The Simultaneous Semi-Random Model for TSP

Eric Balkanski, Yuri Faenza, and Mathieu Kubik

IEOR Department, Columbia University

Abstract. Worst-case analysis is a performance measure that is often too pessimistic to indicate which algorithms we should use in practice. A classical example is in the context of the Euclidean Traveling Salesman Problem (TSP) in the plane, where local search performs extremely well in practice even though it only achieves an $\Omega(\frac{\log n}{\log \log n})$ worst-case approximation ratio. In such cases, a natural alternative approach to worst-case analysis is to analyze the performance of algorithms in semi-random models.

In this paper, we propose and investigate a novel semi-random model for the Euclidean TSP. In this model, called the simultaneous semi-random model, an instance over n points consists of the union of an adversarial instance over $(1-\alpha)n$ points and a random instance over αn points, for some $\alpha \in [0,1]$. As with smoothed analysis, the semi-random model interpolates between distributional (random) analysis when $\alpha = 1$ and worst-case analysis when $\alpha = 0$. In contrast to smoothed analysis, this model trades off allowing some completely random points in order to have other points that exhibit a fully arbitrary structure.

We show that with only an $\alpha = \frac{1}{\log n}$ fraction of the points being random, local search achieves an $\mathcal{O}(\log \log n)$ approximation in the simultaneous semi-random model for Euclidean TSP in fixed dimensions. On the other hand, we show that at least a polynomial number of random points are required to obtain an asymptotic improvement in the approximation ratio of local search compared to its worst-case approximation, even in two dimensions.

Keywords: Traveling Salesman Problem · Semi-random Models · Local Search.

1 Introduction

The Traveling Salesman Problem (TSP) is a cornerstone of integer programming and combinatorial optimization, having been investigated for more than 60 years. Since Dantzig, Fulkerson, and Johnson [10] developed the cutting plane method to solve a (then astonishing) 42 cities instance, the TSP has been at the forefront of research in optimization, pushing the limits of computation in practice (see, e.g., [2,5,9,25]), while at the same time being the test bed for many new ideas in the theory of algorithms (see, e.g., [16,18,24,31,32]).

The simplest non-trivial TSP instances are arguably those that go under the name of *Euclidean*: given a set of points in the d-dimensional unit cube, find a

cycle of minimum total length containing all those points (a *tour*), where the length of an edge between any two points is given by their Euclidean distance. Euclidean TSP is NP-hard [15,26], but in fixed dimension a PTAS can be obtained using approximate dynamic programming ideas [3]. However, in practice, even simple algorithms perform very well on Euclidean instances. Take for instance the 2-*opt* local search algorithm: given the current tour T, orient it arbitrarily and let (a, b) and (c, d) be two edges of T, traversed in this order. Consider the tour T' obtained from T by replacing (a, b), (c, d) with (a, c), (b, d) (i.e., performing a *swap*). If T' has a strictly smaller length than T', let $T = T'$ and iterate; else, attempt to swap two different pairs of edges from T. The algorithm halts when all pairs of edges from the current tour T have been tested for a swap, with none leading to an improved tour.

2-opt is known to perform extremely well on many Euclidean instances, such as those from the TSPLIB library, both in terms of convergence time and quality of the output [12,17,27]. However, classical worst-case analysis does not seem adequate to match these empirical findings with theorems on the performance of 2-opt. For instance, it is known that 2-opt only gives an $\Omega(\frac{\log n}{\log \log n})$-approximation for Euclidean TSP in the plane and may terminate after a number of steps exponential in n [7,8], where n is the number of points. A fundamental quest(ion) is thus to find a theoretical explanation for the empirical performance of 2-opt:

Why does local search perform well on TSP in practice?

A first, natural alternative model assumes that the n points are distributed independently and uniformly at random, instead of being given adversarially. Following [28], we call this model *Distributional*. In the distributional model, the performance of 2-opt – and, more generally, optimal solutions to the Euclidean TSP – are well-understood for fixed dimensions d. The expected number of iterations of 2-opt is polynomial in n [8], while its output obtains, with high probability, a constant factor approximation to the optimal tour. This latter fact holds since the value of the solution found by 2-opt on *any* set of n points in the d-dimensional unit cube is $\mathcal{O}(n^{1-1/d})$ [8] and the length of the optimal tour in the distributional model is, with high probability, $\Omega(n^{1-1/d})$ [30]. However, a main limitation of the distributional model is that random instances have a very particular structure. For example, for a random instance of size n in the unit square, any region of constant size $c \in [0, 1]$ contains, with high probability, $cn \pm \varepsilon$ points.

In order to interpolate between worst-case scenarios and distributional models, much research in optimization has been devoted to define and study *semi-random* models. Such models contain both an adversarial and a random component. A classical example is *smoothed analysis*, where all the input data is perturbed by some noise, and the performance of the algorithm is then studied on the perturbed instance. In the Euclidean TSP case, this perturbation is usually achieved by adding to the positions of each point a value sampled from the same Gaussian distribution $N(0, \sigma)$. It is known that, in this model, the expected running time and approximation ratio of 2-opt are polynomial in σ and logarithmic

in $1/\sigma$, respectively [12,22]. Other common semi-random models for discrete optimization problems first generate a random instance, e.g., a graph, and then adversarialy perturbs it, e.g., by adding / removing edges of the graph, or vice versa (see e.g., [28]).

