). We show $c(v_i, v_\ell) \ge c(v_j, v_k)$, which implies optimality of the spanning tree $T^*[S] \setminus \{e_{\max}[S]\}$ (see, e.g., Theorem 6.3 in [21]). Indeed, $c(v_i, v_\ell) \ge \frac{1}{n} \max\{\frac{\gamma}{p}, \min\{\ell - i, n + i - \ell\}\} \ge \frac{1}{n} \max\{\frac{\gamma}{p}, \min\{k - j, n + \min I - \max I\}\} \ge \min\{c(v_j, v_k), c(e_{\max}[I])\} \ge c(v_j, v_k)$. This concludes the proof of Claim 23. ◁

Next we bound the expected cost of $e_{\max}[S]$, or, to be more precise,

$$\sum_{\{v_0\} \subsetneq U \subseteq V} \mathbb{P}_{S \sim q}[S = U] \cdot c(e_{max}[U]).$$

The probability that $T^*[S]$ contains an edge of cost at least $\varepsilon + \frac{1}{n}$ is at most

$$\sum_{j=0}^{n-1} (1 - \sigma p)^{(\varepsilon + \frac{1}{n}) \cdot n - 1} \;=\; n \cdot (1 - \sigma p)^{\varepsilon n} \;\le\; n \cdot e^{-\sigma p \varepsilon n} \;\le\; n \cdot e^{-p\varepsilon^2 n} \;\le\; \varepsilon - \frac{1}{n}$$

by Lemma 19 (iv) (when choosing $x = n$). As every edge costs less than 1, the expected length of $e_{\max}[S]$ is less than 2ε.

Therefore, using Claim 23, the expected cost of the optimum master tour for S is at least $c(T^*[S]) - 2\varepsilon$, and the expectation cannot increase if we sample every customer, including the depot, with probability σp. Hence, writing $v_i = v_{i-n}$ for $i = n, \ldots, 2n-2$ we get as lower bound

$$\begin{aligned}
\mathbb{E}_{S \sim q}\left[\mathrm{OPT}_{\mathrm{TSP}}(S, c)\right] &\ge \mathbb{E}_{S \sim q}\left[c(T^*[S])\right] - 2\varepsilon \\
&\ge \sum_{i=1}^{n-1} \sum_{j=0}^{n-1} (\sigma p)^2 \cdot (1 - \sigma p)^{i-1} \cdot c(v_j, v_{j+i}) - 2\varepsilon \\
&= \sum_{i=1}^{n-1} (\sigma p)^2 \cdot (1 - \sigma p)^{i-1} \cdot c_i - 2\varepsilon \\
&= \left(\sigma\gamma + (1 - \sigma p)^{\frac{\gamma}{p}} - (1 + \sigma p(n-1))(1 - \sigma p)^{n-1}\right) - 2\varepsilon,
\end{aligned}$$

where we used Lemma 20 for $k = n - 1$ and $\beta = \sigma$ (recall $\varepsilon < \sigma \le \frac{1}{p}$) in the last equation. Using $\sigma \ge \varepsilon$, we can apply Lemma 19 (iv) with $x = \sigma(n-1)$ and obtain

$$(1 + \sigma p(n-1)) \cdot (1 - \sigma p)^{n-1} \;\le\; (1 + \sigma(n-1)) \cdot e^{-\sigma p(n-1)} \;\le\; (1 + \sigma(n-1)) \cdot e^{-p\varepsilon^2 \sigma(n-1)} \;\le\; \varepsilon.$$

This yields

$$\begin{aligned}
\mathbb{E}_{S \sim q}\left[\mathrm{OPT}_{\mathrm{TSP}}(S, c)\right] &\ge \sigma\gamma + (1 - \sigma p)^{\frac{\gamma}{p}} - 3\varepsilon \\
&\ge \sigma\gamma + e^{-\sigma\gamma} - 4\varepsilon \\
&\ge (1 - 4\varepsilon) \cdot (\sigma\gamma + e^{-\sigma\gamma}) \qquad (25)
\end{aligned}$$

using Lemma 19 (ii) for $x = \sigma$ and $y = \gamma$ in the second inequality and $x + e^{-x} \ge 1$ for all $x \in \mathbb{R}$ in the last inequality.

Case 2a: $\sigma \ge \frac{\varepsilon}{p}$.

Then we get that (25) is at least

$$(1-4\varepsilon)\cdot\frac{\gamma\varepsilon}{p} \;\geq\; \frac{1}{\varepsilon} \;\geq\; \frac{e^{-2}}{\varepsilon}-1,$$

where we used $\gamma \geq 1$ and Lemma 19 (i) in the first inequality.

Case 2b: $\varepsilon < \sigma < \frac{\varepsilon}{p}$.

We obtain a lower bound on the connection costs as follows: Let $N := \left\lfloor \frac{\varepsilon n-1}{2} \right\rfloor$. Note that $N \geq \frac{\gamma}{p}$ since $n \geq \frac{2\gamma}{\varepsilon p} + \frac{3}{\varepsilon}$ by Lemma 19 (iii). We only consider vertices v_i with $N+1 \leq i \leq n-1-N$ and connect them to a sampled customer that is at most N hops away on T^*, if such a customer exists. Otherwise, we do not connect them at all. This yields a lower bound for the expected cost of connecting the active customers to the master tour of

$$\begin{aligned}
&\sum_{j=N+1}^{n-1-N}\sum_{i=1}^{N} 2p\cdot(1-\sigma p)^{2i-1}\cdot\mathbb{P}_{S\sim q}\left[S\cap\{v_{j-i},v_{j+i}\}\neq\emptyset\right]\cdot c(v_j,v_{j+i})\\
&= \frac{n-1-2N}{n}\cdot\sum_{i=1}^{N} 2p\cdot(1-\sigma p)^{2i-1}\cdot\sigma p(2-\sigma p)\cdot c_i && \mid N=\left\lfloor\frac{\varepsilon n-1}{2}\right\rfloor\\
&\geq (1-\varepsilon)\cdot\sum_{i=1}^{N} 2p\cdot(1-\sigma p)^{2i-1}\cdot\sigma p(2-\sigma p)\cdot c_i\\
&\geq (1-\varepsilon)\cdot(2-\sigma p)\cdot(1-\sigma p)\cdot\frac{1}{2\sigma}\sum_{i=1}^{N}(2\sigma p)^2\cdot(1-2\sigma p)^{i-1}\cdot c_i && \mid \sigma p\leq\varepsilon\\
&\geq (1-\varepsilon)^3\cdot\frac{1}{\sigma}\cdot\sum_{i=1}^{N}(2\sigma p)^2\cdot(1-2\sigma p)^{i-1}\cdot c_i && \mid \text{Lemma 20}\\
&= (1-\varepsilon)^3\cdot\frac{1}{\sigma}\cdot\left(2\sigma\gamma+(1-2\sigma p)^{\frac{\gamma}{p}}-(1+2\sigma pN)\cdot(1-2\sigma p)^N\right) && \mid \text{Lemma 19 (ii)}\\
&\geq (1-\varepsilon)^3\cdot\frac{1}{\sigma}\cdot\left(2\sigma\gamma+e^{-2\sigma\gamma}-\varepsilon-(1+2\sigma pN)\cdot(1-2\sigma p)^N\right)\\
&\geq (1-\varepsilon)^3\cdot\frac{1}{\sigma}\cdot\left(2\sigma\gamma+e^{-2\sigma\gamma}-2\varepsilon\right)\\
&\geq (1-\varepsilon)^3\cdot(1-2\varepsilon)\cdot\frac{1}{\sigma}\cdot\left(2\sigma\gamma+e^{-2\sigma\gamma}\right).
\end{aligned}$$

Note that for the penultimate inequality we used Lemma 19 (iv) for $x = 2\sigma pN \geq \varepsilon^2 pn - 3$ and

$$(1+x)\cdot\left(1-\frac{x}{N}\right)^N \;\leq\; (1+x)\cdot e^{-x} \;\leq\; (1+x)\cdot e^{-p\varepsilon^2 x}.$$

The last inequality follows since $x + e^{-x} \geq 1$ for all $x \in \mathbb{R}$.

If we compute a TSP tour on the sampled customers that costs α times more than optimal, the computed master tour has expected cost at least α times (25). Hence, adding up the expected cost of the master tour and the expected connection cost yields at least

$$(1-\varepsilon)^3\cdot(1-4\varepsilon)\cdot\left(\alpha(\sigma\gamma+e^{-\sigma\gamma})+2\gamma+\frac{1}{\sigma}e^{-2\sigma\gamma}\right).$$

◀

Proof of Theorem 2. Putting together Lemma 21 and Lemma 22 and considering the limit $\varepsilon \to 0$, the ratio between the expected cost of the master route solution that we get from sampling and the expected cost of an optimum a priori tour is at least

$$\frac{\alpha(\sigma\gamma + e^{-\sigma\gamma}) + 2\gamma + \frac{1}{\sigma}e^{-2\sigma\gamma}}{\gamma + e^{-\gamma}}. \tag{26}$$

For any fixed α and γ, this ratio is minimized for the unique positive σ for which

$$\sigma^2\alpha\gamma(e^{2\sigma\gamma} - e^{\sigma\gamma}) \;=\; 1 + 2\sigma\gamma,$$

but apparently there is no closed-form solution. Optimizing this term numerically gives the bounds shown in Figure 5. Note that due to uniform activation probabilities, the function f in the sampling algorithm is irrelevant except for the value of $f(p)$, which is completely determined by σ. ◀

4 Upper bound on the master route ratio

Our proof of the upper bound on the master route ratio (for normalized A PRIORI TSP instances) follows a similar line as the proof of Theorem 3 in Section 2. However, there are some important differences and further complications. In particular, it is not easy to bound the expected cost for connecting the active customers to the master tour in terms of the buckets introduced in Section 2.2. This will require another level of aggregation. As in Section 2, let β, b_0 be constants that we will choose later.

