# PROBABILISTIC BISECTION SEARCH FOR STOCHASTIC ROOT-FINDING

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Rolf Waeber

January 2013

# PROBABILISTIC BISECTION SEARCH FOR STOCHASTIC ROOT-FINDING

Rolf Waeber, Ph.D.

Cornell University 2013

The goal of a stochastic root-finding algorithm is to locate a point $x^*$ such that $g(x^*) = 0$ for a given function $g$ that can only be observed with noise. In this thesis we investigate the performance of the Probabilistic Bisection Algorithm (PBA), which is a one-dimensional stochastic root-finding algorithm motivated by the well-known bisection search method. In each step, the PBA queries the function $g$ as to whether the root lies to the left or right of a prescribed point $x$. Due to observational noise, the answer to each query has probability $1 - p(x)$ of being incorrect. To account for such possibilities of incorrect observations, the algorithm updates in each iteration a probability density that represents, in some sense, one's belief about the true location of the root $x^*$. The PBA was first introduced in Horstein (1963) under the setting where $p(\cdot)$ is constant and known. While the method works extremely well in this case, very little is known to date about its theoretical properties or potential extensions beyond the current setting for $p(\cdot)$.

The first part of this thesis provides several key findings about the PBA where $p(\cdot)$ is constant and known. Collectively, they lead to the first main conclusion that the expected absolute residuals of successive search results converge to 0 at a geometric rate.

In the second part, we consider the case where $p(\cdot)$ is unknown and varies with $x$. At each query point, the function $g$ is evaluated repeatedly until a lower bound on the probability of obtaining a correct updating signal is achieved. We first

construct a true confidence interval for $x^*$ and prove that its length converges to 0 in the number of query points at a geometric rate. Next, we show that, provided a reasonable conjecture holds, the PBA can be used to construct a sequence of estimators $(\hat{X}_T)_T$ such that the expected absolute residuals $\mathbb{E}[|\hat{X}_T - x^*|]$ converge to 0 at the rate $O(T^{-1/2+\varepsilon})$ for any $\varepsilon > 0$, where $T$ is the number of overall function evaluations. This rate is only slightly slower than $O(T^{-1/2})$, which is the well-established upper bound on the convergence rate of stochastic root-finding problems.

## BIOGRAPHICAL SKETCH

Rolf Waeber was born in Bern, Switzerland on August 26, 1982. At the age of five, he moved with his family to the canton Wallis, where he enjoyed riding his snowboard down the beautiful Swiss Alps—sometimes even on a school day if the snow was fresh and perfect.

After high school, he moved to Zurich to pursue advanced studies at ETH, where he obtained both his Bachelor's degree and Master's degree with distinction in Mathematics. In addition to knowledge in probability spaces and partial differential equations, he collected vivid memories of the daily dinner gatherings, the summer BBQs, and the game nights at his shared apartment in Schwamendingen.

Upon his graduation from ETH, Rolf decided that it is time to leave the "safe-haven" Switzerland and soon joined the School of Operations Research and Information Engineering at Cornell University as a Ph.D. student. Under the guidance of his advisors Peter Frazier and Shane Henderson, he completed his dissertation in the field of Applied Probability, with a focus on simulation-optimization algorithms. Life in Ithaca was packed with fulfilling hard-working days, but never lacked fun and other enjoyable moments (especially after meeting his soon-to-be wife Sophia).

Ever since his first visit to New York City at the age of thirteen, Rolf had wanted to live in this charismatic city for an extended period of time. He is excited to finally realize this dream as he embarks on a career in the financial industry after receiving his doctorate degree from Cornell University.

To my parents:

Reinhard and Esther Waeber-Kalbermatten

## ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors Peter Frazier and Shane Henderson for their guidance and constant support throughout my time at Cornell University. Our weekly meetings were always filled with delightful discussions, where they offered both fresh ideas and necessary tools that helped me overcome various roadblocks in my research. In the meantime, they encouraged me to develop my own thoughts and taught me to be persistent in pursuing my goals. As a Ph.D. student, I really cannot ask for better advisors than them.

I am very grateful to Robert Jarrow and Michael Todd for serving as members of my special committee. They have led two of the greatest classes I took during my graduate studies and helped broadening my academic horizon in many different ways. Furthermore, I would like to thank the entire faculty at the School of Operations Research and Information Engineering. Not only did I have the privilege to attend many interesting classes given by these world-renowned researchers, the doors to their offices were always open for anyone interested in a stimulating discussion. Special thanks go to James Renegar for serving as a proxy during my B-exam, to Stefan Weber for guiding me through an early research project, and to Sasha Stoikov at Cornell Financial Engineering Manhattan for our fruitful collaboration.

Reaching back to my undergraduate studies, I would like to acknowledge my former advisor Paul Embrechts at ETH Zurich, who was truly an inspiration to a young mathematican like me. I would also like to thank Parthanil Roy for his great advices during my master's studies, Peter Bühlmann for several helpful discussions, Lorenz Götte for having me as his research assistant for many years, Klaus Düllmann for the opportunity to work as a visiting research at the Deutsche Bundesbank, as well as Jacob Loveless for providing me with invaluable work

experience in the financial industry.

I would like to thank Gabriel, Brad, Matt, Dmitriy, Martin, Zach, Joyjit, Matthias, Jake, James, Baldur, Tim, Margarita, Younes, Johannes, Collin, Damla, and many other friends and colleagues at Cornell University for making my time as a graduate student fun and memorable. I am also grateful for many wonderful relationships I was able to maintain with friends from Switzerland. The visits, e-mails and Skype conversations allowed me to remain well-connected with my Swiss roots while studying in the States.

Finally, I would like to thank my family: my parents, whose love and support have always been steady and unconditional; my two brothers, Aurel and David, with whom I share many colorful memories; and my fiancée Sophia, who has filled my life with joy and aspirations.

# TABLE OF CONTENTS

CHAPTER 1

# INTRODUCTION

## 1.1 Motivation

A root of a function $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is a point $x^* \in \mathbb{R}^{d_1}$ such that $g(x^*) = 0$. Locating roots of a given function has always been a fundamental problem in mathematics. For example, solving a system of equations is a root-finding problem. Meanwhile, in unconstrained optimization, both global and local extreme points of an objective function $f$ must be roots of its gradient $g = \nabla f$ (given the gradient exists everywhere).

Stochastic root-finding refers to problems where the value of the function $g$ can only be estimated via certain procedures, such as a stochastic simulation. For example, when evaluating $g$ at a point $x$ the response might be $g(x) + \epsilon$, where $\epsilon$ is some stochastic noise term with zero mean. This type of problems appear routinely in many application areas, including sequential statistics (e.g., Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952; Frees and Ruppert, 1990), simulation optimization (e.g., Fu et al., 2005), machine learning (e.g., Bottou, 2004), financial engineering (e.g., Ehrlichman and Henderson, 2007; Cont and Kukanov, 2012), risk management (e.g., Dunkel and Weber, 2010), etc.

In this thesis, we analyze and extend the Probabilistic Bisection Algorithm (PBA) for one-dimensional stochastic root-finding problems ($d_1 = d_2 = 1$). In contrast to popular stochastic root-finding algorithms based on steepest-descent methods (see, for example, survey papers Lai, 2003 or Pasupathy and Kim, 2011, as well as references within), the PBA derives its search mechanism from the well-known bisection method. At each iteration, the algorithm queries the function $g$ at a prescribed point $x$ in an attempt to determine whether the root $x^*$ lies to the left or to the right of $x$. While the stochastic nature of the function evaluations results in incorrect responses with probability $1 - p(x)$, the PBA accounts for such observational noise by updating after each iteration a probability density function on the domain of $g$, which represents, in some sense, one's current belief of the true location of $x^*$. The median of this density provides an estimate of $x^*$ as well as the next query point. The PBA was first introduced in Horstein (1963) under the setting where $p(\cdot)$ is constant and known. Although the algorithm works extremely well for this specific class of stochastic root-finding problems (see, for example, Castro and Nowak, 2008a), very little is known about its theoretical properties or how it might be extended to cover cases where $p(\cdot)$ is nonconstant and unknown.

In Chapter 2 of this thesis we use a Bayesian approach to prove a set of convergence results under the setting where $p(\cdot)$ is constant and known. To this end, we assume that the root is a realization of a random variable $X^*$ with positive density over the domain of $g$. The updating procedure of the PBA, in turn, corresponds to proper Bayesian updating, and techniques from Bayesian analysis can be used to investigate the convergence behavior of the algorithm. More specifically, we show

1. that the PBA is optimal in reducing expected posterior entropy;
2. that $X_n \to X^*$ almost surely as $n \to \infty$, where $(X_n)_n$ corresponds to the sequence of medians generated by the PBA;
3. that the expected absolute residuals $\mathbb{E}[|X_n - X^*|]$ converge to 0 at a geometric rate.

The last result constitutes the main conclusion of Chapter 2. It shows that when $p(\cdot)$ is known and constant, the rate of convergence for the PBA is faster than any polynomial rate and comparable to that of the noise-free bisection search, that is, $O(2^{-n})$[1].

As the PBA with known and constant $p(\cdot)$ has only limited applications in stochastic root-finding, we consider in Chapter 3 a more generalized setting where $p(\cdot)$ is unknown and varies with $x$. In this case, the updating procedure of the PBA no longer corresponds to proper Bayesian updating. Consequently, we rely on frequentist methods to analyze the algorithm's performance, which assumes that the root $x^*$ is a fixed unknown value in the domain of $g$.

Since $p(\cdot)$ is now unknown to us, in order to carry out the updating procedure of the PBA, we first construct a new direction signal $\widetilde{Z}(x)$ by evaluating the underlying function $g$ several times at any prescribed point $x$. While this signal will be correct only with probability $\tilde{p}(x)$, whose exact value remains unknown, we are able to specify a useful lower bound on $\tilde{p}(\cdot)$ that can be used in the updating procedure. Meanwhile, the construction of $\widetilde{Z}(x)$ naturally introduces two time scales into the existing algorithm, namely, a *macro time* $n$ counting the number of updating steps and a *wall-clock time* $T$ counting the total number of function evaluations. We denote with $(X_n)_n$ the sequence of measurement points in macro time, and, with a slight abuse of notation, write $(X_T)_T$ to denote the sequence of measurement points in wall-clock time. The main results in Chapter 3 show

1. that $X_n \to x^*$ almost surely as $n \to \infty$, and $X_T \to x^*$ almost surely as $T \to \infty$;

[1] $f(x) = O(g(x))$ means that $\limsup_{x\to\infty} |f(x)/g(x)| < \infty$, and $f(x) = o(g(x))$ means that $\lim_{x\to\infty} |f(x)/g(x)| = 0$.

2. that a $(1-\alpha)$-confidence interval for $x^*$ can be constructed, whose length converges to 0 at a geometric rate in macro time;
3. that, based on $(X_n)_n$, a sequence of estimators $(\hat{X}_n)_n$ can be constructed that converges to $x^*$ at a geometric rate in macro time;
4. that, in wall-clock time, the sequence of absolute residuals $(|X_T - x^*|)_T$ converges to 0 at a rate slower than $O(T^{-1/2})$;
5. that, based on $(X_T)_T$, a sequence of estimators $(\hat{X}_T)_T$ can be constructed such that $\mathbb{E}[|\hat{X}_T - x^*|] = O\left(T^{-1/2+\varepsilon}\right)$ for any $\varepsilon > 0$, given a reasonable conjecture on the sample paths of $(X_T)_T$ holds true. This conjecture states that the expected absolute residuals of the measurement points $(X_n)_n$ defined by the PBA converge to 0 at a geometric rate in macro time.

As the main competitor of the PBA, Stochastic Approximation (SA) algorithms can at times attain a convergence rate of $O(T^{-1/2})$ in terms of convergence in distribution. While the asymptotic convergence rate of the PBA might be slightly worse than that of SA-type algorithms, it provides the user with more information on the location of the root in the form of a true confidence interval. In addition, we demonstrate empirically that the convergence behavior of the PBA is robust to the choice of input parameters, whereas finite-time as well as asymptotic behavior of SA-type algorithms strongly depends on a chosen tuning sequence. For these reasons, the PBA provides a very appealing way of solving stochastic root-finding problems both from the theoretical and the practical point of view, and is a novel alternative to SA-type algorithms.

The outline of the thesis is as follows. In Chapter 1, we introduce the stochastic root-finding problem and provide some background material on SA-type algorithms and the PBA. In Chapter 2, we prove a set of theoretical properties for the PBA where $p(\cdot)$ is constant and known. In Chapter 3, we extend our analysis of the algorithm to the case when $p(\cdot)$ is unknown and varies with $x$. In Chapter 4, we conduct a series of numerical studies and compare the empirical performance of the PBA to that of its main competitors. In Chapter 5, we conclude and discuss possible future research directions, including the extension to higher dimensional problems.

The following abbreviations are used throughout the remaining part of the thesis: iid for independent and identically distributed, pdf for probability density function, cdf for cumulative distribution function, PBA for Probabilistic Bisection Algorithm, SA for Stochastic Approximation and DP for dynamic program. In addition, we use $\mathbb{1}\{\cdot\}$ to denote the indicator function, which is 1 if the argument is true and 0 otherwise, and $\log(\cdot)$ refers to the natural logarithm.

## 1.2 The Stochastic Root-Finding Problem

In this section, we present the formulation of the stochastic root-finding problem considered in this thesis and show that for a large class of such problems $O(T^{-1/2})$ is an upper bound on the rate of convergence.

A stochastic root-finding problem considers an unknown function $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, whose value can only be estimated at each prescribed point $x \in \mathbb{R}^{d_1}$. The goal is to locate a set $S^* \subseteq \mathbb{R}^{d_1}$ such that $g(x) = 0$ for all $x \in S^*$. A general form of the problem where $d_1 = d_2 = d$ is given in Pasupathy and Kim (2011) as follows:

- **Given:** A procedure that generates, for any $x \in D \subset \mathbb{R}^d$, an estimator $G_m(x)$ of the function $g : D \to \mathbb{R}^d$ such that $G_m(x) \xrightarrow{d} g(x)$ as $m \to \infty$[2].
- **Goal:** Find a root $x^* \in D$ of $g$, that is, find $x^*$ such that $g(x^*) = 0$, assuming one exists.

For the purpose of this thesis, we focus on the case where $d_1 = d_2 = 1$. While extensions to higher-dimensional problems are also important for many applications, their analysis is beyond the scope of this thesis and is deferred to future research. In addition, we assume that $D = [0, 1]$, which can be generalized to any one-dimensional stochastic root-finding problem on an interval through appropriate scaling and shifting.

Throughout this thesis we assume that there exists a unique $x^* \in [0, 1]$ such that $g(x) > 0$ for all $x < x^*$ and $g(x) < 0$ for all $x > x^*$ (though it is not necessarily the case that $g(x^*) = 0$). At any given point $x$, an evaluation of the function $g$ yields $Y(x) = g(x) + \epsilon$, where $\epsilon$ is a stochastic noise term with zero median and independent of $x$ and previous function evaluations. The assumption of zero median is rather unusual for stochastic root-finding problems, which usually assume that the noise term has zero mean. But, as will become clear later, the zero median assumption is required for the PBA to locate $x^*$. Of course, if the noise distribution is symmetric then the assumption of zero median is equivalent of the standard zero mean assumption. We further make the assumption that $\epsilon$ has a probability density function, which is not necessary for all presented results but convenient (in fact, it is often sufficient if $\epsilon$ has a density around the median).

To locate $x^*$, we evaluate the function $g$ at a set of points $(X_n)_n$, which are chosen sequentially based on information obtained from previous function

[2] The notation $\xrightarrow{d}$ stands for *convergence in distribution.*

evaluations. Let $(Y_n(X_n))_n$ be the corresponding results from the function evaluations; $\mathcal{F}_n = \sigma(X_m, Y_m(X_m) : 0 \leq m \leq n)$ be the $\sigma$-algebra generated by $(X_n)_n$ and $(Y_n(X_n))_n$; and $\mathcal{F}_{-1}$ be the trivial $\sigma$-algebra. Based on information contained in $\mathcal{F}_n$, a practitioner needs to make the following decisions for all $n \in \mathbb{N}_0$[3]:

1. What is the current best estimate $\hat{X}_n$ for $x^*$, where $\hat{X}_n$ is an $\mathcal{F}_n$-measurable random variable.
2. At which point $X_{n+1}$ should one evaluate the function $g$ during the next iteration, where $X_{n+1}$ is again an $\mathcal{F}_n$-measurable random variable.

A set of rules used to generate the above decisions is called a *policy* $\pi$, where the set of all policies is denoted by $\Pi$. If a policy does not take into account any newly acquired information for its decision-generating process it is called a *passive sampling* policy. An example is the policy that evaluates the function $g$ over $[0, 1]$ uniformly at random for all $n \in \mathbb{N}_0$. In contrast, if a policy always takes into account all information available at hand when making its decisions (that is, $X_{n+1}$ is $\mathcal{F}_n$-measurable, but not $\mathcal{F}_{n-1}$-measurable), it is called a *fully-sequential* or *active sampling* policy. For most applications, active sampling schemes usually outperform passive sampling schemes. See, for example, Castro and Nowak (2008b) for a discussion on active versus passive sampling for stochastic root-finding problems on a discretized domain. Both the PBA and the popular SA-type algorithms are fully-sequential policies.

When applying a stochastic root-finding algorithm, the user effectively chooses a specific policy $\pi \in \Pi$. If the maximal number of function evaluations is limited to some $n \in \mathbb{N}$ a possible selection criterion would be to find a policy whose

[3] We use the notation $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

performance is close to

$$\inf_{\pi \in \Pi} \mathbb{E}^{\pi}\left[|\hat{X}_n - X^*|\right], \tag{1.1}$$

where

$$\mathbb{E}^{\pi}\left[|\hat{X}_n - X^*|\right] = \int_0^1 \mathbb{E}^{\pi}\left[|\hat{X}_n - x^*|\right] P_0(dx^*),$$

is the expectation over $(X_m, Y_m(X_m) : 0 \leq m \leq n)$ with respect to the probability measure induced by the policy $\pi$ and a prior probability distribution $P_0$ over $x^*$ (this includes the case of a fixed point $x^*$ by setting $P_0$ as a point mass at $x^*$). In general, it is very difficult, if not impossible, to find a policy that solves the above optimization problem. As a result, it is helpful to produce a relative ordering on the set of available policies, where policy $\pi^1$ outperforms policy $\pi^2$ if and only if $\mathbb{E}^{\pi^1}[|\hat{X}_n - X^*|] < \mathbb{E}^{\pi^2}[|\hat{X}_n - X^*|]$. While this average performance of absolute residuals is one possible criterion for comparing the performances of two different policies, the task itself remains challenging as other performance measures should also be considered along with different input scenarios and underlying assumptions. See Waeber et al. (2010, 2012a) for a detailed description on how policies can be compared efficiently combining various performance measures in the case of Ranking and Selection procedures.

Nevertheless, in this thesis we will mostly focus on the performance measure $\mathbb{E}^{\pi}[|\hat{X}_n - X^*|]$, and especially its convergence rate as $n \to \infty$. For this it is informative to first specify general upper bounds on this rate. The optimal convergence rates of deterministic root-finding algorithms for the function $g$ naturally provide such bounds for stochastic root-finding algorithms, since the noise in the observations will generally slow down any algorithm. For any fixed point $x^*$, the residuals $|X_n - x^*|$ of the noise-free bisection algorithm converge to 0 at a rate $\Omega(2^{-n})$[4]

[4] $f(x) = \Omega(g(x))$ means that $\liminf_{x \to \infty} |f(x)/g(x)| > 0$.

and hence the expected residuals of the PBA are not expected to converge faster. As we will see, the expected absolute residuals of the PBA attain a geometric rate of convergence in the number of measurement points (macro time), that is, $\mathbb{E}[|\hat{X}_n - X^*|] = O(C^{-n})$ for some $C > 1$. We then use simulated data to estimate the constant $C$ for the case when $p(\cdot) \equiv p$ is constant and known, and observe that indeed $C \uparrow 2$ as $p \uparrow 1$.

In wall-clock time $T$, that is, the total number of function evaluations, the rate might be significantly worse. In fact, for many stochastic root-finding problems $\Omega(T^{-1/2})$ is an upper bound on the rate of convergence for the expected absolute residuals. To see this, assume there exists a stochastic root-finding algorithm that, for a large class of functions $g$, noise distributions $\epsilon$, and fixed points $x^*$, is able to produce sequences $(\hat{X}_T)_T$ such that $\mathbb{E}[|\hat{X}_T - x^*|] = o(T^{-1/2})$. To be competitive this large class should at least include the simple linear function $g(x) = x^* - x$ and the case when $\epsilon \sim N(0, 1)$. For this function, $Y_T(X_T) = (x^* - X_T) + \epsilon_T$ and $(Y_T(X_T) + X_T) = x^* + \epsilon_T$. Therefore, $Y_T(X_T) + X_T \sim N(x^*, 1)$ independently of the chosen search sequence $(X_T)_T$. If now indeed $\mathbb{E}[|\hat{X}_T - x^*|] = o(T^{-1/2})$, then the estimator $\hat{X}_T$ generated by the stochastic root-finding algorithm is asymptotically more efficient for locating $x^*$ than the sample mean of $(Y_i(X_i) + X_i)_{i=0}^{T}$. This contradicts the asymptotic efficiency of the sample mean for normal random variables (see, for example, Theorem 10.1.12 in Casella and Berger, 2002).

These two upper bounds on the rate of convergence, $\Omega(2^{-n})$ in terms of macro time and $\Omega(T^{-1/2})$ in terms of wall-clock time, already give an indication how different the convergence behavior can be in these two time scales. More specifically, when $p(\cdot)$ is constant then the convergence behaviors in the two time-scales are similar, whereas when $p(x) \to 1/2$ as $x \to x^*$ the convergence is much slower in wall-clock time. In Section 3.6, we make the analogy that the former case corresponds to a stochastic root-finding problem with a discontinuity at $x^*$ and the latter case corresponds to the problem where $g$ is continuous at $x^*$.

## 1.3 Stochastic Approximation

In their seminal paper, Robbins and Monro (1951) introduced the Stochastic Approximation (SA) algorithm to solve stochastic root-finding problems. This algorithm uses an iterative search scheme similar to deterministic steepest-descent methods, such as the Newton-Raphson algorithm. Starting with an initial estimate $X_0 \in D$ of $x^*$ the SA algorithm proceeds for $n = 0, 1, 2, \ldots$

$$X_{n+1} = \Gamma_D(X_n + a_n Y_n(X_n)), \tag{1.2}$$

where $\Gamma_D(x)$ is the projection to the feasible set $D$ and $(a_n)_n$ is a tuning sequence. The SA algorithm does not differentiate into two time scales as the PBA does, that is, $n = T$. Over the past 60+ years schemes based on (1.2) have been studied extensively. See, for example, the monograph Kushner and Yin (2003) for an in-depth analysis of SA algorithms. Such SA-type algorithms are the main competitors for any stochastic root-finding algorithm, including the PBA.

Most theoretical results regarding SA-type algorithms specify conditions on the function $g$, the noise term $\epsilon$, and the tuning sequence $(a_n)_n$ in order to attain the convergence rate $O(n^{-1/2})$[5], and even attempt to optimize the limiting constant. For example, it is known that if

[5] Where the convergence is in distribution, not in $L^1$-norm.

(i) the underlying function $g$ satisfies $g'(x^*) < 0$;

(ii) the tuning sequence $(a_n)_n$ satisfies $na_n \to c$ for some $c > 0$; and $c > -1/(2g'(x^*))$;

(iii) the noise term $\epsilon$ satisfies $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^2] < \infty$;

then $n^{1/2}(X_n - x^*) \overset{d}{\to} N(0, V_1)$, where $V_1 > 0$ is a variance term depending on $(a_n)_n$, $g'(x^*)$ and the variance of $\epsilon$; see Kushner and Yin (2003), Chapter 10, for a proof. Assumptions (i) and (iii) constitute reasonable assumptions for many applications, whereas, Assumption (ii) is hard to verify in practice, since it depends on the (unknown) derivative of $g$ at the root $x^*$. To overcome this assumption, Polyak (1990) and Ruppert (1991) suggest considering the averaging sequence $\overline{X}_n = \frac{1}{n+1}\sum_{i=0}^{n} X_i$ of the iterates $X_i$ as an estimator of $x^*$. This method is referred to as Polyak-Ruppert averaging. Polyak and Juditsky (1992) show that if Assumptions (i) and (iii) hold, and additionally

(ii)' the tuning sequence $(a_n)_n$ satisfies $a_n = O(n^{-\delta})$ for some $\delta \in (1/2, 1)$,

then $n^{1/2}\left(\overline{X}_n - x^*\right) \overset{d}{\to} N(0, V_2)$, where $V_2$ only depends on $g'(x^*)$ and the variance of $\epsilon$.

Even though SA-type algorithms are often able to attain the optimal asymptotic rate of convergence (at least in terms of convergence in distribution), for real-world applications they can be unsatisfactory for several reasons:

- They only provide a point estimate of $x^*$ without specifying any further probabilistic guarantee on the accuracy of this estimate, such as provided by a confidence interval. Hsieh and Glynn (2002) suggest restarting the SA algorithm several times to achieve, based on a central limit argument, at least an approximate confidence interval for $x^*$.

- They do not provide the user with a rule that stops the algorithm once a close estimate of $x^*$ is found (also referred to as a stopping rule).
- Without further adjustments, they may lack robustness since the acquired information is reflected only in the current measurement point $X_n$. For example, when the noise distribution has heavy tails, a single extreme observation of $\epsilon$ can divert the path of $(X_n)_n$ far from $x^*$.
- The choice of tuning sequence usually lacks an intuitive interpretation. While the tuning sequence's effect on the asymptotic behavior is well-understood, its choice also heavily influences the algorithm's finite-time behavior. For example, a poorly chosen tuning sequence can cause the algorithm to require a long time until the optimal asymptotic rate of convergence is attained. In order to improve finite-time properties, recent SA-type algorithms adjust the tuning sequence based on the observed search progress (see, for example, Broadie et al., 2011). While this approach might work well in practice, it requires a significant amount of work from the user.

As we will show, the PBA might have a slightly slower asymptotic rate of convergence than SA-type algorithms, but it is able to overcome the above drawbacks to some extent.