1.1 Our contributions

A new semi-random model. As a step towards answering the motivating question of this paper, we define and study a new semi-random model for Euclidean TSP instances, that we dub *Simultaneous Semi-Random*. In this model, a $1-\alpha$ fraction of the points are chosen by an adversary and an α fraction of the points are uniformly random, for some parameter $\alpha \in [0,1]$. This semi-random model provides an explanation for the approximation performance of algorithms that complements the explanation provided by smoothed analysis. In order to appreciate this complementarity, we distinguish two different levels of the structure of a point set instance in the unit square. Given a parameter $c < 1$, consider a $c^{-1} \times c^{-1}$ grid that partitions the unit square into squares of size $c \times c$ called *local regions*. The *global structure* of an instance is the number of points inside of each local region. The *local structure* of a local region is the positions of the points in that region.

Informally, smoothed instances exhibit an arbitrary global structure and random local structures. In contrast, simultaneous semi-random instances have arbitrary local structures, except for a small random fraction of the local regions. Thus, smoothed analysis explains the performance of local search on instances with specific global structures, e.g., instances where all the points are only in a constant number of local regions. In contrast, our simultaneous semi-random model explains the performance of local search on instances with specific local structures, e.g., points that form perfectly straight lines. In other words, this semi-random model tradeoffs allowing some completely random points in order to capture instances where there is a subset of the points that exhibit a fully arbitrary structure.

Bounds. We show that an $\alpha = 1/\log n$ fraction of random points are sufficient for local search to obtain an $\mathcal{O}(\log\log n)$ approximation ratio in the simultaneous semi-random model in constant dimensions, which improves over the lower bound $\Omega(\log n/\log\log n)$ from worst-case analysis, which holds even in two dimensions [8].

Theorem 1. *For Euclidean TSP in $[0,1]^d$ where d is constant, 2-opt obtains, with probability $1-o(1)$, an $\mathcal{O}(\log\log n)$-approximation ratio in the simultaneous semi-random model, with $\alpha = \frac{1}{\log n}$.*

Theorem 1 is proved in Section 3. This result implies that the hard instances of Euclidean TSP are not robust to the addition of a small number of random points. Combined with smoothed analysis, we get that either a small amount of random noise to all points or a small fraction of completely random points improves the performance of local search. Interestingly, even though the analyses are completely different, the "amount of noise" σ needed for smoothness, and

the fraction of points α needed for the simultaneous model, to improve the approximation to $\mathcal{O}(\log\log n)$ is $1/\log n$ in both cases.

We note that we actually prove a result that is stronger than Theorem 1 in many ways. For instance, one can take $\alpha = \frac{1}{\log^{\delta} n}$ for any constant $\delta > 0$ without changing the approximation ratio. We refer the reader to Section 3 for details. From our proof, it is also easy to see that we obtain the same result in the more challenging model where the adversary may first observe the random points before placing the adversarial points.

Our second main result is that if $\alpha \leq n^{-3/5-\varepsilon}$, then the approximation ratio of local search cannot be improved in the simultaneous semi-random model compared to its worst-case approximation.

Theorem 2. *For Euclidean TSP in* $[0,1]^2$*, 2-opt achieves, with probability* $1 - o(1)$*, an* $\Omega\left(\frac{\log n}{\log\log n}\right)$ *approximation ratio in the simultaneous semi-random model with* $\alpha = n^{-3/5-\varepsilon}$*, for any constant* $\varepsilon > 0$*.*

Theorem 2, which is proved in Section 4, implies that polynomially many random points are required to obtain an approximation that asymptotically improves over the worst-case approximation. We believe that closing the gap between Theorem 1 and Theorem 2 is an intriguing open problem, and in particular resolving whether there is some $\alpha = o(1)$ such that local search obtains a constant approximation. Answering this question could shed further light on the relationship between the simultaneous semi-random model and "real-world" behavior of the local search algorithm. Obtaining bounds on the running time of local search in this model and investigating it in the context of other optimization problems are also interesting paths forward.

1.2 Technical overview

The upper bound. The upper bound consists of two main steps. We first show a new upper bound on the worst-case length of a 2-optimal set of edges over an instance V that gives an $\mathcal{O}(\log \frac{n^{1-1/d}}{\texttt{OPT}_V})$ approximation (here and throughout the paper, an instance V is given by a set of point in the Euclidean space). This bound is useful because it separates adversarial instances V into two regimes. In the first regime, the optimal length of a tour is large ($\texttt{OPT}_V = \Omega(\frac{n^{1-1/d}}{\log n})$) and the approximation of local search on V, even without random points, is $\mathcal{O}(\log\log n)$. In the second regime, $\texttt{OPT}_V$ is small and we get that the optimal tour length $\texttt{OPT}_R$ over the random points R, with $\alpha = 1/\log n$, is such that $\texttt{OPT}_R \geq \texttt{OPT}_V$. We then use our newly proved worst-case bound to analyze the lengths of 2-optimal tours and optimal tours in the simultaneous semi-random model by combining bounds from both worst-case and distributional analysis.

The lower bound. We first present a framework that reduces proving lower bounds in the simultaneous semi-random model to constructing an adversarial instance V and a Hamiltonian path P over V from a point $s \in V$ to a point $t \in V$ that satisfy three parametrized properties: the length $\ell(P)$ of P is such

that $\ell(P) \geq \gamma \texttt{OPT}_V$ (called the γ-*bad property*), $\texttt{OPT}_V \geq L$ (called L-*long*), and finally, $P \cup \{(s,x)\}$, $P \cup \{(x,y)\}$ and $P \cup \{(t,y)\}$ are 2-optimal for some $K \subseteq [0,1]^2$ and any $x, y \in K$ (called K-*resistant*). Existing hard instances of course satisfy the γ-bad property with $\gamma = \Theta(\log n / \log\log n)$, but they do not satisfy the L-long and the K-resistant properties for desirable parameters L and K. These latter two properties cause significant additional challenges.