4.1 An optimization problem for the master route ratio

Consider a normalized instance, and let p denote the (uniform) activation probability. Let T^* be a fixed optimum a priori tour, with customers appearing in the order $v_0, v_1, \ldots, v_{n-1}$; here v_0 denotes the depot. Let $v_i := v_0$ for $i < 0$ or $i > n-1$. As in Section 2.1, we define for $k \in \mathbb{Z}_{\geq 1}$

$$C_k \;:=\; p^2 \cdot \sum_{j\in\mathbb{Z}} c(v_j, v_{j+k}),$$

which are nonnegative variables satisfying the triangle inequality

$$C_{i+j} \;\leq\; C_i + C_j \tag{27}$$

for all $i, j \geq 1$. As seen in Proposition 10, the expected cost of T^* is exactly

$$\sum_{i=1}^{\infty}(1-p)^{i-1} \cdot C_i. \tag{28}$$

We now design master route solutions. For $k \geq 1$, consider a master route solution in which the master tour contains the depot and, with some offset $h \in \{1, \ldots, k\}$, every k-th customer on T^*, i.e., $\{v_j : j \in \mathbb{Z}, j \equiv h \pmod k\}$. Note that this means that for $h \geq n$, the master tour consists only of the depot (recall that $v_j = v_0$ for $j < 0$ and $j > n-1$). We will now show that, for any fixed k, the expected cost of the best such solution is at most

$$\frac{1}{kp^2}\left(C_k + 2p \cdot \sum_{i=1}^{k-1} \min\left\{C_i,\, C_{k-i}\right\}\right). \tag{29}$$

If we choose the offset uniformly at random, the master tour has expected cost $\frac{C_k}{kp^2}$. Indeed, for $k < n$, this directly follows by the definition of C_k, and for $k \geq n$, the master tour has expected cost

$$\mathbb{P}[h < n] \cdot \frac{C_{n-1}}{(n-1)p^2} \;=\; \frac{C_{n-1}}{kp^2} \;=\; \frac{C_k}{kp^2}$$

as claimed.

Now we bound the expected cost of connecting the active customers to the master tour. Again we do this by considering only the following two options for each active customer v: the first sampled customer that we encounter when traversing T^* from v in either direction. Connecting to other sampled customers may be cheaper, but we ignore this again and still obtain an upper bound. Now, for offset h, we obtain an upper bound on the connection cost of

$$2p \cdot \sum_{\substack{j \in \mathbb{Z}: \\ j \equiv h \pmod k}} \sum_{i=1}^{k-1} \min\{c(v_j, v_{j+i}), c(v_{j+i}, v_{j+k})\}$$

since the master tour contains precisely the customers v_j with $j \equiv h \pmod k$. Choosing h uniformly at random, this results in a bound on the connection cost of

$$\begin{aligned}
&\frac{2p}{k} \cdot \sum_{i=1}^{k-1} \sum_{j \in \mathbb{Z}} \min\{c(v_j, v_{j+i}), c(v_{j+i}, v_{j+k})\} \\
&\leq \frac{2p}{k} \cdot \sum_{i=1}^{k-1} \min\left\{ \sum_{j \in \mathbb{Z}} c(v_j, v_{j+i}), \sum_{j \in \mathbb{Z}} c(v_j, v_{j+k-i}) \right\} \\
&= \frac{2}{pk} \cdot \sum_{i=1}^{k-1} \min\{C_i, C_{k-i}\}.
\end{aligned}$$

This concludes the proof that (29) is an upper bound on the best "equidistant" master route solution, and hence on the best master route solution per se.

So for every $k \in \mathbb{N}$, the ratio of (29) to (28) is an upper bound on the master route ratio for that instance. In other words, minimizing (28) subject to the constraints that (29) is equal to 1 and the C_i are nonnegative and satisfy the triangle inequality (27) yields the reciprocal of an upper bound on the master route ratio of all normalized instances with activation probability p. Note that only requiring that (29) is at least 1 does not change the minimum. Again, the number n of customers appears neither in (28) nor in (29). We arrive at the following optimization problem:

$$\min \sum_{i=1}^{\infty} (1-p)^{i-1} \cdot C_i \qquad \text{(Master-Route-Ratio-OP)}$$

$$\begin{aligned}
\text{subject to} \qquad C_i &\geq 0 && \text{for } i \in \mathbb{N} && (30)\\
C_i + C_j &\geq C_{i+j} && \text{for } i, j \in \mathbb{N} && (31)\\
C_k + 2p \cdot \sum_{i=1}^{k-1} \min\{C_i, C_{k-i}\} &\geq kp^2 && \text{for } k \in \mathbb{N}. && (32)
\end{aligned}$$

We have proved:

▶ **Lemma 24.** *Let $p > 0$. The reciprocal of the value of* (Master-Route-Ratio-OP) *is an upper bound on the master route ratio for p-normalized instances.* ◀

4.2 Obtaining a single linear program

The optimization problem (Master-Route-Ratio-OP) has infinitely many variables, but one could omit the terms for, say, $i > \frac{100}{p}$ in the infinite sum while still getting a good upper bound. One could also omit the constraints of type (32) for, say, $k > \frac{100}{p}$ while still getting a good upper bound.

However, we would still have an infinite set of LPs (one for each choice of p), and, by Lemma 9, we should consider the limit for $p \to 0$.

In the following, exactly as in Section 2.2, we require that p is of the form $p = \frac{\beta}{b}$ for some odd integer $b \geq b_0$. Note that $p \to 0$ as $b \to \infty$. In order to obtain a single optimization problem for all such values of p, we again put subsequent C_i's into buckets of size b: We define

$$B_i \;\coloneqq \sum_{j=\max\{1, ib-\frac{b-1}{2}\}}^{ib+\frac{b-1}{2}} C_j \tag{33}$$

for $0 \leq i \leq N$ for some integer N that we will choose later. In the following, we show that we can use the constraints in (Master-Route-Ratio-OP) to generate finitely many (slightly relaxed) constraints that only depend on these buckets. Note that for all $i, j \in \mathbb{Z}_{\geq 1}$ with $i + j \leq N$ we have $B_{i+j} \leq B_i + B_j$ by Proposition 12.

For each $1 \leq i \leq N$, we sum the constraints of type (32) for all k in the i-th bucket (i.e., $k = ib - \frac{b-1}{2}, \ldots, ib + \frac{b-1}{2}$), yielding

$$B_i + 2p \sum_{k=-\frac{b-1}{2}}^{\frac{b-1}{2}} \sum_{j=1}^{ib+k-1} \min\left\{C_j,\, C_{ib+k-j}\right\} \;\geq \sum_{k=-\frac{b-1}{2}}^{\frac{b-1}{2}} (ib+k)p^2 \;=\; i(bp)^2. \tag{34}$$

The left-hand side of (34) still contains the C_i variables, and we want to replace them by the new B_i variables.

Since it is not possible to express the minima in the left-hand side of (34) in terms of the buckets without an additional loss, we will need to be a bit pessimistic here. In order to not lose too much, we form bucket intervals

$$A_j \;\coloneqq \sum_{\ell=\max\{j,0\}}^{\min\{j+a,N\}} B_\ell \tag{35}$$

for $j = -a, \ldots, N$, each consisting of $a+1$ subsequent buckets (except for $j < 0$ or $j > N-a$), where a is some odd integer (think of a large but abp small). Roughly speaking, the purpose of the bucket intervals is the following: When deciding to which sampled neighbor on T^* we connect the active customers, we have to make the same decision for all customers in the same bucket interval. It turns out that this way we will lose only a fraction proportional to $\frac{1}{a}$. See Figure 6 for an illustration.

Then, choosing some offset $h \in \{0, \ldots, a-1\}$, we bound the double sum on the left-hand side of (34) as follows, where we use $C_j = 0$ for $j \leq 0$ and $j > Nb + \frac{b-1}{2}$ (i.e., C_j does not occur in any bucket), and $B_j = 0$ for $j < 0$ and $j > N$ in intermediate steps to simplify the

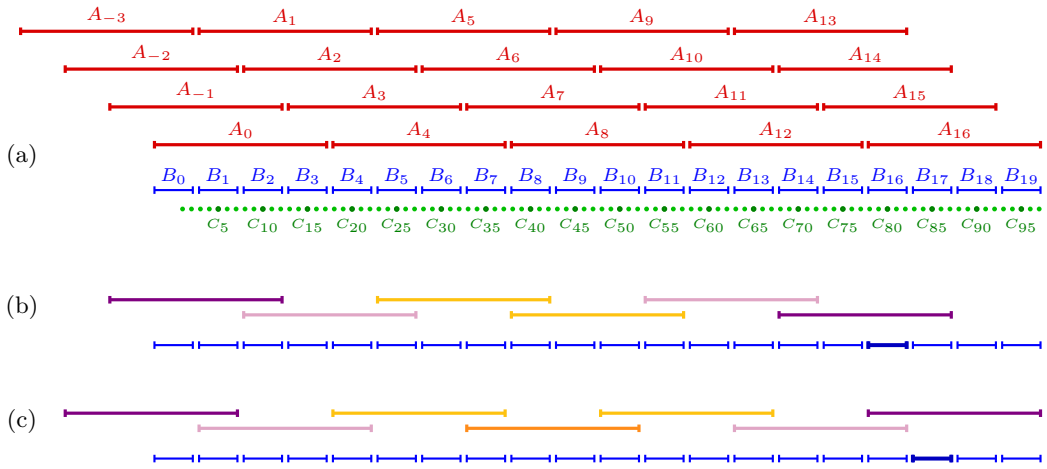


Figure 6 (a): The green dots stand for $C_1, C_2, \ldots$, and the centers of the buckets (C_{ib} for $i \geq 1$) are highlighted. Here the bucket size is $b = 5$, and the blue intervals show the buckets $B_0, B_1, B_2, \ldots$ We now form bucket intervals, shown in red on the top (each consisting of $a + 1 = 4$ buckets; here $a = 3$; we can think of adding empty buckets on the left). (b): Aggregating the constraints (32) for bucket B_{16} (i.e., $k \in \{78, 79, 80, 81, 82\}$) yields (34) and then (36) for $i = 16$; here we collect the terms on the left-hand side according to the bucket intervals shown. (c): The same for bucket B_{17}.

terms.

$$\begin{aligned}
&\sum_{k=-\frac{b-1}{2}}^{\frac{b-1}{2}} \sum_{j=1}^{ib+k-1} \min\left\{C_j,\, C_{ib+k-j}\right\} \\
&\leq \sum_{k=-\frac{b-1}{2}}^{\frac{b-1}{2}} \sum_{j=1}^{ib+k-1+hb} \min\left\{C_{j-hb},\, C_{ib+k-j+hb}\right\} \\
&\leq \sum_{k=-\frac{b-1}{2}}^{\frac{b-1}{2}} \sum_{j=1}^{\left\lceil \frac{h+i+1}{a} \right\rceil} \sum_{\ell=1}^{ab} \min\{C_{(j-1)ab+\ell-hb},\, C_{ib+k-(j-1)ab-\ell+hb}\} \\
&\leq \sum_{k=-\frac{b-1}{2}}^{\frac{b-1}{2}} \sum_{j=1}^{\left\lceil \frac{h+i+1}{a} \right\rceil} \min\left\{\sum_{\ell=1}^{ab} C_{(j-1)ab+\ell-hb},\, \sum_{\ell=1}^{ab} C_{ib+k-(j-1)ab-\ell+hb}\right\} \\
&= \sum_{k=-\frac{b-1}{2}}^{\frac{b-1}{2}} \sum_{j=1}^{\left\lceil \frac{h+i+1}{a} \right\rceil} \min\left\{\sum_{\ell=1}^{ab} C_{(j-1)ab+\ell-hb},\, \sum_{\ell=1}^{ab} C_{ib+k-jab+\ell-1+hb}\right\} \\
&\leq \sum_{k=-\frac{b-1}{2}}^{\frac{b-1}{2}} \sum_{j=1}^{\left\lceil \frac{h+i+1}{a} \right\rceil} \min\left\{\sum_{\ell=0}^{a} B_{(j-1)a+\ell-h},\, \sum_{\ell=0}^{a} B_{i-ja+\ell+h}\right\} \\
&= b \sum_{j=1}^{\left\lceil \frac{h+i+1}{a} \right\rceil} \min\left\{A_{(j-1)a-h},\, A_{i-ja+h}\right\}.
\end{aligned}$$

To minimize the loss, we choose $h_i = (i \bmod a)$ because this implies that different buckets are

counted twice for different values of i, i.e., the buckets B_{aj-h_i} $(j \geq 1)$ are counted twice, as Figure 6 (b) and (c) indicate, but note that this consideration is not needed for correctness. In any case, we conclude that for any $1 \leq i \leq N$,

$$B_i + 2bp \sum_{j=1}^{\left\lceil \frac{h_i+i+1}{a} \right\rceil} \min\left\{A_{(j-1)a-h_i},\, A_{i-ja+h_i}\right\} \;\geq\; i(bp)^2 \tag{36}$$

is a relaxation of (34), and hence of (32) summed for $k = ib - \frac{b-1}{2}, \ldots, ib + \frac{b-1}{2}$.