To finish this section we make a final remark regarding the special case when the function $g$ is discontinuous at $x^*$. While this case has not been covered extensively in the literature, there exist several real-world applications in which it appears naturally. For example, this case arises in simulation-optimization problems where the underlying optimization problem has a discrete domain, but, by means of linear interpolation of the objective function, is solved as a continuous optimization problem. Lim (2011) proves that in this case, under modest technical assumptions, the expected absolute residuals of SA-type algorithms converge to 0 at a rate $O(T^{-1})$. While this rate is faster than the usual optimal rate $O(T^{-1/2})$, the PBA is able to outperform SA-type algorithms in this case significantly since its expected absolute residuals converge to 0 at a geometric rate (see Proposition 11 in Section 3.6). This geometric rate, however, is at the moment only shown for one-dimensional problems, whereas the asymptotic rate for SA-type algorithms is also known to hold for higher-dimensional problems.

## 1.4 The Probabilistic Bisection Algorithm

In this section, we introduce the PBA as originally stated in Horstein (1963). In the next section, we show how the PBA can be used to solve more general stochastic root-finding problems than the ones considered in Horstein (1963).

The deterministic bisection algorithm halves the search space in every iteration based on whether the sign of the function evaluation is positive or negative. Applying such a method directly to a stochastic root-finding problem will fail almost surely, as a single wrong sign will divert the search from the right path. To account for noise the PBA instead updates a probability density at each step reflecting one's current belief about the location of $x^*$ and its policy is to always measure at the median of this distribution. More specifically, the PBA takes a prior density $f_0$ that is positive on $[0, 1]$ and a constant $p_c \in (1/2, 1)$ (also denote $q_c = 1 - p_c$) as input parameters and then, for $n = 0, 1, 2, \ldots$, iterates as follows:

1. Determine the next measurement point: $X_n = F_n^{-1}(1/2)$, where $F_n$ is the cdf of $f_n$. Note that $X_n$ is uniquely defined since $f_n$ is a density with domain $[0, 1]$.

2. Query the function $g$ at the point $X_n$, to obtain the random variable $Z_n = \text{sign}(Y_n(X_n)) \in \{-1, +1\}$, and define $Z_n = +1$ if $Y_n(X_n) = 0$.

3. Update the density:

$$\text{if } Z_n(X_n) = +1, \text{ then } f_{n+1}(y) = \begin{cases} 2p_c f_n(y), & \text{if } y \geq X_n, \\ 2q_c f_n(y), & \text{if } y < X_n, \end{cases} \tag{1.3}$$

$$\text{if } Z_n(X_n) = -1, \text{ then } f_{n+1}(y) = \begin{cases} 2q_c f_n(y), & \text{if } y \geq X_n, \\ 2p_c f_n(y), & \text{if } y < X_n. \end{cases} \tag{1.4}$$

The updating of the density is very natural: Querying at the point $X_n$ divides the posterior distribution into two regions. The posterior probability mass in the region where $x^*$ is believed to be, as indicated by the noisy function evaluation, is increased and the probability mass in the other region, where $x^*$ is believed not to be, is decreased. Furthermore, at each iteration the median $X_n$ provides a point estimate of the root $x^*$. Figure 1.1 shows a sample path of the density $f_n$ after $n = 0, 1, 2, 3, 50, 100$ for $p_c = 0.6$ and $f_0$ being the uniform distribution over $[0, 1]$.

The PBA discards information of the observed value $Y_n(X_n)$ since it only considers the observed sign. This may seem counterproductive, because the size of $Y_n(X_n)$ might contain additional information about the location of the root $x^*$. As we will see, however, this makes a Bayesian-motivated update tractable, and the resulting algorithm produces a more robust estimator of $x^*$, especially when the noise $\epsilon$ is heavy-tailed.

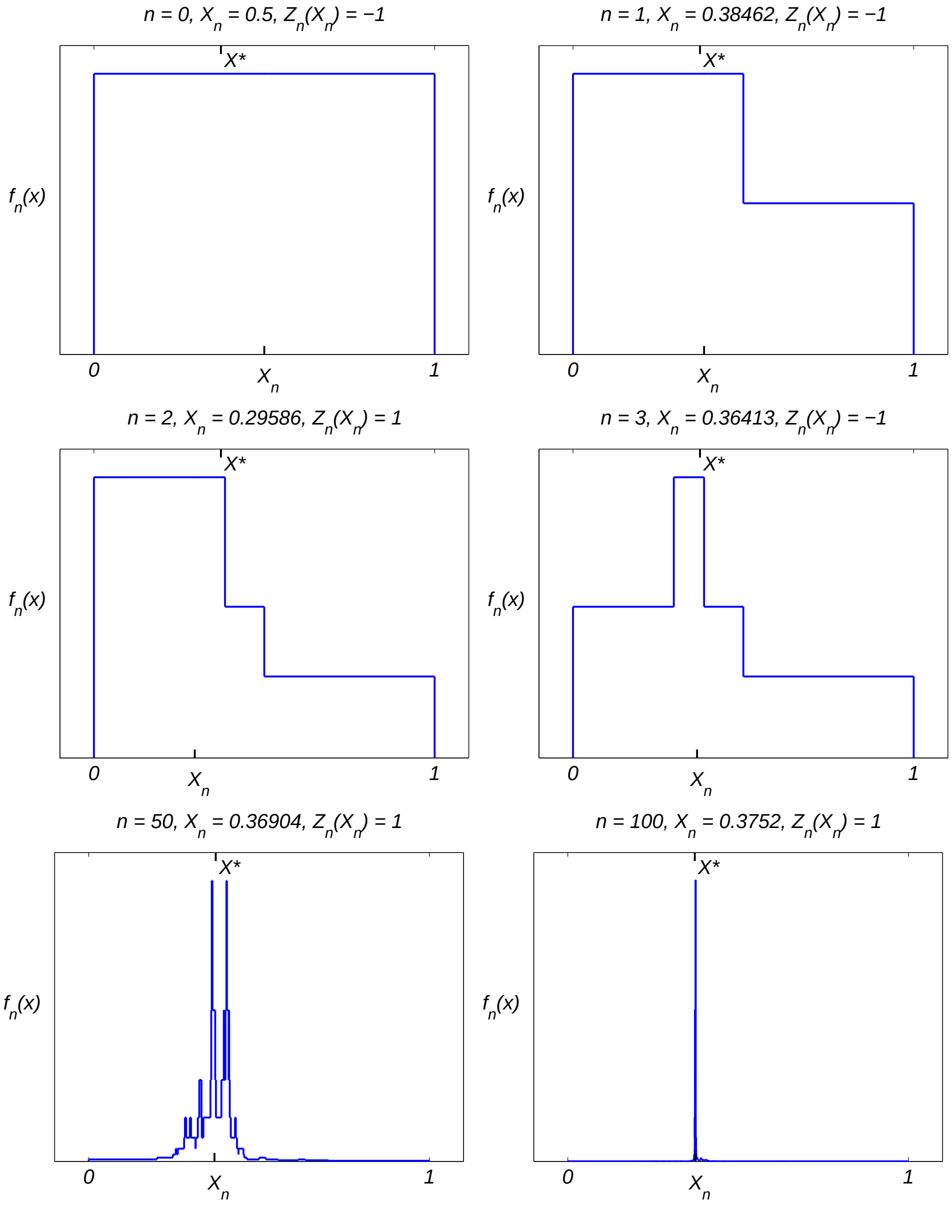

Figure 1.1: The density $f_n$ at time points $n = 0, 1, 2, 3, 50, 100$ on a sample path with input parameters $p_c = 0.6$ and $f_0 = \mathbb{1}\{x \in [0, 1]\}$. The piecewise constant line depicts the posterior density $f_n$. The point $x^*$ is shown on the top of the figures and $X_n$ is shown on the $x$-axis. Above every plot the query point $X_n$ and the observed (noisy) sign $Z_n(X_n)$ are given. Here, the probability of observing a correct sign is $p_c = 0.6$, and is independent of the measurement point $X_n$. The posterior density appears to converge to a point mass at $x^*$ and the sequence $X_n$ seems to converge to $x^*$.

Let us now discuss the input parameters $f_0$ and $p_c$. If some prior knowledge about the root is known, this can be reflected in the prior distribution $f_0$, otherwise the uniform distribution over $[0,1]$ provides a natural choice. In order to discuss the parameter $p_c$, define for $x \in [0,1]$ the function

$$p(x) = \begin{cases} \mathbb{P}\left(Z(x) = +1\right), & \text{if } x < x^*, \\ \mathbb{P}\left(Z(x) = -1\right), & \text{if } x \geq x^*, \end{cases}$$

which corresponds to the *probability of observing a correct sign* when the function $g$ is evaluated at the point $x$. Later, we also use the function $q(\cdot) = 1 - p(\cdot)$. If $p(\cdot) \equiv p$ is constant and known, which is the setting Horstein (1963) considered, then Bayes' rule states that $p_c$ should be chosen as $p$ in order to obtain a proper Bayesian updating for the posterior density process $(f_n)_n$ (see Lemma 8 in Appendix A for a formal proof of this). For the remainder of the current section we discuss this setting. In Section 1.5 we then show how the PBA can be used efficiently when $p(\cdot)$ varies with $x$ and is unknown.

The setting of $p(\cdot)$ constant and known corresponds to the case when $g$ is a step function with a single jump at $x^*$, and the noise distribution as well as the jump height are known. While this only covers a small set of possible stochastic root-finding problems, there exist real-world applications where this setting applies, including

1. Transmission over a noisy channel with noiseless feedback (Horstein, 1963): A real number $x^* \in [0,1]$ should be transmitted from a sender to a receiver. Only one bit of information (0's or 1's) can be sent at each iteration and the signal is sometimes wrong due to corruption by noise. In addition, a noiseless feedback loop informs the sender of what has been recorded by the receiver after each iteration. In this setting, the PBA can be used to

efficiently transmit the number $x^*$.

2. Boundary detection with an airborne radar (Castro and Nowak, 2008a): An airplane equipped with a scanner flies over a pre-determined geographical area several times to locate an edge such as a coast line. At each pass-over, the scanner receives an input as to whether the scanned point is water surface or solid ground but the signal can be wrong. The PBA can be used to determine which point should be scanned at each time so that a good estimate of the edge can be obtained.

3. Zone-detection on a hard disk (Zangenehpour, 1993): A hard disk stores each block of data in one of the disk's several zones, which have different transfer rates. The performance of a filesystem can be improved by accounting for such differences explicitly, but in order to do so, one must be able to identify where each zone begins and ends on the disk. Reading a small collection of data at any location of the disk provides a noisy observation of the transfer rate, and thus the zone identity of that section. The PBA can then efficiently determine the exact zone borders.

Discretized versions of the PBA, which divide the domain $[0, 1]$ into a finite number of intervals, have been studied extensively (Burnashev and Zigangirov, 1974; Rivest et al., 1980; Pelc, 1989; Feige et al., 1994; Karp and Kleinberg, 2007; Ben-Or and Hassidim, 2008; Castro and Nowak, 2008a,b; Nowak, 2008, 2009). However, very little is known about the original PBA with continuous search space $[0, 1]$. Castro and Nowak (2008a) conclude in their review paper: "The Probabilistic Bisection Algorithm seems to work extremely well in practice, but it is hard to analyze and there are few theoretical guarantees for it, especially pertaining error rates of convergence."

In Chapter 2 we provide such convergence guarantees for the PBA when $p(\cdot) \equiv p$. The proof techniques rely on a Bayesian analysis of the updating process, and for this we assume that the root is an absolutely continuous random variable with density $f_0$, denoted $X^*$ (as opposed to $x^*$, which denotes the root as a fixed unknown value). The main result of this chapter shows that the expected absolute residuals $\mathbb{E}[|X_n - X^*|]$ converge to 0 at least at a geometric rate, that is, there exists a constant $c > 1$ such that $\mathbb{E}[|X_n - X^*|] = o(c^{-n})$. This implies that for the case $p(\cdot) \equiv p$ the rate of convergence of the PBA is faster than any polynomial rate and is hence comparable to the rate of the noise-free bisection search, which is $O(2^{-n})$. A consequence of this main result is that the PBA is a consistent method for locating $X^*$, that is, the sequence $(X_n)_n$ generated by the PBA converges almost surely to $X^*$.

The most popular discretized version of the PBA is called the BZ algorithm (Burnashev and Zigangirov, 1974). The algorithm splits the search domain $[0, 1]$ into a finite number of intervals, and aims to locate the interval that contains the point $X^*$. It is known (Burnashev and Zigangirov, 1974) that the BZ algorithm converges geometrically in the number of points queried when $p(\cdot)$ is constant. Our results confirm that a similar rate of convergence holds for the original PBA (without discretization), and effectively closes a gap between the theoretical understandings of the original continuous-space algorithm and that of the corresponding discrete-space version. Although the PBA and the BZ algorithm are conceptually similar, the proof techniques used to analyze the PBA are quite different from the proof techniques usually used to study the BZ algorithm. Such new proof techniques become necessary because the BZ algorithm only samples at breakpoints of the pre-defined intervals, whereas the PBA can sample on the whole domain $[0, 1]$.

There are two reasons for preferring the PBA over the BZ algorithm. First, the PBA is a consistent algorithm, in the sense that it maintains a best estimate $X_n$ of the sought-after point $X^*$, and $X_n$ converges to $X^*$ almost surely as $n \to \infty$ (see Corollary 2 in Section 2.3.2). The BZ algorithm, on the other hand, requires a pre-specified precision (the discretization grid), beyond which no better accuracy can be expected. Since the sequence of estimates $X_n$ does not converge to $X^*$ almost surely the BZ algorithm is not consistent. While one can specify any strictly positive precision, this precision must be specified in advance, which can be inconvenient. Although it might be possible to modify the BZ algorithm to make it consistent, for example, by refining the discretization grid during a run, no such extension has been considered in the literature to the best of our knowledge. The second reason for preferring the PBA to the BZ algorithm is that its implementation is easier (see also Castro and Nowak, 2008a,b). For example, at each step the BZ algorithm requires an additional coin flip to decide which endpoint of the interval containing the median should be queried next. Such "splitting" between the discretization points is not necessary for the PBA.

In addition to the main convergence results, we show that the PBA is optimal in reducing the expected posterior entropy. This result has been proven recently in Jedynak et al. (2012) using concepts from information theory, in particular, the mutual information of the responses $(Z_n(X_n))_n$ and $X^*$. In Chapter 2, we adopt a more direct approach, showing that the PBA minimizes expected posterior entropy using fewer concepts from information theory. To do so, we formulate a dynamic program corresponding to the objective of expected posterior entropy, and solve this dynamic program analytically.

The results of Chaper 2 form a journal paper (Waeber et al., 2012b), which is currently under review at the SIAM Journal on Control and Optimization.

## 1.5 The Probabilistic Bisection Algorithm for General Stochastic Root-Finding Problems

In this section, we show how the PBA as introduced in the previous section can be applied to problems where $p(\cdot)$ varies with $x$ and is unknown.

A first relaxation where $p(\cdot)$ is observable but varies with $x$ has been analyzed in Waeber et al. (2011). In this case, the PBA updating given by equations (1.3) and (1.4), where at each step the constant $p_c$ is replaced by $p(X_n)$, can be used to locate $x^*$. This updating, however, no longer corresponds to proper Bayesian updating, since a probabilistic model on the function $p(\cdot)$ would also be needed. Nevertheless, such a variation of the PBA can still be used as a heuristic. While this heuristic provides a consistent method for locating $x^*$, it hinges on the assumption that $p(\cdot)$ is observable at each step.

Let us now consider the case where $p(\cdot)$ varies with $x$ and is not observable. In order to use an updating method as given by (1.3) and (1.4), it is potentially not necessary to know the value of $p(\cdot)$ exactly. Instead, it might be enough to specify a useful lower bound on $p(\cdot)$ (that is, a lower bound larger than $1/2$). This lower bound can then serve as the constant $p_c$ in the updating procedure.

For many stochastic root-finding problems, namely when $g(x) \to 0$ as $x \to x^*$, it follows that $p(x) \to 1/2$ as $x \to x^*$ (since our setting assumes a constant noise distribution with zero median). So a lower bound on $p(\cdot)$ that is bounded away from $1/2$ does not exist. To overcome this, at each iteration we replace the sign $Z_n(X_n)$

with a new signal $\widetilde{Z}_n(X_n)$ which has probability $1 - \tilde{p}(\cdot)$ of being incorrect. By construction of $\widetilde{Z}_n(X_n)$ it further holds that $\tilde{p}(x) \geq p_c$ for all $x \in [0,1] \setminus \{x^*\}$, where $p_c \in (1/2, 1)$ is a constant chosen by the user. The signal $\widetilde{Z}_n(X_n)$ and the constant $p_c$ can then be used in the updating equations (1.3) and (1.4).

Before we explain in detail how to construct this signal $\widetilde{Z}_n$ by the means of statistical tests of power one let us first introduce a reparameterization of the stochastic root-finding problem. Consider an arbitrary point $x \in [0,1] \setminus \{x^*\}$, and define $s(x) = \mathbb{P}\left(Z(x) = +1\right)$ as well as

$$\tilde{g}(x) = 2s(x) - 1. \tag{1.5}$$

The function $\tilde{g}(x)$, which assumes value in the range $[-1, 1]$ reformulates the original stochastic root-finding problem, that is, $\tilde{g}(x) > 0$ if $g(x) > 0$ and $\tilde{g}(x) < 0$ if $g(x) < 0$. A sufficient condition for the existence of a one-to-one relationship between $g$ and $\tilde{g}$ is that the noise distribution has a density and zero median. Hence, under this assumption $\tilde{g}(x)$ is continuous at $x$ if and only if $g(x)$ is continuous at $x$.

Now, we provide details on the construction of the signal $\widetilde{Z}(x)$ at a prescribed point $x \in [0,1]$. By evaluating the function $g$ at $x$ several times, we can observe a sequence of signs $(Z_i(x))_i$ with $\mathbb{E}[Z_i(x)] = \tilde{g}(x)$. The corresponding simple random walk $S_m(x) = \sum_{i=1}^m Z_i(x)$ has drift $\tilde{g}(x) \in [-1, 1]$ and the goal becomes to detect whether the drift of $S_m(x)$ is positive or negative. Sequential tests of power one provide a powerful tool to decide whether the drift $\theta$ of a random walk satisfies the hypothesis $\theta < \theta_0$ versus $\theta > \theta_0$ for some value $\theta_0$ (for our setting $\theta = \tilde{g}(x)$ and $\theta_0 = 0$).

Such a test of power one for the simple random walk $S_m(x)$ is defined by a positive sequence $(k_i)_i$ and a stopping rule $N(\tilde{g}(x)) = \inf\{m \in \mathbb{N} : |S_m(x)| \geq k_m\}$. The test decides that the drift is positive if $S_{N(\tilde{g}(x))}(x) \geq k_{N(\tilde{g}(x))}$, that the drift is negative if $S_{N(\tilde{g}(x))}(x) \leq -k_{N(\tilde{g}(x))}$ and does not make a decision if $N(\tilde{g}(x)) = \infty$. Furthermore, for a chosen confidence parameter $\gamma \in (0,1)$ such a test satisfies $\mathbb{P}(N(\tilde{g}(x)) < \infty) \leq \gamma$ if $\tilde{g}(x) = 0$ and $\mathbb{P}(N(\tilde{g}(x)) < \infty) = 1$ if $\tilde{g}(x) \neq 0$. In Appendix B, we provide details on the construction of tests of power one for different noise distributions, as well as results on the expected hitting time $\mathbb{E}[N(\tilde{g}(x))]$.

If more information on the noise distribution is known, for example, if it is known that $\epsilon \sim N(0,1)$, then a test of power one for this known noise distribution can be used, that is, to detect wether the drift of the random walk $\sum_{i=1}^{m} Y_i(x)$ is positive or negative. While the finite-time properties of a test designed for a specific noise distribution might outperform a test based only on the signs, the asymptotic behavior between these two tests is usually comparable (see asymptotic results in Appendix B). For this reason, and since it is difficult in practice to verify a specific noise distribution $\epsilon$, we recommend to use the test of power one for the simple random walk $S_m(x) = \sum_{i=1}^{m} Z_m(x)$. This approach is rather robust with respect to the noise distribution as it only requires that the noise distribution has zero median.

Assume now that the PBA measures at some point $X_n \neq x^*$ at the $(n+1)$st iteration. The random walk $S_{n,m} = S_{n,m}(X_n) = \sum_{i=1}^{m} Z_{n,i}(X_n)$ is observed until the test of power one terminates. Denote with $N_n = N_n(\tilde{g}(X_n))$ the stopping time of the power one test which is almost surely finite (since $X_n \neq x^*$), and define the new signal

$$\widetilde{Z}_n(X_n) = \begin{cases} +1, & \text{if } S_{n,N_n} > 0, \\ -1, & \text{if } S_{n,N_n} < 0. \end{cases}$$

Furthermore,

$$\begin{aligned}\mathbb{P}\left(\widetilde{Z}_n(X_n) = +1 \,\middle|\, \tilde{g} < 0\right) &= \mathbb{P}\left(S_{n,N_n} > 0, N_n < \infty \mid \tilde{g}(X_n) < 0\right) \\ &= \mathbb{P}\left(\sum_{i=1}^{N_n} Z_{n,i}(X_n) > 0, N_n < \infty \,\middle|\, \tilde{g}(X_n) < 0\right) \\ &\leq \mathbb{P}\left(\sum_{i=1}^{N_n} Z_{n,i}(X_n) > 0, N_n < \infty \,\middle|\, \tilde{g}(X_n) = 0\right) \\ &\leq \gamma/2, \end{aligned} \tag{1.6}$$

where the first inequality follows by a sample path argument and the second inequality by the property that $\mathbb{P}(N_n < \infty | \tilde{g}(X_n) = 0) \leq \gamma$ and since $S_{n,m}(X_n)$ is a symmetric random walk if $\tilde{g}(X_n) = 0$. An analogous argument shows that

$$\mathbb{P}\left(\widetilde{Z}_n(X_n) = -1 \,\middle|\, \tilde{g}(X_n) > 0\right) \leq \gamma/2. \tag{1.7}$$

So, for $x \in [0,1] \setminus \{x^*\}$, with

$$\tilde{p}(x) = \begin{cases} \mathbb{P}\left(\widetilde{Z}_n(x) = +1\right), & \text{if } x < x^*, \\ \mathbb{P}\left(\widetilde{Z}_n(x) = -1\right), & \text{if } x > x^*, \end{cases}$$

it holds that $\tilde{p}(x) \geq 1-\gamma/2$ for all $x \in [0,1]\setminus\{x^*\}$. Define the constant $p_c = 1-\gamma/2$ and it follows that $\tilde{p}(x) \geq p_c$ for all $x \neq x^*$, where $p_c \in (1/2, 1)$ is a chosen constant (since one can choose $\gamma \in (0,1)$ in the construction of the test of power one).

It remains to discuss the case when $X_n = x^*$. In this case the test of power one might not terminate and the event $\widetilde{Z}_n(x^*)$ is not necessarily defined (the search algorithm might stall). From a theoretical point of view this is not convenient since the sequence $(X_n)_n$ is in this case not well-defined for all $n \in \mathbb{N}_0$. In practice, the stalling of the algorithm can be desirable since in this case the point $x^*$ has successfully been located (this, however, is difficult to justify since its impossible test whether $N_n = \infty$). But, the event that $X_n = x^*$ for any finite $n$ is very unlikely in practice. Consider, for example, the case that $x^*$ is a realization of a random variable with a positive density on $[0, 1]$: Then the probability that $X_n = x^*$ for any $n \in \mathbb{N}$ is zero, since $X_n$ can only assume values on a set of cardinality $2^n$. For the above theoretical and practical reasons, from now on we assume that $X_n \neq x^*$ for all $n \in \mathbb{N}_0$ almost surely and as a consequence the sequence $(X_n)_n$ is well-defined.

Using the PBA with signals $(\widetilde{Z}_n(X_n))_n$ instead of $(Z_n(X_n))_n$ introduces two time scales, namely, the *macro time* $n$ corresponding to the number of different measurement points $(X_n)_n$ and the *wall-clock time* $T$ counting the total number of function evaluations.

Intuitively, the closer the current measurement point $X_n$ is to $x^*$, the closer $p(X_n)$ is to $1/2$, and the longer the test of power one requires to terminate. In fact, it holds that $\lim_{\tilde{g}(x)\to 0} \mathbb{E}[N(\tilde{g}(x))] = \infty$. Moreover, the expected hitting time increases at a faster rate than $O(\tilde{g}(x)^{-2})$ as $\tilde{g}(x) \to 0$, that is, $\lim_{\tilde{g}(x)\to 0} \tilde{g}(x)^2 \mathbb{E}[N(\tilde{g}(x))] = \infty$ (see Appendix B for details). So, if $g$, and thus $\tilde{g}$, are continuous at $x^*$ then the number of function evaluations between two macro iterations is likely to become very large, which explains the discrepancy between the convergence behaviors in the two time scales.

In Chapter 3, we provide convergence results for the PBA in terms of macro and wall-clock time, leading to the main conclusion that the asymptotic rate of convergence of an estimator based on the PBA might be slightly slower than the optimal rate $O(T^{-1/2})$. But, given a reasonable conjecture holds true, we show that there exists a sequence of averaged estimators based on the PBA (similar to Polyak-Ruppert averaging) for which the expected absolute residuals converge to 0 at a near-optimal rate $O(T^{-1/2+\varepsilon})$ for any $\varepsilon > 0$. Empirical examples (presented in Chapter 4) suggest that the asymptotic rate of convergence of such averaged estimators in fact is comparable to the asymptotic rate of convergence of SA-type algorithms. In addition, we show that the PBA provides the simulation analyst with useful information on the location of $x^*$, such as a true confidence interval and a stopping rule once a sufficiently close estimate of $x^*$ is located. Furthermore, the PBA provides a robust and novel alternative to SA-type algorithms.

CHAPTER 2

# CONSTANT AND KNOWN PROBABILITY OF CORRECT RESPONSES

## 2.1 Introduction

In this chapter, we provide convergence guarantees for the PBA when $p(\cdot)$ is constant, that is, the sign of a function evaluation is incorrect with probability $1-p$, where $p$ is a constant in $(1/2,1)$. The main result shows that, if $X^*$ is the realization of an absolutely continuous random variable with density $f_0$, then the expected absolute residuals $\mathbb{E}[|X_n - X^*|]$ of the measurement points generated by the PBA converge to 0 at least at a geometric rate, that is, there exists a constant $c > 1$ such that $\mathbb{E}[|X_n - X^*|] = o(c^{-n})$. This implies that the rate of convergence of the bisection search with noisy responses is faster than any polynomial rate and is hence comparable to the rate of the noise-free bisection search, which is $O(2^{-n})$. Since we are considering residuals under the expectation operator, our result provides an average-case performance guarantee for the PBA. A consequence of this main result is that the PBA is a consistent method to locate $X^*$. This means that the sequence $(X_n)_n$ generated by the PBA converges almost surely to $X^*$.