Our construction starts with the construction from [8] for the $\Omega(\frac{\log n}{\log\log n})$ worst-case lower bound and then consists of three steps that modify it. We first make the construction thinner, so that it fits in $[0,1] \times [0,\varepsilon]$ for some small ε. Then, we stack multiple copies of the thin instance, without incurring any loss in the γ-bad parameter and while keeping the construction relatively thin, to increase the L-long parameter. Finally, the most challenging step is to satisfy the K-resistance property. For that, we carefully add a small number of additional points to the construction so that, with high probability, there is a Hamiltonian path P on the adversarial instance V that can be connected to a 2-optimal Hamiltonian path on the random vertices R to obtain a 2-optimal tour on $V \cup R$.

1.3 Additional related work

Approximation algorithms for Euclidean TSP. For TSP in the plane, Karp [19] showed that a partitioning algorithm that subdivides the points into groups of size t obtains an $\mathcal{O}(\sqrt{n/t})$ approximation, which improves to $\mathcal{O}(t^{-1/2})$ if the points are uniformly random. A seminal result by Arora [3] obtained a PTAS for d-dimensional Euclidean TSP, for any constant d. The approximation ratio of the 2-opt algorithm was recently improved from $\mathcal{O}(\log n)$ [8] to $\mathcal{O}(\frac{\log n}{\log\log n})$ [7] for Euclidean TSP in the plane, which is the best approximation achievable [8]. For additional approximation algorithms results on Euclidean TSP, see e.g. [1,20].

We next discuss three different families of semi-random models.

Semi-random models with a monotone adversary. Seminal work by Blum and Spencer [6] proposed semi-random models for k-coloring. In the colorgame model, edges are first placed at random between pairs of vertices and then an adversary places additional edges. Similar semi-random models where an adversary manipulates randomly generated instances were considered for problems such as minimum bisection and maximum independent set [14].

Smoothed analysis is a semi-random model where random perturbations are applied to an adversarial instance. It was first studied by Spielman and Teng [29] to explain the fast running time of the simplex method in practice. Smoothed analysis of both the running time and approximation of local search (2-opt) for TSP was first studied by [12] who obtained an $O(1/\sigma)$ approximation when Gaussian random variables with mean 0 and standard deviation σ are added to each point. They also obtained more general bounds for any distributions with bounded densities. This approximation was then improved to $O(\log 1/\sigma)$ by [22]. Smoothed analysis of local search has also been studied for general, non-Euclidean, graphs [13] and in the context of clustering [4]. We are not aware of other semi-random models that have been studied for TSP.

Multi-stage semi-random models. More complex semi-random models that generate instances in three or more steps, where some steps are adversarial and the others are made randomly, have also been studied for many problems, including unique games [21], partitioning [23], 3-coloring [11], and clustering mixtures of Gaussians [33].

In contrast to all previous semi-random models, where the randomized and adversarial steps occur sequentially, in the simultaneous semi-random model that we introduce and study in this paper these steps occur simultaneously and independently of each other. We are not aware of previously studied semi-random models that have this property.

2 Preliminaries

In the following, given $n, d \in \mathbb{N}$, an instance of size n and dimension d, or n-instance, is a set of n points in $[0,1]^d$. When the dimension is not mentioned, it is assumed to be 2. For $m \in \mathbb{N}$, the *random instance* $R(m)$ is a set of m points drawn uniformly and independently from $[0,1]^d$. For an instance V, we indifferently call $v \in V$ a *point* or a *vertex*. Given $x \in \mathbb{R}^d$, we let $\|x\|$ be its Euclidean norm. For an edge $e = (v_1, v_2)$, we often write $\|e\| = \|v_1 - v_2\|$. The angle between two edges $e = (v_1, v_2)$ and $e' = (v_1', v_2')$ is the angle between the vectors $u = v_2 - v_1$ and $u' = v_2' - v_1'$, which is equal to $\arccos \frac{u \cdot u'}{\|u\|\|u'\|} \in [0, \pi]$. We let $\sqcup$ denote the disjoint union operator of sets.

Given an instance V and a set of m edges $T = \{(v_{i_1}, v_{i_2}) \mid i \in [m]\}$, the *length* of T is $\ell(T) = \sum_{i=1}^{m} \|v_{i_1} - v_{i_2}\|$. A *tour* on an instance V is a set of $|V|$ edges T that form a cycle. Given an instance V, $\texttt{OPT}_V$ is the length of a tour on V of minimum length. Assume T is an arbitrary collection of edges, a 2-*swap* replaces (v_{i_1}, v_{i_2}) and (v_{j_1}, v_{j_2}) in T with (v_{i_1}, v_{j_1}) and (v_{i_2}, v_{j_2}). We say that T is 2-*optimal* if there is no set T' of strictly smaller length obtained from T via a 2-swap. In particular, when T defines a tour, the concept of 2-optimality coincides with the stopping criterion for 2-opt. We now present some general facts about optimal and 2-optimal TSP tours on Euclidean instances. The first is a bound on the length of any 2-optimal set of edges.

Lemma 1 ([8]). *Let T be a 2-optimal set of n edges on an instance $V \in [0,1]^d$, and assume d to be a constant. Then $\ell(T) = \mathcal{O}\left(n^{1-1/d}\right)$.*

This in particular implies that the optimal tour, up to a constant factor, always has length at most $n^{1-1/d}$. We also know the behavior of $\texttt{OPT}$ on random instances.