Therefore, we obtain a lower bound on the value of (Master-Route-Ratio-OP) by minimizing $\sum_{i=1}^{\infty}(1-p)^{i-1} \cdot C_i$ subject to (35) and (36) and $B_{i+j} \geq B_i + B_j$ for $i, j \geq 1$ with $i + j \leq N$, and $B_i \geq 0$ for $i \geq 0$.

The objective function still contains infinitely many variables and depends on p, and we resolve this in the same way as in Section 2.2. Again, as in (12),

$$\sum_{i=1}^{\infty}(1-p)^{i-1} \cdot C_i \;\geq\; \sum_{i=0}^{\infty}(1-p)^{bi+\frac{b-1}{2}-1} \cdot B_i \;\geq\; \sum_{i=0}^{N}(1-p)^{(i+\frac{1}{2})b} \cdot B_i. \tag{37}$$

Moreover, $p = \frac{\beta}{b}$ and $\lim_{b\to\infty}(1-\frac{\beta}{b})^{(i+\frac{1}{2})b} = e^{-(i+\frac{1}{2})\beta}$ for all $i = 0, \ldots, N$. Again, we will use this to replace the objective function by $\sum_{i=0}^{N} e^{-(i+\frac{1}{2})\beta} \cdot B_i$ and still get an upper bound.

Finally, we omit many triangle inequalities and keep only $i \cdot B_1 \geq B_i$ for $i = 2, \ldots, N$. Moreover, we further introduce auxiliary variables for the minima in (36). Here is our final linear program:

$$\min \sum_{i=0}^{N} e^{-(i+\frac{1}{2})\beta} \cdot B_i \qquad \text{(Master-Route-Ratio-LP)}$$

$$\text{subject to} \qquad i \cdot B_1 \;\geq\; B_i \qquad \text{for } i = 2, \ldots, N \tag{38}$$

$$B_i + 2\beta \sum_{j=1}^{\left\lceil \frac{h_i+i+1}{a} \right\rceil} M_{j,i} \;\geq\; i\beta^2 \qquad \text{for } i = 1, \ldots, N \tag{39}$$

$$A_{(j-1)a-h_i} \;\geq\; M_{j,i} \qquad \text{for } i = 1, \ldots, N \text{ and } j = 1, \ldots, \left\lceil \tfrac{h_i+i+1}{a} \right\rceil \tag{40}$$

$$A_{i-ja+h_i} \;\geq\; M_{j,i} \qquad \text{for } i = 1, \ldots, N \text{ and } j = 1, \ldots, \left\lceil \tfrac{h_i+i+1}{a} \right\rceil \tag{41}$$

$$\sum_{\ell=\max\{i,\,0\}}^{\min\{i+a,\,N\}} B_\ell \;=\; A_i \qquad \text{for } i = -a, \ldots, N \tag{42}$$

$$B, M \;\geq\; 0. \tag{43}$$

We conclude:

▶ **Lemma 25.** *Let N, a be integers with a odd and let $\beta > 0$. Let $h_i \in \{0, \ldots, a-1\}$ for $i = 1, \ldots, N$. The reciprocal of the optimum value of* (Master-Route-Ratio-LP) *is an upper bound on the master route ratio for* A PRIORI *TSP instances with depot.*

Proof. We proceed analogously to the proof of Lemma 15. We compare the value of (Master-Route-Ratio-LP) to the value of (Master-Route-Ratio-OP). We showed that for any feasible solution C to (Master-Route-Ratio-OP) we obtain a feasible solution (A, B, M) to (Master-Route-Ratio-LP) via (33) and (35) and $M_{j,i} = \min\{A_{(j-1)a-h_i}, A_{i-ja+h_i}\}$.

Fix $\delta > 0$. Then there exists $b_0 \in \mathbb{N}$ such that for all odd integers $b \geq b_0$

$$\sum_{i=0}^{N} e^{-(i+\frac{1}{2})\beta} \cdot B_i \;\leq\; (1+\delta) \cdot \sum_{i=0}^{N} \left(1-\tfrac{\beta}{b}\right)^{(i+\frac{1}{2})b} \cdot B_i \;\leq\; (1+\delta) \cdot \sum_{i=0}^{\infty} \left(1-\tfrac{\beta}{b}\right)^{i-1} \cdot C_i.$$

Thus we conclude that the value of (Master-Route-Ratio-LP) is at most $(1+\delta)$ times the value of (Master-Route-Ratio-OP) with $p = \frac{\beta}{b}$ for all odd integers $b \geq b_0$. Hence, by Lemma 11, $(1+\delta)$ times the reciprocal of the value of (Master-Route-Ratio-LP) is an upper bound on the master route ratio for all $\frac{\beta}{b}$-normalized instances for all odd integers $b \geq b_0$. By Lemma 9, the same bound then holds for all instances with depot. Since this bound holds for all $\delta > 0$, it also holds for $\delta = 0$. ◀

4.3 Solving the dual LP

Now we dualize the LP (Master-Route-Ratio-LP). For this, we introduce variables $(x_i)_{i=2,\dots,N}$ for the inequalities of type (38), variables $(y_i)_{i=1,\dots,N}$ for the inequalities of type (39), variables $v_{i,j}$ and $w_{i,j}$ for the inequalities of type (40) and (41), respectively, and variables $(z_i)_{i=-a,\dots,N}$ for the inequalities of type (42).

$$\max \sum_{i=1}^{N} i\beta^2 \cdot y_i \qquad \text{(Dual-Master-Route-Ratio-LP)}$$

$$\begin{aligned}
\text{subject to} \quad y_i + \sum_{j=i-a}^{i} z_j + \mathbb{1}_{i=1} \sum_{j=2}^{N} j \cdot x_j - \mathbb{1}_{2\leq i\leq N} \cdot x_i &\leq e^{-(i+\frac{1}{2})\beta} && \text{for } i = 0,\dots,N \\
2\beta \cdot y_i &\leq v_{j,i} + w_{j,i} && \text{for } i = 1,\dots,N \\
& && \text{and } j = 1,\dots,\left\lceil \tfrac{h_i+i+1}{a} \right\rceil \\
\sum_{i=1}^{N} \sum_{j=1}^{\left\lceil \frac{h_i+i+1}{a} \right\rceil} \left(\mathbb{1}_{k=(j-1)a-h_i} \cdot v_{j,i} + \mathbb{1}_{k=i-ja+h_i} \cdot w_{j,i} \right) &= z_k && \text{for } k = -a,\dots,N \\
x, y, v, w &\geq 0.
\end{aligned}$$

▶ **Corollary 26.** *Let N, a be integers with a odd and let $\beta > 0$. Let $h_i \in \{0,\dots,a-1\}$ for $i = 1,\dots,N$. For any feasible solution (x,y,v,w,z) to* (Dual-Master-Route-Ratio-LP), $\left(\sum_{i=1}^{N} i\beta^2 \cdot y_i\right)^{-1}$ *is an upper bound on the master route ratio restricted to* A PRIORI *TSP instances with depot.* ◀

We have computed a dual solution using Gurobi 10.0.1 with $\beta = \frac{1}{400}$, $a = 199$, $N = 4000$, and $h_i = (i \bmod a)$ for $i = 1,\dots,N$, yielding an upper bound of 2.584 and thus proving Theorem 6. The dual solution and a Python script that verifies that this is indeed a feasible solution to (Dual-Master-Route-Ratio-LP) can be found at `https://doi.org/10.60507/FK2/JCUIRI`.

5 Lower bound on the master route ratio

We now present an example that proves that the master route ratio is larger than 2.541.

▶ **Theorem 27.** *The master route ratio is at least* $\frac{1}{1-e^{-\frac{1}{2}}} > 2.541$.

Proof. We provide a sequence of A PRIORI TSP instances where the master route ratio converges to $\frac{1}{1-e^{-\frac{1}{2}}}$. Consider the complete graph on $n+1$ vertices $d, v_1, \dots, v_n$, and let the distance between any two vertices be 1. The vertex d is our depot, and each other vertex is replaced by a group of m customers that have distance 0 from each other. More formally, we have

$$V = \{d\} \cup \bigcup_{i=1}^{n} V_i, \qquad V_i = \{v_{i,1}, \dots, v_{i,m}\}$$

and a distance function given by $c(d, v_{i,j}) = 1$ and

$$c(v_{i,j}, v_{i',j'}) \;=\; \begin{cases} 1 & i \neq i' \\ 0 & i = i' \end{cases}.$$

All activation probabilities are given by $p(v_{i,j}) = \frac{1}{2m}$, except for $p(d) = 1$.

Consider a master route solution for this instance. We may assume that the depot d belongs to the master tour because otherwise including it can only make the master route solution better. Suppose the master tour visits k of the groups, for some $k \in \{0, \dots, n\}$. The cost of the master tour is then 0 if $k = 0$ and $k+1$ otherwise. Connecting customers to the master tour does not cost anything if there is a customer in the same group that belongs to the master tour. Each of the other $m(n-k)$ customers has to be connected to the master tour with probability $\frac{1}{2m}$ and cost $2 \cdot 1$, yielding a total expected connection cost of $n - k$. Hence the expected cost of this master route solution is

$$\begin{cases} n & \text{if } k = 0 \\ k + 1 + (n-k) & \text{if } k \geq 1 \end{cases}.$$

Thus in the best master route solution, the master tour consists only of the depot ($k = 0$), and the expected cost is n.