In addition to the main convergence results, we show that the PBA is optimal in reducing the expected posterior entropy. This result has been proven recently in Jedynak et al. (2012) using concepts from information theory, in particular, the mutual information of the responses $(Z_n(X_n))_n$ and $X^*$. In this chapter, we adopt a more direct approach, showing that the PBA minimizes expected posterior entropy using fewer concepts from information theory. To do so, we formulate a dynamic program corresponding to the objective of expected posterior entropy, and solve this dynamic program analytically.

The outline of this chapter is as follows. Section 2.2 shows optimality in terms of minimizing expected posterior entropy, whereas Section 2.3 presents and proves the main convergence result of this chapter.

## 2.2 Optimality in Reducing the Expected Posterior Entropy

The result that the PBA is optimal in reducing the expected posterior entropy has recently been proven in Jedynak et al. (2012) using the mutual information of the noisy function evaluations $(Z_n(X_n))_n$ and $X^*$. In Appendix A, we provide a different and more direct proof of this result that borrows fewer concepts from information theory. This proof relies solely on the dynamic programming principle.

The optimality result, stated in Theorem 1 (see below), uses the entropy to measure the information content of the density $f_n$. For a random variable $\psi$ with density $f$ the entropy is defined as $H(f) = \mathbb{E}[-\log_2 f(\psi)]$. The entropy is the predominant measure of uncertainty in information theory, see, for example, Cover and Thomas (1991). Using this measure of uncertainty and given a fixed simulation budget $N \in \mathbb{N}$, the optimality analysis seeks a policy $\pi$ that minimizes the expected entropy of the posterior distribution at time $N$. Here, a policy refers to the allocation rule of the measurements $X_0, \ldots, X_N$, where $X_{n+1}$ has to be $\mathcal{G}_n$-measurable, and $\mathcal{G}_n = \sigma\left(X_m, Z_m(X_m) : 0 \leq m \leq n\right)$ is the $\sigma$-algebra generated by the measurement points $(X_n)_n$ and noisy responses $(Z_n(X_n))_n$, and $\mathcal{G}_{-1}$ is the trivial $\sigma$-algebra. A generic policy is denoted $\pi$ and the space of all possible policies is denoted $\Pi$. This optimization problem can be solved using a dynamic programming (DP) approach. The value function of the DP for fixed $N \in \mathbb{N}$ is

$$V_n(f_n) = \inf_{\pi \in \Pi} \mathbb{E}^{\pi}[H(f_N)|f_n], \qquad \text{for } n = 0, 1, \ldots, N. \tag{2.1}$$

Any policy $\pi$ induces, together with the input density $f_0$ and the parameter $p$, a distribution on $(X_i, Z_i(X_i))_{i=0}^{N-1}$ and through it a distribution on the sequence of pdfs $(f_i)_{i=0}^{N}$. It is under this distribution that $\mathbb{E}^{\pi}$ is taken, and any policy $\pi^*$ attaining the infimum is called optimal, that is, $\mathbb{E}^{\pi^*}[H(f_N)|f_0] = \inf_{\pi \in \Pi} \mathbb{E}^{\pi}[H(f_N)|f_0]$.

The value function (2.1) satisfies Bellman's recursion,

$$V_n(f_n) = \inf_{\pi \in \Pi} \mathbb{E}^{\pi}[V_{n+1}(f_{n+1})|f_n] = \inf_{x \in [0,1]} \mathbb{E}[V_{n+1}(f_{n+1})|X_n = x, f_n], \tag{2.2}$$

where the last equation follows from the fact that the control of a policy $\pi \in \Pi$ is the point at which to evaluate the function. The DP formulated in (2.1) can be solved explicitly.

**Theorem 1.** *For $N \in \mathbb{N}$, the PBA, which always measures at the median of $f_n$, for $n = 0, \ldots, N-1$, minimizes the expected entropy of the density $f_N$. Furthermore, the expected posterior entropy at time $N$ using the PBA is*

$$\begin{aligned} V_n(f_n) &= \mathbb{E}[H(f_N)|f_n] \\ &= H(f_n) - (N-n)(1 + p\log_2 p + (1-p)\log_2(1-p)), \end{aligned} \tag{2.3}$$

*for $n = 0, \ldots, N$.*

The key step in the proof of Theorem 1 is the analysis of the knowledge-gradient policy to the DP formulated in (2.1). A knowledge-gradient policy is a policy that acts optimally if there is only *one* measurement remaining, that is, when $n = N-1$. See Frazier et al. (2008) for more details on knowledge-gradient policies. For this knowledge-gradient policy the value attained by the infimum is equal to the entropy of $f_n$ minus an additional amount which may be interpreted as the maximum information content of a single measurement. The fact that this amount does not depend on $f_n$ is important in proving that the knowledge-gradient policy in fact is the optimal policy in general when more than just one measurement is remaining. The next proposition shows that the PBA is indeed the knowledge-gradient policy for the problem stated in (2.1).

**Proposition 1.** *For any $N \in \mathbb{N}$,*

$$\begin{aligned}
&\inf_{x\in[0,1]} \mathbb{E}[V_N(f_N)|X_{N-1} = x, f_{N-1}] \\
&\qquad = \inf_{x\in[0,1]} \mathbb{E}[H(f_N)|X_{N-1} = x, f_{N-1}] \\
&\qquad = H(f_{N-1}) - p\log_2 p - (1-p)\log_2(1-p) - 1,
\end{aligned}$$

*and the infimum is achieved by choosing $X_{N-1}$ to be the median of $f_{N-1}$.*

We provide proofs of Theorem 1 and Proposition 1 in Appendix A.

## 2.3 Geometric Rate of $L^1$-Convergence

In this section we present and prove the main result of this chapter, which is that the expected absolute residuals of the PBA converge to 0 at a rate $o(c^{-n})$ for some $c > 1$. This, in particular, implies that the asymptotic rate of convergence is faster than any polynomial rate and is comparable to the rate of convergence of the noise-free bisection algorithm which has rate $O(2^{-n})$. Such a geometric rate of convergence is known to hold for discretized versions (e.g., the BZ algorithm) of the PBA; see Burnashev and Zigangirov (1974), Castro and Nowak (2008a,b). But, to the best of our knowledge, it is a new result for the original PBA.

**Theorem 2.** *There exists a constant $c(p) > 1$ such that $\mathbb{E}[|X_n - X^*|] = o(c(p)^{-n})$, where $(X_n)_n$ is the sequence of query points generated by the PBA.*

Before developing the proof of Theorem 2, we first discuss the constant $c(p)$, introduce some simplified notation and provide a sketch of the proof.

The constant $c(p)$ can be any fixed value in the open interval $(1, C(p))$, where $\log(C(p))$ is the smaller solution to the quadratic equation (2.5) given in Lemma 1 (see below). For the most part, it suffices to know that $C(p)$ is a constant only depending on the parameter $p$, and that $C(p) > 1$. From the rate of convergence of the noise-free bisection algorithm we know that $C(p) \leq 2$. In fact, $C(p)$ is often much smaller than 2 and is usually quite close to 1. This, however, does not necessarily imply that the rate of convergence of the PBA is much slower (in terms of the constant $c(p)$) than the rate of convergence of the noise-free bisection algorithm since our result only provides a lower bound on the rate of convergence. We leave for future work the problem of identifying the exact rate of convergence, but the empirical results in Chapter 4 (Section 4.1.1) supports our expectation that the true rate $\widetilde{C}(p)$, that is, the constant $\widetilde{C}(p)$ such that $\mathbb{E}[|X_n - X^*|] \sim \left(\widetilde{C}(p)\right)^{-n}$, satisfies $\widetilde{C}(p) \uparrow 2$ as $p \uparrow 1$[1].

We now introduce some simplified notation. Define $q = 1 - p$ and $D(p) = (\log(2p) + \log(2q))/2$. The fact that $D(p) < 0$ for $p \in (1/2, 1)$ will be important in the upcoming proofs. From now on we will often simply write $c, C$, and $D$ when the context allows it and keep in mind that all these constants only depend on the parameter $p$. We further denote by $\mathbb{P}_n(\cdot)$ the probability measure defined by the density $f_n$, and by $\mathbb{E}_n[\cdot]$ the expectation under this measure.

[1] Here, the notation $g(x) \sim f(x)$ means that $\lim_{x \to x_0} f(x)/g(x) = a$ for some constant $a > 0$.

*Sketch of Proof of Theorem 2.* The proof of Theorem 2 consists of two major steps. Each is formulated in the next subsection as a separate proposition. In Proposition 5 we show that the stochastic process $\left(c^n \mathbb{E}_n[|X^* - X_n|]\right)_n$ converges to 0 in probability. We then show the uniform integrability of this process in Proposition 6, and Theorem 2 follows from the fact that a sequence of uniformly integrable random variables converges in $L^1$ if and only if it converges in probability.

The key to prove these two propositions is to analyze the stochastic process $\left(\mathbb{E}_n[|X^* - X_n|]\right)_n$. We now give an intuitive outline why this process converges at a geometric rate. All the arguments are made precise in the next subsection.

Using integration by parts it holds that

$$\begin{aligned}\mathbb{E}_n[|X^* - X_n|] &= \int_0^1 \mathbb{P}_n(|X^* - X_n| > h)\, dh \\ &\leq h + \mathbb{P}_n(|X^* - X_n| > h),\end{aligned}$$

for any $h \in (0,1)$. The inequality holds since $\mathbb{P}_n(|X^* - X_n| > h) \leq 1$ and is decreasing in $h$. It is then enough to show that the process $\mathbb{P}_n(|X^* - X_n| > h)$ converges to 0 at a geometric rate and consider the case $h \to 0$. Fix for now an $h \in (0,1)$. At time $n$ there exists an integer $K_n$ such that $X_n \in [(K_n - 1)h, K_n h)$ and

$$\mathbb{P}_n(|X^* - X_n| > h) \leq \mathbb{P}_n(X^* \in [0, (K_n - 1)h)) + \mathbb{P}_n(X^* \in [K_n h, 1]).$$

We then focus on the process $(A_n)_n$, where $A_n = \mathbb{P}_n(X^* \in [0, (K_n - 1)h))$ (the analysis of the process $\mathbb{P}_n(X^* \in [K_n h, 1])$ follows analogously). After querying the function at $X_n$ the quantity $A_n$ is multiplied by either $2p$ or $2q$. Also, since $X_n$ is the median of $f_n$, either multiplication happens with probability 1/2. If we (for now) ignore the fact that $K_n$ depends on $n$ then $A_n$ behaves like a geometric random walk with drift $e^D$ and hence converges to 0 at a geometric rate. This is the basic argument why the geometric rate of convergence holds. Most of the proof is then devoted to the fact that $K_n$ depends on $n$ and is a stochastic process itself. It turns out that $A_n$ is not a true geometric random walk (which can already be seen since $A_n$ is always smaller than 1/2), but that $A_n$ can be dominated by a collection of dependent geometric random walks and each of these random walks has drift $e^D$. Using this dominating argument and results from random walk theory we can then show that the geometric rate of convergence indeed holds for $\mathbb{P}(|X_n - X^*| > h)$. By letting $h \to 0$ this geometric rate also holds for $\mathbb{E}_n[|X_n - X^*|]$, and for $\mathbb{E}[|X_n - X^*|]$ by applying the tower property of conditional expectations.

### 2.3.1 Proof of the Geometric Rate of Convergence

We start with a lemma which is an application of random walk theory. This lemma defines the constant $C$ and will also be useful for later proofs.

**Lemma 1.** *Let $p \in (1/2, 1)$, $q = 1 - p$, and $(R_n)_n$ be a random walk with starting point $R_0 \leq \log(1/2)$ and iid increments $(\psi_n)_n$, that is, $R_n = R_0 + \sum_{j=1}^{n} \psi_j$, with increment distribution $\mathbb{P}(\psi_j = \log(2p)) = \mathbb{P}(\psi_j = \log(2q)) = 1/2$. Then*

$$\mathbb{P}\left(e^{R_n} > C^{-n}/2\right) \leq C^{-2n}, \tag{2.4}$$

*for all $n \in \mathbb{N}$. Here, $C = e^{\tilde{u}}$, where $\tilde{u}$ is the smaller solution to*

$$\left(\frac{u + D}{\log(2p) - \log(2q)}\right)^2 - u = 0. \tag{2.5}$$

*Furthermore, $\tilde{u} > 0$.*

Equation (2.5) is a quadratic equation and it is possible to write down an explicit formula for $C$. However, the explicit form of $C$ is cumbersome and not informative, hence is omitted. The proof of Lemma 1 is given in Appendix A.

The next result, which studies the stochastic process $A_n = \mathbb{P}_n(X^* \in [0, a))$ for some $a \in [0, 1]$, is a first key ingredient to show the geometric rate of convergence. Define $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

**Proposition 2.** *Let $C$ be the constant defined in Lemma 1. For $a \in [0, 1]$ define $A_n = \mathbb{P}_n(X^* \in [0, a)) = \int_0^a f_n(y)dy$. Then*

$$\mathbb{P}(A_n \wedge (1 - A_n) > C^{-n}/2) \leq C^{-2n},$$

*for all $n \in \mathbb{N}$.*

*Proof.* The claim holds trivially for $a = 0$ or $a = 1$ since for all $n \in \mathbb{N}$ the probability measure $\mathbb{P}_n(\cdot)$ has a density.

Now fix an arbitrary $a \in (0, 1)$ and consider the stochastic process $(A_n)_n$. If $A_n \leq 1/2$, then $X_n \geq a$ and $A_n$ will be either multiplied by $2p$ or $2q$ in the next iteration, behaving like an iteration of a geometric random walk. If, on the other hand, $A_n > 1/2$ then $X_n \in [0, a)$ and $A_n$ does not behave like an iteration of a geometric random walk anymore, but $(1 - A_n)$ does. We next make this argument precise. To simplify notation we take logarithms and consider the process $(\log(A_n))_n$. The stochastic driver of this process is the sequence of noisy signs $(Z_n(X_n))_n$. If we condition on the available information up to time $n$, then, by Lemma 8 given in Appendix A, $\eta(X_n) = \mathbb{P}(Z_n(X_n) = +1|\mathcal{G}_{n-1}) = 1/2$ for the PBA. Moreover, the only random source that drives the stochastic process $(Z_{n+1}(X_{n+1})|\mathcal{G}_n)_n$ is the sequence $(Q_n)_n$, a sequence of iid Bernoulli($p$) random variables that determines whether the sign is correct or not, hence the sequence $(Z_{n+1}(X_{n+1})|\mathcal{G}_n)_n$ is itself a sequence of independent random variables. At time $n$, the random variable $\log(A_{n+1})|\mathcal{G}_n$ can be constructed as follows: if $\log(A_n) \leq \log(1/2)$, then

$$\log(A_{n+1})|\mathcal{G}_n = \log(A_n) + \begin{cases} \log(2q), & \text{if } Z_n(X_n) = +1, \\ \log(2p), & \text{if } Z_n(X_n) = -1, \end{cases}$$

and if $\log(A_n) > \log(1/2)$, then

$$\log(1 - A_{n+1})|\mathcal{G}_n = \log(1 - A_n) + \begin{cases} \log(2p), & \text{if } Z_n(X_n) = +1, \\ \log(2q), & \text{if } Z_n(X_n) = -1. \end{cases}$$

Now consider the process $M_n = \log(A_n) \wedge \log(1 - A_n)$. The only times when the dynamics of $(M_n)_n$ are different from a random walk is when it crosses the boundary $\log(1/2)$, that is, when there is a switch from the process $(\log(A_n))_n$ to the process $(\log(1 - A_n))_n$ in the definition of $(M_n)_n$. To overcome this difficulty we construct a true random walk $(S_n)_n$ that is *coupled* with $(M_n)_n$ and dominates $(M_n)_n$.

We first define the coupling sequence

$$W_n = \begin{cases} Z_n(X_n), & \text{if } \log(A_n) > \log(1/2), \\ -Z_n(X_n), & \text{if } \log(A_n) \leq \log(1/2), \end{cases}$$

and then the process

$$S_{n+1} = S_n + \begin{cases} \log(2p), & \text{if } W_n = +1, \\ \log(2q), & \text{if } W_n = -1, \end{cases}$$

for $n \in \mathbb{N}$ and starting point $S_0 = M_0$. The process $(S_n)_n$ is a random walk with iid increments $(\xi_n)_n$ and $\mathbb{P}(\xi_n = \log(2p)) = \mathbb{P}(\xi_n = \log(2q)) = 1/2$.

The processes $(M_n)_n$ and $(S_n)_n$ have the same starting point and are driven by the same sequence of random variables $(W_n)_n$. Assume that $M_0 = \log(1 - A_0)$, and define $\tau = \inf\{n \geq 1 : 1 - A_n \geq 1/2\}$ (if $M_0 = \log(A_0)$ then the definition of $\tau$ and the following arguments can be adapted accordingly). For $n < \tau$ it holds that $M_n = S_n$. At time $\tau$ the processes $\log(1-A_n)_n$ and $(S_n)_n$ increase by $\log(2p)$. On the other hand, the process $(M_n)_n$ switches from being defined by $\log(1-A_n)$ to being defined by $\log(A_n)$ and may increase or decrease, that is,

$$M_\tau - M_{\tau-1} = \log(A_\tau) - M_{\tau-1} \leq \log(1 - A_\tau) - M_{\tau-1} = S_\tau - S_{\tau-1},$$

and hence $M_\tau \leq S_\tau$. (See Figure 2.1.) After time $\tau$ this argument carries over in the following sense: Each time $(S_n)_n$ decreases by $\log(2q)$, then also $(M_n)_n$ decreases by $\log(2q)$. However, when $(S_n)_n$ increases by $\log(2p)$ then $(M_n)_n$ increases by a quantity smaller than or equal to $\log(2p)$ (the increase might also be negative). It follows that $M_n \leq S_n$, and hence $A_n \wedge (1 - A_n) \leq e^{S_n}$ for all $n \in \mathbb{N}$. Then

$$\mathbb{P}(A_n \wedge (1 - A_n) > C^{-n}/2) \leq \mathbb{P}(e^{S_n} > C^{-n}/2) \leq C^{-2n},$$

where the last inequality follows from Lemma 1 since $(S_n)_n$ is a random walk as considered in that lemma. □

We can now use the previous result to bound the probability of observing a large posterior probability mass away from the current best estimate $X_n$.

**Proposition 3.** *Let $C$ be the constant defined in Lemma 1. Then*

$$\mathbb{P}(\mathbb{P}_n(|X_n - X^*| > h) > C^{-n}) \leq h^{-1}C^{-2n}$$

*for all $h \in (0, 1)$ and $n \in \mathbb{N}$.*

*Proof.* Fix an arbitrary $h \in (0, 1)$ and denote $\overline{K} = \lfloor h^{-1} \rfloor$. Define intervals $I(k) = [(k-1)h, kh)$ for $k = 1, \ldots, \overline{K}$ and $I(\overline{K}+1) = [\overline{K}h, 1]$. These $\overline{K}+1$ intervals are pairwise disjoint and cover the domain $[0, 1]$. Further define the stochastic

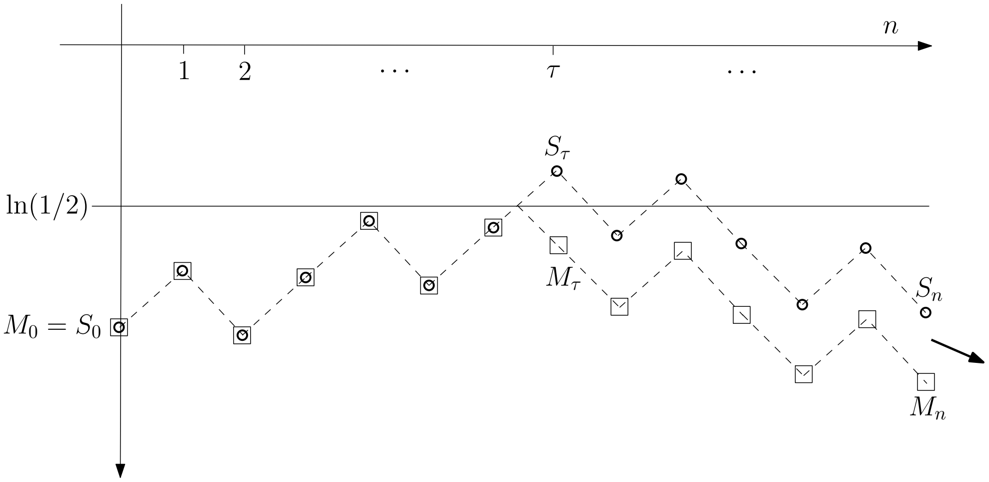

Figure 2.1: The process $(S_n)_n$ (circles) dominates the process $(M_n)_n$ (squares) for all $n \in \mathbb{N}$. The process $(S_n)_n$ is a random walk with negative drift, so by the law of large numbers $S_n \to -\infty$ almost surely as $n \to \infty$, as indicated by the arrow at the right side of the figure. (Both processes are defined in discrete time. We draw a dashed line between time steps for better visibility.)

processes

$$A_n(k) = \mathbb{P}_n\left(X^* \in \bigcup_{j=1}^{k} I(j)\right),$$

for $k = 1, \dots, \overline{K} + 1$ and the trivial process $A_n(0) = 0$ for all $n \in \mathbb{N}$.

At time $n \in \mathbb{N}$ let $K_n$ be the index such that $X_n \in I(K_n)$. Then

$$\begin{aligned}\mathbb{P}_n(|X_n - X^*| > h) &\leq \mathbb{P}_n(X^* \in [0, (K_n - 1)h)) + \mathbb{P}_n(X^* \in [K_n h, 1]) \\ &= A_n(K_n - 1) + (1 - A_n(K_n)) \\ &= [A_n(K_n - 1) \wedge (1 - A_n(K_n - 1))] \\ &\qquad + [A_n(K_n) \wedge (1 - A_n(K_n))],\end{aligned}$$

where the last equation holds since $X_n \in I(K_n)$ implies $A_n(K_n - 1) \leq 1/2$ and $1 - A_n(K_n) \leq 1/2$. The index $K_n$ is a random variable taking values in $\{1, \ldots, \overline{K}+1\}$, hence

$$\begin{aligned}
\mathbb{P}_n(|X_n - X^*| > h) &\leq \max_{k \in \{1,\ldots,\overline{K}+1\}} \Big( [A_n(k-1) \wedge (1 - A_n(k-1))] \\
&\qquad\qquad + [A_n(k) \wedge (1 - A_n(k))] \Big) \\
&\leq \max_{k \in \{1,\ldots,\overline{K}\}} 2\, [A_n(k) \wedge (1 - A_n(k))] \,,
\end{aligned}$$

since $A_n(0) = 0$ and $A_n(\overline{K}+1) = 1$ for all $n \in \mathbb{N}$. Then

$$\begin{aligned}
&\mathbb{P}(\mathbb{P}_n\, (|X_n - X^*| > h) > C^{-n}) \\
&\qquad \leq \mathbb{P}\left( \max_{k \in \{1,\ldots,\overline{K}\}} 2\, [A_n(k) \wedge (1 - A_n(k))] > C^{-n} \right) \\
&\qquad \leq \mathbb{P}\left( \max_{k \in \{1,\ldots,\overline{K}\}} [A_n(k) \wedge (1 - A_n(k))] > C^{-n}/2 \right) \\
&\qquad = \mathbb{P}\left( \bigcup_{k=1}^{\overline{K}} \big\{ [A_n(k) \wedge (1 - A_n(k))] > C^{-n}/2 \big\} \right) \\
&\qquad \leq \sum_{k=1}^{\overline{K}} \mathbb{P}\big( [A_n(k) \wedge (1 - A_n(k))] > C^{-n}/2 \big) \\
&\qquad \leq \overline{K} C^{-2n}.
\end{aligned}$$

The last inequality follows by Proposition 2 since the processes $(A_n(k))_n$ are exactly of the form required for that proposition. Note that $\overline{K} = \lfloor h^{-1} \rfloor \leq h^{-1}$ and the claim follows. □

The next proposition provides an upper bound on $\mathbb{P}\big(c^n \mathbb{E}_n[|X_n - X^*|] > \varepsilon\big)$ for $\varepsilon > 0$ and large $n$. This result is the last step before we can prove convergence in probability and uniform integrability of the stochastic process $\big(c^n \mathbb{E}_n[|X_n - X^*|]\big)_n$, and is also interesting by itself. It provides a large deviation result for the stochastic process $\big(c^n \mathbb{E}_n[|X_n - X^*|]\big)_n$. In contrast to Theorem 2, it provides a finite-time guarantee for large $n \in \mathbb{N}$, instead of an asymptotic convergence guarantee.

**Proposition 4.** *Let $C$ be the constant defined in Lemma 1, $c \in (1, C)$ and $\varepsilon > 0$. Then*

$$\mathbb{P}\big(c^n \mathbb{E}_n[|X_n - X^*|] > \varepsilon\big) \leq C^{-n}, \quad \textit{for } n \geq 0 \vee \widetilde{N}(\varepsilon, c, C),$$

*where*

$$\widetilde{N}(\varepsilon, c, C) = \frac{\log(2/\varepsilon)}{\log(C/c)}. \tag{2.6}$$

*Proof.* Fix $n \geq 0 \vee \widetilde{N}(\varepsilon, c, C)$. Consider

$$c^n \mathbb{E}_n[|X_n - X^*|] = c^n \int_0^1 \mathbb{P}_n(|X_n - X^*| > h)\, dh,$$

which follows from integration by parts of the right-hand side. The random function $\mathbb{P}_n(|X_n - X^*| > h)$ is non-increasing in $h$ and $\mathbb{P}_n(|X_n - X^*| > h) \leq 1$ for all $h \in (0, 1)$. Then for any $h \in (0, 1)$

$$\begin{aligned} c^n \mathbb{E}_n[|X_n - X^*|] &\leq c^n \left(h + (1 - h)\mathbb{P}_n(|X_n - X^*| > h)\right) \\ &\leq c^n \left(h + \mathbb{P}_n(|X_n - X^*| > h)\right). \end{aligned}$$

So we can choose $h = C^{-n} \in (0, 1)$ and get

$$c^n \mathbb{E}_n[|X_n - X^*|] \leq c^n \left(C^{-n} + \mathbb{P}_n\left(|X_n - X^*| > C^{-n}\right)\right).$$

Note that on the event $\{\mathbb{P}_n(|X_n - X^*| > C^{-n}) \leq C^{-n}\}$:

$$c^n \mathbb{E}_n[|X_n - X^*|] \leq c^n \left(2C^{-n}\right) = 2(c/C)^n \leq \varepsilon,$$

where the last inequality follows since $n \geq \widetilde{N}(\varepsilon, c, C)$.