Lemma 2 ([30]). *With probability $1 - o(1)$, we have $\texttt{OPT}_{R(n)} = \Theta\left(n^{1-1/d}\right)$.*

From Lemma 1 and Lemma 2, we immediately deduce the following corollary.

Corollary 1. *On a random instance, the approximation ratio of 2-opt is constant with probability $1 - o(1)$.*

Last, we recall the following best-known upper bounds on the performance of 2-opt on general instances.

Lemma 3 ([7,8]). *Let V be an arbitrary d-dimensional instance of size n with d constant. Let $T \subset V^2$ be a 2-optimal set of edges. Then $\ell(T) = \mathcal{O}(\log n)\texttt{OPT}_V$. Moreover, if $d = 2$, then $\ell(T) = \mathcal{O}\left(\frac{\log n}{\log\log n}\right)\texttt{OPT}_V$.*

3 An Improved Approximation for Local Search in the Simultaneous Semi-Random Model

In this section, we show that an $\alpha = 1/\log n$ fraction of random points is sufficient to improve the approximation achieved by local search to $\mathcal{O}(\log\log n)$ in the simultaneous semi-random model.

3.1 An improved worst-case approximation for local search

We first show a new upper bound on the worst case approximation of 2-opt.

Lemma 4. *Let V be an arbitrary d-dimensional instance of size n with d constant. Let $T \subset V^2$ be a 2-optimal set of edges. Then $\ell(T) = \mathcal{O}\left(\texttt{OPT}_V \log \frac{n^{1-1/d}}{\texttt{OPT}_V}\right)$.*

This new bound is helpful because it separates instances into two regimes. The first is when the length of the optimal tour is large, when $\texttt{OPT}_V = \Omega(\frac{n^{1-1/d}}{\log n})$. In this regime, we immediately get from Lemma 4 that $\ell(T) = \mathcal{O}(\log\log n \cdot \texttt{OPT}_V)$ for any 2-optimal set of edges T, so a locally optimal tour performs well on the adversarial instance, without even needing random points.

In the second regime, when $\texttt{OPT}_V = o(\frac{n^{1-1/d}}{\log n})$, we have that for $\alpha = 1/\log n$, the length of the optimal tour on the random points R dominates the length of the optimal tour on the adversarial instance V: $\texttt{OPT}_R = \Theta(|R|^{1-1/d}) = \Theta((\frac{n}{\log n})^{1-1/d}) \geq \texttt{OPT}_V$ where the first equality is by Lemma 2. We later combine the constant approximation obtained by 2-opt on random instances and the fact that $\texttt{OPT}_R \geq \texttt{OPT}_V$ to get that 2-opt obtains a constant approximation on $V \sqcup R$ in that regime. In summary, Lemma 4 is helpful because it shows that it is only when the length of the optimal tour is small that 2-opt performs poorly on adversarial instances. We will show that, in this regime, adding random points to an adversarial instance improves the approximation obtained by 2-opt.

The remainder of Section 3.1 is devoted to the proof of Lemma 4. We first introduce the concepts of similarly-oriented edges and similar-length edges that will be used in the proof.

Similarly-oriented edges. We use the notion of similarly-oriented edges from [8].

Definition 1 ([8]). *Edges e and e' are* similarly-oriented *if the angle between e and e' is at most* $\arctan\frac{1}{4}$.

Edges can be partitioned into a constant number of families of edges such that every pair of edges in a same family are similarly-oriented. For a vector $u \in \mathbb{R}^d$, we denote by T^u the collection of all vectors $u' \in T$ such that the angle between u and u' is at most $\frac{1}{2}\arctan\frac{1}{4}$ and by $\mathbb{S}^{d-1}$ the unit sphere in $\mathbb{R}^d$. Thus, for any $u \in \mathbb{R}^d$, every pair of edges in T^u are similarly-oriented. Using the topological definition of compactness, we know that there exists a constant I and $u_1, \dots u_I \in \mathbb{S}^{d-1}$ such that $T = \cup_{i=1}^{I} T^{u_i}$. Hence, up to a constant loss, it is sufficient to bound the total length of all edges in T^{u_i} for an arbitrary i. For the remainder of this section, we abuse notation and write T^i instead of T^{u_i}. Similarly oriented edges have the following useful property.

Lemma 5 ([8]). *Let $e = (v_1, v_2)$ and $e' = (v_1', v_2')$ be two similarly-oriented edges which form a 2-optimal set. Then $\|v_1' - v_1\| \geq \frac{1}{2}\min(\|e\|, \|e'\|)$.*

Similar-length edges. In addition to being partitioned into families of similarly-oriented edges, edges are also partitioned into families of similar length edges. In particular, let $1 > \eta > \varepsilon > 0$, we define $T_< = \{e \in T \mid \|e\| < \varepsilon\}$, $T_> = \{e \in T \mid \|e\| \geq \eta\}$, and, for any $j \geq 0$ such that $2^j\varepsilon \leq \eta$, $T_j := \{e \in T \mid 2^j\varepsilon \leq \|e\| < 2^{j+1}\varepsilon\}$. Thus writing $J = \lfloor \log_2 \frac{\eta}{\varepsilon} \rfloor$, we have $T = T_< \sqcup T_> \sqcup \bigsqcup_{j=0}^{J} T_j$. The following result is known for long edges.