The optimum a priori tour visits the customers in each group consecutively. Its cost only depends on the number of groups in which at least one customer is active. The probability that a group V_i contains an active customer is $q \coloneqq 1-(1-\frac{1}{2m})^m$, which converges to $1-e^{-\frac{1}{2}}$ as $m \to \infty$. For every fixed n there is thus a sufficiently large m such that the expected number of active groups is at most

$$nq \;\leq\; n\left(1 - e^{-\frac{1}{2}}\right) + 1.$$

The cost of the resulting tour visiting the active customers is at most the number of active groups plus one (for the depot d). Hence the expected cost of the optimum a priori tour is at most

$$nq + 1 \;\leq\; n\left(1 - e^{-\frac{1}{2}}\right) + 2.$$

This implies that the master route ratio is at least

$$\frac{n}{n\left(1 - e^{-\frac{1}{2}}\right) + 2} \xrightarrow{n\to\infty} \frac{1}{1 - e^{-\frac{1}{2}}}.$$

◀

6 Introducing a depot (Proof of Theorem 1)

In this section, we show that assuming that the instance contains a *depot*, i.e., an element $d \in V$ with $p(d) = 1$, is no severe restriction. We remark that although we can condition on the event that at least two customers are active (because otherwise the cost is zero regardless of the a priori tour), we cannot simply guess these a priori and declare one of them to be the depot. As mentioned earlier, the previous works [25] and [28] provided ad hoc proofs that *their* algorithms generalize to the non-depot case. Theorem 1 yields a general reduction, losing only an arbitrarily small constant.

The proof consists of two parts. First we show that for instances whose total activation probability is bounded by a constant, we can find an almost optimal master route solution,

yielding a $(3+\varepsilon)$-approximation. The second part is to prove that otherwise it does not harm much to declare one customer to be the depot.

For instances in which the expected number of active customers is small (bounded by a constant) it is crucial to take into account that the cost of an a priori tour cut short to fewer than two customers is zero: For example, if $c(v,w) = 1$ for all $v, w \in V$ with $v \neq w$ and $p(v) = \frac{\varepsilon}{n}$ for all $v \in V$ (where $n = |V|$ and $\varepsilon > 0$ tends to zero), then OPT $\approx \varepsilon^2$ but the standard analysis of any master route solution (even if consisting only of a single point) yields cost at least roughly 2ε. However, using the fact that the cost is zero if fewer than two customers are active, Shmoys and Talwar [25] show that the a priori tour resulting from an optimum master route solution is *always* at most a factor 3 worse than an optimum a priori tour. The other good news is that a near-optimal master route solution can be found in polynomial time if the expected number of active customers is bounded by a constant, which has some similarities to a PTAS found by Eisenbrand, Grandoni, Rothvoß and Schäfer [7] for connected facility location in a bounded case. We need the following fact on the sum of independent Bernoulli random variables:

▶ **Lemma 28.** *Let $\varepsilon > 0$ and $k > 0$ be constants. Then there is a constant $N = N(k,\varepsilon)$ such that the following holds. Let $X_1, \dots, X_n$ be independent Bernoulli variables and $X = \sum_{i=1}^n X_i$ with $\mathbb{P}[X \geq 2] > 0$ and $\mathbb{E}[X] \leq k$. Then $\mathbb{P}[X > N \mid X \geq 2] \leq \varepsilon$.*

Proof. Let $p_i := \mathbb{P}[X_i = 1]$ for $i = 1, \dots, n$. We have $\sum_{i=1}^n p_i = \mathbb{E}[X] \leq k$.

Let $S := \{i \in \{1,\dots,n\} : p_i \leq \frac{1}{2}\}$ be the indices of the Bernoulli variables with small success probability. Let $X' = \sum_{i \in S} X_i$. We always have

$$X - X' \;\leq\; \left|\left\{i \in \{1,\dots,n\} : p_i > \frac{1}{2}\right\}\right| \;<\; 2k.$$

To bound the probability that X' is large, we compute, for $m > 2$,

$$\begin{aligned}
\mathbb{P}[X' = m] &= \sum_{T \in \binom{S}{m}} \prod_{j \in T} p_j \cdot \prod_{j \in S \setminus T} (1 - p_j) \\
&= \frac{1}{m} \sum_{i \in S} p_i \cdot \sum_{T \in \binom{S \setminus \{i\}}{m-1}} \prod_{j \in T} p_j \cdot \prod_{j \in S \setminus (T \cup \{i\})} (1 - p_j) \\
&= \frac{1}{m} \sum_{i \in S} \frac{p_i}{1 - p_i} \cdot \sum_{T \in \binom{S \setminus \{i\}}{m-1}} \prod_{j \in T} p_j \cdot \prod_{j \in S \setminus T} (1 - p_j) \\
&\leq \frac{1}{m} \sum_{i \in S} \frac{p_i}{1 - p_i} \cdot \sum_{T \in \binom{S}{m-1}} \prod_{j \in T} p_j \cdot \prod_{j \in S \setminus T} (1 - p_j) \\
&= \frac{1}{m} \sum_{i \in S} \frac{p_i}{1 - p_i} \cdot \mathbb{P}[X' = m - 1] \\
&\leq \frac{1}{m} \sum_{i \in S} 2p_i \cdot \mathbb{P}[X' = m - 1] \\
&\leq \frac{2k}{m} \cdot \mathbb{P}[X' = m - 1],
\end{aligned}$$

where we used $p_i \leq \frac{1}{2}$ for $i \in S$ in the second inequality. An iterative application yields

$$\mathbb{P}[X' = m] \;\leq\; \frac{2 \cdot (2k)^{m-2}}{m!} \cdot \mathbb{P}[X' = 2].$$

Hence, for any integer $\ell \geq 2ek$,

$$\begin{aligned}\mathbb{P}[X \geq 2k+\ell] \;&\leq\; \mathbb{P}[X' \geq \ell+1] \;\leq\; \sum_{m=\ell+1}^{\infty} \frac{2\cdot(2k)^{m-2}}{m!}\cdot\mathbb{P}[X'=2] \\ &\leq\; \frac{2}{\ell}\cdot\mathbb{P}[X'=2] \;\leq\; \frac{2}{\ell}\cdot\mathbb{P}[X\geq 2],\end{aligned}$$

where the third inequality follows (with Stirling's formula) from

$$\begin{aligned}\sum_{m=\ell+1}^{\infty}\frac{(2k)^{m-2}}{m!} \;&\leq\; \sum_{m=\ell+1}^{\infty}\frac{(\frac{\ell}{e})^{m-2}}{m!} \;\leq\; \frac{e(\frac{\ell}{e})^{\ell}}{\ell(\ell+1)\ell!}\cdot\sum_{i=0}^{\infty}e^{-i} \\ &=\; \frac{e^2}{(e-1)\ell(\ell+1)}\cdot\frac{(\frac{\ell}{e})^{\ell}}{\ell!} \;\leq\; \frac{e^2}{\sqrt{2\pi\ell}(e-1)\ell(\ell+1)} \;\leq\; \frac{1}{\ell}\end{aligned}$$

for $\ell \geq 1$. Setting $\ell = \lceil \max\{2ek, \frac{2}{\varepsilon}\}\rceil$ and $N = 2k+\ell$ finishes the proof. ◀

Now we are ready to prove the following:

▶ **Lemma 29.** *Let $k > 0$ and $\varepsilon > 0$ be constants. Then we can find an a priori tour with expected cost at most $(3+\varepsilon)\cdot\mathrm{OPT}$ for any given A PRIORI TSP instance (V,c,p) with $\sum_{v\in V} p(v) \leq k$ in polynomial time.*

Proof. Shmoys and Talwar [25] showed that if we randomly sample a subset $S \subseteq V$ with at least two elements, where S is chosen with probability $(\mathbb{P}_{A\sim p}[A=S])/(\mathbb{P}_{A\sim p}[|A|\geq 2])$, then any a priori tour T_S corresponding to the master route solution with an optimum TSP tour for S as master tour has expected cost at most $3\cdot\mathrm{OPT}$.

Now we exploit that $\sum_{v\in V} p(v) \leq k$. We may assume $\varepsilon \leq 1$. By Lemma 28 there is a constant $N = N(k, \frac{\varepsilon}{4})$ such that, for the randomly chosen S, the probability that $|S| > N$ is at most $\frac{\varepsilon}{4}$. Hence for at least one set S with $2 \leq |S| \leq N$ we have

$$\mathbb{E}_{A\sim p}\left[c(T_S[A])\right] \;\leq\; \frac{3}{1-\frac{\varepsilon}{4}}\cdot\mathrm{OPT} \;\leq\; (3+\varepsilon)\cdot\mathrm{OPT}.$$

Since N is a constant, we can find such a set S and an optimum TSP tour for S by complete enumeration. ◀

By Lemma 29 we can assume that the expected activity is high. Then assuming a depot incurs an arbitrarily small loss only:

▶ **Lemma 30.** *Let $\varepsilon > 0$ and $\rho \geq 3$ be constants. If there is a (randomized) polynomial-time ρ-approximation algorithm for instances of the A PRIORI TSP that have a depot d with $p(d) = 1$, then there is a (randomized) polynomial-time $(\rho+\varepsilon)$-approximation algorithm for instances (V,c,p) with $\sum_{v\in V} p(v) \geq \frac{2\rho}{\varepsilon}$.*

Proof. Let $p(V) \coloneqq \sum_{v\in V} p(v)$ and assume $p(V) \geq \frac{2\rho}{\varepsilon}$. We try all $v \in V$, redefine $p(v) = 1$ (i.e., make v the depot), and call the ρ-approximation algorithm on the resulting instance.

More precisely, for $v \in V$, let $p^{v=d}$ be defined by $p^{v=d}(u) = p(u)$ for $u \in V\setminus\{v\}$ and $p^{v=d}(v) = 1$. In addition, let $\mathrm{OPT}^{v=d}$ denote the expected cost of an optimum a priori tour for the modified instance where we replace p by $p^{v=d}$. Calling the ρ-approximation for instances with depot yields a solution T^v with expected cost $\mathbb{E}_{A\sim p^{v=d}}[c(T^v[A])] \leq \rho\cdot\mathrm{OPT}^{v=d}$.

