Electronic Colloquium on Computational Complexity, Report No. 39 (2007)

Smoothed Analysis of Binary Search Trees and Quicksort Under Additive Noise

Bodo Manthey[*]
Yale University
Department of Computer Science
P. O. Box 208 285
New Haven, CT 06520-8285
manthey@cs.yale.edu

Till Tantau
Universität zu Lübeck
Institut für Theoretische Informatik
Ratzeburger Allee 160
23538 Lübeck, Germany
tantau@tcs.uni-luebeck.de

March 27, 2007

Abstract

Binary search trees are a fundamental data structure and their height plays a key role in the analysis of divide-and-conquer algorithms like quicksort. Their worst-case height is linear; their average height, whose exact value is one of the best-studied problems in average-case complexity, is logarithmic. We analyze their smoothed height under additive noise: An adversary chooses a sequence of n real numbers in the range $[0, 1]$; each number is individually perturbed by adding a random value from an interval of size d; and the resulting numbers are inserted into a search tree. The expected height of this tree is called smoothed tree height. If d is very small, namely for $d \leq 1/n$, the smoothed tree height is the same as the worst-case height; if d is very large, the smoothed tree height approaches the logarithmic average-case height. An analysis of what happens between these extremes lies at the heart of our paper: We prove that the smoothed height of binary search trees is $\Theta(\sqrt{n/d} + \log n)$, where $d \geq 1/n$ may depend on n. This implies that the logarithmic average-case height becomes manifest only for $d \in \Omega(n/\log^2 n)$. For the analysis, we first prove that the smoothed number of left-to-right maxima in a sequence is also $\Theta(\sqrt{n/d} + \log n)$. We apply these findings to the performance of the quicksort algorithm, which needs $\Theta(n^2)$ comparisons in the worst case and $\Theta(n \log n)$ on average, and prove that the smoothed number of comparisons made by quicksort is $\Theta\big(\frac{n}{d+1}\sqrt{n/d} + n \log n\big)$. This implies that the average-case becomes manifest already for $d \in \Omega\big(\sqrt[3]{n/\log^2 n}\big)$.

Keywords: Smoothed analysis, binary search trees, quicksort.

1 Introduction

To explain the discrepancy between average-case and worst-case behavior of the simplex algorithm, Spielman and Teng introduced the notion of *smoothed analysis* [11]. Smoothed analysis interpolates between average-case and worst-case analysis: Instead of taking a worst-case instance or, as in average-case analysis, choosing an instance completely at random, we analyze the complexity of

[*] Supported by the Postdoc-Program of the German Academic Exchange Service (DAAD). On leave from Saarland University, Department of Computer Science, Postfach 151150, 66041 Saarbrücken, Germany.

ISSN 1433-8092

(possibly worst-case) objects subject to slight random perturbations. On the one hand, perturbations model that nature is not (or not always) adversarial. On the other hand, perturbations reflect the fact that data is often subject to measurement or rounding errors; even if the instance at hand was initially a worst-case instance, due to measurement and rounding errors we would probably get a less difficult instance in practice. Spielman and Teng [12] give a comprehensive survey on results and open problems in smoothed analysis.

Binary search trees are one of the most fundamental data structures in computer science and they are the building blocks for a large variety of data structures. The most important parameter of binary search trees is their *height* since this parameter heavily influences the performance of algorithms that use binary search trees. The worst-case height of a binary tree for n numbers is n. The average-case behavior has been the subject of a considerable amount of research, culminating in the result that the average-case height is $\alpha \ln n + \beta \ln \ln n + O(1)$, where $\alpha \approx 4.311$ is the larger root of $\alpha \ln(2e/\alpha) = 1$ and $\beta = 3/(2\ln(\alpha/2)) \approx 1.953$ [8]. Furthermore, the variance of the height is constant, as was proved independently by Drmota [2] and Reed [8], and it is conjectured that all moments are bounded by constants as well [9]. Drmota [3] gives a recent survey.

Beyond being an important data structure, binary search trees play a central role in the analysis of divide-and-conquer algorithms like quicksort [5, Section 5.2.2]. While quicksort needs $\Theta(n^2)$ comparisons in the worst case, the average number of comparisons is $2n \log n - \Theta(n)$ with a variance of $(7 - \frac{2}{3}\pi^2) \cdot n^2 - 2n \log n + O(n)$ as mentioned by Fill and Janson [4]. The relationship of quicksort to binary search tree height is the following: The height of the tree $T(\sigma)$ obtained from a sequence σ is equal to the number of levels of recursion required by quicksort to sort σ. The number of comparisons, which corresponds to the total path length of $T(\sigma)$, is at most n times the height of $T(\sigma)$.

Binary search trees are also related to the number of left-to-right maxima of a sequence, which is the number of new maxima seen while scanning a sequence from left to right. The number of left-to-right maxima of σ is equal to the length of the rightmost path of the tree $T(\sigma)$, which means that left-to-right maxima provide an easy-to-analyze lower bound for the the height of binary search trees. Furthermore, left-to-right maxima play a role in the analysis of quicksort [10]. In the worst-case, the number of left-to-right maxima is n, while it is $\sum_{i=1}^{n} 1/i \in \Theta(\log n)$ on average.

Given the discrepancies between average-case and worst-case behavior of binary search trees, quicksort, and the number of left-to-right maxima, the question arises of what happens in between when the randomness is limited.

Previously Studied Perturbation Models. The perturbation model introduced by Spielman and Teng for the smoothed analysis of continuous problems like linear programming is appropriate for algorithms that process real numbers. In their model, each of the real numbers in the adversarial input is perturbed by adding a small Gaussian noise. This model of perturbation favors instances in the neighborhood of the adversarial input for a fairly natural and realistic notion of "neighborhood."

The first smoothed analysis of quicksort, due to Banderier, Beier, and Mehlhorn [1], uses a different perturbation model, namely a *discrete perturbation model*. Such models take discrete objects like permutations as input and again yield discrete objects like another permutation. Banderier, Beier, and Mehlhorn used *p-partial permutations*, which work as follows: An adversary chooses a permutation of the numbers $\{1, \dots, n\}$ as sequence, every element of the sequence is marked independently with a probability of p, and then the marked elements are randomly permuted. Banderier et al. showed that the number of comparisons subject to p-partial permutations is $O\big(\frac{n}{p} \cdot \log n\big)$. Furthermore, they proved bounds on the smoothed number of left-to-right maxima subject to this model.

Manthey and Reischuk [6] analyzed the height of binary search trees under p-partial permutations. They proved a lower bound of $0.8 \cdot (1-p) \cdot \sqrt{n/p}$ and an asymptotically matching upper bound of $6.7 \cdot (1-p) \cdot \sqrt{n/p}$ for smoothed tree height. For the number of left-to-right maxima they showed that it lies between $0.6 \cdot (1-p) \cdot \sqrt{n/p}$ and $3.6 \cdot (1-p) \cdot \sqrt{n/p}$.

Special care must be taken when defining perturbation models for discrete inputs: The perturbation should favor instances in the neighborhood of the adversarial instance, which requires a suitable definition of neighborhood in the first place, and the perturbation should preserve the global structure of the adversarial instance. Partial permutations have the first feature [6, Lemma 3.2], but destroy much of the global order of the adversarial sequence.

Our Perturbation Model and Our Results. In the present paper we continue the smoothed analysis of binary search trees and quicksort begun by Banderier et al. [1] and Manthey and Reischuk [6]. However, we return to the original idea of smoothed analysis, namely that input numbers are perturbed by measurement errors. In our model the adversarial input sequence consists of n real numbers in the interval $[0, 1]$, rather than a permutation of the numbers $\{1, \ldots, n\}$. Then, each of the real numbers is individually perturbed by adding a random number drawn uniformly from an interval of size d. If we perturb sequences by a d less than $1/n$, then the sorted sequence $(1/n, 2/n, 3/n, \ldots, n/n)$ stays a sorted sequence. This means that for $d < 1/n$ the smoothed height of binary search trees (as well as the performance of quicksort and the number of left-to-right maxima) is the same as in the worst-case. For this reason, we always assume $d \geq 1/n$ in the following.

For the additive noise model, we study the smoothed height of binary search trees, the smoothed number of comparisons made by quicksort, and the smoothed number of left-to-right maxima. In each case we prove tight upper and lower bounds:

1. The smoothed number of left-to-right maxima is $\Theta(\sqrt{n/d} + \log n)$ as shown in Section 3. This result will be exploited in the subsequent sections.

2. The smoothed height of binary search trees is $\Theta(\sqrt{n/d} + \log n)$ as shown in Section 4.

3. The smoothed number of comparisons made by quicksort is $\Theta\big(\frac{n}{d+1}\sqrt{n/d} + n \log n\big)$ as shown in Section 5. Thus, the perturbation effect of non-constant d is stronger than for constant d.