Lemma 6 ([8]). *For any $\eta > 0$ and constant dimension d, $\ell(T_>) = \mathcal{O}(\eta^{1-d})$.*

We denote families of edges that are both similarly-oriented and of similar-length by $T_j^i = T^i \cap T_j$. We similarly denote $T_<^i = T^i \cap T_<$ and $T_>^i = T^i \cap T_>$. Now we are ready to prove Lemma 4.

Proof (of Lemma 4). Let $i \in [I]$ and consider the family $T^i \subset T$ of similarly-oriented edges. First, we have $\ell(T_<^i) \leq n\varepsilon$. Second, by Lemma 6, we have $\ell(T_>^i) \leq \mathcal{O}(\eta^{1-d})$. To bound $\ell(T^i) = \ell(T_<^i) + \ell(T_>^i) + \sum_{j=0}^{J} \ell(T_j^i)$, it remains to bound the length of the family of similarly-oriented and similar length edges T_j^i for an arbitrary $j \in [J]$.

Let T^* be an optimal tour on V: if we fix any point to be the first one, T^* defines an order on V, and we can use it to order T_j^i by saying that $(v_1, v_2) < (v_1', v_2')$ if $v_1 < v_1'$ in T^*. Hence, for fixed i and j, we can enumerate $T_j^i = \{e^l = (v_1^l, v_2^l) \mid 1 \leq l \leq N\}$ (where $N = |T_j^i|$), such that v_1^l appears before v_1^{l+1} in T^*.

Let $\widetilde{T}^* = \{(v_1^l, v_1^{l+1}) \mid 1 \leq l \leq N\}$ (where we let $v_1^{N+1} = v_1^1$). By the triangular inequality, we have $\ell(\widetilde{T}^*) \leq \ell(T^*)$. Moreover, for any l , since $T_j^i = \{e^l = (v_1^l, v_2^l) \mid 1 \leq l \leq N\} \subseteq T$ and is therefore 2-optimal, we have by Lemma 5 $||v_1^l - v_1^{l+1}|| \geq \frac{1}{2}\min\left(||e^l||, ||e^{l+1}||\right)$. As $e^l, e^{l+1} \in T_j^i$, and the length of every vector in T_j^i is between $2^j\varepsilon$ and $2^{j+1}\varepsilon$, the longest of $||e^l||, ||e^{l+1}||$ is at most two times larger than the shortest. Thus, we get $||v_1^l - v_1^{l+1}|| \geq \frac{1}{4}||e^l||$. Summing for $l \in [N]$, we obtain $\ell(T_j^i) = \sum_{l=1}^{N} ||e^l|| \leq 4\sum_{l=1}^{N} ||v_1^l - v_1^{l+1}|| = 4\ell(\widetilde{T}^*) \leq 4\ell(T^*)$. Hence, putting the bounds on $\ell(T_<^i)$, $\ell(T_>^i)$, $\sum_{j=0}^{J} \ell(T_j^i)$ together, we have

$$\ell(T^i) \leq n\varepsilon + \mathcal{O}(\eta^{1-d}) + \lfloor \log_2 \frac{\eta}{\varepsilon} \rfloor 4\texttt{OPT} = \mathcal{O}\left(n\varepsilon + \eta^{1-d} + \texttt{OPT}\log\frac{\eta}{\varepsilon}\right).$$

Summing over all families of similarly-oriented edges and letting $\varepsilon = \frac{\mathtt{OPT}}{n}$ and $\eta = \mathtt{OPT}^{\frac{1}{1-d}}$, we get

$$\ell(T) \leq \sum_{i=1}^{I} \ell(T^i) = \mathcal{O}\left(n\varepsilon + \eta^{1-d} + \mathtt{OPT}\log\frac{\eta}{\varepsilon}\right) = \mathcal{O}\left(\mathtt{OPT}\log\frac{n^{1-\frac{1}{d}}}{\mathtt{OPT}}\right). \qquad \square$$

3.2 Proof of Theorem 1

By combining the new worst-case bound from Section 3.1 together with the bound in Lemma 1, one obtains the following upper bound on the length of a 2-optimal tour on the union of two instances.

Lemma 7. *Let V and U be disjoint instances of sizes n and m. Then, if d is constant, for any 2-optimal tour T on $V \sqcup U$, $\ell(T) = \mathcal{O}(\mathtt{OPT_V}\log\frac{n^{1-\frac{1}{d}}}{\mathtt{OPT_V}} + m^{1-\frac{1}{d}})$.*

The last lemma needed is a bound on the optimal length of a tour on an instance of the simultaneous semi-random model.

Lemma 8. *Let $n, m \in \mathbb{N}$ and V be a d-dimensional n-instance with d constant. With probability $1 - o(1)$, $\mathtt{OPT}_{V\sqcup R(m)} = \Omega\left(\max\left(\mathtt{OPT}_V, m^{1-\frac{1}{d}}\right)\right)$.*

We are now ready to prove the main result for this section, from which Theorem 1 follows.