Note that for every TSP tour T for V and any $v \in V$, we have

$$\begin{aligned}
&\sum_{U \subseteq V} \mathbb{P}_{A \sim p}[A = U] \cdot c(T[U \cup \{v\}]) \\
&= \sum_{U \subseteq V : v \in U} \left(\mathbb{P}_{A \sim p}[A = U \setminus \{v\}] + \mathbb{P}_{A \sim p}[A = U]\right) \cdot c(T[U]) \\
&= \sum_{U \subseteq V : v \in U} \prod_{u \in U \setminus \{v\}} p(u) \cdot \prod_{w \in V \setminus U} (1 - p(w)) \cdot c(T[U]) \\
&= \sum_{U \subseteq V : v \in U} \mathbb{P}_{A \sim p^{v=d}}[A = U] \cdot c(T[U]) \\
&= \mathbb{E}_{A \sim p^{v=d}}[c(T[A])].
\end{aligned}$$

We will now apply this equation twice, first to the tours T^v and then to an optimum solution T^* to the original instance. This way we can evaluate how good the solutions T^v are for the original instance:

$$\begin{aligned}
&\mathbb{E}_{A \sim p}[c(T^v[A])] \\
&= \sum_{U \subseteq V} \mathbb{P}_{A \sim p}[A = U] \cdot c(T^v[U]) \\
&\leq \sum_{U \subseteq V} \mathbb{P}_{A \sim p}[A = U] \cdot c(T^v[U \cup \{v\}]) \\
&= \mathbb{E}_{A \sim p^{v=d}}[c(T^v[A])] \\
&\leq \rho \cdot \mathrm{OPT}^{v=d} \\
&\leq \rho \cdot \mathbb{E}_{A \sim p^{v=d}}[c(T^*[A])] \\
&= \rho \cdot \sum_{U \subseteq V} \mathbb{P}_{A \sim p}[A = U] \cdot c(T^*[U \cup \{v\}]) \\
&\leq \rho \cdot \sum_{U \subseteq V} \mathbb{P}_{A \sim p}[A = U] \cdot c(T^*[U]) + \rho \cdot \sum_{\substack{U \subseteq V \\ U \setminus \{v\} \neq \emptyset}} \mathbb{P}_{A \sim p}[A = U] \cdot \left(c(v_U^-, v) + c(v, v_U^+)\right),
\end{aligned}$$

where v_U^- and v_U^+ are the first customers in $U \setminus \{v\}$ that we encounter when we traverse a fixed orientation of T^* from v in backward and in forward direction, respectively. Given $u \neq w \in T^*$, we denote the set of customers (other than u and w) that we visit when traversing T^* in the given orientation from u to w by $T^*_{(u,w)}$. Taking the weighted average of these bounds, where the bound for T^v is weighted with $\frac{p(v)}{p(V)}$, we conclude that the expected cost of the best such solution T^v is at most

$$\rho \cdot \sum_{U \subseteq V} \mathbb{P}_{A \sim p}[A = U] \cdot c(T^*[U]) + \sum_{v \in V} \frac{\rho \cdot p(v)}{p(V)} \cdot \sum_{\substack{U \subseteq V \\ U \setminus \{v\} \neq \emptyset}} \mathbb{P}_{A \sim p}[A = U] \cdot \left(c(v_U^-, v) + c(v, v_U^+)\right).$$

Now for any ordered pair $(u, w) \in V^2$, the term $c(u, w)$ appears in the right-hand sum-

mand with coefficient

$$\begin{aligned}
&\frac{\rho \cdot p(w)}{p(V)} \cdot \mathbb{P}_{A\sim p}[w_A^- = u] + \frac{\rho \cdot p(u)}{p(V)} \cdot \mathbb{P}_{A\sim p}[u_A^+ = w] \\
&= \frac{2\rho}{p(V)} \cdot p(u) \cdot p(w) \cdot \prod_{v \in T^*_{(u,w)}} (1 - p(v)) \\
&= \frac{2\rho}{p(V)} \cdot \mathbb{P}_{A\sim p}[u \in A,\, w = u_A^+] \\
&\leq \varepsilon \cdot \mathbb{P}_{A\sim p}[u \in A,\, w = u_A^+]
\end{aligned}$$

because $p(V) \geq \frac{2\rho}{\varepsilon}$. Summing over all ordered pairs (u, w) yields a bound of $\varepsilon \cdot \mathbb{E}_{A\sim p}[c(T^*[A])]$. Hence the expected cost of the best T^v is at most

$$(\rho + \varepsilon) \cdot \mathbb{E}_{A\sim p}[c(T^*[A])]$$

as required.

It is easy to compute the exact expected cost of each T^v deterministically in $O(|V|^3)$ time (and thus choose the best) via

$$\mathbb{E}_{A\sim p}[c(T^v[A])] = \sum_{\{u,w\}\in\binom{V}{2}} p(u) \cdot p(w) \cdot \left(\prod_{t \in T^v_{(u,w)}} (1 - p(t)) + \prod_{t \in T^v_{(w,u)}} (1 - p(t)) \right) \cdot c(u, w).$$

◀

The proof of Theorem 1 follows from combining Lemma 29 and Lemma 30.

7 Reducing to normalized instances (Proof of Lemma 9)

In this section we prove Lemma 9, i.e., that upper bounds on the master route ratio and the approximation ratio of the sampling algorithm for normalized instances with low uniform activation probability yield the same bounds for all instances of the A PRIORI TSP with a depot. The high-level idea of this reduction is rather simple: we replace each customer by many copies, each with the same very low activation probability, such that for each customer the probability that at least one of its copies is active roughly matches the probability that the original customer is active. However, as we will see, it requires quite some technical care to prove Lemma 9 formally.

We first prove a few auxiliary statements that will be useful later.

▶ **Proposition 31.** *Let $0 < x < y < 1$. Then*

$$\frac{x}{y} - x < \frac{\ln(1-x)}{\ln(1-y)} < \frac{x}{y}.$$

Proof. For the first inequality, we calculate

$$-\ln(1-x) = \int_{1-x}^{1} \frac{1}{t}\, dt > x \quad \text{and} \quad -\ln(1-y) = \int_{1-y}^{1} \frac{1}{t}\, dt < \frac{y}{1-y}.$$

For the second inequality, we compute

$$\begin{aligned}
-\ln(1-y) &= \int_{1-y}^{1} \frac{1}{t}\,dt \\
&> \int_{1-y}^{1-x} \frac{1}{1-x}\,dt + \int_{1-x}^{1} \frac{1}{t}\,dt \\
&= \frac{y-x}{1-x} + \int_{1-x}^{1} \frac{1}{t}\,dt \\
&> \frac{y-x}{x} \cdot \int_{1-x}^{1} \frac{1}{t}\,dt + \int_{1-x}^{1} \frac{1}{t}\,dt \\
&= \frac{y}{x} \cdot \int_{1-x}^{1} \frac{1}{t}\,dt \\
&= -\frac{y}{x} \cdot \ln(1-x).
\end{aligned}$$

◀

▶ **Lemma 32.** *Let $\sigma \in (0,1)$, $\lambda > 0$ and $p \in (0,1]$. Then there is $\varepsilon_p \in (0,1)$ with the following property: For every $\varepsilon \in (0, \varepsilon_p]$ there is $k \in \mathbb{N}$ such that*

(i) $1-(1-\varepsilon)^k \le p \le \varepsilon k$,
(ii) $(1-p)^\sigma \le (1-\sigma\varepsilon)^k$,
(iii) $1-(1-p)^\sigma \le (1+\lambda)\cdot(1-(1-\sigma\varepsilon)^k)$,
(iv) $(1-\varepsilon)^k \le (1+\lambda)\cdot(1-p)$ *if* $p<1$, *and* $(1-\varepsilon)^k \le \lambda$ *if* $p=1$.

Proof. For $\varepsilon \in (0,1)$ choose

$$k := \begin{cases} \left\lfloor \frac{\ln(1-p)}{\ln(1-\varepsilon)} \right\rfloor & \text{if } p<1 \\ \left\lceil \max\left\{ \frac{1}{\varepsilon}, \frac{\ln(\frac{\lambda}{1+\lambda})}{\ln(1-\sigma\varepsilon)} \right\} \right\rceil & \text{if } p=1 \end{cases}.$$

We first handle the easier case $p=1$, in which the inequalities hold for any fixed $\varepsilon \in (0,1)$. Indeed, the first inequality in (i) and (ii) follow directly from $p=1$. Moreover, the second inequality in (i) is implied by our choice of k. Finally, the definition of k yields

$$(1-\varepsilon)^k \le (1-\sigma\varepsilon)^k \le \frac{\lambda}{1+\lambda} < \lambda,$$

which implies (iii) and (iv).

Now, we deal with the more interesting case $p<1$. Applying Proposition 31 with $x = \frac{p}{2}$ and $y = p$ yields

$$-\ln(1-p) \;>\; -2\ln(1-\tfrac{p}{2}) \;>\; 0.$$

Thus, there is $\varepsilon_1 > 0$ such that

$$-\ln(1-p) \;\ge\; (1+\varepsilon_1)\cdot(-2)\cdot\ln(1-\tfrac{p}{2}).$$

As $p>0$, we have $(1-p)^\sigma < 1$, so there is $\kappa > 0$ with

$$1-(1+\kappa)\cdot(1-p)^\sigma \;\ge\; (1+\lambda)^{-1}\cdot(1-(1-p)^\sigma). \tag{44}$$

Pick $\varepsilon_2 > 0$ such that

$$(1-\sigma\varepsilon_2)^{-1} \cdot (1-p)^{-\sigma\varepsilon_2} \;\leq\; 1+\kappa.$$

Finally, let $\varepsilon_p \coloneqq \min\{p \cdot \varepsilon_1, \varepsilon_2, \frac{p}{3}, \frac{\lambda}{1+\lambda}\}$. Now let $0 < \varepsilon \leq \varepsilon_p$. Then we have

$$\frac{\ln(1-p)}{\ln(1-\varepsilon)} \;\geq\; \frac{p}{\varepsilon} \cdot \left(1 + \frac{\varepsilon}{p}\right) \tag{45}$$

because

$$\left(1+\frac{\varepsilon}{p}\right) \cdot \frac{-\ln(1-\varepsilon)}{\varepsilon} \;\leq\; (1+\varepsilon_1) \cdot \frac{-\ln(1-\varepsilon)}{\varepsilon} \;<\; (1+\varepsilon_1) \cdot \frac{-\ln(1-\frac{p}{2})}{\frac{p}{2}} \leq \frac{-\ln(1-p)}{p},$$

where we used $\varepsilon_p \leq p \cdot \varepsilon_1$ in the first inequality, Proposition 31 with $x = \varepsilon \leq \frac{p}{3} < \frac{p}{2} = y$ in the second inequality, and the definition of ε_1 in the third inequality. By definition of ε_p and ε_2, we also have

$$(1-\sigma\varepsilon)^{-1} \cdot (1-p)^{-\sigma\varepsilon} \;\leq\; 1+\kappa. \tag{46}$$

To prove (i), we compute

$$\begin{aligned}
1-(1-\varepsilon)^k \;&\leq\; 1-(1-\varepsilon)^{\frac{\ln(1-p)}{\ln(1-\varepsilon)}} \\
&=\; p \\
&=\; \varepsilon \cdot \left(\frac{p}{\varepsilon} \cdot \left(1+\frac{\varepsilon}{p}\right) - 1\right) \\
&\overset{(45)}{\leq}\; \varepsilon \cdot \left(\frac{\ln(1-p)}{\ln(1-\varepsilon)} - 1\right) \\
&\leq\; \varepsilon \cdot k.
\end{aligned}$$

To prove (ii), we calculate

$$(1-\sigma\varepsilon)^k \;\geq\; (1-\sigma\varepsilon)^{\frac{\ln(1-p)}{\ln(1-\varepsilon)}} \;=\; (1-p)^{\frac{\ln(1-\sigma\varepsilon)}{\ln(1-\varepsilon)}} \;\geq\; (1-p)^{\frac{\sigma\varepsilon}{\varepsilon}} = (1-p)^{\sigma}.$$

Here, the last inequality follows since $1-p \in [0,1)$ and $\frac{\ln(1-\sigma\varepsilon)}{\ln(1-\varepsilon)} < \frac{\sigma\cdot\varepsilon}{\varepsilon}$ by Proposition 31.