We believe that the additive noise model is more realistic than the previously studied p-partial permutations. For instance, when we need to sort data obtained from, say, temperature measurements or from stock prices, the input will not be distributed according to a discrete model like p-partial permutations.

2 Preliminaries

Intervals of the real axis are denoted by $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$. To denote an interval that does not include an endpoint, we replace the square bracket next to the endpoint by a parenthesis. We call $\sigma = (\sigma_1, \ldots, \sigma_n)$ with $\sigma_i \in \mathbb{R}$ a *sequence*. For $U = \{i_1, \ldots, i_\ell\} \subseteq \{1, \ldots, n\}$ with $i_1 < i_2 < \cdots < i_\ell$ let $\sigma_U = (\sigma_{i_1}, \sigma_{i_2}, \ldots, \sigma_{i_\ell})$ denote the *subsequence* of σ of the elements at positions in U.

We denote probabilities by $\mathbb{P}$ and expected values by $\mathbb{E}$. To bound large deviations from the expected value, we will use the Chernoff bound [7, Sect. 4.1] a couple of times: Let $X_1, \ldots, X_n$ be random variables with $\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0)$. Let $X = \sum_{i=1}^{n} X_i$. Then $\mathbb{E}(X) = pn$ and, for every $\delta > 0$, we have

$$\mathbb{P}\big(X > (1+\delta) \cdot pn\big) < \left(\frac{\exp(\delta)}{(1+\delta)^{1+\delta}}\right)^{pn} .$$

Throughout the paper, we will assume for the sake of clarity that numbers like $\sqrt{d}$ are integral and we do not write down the tedious floor and ceiling functions that are actually necessary. Since we are interested in asymptotic bounds, this does not affect the validity of the proofs.

2.1 Binary Search Trees, Left-To-Right Maxima, and Quicksort

Let σ be a sequence of length n consisting of pairwise distinct elements. For the following definitions, let $G = \{i \in \{1, \ldots, n\} \mid \sigma_i > \sigma_1\}$ be the set of positions of elements greater than σ_1, and let $S = \{i \in \{1, \ldots, n\} \mid \sigma_i < \sigma_1\}$ be the set of positions of elements smaller than σ_1.

From σ, we obtain a *binary search tree* $T(\sigma)$ by iteratively inserting the elements $\sigma_1, \sigma_2, \ldots, \sigma_n$ into the initially empty tree as follows: The root of $T(\sigma)$ is σ_1. The left subtree of the root σ_1 is $T(\sigma_S)$, and the right subtree of σ_1 is $T(\sigma_G)$. The *height of* $T(\sigma)$ is the maximum number of nodes on any root-to-leaf path of $T(\sigma)$: Let $\text{height}(\sigma) = 1 + \max\{\text{height}(\sigma_S) + \text{height}(\sigma_G)\}$, and let $\text{height}(\sigma) = 0$ when σ is the empty sequence.

The number of *left-to-right maxima* of σ is the number of maxima seen when scanning σ from left to right: let $\text{ltrm}(\sigma) = 1 + \text{ltrm}(\sigma_G)$, and let $\text{ltrm}(\sigma) = 0$ when σ is the empty sequence. The number of left-to-right maxima of σ is equal to the length of the rightmost path of $T(\sigma)$. Thus, $\text{ltrm}(\sigma) \leq \text{height}(\sigma)$, which follows also immediately from the definition.

Quicksort is the following sorting algorithm: Given σ, we construct σ_S and σ_G. To do this, all elements of $(\sigma_2, \ldots, \sigma_n)$ have to be compared to σ_1, which is called the *pivot element*. Then we sort σ_S and σ_G recursively to obtain τ_S and τ_G, respectively. Finally, we output $\tau = (\tau_S, \sigma_1, \tau_G)$. The number of comparisons $\text{qs}(\sigma)$ needed to sort σ is thus $\text{qs}(\sigma) = (n-1) + \text{qs}(\sigma_S) + \text{qs}(\sigma_G)$ if σ has a length of $n \geq 1$, and $\text{qs}(\sigma) = 0$ when σ is the empty sequence.

2.2 Perturbation Model

The perturbation model of *additive noise* is defined as follows: Let $d = d(n) \geq 0$ be the perturbation parameter (d may depend on n). Given a sequence σ of n numbers chosen by an adversary from the interval $[0, 1]$, we draw a *noise* ν_i for each $i \in \{1, \ldots, n\}$ uniformly at random from the interval $[0, d]$. Then we obtain the perturbed sequence $\overline{\sigma} = (\overline{\sigma}_1, \ldots, \overline{\sigma}_i)$ by adding ν_i to σ_i, that is, $\overline{\sigma}_i = \sigma_i + \nu_i$. Note that $\overline{\sigma}_i$ need no longer be an element of $[0, 1]$, but $\overline{\sigma}_i \in [0, d+1]$. For $d > 0$ all elements of $\overline{\sigma}$ are distinct with a probability of one. The choice of the interval sizes is arbitrary since the model is invariant under scaling if we scale the perturbation parameter accordingly: If the adversary may draw n numbers from the interval $[a, b]$ and the noise is uniformly distributed in the interval $[c, d]$, we get exactly the same results when the adversary must choose the numbers from the interval $[0, 1]$ and the noise is from the interval $[0, (d-c)/(b-a)]$.

For this model, we define the functions $\text{height}_d(\sigma)$, $\text{qs}_d(\sigma)$, and $\text{ltrm}_d(\sigma)$, which denote the smoothed search tree height, smoothed number of quicksort comparisons, and smoothed number of left-to-right maxima, respectively, when σ is perturbed by d-noise. Since the adversary chooses σ, our goal are bounds for $\max_{\sigma\in[0,1]^n} \mathbb{E}\bigl(\text{height}_d(\sigma)\bigr)$, $\max_{\sigma\in[0,1]^n} \mathbb{E}\bigl(\text{qs}_d(\sigma)\bigr)$, as well as $\max_{\sigma\in[0,1]^n} \mathbb{E}\bigl(\text{ltrm}_d(\sigma)\bigr)$.

As argued earlier, if $d < 1/n$, the adversary can specify $\sigma = (1/n, 2/n, 3/n, \ldots, n/n)$ and adding the noise terms does not affect the order of the elements. This means that we get the worst-case height, number of comparisons, and number of left-to-right maxima. Because of this observation we will restrict our attention to $d \geq 1/n$.

If d is large, the noise will swamp out the original instance, and the order of the elements of $\overline{\sigma}$ will depend only on the noise rather than the original instance. For intermediate d, additive noise interpolates between average and worst case.

3 Smoothed Number of Left-To-Right Maxima

We start our analyses with the smoothed number of left-to-right maxima, which provides us with a lower bound on the height of binary search trees as well. Our aim for the present section is to prove the following theorem.

Theorem 3.1. *For $d \geq 1/n$, we have*

$$\max_{\sigma\in[0,1]^n} \mathbb{E}\bigl(\text{ltrm}_d(\sigma)\bigr) \in \Theta\bigl(\sqrt{n/d} + \log n\bigr).$$

The theorem is proved in the rest of this section by first proving an upper bound of $O\bigl(\sqrt{n/d} + \log n\bigr)$ and then proving a lower bound of $\Omega\bigl(\sqrt{n/d} + \log n\bigr)$. In the proofs, the following notations will be helpful: For $j \leq 0$, let $\sigma_j = \nu_j = 0$. This allows us to define $\delta_i = \sigma_i - \sigma_{i-\sqrt{nd}}$ for all $i \in \{1, \ldots, n\}$. We define $I_i = \{j \in \{1, \ldots, n\} \mid i - \sqrt{nd} \leq j < i\}$ to be the set of the $|I_i| = \min\{i-1, \sqrt{nd}\}$ positions that precede i.

3.1 Upper Bound on the Smoothed Number of Left-To-Right Maxima

To prove the upper bound for the smoothed number of left-to-right maxima, we proceed in two steps: First, we show that the adversary should choose a sorted sequence. Second, we prove that the expected number of left-to-right maxima of sorted sequences is $O\bigl(\sqrt{n/d} + \log n\bigr)$.

Lemma 3.2. *For every sequence σ and its sorted version τ, we have*

$$\mathbb{E}\bigl(\text{ltrm}_d(\sigma)\bigr) \leq \mathbb{E}\bigl(\text{ltrm}_d(\tau)\bigr).$$

Proof. We prove the lemma by "bubble-sorting" σ. If σ is already sorted, then there is nothing to show. Otherwise, there exist adjacent σ_i and σ_{i+1} with $\sigma_i > \sigma_{i+1}$. Our aim is to show that $\mathbb{E}\bigl(\text{ltrm}_d(\sigma)\bigr) \leq \mathbb{E}\bigl(\text{ltrm}_d(\tau)\bigr)$, where τ is obtained from σ by swapping σ_i and σ_{i+1}. Then the claim follows by iteratively applying this argument.