Then

$$\mathbb{P}(c^n \mathbb{E}_n[|X_n - X^*|] \leq \varepsilon) \geq \mathbb{P}(\mathbb{P}_n(|X_n - X^*| > C^{-n}) \leq C^{-n})$$

and

$$\mathbb{P}(c^n \mathbb{E}_n[|X_n - X^*|] > \varepsilon) \leq \mathbb{P}(\mathbb{P}_n(|X_n - X^*| > C^{-n}) > C^{-n}) \leq C^{-n},$$

where the last step follows by Proposition 3. □

Now we are ready to prove convergence in probability and uniform integrability of the process $\big(c^n\mathbb{E}[|X_n - X^*|]\big)_n$ and with that prove Theorem 2.

**Proposition 5.** *Let $C$ be the constant defined in Lemma 1. Then*

$$\mathbb{E}_n[|X_n - X^*|] = o_p(c^{-n})$$

*for all $c \in (1, C)^2$.*

*Proof.* Choose arbitrary $c \in (1, C)$, which exists since $C > 1$. Fix $\varepsilon > 0$. Then, by Proposition 4, $\mathbb{P}(c^n\mathbb{E}_n[|X_n - X^*|] > \varepsilon) \leq 2C^{-n}$ for large $n$, that is, for $n > \widetilde{N}(\varepsilon, c, C)$. Thus,

$$\lim_{n\to\infty} \mathbb{P}\big(c^n\mathbb{E}_n[|X_n - X^*|] > \varepsilon\big) = 0,$$

which holds for any chosen $\varepsilon > 0$, and hence $\big(c^n\mathbb{E}_n[|X_n - X^*|]\big)_n$ converges to 0 in probability. □

**Proposition 6.** *Let $C$ be the constant defined in Lemma 1. Then the stochastic process $\big(c^n\mathbb{E}_n[|X_n - X^*|]\big)_n$ is uniformly integrable for all $c \in (1, C)$.*

*Proof.* By definition a sequence of random variables $(Y_n)_n$ is uniformly integrable if $\sup_{n\in\mathbb{N}} \mathbb{E}\big[|Y_n|\mathbb{1}\{|Y_n| > t\}\big] \to 0$ as $t \to \infty$.

Choose arbitrary $c \in (1, C)$ and consider $\widetilde{N}(1, c, C) = (\log 2)/(\log(C/c))$, which is strictly positive (the function $\widetilde{N}(\varepsilon, c, C)$ is defined in Proposition 4). Note that $\widetilde{N}(t, c, C) \leq \widetilde{N}(1, c, C)$ for $t \geq 1$. Define $T(c, C) = c^{\widetilde{N}(1,c,C)} > 1$ and consider arbitrary $t \geq T(c, C) > 1$. It follows that $\mathbb{P}\big(c^n\mathbb{E}_n[|X_n - X^*|] > t\big) = 0$ for $n \leq \widetilde{N}(1, c, C)$, since $\mathbb{E}_n[|X_n - X^*|] \leq 1$ and $c^n \leq t$ for $n \leq \widetilde{N}(1, c, C)$. By Proposition 4, $\mathbb{P}\big(c^n\mathbb{E}_n[|X_n - X^*|] > t\big) \leq C^{-n}$ for $n \geq \widetilde{N}(1, c, C) \geq \widetilde{N}(t, c, C)$.

[2] $f(x) = o_p\,(g(x))$ means $f(x)/g(x) \to 0$ in probability as $x \to \infty$.

Hence $\mathbb{P}\big(c^n\mathbb{E}_n[|X_n - X^*|] > t\big) \leq C^{-n}$ *for all* $n \in \mathbb{N}$ and all $t > T(c, C)$. Using $\mathbb{E}_n[|X_n - X^*|] \leq 1$ shows that for all $n \in \mathbb{N}_0$,

$$\begin{aligned}
\mathbb{E}\big[c^n\mathbb{E}_n[|X_n - X^*|]\mathbb{1}\{c^n\mathbb{E}_n[|X_n - X^*|] > t\}\big]& \\
&\leq c^n\mathbb{E}\big[\mathbb{1}\{c^n\mathbb{E}_n[|X_n - X^*|] > t\}\big] \\
&= c^n\mathbb{E}\big[\mathbb{1}\{c^n > t\}\mathbb{1}\{c^n\mathbb{E}_n[|X_n - X^*|] > t\}\big] \\
&= c^n\mathbb{1}\{c^n > t\}\mathbb{P}\big(c^n\mathbb{E}_n[|X_n - X^*|] > t\big) \\
&\leq c^n\mathbb{1}\{c^n > t\}C^{-n} = (c/C)^n\,\mathbb{1}\{n > \log_c t\}.
\end{aligned}$$

Now we take on both sides the supremum over $n \in \mathbb{N}_0$:

$$\begin{aligned}
\sup_{n\in\mathbb{N}_0} \mathbb{E}\big[c^n\mathbb{E}_n[|X_n - X^*|]\mathbb{1}\{c^n\mathbb{E}_n[|X_n - X^*|] > t\}\big] &\leq \sup_{n\in\mathbb{N}_0} (c/C)^n\,\mathbb{1}\{n > \log_c t\} \\
&= (c/C)^{\log_c t},
\end{aligned}$$

and uniform integrability follows by letting $t$ go to $+\infty$. □

*Proof of Theorem 2.* By Propositions 5 and 6 we can choose an arbitrary constant $c \in (1, C)$ such that the sequence $\big(c^n\mathbb{E}_n[|X_n - X^*|]\big)_n$ converges to 0 in probability and is uniformly integrable. Then

$$\mathbb{E}\left[c^n\mathbb{E}_n\left[|X_n - X^*|\right]\right] \to 0, \quad \text{as } n \to \infty,$$

since convergence in probability and uniform integrability is a necessary and sufficient condition for convergence in $L^1$; see, for example, Theorem 4.5.2 in Durrett (2005). Finally, by the tower property of conditional expectation, $c^n\mathbb{E}[|X_n - X^*|] = \mathbb{E}[c^n\mathbb{E}_n[|X_n - X^*|]]$ and hence $\mathbb{E}[|X_n - X^*|] = o(c^{-n})$. □

### 2.3.2 Consistency and Robustness

Almost immediate consequences of the preceding analysis are that the posterior absolute residuals converge to 0 almost surely and that the posterior density $f_n$ converges to a point mass at $X^*$. Hence the PBA is a consistent method for locating $X^*$.

**Theorem 3.** $\mathbb{E}_n[|X_n - X^*|] \to 0$ *almost surely as* $n \to \infty$.

**Corollary 1.** *With probability one the posterior distribution* $F_n$ *converges weakly to a point mass at* $X^*$*, that is,* $\lim_{n\to\infty} F_n(x) = \mathbb{1}\{x \geq X^*\}$ *for all* $x \neq X^*$ *almost surely.*

**Corollary 2.** *The sequence of medians* $(X_n)_n$ *generated by the PBA converges to* $X^*$ *almost surely, that is,* $\mathbb{P}(\lim_{n\to\infty} X_n = X^*) = 1$.

The proofs of Theorem 3 and Corollaries 1 – 2 are provided in Appendix A.

As a final remark, we show that in some cases the geometric rate of convergence shown in Theorem 2 still holds even if the density of the average-case performance measure is different from the density used in the updating process of the PBA. Suppose that the random variable $X^*$ has a density $g_0$ on $[0, 1]$ and let $(X_n)_n$ be the sequence of medians generated by the PBA using some other initial prior density $f_0$ (which has to be positive on $[0, 1]$). Then a sufficient condition that the geometric rate of convergence of the expected absolute residuals still holds is that the likelihood ratio between $g_0$ and $f_0$ is bounded, that is, there exists a constant $L \in \mathbb{R}$ such that $g_0(x)/f_0(x) \leq L$ for all $x \in [0,1]$. In this case,

$$\begin{aligned}
\mathbb{E}[|X_n - X^*|] &= \int_0^1 g_0(x)\mathbb{E}\left[|X_n - x| \;\middle|\; X^* = x\right] dx \\
&= \int_0^1 f_0(x)\frac{g_0(x)}{f_0(x)}\mathbb{E}\left[|X_n - x| \;\middle|\; X^* = x\right] dx \\
&\leq L\int_0^1 f_0(x)\mathbb{E}\left[|X_n - x| \;\middle|\; X^* = x\right] dx = L\mathbb{E}\left[|X_n - X_f^*|\right],
\end{aligned}$$

where $X_f^* \sim f_0$, and thus Theorem 2 implies $\mathbb{E}[|X_n - X^*|] = o(c^{-n})$. In the case that the performance measure has an unbounded likelihood ratio with respect to $f_0$, for example, when $g_0$ is a point mass at a given point, it remains an open research question whether or not the geometric rate of convergence still holds. (See also Conjecture 2 in Section 3.5.)

This concludes the discussion of the case when $p(\cdot)$ is constant and known. In the next chapter we analyze the behavior of the PBA when $p(\cdot)$ is unknown and varies with $x$.

CHAPTER 3

# VARYING AND UNKNOWN PROBABILITY OF CORRECT RESPONSES

## 3.1 Introduction

In the previous chapter, we analyzed the PBA under the setting where the probability $p(\cdot)$ of obtaining a correct sign at every iteration is constant and known. While these are realistic assumptions for some real-world applications, such as signal transmission over a noisy channel (Horstein, 1963) and edge detection (Castro and Nowak, 2008a), they do not hold for many stochastic root-finding problems. In Section 1.5, we demonstrated how tests of power one can be used so that the PBA updating remains reasonable even when $p(\cdot)$ is nonconstant and unknown. In the current chapter, we investigate the convergence behavior of this modified PBA via a frequentist approach, where the root $x^*$ is a fixed unknown value in $[0, 1]$. Several key results including consistency, finite-time confidence intervals and asymptotic rates of convergence are provided.

Recall from Section 1.5 that, when $p(\cdot)$ is nonconstant and unknown, at each measurement point $X_n$ a signal $\widetilde{Z}_n(X_n)$ is constructed by means of the test of power one (if the context allows it, we write $\widetilde{Z}_n$). For each $n \in \mathbb{N}$, this signal has probability $\tilde{p}(X_n)$ of being correct. As long as $X_n \neq x^*$ we have that $\tilde{p}(X_n) \geq p_c$, where $p_c$ is an input parameter chosen by the user. Both, $\widetilde{Z}_n(X_n)$ and the constant $p_c$ can then be used in the updating equations of the PBA, that is, (1.3) and (1.4). Since the test of power one requires a random number of function evaluations at each step (denoted $N_n(\tilde{g}(X_n))$ or simply $N_n$), it effectively introduces two time scales to the modified PBA, namely, a macro time scale counting the number of

PBA iterations, and a wall-clock time scale counting the total number of function evaluations $T$. By definition, $T_n = \sum_{i=0}^{n} N_i$ represents the total wall-clock time across the first $(n+1)$ iterations.

To better understand the modified PBA, we first study in Sections 3.2–3.5 its behavior in macro time. As shown in Theorem 2, when $p(\cdot) \equiv p_c$, the expected absolute residuals of the PBA converge to 0 at a geometric rate in macro time when $X^* \sim f_0$. The same result holds also for fixed values $x^* \in B$, where $B$ is a set of Lebesgue measure one. While it may seem natural to have at least a similar rate of convergence in macro time when $\tilde{p}(\cdot) \geq p_c$ instead of $p(\cdot) \equiv p_c$, it becomes challenging to turn this intuition into a rigorous proof. First of all, the proof of Theorem 2 cannot be extended to the case $\tilde{p}(\cdot) \geq p_c$ easily as it relies heavily on a symmetric random walk argument that only holds when $p(\cdot) \equiv p_c$. Moreover, when $\tilde{p}(\cdot) \geq p_c$, the PBA no longer guarantees a proper Bayesian updating. To avoid creating a separate probabilistic model for $\tilde{p}(\cdot)$, which is a prerequisite for the construction of true Bayesian dynamics when $\tilde{p}(\cdot) \geq p_c$, we adopt a frequentist approach to prove a set of results similar to those in Theorem 2. In addition, we show how for any finite $n \in \mathbb{N}$ a true confidence interval of the root $x^*$ can be specified, providing a useful finite-time guarantee. More specifically, the main results in macro time show

1. that for any $n \in \mathbb{N}$ and $\alpha \in (0,1)$ an interval $J_n(\alpha) \subseteq [0,1]$ can be constructed such that $\mathbb{P}(x^* \in J_n(\alpha)) \geq 1-\alpha$; moreover the length of $J_n(\alpha)$ converges to 0 at a geometric rate;
2. that a point of maximum posterior density, that is, $X_n^M \in \operatorname{argmax}_{x \in [0,1]} f_n(x)$, is an estimator of $x^*$ whose expected absolute residuals converge to 0 at a geometric rate.

To show these results we will use the following set of standing assumptions:

**Assumption $\mathcal{A}$.** (i) $x^*$ is a fixed and unknown point in $[0,1]$;

(ii) $X_n \neq x^*$ for any $n \in \mathbb{N}_0$;

(iii) $f_0(x) = \mathbb{1}_{[0,1]}(x)$, that is, the PBA starts with a uniform prior density;

(iv) $\tilde{p}(x) \geq p_c$ for all $x \neq x^*$, where $p_c \in (1/2, 1)$ is an input parameter chosen by the user, and $\tilde{p}(x)$ is the probability of receiving a correct signal $\widetilde{Z}(x)$ when using a test of power one at the measurement point $x \in [0,1] \setminus \{x^*\}$. We further use the notation $\tilde{q}(\cdot) = 1 - \tilde{p}(\cdot)$ and $q_c = 1 - p_c$. Moreover, $\mathbb{P}(N(x) < \infty) = 1$ for all $x \neq x^*$, where $N(x)$ is the stopping time of the test of power one.

Assumption (ii) is necessary in order for the sequence $(X_n)_n$ to be well-defined. If for some finite $n$ we actually measure exactly at the root, that is, $X_n = x^*$, then the test of power one has a positive chance of never terminating and no further measurement $X_{n+1}$ is taken. As discussed in Section 1.5, this is a very unlikely event in practice. More precisely, for any fixed constant $p_c$, this can only happen for $x^* \in E(p_c)$, where $E(p_c)$ is a set of Lebesgue measure 0. One way to ensure that (ii) holds with probability one is to extend the starting interval from $[0,1]$ to $[0, 1+\delta]$, where $\delta$ is chosen uniformly at random from the interval $[0, \varepsilon]$, for some $\varepsilon > 0$. If the PBA queries at a point in $(1, 1+\delta]$ then return $\widetilde{Z}_n(X_n) = -1$, indicating that the root is further to the left.

Assumption (iii) is not necessary for most presented results, but convenient. If a different distribution $f_0$ is used as a input density for the PBA, then all presented results still hold as long as $\inf_{x\in[0,1]} f_0(x) > 0$ (but the proofs would have to be modified accordingly).

In Section 3.6, we then analyze the PBA in wall-clock time. If the function $g$, of which we want to locate the root $x^*$, is continuous at $x^*$, then the test of power one requires more function evaluations the closer the measurement point $X_n$ is to $x^*$. In other words, as the sequence $(X_n)_n$ generated by the PBA updating approaches the sought-after point $x^*$ the test of power one slows down the rate in wall-clock time significantly. We show that the sequence $(X_n)_n$ fails to achieve the optimal rate of convergence in wall-clock time, that is, $O(T^{-1/2})$, but, that an averaging-scheme of the measurement points $(X_n)_n$, similar to Polyak-Ruppert averaging, recovers a near-optimal rate of convergence. The proof of this last statement requires that a reasonable conjecture holds true. While we provide empirical evidence that the conjecture holds, a formal proof is still missing. Furthermore, empirical results suggest that such an averaging of the sequence $(X_n)_n$ might even recover the same asymptotic rate of convergence as SA-type algorithms. See Chapter 4 for details on the numerical results. The main results regarding the convergence behavior of the PBA in wall-clock time show

1. that the measurement points defined by the PBA converge to 0 at a rate that is slower than $O(T^{-1/2})$, that is, the sequence $\left(|X_n - x^*|(T_n)^{1/2}\right)_n$ is not tight, where $T_n = \sum_{i=0}^{n} N_i$ is the wall-clock time as a function of macro iterations;
2. that, if there exists an $r > 0$ such that $\mathbb{E}[|X_n - x^*|] = O(e^{-rn})$, then an estimator $\hat{X}_n$ based on the PBA can be constructed such that $\mathbb{E}[(T_n)^{1/2-\varepsilon}|X_n - x^*|] = O(1)$ for any $\varepsilon > 0$.

The rest of this chapter is organized as follows. In Section 3.2, we show that the PBA provides a consistent estimator for the setting considered. In Section 3.3, we introduce and analyze the construction of the confidence interval for $x^*$. In

Section 3.4, we extend this construction to provide sequential confidence intervals, which are sequences of intervals that contain $x^*$ for all $n \in \mathbb{N}$ with high probability. In Section 3.5, we analyze the asymptotic rate of convergence of the PBA in macro time. In Section 3.6, we investigate the asymptotic rate of convergence in wall-clock time.

## 3.2 Consistency

The main result of this section shows that the PBA using tests of one provides a consistent method for locating the root $x^*$.

**Theorem 4.** *Suppose that Assumption $\mathcal{A}$ in Section 3.1 holds. Let $(X_n)_n$ be the sequence of measurement points of the PBA in macro time. Then $X_n \to x^*$ almost surely as $n \to \infty$.*

*Proof.* We first show, in Lemma 2 below, that if intervals $A$ and $B$ are such that $A$ lies completely to the left of $B$ and $B$ lies completely to the left of $x^*$, then the stochastic process

$$V_n = \frac{\mu_n(A)}{\mu_n(B)} \tag{3.1}$$

is a supermartingale. Here, $\mu_n(\cdot)$ denotes the probability measure defined by the pdf $f_n$. We then show (Lemma 3) that this implies that the conditional distribution $\nu_n$ of $\mu_n$, that is, $\mu_n$ restricted to $[0, x^*]$ and normalized by $\mu_n([0, x^*])$, converges weakly to a distribution $\nu$ almost surely. Furthermore, $\nu([x^* - \varepsilon, x^*]) > 0$ for any $\varepsilon > 0$ almost surely. These facts are then used to show (Lemma 4) that $\liminf_{n \to \infty} X_n \geq x^*$ almost surely.

A symmetric argument on the interval $[x^*, 1]$ (omitted) shows that $\limsup_{n\to\infty} X_n \leq x^*$ almost surely. Therefore $X_n \to x^*$ as $n \to \infty$ almost surely. □

We denote by $\widetilde{\mathbb{G}} = (\widetilde{\mathcal{G}}_n)_{n\geq -1}$ the filtration generated by the measurement points $(X_n)_n$ and signals $\widetilde{Z}_n$ observed so far, that is, $\widetilde{\mathcal{G}}_n = \sigma\big(X_m, \widetilde{Z}_m : 0 \leq m \leq n\big)$ for $n \geq 0$ and $\widetilde{\mathcal{G}}_{-1}$ is the trivial $\sigma$-algebra.

**Lemma 2.** *Let $A$ be the interval $[a_1, a_2)$ and $B$ be the interval $[b_1, b_2]$, where $0 \leq a_1 < a_2 \leq b_1 < b_2 \leq x^*$ and consider the process $V_n = \mu_n(A)/\mu_n(B)$. Then $(V_{n+1})_{n\geq 0}$ is a supermartingale with respect to the filtration $\widetilde{\mathbb{G}}$.*

*Proof.* We redefine $B = [b_1, b_2)$ and prove the supermartingale property under this modified definition. This implies the originally stated result since $\mu_n(\{x^*\}) = 0$ for all $n$, and is convenient because now all points of $B$ lie strictly to the left of the root.

Consider arbitrary $n \in \mathbb{N}_0$. Since $V_n$ is $\widetilde{\mathcal{G}}_{n-1}$ measurable, the only property that requires verification is the supermartingale inequality, that is, $\mathbb{E}\left[V_{n+1}|\widetilde{\mathcal{G}}_{n-1}\right] \leq V_n$.

If $X_n \geq b_2$ or $X_n < a_1$, then $V_{n+1} = V_n$ so $\mathbb{E}[V_{n+1}|\widetilde{\mathcal{G}}_{n-1}] = V_n$ on this event.

If $X_n \in A$, and $c_n = \int_{a_1}^{X_n} \mu_n(dx), d_n = \int_{X_n}^{a_2} \mu_n(dx)$, then, conditional on $\widetilde{\mathcal{G}}_{n-1}$,

$$V_{n+1} = \begin{cases} \frac{2q_c c_n + 2p_c d_n}{2p_c \mu_n(B)}, & \text{with probability } \tilde{p}(X_n), \\ \frac{2p_c c_n + 2q_c d_n}{2q_c \mu_n(B)}, & \text{with probability } \tilde{q}(X_n). \end{cases}$$

Since $p_c < 1$ it follows that $V_n > 0$ for all $n \in \mathbb{N}_0$, and therefore

$$\begin{aligned}\mathbb{E}\left[\frac{V_{n+1}}{V_n}\middle|\widetilde{\mathcal{G}}_{n-1}\right] &= \frac{\tilde{p}(X_n)}{p_c}\left[\frac{q_c c_n + p_c d_n}{\mu_n(A)}\right] + \frac{\tilde{q}(X_n)}{q_c}\left[\frac{p_c c_n + q_c d_n}{\mu_n(A)}\right] \\ &= \frac{1}{\mu_n(A)}\left[\frac{q_c}{p_c}\tilde{p}(X_n)c_n + \tilde{p}(X_n)d_n + \frac{p_c}{q_c}\tilde{q}(X_n)c_n + \tilde{q}(X_n)d_n\right] \\ &= \frac{1}{\mu_n(A)}\left[\left(\frac{q_c}{p_c}\tilde{p}(X_n) + \frac{p_c}{q_c}\tilde{q}(X_n)\right)c_n + d_n\right].\end{aligned}$$

It remains to show that the factor multiplying $c_n$ is smaller than 1. To that end,

$$\begin{aligned}\frac{q_c}{p_c}\tilde{p}(X_n) + \frac{p_c}{q_c}\tilde{q}(X_n) &= \frac{q_c^2\tilde{p}(X_n) + p_c^2\tilde{q}(X_n)}{p_c q_c} \\ &= \frac{q_c^2\tilde{p}(X_n) + p_c^2(1-\tilde{p}(X_n))}{p_c q_c} \\ &= \frac{p_c^2 - (p_c^2 - q_c^2)\tilde{p}(X_n)}{p_c q_c} \\ &= \frac{p_c^2 - (p_c - q_c)(p_c + q_c)\tilde{p}(X_n)}{p_c q_c} \\ &= \frac{p_c^2 - (p_c - q_c)\tilde{p}(X_n)}{p_c q_c} \le 1, \qquad (3.2)\end{aligned}$$

where the last inequality holds since $\tilde{p}(X_n) \ge p_c$, and it would hold with equality if $\tilde{p}(X_n) = p_c$. Hence $\mathbb{E}[V_{n+1}|\widetilde{\mathcal{G}}_{n-1}] \le V_n$ on the event that $X_n \in A$.

If $a_2 \le X_n < b_1$, that is, the median lies between $A$ and $B$ then, conditional on $\widetilde{\mathcal{G}}_{n-1}$,

$$\frac{V_{n+1}}{V_n} = \begin{cases} \frac{q_c}{p_c}, & \text{with probability } \tilde{p}(X_n), \\ \frac{p_c}{q_c}, & \text{with probability } \tilde{q}(X_n), \end{cases}$$

and

$$\mathbb{E}\left[\frac{V_{n+1}}{V_n}\middle|\widetilde{\mathcal{G}}_{n-1}\right] = \frac{q_c}{p_c}\tilde{p}(X_n) + \frac{p_c}{q_c}\tilde{q}(X_n) \le 1$$

as shown in (3.2). Hence $\mathbb{E}[V_{n+1}|\widetilde{\mathcal{G}}_{n-1}] \le V_n$ on the event that $a_2 \le X_n \le b_1$.