Theorem 3. *Let V be any d-dimensional n-instance, with d a fixed constant. With probability $1-o(1)$, the approximation ratio of 2-opt on $V \sqcup R(m)$ is $\mathcal{O}(1)$ if $m^{1-\frac{1}{d}} > \mathtt{OPT}_V \log\frac{n^{1-\frac{1}{d}}}{\mathtt{OPT}_V}$ and $\mathcal{O}(\log\frac{n^{1-\frac{1}{d}}}{\mathtt{OPT}_V})$ otherwise. In particular, for $m = \frac{n}{\log^c n}$, for any constant $c > 0$, the approximation ratio is $\mathcal{O}(\log\log n)$.*

Proof. Let T be any 2-optimal tour on $V \sqcup R(m)$. If $m^{1-\frac{1}{d}} > \mathtt{OPT}_V \log\frac{n^{1-\frac{1}{d}}}{\mathtt{OPT}_V}$, by Lemma 7 we obtain that $\ell(T) = \mathcal{O}(m^{1-\frac{1}{d}})$ while by Lemma 8 we have $\mathtt{OPT}_{V\sqcup R(m)} = \Omega(m^{1-\frac{1}{d}})$. Hence, the approximation ratio of 2-opt on this instance is $\mathcal{O}(1)$. However, when $m^{1-\frac{1}{d}} \leq \mathtt{OPT}_V \log\frac{n^{1-\frac{1}{d}}}{\mathtt{OPT}_V}$ Lemma 7 tells us that $\ell(T) = \mathcal{O}(\mathtt{OPT}_V \log\frac{n^{1-\frac{1}{d}}}{\mathtt{OPT}_V}) = \mathcal{O}(\mathtt{OPT}_{V\sqcup R(m)} \log\frac{n^{1-\frac{1}{d}}}{\mathtt{OPT}_V})$, hence we can deduce that the approximation ratio is $\mathcal{O}(\log\frac{n^{1-\frac{1}{d}}}{\mathtt{OPT}_V})$.

Now take $m = \frac{n}{\log^c n}$, for any $c > 0$. For $\mathtt{OPT}_V = \mathcal{O}((\frac{n}{\log^c n})^{1-\frac{1}{d}}/\log n)$, the first regime applies and 2-opt gives a constant approximation. Else, using the second regime, we obtain a $\mathcal{O}(\log\log n)$-approximation. $\square$

4 Improved Approximations Require poly(n) Random Points

In this section, we complement the upper bound from the previous section by showing that, with $\alpha = n^{-3/5+\varepsilon}$ for any constant $\varepsilon > 0$, local search obtains an

$\Omega(\frac{\log n}{\log \log n})$ approximation ratio in the simultaneous semi-random model when $d = 2$. This lower bound implies that more than $n^{2/5-\varepsilon}$ random points are required to obtain an asymptotic improvement over the $\mathcal{O}(\frac{\log n}{\log \log n})$ worst-case approximation in the simultaneous semi-random model.

To show this lower bound, we construct a 2-optimal tour T over an adversarial instance V that is far from optimal and is such that, with high probability, T can be augmented to obtain a 2-optimal tour $T_{V \sqcup R}$ over $V \sqcup R$ that contains T, where R consists of m random points. The length of the optimal tour on $V \sqcup R$ is upper bounded by combining the lengths of the optimal tours on both V and R. We first develop a framework for proving lower bounds in the simultaneous semi-random model, see Section 4.1. We then sketch the construction of the bad instance, which builds upon the construction from [8], in Section 4.2, and then we analyze it using the framework from Section 4.1. Complete details can be found in the full version of the paper.

4.1 A framework for simultaneous semi-random lower bounds

In this section, we define parametrized properties of an instance that, if satisfied, guarantee a lower bound. In other words, this section reduces the problem of showing a lower bound to constructing an instance that satisfies the following properties. A path of an instance V is called *Hamiltonian* if it passes exactly once through each vertex of V.

Definition 2. *An instance V and a 2-optimal Hamiltonian path P over V from $s \in V$ to $t \in V$ are*
- *L-**long** if $\texttt{OPT}_V \geq L$;*
- *γ-**bad** if $\ell(P) \geq \gamma \texttt{OPT}_V$;*
- *K-**resistant**, $K \subseteq [0,1]^2$, if for any $x, y \in K$, $P \cup \{(s,x)\}$, $P \cup \{(x,y)\}$ and $P \cup \{(t,y)\}$ are 2-optimal.*

We now give some intuition on why we care about the above properties. If K is sufficiently big and α sufficiently small, then, with high probability, the αn random points R all lie in K. Combined with the K-resistance condition, this implies that a 2-optimal Hamiltonian path P over V can be extended to obtain a 2-optimal tour over $V \sqcup R$. More precisely, we have the following lemma.

Lemma 9. *Let V be an instance and $K \subseteq [0,1]^2$ be any region. If V has a Hamiltonian path P that is K-resistant, then for any instance $U \subseteq K$ disjoint from V, there exists a 2-optimal tour T on $U \sqcup V$ which extends P, i.e., such that P is a subpath of T.*

The γ-bad condition guarantees that there is a bad 2-opt tour on V, which we care about because the tour we want to expand must have a bad approximation ratio for the lower bound to be effective. The L-long condition guarantees that the length of the part of the optimal tour on $V \sqcup R$ that connects vertices in V dominates the length of the part that connects the random vertices R, which is important since a 2-opt tour on random vertices R performs well compared to OPT_R. The intuition discussed above is summarized in the next lemma.