After perturbation with ν, we obtain $\overline{\sigma}$ and $\overline{\tau}$, where $\overline{\tau}_i = \tau_i + \nu_{i+1} = \sigma_{i+1} + \nu_{i+1}$ and $\overline{\tau}_{i+1} = \tau_{i+1} + \nu_i = \sigma_i + \nu_i$. Now we analyze the number of left-to-right maxima of $\overline{\sigma}$ and $\overline{\tau}$. To do this, let $\delta = \sigma_i - \sigma_{i+1} > 0$. We distinguish two cases. First, we condition on $\nu_i \in [0, d-\delta]$ and $\nu_{i+1} \in [\delta, d]$. In this case, both $(\overline{\sigma}_i, \overline{\sigma}_{i+1})$ and $(\overline{\tau}_i, \overline{\tau}_{i+1})$ are pairs of random numbers, all of which lie uniformly in the interval $[\sigma_i, \sigma_{i+1} + d]$. Then the expected numbers of left-to-right maxima of $\overline{\sigma}$ and $\overline{\tau}$ are equal. Second, we condition on the event that $\nu_i \in (d-\delta, d]$ or $\nu_{i+1} \in [0, \delta)$. In either case, both

$\overline{\sigma}_i > \overline{\sigma}_{i+1}$ and $\overline{\tau}_i < \overline{\tau}_{i+1}$ hold. Then $\overline{\sigma}_{i+1}$ cannot be a left-to-right maximum in $\overline{\sigma}$, and if $\overline{\sigma}_i$ is a left-to-right maximum in $\overline{\sigma}$, then so is $\overline{\tau}_{i+1}$ in $\overline{\tau}$.

Since the case distinction is exhaustive, the lemma is proved. □

Lemma 3.3. *For every sequence σ of length n and all $d \geq 1/n$, we have*

$$\mathbb{E}\big(\mathrm{ltrm}_d(\sigma)\big) \in O\big(\sqrt{n/d} + \log n\big).$$

Proof. By Lemma 3.2 we can restrict ourselves to proving the lemma for sorted sequences σ. We estimate the probability that a given $\overline{\sigma}_i$ for $i \in \{1, \ldots, n\}$ is a left-to-right maximum. Then the bound follows by the linearity of expectation. To bound the probability that $\overline{\sigma}_i$ is a left-to-right maximum (ltrm), consider the following computation:

$$\begin{aligned}
\mathbb{P}\big(\overline{\sigma}_i \text{ is an ltrm}\big) \;&\leq\; \mathbb{P}\big(\forall j \in I_i \colon \nu_j < \overline{\sigma}_i - \sigma_{i-\sqrt{nd}}\big) && (1)\\
&\leq\; \mathbb{P}(d < \overline{\sigma}_i - \sigma_{i-\sqrt{nd}}) + \int_0^{d-\delta_i} \mathbb{P}\big(\forall j \in I_i \colon \nu_j < \sigma_i + x - \sigma_{i-\sqrt{nd}}\big)\,\mathrm{d}x && (2)\\
&\leq\; \frac{\delta_i}{d} + \int_0^{d} \mathbb{P}\big(\forall j \in I_i \colon \nu_j < x\big)\,\mathrm{d}x && (3)\\
&\leq\; \frac{\delta_i}{d} + \mathbb{P}\big(\forall j \in I_i \colon \nu_j < \nu_i\big) \;=\; \frac{\delta_i}{d} + \frac{1}{|I_i| + 1}. && (4)
\end{aligned}$$

To see that (1) holds, assume that $\overline{\sigma}_i$ is a left-to-right maximum. Then $\overline{\sigma}_i - \sigma_{i-\sqrt{nd}}$ must be larger than the noises of all the elements in the index range I_i, for if the noise ν_j of some element σ_j were larger than $\overline{\sigma}_i - \sigma_{i-\sqrt{nd}}$, then $\overline{\sigma}_j = \sigma_j + \nu_j$ would be larger than $\sigma_j + \overline{\sigma}_i - \sigma_{i-\sqrt{nd}}$. Since the sequence is sorted, $\sigma_j + \overline{\sigma}_i - \sigma_{i-\sqrt{nd}} \geq \overline{\sigma}_i$ and $\overline{\sigma}_i$ would not be a left-to-right maximum.

For (2), first observe that $\nu_j < \overline{\sigma}_i - \sigma_{i-\sqrt{nd}}$ is surely the case for all $j \in I_i$ if $d < \overline{\sigma}_i - \sigma_{i-\sqrt{nd}}$. So, consider the case $d \geq \overline{\sigma}_i - \sigma_{i-\sqrt{nd}} = \delta_i + \nu_i$. Then $\nu_i \in [0, d - \delta_i]$ and we can rewrite $\mathbb{P}(\forall j \in I_i \colon \nu_j < \delta_i + \nu_i)$ as $\int_0^{d-\delta_i} \mathbb{P}(\forall j \in I_i \colon \nu_j < \delta_i + x)\,\mathrm{d}x$. For (3) observe that $d < \overline{\sigma}_i - \sigma_{i-\sqrt{nd}}$ is equivalent to $d - \delta_i < \nu_i$ and the probability of this is δ_i/d. Furthermore, we performed an index shift in the integral. In (4), we replaced the integral by a probability once more and get the final result.

Let us bound $\sum_{i=1}^n \delta_i$. We have $\sum_{i=1}^n \delta_i = \sum_{i=1}^n (\sigma_i - \sigma_{i-\sqrt{nd}}) = \sum_{i=n-\sqrt{nd}+1}^n \sigma_i \leq \sqrt{nd}$. The second equality holds since most σ_i cancel themselves out and $\sigma_i = 0$ for $i \leq 0$. The inequality holds since there are $\sqrt{nd}$ summands. We complete the proof by bounding $1/(|I_i| + 1) = 1/\min\{i, \sqrt{nd}\}$ by $1/i + 1/\sqrt{nd}$ and summing over all i:

$$\mathbb{E}\big(\mathrm{ltrm}_d(\sigma)\big) \leq \sum_{i=1}^n \frac{\delta_i}{d} + \sum_{i=1}^n \frac{1}{|I_i|} \leq \frac{\sqrt{nd}}{d} + \sum_{i=1}^n \frac{1}{i} + n\frac{1}{\sqrt{nd}} \in O\left(\sqrt{n/d} + \log n\right).$$ □

3.2 Lower Bound on the Smoothed Number of Left-To-Right Maxima

Let us now show a lower bound that matches the upper bound proved in the previous section. Although the sequences may consist of n arbitrary numbers from the interval $[0, 1]$, it suffices to consider sorted sequences $(1/n, 2/n, \ldots, n/n)$.

Lemma 3.4. *For the sequence $\sigma = (1/n, 2/n, \ldots, n/n)$ and all $d \geq 1/n$, we have*

$$\mathbb{E}\big(\mathrm{ltrm}_d(\sigma)\big) \in \Omega\big(\sqrt{n/d} + \log n\big).$$

Proof. We assume that $d \geq 4/n$. This is no restriction since for $d < 4/n$, we immediately get a lower bound of $n/4 \in \Omega(n)$ in compliance with the theorem.

We give two estimates for the probability that a given $\overline{\sigma}_i$ is a left-to-right maximum; only this time, we need to bound this probability from below. The first estimate is simple:

$$\mathbb{P}\big(\overline{\sigma}_i \text{ is an ltrm}\big) = \mathbb{P}\big(\forall j < i \colon \nu_j + j/n < \nu_i + i/n\big) \geq \mathbb{P}\big(\forall j < i \colon \nu_j < \nu_i\big) = 1/i.$$

For the second estimate, assume $\nu_i > d - \sqrt{d/n}$ for a given $i \in \{1, \ldots, n\}$. Then $\overline{\sigma}_j < \overline{\sigma}_i$ for all $j \in \{1, \ldots, i - \sqrt{nd} - 1\}$ since the noise of σ_i is so large that σ_j before $\sigma_{i-\sqrt{nd}}$ can never reach $\overline{\sigma}_i$ even when a noise of d is added. This shows that $\overline{\sigma}_i$ is a left-to-right maximum if (a) we have $\nu_i > d - \sqrt{d/n}$ and (b) we have $\nu_j < d - \sqrt{d/n}$ for all $j \in I_i$. The probability of (a) is $1/\sqrt{nd}$ and of (b) is $(1 - 1/\sqrt{nd})^{|I_i|} \geq (1 - 1/\sqrt{nd})^{\sqrt{nd}}$. Since $d \geq 4/n$, this yields

$$\mathbb{P}\big(\overline{\sigma}_i \text{ is an ltrm}\big) \geq \frac{1}{\sqrt{nd}} \left(1 - \frac{1}{\sqrt{nd}}\right)^{\sqrt{nd}} \geq \frac{1}{4 \cdot \sqrt{nd}}.$$

The two estimates together yield $\mathbb{P}\big(\overline{\sigma}_i \text{ is an ltrm}\big) \geq \max\{1/i, 1/(4\sqrt{nd})\} \geq \frac{1}{2}\big(1/i + 1/(4\sqrt{nd})\big)$. By the linearity of expectation we get

$$\mathbb{E}\big(\mathrm{ltrm}_d(\sigma)\big) \geq \frac{1}{2}\sum_{i=1}^{n} \frac{1}{i} + \frac{1}{2}\sum_{i=1}^{n} \frac{1}{4\sqrt{nd}} \in \Theta(\log n + \sqrt{n/d}). \qquad \square$$

4 Smoothed Height of Binary Search Trees

In this section we prove our first main result, an exact bound on the smoothed height of binary search trees under additive noise. The bound is the same as for left-to-right maxima, as stated in the following theorem.