Finally, if $X_n \in B$, and $c_n = \int_{b_1}^{X_n} \mu_n(dx), d_n = \int_{X_n}^{b_2} \mu_n(dx)$, then, conditional on $\widetilde{\mathcal{G}}_{n-1}$,

$$\frac{V_{n+1}}{V_n} = \begin{cases} \frac{q_c \mu_n(B)}{q_c c_n + p_c d_n}, & \text{with probability } \tilde{p}(X_n), \\ \frac{p_c \mu_n(B)}{p_c c_n + q_c d_n}, & \text{with probability } \tilde{q}(X_n), \end{cases}$$

and

$$\begin{aligned} \mathbb{E}\left[\frac{V_{n+1}}{V_n}\middle|\widetilde{\mathcal{G}}_{n-1}\right] &= \mu_n(B)\left(\frac{\tilde{p}(X_n)q_c}{q_c c_n + p_c d_n} + \frac{\tilde{q}(X_n)p_c}{p_c c_n + q_c d_n}\right) \\ &= \mu_n(B)\left(\frac{\tilde{p}(X_n)q_c(p_c c_n + q_c d_n) + \tilde{q}(X_n)p_c(q_c c_n + p_c d_n)}{p_c q_c c_n^2 + q_c^2 c_n d_n + p_c^2 c_n d_n + p_c q_c d_n^2}\right) \\ &= \mu_n(B)\left(\frac{\tilde{p}(X_n)q_c p_c c_n + \tilde{p}(X_n)q_c^2 d_n + \tilde{q}(X_n)p_c q_c c_n + \tilde{q}(X_n)p_c^2 d_n}{p_c q_c(c_n^2 + d_n^2) + (q_c^2 + p_c^2)c_n d_n}\right) \\ &= \mu_n(B)\left(\frac{p_c q_c c_n + d_n(\tilde{p}(X_n)q_c^2 + \tilde{q}(X_n)p_c^2)}{p_c q_c(c_n^2 + d_n^2) + (q_c^2 + p_c^2)c_n d_n}\right) \\ &= \mu_n(B)\left(\frac{c_n + d_n(\tilde{p}(X_n)\frac{q_c}{p_c} + \tilde{q}(X_n)\frac{p_c}{q_c})}{c_n^2 + 2c_n d_n + d_n^2 - 2c_n d_n + \frac{(q_c^2+p_c^2)}{p_c q_c}c_n d_n}\right) \\ &= \mu_n(B)\left(\frac{c_n + d_n(\tilde{p}(X_n)\frac{q_c}{p_c} + \tilde{q}(X_n)\frac{p_c}{q_c})}{(c_n + d_n)^2 + \frac{(q_c-p_c)^2}{p_c q_c}c_n d_n}\right) \leq 1, \end{aligned}$$

where the last inequality follows from (3.2) and the fact that $\mu_n(B) = c_n + d_n$. Here, equality does not hold even if $\tilde{p}(X_n) = p_c$ unless either $c_n$ or $d_n$ is 0, that is, unless $X_n$ is equal to one of the endpoints of the interval $B$. Hence also $\mathbb{E}[V_{n+1}|\widetilde{\mathcal{G}}_{n-1}] \leq V_n$ on the event that $X_n \in B$ and this completes the proof. □

**Lemma 3.** *Let $\nu_n(\cdot) = \mu_n(\cdot \cap [0, x^*])/\mu_n([0, x^*])$ be the conditional probability distribution to the left of $x^*$. Then $\nu_n(\cdot)$ converges weakly to a probability measure $\nu(\cdot)$, where $\nu([x^* - \varepsilon, x^*]) > 0$ for any $\varepsilon > 0$, almost surely.*

*Proof.* Let $x \in (0, x^*)$ and define the interval $A(x) = [0, x)$ and the interval $B(x) = [x, x^*)$. Then, by Lemma 2, the process $(\mu_n(A(x))/\mu_n(B(x))_n$ is a supermartingale and the martingale convergence theorem implies that $\mu_n(A(x))/\mu_n(B(x))$

converges almost surely to a finite-valued random variable $R(A(x))$ say. Define $D(x) \subseteq \Omega$ to be the set of sample paths where

$$\nu_n(A(x), \omega) = \frac{\mu_n(A(x), \omega)}{\mu_n(A(x), \omega) + \mu_n(B(x), \omega)} \to \frac{R(A(x), \omega)}{R(A(x), \omega) + 1}$$

as $n \to \infty$ holds, then $D(x)$ is a set of probability one. (We have used the fact that $\mu_n$ has a density for all $n$ so that one can include or exclude the endpoints of intervals at will.)

Let $\mathbb{Q}[0, x^*]$ be the set of rational numbers in the interval $[0, x^*]$ and let $\mathscr{A} = \{A(x) : x \in \mathbb{Q}[0, x^*]\}$. Since $\mathscr{A}$ consists of a countable number of sets, it follows that $\nu_n(A(x), \omega) \to R(A(x), \omega)/(R(A(x), \omega) + 1)$ for all $x \in \mathbb{Q}[0, x^*]$ and for all $\omega$ in the set $D$ of probability one, where

$$D = \bigcap_{x \in \mathbb{Q}[0, x^*]} D(x).$$

Fix $\omega \in D$. The sequence of probability measures $(\nu_n(\cdot, \omega))_n$ is trivially tight since the measures are all defined on the bounded interval $[0, x^*]$. Let $\nu^a(\cdot, \omega)$ and $\nu^b(\cdot, \omega)$ denote two weak limits of subsequences of $(\nu_n(\cdot, \omega))_n$. These probability measures agree on the class $\mathscr{A}$ since the full sequence $\nu_n(A, \omega)$ converges to $R(A, \omega)$ for all $A \in \mathscr{A}$. But the family of intervals $\mathscr{A}$ is a $\pi$-system that generates the Borel field on $[0, x^*]$, and is therefore a separating class (Billingsley, 1999, p. 9). Therefore $\nu^a(\cdot, \omega) = \nu^b(\cdot, \omega)$. Define $\nu(\cdot, \omega)$ to be the common limiting measure. Prohorov's theorem (Billingsley, 1999, p.59) then shows that $\nu_n(\cdot, \omega)$ converges weakly to $\nu(\cdot, \omega)$. This holds for all $\omega \in D$, a set of probability one, hence the first part of the claim follows.

It remains to show that $\nu([x^* - \varepsilon, x^*]) > 0$ for any $\varepsilon > 0$, almost surely. Let $\varepsilon \in (0, x^*)$ be fixed and define $A = [0, x^* - \varepsilon), B = [x^* - \varepsilon, x^*]$. Lemma 2 then implies that

$$\mathbb{P}\left(\limsup_{n\to\infty} \nu_n(B) = 0\right) = \mathbb{P}\left(\liminf_{n\to\infty} \frac{\nu_n(A)}{\nu_n(B)} = \infty\right) = \mathbb{P}\left(\liminf_{n\to\infty} \frac{\mu_n(A)}{\mu_n(B)} = \infty\right) = 0.$$

But $\nu_n$ converges to $\nu$ weakly as $n \to \infty$ almost surely and $B$ is closed, so that $\nu(B) \geq \limsup_{n\to\infty} \nu_n(B) > 0$ almost surely (see Theorem 2.1, p. 16 of Billingsley (1999)). Hence $\nu(B) > 0$ almost surely. □

**Lemma 4.** *It holds that* $\liminf_{n\to\infty} X_n \geq x^*$ *almost surely.*

*Proof.* Suppose that $\liminf_{n\to\infty} X_n < x^*$ on a set of positive probability. Fix a sample path $\omega$ within this set that also belongs to the set of probability one where the properties of Lemma 3 hold. Let $\varepsilon > 0$ be such that $X_n < x^* - \varepsilon$ infinitely often and consider the set $A = [x^* - \varepsilon, x^*]$. By Lemma 3 it holds that $\nu_n(A) \to \nu(A)$ as $n \to \infty$ and $\nu(A) > 0$. Let $\eta > 0$, $\varepsilon' > 0$ and $N \in \mathbb{N}$ such that $\nu_n(A) \geq \eta$ and

$$|\nu_n(A) - \nu(A)| < \varepsilon', \tag{3.3}$$

for all $n \geq N$.

Consider the updating of $\nu_n(A)$ at a time $n \geq N$ where also $X_n < x^* - \varepsilon$ (such an $n$ always exists since $X_n < x^* - \varepsilon$ infinitely often). At this time, $\nu_n(A) \geq \eta$ and $\nu_n([0, X_n])) \geq \mu_n([0, X_n]) \geq 1/2$. Let us outline how this will lead to a contradiction: Since the conditional distribution $\nu_n$ has a "large" amount of probability mass to the left and right of the measurement point $X_n$ it follows that, independently of the outcome of $\widetilde{Z}_n$, the updated distribution $\nu_{n+1}$ differs significantly from the distribution $\nu_n$ in the sense that $|\nu_n(A) - \nu(A)| < \varepsilon'$ cannot hold for $n$ and $n + 1$, contradicting (3.3).

Let us now make this argument precise by considering the exact updating procedure of $\nu_n$. Since $\mu_n([0, X_n]) = 1/2$ it follows that $\nu_n([0, X_n]) > 1/2$ and trivially it also holds that $\nu_n([0, X_n]) \leq 1$. Thus, since $p_c > q_c$,

$$\frac{q_c}{p_c}\nu_n([0, X_n]) + \nu_n([X_n, x^*]) \leq 1 - \frac{p_c - q_c}{p_c}, \quad \text{and} \tag{3.4}$$
$$\frac{p_c}{q_c}\nu_n([0, X_n]) + \nu_n([X_n, x^*]) \geq 1 + \frac{p_c - q_c}{2q_c}. \tag{3.5}$$

Now define $\delta = \min\{(p_c - q_c)/p_c, (p_c - q_c)/(2q_c)\}$ and let $\varepsilon' > 0$ be smaller than $\delta\nu(A)/(2 + \delta)$ (such an $\varepsilon' > 0$ always exists since $\nu(A) > 0$). Since $\nu_n(A) \to \nu(A)$, there exists an $N > 0$ such that for all $n \geq N$,

$$|\nu_n(A) - \nu(A)| < \varepsilon'. \tag{3.6}$$

Let $n \geq N$ be such that $X_n < x^* - \varepsilon$, which exists since by assumption $X_n < x^* - \varepsilon$ infinitely often. If the signal at the $(n+1)$st iteration is negative, that is, $\widetilde{Z}_n = -1$, then

$$\begin{aligned}\nu_{n+1}(A) &= \frac{\mu_{n+1}(A)}{\mu_{n+1}([0, x^*])} \\ &= \frac{2q_c\mu_n(A)}{2p_c\mu_n([0, X_n]) + 2q_c\mu_n([X_n, x^*])} \\ &= \frac{\nu_n(A)}{\frac{p_c}{q_c}\nu_n([0, X_n]) + \nu_n([X_n, x^*])}.\end{aligned}$$

From (3.5), the denominator is bounded below by $1 + \delta$, and so

$$\nu_{n+1}(A) \leq \frac{\nu_n(A)}{1 + \delta} \leq \frac{\nu(A) + \varepsilon'}{1 + \delta} \leq \nu(A) - \varepsilon'$$

because of the way we chose $\varepsilon'$, and this contradicts (3.6).

If, on the other hand, the signal at the $(n + 1)$st iteration is positive, that is, $\widetilde{Z}_n = +1$, then

$$\begin{aligned}\nu_{n+1}(A) &= \frac{2p_c\mu_n(A)}{2q_c\mu_n([0, X_n]) + 2p_c\mu_n([X_n, x^*])} \\ &= \frac{\nu_n(A)}{\frac{q_c}{p_c}\nu_n([0, X_n]) + \nu_n([X_n, x^*])}.\end{aligned}$$

From (3.4), the denominator is bounded above by $1-\delta$, so that

$$\nu_{n+1}(A) \geq \frac{\nu_n(A)}{1-\delta} \geq \frac{\nu(A)-\varepsilon'}{1-\delta} \geq \nu(A)+\varepsilon'$$

because of the way we chose $\varepsilon'$, and this again contradicts (3.6). So irrespective of the outcome of $\widetilde{Z}_n$ we arrive at a contradiction. Hence $\liminf_{n\to\infty} X_n < x^*$ can only hold on a set of probability 0. □

## 3.3 Confidence Intervals

As shown in the previous section, the PBA provides a consistent method for locating the root $x^* \in [0,1]$. For a stochastic root-finding method to be successful in practice, however, more is needed. It is important that the root $x^*$ (or a close estimate of it) is found with as few function evaluations as possible. Since $x^*$ can take any real value in $[0,1]$, finding the root $x^*$ exactly in finite time seems impossible, raising the question: Can we provide a statistical guarantee, such as a confidence interval, on the location of the root after $n$ function evaluations?

Surprisingly, no stochastic root-finding algorithm exists—to the best of our knowledge—that provides the simulation analyst with such a guarantee. One method of constructing at least approximate confidence intervals for stochastic root-finding problems is to restart the search algorithm several times and then use a central limit theorem approximation of $x^*$; see Hsieh and Glynn (2002). Whereas this method might work well in practice, no strict guarantees on the coverage probability can be provided. In this section, we show that the PBA provides not only a point estimate of $x^*$ after $n$ observed signals, but also a true confidence interval for $x^*$. Moreover, the width of of this confidence interval converges to 0 at a geometric rate in macro time.

Let $\left(X_{(i)}\right)_{i=0}^{n}$ denote the order statistics of the query points $(X_i)_{i=0}^n$, that is, $X_{(0)} \leq X_{(1)} \leq \cdots \leq X_{(n)}$, and denote the intervals defined by $\left(X_{(i)}\right)_{i=0}^{n}$ as $I_{i,n}$ for $i = 0, \ldots, n+1$. These are at most $n+2$ non-empty intervals (it can be less than $n+2$ intervals if the algorithm measures at the same point more than once). These intervals are

$$\begin{aligned} I_{0,n} &= \left[0, X_{(0)}\right), \\ I_{i,n} &= \left[X_{(i-1)}, X_{(i)}\right), \quad \text{for } i = 1, \ldots, n, \text{ and} \\ I_{n+1,n} &= \left[X_{(n)}, 1\right]. \end{aligned}$$

Let us denote the height of the (piecewise constant) density $f_n$ on the $i$th interval by $h_n(I_{i,n})$, for $i = 0, \ldots, n+1$. We also use the notation $d = d(p_c) = p_c \log(2p_c) + q_c \log(2q_c)$ and $\beta = \beta(p_c) = \log(p_c/q_c)$. Later, the fact that $d(p_c) > 0$ for all $p_c \in (1/2, 1)$ will be important.

Let $\alpha \in (0,1)$ and define

$$b_n = b_n(\alpha) = b_n(\alpha, p_c) = nd(p_c) - n^{1/2}(-1/2 \log(\alpha/2))^{1/2}\beta(p_c); \tag{3.7}$$

the intervals

$$\widetilde{I}_{i,n}(\alpha) = \begin{cases} I_{i,n}, & \text{if } h_n(I_{i,n}) > e^{b_n(\alpha)}, \\ \emptyset, & \text{otherwise;} \end{cases}$$

the set $G_n(\alpha) = \bigcup_{i=0}^{n+1} \widetilde{I}_{i,n}(\alpha)$, which consists of all points in $[0,1]$ whose density $f_n$ is above $e^{b_n}$ at time $n$; and the interval

$$J_n(\alpha) = \operatorname{conv}(G_n(\alpha)), \tag{3.8}$$

which is the convex hull of the set $G_n(\alpha)$. The next proposition shows that the set $G_n(\alpha)$ (the interval $J_n(\alpha)$) provides a $(1 - \alpha/2)$-confidence set (interval) for the sought-after point $x^*$.

**Proposition 7.** *Suppose that Assumption $\mathcal{A}$ in Section 3.1 holds and $\alpha \in (0,1)$. Then, for all $n \in \mathbb{N}_0$,*

$$\mathbb{P}(x^* \notin J_n(\alpha)) \leq \mathbb{P}(x^* \notin G_n(\alpha)) \leq \alpha/2.$$

*Proof.* First note that $G_n(\alpha) \subseteq J_n(\alpha)$, since $J_n(\alpha)$ is the convex hull of $G_n(\alpha)$, and so it is sufficient to show that $\mathbb{P}(x^* \notin G_n(\alpha)) \leq \alpha/2$.

Assume first that $\tilde{p}(\cdot) \equiv p_c$. Consider the random variable $A_n = f_n(x^*)$. By definition of the set $G_n(\alpha)$, it holds that

$$\mathbb{P}(x^* \notin G_n(\alpha)) = \mathbb{P}(A_n \leq e^{b_n}),$$

and it is enough to show that $\mathbb{P}(\log A_n \leq b_n) \leq \alpha/2$. The random variable $A_n$ is a product of iid random variables, that is, $A_n = \prod_{i=1}^n (2p_c)^{C_i}(2q_c)^{1-C_i}$, where $C_i = 1$ if the signal $\widetilde{Z}_{i-1}$ at the $i$th iteration of the PBA is correct and 0 otherwise ($C$ is mnemonic for "correct"). By taking logarithms we get $\log A_n = \sum_{i=1}^n \xi_i$, where $\xi_i = \log(2p_c)$ if $C_i = 1$ and $\xi_i = \log(2q_c)$ otherwise, and $\mathbb{P}(\xi_i = \log(2p_c)) = p_c$ and $\mathbb{P}(\xi_i = \log(2q_c)) = q_c$. Then

$$\begin{aligned}
\mathbb{P}\left(\log A_n \leq b_n\right) &= \mathbb{P}\left(\sum_{i=1}^n \xi_i \leq b_n\right) \\
&= \mathbb{P}\left(n^{-1}\sum_{i=1}^n \xi_i - d \leq n^{-1}b_n - d\right) \\
&\leq \exp\left(-2\frac{(b_n/n - d)^2 n}{\beta^2}\right), \qquad (3.9)
\end{aligned}$$

which follows by Hoeffding's inequality[1]. The claim follows by the definition of $b_n$ given in (3.7).

[1] Let $X_1, \ldots, X_n$ be iid bounded random variables, that is, $\mathbb{P}(X_i \in [a,b]) = 1$. Then for the empirical mean $\overline{X} = n^{-1}\sum_{i=1}^n X_i$ the inequality $\mathbb{P}\big(\overline{X} - \mathbb{E}[\overline{X}] \geq t\big) \leq \exp\big(-2t^2 n(b-a)^{-2}\big)$ holds when $t \geq 0$. See Hoeffding (1963).

Now consider the case where $\tilde{p}(\cdot) \geq p_c$. Again, we consider the random variable $\log A_n = \log\big(f_n(x^*)\big)$. This random variable is no longer a sum of iid random variables, but stochastically dominates a random variable that is a sum of iid random variables as defined in the previous case. More precisely, it holds that $\log A_n \overset{d}{\sim} \sum_{i=1}^n \xi_i$ (the notation $\overset{d}{\sim}$ stands for equality in distribution), and the random variables $(\xi_i)_i$ are defined by

$$\xi_i = \begin{cases} \log(2p_c), & \text{if } U_i \leq \tilde{p}(X_{i-1}), \\ \log(2q_c), & \text{otherwise,} \end{cases}$$

where $(U_i)_i$ is a sequence of iid $U(0,1)$ random variables. Use pathwise the same sequence $(U_i)_i$ to also define

$$\phi_i = \begin{cases} \log(2p_c), & \text{if } U_i \leq p_c, \\ \log(2q_c), & \text{otherwise.} \end{cases}$$

By construction, $\sum_{i=1}^n \phi_i \leq \sum_{i=1}^n \xi_i$. Thus,

$$\begin{aligned} \mathbb{P}(x^* \notin G_n) &= \mathbb{P}(\log A_n \leq b_n) \\ &= \mathbb{P}\left(\sum_{i=1}^n \xi_i \leq b_n\right) \leq \mathbb{P}\left(\sum_{i=1}^n \phi_i \leq b_n\right) \leq \alpha/2, \end{aligned}$$

where the last inequality follows as in the case where $p(\cdot) \equiv p_c$. □

Such a true confidence interval $J_n(\alpha)$ is only useful if its size decreases as $n$ increases. As an extreme example, the trivial interval $[0,1]$ will always provide a true confidence interval. The next proposition shows that the length of the confidence interval $J_n(\alpha)$ converges to 0 at a geometric rate with high probability. The proof requires that the input parameter satisfies $p_c \geq 0.85$, but we conjecture that the result should hold for any $p_c > 1/2$, which is supported by empirical results in Chapter 4.

**Proposition 8.** *Suppose that Assumption $\mathcal{A}$ in Section 3.1 holds and let $p_c \in [0.85, 1)$. Then $d - q_c\beta > 0$. If $\alpha \in (0,1)$ and $r \in (0, d - q_c\beta)$, then*

$$\mathbb{P}(|J_n(\alpha)| > e^{-rn}) \leq \alpha/2, \tag{3.10}$$

*for all $n \geq N_1$, where*

$$N_1 = N_1(\alpha, r, p_c) = 2\log(2/\alpha)\left(\frac{\beta}{d - r - q_c\beta}\right)^2 \in \mathbb{R}. \tag{3.11}$$

*Proof.* For the moment assume that $d - q_c\beta > 0$ for $p_c \in [0.85, 1)$. At the end of the proof we show that this is indeed true. Also, choose arbitrary $r \in (0, d - q_c\beta)$ and $\alpha \in (0, 1)$. In this proof we often just write $J_n$ instead of $J_n(\alpha)$.

Consider fixed $n \geq N_1$. Let $A_n = \inf J_n$ and $B_n = \sup J_n$ be the endpoints of the interval $J_n$[2]. We define the event $\mathcal{B} = \{|J_n| > e^{-rn}\}$, which is, by definition of the interval $J_n$, equal to

$$\mathcal{B} = \big\{(B_n - A_n) > e^{-rn}, f_n(A_n) \geq e^{b_n}, f_n(B_n-) \geq e^{b_n}\big\}.$$

Then

$$\mathcal{B} \subseteq \{\hat{f}_n < e^{rn}\}, \tag{3.12}$$

where $\hat{f}_n = \inf_{x \in J_n} f_n(x)$ is the lowest posterior density in the interval $J_n$. To verify this, assume there exists a sample path in $\mathcal{B}$ with $\hat{f}_n \geq e^{rn}$, then, on this sample path,

$$\int_{A_n}^{B_n} f_n(y)dy \geq (B_n - A_n)\hat{f}_n \geq (B_n - A_n)e^{rn} > e^{-rn}e^{rn} = 1,$$

which is a contradiction to $f_n$ being a probability density function.

Now consider a fixed sample path $\omega \in \mathcal{B}$. Define

$$X_{\hat{f}} = \inf\big\{x \in J_n : f_n(x) \leq \hat{f}_n\big\}$$

[2] We use the standard convention that $\sup \emptyset = -\infty$ and $\inf \emptyset = +\infty$.

to be the leftmost point with the lowest posterior density in $J_n$. Since $n \geq N_1$, it holds that $rn < b_n$ (see, for example, proof of Lemma 5), and consequently $X_{\hat{f}} \notin G_n(\alpha)$, that is, $\hat{f}_n < e^{b_n}$. Thus $X_{\hat{f}} \notin \{0, 1\}$, instead $X_{\hat{f}}$ was a previous query point and hence $X_{\hat{f}} \neq x^*$ by Assumption $\mathcal{A}$(ii). Furthermore, the density $f_n$ needs to increase from $\hat{f}_n$ up to $e^{b_n}$ to the right *and* left of $X_{\hat{f}}$ because otherwise $X_{\hat{f}} \notin J_n$.

Assume now that $X_{\hat{f}} < x^*$, the arguments (omitted) for the case $X_{\hat{f}} > x^*$ hold analogously. In this case an increase of the density from $\hat{f}_n$ at $X_{\hat{f}}$ to at least $e^{b_n}$ at $A_n$ can only occur when incorrect signals were observed in the interval $[A_n, X_{\hat{f}}]$. More specifically, consider the dynamics of the density at points $A_n$ and $X_{\hat{f}}$ during all previous $n$ iterations separately. First, for all $X_i \notin [A_n, X_{\hat{f}}]$ the densities at $A_n$ and $X_{\hat{f}}$ were multiplied by the same factor independent of the observed signal, for $i = 0, 1, \ldots, n-1$. So, to achieve a difference in the density at these two points it is necessary that previous measurement points fell into the interval $[A_n, X_{\hat{f}}]$. Next, consider the measurements $X_i \in [A_n, X_{\hat{f}}]$ for $i = 0, 1, \ldots, n-1$. On these events the density at $A_n$ was multiplied by $2p_c$ if $\widetilde{Z}_i = -1$, which is an incorrect signal since $X_{\hat{f}} < x^*$, and the density at the point $X_{\hat{f}}$ was multiplied by $2q_c$. Thus, the density at $A_n$ increases by a factor $p_c/q_c$ relative to the density at $X_{\hat{f}}$ for each incorrect response in the interval $[A_n, X_{\hat{f}}]$. Each correct signal in $[A_n, X_{\hat{f}}]$, on the other hand, decreases the density at $A_n$ relative to the density at $X_{\hat{f}}$ by a factor $q_c/p_c$. This shows, in order for the density to grow from $\hat{f}_n$ at the point $X_{\hat{f}}$ to at least $e^{b_n}$ at the point $A_n$ there must have been at least $H_n$ incorrect replies in the interval $[A_n, X_{\hat{f}}]$ up to time $n$, where $H_n$ needs to satisfy

$$(p_c/q_c)^{H_n} \hat{f}_n \geq e^{b_n}. \tag{3.13}$$

If $W_n$ denotes the total number of incorrect replies up to time $n$ ($W$ is mnemonic for "wrong") then naturally $W_n \geq H_n$ and the bound (3.13) implies that

$$(p_c/q_c)^{W_n} \hat{f}_n \geq e^{b_n}, \tag{3.14}$$

which needs to hold for all $\omega \in \mathcal{B}$. This bound is not very tight, since it ignores the positive effects of correct responses in the interval $[A_n, X_{\hat{f}}]$ as well as the fact that incorrect responses can occur outside of this interval. Nevertheless, it is sufficient for the current proof.

Bounds (3.12) and (3.13) show that, for all sample paths in $\mathcal{B}$,

$$\begin{aligned}(p_c/q_c)^{W_n} e^{rn} &\geq (p_c/q_c)^{W_n} \hat{f}_n \geq e^{b_n} \\ (p_c/q_c)^{W_n} &\geq e^{b_n - rn} \\ W_n &\geq \frac{b_n - rn}{\log(p_c/q_c)} = \frac{b_n - rn}{\beta},\end{aligned}$$

and since this holds for every sample path in $\mathcal{B}$ it follows that

$$\mathcal{B} \subseteq \left\{ W_n \geq \frac{b_n - rn}{\beta} \right\}. \tag{3.15}$$

In order to prove the statement we show that the probability of the event on the right-hand side in (3.15) is smaller than $\alpha/2$. If $\overline{W} \sim \text{Binomial}(n, q_c)$ then $\overline{W}$ stochastically dominates $W_n$, that is

$$\mathbb{P}(W_n \geq x) \leq \mathbb{P}(\overline{W}_n \geq x), \quad \text{for all } x \in \mathbb{R},$$

and hence

$$\mathbb{P}(|J_n| > e^{-rn}) = \mathbb{P}(\mathcal{B}) \leq \mathbb{P}\left( W_n \geq \frac{b_n - rn}{\beta} \right) \leq \mathbb{P}\left( \overline{W}_n \geq \frac{b_n - rn}{\beta} \right) \leq \alpha/2,$$

where the last step follows by Lemma 5 (below).