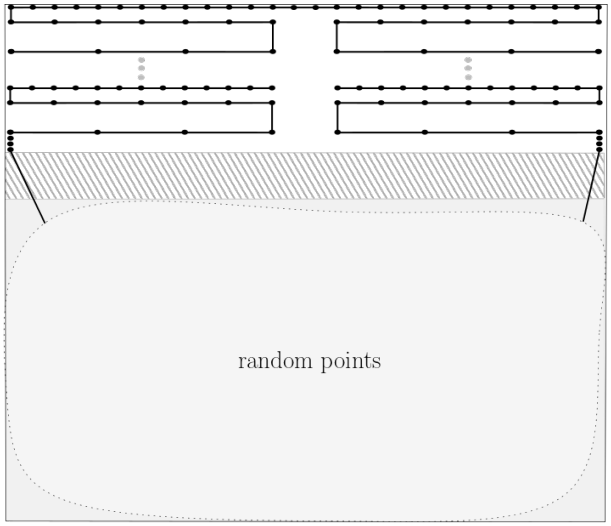

Fig. 1. High-level illustration of the construction from Section 4 (image not to scale). The gray area K accounts for most of the area of the unit square, hence w.h.p. all random points v are such that $v \in K$. The adversarial construction is in the top part of the square. A 2-optimal tour is given by the bold path P (which is a 2-optimal Hamiltonian path on the adversarial instance and far from optimal), plus an optimal Hamiltonian path covering the random points. The striped area is w.h.p. empty, and serves as a buffer between the deterministic and the random points, in order to ensure K-resistance of P in the adversarial instance.

Lemma 10. *Let $\alpha \in (0,1)$ be some parameter that can depend on n. If there exists a $\sqrt{\alpha n}$-long n-instance V with a γ-bad, K-resistant Hamiltonian path P, for some region K with area $1 - o(1/(\alpha n))$, then with probability $1 - o(1)$ the approximation ratio of 2-opt on the α-semi-random instance $R(\alpha n) \sqcup V$ is $\Omega(\gamma)$.*

4.2 The construction

In this section, we give a sketch of the construction of an instance that satisfies Lemma 10. The full construction can be found in the full version of the paper. Our starting point is the construction of [8], which does not satisfy the γ-bad and K-resistant properties for desirable parameters γ and K. We gradually modify and extend it so that it acquires the desired properties:

1. first, we modify it by making it fit in a small region and having a "bad" Hamiltonian path that only consists of "short" edges. Hence, this ensures a property useful for the separation condition, i.e., the adversarial instance is w.h.p. separated from the random points.
2. then, we modify it again to ensure the longness condition by stacking and connecting multiple copies of the thin instance;
3. finally, for the resistance condition, we add a small number of additional points so that the bad Hamiltonian path P over the adversarial instance V

can be extended to a 2-optimal tour $T_{V \sqcup R}$ over the semi-random instance $V \sqcup R$ with random points R such that P is a subpath of $T_{V \sqcup R}$.

A high-level sketch of the construction is given in Figure 1. The properties of the resulting instance V are summarized in the following lemmas. We let $\varepsilon > 0$ be a rational number such $0 < \varepsilon < \frac{1}{4}$ and $p \geq 3$ such that $p/4$ and εp are integers. We also assume that εp is odd. Since there are infinitely many of such p, all limits are understood as when p goes to infinity. Let $z = p/4$ and $s = (1-\varepsilon)p$.

Lemma 11. *V is included in $S = [0,1] \times [1 - p^{2(z-s)} - 2p^{-s}, 1]$ and has $n = \Theta(p^{2(z+p)}) = \Theta(p^{\frac{5}{2}p})$ points. In particular, we have $p = \Theta\left(\frac{\log n}{\log\log n}\right)$.*

Lemma 12. *The optimal tour on V has length $\Theta(p^{2z}) = \Theta(n^{1/5})$.*

Lemma 13. *There exists a 2-optimal Hamiltonian path P on the adversarial instance V of length $\ell(P) \geq \frac{2}{3}\varepsilon p^{2z+1} = \Theta\left(\varepsilon \frac{\log n}{\log\log n} n^{1/5}\right)$.*

Let $K = [0,1] \times [0, 1 - p^{2(z-s)} - 4p^{-s}]$. The following is the main technical lemma of this section. Its proof requires a careful geometric analysis of V.

Lemma 14. *The instance V and the path P are K-resistant.*

Combining the previous lemmas, we get the following.

Lemma 15. *The n-instance V is $n^{1/5}$-long, and the Hamiltonian path P is $\Omega\left(\varepsilon \frac{\log n}{\log\log n}\right)$-bad and K-resistant, for some region K of area $1 - o(n^{-(2/5-\varepsilon)})$.*

Proof. By Lemma 11, Lemma 12, and Lemma 13, V with P are $n^{1/5}$-long and $\Omega\left(\varepsilon \frac{\log n}{\log\log n}\right)$-bad. Moreover, for $K = [0,1] \times [0, 1 - p^{2(z-s)} - 4p^{-s}]$, V with P are K-resistant by Lemma 14. Note that $p^{2(z-s)} = p^{p/2-(1-\varepsilon)p} p^{-s}$, and since $\varepsilon < 1/4$, $p^{2(z-s)} = o(p^{-s})$. Finally, $p^{-s} = \Theta(n^{-\frac{2}{5}(1-\varepsilon)}) = o(n^{-(\frac{2}{5}-\varepsilon)})$, thus K has area $1 - o(n^{-(\frac{2}{5}-\varepsilon)})$. □

The previous lemma combined with Lemma 10 immediately give us the proof of Theorem 2. Indeed, let $1/4 > \varepsilon > 0$ be constant. Let n be a number and V an instance as above (recall that there are infinitely many of them). By Lemma 15, the instance V verifies the hypothesis of Lemma 10 with $\alpha = n^{-3/5-\varepsilon}$ and $\gamma = \Omega\left(\frac{\log n}{\log\log n}\right)$; thus by Lemma 10 the approximation ratio of 2-opt on V is $\Omega\left(\frac{\log n}{\log\log n}\right)$ with probability $1 - o(1)$.