Theorem 4.1. *For $d \geq 1/n$, we have*

$$\max_{\sigma \in [0,1]^n} \mathbb{E}\big(\mathrm{height}_d(\sigma)\big) \in \Theta\big(\sqrt{n/d} + \log n\big).$$

In the rest of this section, we prove this theorem. We have to prove an upper and a lower bound, but the lower bound follows directly from the lower bound of $\Omega\big(\sqrt{n/d} + \log n\big)$ for the smoothed number of left-to-right maxima (recall that the number of left-to-right maxima in a sequence is the length of the rightmost path of the sequence's search tree). Thus, we only need to focus on the upper bound.

To prove the upper bound of $O\big(\sqrt{n/d} + \log n\big)$ on the smoothed height of binary search trees, we need some preparations. In the next subsection we introduce the concept of *increasing and decreasing runs* and show how they are related to binary search tree height. As we will see, bounding the length of these runs implicitly bounds the height of binary search trees. This allows us to prove the upper bound on the smoothed height of binary search trees in the main part of this section.

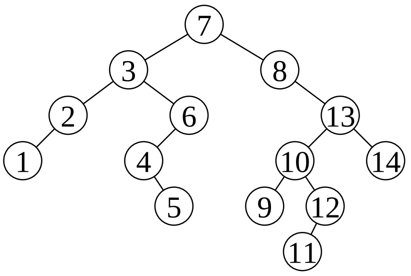

Figure 1: The tree $T(\sigma)$ obtained from the sequence $\sigma = (7, 8, 13, 3, 2, 10, 9, 6, 4, 12, 14, 1, 5, 11)$. We have $\text{height}(\sigma) = 6$. The root-to-leaf path ending at 11 can be divided into the increasing run $\alpha = (7, 8, 10, 11)$ and the decreasing run $\beta = (13, 12, 11)$.

4.1 Increasing and Decreasing Runs

In order to analyze the smoothed height of binary search trees, we introduce a related measure for which an upper bound is easier to obtain. Given a sequence σ, consider a root-to-leaf path of the tree $T(\sigma)$. We extract two subsequences $\alpha = (\alpha_1, \ldots, \alpha_k)$ and $\beta = (\beta_1, \ldots, \beta_\ell)$ from this path according to the following algorithm: We start at the root. If we are at an element σ_i of the path, we look at the direction in which the path continues from σ_i. If it continues with the right child of σ_i, we append σ_i to α; if it continues with the left child, we append σ_i to β; and if σ_i is a leaf (has no children), then we append σ_i to both α and β. This construction ensures $\alpha_1 < \cdots < \alpha_k = \beta_\ell < \cdots < \beta_1$ and the length of σ is $k + \ell - 1$. Figure 1 shows an example of how α and β are constructed.

A crucial property of the sequence α is the following: Let $\alpha_i = \sigma_{j_i}$ for $i \in \{1, \ldots, k\}$ with $j_1 < j_2 < \cdots < j_k$. Then none of $\sigma_1, \ldots, \sigma_{j_i - 1}$ lies in the interval (α_i, α_{i+1}), for otherwise α_i and α_{i+1} cannot be on the same root-to-leaf path. A similar property holds for the sequence β: No element of σ prior to β_i lies in the interval (β_{i+1}, β_i). We introduce a special name for sequences with this property.

Definition 4.2. *Let σ be a sequence. An* increasing run *of σ is a subsequence $(\sigma_{i_1}, \sigma_{i_2}, \ldots, \sigma_{i_k})$ of σ with $\sigma_{i_1} < \cdots < \sigma_{i_k}$ with the following property: No element of σ prior to σ_{i_k} lies in the interval $(\sigma_{i_k}, \sigma_{i_{k+1}})$. Analogously, a* decreasing run *of σ is a subsequence $(\sigma_{i_1}, \ldots, \sigma_{i_\ell})$ with $\sigma_{i_1} > \cdots > \sigma_{i_\ell}$ such no element prior to σ_{i_k} lies in the interval $(\sigma_{i_{k+1}}, \sigma_{i_k})$.*

Let $\text{inc}(\sigma)$ and $\text{dec}(\sigma)$ denote the length of the longest increasing and decreasing run of σ, respectively. Furthermore, let $\text{dec}_d(\sigma)$ and $\text{inc}_d(\sigma)$ denote the length of the longest runs under d-noise. In Figure 1, we have $\text{inc}(\sigma) = 4$ because of $(7, 8, 10, 12)$ or $(7, 8, 13, 14)$ and $\text{dec}(\sigma) = 4$ because of $(7, 3, 2, 1)$.

Since every root-to-leaf path can be divided into an increasing and a decreasing run, we immediately obtain the following lemma.

Lemma 4.3. *For every sequence σ and all d we have*

$$\text{height}(\sigma) \leq \text{dec}(\sigma) + \text{inc}(\sigma),$$
$$\mathbb{E}\big(\text{height}_d(\sigma)\big) \leq \mathbb{E}\big(\text{dec}_d(\sigma) + \text{inc}_d(\sigma)\big).$$

In terms of upper bounds, $\operatorname{dec}(\sigma)$ and $\operatorname{inc}(\sigma)$ as well as $\operatorname{dec}_d(\sigma)$ and $\operatorname{inc}_d(\sigma)$ behave equally. The reason is that given a sequence σ, the sequence τ with $\tau_i = 1-\sigma_i$ has the properties $\operatorname{dec}(\sigma) = \operatorname{inc}(\tau)$ and $\mathbb{E}\bigl(\operatorname{dec}_d(\sigma)\bigr) = \mathbb{E}\bigl(\operatorname{inc}_d(\tau)\bigr)$. This observation together with Lemma 4.3 proves the next lemma.

Lemma 4.4. *For all d, we have*

$$\max_{\sigma\in[0,1]^n} \mathbb{E}\bigl(\operatorname{height}_d(\sigma)\bigr) \leq 2 \cdot \max_{\sigma\in[0,1]^n} \mathbb{E}\bigl(\operatorname{inc}_d(\sigma)\bigr).$$

The lemma states that in order to bound the smoothed height of search trees from above we can instead bound the smoothed length of increasing or decreasing runs. To simplify the analysis even further, we show that we can once more restrict our attention to sorted sequences.

Lemma 4.5. *For every sequence σ and its sorted version τ, we have*

$$\mathbb{E}\bigl(\operatorname{inc}_d(\sigma)\bigr) \leq \mathbb{E}\bigl(\operatorname{inc}_d(\tau)\bigr).$$

Proof. We sort σ successively as we already did for Lemma 3.2. Assume that $\sigma_i > \sigma_{i+1}$ for some i and let $\delta = \sigma_i - \sigma_{i+1} > 0$. We show $\mathbb{E}(\operatorname{inc}_d(\sigma)) \leq \mathbb{E}(\operatorname{inc}_d(\tau))$, where τ is obtained from σ by swapping σ_i and σ_{i+1}. Let ν denote the noise vector added to σ and τ. Then $\overline{\tau}_i = \tau_i + \nu_{i+1} = \sigma_{i+1} + \nu_{i+1}$ and $\overline{\tau}_{i+1} = \tau_{i+1} + \nu_i = \sigma_i + \nu_i$.

We distinguish two cases. First, we condition on $\nu_i \in [0, d-\delta]$ and $\nu_{i+1} \in [\delta, d]$. Similar to the argument in Lemma 3.2, both $(\overline{\sigma}_i, \overline{\sigma}_{i+1})$ and $(\overline{\tau}_i, \overline{\tau}_{i+1})$ are pairs of random numbers, all of which lie uniformly in the interval $[\sigma_i, \sigma_{i+1} + d]$, and the expected values of $\operatorname{inc}_d(\sigma)$ and $\operatorname{inc}_d(\tau)$ are equal.