It remains to show that $d - q_c\beta > 0$ for all $p_c \in [0.85, 1)$. For this consider the function

$$\begin{aligned} v(p) &= d(p) - (1-p)\beta(p) \\ &= p\log(2p) + (1-p)\log(2(1-p)) - (1-p)\log(p/(1-p)) \\ &= (2p-1)\log p + \log 2 + 2(1-p)\log(1-p). \end{aligned}$$

Then

(1) $v(p^*) = 0$ for some $p^* > 0$,

(2) $v'(p) = 2\log p - 2\log(1-p) - p^{-1}$, and $v'(p^*) > 0$,

(3) $v''(p) = 2p^{-1} + (1-p)^{-1} + p^{-2} > 0$ for all $p \in [0,1]$,

hence $v$ is a convex function that is positive at $0.85 > p^* \approx 0.8455$ and has positive slope at $p^*$, which implies that $v(p) > 0$ for all $p > p^*$ (see Figure 3.1). For notational convenience the statement of the proposition requires $p_c \geq 0.85$, instead of the weaker condition $p_c \geq p^*$. □

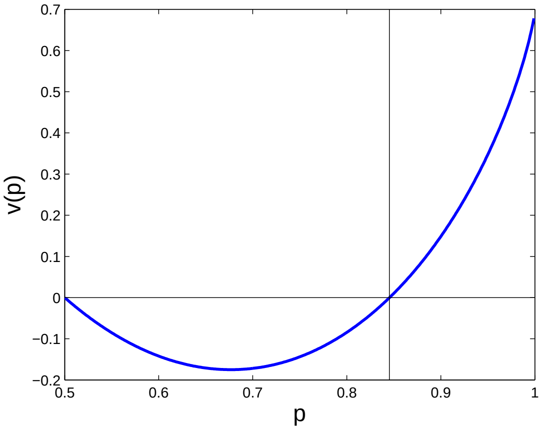

Figure 3.1: The function $v(p) = d(p) - (1-p)\beta(p)$ is convex with one root at 0 and the other root $p^*$ at approximately 0.8455, thus $v(p) > 0$ for all $p \geq 0.85$.

**Lemma 5.** *Let $p \in [0.85, 1)$, $q = 1 - p$, $d = p \log(2p) + q \log(2q)$, $\beta = \log(p/q)$, $r \in (0, d - q\beta)$, $\alpha \in (0, 1)$, $b_n$ as defined in* (3.7)*, and $W_n \sim \text{Binomial}(n, q)$. Then*

$$\mathbb{P}\left(W_n \geq \frac{b_n - rn}{\beta}\right) \leq \frac{\alpha}{2} \tag{3.16}$$

*for all $n \geq N_1$, where $N_1$ is defined by* (3.11) *with $q_c$ replaced by $q$.*

The proof of Lemma 5 is provided in the Appendix A.

The next theorem combines Propositions 7 and 8 into one statement. The proof is a direct application of Boole's inequality and is omitted.

**Theorem 5.** *Suppose that Assumption $\mathcal{A}$ in Section 3.1 holds. Let $p_c \in [0.85, 1)$, $r \in (0, d - q_c\beta)$ and $\alpha \in (0, 1)$. Then*

$$\mathbb{P}\left(x^* \in J_n(\alpha), |J_n(\alpha)| \leq e^{-rn}\right) \geq 1 - \alpha$$

*for all $n \geq N_1$, where $N_1$ is defined by* (3.11)*.*

## 3.4 Sequential Confidence Intervals and Stopping Rules

In this section we construct sequential confidence intervals for the root $x^*$. These are intervals $(K_n(\alpha))_n$ such that $x^*$ is contained in the whole sequence with high probability. In order to achieve this, we again locate the process $A_n = f_n(x^*)$, but, instead of using Hoeffding's bound for a fixed $n \in \mathbb{N}_0$, we now use a statistical test of power one for Bernoulli trials (see Appendix B for details on such tests of power one). This test of power one is performed on the density $f_n(x)$ for every $x \in [0, 1]$ in macro time and is not to be confused with the test of power power one used at each iteration of the PBA producing the signal $\widetilde{Z}_n$.

Let $\alpha \in (0,1)$ and define

$$a_n = a_n(\alpha) = a_n(\alpha, p_c) = nd(p_c) - n^{1/2}\left[-1/2\log\left(\frac{\alpha}{n+1}\right)\right]^{1/2}\beta(p_c), \tag{3.17}$$

and

$$\widetilde{I}^s_{i,n}(\alpha) = \begin{cases} I_{i,n}, & \text{if } h_n(I_{i,n}) > e^{a_n(\alpha)}, \\ \emptyset, & \text{otherwise.} \end{cases}$$

Let $L_n(\alpha) = \bigcap_{j=0}^{n}\bigcup_{i=0}^{j+1}\widetilde{I}^s_{i,n}$ and $K_n(\alpha) = \text{conv}(L_n(\alpha))$. The sequence of intervals $(K_n(\alpha))_n$ is decreasing, that is, $K_0(\alpha) \supseteq K_1(\alpha) \supseteq K_2(\alpha) \supseteq \cdots$. Thus, if the sequence $(K_n(\alpha))_n$ at some time $n$ fails to contain $x^*$, then from this time onwards the sequence $(K_n(\alpha))_n$ will not be able to locate $x^*$ anymore. But, this event can only happen with at most probability $\alpha/2$. Such decreasing and sequential behavior is not guaranteed by the confidence intervals $(J_n(\alpha))_n$ defined in the previous section, which provide the statistical guarantee only for a fixed $n$.

**Proposition 9.** *Suppose that Assumption $\mathcal{A}$ in Section 3.1 holds and $\alpha \in (0,1)$. Then*

$$\mathbb{P}\left(x^* \notin K_n(\alpha) \textit{ for some } n \geq 1\right) \leq \mathbb{P}\left(x^* \notin L_n(\alpha) \textit{ for some } n \geq 1\right) \leq \alpha/2. \tag{3.18}$$

*Proof.* By construction $L_n(\alpha) \subseteq K_n(\alpha)$ for all $n \in \mathbb{N}$ and so it is sufficient to show that $\mathbb{P}(x^* \notin L_n(\alpha) \text{ for some } n \geq 1) \leq \alpha/2$.

Assume first that $\tilde{p}(\cdot) \equiv p_c$. Consider the stochastic process $(A_n)_n$, where $A_n = f_n(x^*)$. Then, by the definition of the set $L_n(\alpha)$,

$$\mathbb{P}(x^* \notin L_n(\alpha) \text{ for some } n \geq 1) = \mathbb{P}(A_n \leq e^{a_n} \text{ for some } n \geq 1),$$

and it is enough to show that $\mathbb{P}(\log A_n \leq a_n \text{ for some } n \geq 1) \leq \alpha/2$. Let $C_n$ be the number of correct signals up to time $n$. Then

$$C_n = \frac{\log A_n - n\log 2q_c}{\beta}, \tag{3.19}$$

and $C_n \sim \text{Binomial}(n, p_c)$. By definition of $a_n$ in (3.17) and $k_n$ in (B.5) (replace $\alpha$ with $\alpha/2$),

$$\begin{aligned}\{\log A_n \leq a_n \text{ for some } n \geq 1\} &= \{C_n - p_c \leq -k_n \text{ for some } n \geq 1\} \qquad (3.20)\\ &\subseteq \{|C_n - np| \geq k_n \text{ for some } n \geq 1\}\end{aligned}$$

and the claim for the case $\tilde{p}(\cdot) \equiv p_c$ follows by the construction of the test of power one for Bernoulli random variables, which guarantees that $\mathbb{P}\left(|C_n - np| \geq k_n \text{ for some } n \geq 1\right) \leq \alpha/2$ (see Appendix B).

Now assume that $\tilde{p}(\cdot) \geq p_c$. Then $C_n$ is not necessarily a $\text{Binomial}(n, p_c)$ random variable anymore, but pathwise dominates a $\text{Binomial}(n, p_c)$ random variable. More precisely, it holds that $C_n \overset{d}{\sim} \sum_{i=1}^n \psi_i$, and the random variables $(\psi_i)_i$ are defined by

$$\psi_i = \begin{cases} 1, & \text{if } U_i \leq \tilde{p}(X_{i-1}), \\ 0, & \text{otherwise,} \end{cases}$$

where $(U_i)_i$ is a sequence of $\text{U}(0,1)$ random variables. Use pathwise the same sequence $(U_i)_i$ to also define

$$\varphi_i = \begin{cases} 1, & \text{if } U_i \leq p_c, \\ 0, & \text{otherwise.} \end{cases}$$

By construction, $\sum_{i=1}^n \varphi_i \leq \sum_{i=1}^n \psi_i$. Then, using (3.20), it follows that

$$\begin{aligned}\mathbb{P}\left(\log A_n \leq a_n \text{ for some } n \geq 1\right) &= \mathbb{P}\left(C_n - np_c \leq -k_n \text{ for some } n \geq 1\right)\\ &= \mathbb{P}\left(\sum_{i=1}^n \psi_i - np_c \leq -k_n \text{ for some } n \geq 1\right)\\ &\leq \mathbb{P}\left(\sum_{i=1}^n \varphi_i - np_c \leq -k_n \text{ for some } n \geq 1\right)\\ &\leq \frac{\alpha}{2},\end{aligned}$$

where the last inequality follows as in the case when $\tilde{p}(\cdot) \equiv p_c$. □

Analogous to the confidence intervals $(J_n(\alpha))_n$, the sequential confidence intervals $(K_n(\alpha))_n$ are only useful when their lengths converge to 0 reasonably fast. The next proposition shows that $(|K_n(\alpha)|)_n$ converges to 0 at a geometric rate with high probability.

**Proposition 10.** *Suppose that Assumption $\mathcal{A}$ in Section 3.1 holds. Let $p_c \in [0.85, 1)$, $r \in (0, d - q_c\beta)$ and $\alpha \in (0, 1)$. Then there exists a constant $N_2 \in \mathbb{N}$ such that*

$$\mathbb{P}(|K_n(\alpha)| > e^{-rn} \textit{ for some } n \geq N_2) \leq \alpha/2.$$

*Proof.* Proposition 8 shows that $d - q_c\beta > 0$ for $p_c \geq 0.85$, so fix arbitrary $r \in (0, d - q_c\beta)$.

Analogous to the proof of Proposition 8 it follows that

$$\begin{aligned}\mathbb{P}\left(|K_n(\alpha)| > e^{-rn}\right) &\leq \mathbb{P}\left(W_n \geq \frac{a_n - rn}{\beta} \text{ for some } n \geq 1\right)\\ &\leq \mathbb{P}\left(\overline{W}_n \geq \frac{a_n - rn}{\beta} \text{ for some } n \geq 1\right),\end{aligned}$$

where $W_n$ is the number of incorrect signals observed up to time $n$ and $\overline{W}_n \sim \text{Bernoulli}(n, q_c)$, and

$$\begin{aligned}\mathbb{P}\left(\overline{W}_n \geq \frac{a_n - rn}{\beta} \text{ for some } n \geq 1\right) &\leq \mathbb{P}\left(\overline{W}_n \geq \frac{a_n - rn}{\beta} \text{ for some } n \geq N_2\right)\\ &\leq \frac{\alpha}{2},\end{aligned}$$

where the second inequality follows by Lemma 6 (below). □

**Lemma 6.** *Let $p \in [0.85, 1)$, $q = 1 - p$, $d = p\log(2p) + q\log(2q)$, $\beta = \log(p/q)$, $r \in (0, d - q\beta)$, $\alpha \in (0, 1)$, $a_n$ as defined in (3.17) and $W_n \sim$ Binomial$(n, q)$. Then there exists a constant $N_2 \in \mathbb{N}$ such that*

$$\mathbb{P}\left(W_n \geq \frac{a_n - rn}{\beta} \textit{ for some } n \geq N_2\right) \leq \frac{\alpha}{2}.$$

The proof of Lemma 6 is provided in Appendix A.

Similarly to Theorem 5, the next theorem combines the previous two propositions into one statement. The proof is again a direct application of Boole's inequality and is omitted.

**Theorem 6.** *Suppose that Assumption $\mathcal{A}$ in Section 3.1 holds. Let $p_c \in [0.85, 1)$, $\alpha \in (0,1)$ and $r \in (0, d - q_c\beta)$. Then there exists a constant $N_2 \in \mathbb{N}$ such that*

$$\mathbb{P}\big(x^* \in K_n(\alpha), |K_n(\alpha)| \leq e^{-rn} \text{ for all } n \geq N_2\big) \geq 1 - \alpha.$$

Often a user of a stochastic root-finding algorithm is interested in how much simulation effort is needed to achieve a certain confidence in the algorithm's output. For example, it would be convenient if the user could specify an accuracy $\delta > 0$ as well as a confidence parameter $\alpha \in (0,1)$, and the search algorithm would terminate automatically as soon as it locates an estimate $\hat{X}_n$ that is contained in a $\delta$-ball of $x^*$ with at least probability $1-\alpha$. Proposition 9 provides exactly a method for achieving this. Let $\hat{X}_n$ be the midpoint of $K_n(\alpha)$ after $n$ PBA iterations. If $\tau = \inf\{n \geq 1 : |K_n(\alpha)| \leq 2\delta\}$ then Proposition 9 implies that

$$\mathbb{P}\big(|\hat{X}_\tau - x^*| > \delta\big) \leq \alpha/2.$$

Proposition 10 furthermore shows that if $p_c \in [0.85, 1)$ then, for $r \in (0, d - q_c\beta)$,

$$\mathbb{P}\left(\tau \leq -\frac{\log 2\delta}{r} \wedge N_2\right) \leq \frac{\alpha}{2}.$$

For fixed $r, p_c$ and $\alpha$, $N_2$ is a fixed constant, and

$$\mathbb{P}\left(\tau \leq -\frac{\log 2\delta}{r}\right) \leq \frac{\alpha}{2}, \qquad \text{as } \delta \to 0.$$

This provides a probabilistic guarantee that the PBA returns, in geometric time, an estimate $\hat{X}_\tau$ that is inside of a $\delta$-ball of $x^*$.

## 3.5 Asymptotic Rate of Convergence in Macro Time

In the previous sections, we showed that the PBA provides finite-time guarantees in the form of true confidence intervals for $x^*$. While this is an exciting new development for stochastic root-finding algorithms, it is also important to better understand the asymptotic behavior of the PBA. The results from previous sections (Theorem 2, 5 and 6) suggest that the PBA produces an estimator $\hat{X}_n$ that converges towards $x^*$ at a geometric rate in macro time. We now confirm this explicitly.

One simple estimator that achieves this convergence rate is choosing a point of maximum posterior density at each iteration, that is, $X_n^M \in \text{argmax}_{x\in[0,1]} f_n(x)$. Such a point $X_n^M$ always exists, though is usually not unique, in which case we can choose a point at random out of the set of possible candidates. To determine $X_n^M$ it is not necessary to construct the confidence interval $J_n(\alpha)$, which can save computational power if a user is only interested in a quickly converging point estimator of $x^*$.

**Theorem 7.** *Suppose that Assumption $\mathcal{A}$ in Section 3.1 holds and $p_c \in [0.85, 1)$. For any $r > 0$ satisfying*

$$\left(\frac{\sqrt{2r}\beta}{d - r - q_c\beta}\right)^2 < 1, \tag{3.21}$$

*the following convergence properties hold:*

*(a) There exists a set $B \subseteq \Omega$ of probability one, such that for all $\omega \in B$ there exists $N(\omega) \in \mathbb{N}$ such that $x^* \in J_n(e^{-rn})(\omega)$ for all $n \geq N(\omega)$, that is,*

$$\mathbb{P}\left(\lim_{n\to\infty} \mathbb{1}\left\{x^* \in J_n(e^{-rn})\right\} = 1\right) = 1;$$

*(b) $\mathbb{E}\left[|J_n(e^{-rn})|\right] = O(e^{-rn})$;*

*(c)* $e^{rn}|J_n(e^{-rn})| \to 0$ *almost surely as* $n \to \infty$*;*

*(d) For every* $n \in \mathbb{N}$*, let* $\hat{X}_n$ *be an arbitrary point in* $J_n(e^{-rn})$*, then* $\mathbb{E}[|\hat{X}_n - x^*|] = O(e^{-rn})$*;*

*(e)* $e^{rn}|\hat{X}_n - x^*| \to 0$ *almost surely as* $n \to \infty$*;*

*(f) For every* $n \in \mathbb{N}$*, let* $X_n^M \in \operatorname{argmax}_{x\in[0,1]} f_n(X)$*, then* $\mathbb{E}[|X_n^M - x^*|] = O(e^{-rn})$*;*

*(g)* $e^{rn}|X_n^M - x^*| \to 0$ *almost surely as* $n \to \infty$*.*

*Proof. Part (a):* By Proposition 7 it holds that $\mathbb{P}(x^* \notin J_n(e^{-rn})) \leq e^{-rn}$ for all $n \in \mathbb{N}$, hence $\sum_{n=0}^{\infty} \mathbb{P}\left(x^* \notin J_n(e^{-rn})\right) \leq \sum_{n=0}^{\infty} e^{-rn} < \infty$ and, by the Lemma of Borel-Cantelli, it follows that

$$\mathbb{P}\left(\lim_{n\to\infty} \mathbb{1}\{x^* \in J_n(e^{-rn})\} = 0\right) = 0.$$

*Part (b):* By the law of total probability,

$$\begin{aligned}\mathbb{E}[|J_n(e^{-rn})|] &= \mathbb{E}\left[|J_n(e^{-rn})| \middle| |J_n(e^{-rn})| \leq e^{-rn}\right] \mathbb{P}\left(|J_n(e^{-rn})| \leq e^{-rn}\right) \\ &\quad + \mathbb{E}\left[|J_n(e^{-rn})| \middle| |J_n(e^{-rn})| > e^{-rn}\right] \mathbb{P}\left(|J_n(e^{-rn})| > e^{-rn}\right) \\ &\leq e^{-rn} + \mathbb{P}\left(|J_n(e^{-rn})| > e^{-rn}\right),\end{aligned}$$

which follows by the trivial bounds $\mathbb{P}(\cdot) \leq 1$ and $|J_n(e^{-rn})| \leq 1$. It remains to show that

$$\mathbb{P}\left(|J_n(e^{-rn})| > e^{-rn}\right) \leq e^{-rn} \tag{3.22}$$

for all $n \geq N_3(r)$, where $N_3(r) \in \mathbb{N}$ is a constant depending on $r$ (we often just write $N_3$). Since $p_c \in [0.85, 1)$ and $r < (d - q_c\beta)$ Proposition 8 shows that (3.22) holds for all $n \geq N_1(e^{-rn}, r, p_c)$, where $N_1$ is defined by (3.11). Consequently, to show that there exists a constant $N_3$ such that (3.22) holds for all $n \geq N_3$ it is sufficient to show that $N_1(e^{-rn}, r, p_c)$ grows slower than $n$ in $n$, that is,

$$\limsup_{n\to\infty} \frac{N_1(e^{-rn}, r, p_c)}{n} < 1. \tag{3.23}$$

Using the definition of $N_1$,

$$\begin{aligned} \frac{N_1(e^{-rn}, r, p_c)}{n} &= \frac{2\log(2/e^{-rn})\left(\frac{\beta}{d-r-q_c\beta}\right)^2}{n} \\ &= \frac{2\log 2\left(\frac{\beta}{d-r-q_c\beta}\right)^2}{n} + 2r\left(\frac{\beta}{d-r-q_c\beta}\right)^2, \end{aligned}$$

and (3.23) holds if

$$2r\left(\frac{\beta}{d-r-q_c\beta}\right)^2 < 1,$$

which is assured by the assumption that $r > 0$ satisfies (3.21).

*Part (c):* Let $r^* > r' > r > 0$ be three constants satisfying (3.21). Then, as seen in the proof of part *(b)*, it holds that $\mathbb{P}(|J_n(e^{-r'n})| > e^{-r'n}) \leq e^{-r'n}$ for large $n \geq N_3$. Also, $|J_n(e^{-rn})| \leq |J_n(e^{-r'n})|$ for all $n \in \mathbb{N}$. Hence, for $n \geq N_3$,

$$\begin{aligned} \mathbb{P}\big(|J_n(e^{-rn})| > e^{-r'n}\big) &\leq \mathbb{P}\big(|J_n(e^{-r'n})| > e^{-r'n}\big) \leq e^{-r'n} \\ \mathbb{P}\big(|J_n(e^{-rn})| > e^{rn-rn-r'n}\big) &\leq e^{-r'n} \\ \mathbb{P}\big(e^{rn}|J_n(e^{-rn})| > e^{-(r'-r)n}\big) &\leq e^{-r'n}. \end{aligned}$$

Now consider arbitrary $\varepsilon > 0$ and define $N_\varepsilon = \log(1/\varepsilon)/(r' - r)$. Then

$$\mathbb{P}\big(e^{rn}|J_n(e^{-rn})| \geq \varepsilon\big) \leq e^{-r'n}$$

for all $n \geq N_\varepsilon \vee N_3$. Therefore,

$$\sum_{n=1}^{\infty} \mathbb{P}\big(e^{rn}|J_n(e^{-rn})| > \varepsilon\big) \leq N_\varepsilon \vee N_3 + \frac{1}{1-e^{-r}} < \infty.$$

By the Lemma of Borel-Cantelli it follows that $\mathbb{P}\big(e^{rn}|J_n(e^{-rn})| > \varepsilon \text{ i.o.}\big) = 0$. Since $\varepsilon > 0$ was chosen arbitrarily the claim follows.

*Part (d):* Note that $\big\{|\hat{X}_n - x^*| \leq e^{-rn}\big\} \subseteq \big\{|J_n(e^{-rn})| \leq e^{-rn}, x^* \in J_n(e^{-rn})\big\}$, and Theorem 5 shows that

$$\mathbb{P}\big(|\hat{X}_n - x^*| \leq e^{-rn}\big) \geq 1 - e^{-rn}, \tag{3.24}$$

for $n \geq N_1(e^{-rn}, r, p_c)$. As seen in the proof of part *(b)*, there exists a constant $N_3$ such that (3.24) holds for all $n \geq N_3$. Then, for all $n \geq N_3$,

$$\begin{aligned}\mathbb{E}\big[|\hat{X}_n - x^*|\big] &= \mathbb{E}\big[|\hat{X}_n - x^*| \big| |\hat{X}_n - x^*| \leq e^{-rn}\big]\mathbb{P}\big(|\hat{X}_n - x^*| \leq e^{-rn}\big) \\ &\qquad + \mathbb{E}\big[|\hat{X}_n - x^*| \big| |\hat{X}_n - x^*| > e^{-rn}\big]\mathbb{P}\big(|\hat{X}_n - x^*| > e^{-rn}\big) \\ &\leq e^{-rn} + e^{-rn},\end{aligned}$$

where we use the trivial bounds $|\hat{X}_n - x^*| \leq 1$ and $\mathbb{P}(\cdot) \leq 1$.

*Part (e):* Let $r^* > r' > r > 0$ be three constants satisfying (3.21). Then, as seen in the proof of part *(d)*,

$$\mathbb{P}\big(|\hat{X}_n - x^*| > e^{-r'n}\big) \leq e^{-r'n}, \tag{3.25}$$

for all $n \geq N_3$. By the same arguments as in part *(c)* the claim follows.

*Parts (f) and (g):* Note that $X_n^M \in J_n(\alpha)$ for all $\alpha \in (0, 1)$ and hence *(f)* and *(g)* follow immediately by *(d)* and *(e)*. □

This shows that the PBA produces an estimator $\hat{X}_n$ that converges to $x^*$ at a geometric rate. This estimator, however, is not equal to $X_n$, that is, the median of the posterior density $f_n$. Although Theorem 2 and 7 as well as empirical results suggest that the sequence of medians $(X_n)_n$ also converges to $x^*$ at a geometric rate in macro time, there does not exist a formal proof of this result at the moment, and we state it as a conjecture.

**Conjecture 1.** *Suppose that Assumption $\mathcal{A}$ in Section 3.1 holds. Denote with $(X_n)_n$ the sequence of medians defined by the PBA. Then there exists $r > 0$ such that $\mathbb{E}[|X_n - x^*|] = O(e^{-rn})$.*

At first sight, Conjecture 1 does not seem relevant since Theorem 7 already provides a simple estimator, namely $X_n^M$, that converges to $x^*$ at a geometric rate. However, when investigating the rate of convergence in wall-clock time in the following section, the difference between $X_n^M$ and $X_n$ will be rather significant as the number of function evaluations required by the test of power one between two macro iterations and, in turn, the convergence rate in wall-clock time, strongly depends on the location of the actual measurement point $X_n$. Furthermore, altering the PBA in such a way that it always evaluates the function $g$ at $X_n^M$ instead of the median $X_n$ will not provide a useful stochastic root-finding algorithm as this variation mimics noise-free bisection search and a single wrong signal $\widetilde{Z}_n$ leads the search astray.

As a final remark, the proof of a slightly weaker conjecture on the sample path behavior of $(X_n)_n$ would be sufficient to provide useful convergence results in wall-clock time.

**Conjecture 2.** *Suppose that Assumption $\mathcal{A}$ in Section 3.1 holds. Denote with $(X_n)_n$ the sequence of medians defined by the PBA. For $r > 0$ define $M_r = \sum_{i=0}^{\infty} \mathbb{1}\{|X_i - x^*| > e^{-ri}\}$. Then there exists $r > 0$, such that $\mathbb{E}[N_r] < \infty$.*

Let us show that Conjecture 1 implies Conjecture 2. By Conjecture 1 there exists $r > 0$ and constant $N_4$ such that $\mathbb{E}[|X_n - x^*|] \leq e^{-rn}$ for all $n \geq N_4$.

Consider $0 < r' < r$. Then

$$\begin{aligned}\mathbb{E}\big[M_{r'}\big] &= \mathbb{E}\left[\sum_{i=0}^{\infty} \mathbb{1}\big\{|X_i - x^*| > e^{-r'i}\big\}\right] \\ &= \sum_{i=0}^{\infty} \mathbb{P}\big(|X_i - x^*| > e^{-r'i}\big) \\ &\leq \sum_{i=0}^{\infty} e^{r'i}\mathbb{E}\big[|X_i - x^*|\big] \leq \sum_{i=0}^{N_4-1} e^{r'i} + \sum_{i=N_4}^{\infty} e^{r'i}e^{-ri} < \infty,\end{aligned}$$

where the first inequality follows by Markov's inequality, the second inequality by the trivial bound $\mathbb{E}[|X_i - x^*|] \leq 1$ and the third by the condition that $0 < r' < r$.

## 3.6 Asymptotic Rate of Convergence in Wall-Clock Time

In this section we study the rate of convergence of the PBA in wall-clock time, that is, $T_n = \sum_{i=0}^{n} N_i$, where $N_i$ is the stopping time of the test of power one used to generate the signal $\widetilde{Z}_i(X_i)$.