To prove (iii), we compute

$$\begin{aligned}
(1-\sigma\varepsilon)^k \;&\leq\; (1-\sigma\varepsilon)^{-1} \cdot (1-\sigma\varepsilon)^{\frac{\ln(1-p)}{\ln(1-\varepsilon)}} \\
&=\; (1-\sigma\varepsilon)^{-1} \cdot (1-p)^{\frac{\ln(1-\sigma\varepsilon)}{\ln(1-\varepsilon)}} \\
&\leq\; (1-\sigma\varepsilon)^{-1} \cdot (1-p)^{\sigma\cdot(1-\varepsilon)} \\
&\overset{(46)}{\leq}\; (1+\kappa) \cdot (1-p)^{\sigma},
\end{aligned}$$

where the second inequality follows from Proposition 31 with $x = \sigma\varepsilon$ and $y = \varepsilon$. Hence,

$$1-(1-\sigma\varepsilon)^k \;\geq\; 1-(1+\kappa)\cdot(1-p)^{\sigma} \;\overset{(44)}{\geq}\; (1+\lambda)^{-1} \cdot (1-(1-p)^{\sigma}).$$

Finally, to prove (iv), we calculate

$$(1+\lambda)^{-1} \cdot (1-\varepsilon)^k \;=\; \left(1-\frac{\lambda}{1+\lambda}\right) \cdot (1-\varepsilon)^k \;\leq\; (1-\varepsilon)^{k+1} \;\leq\; (1-\varepsilon)^{\frac{\ln(1-p)}{\ln(1-\varepsilon)}} \;=\; 1-p.$$

◀

▶ **Lemma 33.** *Let (V, c, p) be an instance of* A PRIORI TSP *with depot d and let T^* be an optimum a priori tour for this instance. Let $p' : V \to (0, 1]$ such that $p'(v) \leq p(v)$ for all $v \in V$ and $p'(d) = p(d) = 1$. Then*

$$\mathrm{OPT}(V, c, p') \;\leq\; \mathbb{E}_{A\sim p'}[c(T^*[A])] \;\leq\; \mathbb{E}_{A\sim p}[c(T^*[A])] \;=\; \mathrm{OPT}(V, c, p).$$

Proof. It suffices to prove the lemma for the special case where p and p' only differ in exactly one customer because the general case follows by induction. Thus, let (V, c, p) be an instance of A PRIORI TSP with depot d, let $w \in V \setminus \{d\}$ and let $p' : V \to (0, 1]$ such that $p'(w) \leq p(w)$ and $p'(v) = p(v)$ for all $v \in V \setminus \{w\}$. Pick an optimum a priori tour T^* for (V, c, p). Then T^* is also feasible for (V, c, p'). We compute

$$\begin{aligned}
\mathrm{OPT}(V, c, p') \;&\leq\; \mathbb{E}_{A\sim p'}[c(T^*[A])] \\
&= \sum_{S\subseteq V\setminus\{w\}} \mathbb{P}_{A\sim p'}[A \setminus \{w\} = S] \cdot \big(p'(w) \cdot c(T^*[S \cup \{w\}]) + (1 - p'(w)) \cdot c(T^*[S])\big) \\
&= \sum_{S\subseteq V\setminus\{w\}} \mathbb{P}_{A\sim p}[A \setminus \{w\} = S] \cdot \big(p'(w) \cdot c(T^*[S \cup \{w\}]) + (1 - p'(w)) \cdot c(T^*[S])\big) \\
&\leq \sum_{S\subseteq V\setminus\{w\}} \mathbb{P}_{A\sim p}[A \setminus \{w\} = S] \cdot \big(p(w) \cdot c(T^*[S \cup \{w\}]) + (1 - p(w)) \cdot c(T^*[S])\big) \\
&= \mathbb{E}_{A\sim p}[c(T^*[A])] \;=\; \mathrm{OPT}(V, c, p).
\end{aligned}$$

The last inequality follows from $p'(w) \leq p(w)$ and $c(T^*[S]) \leq c(T^*[S \cup \{w\}])$. ◀

▶ **Lemma 34.** *Let (V, c, p) be an instance of* A PRIORI TSP *with depot d and let $d \in S \subseteq V$. Then*

$$\mathrm{MR}(S) \;=\; \left(1 - \prod_{v\in V\setminus\{d\}} (1 - p(v))\right) \cdot \mathrm{OPT}_{TSP}(S, c) + 2 \cdot \sum_{v\in V\setminus\{d\}} p(v) \cdot c(v, S).$$

Proof. By definition, we have

$$\begin{aligned}
\mathrm{MR}(S) \;&=\; \mathbb{E}_{A\sim p}\left[\mathbb{1}_{|A|\geq 2} \cdot \left(\mathrm{OPT}_{\mathrm{TSP}}(S, c) + 2 \cdot \sum_{v\in A} c(v, S)\right)\right] \\
&=\; \mathbb{P}_{A\sim p}[|A| \geq 2] \cdot \mathrm{OPT}_{\mathrm{TSP}}(S, c) + 2 \cdot \mathbb{E}_{A\sim p}\left[\mathbb{1}_{|A|\geq 2} \cdot \sum_{v\in A} c(v, S)\right].
\end{aligned}$$

As the depot is always active, we have $|A| < 2$ if and only if every customer in $V \setminus \{d\}$ is inactive, which happens with probability $\prod_{v\in V\setminus\{d\}}(1 - p(v))$. Hence,

$$\mathbb{P}_{A\sim p}[|A| \geq 2] \;=\; 1 - \prod_{v\in V\setminus\{d\}} (1 - p(v)).$$

We further observe that

$$\mathbb{E}_{A\sim p}\left[\mathbb{1}_{|A|\geq 2} \cdot \sum_{v\in A} c(v, S)\right] \;=\; \mathbb{E}_{A\sim p}\left[\sum_{v\in A\setminus\{d\}} c(v, S)\right] = \sum_{v\in V\setminus\{d\}} p(v) \cdot c(v, S),$$

where the first equality follows from the facts that $\sum_{v\in A} c(v, S) = \sum_{v\in A\setminus\{d\}} c(v, S)$ since $d \in S$ and that this sum can only be nonzero if $|A| \geq 2$ because d is always active. ◀

Now we are ready to prove Lemma 9.

Proof of Lemma 9. Let (V, c, p) be an instance of A PRIORI TSP with depot d. We show that for every $\delta > 0$, there exist $i \in \mathbb{N}$ and an ε_i-normalized instance (V', c', p') with depot d such that

(i) $\mathrm{OPT}(V', c', p') \leq \mathrm{OPT}(V, c, p)$;
(ii) $(1+\delta) \cdot \min_{S' \subseteq V' : d \in S'} \mathrm{MR}(S') \geq \min_{S \subseteq V : d \in S} \mathrm{MR}(S)$;
(iii) Let $f(q) \coloneqq 1-(1-q)^\sigma$ and $f'(q) \coloneqq \sigma \cdot q$ for all $q \in [0,1)$, and $f(1) \coloneqq f'(1) \coloneqq 1$. Then

$$(1+\delta) \cdot \mathbb{E}_{S \sim f' \circ p'} \left[\alpha \cdot \mathrm{OPT}_{\mathrm{TSP}}(S, c') + 2 \cdot \sum_{v \in V'} p(v) \cdot c'(v, S) \right]$$
$$\geq \mathbb{E}_{S \sim f \circ p} \left[\alpha \cdot \mathrm{OPT}_{\mathrm{TSP}}(S, c) + 2 \cdot \sum_{v \in V} p(v) \cdot c(v, S) \right].$$

Note that this immediately gives Lemma 9. Let $n \coloneqq |V|$ and pick $\lambda > 0$ such that $1 - \lambda \geq (1+\delta)^{-1}$, $(1+\lambda)^{n-1} \leq 1+\delta$ and

$$(1+\delta) \cdot \left(1 - (1+\lambda)^{n-1} \cdot \prod_{v \in V \setminus \{d\}} (1-p(v)) \right) \;\geq\; 1 - \prod_{v \in V \setminus \{d\}} (1-p(v)). \tag{47}$$

Observe that this is possible because $\prod_{v \in V \setminus \{d\}} (1-p(v)) < 1$ since $p(v) > 0$ for all $v \in V$, meaning that (47) is a strict inequality for $\lambda = 0$. Apply Lemma 32 to obtain constants $\varepsilon_{p(v)} > 0$ for all $v \in V$. Pick i such that $\varepsilon_i \leq \min_{v \in V} \varepsilon_{p(v)}$. Let (V', c', p') result from (V, c, p) by keeping the depot d with $p(d) = 1$ unchanged, and replacing each other $v \in V$ by k_v copies of v, each with activation probability ε_i, where k_v is chosen as the number k in Lemma 32 for $p = p(v)$ and $\varepsilon = \varepsilon_i$. Denote the projection of V' onto V by π. Then

$$\begin{aligned} \mathrm{OPT}_{\mathrm{TSP}}(S', c') &= \mathrm{OPT}_{\mathrm{TSP}}(\pi(S'), c) && \text{and} \\ c'(v', S') &= c(\pi(v'), \pi(S')) && \text{for all } S' \subseteq V', v' \in V'. \end{aligned} \tag{48}$$

First, we show (i). Let T^* be an optimum a priori tour for (V, c, p) and let T' arise from T^* by replacing each $v \in V \setminus \{d\}$ by its k_v copies. Then T' is a feasible solution for (V', c', p'). Now, sample $S' \subseteq V'$ according to the activation probabilities p'. We always have $c'(T'[S']) = c(T^*[\pi(S')])$. Define $\bar{p}(d) \coloneqq 1$ and $\bar{p}(v) \coloneqq 1 - (1-\varepsilon_i)^{k_v}$ for each $v \in V \setminus \{d\}$. Note that $\bar{p}(v) \leq p(v)$ by Lemma 32 (i). Moreover, for $v \in V$, the probability that $v \in \pi(S')$ equals $\bar{p}(v)$. Hence, the expected cost of T' equals the expected cost of T^* for the instance $(V, c, \bar{p})$. Thus, Lemma 33 allows us to conclude that $\mathrm{OPT}(V', c', p') \leq \mathrm{OPT}(V, c, p)$.