Second, we condition on the events that $\nu_i \in (d-\delta, d]$ or $\nu_{i+1} \in [0, \delta)$. In either case, $\overline{\sigma}_i > \overline{\sigma}_{i+1}$ and $\overline{\tau}_i < \overline{\tau}_{i+1}$. Thus, every increasing run of $\overline{\sigma}$ corresponds to an increasing run of $\overline{\tau}$: If the run of $\overline{\sigma}$ uses neither $\overline{\sigma}_i$ nor $\overline{\sigma}_{i+1}$, this is obvious. If the run of $\overline{\sigma}$ uses $\overline{\sigma}_i$, then we get the same run of $\overline{\tau}$, where now $\overline{\tau}_{i+1}$ is used. The run cannot be interrupted by $\overline{\tau}_i$ because $\overline{\tau}_i < \overline{\tau}_{i+1}$. If the run of $\overline{\sigma}$ uses $\overline{\sigma}_{i+1}$, then we obtain a run of the same length using $\overline{\tau}_i$. This run is also an increasing run since the only difference of $\overline{\sigma}$ and $\overline{\tau}$ is that now the larger element $\overline{\tau}_{i+1}$ appears after $\overline{\tau}_i$. Finally, the run of $\overline{\sigma}$ cannot use both $\overline{\sigma}_i$ and $\overline{\sigma}_{i+1}$ because of $\overline{\sigma}_{i+1} < \overline{\sigma}_i$. Thus, we have $\operatorname{inc}(\overline{\sigma}) \leq \operatorname{inc}(\overline{\tau})$, which proves the lemma. □

4.2 Upper Bound on the Smoothed Height of Binary Search Trees

We are now ready to prove the upper bound for binary search trees by proving an upper bound on the smoothed length of increasing runs of sorted sequences. For this, we prove four lemmas, the last of which claims exactly the desired upper bound.

Lemma 4.6 deals with $d = 1$ and states that $\mathbb{E}\bigl(\operatorname{height}_1(\sigma)\bigr) \in O(\sqrt{n})$ for every sequence σ.

Lemma 4.7 states that in order to bound tree heights, we can divide sequences into (possibly overlapping) parts and consider the height of the trees induced by the subsequences individually. A less general form of the lemma has already been shown by Manthey and Reischuk [6, Lemma 4.1].

Lemma 4.8 establishes that if $d = n/\log^2 n$, a perturbed sequence behaves the same way as a completely random sequence with respect to the smoothed length of its longest increasing run. The core idea is to partition the sequence into a set of "completely random" elements, which behave as expected, and two sets of more bothersome elements lying in a small range. As we will see, the number of bothersome elements is roughly $\log^2 n$ and since the range of values of these elements is small, we can use the result $\mathbb{E}\bigl(\operatorname{height}_1(\sigma)\bigr) \in O(\sqrt{n})$ to show that their contribution to the length on an increasing run is just $O(\log n)$.

Finally, in Lemma 4.9 we allow general $d \geq 1/n$. This case turns out to be reducible to the case $d = n/\log^2 n$ by a scaling argument.

For the proofs of the lemmas, two technical terms will be helpful: For a given real interval $I = [a, b]$, we say that a position i of σ is *eligible* for I if $\overline{\sigma}_i$ can assume any value in I. In other words, i is eligible for $[a, b]$ if $\sigma_i \leq a$ and $\sigma_i + d \geq b$. Furthermore, we say furthermore that i is *regular* if $\overline{\sigma}_i$ actually lies inside I.

Lemma 4.6. *For every sequence σ, we have*

$$\mathbb{E}\big(\mathrm{inc}_1(\sigma)\big) \in O(\sqrt{n}).$$

Proof. Let us take a closer look at increasing runs. Every increasing run of a sequence $\overline{\sigma}$ starts with a number of left-to-right maxima. However, after the first element that is not a left-to-right maximum, the run does not contain any more left-to-right maxima. More formally: If $\alpha_1, \ldots, \alpha_\ell$ is an increasing run of $\overline{\sigma}$, then there exists a $k \in \{0, 1, \ldots, \ell\}$ such that all elements $\alpha_1, \ldots, \alpha_k$ are left-to-right maxima of $\overline{\sigma}$ while α_{k+1} is not (or $k = \ell$). Furthermore, if $\overline{\sigma}_i$ is the first left-to-right maximum after α_k, then the remaining elements $\alpha_{k+1}, \ldots, \alpha_\ell$ all lie in the interval $[\alpha_k, \overline{\sigma}_i]$.

Due to this property, it suffices to a) bound $\mathbb{E}\big(\mathrm{ltrm}_1(\sigma)\big)$ and b) bound the maximum length of an increasing run such that the values of the elements of the run lie between the values of two consecutive left-to-right maxima. By Lemma 3.3, we have $\mathbb{E}\big(\mathrm{ltrm}_1(\sigma)\big) \in O(\sqrt{n})$, so let us focus on b). For the bound we prove two claims and for the formulation of these claims the following definition is helpful: We say that a set of numbers x_i is *ϵ-dense for an interval I* if every interval $J \subseteq I$ of length ϵ contains at least one x_i.

Claim 1. *Let $a \leq s < b$, let $y_1, \ldots, y_{\sqrt{n}} \in [a, s]$, and let $x_1, \ldots, x_{\sqrt{n}}$ be random variables where x_i is uniformly distributed in the interval $[y_i, b]$. Then the x_i are $\big((b - a)n^{-1/4}\big)$-dense for the interval $[s, b]$ with a probability of at least $1 - \exp\big(-\Theta(n^{1/4})\big)$.*

Proof. By a scaling argument, it suffices to consider the case $a = 0$, $b = 1$, and $s \in [0, 1]$. We divide the interval $[s, 1]$ into subintervals $I_1, I_2, \ldots, I_k$, where

$$I_j = \Big[s + (j - 1) \cdot n^{-1/4}/2, s + j \cdot n^{-1/4}/2\Big]$$

and $k = \lfloor (1 - s) \cdot 2n^{1/4} \rfloor$. Every interval J of length $n^{1/4}$, which is twice the length of an I_j, contains at least one I_j as a subinterval. Thus, if every I_j contains at least one x_i, no interval of length $n^{1/4}$ is empty. The probability that any fixed element x_i does not assume a value in I_j is at most $1 - n^{-1/4}/2$. Thus, the probability that I_j does not contain any x_i is at most

$$\left(1 - \frac{n^{-1/4}}{2}\right)^{\sqrt{n}} = \left(\left(1 - \frac{1}{2n^{1/4}}\right)^{2n^{1/4}}\right)^{n^{1/4}/2} \leq \exp\big(-n^{1/4}/2\big).$$

Using the union bound, we obtain that the probability that there exists an empty I_j is at most $k \cdot \exp\big(-n^{1/4}/2\big) \in \exp\big(-\Theta(n^{1/4})\big)$. □

Claim 2. *Let $\tau = (\tau_1, \ldots, \tau_k)$ be the sequence obtained by sorting $(\overline{\sigma}_1, \ldots, \overline{\sigma}_k)$, and let $\tau_0 = 0$. Let $j \in \{1, \ldots, k\}$, and let $I = [\tau_{j-1}, \tau_j]$. Let $\overline{\sigma}_{\ell_1}, \ldots, \overline{\sigma}_{\ell_{\sqrt{n}}}$ be the first $\sqrt{n}$ elements of $\overline{\sigma}$ that fall into I. Let $s = \max\{\tau_{j-1}, \sigma_{\ell_{\sqrt{n}}}\}$. Then the set $\{\overline{\sigma}_{\ell_1}, \ldots, \overline{\sigma}_{\ell_{\sqrt{n}}}\}$ is $\big((\tau_j - \tau_{j-1})n^{-1/4}\big)$-dense for the interval $[s, \tau_j]$ with a probability of at least $1 - \exp\big(-\Theta(n^{1/4})\big)$.*

Proof. This follows from the first claim by setting $a = \tau_{j-1}$, $b = \tau_j$, and $y_i = \max\{a, \sigma_{\ell_i}\}$ for all $i \in \{1, \ldots, \sqrt{n}\}$. To see this, first note that we added the element $\tau_0 = 0$ so that the interval $[0, \tau_1]$ does not require special attention. Second, note that, indeed, each $\overline{\sigma}_{\ell_i}$ is uniformly distributed in $[y_i, b] = \big[\max\{a, \sigma_{\ell_i}\}, \tau_j]$. Finally, note that $y_i = \max\{a, \sigma_{\ell_i}\} \in [a, s] = [\tau_{j-1}, \max\{\tau_{j-1}, \sigma_{\ell_{\sqrt{n}}}\}]$ holds since σ is a sorted sequence. □

Let us return to our original aim: We wish to bound the length of an increasing run for which the values of the run all lie between the values of two consecutive left-to-right maxima. The idea is to apply Claim 2 twice. Each time, with high probability, if we consider additional $\sqrt{n}$ elements of the run, their values must be dense for a smaller and smaller interval. After having applied the claim twice, the interval is so small that, again with high probability, the interval can only contain a small number of elements – which proves that the total number of elements in the run cannot be large. In the following, we detail this argument.