References

1. Antoniadis, A., Fleszar, K., Hoeksma, R., Schewior, K.: A ptas for euclidean tsp with hyperplane neighborhoods. In: SODA. pp. 1089–1105. SIAM (2019)
2. Applegate, D.L., Bixby, R.E., Chvátal, V., Cook, W.J.: The traveling salesman problem. Princeton university press (2011)
3. Arora, S.: Polynomial time approximation schemes for euclidean traveling salesman and other geometric problems. Journal of the ACM (JACM) **45**(5), 753–782 (1998)
4. Arthur, D., Vassilvitskii, S.: Worst-case and smoothed analysis of the icp algorithm, with an application to the k-means method. In: FOCS. pp. 153–164. IEEE (2006)
5. Asadpour, A., Goemans, M.X., Madry, A., Gharan, S.O., Saberi, A.: An $O(\log n/\log\log n)$-approximation algorithm for the asymmetric traveling salesman problem. Operations Research **65**(4), 1043–1061 (2017)
6. Blum, A., Spencer, J.: Coloring random and semi-random k-colorable graphs. Journal of Algorithms **19**(2), 204–234 (1995)
7. Brodowsky, U.A., Hougardy, S.: The approximation ratio of the 2-opt heuristic for the euclidean traveling salesman problem. arXiv preprint arXiv:2010.02583 (2020)
8. Chandra, B., Karloff, H., Tovey, C.: New results on the old k-opt algorithm for the traveling salesman problem. SIAM Journal on Computing **28**(6), 1998–2029 (1999)
9. Christofides, N.: Worst-case analysis of a new heuristic for the travelling salesman problem. Tech. rep., Carnegie-Mellon Univ Pittsburgh Pa Management Sciences Research Group (1976)
10. Dantzig, G., Fulkerson, R., Johnson, S.: Solution of a large-scale traveling-salesman problem. Journal of the operations research society of America **2**(4), 393–410 (1954)
11. David, R., Feige, U.: On the effect of randomness on planted 3-coloring models. In: Proceedings of the forty-eighth annual ACM symposium on Theory of Computing. pp. 77–90 (2016)
12. Englert, M., Röglin, H., Vöcking, B.: Worst case and probabilistic analysis of the 2-opt algorithm for the TSP. Algorithmica **68**(1), 190–264 (2014)
13. Englert, M., Röglin, H., Vöcking, B.: Smoothed analysis of the 2-opt algorithm for the general tsp. TALG **13**(1), 1–15 (2016)
14. Feige, U., Kilian, J.: Heuristics for semirandom graph problems. Journal of Computer and System Sciences **63**(4), 639–671 (2001)
15. Garey, M.R., Graham, R.L., Johnson, D.S.: Some np-complete geometric problems. In: Proceedings of the eighth annual ACM symposium on Theory of computing. pp. 10–22 (1976)
16. Held, M., Karp, R.M.: A dynamic programming approach to sequencing problems. Journal of the Society for Industrial and Applied mathematics **10**(1), 196–210 (1962)
17. Johnson, D.S., McGeoch, L.A.: 8. the traveling salesman problem: a case study. In: Local search in combinatorial optimization, pp. 215–310. Princeton University Press (2018)
18. Karlin, A.R., Klein, N., Gharan, S.O.: A (slightly) improved approximation algorithm for metric tsp. In: Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing. pp. 32–45 (2021)
19. Karp, R.M.: Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane. Mathematics of operations research **2**(3), 209–224 (1977)

20. Klein, P.N.: A linear-time approximation scheme for tsp in undirected planar graphs with edge-weights. SIAM Journal on Computing **37**(6), 1926–1952 (2008)
21. Kolla, A., Makarychev, K., Makarychev, Y.: How to play unique games against a semi-random adversary: Study of semi-random models of unique games. In: 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science. pp. 443–452. IEEE (2011)
22. Künnemann, M., Manthey, B.: Towards understanding the smoothed approximation ratio of the 2-opt heuristic. In: International Colloquium on Automata, Languages, and Programming. pp. 859–871. Springer (2015)
23. Makarychev, K., Makarychev, Y., Vijayaraghavan, A.: Approximation algorithms for semi-random partitioning problems. In: STOC. pp. 367–384 (2012)
24. Mömke, T., Svensson, O.: Removing and adding edges for the traveling salesman problem. Journal of the ACM (JACM) **63**(1), 1–28 (2016)
25. Padberg, M., Rinaldi, G.: A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems. SIAM review **33**(1), 60–100 (1991)
26. Papadimitriou, C.H.: The euclidean travelling salesman problem is np-complete. Theoretical computer science **4**(3), 237–244 (1977)
27. Reinelt, G.: TSPLIB—a traveling salesman problem library. ORSA journal on computing **3**(4), 376–384 (1991)
28. Roughgarden, T.: Beyond the Worst-Case Analysis of Algorithms. Cambridge University Press (2021)
29. Spielman, D.A., Teng, S.H.: Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. Journal of the ACM (JACM) **51**(3), 385–463 (2004)
30. Steele, J.M.: Probability theory and combinatorial optimization. SIAM (1997)
31. Svensson, O., Tarnawski, J., Végh, L.A.: A constant-factor approximation algorithm for the asymmetric traveling salesman problem. Journal of the ACM (JACM) **67**(6), 1–53 (2020)
32. Traub, V., Vygen, J.: Approaching 3/2 for the s-t-path tsp. Journal of the ACM (JACM) **66**(2), 1–17 (2019)
33. Vijayaraghavan, A., Awasthi, P.: Clustering semi-random mixtures of gaussians. In: ICML. pp. 5055–5064. PMLR (2018)