The convergence behavior of the PBA in wall-clock time strongly depends on the form of the function $g$, especially its behavior at the root $x^*$. In Section 3.6.1 we first investigate the rate of convergence when the function $g$ is discontinuous at the root $x^*$. In Section 3.6.2 we then consider the case when $g(x) \to 0$ as $x \to x^*$. Recall, there is a strong relationship between the functions $g$ and $\tilde{g}$ and we always assume that if a property (such as a linear growth condition, or continuity) holds for $g$, then it also holds for $\tilde{g}$, and vice versa.

### 3.6.1 The Case where $g(x)$ is Discontinuous at $x^*$

The case where $g(x)$ has a discontinuity at $x^*$ appears in settings such as edge detection (Castro and Nowak, 2008a) or in simulation-optimization problems with nonsmooth objective functions. For the latter case, Lim (2011) shows that an SA algorithm (under modest technical assumptions) produces estimates of $x^*$ that satisfy $\mathbb{E}[|X_T - x^*|] = O(T^{-1})$. We show in Corollary 3 that—at least for the one-dimensional case—this rate is geometric when using the PBA. To this end, we extend Theorem 7 to show that the length of the confidence intervals $J(e^{-rn})$ converges to 0 at a geometric rate in wall clock time as well.

**Proposition 11.** *Suppose that Assumption $\mathcal{A}$ in Section 3.1 holds. Let $p_c \in [0.85, 1)$, and $r > 0$ such that* (3.21) *holds. If there exists $c > 0$ such that $|g(x)| \geq c$ for all $x \neq x^*$, then there exists $r' > 0$ such that the sequence $\left(|J_n(e^{-rn})|e^{r'T_n}\right)_n$ is tight, where $J_n(e^{-rn})$ is the $(1 - e^{-rn})$-confidence interval of $x^*$ defined by* (3.8).

*Proof.* First note that the condition $|g(x)| > c$ for all $x \neq x^*$ implies that $|\tilde{g}(x)| > c'$ for all $x \neq x^*$ where $c' > 0$ is some constant, since the noise distribution has a density and 0 median.

Consider a measurement point $X_i$ and assume that $\tilde{g}(X_i) > 0$ (if $\tilde{g}(X_i) < 0$ the arguments follow analogously). Define $N_i^1 = \inf\{m \geq 1 : S_{i,m} \geq k_m\}$ (this stopping time differs from $N_i$ as it only considers exits through the upper boundary), where the boundary $(k_m)_m$ is defined in (B.6), and

$$N_i^c = \inf\Big\{m \geq 1 : S_{i,m} - 2\sum_{j=1}^{m} Q_j \geq 2k_n\Big\},$$

where $(Q_j)_j$ is a sequence of independent Bernoulli$((\tilde{g}(X_i) - c')/2)$ random variables. By construction, $N_i \leq N_i^1 \leq N_i^c$ for all $i \in \mathbb{N}_0$ and it follows that

$$T_n \leq \sum_{i=0}^{n} N_i^c. \tag{3.26}$$

Note that $(N_i^c)_i$ forms an iid sequence of random variables, each corresponding to an independent first hitting time of the upper boundary $(k_m)_m$ by a simple random walk with drift $c$. Furthermore, $\mathbb{E}[N_i^c] < \infty$ (see, for example, Theorem 4.5.1 in Gut, 2009).

Let $\varepsilon > 0$ and define $N_5$ large enough such that

$$\mathbb{P}\big(|J_n(e^{-rn})| < e^{-rn}\big) \geq 1 - \varepsilon/2, \tag{3.27}$$

for all $n \geq N_5$ (such an $N_5$ always exists as shown in (3.22)). Next, choose $0 < r' < r/\mathbb{E}[N_i^c]$, which always exists since $r > 0$ and $0 < \mathbb{E}[N_i^c] < \infty$. The bound (3.26) implies that

$$\mathbb{P}\big(|J_n(e^{-rn})|e^{r'T_n} \leq 1\big) \geq \mathbb{P}\big(|J_n|e^{r'\sum_{i=0}^{n} N_i^c} \leq 1\big),$$

then, by the law of total probability and for arbitrary $n \geq N_5$,

$$\begin{aligned}
&\mathbb{P}\big(|J_n(e^{-rn})|e^{r'\sum_{i=0}^n N_i^c} \leq 1\big) \\
&\quad = \mathbb{P}\left(|J_n(e^{-rn})|e^{r'\sum_{i=0}^n N_i^c} \leq 1 \middle| |J_n(e^{-rn})| < e^{-rn}\right) \mathbb{P}\left(|J_n(e^{-rn})| < e^{-rn}\right) \\
&\qquad + \mathbb{P}\left(|J_n(e^{-rn})|e^{r'\sum_{i=0}^n N_i^c} \leq 1 \middle| |J_n(e^{-rn})| \geq e^{-rn}\right) \mathbb{P}\left(|J_n(e^{-rn})| \geq e^{-rn}\right) \\
&\quad \geq \mathbb{P}\left(|J_n(e^{-rn})|e^{r'\sum_{i=0}^n N_i^c} \leq 1 | |J_n(e^{-rn})| < e^{-rn}\right) (1 - \varepsilon/2) \quad \text{(by (3.27))} \\
&\quad \geq \mathbb{P}\left(e^{-rn} e^{r'\sum_{i=0}^n N_i^c} \leq 1\right) (1 - \varepsilon/2) \\
&\quad = \mathbb{P}\left(-rn + r'\sum_{i=0}^n N_i^c \leq 0\right) (1 - \varepsilon/2) \\
&\quad = \mathbb{P}\left(r'\sum_{i=0}^n N_i^c \leq rn\right) (1 - \varepsilon/2) \\
&\quad = \mathbb{P}\left(\frac{1}{n}\sum_{i=0}^n N_i^c \leq \frac{r}{r'}\right) (1 - \varepsilon/2) \\
&\quad \geq \mathbb{P}\left(\frac{1}{n}\sum_{i=0}^n N_i^c \leq \mathbb{E}[N_i^c] + \delta\right) (1 - \varepsilon/2),
\end{aligned}$$

where $\delta = r/r' - \mathbb{E}[N_i^c] > 0$. By the strong law of large numbers, there exists $N_6 \in \mathbb{N}$ such that $\mathbb{P}(n^{-1}\sum_{i=0}^n N_i^c \leq \mathbb{E}[N_i^c] + \delta) \geq (1 - \varepsilon/2)$ for all $n \geq N_6$ (even though the denominator is $n$ instead of $(n+1)$ the bound still holds since $\mathbb{E}[N_i^c] < \infty$). Now, for $n \geq N_5 \vee N_6$, it holds that

$$\mathbb{P}\big(|J_n(e^{-rn})|e^{r'T_n} \leq 1\big) \geq (1 - \varepsilon/2)(1 - \varepsilon/2) > 1 - \varepsilon, \tag{3.28}$$

(here we assume that $\varepsilon < 1$, but if $\varepsilon \geq 1$ then the bound (3.28) holds trivially) and hence $\liminf_{n\to\infty} \mathbb{P}(|J_n|(e^{-rn})e^{r'T_n} \leq 1) > 1 - \varepsilon$. Since this holds for any chosen $\varepsilon > 0$ the sequence $(|J_n(e^{-rn})|e^{r'T_n})_n$ is tight. □

As an almost immediate consequence, $(X_n^M)_n$, that is, a sequence of points with maximum posterior density, converges to $x^*$ at a geometric rate in wall-clock time if $g$ is discontinuous at $x^*$.

**Corollary 3.** *Assume the same setting as in Proposition 11.*

*(a) Let* $\hat{X}_n$ *be an arbitrary point in* $J_n(e^{-rn})$*. Then there exists* $r' > 0$ *such that* $\left(|\hat{X}_n - x^*|e^{r'T_n}\right)_n$ *is tight;*

*(b) Let* $X_n^M \in \operatorname{argmax}_{x\in[0,1]} f_n(x)$*. Then there exists* $r' > 0$ *such that* $\left(|X_n^M - x^*|e^{r'T_n}\right)_n$ *is tight.*

*Proof. Part (a):* In the proof of Proposition 11, replace (3.27) by

$$\mathbb{P}(|\hat{X}_n - x^*| < e^{-rn}) \geq 1 - \varepsilon/2,$$

for all $n \geq N_7$ (such a constant $N_7$ always exists as shown in (3.24)). By the same arguments given in the proof of Proposition 11 the claim follows.

*Part (b):* As $X_n^M \in J_n(\alpha)$ for all $\alpha \in (0,1)$, the claim follows by *(a).* □

### 3.6.2 The Case where $g(x)$ is Continuous at $x^*$

We now consider the case where $g(x)$, and in turn $\tilde{g}(x)$, are continuous at $x^*$, that is, $g(x) \to 0$ and $\tilde{g}(x) \to 0$ as $x \to x^*$. We believe this to be the dominant setting in applications, and hence the more important of the two cases.

Using the PBA with power one tests in this setting results in complex behavior arising from the fact that the power one test requires more samples the closer to $x^*$ we are measuring. Let $\theta$ denote the drift of a simple random walk and $N(|\theta|)$ the stopping time of the corresponding test of power one. A sample path argument shows that $\mathbb{E}[N(|\theta|)]$ is decreasing in $|\theta|$, and it can be shown that $\mathbb{E}[N(|\theta|)] \to \infty$ as $|\theta| \to 0$. More precisely, Farrell (1964) shows that for *any* test of power one for the hypothesis $\theta > 0$ versus $\theta < 0$ it holds that $\lim_{\theta \to 0} |\theta|^2 \mathbb{E}[N(|\theta|)] = \infty$ (as long as the noise distribution belongs to the exponential family). So, in wall-clock time the sequence $(X_n)_n$ slows down dramatically as $g(X_n) \to 0$, posing the following question: What kind of rate of convergence can we still achieve in this setting?

To at least partly answer the above question we first show that $(X_n)_n$ converges to $x^*$ at a rate that is slower than $O(T_n^{-1/2})$, that is, the PBA fails to achieve the same asymptotic rate of convergence as optimal SA-type algorithms are able to attain. While this result is somewhat discouraging, we then show that by averaging the samples $(X_n)_n$ a sequence of estimators can be constructed whose expected absolute residuals converge at an asymptotic rate that is only slightly slower than the optimal $O(T_n^{-1/2})$ rate. This result, however, assumes that Conjecture 2 as stated in Section 3.5 is true.

**Theorem 8.** *Suppose that*

*(i)* $\tilde{g}(x)$ *is bounded away from 0, when* $x$ *is bounded away from* $x^*$*;*

*(ii) there exists an* $\varepsilon > 0$ *and a constant* $c > 0$*, such that* $|\tilde{g}(x)| \leq c|x - x^*|$ *for all* $x \in (x^* - \varepsilon, x^* + \varepsilon)$*;*

*(iii)* $(X_n)_n$ *is a sequence of random variables in* $[0,1]$ *such that* $\mathbb{P}(X_n \neq x^*) = 1$ *for all* $n \in \mathbb{N}$ *and*

$$X_n \to x^* \text{ almost surely as } n \to \infty;$$

*(iv) at each measurement point* $X_n$ *a test of power one is used to construct the signal* $\widetilde{Z}_n(X_n)$ *and* $N_n$ *denotes the hitting time of the corresponding test of power one.*

*Let* $T_n = \sum_{i=0}^{n} N_i$ *be the total simulation budget spend up to and including the* $(n+1)$*st measurement point. Then the sequence* $\left(|X_n - x^*|(T_n)^{1/2}\right)_n$ *is not tight.*

In order to prove this theorem we need to better understand the distribution of the hitting time $N_n$, especially as $X_n$ approaches $x^*$. More specifically, we need to be able to say something about $\mathbb{P}(m \leq N_n)$ as $m \to \infty$ and $n \to \infty$. The next lemma (Farrell, 1964) provides the necessary insight into the distribution of $N_n$. Let us motivate its statement intuitively. The hitting time $N_n(|\tilde{g}(X_n)|)$ of the test of power one depends on the drift of the random walk $S_{n,m}(X_n)$, that is, $\tilde{g}(X_n)$. Since $X_n \to x^*$ and $\tilde{g}(x) \to 0$ as $x \to x^*$, the sequence of hitting times $(N_n(X_n))_n$ behaves pathwise like a sequence $(N(\theta_n))_n$, where $\theta_n \to 0$. In other words, providing a statement on $\mathbb{P}(m \leq N_n(X_n))$ is, in some sense, equivalent of providing a statement on $\mathbb{P}(m \leq N(\theta_n))$ for an arbitrary sequence that satisfies $\theta_n \to 0$ as $n \to \infty$. Next, if $m$ is very small and $n$ very large, then it is reasonable that $\mathbb{P}(m \leq N(\theta_n)) \geq 1 - \gamma$, since $\mathbb{P}(m \leq N(0)) \geq 1 - \gamma$ and $\theta_n \to 0$ as $n \to \infty$. Based on this observation it is now possible to make $n(m)$ a function of $m$ that satisfies $n(m) \to \infty$ as $m \to \infty$ (in fact, $n(m)$ converges to $\infty$ much faster than $n$), such that

$$\liminf_{m \to \infty} \mathbb{P}\left(m \leq N(\theta_{n(m)})\right) \geq 1 - \gamma.$$

**Lemma 7** (Farrell, 1964). *Let $S_n(\theta) = \sum_{i=1}^{n} \xi_i(\theta)$ be the random walk with increments $(\xi_i)_i$ and the distribution of the increments $(\xi_i)_i$ belongs to the exponential family. Let $N(\theta) = \inf\{n \geq 1 : S_n(\theta) \geq |k_n|\}$ be the stopping variable of a test of power one of the hypothesis $\theta > 0$ versus the alternative $\theta < 0$, where $(k_n)_n$ is the curved boundary of this test, such that $\mathbb{P}(N(0) < \infty) = \gamma < 1$. Let $0 < \rho < 1$, and $(\theta_n)_n$ be a sequence such that $0 < \theta_n \leq \rho k_n / n$ for all $n \geq n_0$, where $n_0 \in \mathbb{N}$ is a constant, then*

$$\liminf_{n \to \infty} \mathbb{P}(n \leq N(\theta_n)) \geq 1 - \gamma.$$

Now, we are ready to provide the proof of Theorem 8.

*Proof of Theorem 8.* Since $(|X_n - x^*|^2 T_n) \geq (|X_n - x^*|^2 N_n)$ it is enough to show that $(|X_n - x^*|^2 N_n)_n$ is not tight. Let $D \subseteq \Omega$ be the set of probability one where $\lim_{n\to\infty} X_n = x^*$ and $X_n \neq x^*$ for all $n \in \mathbb{N}$. Such a set exists by Theorem 4 and Assumption (iii). Then, by assumption (ii), for every $\omega \in D$ there exists $N^\varepsilon(\omega)$ such that $|X_n - x^*|c \geq |\tilde{g}(X_n)|$ for all $n \geq N^\varepsilon(\omega)$. Hence it is enough to show that the sequence $(\tilde{g}(X_n)^2 N_n)_n$ is not tight.

For $\omega \in D$ it holds that $|\tilde{g}(X_n)(\omega)| \to 0$ as $n \to \infty$ and $|\tilde{g}(X_n)(\omega)| > 0$ for all $n \in \mathbb{N}$. Choose $0 < \rho < 1$ and let $(m_i(\omega))_i$ and $(q_i(\omega)_i)$ be integer sequences such that $m_i(\omega) \to \infty$ and $q_i(\omega) \to \infty$ as $i \to \infty$, and such that

$$\frac{\rho k_{m_i(\omega)}}{m_i(\omega)} \geq |\tilde{g}(X_{q_i(\omega)})(\omega)| \geq \frac{\rho k_{m_i(\omega)+1}}{m_i(\omega)+1}, \tag{3.29}$$

where $(k_n)_n$ is the curved boundary of the test of power one used at each iteration. Then, by (3.29) and the fact that $\lim_{n\to\infty} n^{1/2} k_n = \infty$ (this is a necessary condition for the test of power one, since otherwise $\mathbb{P}(N(0) < \infty) = 1$ by the law of the iterated logarithm),

$$\lim_{i\to\infty} (m_i(\omega)+1)^{1/2} |\tilde{g}(X_{q_i(\omega)})(\omega)| = \infty. \tag{3.30}$$

So, for arbitrary $M > 0$ there exists $N_8(\omega)$ such that $m_i(\omega)(\tilde{g}(X_{q_i(\omega)}))^2 > M$, and hence

$$m_i \geq \frac{M}{(\tilde{g}(X_{q_i(\omega)}))^2}, \tag{3.31}$$

for all $i \geq N_8(\omega)$. This implies that for $i \geq N_8(\omega)$

$$\mathbb{P}\left(N_{q_i} > \frac{M}{(\tilde{g}(X_{q_i}))^2} \;\middle|\; |\tilde{g}(X_{q_i})|\right)(\omega) \geq \mathbb{P}\big(N_{q_i} \geq m_i \mid |\tilde{g}(X_{q_i})|\big)(\omega), \tag{3.32}$$

which follows since we make the right-hand side in the conditional probability larger and hence reduce this probability.

Therefore, for almost every $\omega \in D$,

$$\mathbb{P}\big(\tilde{g}(X_{q_i})^2 N_{q_i} > M \mid |\tilde{g}(X_{q_i})|\big)(\omega) = \mathbb{P}\bigg(N_{q_i} > \frac{M}{(\tilde{g}(X_{q_i}))^2} \;\bigg|\; |\tilde{g}(X_{q_i})|\bigg)(\omega),$$

and (3.32) shows that

$$\liminf_{i\to\infty} \mathbb{P}\big(\tilde{g}(X_{q_i(\omega)})^2 N_{q_i(\omega)} > M \mid |\tilde{g}(X_{q_i(\omega)})\big) \geq \liminf_{i\to\infty} \mathbb{P}\big(N_{q_i(\omega)} \geq m_i \mid |\tilde{g}(X_{q_i(\omega)})|\big).$$

Recall that Lemma 7 shows that

$$\liminf_{i\to\infty} \mathbb{P}\big(N_{q_i} \geq m_i \mid |\tilde{g}(X_{q_i})|\big)(\omega) \geq 1-\gamma > 0.$$

Since this holds for almost all $\omega \in D$, it follows that

$$\liminf_{i\to\infty} \mathbb{P}((\tilde{g}(X_{q_i}))^2 N_{q_i} > M) > 1-\gamma,$$

and hence $\liminf_{n\to\infty} \mathbb{P}((\tilde{g}(X_n))^2 N_n > M) > 1-\gamma$. Since $M$ was chosen arbitrarily and $\gamma < 1$ it follows that $(\tilde{g}(X_n)^2 N_n)_n$ is not tight. □

The geometric rate of convergence for the PBA is shown in terms of the width of the confidence interval $J_n(e^{-rn})$, respectively, $X_n^M$, in Theorem 7. Theorem 8, on the other hand, shows that $\big(|X_n - x^*|(T_n)^{1/2}\big)_n$ is not tight, where $X_n$ is the median of the posterior density. As it is not necessarily the case that the median $X_n$ falls into the confidence interval $J_n(e^{-rn})$ at every iteration, the next proposition shows that $\big(|J_n(\alpha_n)|(T_n)^{1/2}\big)_n$ cannot be tight either. Whether the sequence $\big(|\hat{X}_n - x^*|(T_n)^{1/2}\big)_n$, where $\hat{X}_n$ is an arbitrary point in $J_n(\alpha_n)$ and the sequence $\big(|X_n^M - x^*|(T_n)^{1/2}\big)_n$, are tight or not is still an open question.

**Proposition 12.** *Consider $r > 0$ such that* (3.21) *holds. Under the same setting as in Theorem 8 the sequence $(|J_n(e^{-rn}))|(T_n)^{1/2})_n$ is not tight.*

*Proof.* In this proof we write $J_n$ instead of $J_n(e^{-rn})$. Assume that $(|J_n|(T_n)^{1/2})_n$ is tight. Then, by the same arguments as in the proof of Theorem 8 this implies that

$$\mathbb{P}(\liminf_{n\to\infty} |J_n|^2 N_n = \infty) = 0.$$

Further, the proof of Theorem 8 shows that there exists a set $D_1 \subseteq \Omega$ with positive probability mass on which $\liminf_{n\to\infty}(X_n - x^*)^2 N_n = \infty$. This implies that for all $\omega \in D_1$, there exists $N_9(\omega) \in \mathbb{N}$ such that $|J_n| < |X_n - x^*|$ for all $n \geq N_9(\omega)$.

Next, Proposition 7 part *(a)* shows that on a set of probability one, say $D_2$, $\lim_{n\to\infty} \mathbb{1}\{x^* \in J_n\} = 1$ holds. So, for every $\omega \in D_2$ there exists $N_{10}(\omega)$ such that $x^* \in J_n$ for all $n \geq N_{10}(\omega)$.

Consequently, (1) $x^* \in J_n(\omega)$ and (2) $X_n(\omega) \in J_n(\omega)$ for all $n \geq N_{11}(\omega) = N_9(\omega) \vee N_{10}(\omega)$ and $\omega \in D_3 = D_1 \cap D_2$ (the set $D_3$ has positive probability). In other words, no more measurements are taken inside of the interval $J_n$ after time $N_{11}(\omega)$ and the interval $J_n$ contains from time $N_{11}(\omega)$ onwards the sought-after point $x^*$.

Based on the above observations we now construct a contradiction to $X_n \to x^*$ almost surely. For this, recall how the interval $J_n$ is constructed: at time $N_{11}(\omega)$ it is a union of a finite number of disjoint intervals defined by the measurement points $(X_i)_{i=0}^{N_{11}(\omega)-1}$ and exactly one such interval contains $x^*$. Let us denote this interval that contains $x^*$ at time $N_11(\omega)$ as $I^*(\omega)$. Now, consider what can happen to this interval from time $N_{11}(\omega)$ onwards:

(a) The whole interval $I^*(\omega)$ cannot leave the set $J_n(\alpha)$, since otherwise (1) would be violated,

(b) The interval $I^*(\omega)$ cannot shrink further, since such a shrinking would require that $X_n \in I^*(\omega)$ violating (2) (since $I^*(\omega) \subseteq J_n$).

Finally, (a) and (b) imply that for all $n \geq N_{11}(\omega)$:

(i) $|I^*(\omega)| > 0$ stays constant, and $x^*$ is in the interior of $I^*$. (Since, by Assumption $\mathcal{A}$(ii), $x^*$ cannot be on the endpoints of $I^*$ which correspond to previous measurement points. If one of the endpoints is either 0 or 1, then we allow that $x^* = 0$ or $x^* = 1$ can occur.)

(ii) $X_n \notin I^*$.

Since this holds for all $\omega \in D_3$, which is a set of positive probability, this contradicts $X_n \to x^*$ almost surely (Theorem 4). □

The last two propositions show that if we consider the sequence $(X_n)_n$ or the sequence of confidence intervals $(J_n)_n$ we cannot recover the optimal rate of convergence $O(T_n^{-1/2})$. But, an averaging of the sequence $(X_n)_n$, similar to Polyak-Ruppert averaging for SA algorithms (Polyak, 1990; Ruppert, 1991), might be helpful for improving the asymptotic rate of convergence.

For the PBA a natural averaging estimator of $x^*$ would be $\overline{X}_T = 1/T_n \sum_{i=0}^{n} N_i X_i$. Intuitively, such an averaging estimator is very appealing, since the estimator is likely to spend a lot of time at measurement points close to $x^*$, and hence the average should be dominated by such points.

Analyzing such an estimator seems to be extremely challenging, and it is still an open question if this averaging (or any other averaging scheme) is able to recover the same asymptotic convergence rate as SA-type algorithms are able to attain, that is, $O(T^{-1/2})$ in terms of convergence in distribution. Empirical experiments show that this might be the case; see Chapter 4.

The next theorem shows that, under mild assumptions on the underlying function $g$, respectively, $\tilde{g}$, a slightly different averaging scheme can be used to almost recover this rate of convergence, that is, the expected absolute residuals converge at the rate $O\left(T_n^{-\frac{1}{2}+\varepsilon}\right)$ for any $\varepsilon > 0$ (which also implies convergence in distribution at this rate). This theorem, however, assumes that Conjecture 2 holds, something that is still an open question.

**Theorem 9.** *Suppose that*

*(a) Assumption $\mathcal{A}$ in Section 3.1 holds;*

*(b) Conjecture 2 in Section 3.5 is true, that is, there exists a constant $r > 0$ such that $\mathbb{E}\left[M_r\right] < \infty$, where*

$$M_r = \sum_{i=0}^{\infty} \mathbb{1}\left\{|X_i - x^*| > e^{-ri}\right\};$$

*(c) there exists a constant $k > 0$, such that $|\tilde{g}(x)| \geq k|x - x^*|$ for all $x \in [0, 1]$.*

*For $\varepsilon > 0$ define*

$$\hat{X}_n(\varepsilon) = \frac{1}{\sum_{i=0}^{n} N_i^{\frac{1}{2}-\varepsilon}} \sum_{i=0}^{n} N_i^{\frac{1}{2}-\varepsilon} X_i. \tag{3.33}$$

*Then $\mathbb{E}\left[|\hat{X}_n(\varepsilon) - x^*| T_n^{\frac{1}{2}-\varepsilon}\right] = O(1)$.*

Assumption (c) is formulated in terms of $\tilde{g}$, but for reasonable noise distributions it translates to a similar assumption on the function $g$. This assumption is necessary such that $\tilde{g}(x)$ is not too "flat" around the root $x^*$, which is considered a difficult case for stochastic root-finding problem in terms of asymptotic convergence behavior. Similar assumptions are needed to prove rate of convergence results for SA-type algorithms, for example, they often assume that $g'(x^*) \neq 0$ (see discussion in Section 1.3).

*Proof.* The claim follows trivially if $\varepsilon \geq 1/2$ by the fact that $|\hat{X}_n(\varepsilon) - x^*| \leq 1$. So assume that $0 < \varepsilon < 1/2$ and define $\gamma > 0$ and $\delta < 1/2$ such that

$$\delta = \frac{1}{2(1+\gamma)} = \frac{1}{2} - \varepsilon. \tag{3.34}$$

It is sufficient to show that $\limsup_{n\to\infty} \mathbb{E}\big[|\hat{X}_n(\varepsilon) - x^*| T_n^\delta\big] < \infty$.