Let $a = \overline{\sigma}_i$ be a left-to-right maximum of $\overline{\sigma}$, and let $b = \overline{\sigma}_k > a$ be the next left-to-right maximum after a, that is, $k > i$ is chosen minimally such that $\overline{\sigma}_k > a$. If less than $\sqrt{n}$ elements assume a value in $[a, b]$, then these elements contribute at most $\sqrt{n}$ to the length of an increasing run (recall that all elements of an increasing run must lie inside the interval $[a, b]$ if a and b are consecutive left-to-right maxima). Otherwise, we apply Claim 2 for $j = k$: The first $\sqrt{n}$ elements of $\overline{\sigma}$ that are inserted into $[a, b]$ lie $\big((b - a)n^{-1/4}\big)$-densely in $[s, b]$ with a probability of at least $1 - \exp\big(-\Theta(n^{1/4})\big)$. If this not the case (despite the high probability), we call the situation a *failure*, which will be dealt with later.

Among the first $\sqrt{n}$ elements in $[a, b]$, let $c_1 < \ldots < c_m$ be the elements in $[s, b]$ in increasing order. Let $c_0 = s$ and $c_{m+1} = b$. Then, after inserting these m elements into $[s, b]$, any increasing run whose elements assume values in $[s, b]$ can only be continued with elements of a subinterval $J = [c_{i-1}, c_i]$. The length $c_i - c_{i-1}$ of J is at most $(b - a)n^{-1/4} \leq 2n^{-1/4}$ since the first $\sqrt{n}$ elements are $\big((b - a)n^{-1/4}\big)$-dense in $[s, b]$. We apply the same argument once more: If less than $\sqrt{n}$ elements fall into J, then they contribute at most $\sqrt{n}$ to the length of an increasing run. Otherwise, we can apply Claim 2 again: The probability that after inserting $\sqrt{n}$ elements into J, there is a subinterval of J of length at least $(c_i - c_{i-1})n^{-1/4}$ that does not contain an element is at most $\exp\big(-\Theta\big(n^{1/4}\big)\big)$. If this nevertheless happens, we again call this a failure and deal with it later on.

We can now conclude that the values of all elements of an increasing run that we have not yet dealt with must lie in an interval of size at most $(c_i - c_{i-1})n^{-1/4} \leq 2n^{-1/2}$. The expected number of elements that fall into such an interval is at most $2\sqrt{n}$. By the Chernoff bound, the probability that such an interval contains more than $4\sqrt{n}$ elements is at most $\exp\big(-\Omega\big(\sqrt{n}\big)\big)$. Again, we call it a failure if this nevertheless happens.

To finish the proof, we first estimate the length of the longest increasing run given that we have no failures. Second, we show that failures happen only with a negligible probability, thus contribute not too much to the expected value of $\mathrm{inc}(\overline{\sigma})$.

If we have no failure, then the expected length of any increasing run is at most $\mathbb{E}\big(\mathrm{ltrm}(\overline{\sigma})\big)$ plus $\sqrt{n}$ for the first $\sqrt{n}$ elements that fall between two consecutive left-to-right maxima plus $\sqrt{n}$ for the first $\sqrt{n}$ elements that fall into one interval of length at most $2n^{-1/4}$ plus $4 \cdot \sqrt{n}$, which is the maximum number of elements in an interval of length $2/\sqrt{n}$. Altogether, we have $\mathrm{inc}(\overline{\sigma}) \in O(\sqrt{n})$ in this case.

Failure can happen due to the following three events: First, there is an interval of length at most $2n^{-1/4}$ between two left-to-right maxima that does not contain one of the first $\sqrt{n}$ elements that fall between these left-to-right maxima. Second, there is an interval of length at most $2/\sqrt{n}$

without any of the first $\sqrt{n}$ elements after the second phase. Third, any interval of length $2/\sqrt{n}$ contains more than $4\sqrt{n}$ elements.

Overall, there are at most $O(n)$ pairs of consecutive left-to-right maxima between which a failure can happen. Furthermore, there are at most $O(n)$ pairs of consecutive elements a and b between failure can happen. Finally, there are at most $O(n)$ intervals of length at most $2/\sqrt{n}$ where failure can happen. (These are very rough estimates.) By taking a union bound, the overall probability of failure is thus at most $O(n \cdot \exp(-\Omega(n^{1/4})))$. If we have failure, then we bound $\mathrm{inc}(\overline{\sigma})$ by the trivial upper bound of n. This contributes only $O(n^2 \cdot \exp(-\Omega(n^{1/4}))) \subseteq o(1)$ to the expected value, which completes the proof. □

Lemma 4.7. *For every sequence σ, all d, and every covering $U_1, \ldots, U_k$ of $\{1, \ldots, n\}$ (which means $\bigcup_{i=1}^k U_i = \{1, \ldots, n\}$), we have*

$$\mathrm{height}(\sigma) \leq \textstyle\sum_{i=1}^k \mathrm{height}(\sigma_{U_i}),$$
$$\mathbb{E}\big(\mathrm{height}_d(\sigma)\big) \leq \textstyle\sum_{i=1}^k \mathbb{E}\big(\mathrm{height}_d(\sigma_{U_i})\big).$$

Proof. Let $U_1, \ldots, U_k$ cover $\{1, \ldots, n\}$. For a fixed i, let a and b with $a < b$ be two elements of σ_{U_i} that do not lie on the same root-to-leaf path in $T(\sigma_{U_i})$. Then there exists a c prior to a and b in σ_{U_1} with $a < c < b$, which implies that a and b do no lie on the same root-to-leaf path in the tree $T(\sigma)$ either. Now consider a root-to-leaf path p of $T(\sigma)$ that has a length of $\mathrm{height}(\sigma)$. Let p_{U_i} be p restricted to elements of σ_{U_1} and let ℓ_{U_i} its length. Then $\sum_{i=1}^k \mathrm{height}(\sigma_{U_1}) \geq \sum_{i=1}^k \ell_{U_i} \geq \mathrm{height}(\sigma)$, because the U_i cover $\{1, \ldots, n\}$.

The second inequality follows directly from the first since both taking expectation and smoothing are monotone operations. □

Lemma 4.8. *For every sequence $\sigma \in [0, 1]^n$ and for $d = n/(\log n)^2$, we have*

$$\mathbb{E}\big(\mathrm{height}_d(\sigma)\big) \in O(\log n).$$

Proof. Recall that a position i is called *eligible* for an interval if $\overline{\sigma}_i$ could be any value in the interval, and it is called *regular* if it actually lies in the interval. All positions are eligible for $[1, d]$.

Let $\overline{\sigma}$ be a perturbed sequence and let R be the set of regular positions. A sufficient condition for i being regular is $\nu_i \in [1, d-1]$. Let $F = \{i \in \{1, \ldots, n\} \mid \nu_i \leq 1\}$ and $B = \{i \in \{1, \ldots, n\} \mid \nu_i \geq d-1\}$ denote the sets of positions i that are possibly not regular because ν_i is either too small or too large.

The three sets R, F, and B are usually not disjoint, but they cover $\{1, \ldots, n\}$, which allows us to apply Lemma 4.7: If we can individually bound the expected values of $\mathrm{height}_d(\sigma_R)$, $\mathrm{height}_d(\sigma_F)$ and $\mathrm{height}_d(\sigma_B)$ by $O(\log n)$, we are done.

Let us start with $\mathbb{E}\big(\mathrm{height}(\sigma_R)\big)$. Given that a position i is regular, the element $\overline{\sigma}_i$ is uniformly distributed in $[1, d]$ and, thus, the order of the elements of $\overline{\sigma}_R$ is random with all permutations being equally likely. This implies that $\mathbb{E}\big(\mathrm{height}(\overline{\sigma}_R)\big) \in O(\log |R|) \subseteq O(\log n)$.

It remains to deal with $\overline{\sigma}_F$ and $\overline{\sigma}_B$. The distributions of $\mathrm{height}(\overline{\sigma}_F)$ and $\mathrm{height}(\overline{\sigma}_B)$ are clearly identical, so it suffices to analyze $\mathrm{height}(\overline{\sigma}_F)$. For this, take a different look at how $\overline{\sigma}_F$ is generated: We can think of this as first flipping a coin for every $i \in \{1, \ldots, n\}$ to determine $i \in F$ (with the coin being extremely biased so that $\mathbb{P}(i \in F) = 1/d = (\log^2 n)/n$ holds). After we have chosen F, we draw ν_i for each $i \in F$ uniformly at random from the interval $[0, 1]$.

By the Chernoff bound, the probability that $\overline{\sigma}_F$ contains more than $2(\log n)^2$ elements is less than $n^{-(\log n)/3}$. If $\overline{\sigma}_F$ indeed contains more elements, we bound height$(\overline{\sigma}_F)$ by n. This contributes only $n^{-(\log n)/3} \cdot n \in o(1)$ to the expected value of height$(\overline{\sigma}_F)$.