Next we show (ii). Let $S' \subseteq V'$ with $d \in S'$ and let $S \coloneqq \pi(S')$. Note that $\mathrm{OPT}_{\mathrm{TSP}}(S, c) = \mathrm{OPT}_{\mathrm{TSP}}(S', c')$. We claim that

$$(1+\delta)^{-1} \cdot \left(1 - \prod_{v \in V \setminus \{d\}} (1-p(v)) \right) \;\leq\; 1 - \prod_{v' \in V' \setminus \{d\}} (1-p'(v')). \tag{49}$$

Indeed, if there is $w \in V \setminus \{d\}$ with $p(w) = 1$, then Lemma 32 (iv) tells us that

$$1 - \prod_{v' \in V' \setminus \{d\}} (1-p'(v')) \;\geq\; 1-(1-\varepsilon_i)^{k_w} \;\geq\; 1-\lambda \;\geq\; (1+\delta)^{-1}$$
$$= (1+\delta)^{-1} \cdot \left(1 - \prod_{v \in V \setminus \{d\}} (1-p(v)) \right).$$

On the other hand, if $p(v) < 1$ for all $v \in V \setminus \{d\}$, then Lemma 32 (iv) and (47) tell us that

$$\begin{aligned}
&(1+\delta)^{-1} \cdot \left(1 - \prod_{v \in V \setminus \{d\}} (1 - p(v))\right) \\
&\leq\ 1 - \prod_{v \in V \setminus \{d\}} (1+\lambda) \cdot (1 - p(v)) \\
&\leq\ 1 - \prod_{v \in V \setminus \{d\}} (1 - \varepsilon_i)^{k_v} \\
&=\ 1 - \prod_{v' \in V' \setminus \{d\}} (1 - p'(v')).
\end{aligned}$$

By Lemma 34 and Lemma 32 (i), we obtain

$$\begin{aligned}
\mathrm{MR}(S) \ &=\ \left(1 - \prod_{v \in V \setminus \{d\}} (1 - p(v))\right) \cdot \mathrm{OPT}_{\mathrm{TSP}}(S, c) + 2 \cdot \sum_{v \in V \setminus \{d\}} p(v) \cdot c(v, S) \\
&\overset{(49)}{\leq}\ (1+\delta) \cdot \left(1 - \prod_{v' \in V' \setminus \{d\}} (1 - p'(v'))\right) \cdot \mathrm{OPT}_{\mathrm{TSP}}(S', c') + 2 \cdot \sum_{v \in V \setminus \{d\}} p(v) \cdot c(v, S) \\
&\leq\ (1+\delta) \cdot \left(1 - \prod_{v' \in V' \setminus \{d\}} (1 - p'(v'))\right) \cdot \mathrm{OPT}_{\mathrm{TSP}}(S', c') + 2 \cdot \sum_{v \in V \setminus \{d\}} k_v \cdot \varepsilon_i \cdot c(v, S) \\
&=\ (1+\delta) \cdot \left(1 - \prod_{v' \in V' \setminus \{d\}} (1 - p'(v'))\right) \cdot \mathrm{OPT}_{\mathrm{TSP}}(S', c') + 2 \cdot \sum_{v' \in V' \setminus \{d\}} p'(v') \cdot c'(v', S') \\
&=\ (1+\delta) \cdot \mathrm{MR}(S').
\end{aligned}$$

Thus, $(1+\delta) \cdot \mathrm{MR}(S') \geq \mathrm{MR}(S)$.

Finally, we bound the sampling costs as in (iii). For every $U \subseteq V$ with $d \in U$, we have

$$\begin{aligned}
\mathbb{P}_{S \sim f \circ p}[S = U] \ &=\ \prod_{v \in U \setminus \{d\}} (1 - (1 - p(v))^{\sigma}) \cdot \prod_{v \in V \setminus U} (1 - p(v))^{\sigma} \\
&\leq\ \prod_{v \in U \setminus \{d\}} (1+\lambda) \cdot (1 - (1 - \sigma \cdot \varepsilon_i)^{k_v}) \cdot \prod_{v \in V \setminus U} (1 - \sigma \cdot \varepsilon_i)^{k_v} \\
&\leq\ (1+\lambda)^{n-1} \cdot \mathbb{P}_{S' \sim f' \circ p'}[\pi(S') = U] \\
&\leq\ (1+\delta) \cdot \mathbb{P}_{S' \sim f' \circ p'}[\pi(S') = U],
\end{aligned}$$

where we used Lemma 32 (ii) and (iii) in the first inequality. We use this to compare the expected costs of the master tours:

$$\begin{aligned}
\alpha \cdot \mathbb{E}_{S \sim f \circ p}[\mathrm{OPT}_{\mathrm{TSP}}(S, c)] \ &=\ \alpha \cdot \sum_{U \subseteq V} \mathbb{P}_{S \sim f \circ p}[S = U] \cdot \mathrm{OPT}_{\mathrm{TSP}}(U, c) \\
&\leq\ (1+\delta) \cdot \alpha \cdot \sum_{U \subseteq V} \mathbb{P}_{S' \sim f' \circ p'}[\pi(S') = U] \cdot \mathrm{OPT}_{\mathrm{TSP}}(U, c) \\
&\overset{(48)}{=}\ (1+\delta) \cdot \alpha \cdot \sum_{U' \subseteq V'} \mathbb{P}_{S' \sim f' \circ p'}[S' = U'] \cdot \mathrm{OPT}_{\mathrm{TSP}}(U', c') \\
&=\ (1+\delta) \cdot \alpha \cdot \mathbb{E}_{S' \sim f' \circ p'}[\mathrm{OPT}_{\mathrm{TSP}}(S', c')].
\end{aligned}$$

For the connection costs, we compute a bound of

$$\begin{aligned}
&\mathbb{E}_{S\sim f\circ p}\left[\sum_{v\in V} 2\cdot p(v)\cdot c(v,S)\right]\\
&= \sum_{U\subseteq V} \mathbb{P}_{S\sim f\circ p}[S=U]\cdot\sum_{v\in V} 2\cdot p(v)\cdot c(v,U)\\
&\le (1+\delta)\cdot\sum_{U\subseteq V} \mathbb{P}_{S'\sim f'\circ p'}[\pi(S')=U]\cdot\sum_{v\in V} 2\cdot p(v)\cdot c(v,U)\\
&\le (1+\delta)\cdot\sum_{U\subseteq V} \mathbb{P}_{S'\sim f'\circ p'}[\pi(S')=U]\cdot\sum_{v\in V} 2\cdot \varepsilon_i\cdot k_v\cdot c(v,U)\\
&\overset{(48)}{=} (1+\delta)\cdot\sum_{U'\subseteq V'} \mathbb{P}_{S'\sim f'\circ p'}[S'=U']\cdot\sum_{v\in V'} 2\cdot p'(v)\cdot c'(v,U')\\
&= (1+\delta)\cdot\mathbb{E}_{S'\sim f'\circ p'}\left[\sum_{v\in V'} 2\cdot p'(v)\cdot c'(v,S')\right]
\end{aligned}$$

where we used Lemma 32 (i) in the second inequality. This proves (iii). ◀

8 Analysis of the deterministic algorithm via the master route ratio

Our new upper bound on the master route ratio implies a better deterministic approximation algorithm as the following theorem shows. This is essentially due to van Zuylen in [28], although she did not formally define the master route ratio.

▶ **Theorem 35.** *Let ρ denote the master route ratio for A PRIORI TSP instances with depot. Suppose we have an algorithm for (metric) TSP that always computes a tour of cost at most α times the value of the subtour elimination LP. Then there is a deterministic $(2+\alpha\rho)$-approximation algorithm for A PRIORI TSP instances with depot.*

By Theorem 6 we have $\rho < 2.6$. Together with Theorem 1, plugging in $\alpha = 1.5$ [27] or $\alpha = 1.5 - 10^{-36}$ [20], yields Corollary 7. The proof of Theorem 35 follows van Zuylen [28]:

Proof. In order to derandomize the approximation algorithm for A PRIORI TSP sketched in Section 1.2, it suffices to determine the nonempty set S on which we build the master tour in a deterministic way. In light of Lemma 34, our goal is to find a nonempty set S and a TSP tour T for S minimizing

$$\left(1-\prod_{v\in V\setminus\{d\}}(1-p(v))\right)\cdot c(T)+\sum_{v\in V}2p(v)c(v,S).$$

Using the method of conditional expectation, we decide for each customer one by one whether it should be part of S. We maintain a set P of customers chosen to be in S and a set $\overline{P}$ of customers that will not be part of S. Initially, $P=\{d\}$ and $\overline{P}=\emptyset$. Considering a customer $v\notin P\cup\overline{P}$, we compute a pessimistic estimator for the conditional expectation value of

$$\left(1-\prod_{v\in V\setminus\{d\}}(1-p(v))\right)\cdot c(T)+\sum_{v\in V}2p(v)c(v,S)$$

when drawing S according to the activation probabilities – once under the condition that $P\cup\{v\}\subseteq S$ and $\overline{P}\cap S=\emptyset$ and once under the condition that $P\subseteq S$ and $(\overline{P}\cup\{v\})\cap S=\emptyset$. We add v to P or $\overline{P}$ depending on which estimator for the conditional expectation is smaller.

The pessimistic estimator has two components. First, the expected cost for connecting the active customers to the master tour on S, where $P \subseteq S \subseteq V \setminus \overline{P}$ and every $v \in V \setminus (P \cup \overline{P})$ is independently included into S with probability $p(v)$, is

$$\sum_{v \in V} 2p(v) \cdot \mathbb{E}_{S \sim p} \left[c(v, S) \mid P \subseteq S \subseteq V \setminus \overline{P} \right]. \tag{50}$$

It is not hard to see that this can be computed exactly in polynomial time for any $P, \overline{P}$. To bound the conditional expectation of the cost of the master tour, let $E = \binom{V}{2}$ and consider the following linear program, where $Q \coloneqq \left(1 - \prod_{v \in V \setminus \{d\}} (1 - p(v))\right)$ is the probability that the depot is not the only active customer:

$$\begin{aligned} \min \sum_{e \in E} \left(Q \cdot c(e) b_e + \sum_{v \in V \setminus \{d\}} p(v) c(e) r_e^v \right) & & \text{(Master-Route-Solution-LP)} \\ \text{subject to} \sum_{e \in \delta(U)} (b_e + r_e^v) \geq 2 & & \text{for } v \in U \subseteq V \setminus \{d\} \\ b_e, r_e^v \geq 0 & & \text{for } e \in E, v \in V \setminus \{d\}. \end{aligned}$$

The variables b_e represent the edges of the master tour (think of them as edges that we **b**uy) and the variables r_e^v stand for the edges we only use to connect v to the master tour if v is active (think of them as edges that we **r**ent for every customer separately). By Lemma 34, (Master-Route-Solution-LP) is an LP relaxation of the problem of finding the best master route solution. From the definition of the master route ratio it follows that the cost of an optimum solution to (Master-Route-Solution-LP) is most $\rho \cdot \mathrm{OPT}$, where OPT again denotes the expected cost of an optimum a priori tour.

The LP can be solved in polynomial time by standard techniques. Given an optimum solution $\hat{b}, \hat{r}$ to (Master-Route-Solution-LP), we set $Q_{P,\overline{P}} \coloneqq \mathbb{P}_{S \sim p} \left[|S| \geq 2 \mid P \subseteq S \subseteq V \setminus \overline{P} \right]$ and claim that

$$\alpha \cdot \sum_{e \in E} c(e) \left(Q_{P,\overline{P}} \cdot \hat{b}_e + \sum_{v \in P \setminus \{d\}} \hat{r}_e^v + \sum_{v \in V \setminus (P \cup \overline{P})} p(v) \hat{r}_e^v \right) \tag{51}$$

is an upper bound on the conditional expected cost of the master tour. To show this, note that (51) is

$$\mathbb{E}_{S \sim p} \left[\alpha \cdot \sum_{e \in E} c(e) y_e^S \;\middle|\; P \subseteq S \subseteq V \setminus \overline{P} \right],$$

where $y^{\{d\}} \coloneqq 0$ and y^S for $\{d\} \subsetneq S \subseteq V$ is defined by $y_e^S = \hat{b}_e + \sum_{v \in S \setminus \{d\}} \hat{r}_e^v$ for $e \in E$. For all S with $|S| \geq 2$, we have that y^S is a feasible solution to the subtour elimination LP

$$\begin{aligned} \min \sum_{e \in E} c(e) y_e & & \\ \text{subject to} \sum_{e \in \delta(U)} y_e \geq 2 & & \text{for } U \subset V : S \setminus U \neq \emptyset, S \cap U \neq \emptyset \\ y_e \geq 0 & & \text{for } e \in E. \end{aligned}$$

Hence for given S we can find (in polynomial time) a TSP tour for S that has cost at most $\alpha \cdot \sum_{e \in E} c(e) y_e^S$, and (51) is the conditional expectation of this upper bound. Therefore we can use the sum of (51) and (50) as pessimistic estimator.

For $P = \{d\}$ and $\overline{P} = \emptyset$, we note that $Q = Q_{\{d\},\emptyset}$. Moreover, in this case, (50) is at most $2 \cdot \mathrm{OPT}$ (as in Section 1.1), and (51) is $\alpha \sum_{e \in E} \left(Q \cdot c(e) \hat{b}_e + \sum_{v \in V \setminus \{d\}} p(v) c(e) \hat{r}_e^v \right)$, which is at most $\alpha\rho \cdot \mathrm{OPT}$ as we observed above. The conditional expectation never increases during the described procedure as the current value of the conditional expectation is always a convex combination of the two possible next values. Thus the described deterministic algorithm results in an $(\alpha\rho + 2)$-approximation. ◀

9 Discussion

We conjecture (but could not prove) that our lower bound examples are really worst-case examples, and that the values of our linear programs converge to these bounds.

Another question is whether the master route ratio is $\frac{1}{1-e^{-1/2}}$ even for low-activity instances. Currently we only know the upper bound of 3 from [25], but know no example with master route ratio larger than $\frac{1}{1-e^{-1/2}}$ (and this value is attained by our example only as the activity tends to infinity). The analogous question applies to the sampling algorithm: whether we need to consider the low-activity case separately is an open question.

Finally, we hope that our approach can also help for proving a better bound for related problems where similar random sampling techniques are used, or for showing that known bounds are best possible.

References

1 Mohamed Abdellahi Amar, Walid Khaznaji, and Monia Bellalouna. An exact resolution for the probabilistic traveling salesman problem under the a priori strategy. *Procedia Computer Science*, 108:1414–1423, 2017. doi:10.1016/j.procs.2017.05.068.

2 Dimitris Bertsimas. *Probabilistic combinatorial optimization problems*. PhD thesis, Massachusetts Institute of Technology, 1988. URL: https://dspace.mit.edu/handle/1721.1/14386.

3 Dimitris J. Bertsimas, Patrick Jaillet, and Amedeo R. Odoni. A priori optimization. *Operations Research*, 38(6):1019–1033, 1990. doi:10.1287/opre.38.6.1019.

4 Neill E. Bowler, Thomas M. A. Fink, and Robin C. Ball. Characterization of the probabilistic traveling salesman problem. *Phys. Rev. E*, 68:036703, 2003. doi:10.1103/PhysRevE.68.036703.

5 Martijn van Ee, Leo van Iersel, Teun Janssen, and René Sitters. A priori TSP in the scenario model. *Discrete Applied Mathematics*, 250:331–341, 2018. doi:10.1016/j.dam.2018.04.002.

6 Martijn van Ee and René Sitters. The a priori traveling repairman problem. *Algorithmica*, 80(10):2818–2833, 2018. doi:10.1007/s00453-017-0351-z.

7 Friedrich Eisenbrand, Fabrizio Grandoni, Thomas Rothvoß, and Guido Schäfer. Connected facility location via random facility sampling and core detouring. *Journal of Computer and System Sciences*, 76(8):709–726, 2010. doi:10.1016/j.jcss.2010.02.001.

8 Finn Fernstrøm and Teresa Anna Steiner. A constant approximation algorithm for the uniform a priori capacitated vehicle routing problem with unit demands. *Information Processing Letters*, 159-160:105960, 2020. doi:10.1016/j.ipl.2020.105960.

9 Arun Ganesh, Bruce M. Maggs, and Debmalya Panigrahi. Robust algorithms for TSP and Steiner Tree. *ACM Trans. Algorithms*, 19(2), 2023. doi:10.1145/3570957.

10 Naveen Garg, Anupam Gupta, Stefano Leonardi, and Piotr Sankowski. Stochastic analyses for online combinatorial optimization problems. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, page 942–951. SIAM, 2008. URL: https://dl.acm.org/doi/10.5555/1347082.1347185.

11 Michel Goemans and Jon Kleinberg. An improved approximation ratio for the minimum latency problem. *Mathematical Programming*, 82(1):111–124, 1998. doi:10.1007/BF01585867.

12 Igor Gorodezky, Robert D. Kleinberg, David B. Shmoys, and Gwen Spencer. Improved lower bounds for the universal and a priori TSP. In Maria Serna, Ronen Shaltiel, Klaus Jansen, and José Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 178–191. Springer, 2010. doi:10.1007/978-3-642-15369-3_14.

13 Anupam Gupta, Amit Kumar, Martin Pál, and Tim Roughgarden. Approximation via cost sharing: Simpler and better approximation algorithms for network design. *J. ACM*, 54(3), 2007. doi:10.1145/1236457.1236458.

14 Anupam Gupta, Amit Kumar, and Tim Roughgarden. Simpler and better approximation algorithms for network design. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, page 365–372. ACM, 2003. doi:10.1145/780542.780597.

15 Anupam Gupta, Martin Pál, R. Ravi, and Amitabh Sinha. Boosted sampling: Approximation algorithms for stochastic optimization. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, page 417–426. ACM, 2004. doi:10.1145/1007352.1007419.

16 Patrick Jaillet. Probabilistic traveling salesman problems. PhD thesis, Massachusetts Institute of Technology, 1985. URL: https://dspace.mit.edu/handle/1721.1/15231.

17 Patrick Jaillet. A priori solution of a traveling salesman problem in which a random subset of the customers are visited. *Operations Research*, 36(6):929–936, 1988. doi:10.1287/opre.36.6.929.

18 Kamal Jain, Mohammad Mahdian, Evangelos Markakis, Amin Saberi, and Vijay V. Vazirani. Greedy facility location algorithms analyzed using dual fitting with factor-revealing lp. *J. ACM*, 50(6):795–824, 2003. doi:10.1145/950620.950621.

19 Anna R. Karlin, Nathan Klein, and Shayan Oveis Gharan. A (slightly) improved approximation algorithm for metric TSP. In *Proceedings of the 53rd Annual ACM Symposium on Theory of Computing*, page 32–45. ACM, 2021. doi:10.1145/3406325.3451009.

20 Anna R. Karlin, Nathan Klein, and Shayan Oveis Gharan. A deterministic better-than-3/2 approximation algorithm for metric TSP. In Alberto Del Pia and Volker Kaibel, editors, *Integer Programming and Combinatorial Optimization*, pages 261–274. Springer, 2023. doi:10.1007/978-3-031-32726-1_19.

21 Bernhard Korte and Jens Vygen. *Combinatorial Optimization: Theory and Algorithms*. Springer, 6th edition, 2018.

22 Fatemeh Navidi, Inge Li Gørtz, and Viswanath Nagarajan. Approximation algorithms for the a priori traveling repairman. *Operations Research Letters*, 48(5):599–606, 2020. doi:10.1016/j.orl.2020.07.009.

23 Christos H. Papadimitriou and Mihalis Yannakakis. The traveling salesman problem with distances one and two. *Mathematics of Operations Research*, 18(1):1–11, 1993. doi:10.1287/moor.18.1.1.

24 Frans Schalekamp and David B. Shmoys. Algorithms for the universal and a priori TSP. *Operations Research Letters*, 36(1):1–3, 2008. doi:10.1016/j.orl.2007.04.009.

25 David Shmoys and Kunal Talwar. A constant approximation algorithm for the a priori traveling salesman problem. In Andrea Lodi, Alessandro Panconesi, and Giovanni Rinaldi, editors, *Integer Programming and Combinatorial Optimization*, pages 331–343. Springer, 2008. doi:10.1007/978-3-540-68891-4_23.

26 Alejandro Toriello, William B. Haskell, and Michael Poremba. A dynamic traveling salesman problem with stochastic arc costs. *Operations Research*, 62(5):1107–1125, 2014. doi:10.1287/opre.2014.1301.

27 Laurence A. Wolsey. Heuristic analysis, linear programming and branch and bound. *Mathematical Programming Study*, 13:121–134, 1980. doi:10.1007/BFb0120913.

28 Anke van Zuylen. Deterministic sampling algorithms for network design. *Algorithmica*, 60(1):110–151, 2011. doi:10.1007/s00453-009-9344-x.