We can now apply Lemma 4.6 to σ_F, where $n' \leq 2(\log n)^2$ and $d' = 1$, and get $\mathbb{E}\big(\text{height}_1(\tau)\big) \in O(\sqrt{n'}) \subseteq O(\log n)$. □

Lemma 4.9. *For every sequence σ and all $d \geq 1/n$ we have*

$$\mathbb{E}\big(\text{height}_d(\sigma)\big) \in O\big(\sqrt{n/d} + \log n\big).$$

Proof. If $d \in \Omega\big(n/(\log n)^2\big)$, then $\mathbb{E}\big(\text{height}_d(\sigma)\big) \in O(\log n)$ by Lemma 4.8.

To prove the theorem for smaller values of d, we divide the sequence into subsequences. Let N solve the equation $N^2/\log^2 N = nd$. Then $\log N \in \Theta(\log(nd))$, and thus $N = c \cdot \sqrt{nd} \cdot \log(nd)$ for some $c \in \Theta(1)$. Let n_j be the number of elements of σ with $\sigma_i \in [(j-1) \cdot N/n, j \cdot N/n]$. Choose $k_j \in \mathbb{N}$ such that $(k_j - 1) \cdot N < n_j \leq k_j N$. We divide the n_j elements of the interval $[(j-1) \cdot N/n, j \cdot N/n]$ into k_j subsequences $\sigma^{j,1}, \ldots, \sigma^{j,k_j}$ such that no subsequence contains more than N elements. Since

$$\sum_{j=1}^{n/N} k_j \leq \sum_{j=1}^{n/N} \frac{n_j + N}{N} \leq 2n/N,$$

we obtain at most $2n/N$ such subsequences. Each subsequence spans at most an interval of length N/n and contains at most N elements. Thus, by Lemma 4.8, we have height$_d(\sigma^{j,\ell}) \in O(\log(N))$. Finally, Lemma 4.7 yields

$$\mathbb{E}\big(\text{height}_d(\sigma)\big) \leq \sum_{j=1}^{n/N} \sum_{\ell=1}^{k_j} \mathbb{E}\big(\text{height}_d(\sigma^{j,\ell})\big) \in O\left(\frac{n \cdot \log N}{N}\right) = O\left(\sqrt{n/d}\right).$$

□

5 Smoothed Number of Quicksort Comparisons

In this section, we apply our results about binary search trees and left-to-right maxima to analyze the performance of the quicksort algorithm. The following theorem summarizes the findings.

Theorem 5.1. *For $d \geq 1/n$ we have*

$$\max_{\sigma \in [0,1]^n} \mathbb{E}\big(\text{qs}_d(\sigma)\big) \in \Theta\big(\tfrac{n}{d+1}\sqrt{n/d} + n \cdot \log n\big).$$

In other words, for $d \in O(1)$, the number of comparisons is at most $O(n\sqrt{n/d})$, while for $d \in \Omega(1)$, it is at most $O(\frac{n}{d}\sqrt{n/d})$. This means that d has a stronger influence for $d \in \Omega(1)$.

5.1 Upper Bound on the Smoothed Number of Quicksort Comparisons

To prove the upper bound, we first need a lemma similar to Lemma 4.7 that allows us to estimate the number of comparisons of subsequences.

Lemma 5.2. *For every sequence σ, all d, and every covering $U_1, \ldots, U_k$ of $\{1, \ldots, n\}$, we have*

$$\mathrm{qs}(\sigma) \leq \textstyle\sum_{i=1}^{k} \mathrm{qs}(\sigma_{U_i}) + Q,$$
$$\mathrm{qs}_d(\sigma) \leq \textstyle\sum_{i=1}^{k} \mathrm{qs}_d(\sigma_{U_i}) + \overline{Q},$$

where Q is the number of comparisons of elements of σ_{U_i} with elements of $\sigma_{\{1,\ldots,n\}\setminus U_i}$ for any i and the random variable $\overline{Q}$ is defined analogously for $\overline{\sigma}$.

The proof goes along the same lines as the proof of Lemma 4.7 and is omitted.

Lemma 5.3. *For every sequence σ and all $d \geq 1/n$, we have*

$$\mathbb{E}\big(\mathrm{qs}_d(\sigma)\big) \in O\big(\tfrac{n}{d+1}\sqrt{n/d} + n \log n\big).$$

Proof. Given a sequence σ, first observe the quicksort will make at most $O(n\sqrt{n/d} + n \log n)$ comparisons, which follows directly from Lemma 4.9 and the observation that $\mathrm{qs}(\overline{\sigma}) \leq n \cdot \mathrm{height}(\overline{\sigma})$ for every sequence $\overline{\sigma}$: Every level of recursion of quicksort contributes at most $n-1$ comparisons, and we have $\mathrm{height}(\overline{\sigma})$ levels of recursion. Thus, the claim of the theorem is correct for $d \in O(1)$.

Let us now consider the case $d \in \omega(1)$. Furthermore, we assume that $d \in O\big(\sqrt[3]{n/\log^2 n}\big)$. This is no restriction since we obtain the average-case bound of $O(n \log n)$ already for $d \in \Theta\big(\sqrt[3]{n/\log^2 n}\big)$, thus also for larger d.

Similar to the proof of Lemma 4.9, we divide the sequence $\overline{\sigma}$ into three parts. The set $R = \{i \in \{1, \ldots, n\} \mid \overline{\sigma}_i \in [1, d]\}$ of regular elements for the interval $[1, d]$ is defined as before. The set F is defined slightly differently, namely as $F = \{i \in \{1, \ldots, n\} \mid \nu_i \leq 3\}$. This means that F contains all i for which ν_i is too small, plus some extra elements. Similarly $B = \{i \in \{1, \ldots, n\} \mid \nu_i \geq d - 3\}$.

As in Lemma 4.9, the regular elements are easy to handle since they are uniformly distributed in $[1, d]$ and, thus, $\mathbb{E}\big(\mathrm{qs}_d(\sigma_R)\big) \in O(n \log n)$.

We have $\mathbb{E}\big(\mathrm{height}_d(\sigma_F)\big) = \mathbb{E}\big(\mathrm{height}_d(\sigma_B)\big) \in O(\sqrt{n/d})$, which follows from the same argument as the one used in Lemma 4.9: The probability that σ_F contains more than $6n/d$ elements is at most $(e/4)^{3n/d} \in O\big((e/4)^{\sqrt{n}}\big)$ due to the Chernoff bound and $d \in O\big(\sqrt[3]{n/\log^2 n}\big)$. The same holds for σ_B. If either contains more element, we bound the height by n, which contributes at most $o(1)$ to the expectation. Otherwise, we have sequences with $O(n/d)$ elements that are perturbed with a perturbation parameter of 3. We obtain

$$\mathbb{E}\big(\mathrm{qs}(\overline{\sigma}_F)\big) = \mathbb{E}\big(\mathrm{qs}(\overline{\sigma}_B)\big) \in O\big(\mathbb{E}\big(\mathrm{height}_3(\overline{\sigma}_F)\big) \cdot n/d\big) \subseteq O\big(\tfrac{n}{d}\sqrt{n/d}\big).$$

By Lemma 5.2, what remains to be estimated is the number of comparisons of elements $\overline{\sigma}_i$ and $\overline{\sigma}_j$ where i and j are in two different sets of R, F, and B.

Due to the symmetry between $\overline{\sigma}_F$ and $\overline{\sigma}_B$, it suffices to restrict ourselves to estimating the number of comparisons of elements in $\overline{\sigma}_F$ with elements in $\overline{\sigma}_R$ and $\overline{\sigma}_B$. This boils down to count the number of comparisons of elements $\overline{\sigma}_i$ with $\nu_i \leq 3$ to elements $\overline{\sigma}_j$ with $\overline{\sigma}_j \geq 1$.

The number of comparisons between elements $\overline{\sigma}_i$ and $\overline{\sigma}_j$ with $i \in F$ and $j \in F \cap R$ can be bounded by the total number of comparisons between elements in F, but this number is $\mathbb{E}\big(\mathrm{qs}(\overline{\sigma}_F)\big) \in O\big(\frac{n}{d}\sqrt{n/d}\big)$. Similarly, since $\mathbb{E}\big(\mathrm{qs}(\overline{\sigma}_R)\big) \in O(n \log n)$, the expected number of comparisons between positions $i \in F \cap R$ and $j \in R$ is at most $O(n \log n)$.

Thus, we can concentrate on $i \in F$ with $\nu_i \leq 1$ and $j \in R$ with $\overline{\sigma}_i \geq 3$, which includes all $i \in F \setminus R$ and $j \in R \setminus F$.

We distinguish two cases: First, we estimate the expected number of such comparisons with $\overline{\sigma}_i$ being the pivot element. Second, we consider the case that $\overline{\sigma}_j$ is the pivot element.

The two elements $\overline{\sigma}_i \leq \sigma_i + 1 \leq 2$ and $\overline{\sigma}_j \geq 3$ will be compared with $\overline{\sigma}_i$ being the pivot only if $i < j$ and $\overline{\sigma}$ contains no element $\overline{\sigma}_k \in [\overline{\sigma}_i, \overline{\sigma}_j]$ for $k < i$. In particular, $\overline{\sigma}$ must not contain an element $\overline{\sigma}_k \in [2, 3]$ with $k < i$.

Since $d \in \omega(1)$, every element is eligible for the interval $[2, 3]$. Furthermore, for every $i \in \{1, \ldots, n\}$, we have $\mathbb{P}(\nu_i \leq 1) = \mathbb{P}(\overline{\sigma}_k \in [2, 3]) = 1/d$ and these two events are disjoint. (If $\sigma_i = 1$, then this is not true since it might be $\nu_i = 1$. However, the probability of this is 0.) Thus, the probability that $\overline{\sigma}$ contains more than $O(\log n)$ elements with $\nu_i \leq 1$ prior to the first element $\overline{\sigma}_k \in [2, 3]$ is $O(1/n)$. If this happens nevertheless, we bound the number of comparisons by the trivial upper bound of n^2, which contributes only $O(n^2 \cdot 1/n) = O(n)$ to the expected value.

Otherwise, at most $O(\log n)$ elements $\overline{\sigma}_i$ with $\nu_i \leq 1$ are compared to elements $\overline{\sigma}_j$ with $\overline{\sigma}_j \geq 3$ with $\overline{\sigma}_i$ being the pivot, which contributes $O(n \log n)$ comparisons.

Now we consider the second case: How many comparisons of elements $\overline{\sigma}_j \geq 3$ with elements $\overline{\sigma}_i \leq \sigma_i + 1$ with $\overline{\sigma}_j$ being the pivot element do we have to expect? The element $\overline{\sigma}_j$ is compared to $\overline{\sigma}_i$ only if $j < i$ and there is no $k < j$ with $\overline{\sigma}_k \in [\overline{\sigma}_i, \overline{\sigma}_j]$. Thus, it is necessary that $\overline{\sigma}_j$ is the minimal among all elements $\overline{\sigma}_k \geq 3$ with $k \leq j$.

If we restrict ourselves to $\overline{\sigma}_k \in [3, d]$, then this corresponds just to the average number of left-to-right *minima*, which is $O(\log n)$. (The average number of left-to-right minima is equal to the average number of left-to-right maxima.) Thus, the expected number of elements $\overline{\sigma}_j \in [3, d]$ that, when being the pivot element, are compared to any element $\overline{\sigma}_i \leq \sigma_i + 1$, is $O(\log n)$. This contributes at most $O(n \cdot \log n)$ to the expected number of comparisons.

Elements $\overline{\sigma}_k \geq d$ remain to be considered. Since $d \in \omega(1)$, there are at most $O(\log n)$ such elements prior to the first element of the interval $[3, d]$ with high probability. Furthermore, there are at most $O(\log n)$ elements of $\overline{\sigma}_F$ prior to the first element of $[1, d]$ with high probability. Thus, the contribution to the number of comparisons is only $O(\log^2 n)$. □

5.2 Lower Bound on the Smoothed Number of Quicksort Comparisons

Now we show that the upper bound proved in the previous section is tight. The standard sorted sequence provides a worst case, but in the following lemma we use a sequence that is slightly easier to handle technically.

Lemma 5.4. *For the sequence $\sigma = (1/n, 2/n, 3/n, \ldots, \frac{n}{2}/n, 1, 1, \ldots, 1)$ and all $d \geq 1/n$, we have*

$$\mathbb{E}\big(\mathrm{qs}_d(\sigma)\big) \in \Omega\big(\tfrac{n}{d+1}\sqrt{n/d} + n \log n\big).$$

Proof. In the perturbed sequence $\overline{\sigma}$ the first $n/2$ elements contain an expected number of $\Omega\big(\sqrt{n/d}\big)$ left-to-right maxima according to Lemma 3.4. Every left-to-right maximum $\overline{\sigma}_i$ of $\overline{\sigma}$ has to be compared to all the elements that come later and are greater than $\overline{\sigma}_i$.

If $d \in o(1)$, all $n/2$ elements of the second half of $\overline{\sigma}$ are greater than any left-to-right maximum of the first half of $\overline{\sigma}$. Thus, the expected number of comparisons is at least $\Omega\big(n\sqrt{n/d}\big) = \Omega\big(\frac{n}{d+1}\sqrt{n/d} + n \log n\big)$.

If $d \in \Omega(1)$, then the probability that an element $\overline{\sigma}_i$ of the second half of $\overline{\sigma}$ is greater than all left-to-right maxima of the first half of $\overline{\sigma}$ is

$$\mathbb{P}\big(\forall j \leq n/2 \colon 1 + \nu_i \geq \overline{\sigma}_j\big) \geq \mathbb{P}\big(1 + \nu_i \geq 1/2 + d\big) = \frac{1}{2d}.$$

Thus, the expected number of elements that are greater than all left-to-right maxima of the first half is $\Omega\big(n/d\big)$. Multiplying this with the expected number of left-to-right maxima of the first half yields that at least an expected number of $\Omega\big(\frac{n}{d}\sqrt{n/d}\big) \subseteq \Omega\big(\frac{n}{d+1}\sqrt{n/d}\big)$ comparisons are needed. Since quicksort always needs at least $\Omega(n \log n)$ comparisons, we get the claim. □

6 Conclusion

We have analyzed the smoothed height of binary search trees and the smoothed number of comparisons made by the quicksort algorithm under additive noise. The smoothed height of binary search trees and also the smoothed number of left-to-right maxima are $\Theta(\sqrt{n/d} + \log n)$; the smoothed number of quicksort comparisons is $\Theta(\frac{n}{d+1}\sqrt{n/d} + n \log n)$.

While we obtain the average-case height of $\Theta(\log n)$ for binary search trees only for $d \in \Omega(n/\log^2 n)$ – which is large compared to the interval size $[0, 1]$ from which the numbers are drawn –, for the quicksort algorithm already $d \in \Omega\big(\sqrt[3]{n/\log^2 n}\big)$ suffices so that the expected number of comparisons equals the average-case number of $\Theta(n \cdot \log n)$. On the other hand, the recursion depth of quicksort, which is equal to the tree height, can be as large as $\Omega\big(\sqrt{n/d}\big)$. Thus, although the average number of comparisons is already obtained for $d \in \Omega\big(\sqrt[3]{n/\log^2 n}\big)$, the recursion depth remains asymptotically larger than its average value for $d \in o\big(n/(\log n)^2\big)$.

A natural question arising from our results is, what happens when the noise is drawn according to distributions other than the uniform distribution? In a more general additive noise model, the adversary can not only specify the sequence σ, but also a density function f according to which the noise is drawn. We conjecture that if $\max_{x \in \mathbb{R}} f(x) = \phi$, then the expected tree height and the expected number of left-to-right maxima are $\Theta(\sqrt{n\phi} + \log n)$ while the expected number of quicksort comparisons is $\Theta\big(\frac{\phi n}{\phi+1}\sqrt{n\phi} + n \log n\big)$. These bounds would be in compliance with our bounds for uniformly distributed noise, where $\phi = 1/d$.

References

[1] Cyril Banderier, René Beier, and Kurt Mehlhorn. Smoothed analysis of three combinatorial problems. In Branislav Rovan and Peter Vojtáš, editors, *Proc. of the 28th Int. Symp. on Mathematical Foundations of Computer Science (MFCS)*, volume 2747 of *Lecture Notes in Computer Science*, pages 198–207. Springer, 2003.

[2] Michael Drmota. An analytic approach to the height of binary search trees II. *Journal of the ACM*, 50(3):333–374, 2003.

[3] Michael Drmota. Profile and height of random binary search trees. *Journal of the Iranian Statistical Society*, 3(2):117–138, 2004.

[4] James Allen Fill and Svante Janson. Quicksort asymptotics. *Journal of Algorithms*, 44(1):4–28, 2002.

[5] Donald E. Knuth. *Sorting and Searching*, volume 3 of *The Art of Computer Programming.* Addison-Wesley, 2nd edition, 1998.

[6] Bodo Manthey and Rüdiger Reischuk. Smoothed analysis of binary search trees. *Theoretical Computer Science*, to appear. A preliminary version appeared in *Proc. of the 16th Int. Symp.*

on Algorithms and Computation (ISAAC), vol. 3827 of *Lecture Notes in Computer Science*, pp. 483–492, Springer, 2005.

[7] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[8] Bruce Reed. The height of a random binary search tree. *Journal of the ACM*, 50(3):306–332, 2003.

[9] John Michael Robson. Constant bounds on the moments of the height of binary search trees. *Theoretical Computer Science*, 276(1–2):435–444, 2002.

[10] Robert Sedgewick. The analysis of quicksort programs. *Acta Informatica*, 7(4):327–355, 1977.

[11] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3):385–463, 2004.

[12] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms and heuristics: Progress and open questions. In Luis M. Pardo, Allan Pinkus, Endre Süli, and Michael J. Todd, editors, *Foundations of Computational Mathematics, Santander 2005*, pages 274–342. Cambridge University Press, 2006.