Observe that

$$\begin{aligned}
&\left|\hat{X}_n(\varepsilon) - x^*\right| T_n^\delta \\
&\quad = \left(\left|\hat{X}_n(\varepsilon) - x^*\right|^{1/\delta} T_n\right)^\delta \\
&\quad = \left(\left|\frac{1}{\sum_{i=0}^n N_i^\delta} \sum_{i=0}^n N_i^\delta X_i - x^*\right|^{1/\delta} T_n\right)^\delta \\
&\quad = \left(\left|\frac{1}{\sum_{i=0}^n N_i^\delta} \sum_{i=0}^n N_i^\delta \left(X_i - x^*\right)\right|^{1/\delta} T_n\right)^\delta \\
&\quad = \left(\left(\frac{1}{\sum_{i=0}^n N_i^\delta}\right)^{1/\delta} \left|\sum_{i=0}^n N_i^\delta \left(X_i - x^*\right)\right|^{1/\delta} T_n\right)^\delta \\
&\quad \leq \left(\left(\frac{1}{\sum_{i=0}^n N_i^\delta}\right)^{1/\delta} \left(\sum_{i=0}^n N_i^\delta \left|X_i - x^*\right|\right)^{1/\delta} T_n\right)^\delta \\
&\quad \leq \left(\left(\frac{1}{\left(\sum_{i=0}^n N_i\right)^\delta}\right)^{1/\delta} \left(\sum_{i=0}^n N_i^\delta \left|X_i - x^*\right|\right)^{1/\delta} T_n\right)^\delta,
\end{aligned}$$

since $\|x\|_l \geq \|x\|_1$ for $0 < l < 1$. Then, by the definition of $T_n$,

$$= \left( \left( \sum_{i=0}^{n} N_i^\delta |X_i - x^*| \right)^{1/\delta} \right)^\delta$$
$$= \sum_{i=0}^{n} N_i^\delta |X_i - x^*|,$$

and therefore

$$\begin{aligned} \mathbb{E}\left[ \left| \hat{X}_n(\varepsilon) - x^* \right| T_n^\delta \right] &\leq \mathbb{E}\left[ \sum_{i=0}^{n} N_i^\delta |X_i - x^*| \right] \\ &= \mathbb{E}\left[ \sum_{i=0}^{n} \mathbb{E}\left[ N_i^\delta |X_i - x^*| \middle| \widetilde{\mathcal{G}}_{i-1} \right] \right], \end{aligned}$$

where the last equality follows by the tower property of the conditional expectation. Recall that $\widetilde{\mathcal{G}}_n$ is the $\sigma$-algebra generated by the measurement points $(X_i)_{i=0}^n$ and the signals $(\widetilde{Z}_i)_{i=0}^n$ for $n \in \mathbb{N}_0$, and $\widetilde{\mathcal{G}}_{-1}$ is the trivial $\sigma$-algebra. The remainder of the proof shows that the limit superior of the right-hand side is bounded.

As before denote with $N(\theta)$ the stopping time of the test of power one for a simple random walk with drift $\theta$ and curved boundary $(k_n)_n$ given in (B.6) (this corresponds to the test used to generate the signals $(\widetilde{Z}_n)_n$). As discussed in (B.7), this test satisfies

$$\limsup_{\theta \to 0} \mathbb{E}\left[ N(\theta) \right] \theta^{-2} \log(|\theta|^{-1}) < \infty.$$

Therefore, there exists a constant $\tilde{\theta} > 0$ such that

$$\mathbb{E}[N(\theta)] \leq \begin{cases} \tilde{\theta}^{-(2+\gamma)}, & \text{if } |\theta| > \tilde{\theta}, \\ |\theta|^{-(2+\gamma)}, & \text{if } 0 < |\theta| \leq \tilde{\theta}. \end{cases} \tag{3.35}$$

Let $\bar{r} > 0$ such that $\mathbb{E}[M_{\bar{r}}] < \infty$ (which exists by Assumption (b)) and consider $r \in (0, \bar{r})$, then

$$\mathbb{E}\left[ \sum_{i=0}^{\infty} \mathbb{1}\left\{ k|X_i - x^*| > e^{-ri} \right\} \right] < \infty. \tag{3.36}$$

Recall that $k$ is the slope of a lower linear bound on $|\tilde{g}(x)|$, see Assumption (c). For every fixed realization $\omega \in \Omega$ define three disjoint index sets:

$$\begin{aligned}
I_n^1(\omega) &= \Big\{ i \in \{0, \ldots, n\} : k|X_i(\omega) - x^*| > \tilde{\theta} \Big\} \\
I_n^2(\omega) &= \big\{ i \in \{0, \ldots, n\} : X_i(\omega) \notin I_n^1(\omega), k|X_i(\omega) - x^*| > e^{-ri} \big\} \\
I_n^3(\omega) &= \big\{ i \in \{0, \ldots, n\} : X_i(\omega) \notin I_n^1(\omega), k|X_i(\omega) - x^*| \le e^{-ri} \big\} .
\end{aligned}$$

Then

$$\sum_{i=0}^{n} \mathbb{E}\big[ N_i^\delta |X_i - x^*| \,\big|\, \widetilde{\mathcal{G}}_{i-1} \big](\omega) = \sum_{j=1}^{3} \sum_{i=0}^{n} \mathbb{E}\big[ N_i^\delta |X_i - x^*| \mathbb{1}\{i \in I_n^j\} \,\big|\, \widetilde{\mathcal{G}}_{i-1} \big](\omega), \quad (3.37)$$

and we show that the limit superior of the expectation is finite for each of the three outer summands on the right-hand side separately.

Consider the first summand in (3.37). Then, for almost every $\omega \in \Omega$,

$$\begin{aligned}
\sum_{i=0}^{n} \mathbb{E}\big[ N_i^\delta |X_i - x^*| \mathbb{1}\{i \in I_n^1\} \,\big|\, \widetilde{\mathcal{G}}_{i-1} \big](\omega) &\le \sum_{i=0}^{n} \mathbb{E}\big[ N_i^\delta \mathbb{1}\{i \in I_n^1\} \,\big|\, \widetilde{\mathcal{G}}_{i-1} \big](\omega) \\
&\le \sum_{i=0}^{n} \mathbb{E}\big[ N_i \mathbb{1}\{i \in I_n^1\} \,\big|\, \widetilde{\mathcal{G}}_{i-1} \big](\omega), \qquad (3.38)
\end{aligned}$$

where the first inequality follows by the trivial bound $|X_n - x^*| \le 1$ and the second inequality by the fact that $\delta < 1/2$. Note that $\mathbb{1}\{i \in I_n^1\}$ is $\widetilde{\mathcal{G}}_{i-1}$-measurable, and it holds that $\mathbb{E}[N_i \mathbb{1}\{i \in I_n^1\} | \widetilde{\mathcal{G}}_{i-1}] \le \tilde{\theta}^{-(2+\gamma)}$, which follows by Assumption (c) and (3.35). Therefore

$$\begin{aligned}
\sum_{i=0}^{n} \mathbb{E}\big[ N_i \mathbb{1}\{i \in I_n^1\} \,\big|\, \widetilde{\mathcal{G}}_{i-1} \big](\omega) &\le \sum_{i=0}^{n} \mathbb{1}\{i \in I_n^1\} \mathbb{E}\big[ N_i \mathbb{1}\{i \in I_n^1\} \,\big|\, \widetilde{\mathcal{G}}_{i-1} \big](\omega) \\
&\le \sum_{i=0}^{n} \mathbb{1}\{i \in I_n^1\} \tilde{\theta}^{-(2+\gamma)}(\omega). \qquad (3.39)
\end{aligned}$$

Combining (3.38) and (3.39) shows that, for almost all $\omega \in \Omega$,

$$\sum_{i=0}^{n} \mathbb{E}\big[ N_i^\delta |X_i - x^*| \mathbb{1}\{i \in I_n^1\} \,\big|\, \widetilde{\mathcal{G}}_{i-1} \big](\omega) \le \sum_{i=0}^{n} \mathbb{1}\{i \in I_n^1\} \tilde{\theta}^{-(2+\gamma)}(\omega),$$

and taking expectations on both sides yields

$$\mathbb{E}\Bigg[\sum_{i=0}^{n}\mathbb{E}\big[N_i^\delta|X_i - x^*|\mathbb{1}\{i \in I_n^1\}\;\big|\;\widetilde{\mathcal{G}}_{i-1}\big]\Bigg] \leq \widetilde{\theta}^{-(2+\gamma)}\mathbb{E}\big[|I_n^1|\big].$$

Now let $\tilde{n} \in \mathbb{N}$ be such that $e^{-rn} < \tilde{\theta}$ for all $n \geq \tilde{n}$, then $|I_n^1| \leq \tilde{n} + M_r$ for all $n \in \mathbb{N}$, and hence $\mathbb{E}[|I_n^1|] \leq \tilde{n} + \mathbb{E}[M_r]$ for all $n \in \mathbb{N}$, which shows that

$$\lim_{n\to\infty}\mathbb{E}[|I_n^1|] < \infty, \tag{3.40}$$

since $\mathbb{E}[M_r] < \infty$ by Assumption (b). This shows that

$$\limsup_{n\to\infty}\mathbb{E}\Bigg[\sum_{i=0}^{n}\mathbb{E}\big[N_i^\delta|X_i - x^*|\mathbb{1}\{i \in I_n^1\}\,\big|\,\widetilde{\mathcal{G}}_{i-1}\big]\Bigg] < \infty,$$

that is, the limit superior of the expectation of the first summand in (3.37) is bounded.

Consider the second summand in (3.37). Since $\mathbb{1}\{i \in I_n^2\}$ and $X_i$ are $\widetilde{\mathcal{G}}_{i-1}$-measurable it holds that, for almost all $\omega \in \Omega$,

$$\begin{aligned}
&\sum_{i=0}^{n}\mathbb{E}\big[N_i^\delta|X_i - x^*|\mathbb{1}\{i \in I_n^2\}\,\big|\,\widetilde{\mathcal{G}}_{i-1}\big](\omega)\\
&\qquad\qquad = \sum_{i=0}^{n}\mathbb{1}\{i \in I_n^2\}|X_i - x^*|\mathbb{E}\big[N_i^\delta\mathbb{1}\{i \in I_n^2\}\,\big|\,\widetilde{\mathcal{G}}_{i-1}\big](\omega)\\
&\qquad\qquad \leq \sum_{i=0}^{n}\mathbb{1}\{i \in I_n^2\}|X_i - x^*|\mathbb{E}\big[N_i\mathbb{1}\{i \in I_n^2\}\,\big|\,\widetilde{\mathcal{G}}_{i-1}\big]^\delta(\omega),
\end{aligned} \tag{3.41}$$

which follows by Jensen's inequality. Next,

$$\begin{aligned}
\mathbb{E}\big[N_i\mathbb{1}\{i \in I_n^2\}\,\big|\,\widetilde{\mathcal{G}}_{i-1}\big] &= \mathbb{E}\big[N(|\tilde{g}(X_i)|)\mathbb{1}\{i \in I_n^2\}\,\big|\,\widetilde{\mathcal{G}}_{i-1}\big]\\
&\leq \mathbb{E}\big[N(k|X_i - x^*|)\mathbb{1}\{i \in I_n^2\}\,\big|\,\widetilde{\mathcal{G}}_{i-1}\big],
\end{aligned}$$

which follows by Assumption (c) and the fact that $\mathbb{E}\left[N(|\theta|)\right]$ is non-increasing in the drift $|\theta|$, which can be seen by a sample path argument. By definition of the set $I_n^2$ it holds that $k|X_i - x^*|\mathbb{1}\{i \in I_n^2\} \leq \tilde{\theta}$, and (3.35) shows that

$$\mathbb{E}\big[N_i\mathbb{1}\{i \in I_n^2\}\,\big|\,\widetilde{\mathcal{G}}_{i-1}\big] \leq (k|X_i - x^*|)^{-(2+\gamma)}.$$

Combining this with (3.41) yields

$$
\begin{aligned}
\sum_{i=0}^{n} \mathbb{E}\big[N_i^{\delta}|X_i - x^*|\mathbb{1}\{i \in I_n^2\} \,\big|\, \widetilde{\mathcal{G}}_{i-1}\big](\omega) \\
&\le \sum_{i=0}^{n} \mathbb{1}\{i \in I_n^2\}|X_i - x^*| \left(k|X_i - x^*|\right)^{-\delta(2+\gamma)}(\omega) \\
&= k^{-\delta(2+\gamma)} \sum_{i=0}^{n} \mathbb{1}\{i \in I_n^2\}|X_i - x^*|^{1-\delta(2+\gamma)}(\omega) \qquad (3.42) \\
&\le k^{-\delta(2+\gamma)}|I_n^2|(\omega), \qquad (3.43)
\end{aligned}
$$

where the last inequality follows by the trivial bound $|X_i - x^*| \le 1$ and the fact that $1 - \delta(2+\gamma) > 0$. Since $|I_n^2| \le M_r$ for all $n \in \mathbb{N}$ it follows that $\mathbb{E}[|I_n^2|] \le \mathbb{E}[M_r] < \infty$ for all $n \in \mathbb{N}$ and so

$$
\lim_{n\to\infty} \mathbb{E}\left[|I_n^2|\right] < \infty. \qquad (3.44)
$$

Now taking expectations on both sides in (3.43), and since (3.44) holds, it follows that the limit superior of the expectation of the second summand in (3.37) is bounded.

It remains to show that the limit superior of the expectation of the third term in (3.37) is bounded as well. By the same derivation as for the second summand, it holds that, for almost all $\omega \in \Omega$,

$$
\sum_{i=0}^{n} \mathbb{E}\big[N_i^{\delta}|X_i - x^*|\mathbb{1}\{i \in I_n^3\} \,\big|\, \widetilde{\mathcal{G}}_{i-1}\big](\omega) \le k^{-\delta(2+\gamma)} \sum_{i=0}^{n} \mathbb{1}\{i \in I_n^3\}|X_i - x^*|^{1-\delta(2+\gamma)},
$$

see (3.42). Define $\sigma = 1 - \delta(2+\gamma) > 0$. For $i \in I_n^3$ it holds that $k|X_i - x^*| \le e^{-ri}$, therefore

$$\begin{aligned}\sum_{i=0}^{n} \mathbb{E}\big[N_i^\delta |X_i - x^*| \mathbb{1}\{i \in I_n^3\} \,\big|\, \widetilde{\mathcal{G}}_{i-1}\big](\omega) &\le k^{-\delta(2+\gamma)} \sum_{i=0}^{n} \big(e^{-ri}/k\big)^\sigma \\ &= k^{-\delta(2+\gamma)-\sigma} \sum_{i=0}^{n} e^{-r\sigma i} \\ &\le k^{-\delta(2+\gamma)-\sigma} \sum_{i=0}^{\infty} e^{-r\sigma i} \\ &= k^{-\delta(2+\gamma)-\sigma} \frac{1}{1-e^{-r\sigma}}.\end{aligned}$$

Taking expectations on both sides shows that also the limit superior of the expectation of the third summand in (3.37) is bounded, which finishes the proof. □

This concludes our the theoretical analysis of the PBA in the context of stochastic root-finding problems. In the next chapter, we provide numerical results which analyze empirically the convergence behavior in macro time as well as in wall-clock time for different test functions $g$.

CHAPTER 4

# NUMERICAL RESULTS

In this chapter, we present a series of numerical results based on Monte-Carlo simulations in order to empirically investigate the performance of the PBA for different scenarios. In addition to empirically confirming some of the previous results, we also provide evidence for some of the stated conjectures as well as a direct comparison of the PBA and SA-type algorithms.

We assume that $X^* \sim U(0,1)$. While conditioning on a specific realization of $X^*$ covers the frequentist setting, we are mostly interested in an average-case performance, such as the behavior of $\mathbb{E}[|X_n - X^*|]$ as $n \to \infty$. In all empirical examples we start the PBA with a uniform prior distribution $f_0$.

This chapter is organized as follows. In Section 4.1 we analyze the convergence behavior of the residuals and confidence intervals of the PBA for the $p(\cdot)$ constant and known case. In Section 4.2 we extend the analysis to the more realistic case when $p(\cdot)$ is nonconstant and unknown, and compare five stochastic root-finding algorithms (thee PBA-type algorithms and two SA-type algorithms) on four different test functions. We also provide a sensitivity analysis as well as a discussion of confidence intervals in wall-clock time.

## 4.1 Convergence Behavior in Macro Time

### 4.1.1 Absolute Residuals

Let us discuss the residuals $|X_n - X^*|$ of the PBA for the setting when the probability of a correct sign is constant and known, that is, $p(\cdot) \equiv p_c$.

Figure 4.1 shows ten sample paths and the estimated mean and median path of 10,000 simulated runs where $p_c = 0.70$, where $X^* = 0.6647$. We have displayed the $y$-axis in log-scale to investigate the rate of convergence. If the rate of convergence is indeed geometric in $n$ then the sample paths should be linear under this transformation. The geometric convergence can be observed for the mean and median paths as well as, to some extent, for each single sample path. This figure also shows that the mean sample path is dominated by a few bad sample paths: Out of the ten chosen sample paths only one is clearly above the mean sample path and the median sample path significantly outperforms the mean sample path. The histogram of the 10,000 residuals for $n = 100$ shown in Figure 4.2 confirms this.

Before further discussing the empirical behavior of the PBA, let us quickly discuss a challenge that arises when implementing the PBA with finite floating point arithmetic. When starting with a uniform prior it is sufficient to keep track of an ordered list containing the measurement points $(X_n)_n$ and the corresponding heights of the density $(h_n)_n$ between the measurement points. At each iteration the median of $f_n$ is determined and inserted in this sorted list, and, after observing the sign, the heights $(h_n)_n$ are updated according the PBA. Finding the median and updating the heights can fail due to the finite precision of floating point arithmetics when measurement points in $(X_n)_n$ become very close to each other (within $10^{-15}$).

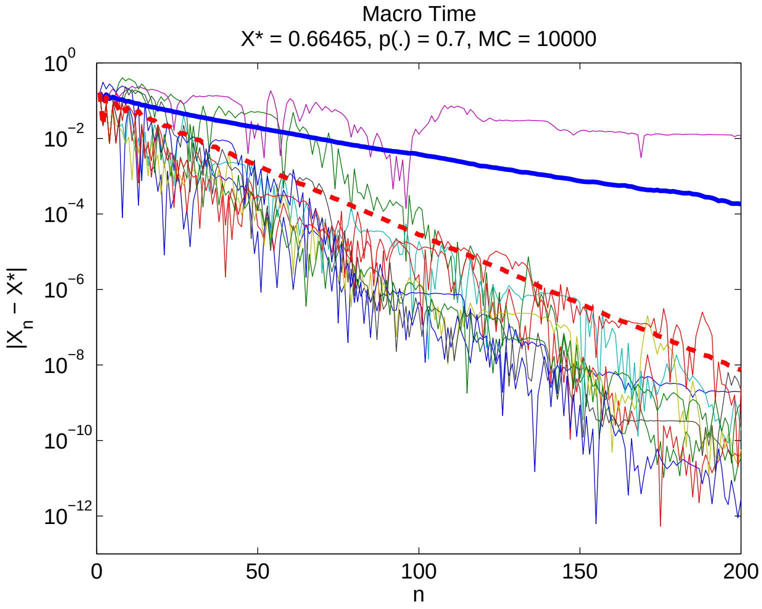

Figure 4.1: The estimated mean (thick solid blue) and median (thick dashed red) path of 10,000 simulated runs together with ten individual sample paths of the PBA, where $p(\cdot) \equiv 0.7$. To analyze the geometric rate of convergence behavior the $y$-axis is shown in log-scale.

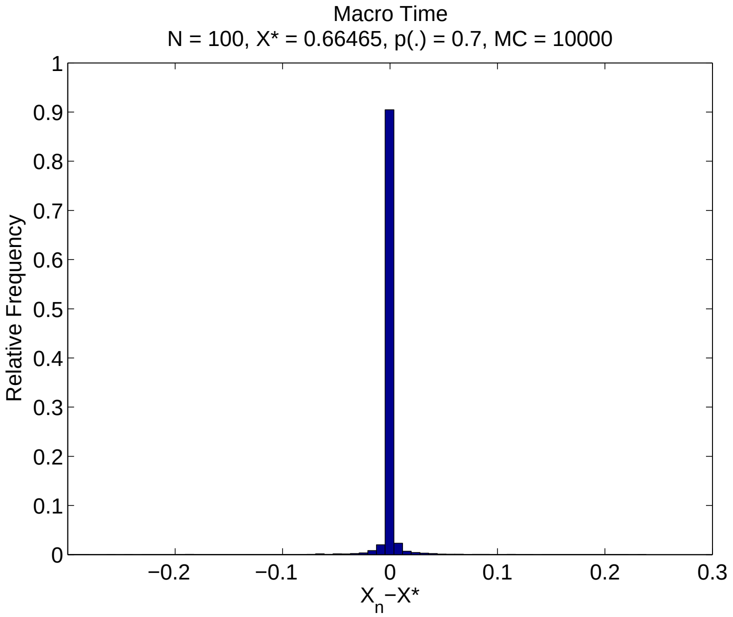

Figure 4.2: Histogram of 10,000 residuals $X_n - X^*$ at time $n = 100$. About 90% of all samples are extremely close to the sought-after point $X^*$, but a few sample paths are still far away.

This failure in the updating arises more quickly for large values of $p_c$ since in this case the measurement points $(X_n)_n$ approach $X^*$ very quickly. In the presented examples, sample paths are discarded once the computational accuracy fails (the density $f_n$ does not integrate to one anymore). This, however, introduces an upwards bias in the estimation of $\mathbb{E}[|X_n - X^*|]$ for large $n$ since a discarded path has usually located $X^*$ with very high precision. While a more robust implementation of the algorithm would potentially improve the estimation results, the current results are sufficient to provide insight into the behavior of the PBA.

Figures 4.3 and 4.4 show the rate of convergence of the mean and median path for different parameters $p_c$. Here, for every sample run the root $X^*$ is an independent realization of a $U(0,1)$ random variable. We use linear regression to estimate the slope of $\log(\mathbb{E}[|X_n - X^*|])$, where the range of $n$ is chosen to avoid issues due to numerical precision. To be specific, for every $p_c$ we let $N^u$ be the largest $n$ such that not more than 25% of all sample path have been eliminated due to computational inaccuracy. In order to account for the asymptotic convergence behavior we let $N^l = \lfloor 0.5N^u \rfloor$ and use samples $n \in \left\{N^l, N^l + 1, \ldots, N^u\right\}$ for the linear regression estimation. The fitted lines (thin black lines) are also shown in Figures 4.3 and 4.4. Figure 4.5 shows the estimated rate $r$ and $C = e^r$ as a function of $p_c$ for the mean and median paths together with the lower bound on this rate proven in Theorem 2. The trivial upper bound $r = \log 2$ implied by the noise-free bisection search is also shown. The right-hand side plot confirms the intuition that $C \uparrow 2$ as $p_c \uparrow 1$ for the median as well as the mean sample path of the expected residuals.

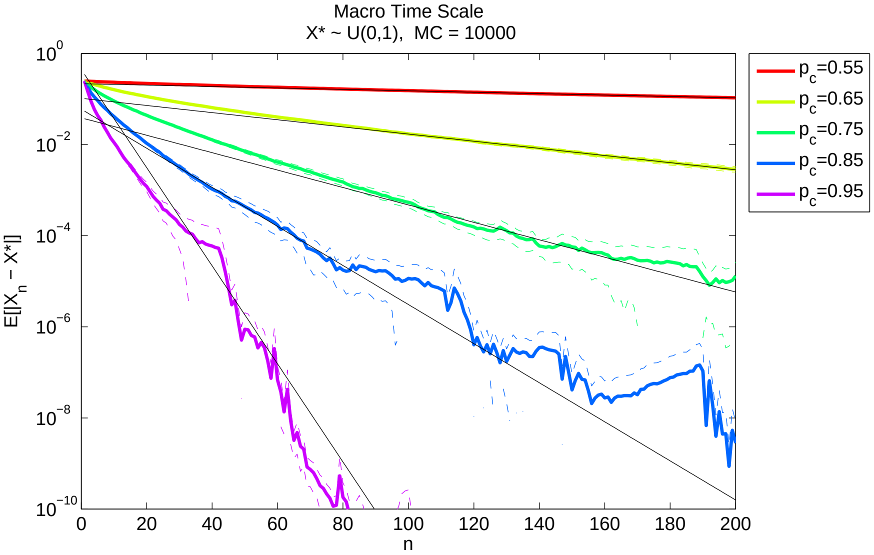

Figure 4.3: Estimated expected residuals $\mathbb{E}[|X_n - X^*|]$ for $X^* \sim U(0,1)$ and parameters $p_c = 0.55, 0.65, 0.75, 0.85, 0.95$ (thick lines) with fitted lines (thin black lines). The normal approximation 95%-confidence intervals are also given (dashed). For large $p_c$ and large $n$ the estimation becomes less reliable due to computational accuracy conflicts—for some values the lower bound on the approximate confidence interval even became negative, a not uncommon effect when estimating small positive quantities.

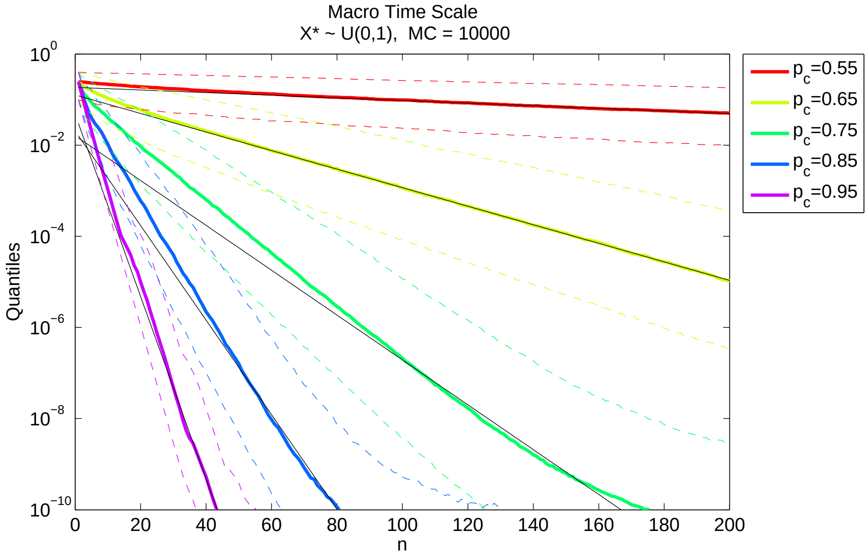

Figure 4.4: Estimated median of $|X_n - X^*|$ for $X^* \sim U(0,1)$ and parameters $p_c = 0.55, 0.65, 0.75, 0.85, 0.95$ (thick lines) with fitted lines (thin black lines). The 20% and 80% quantiles are also given (dashed).

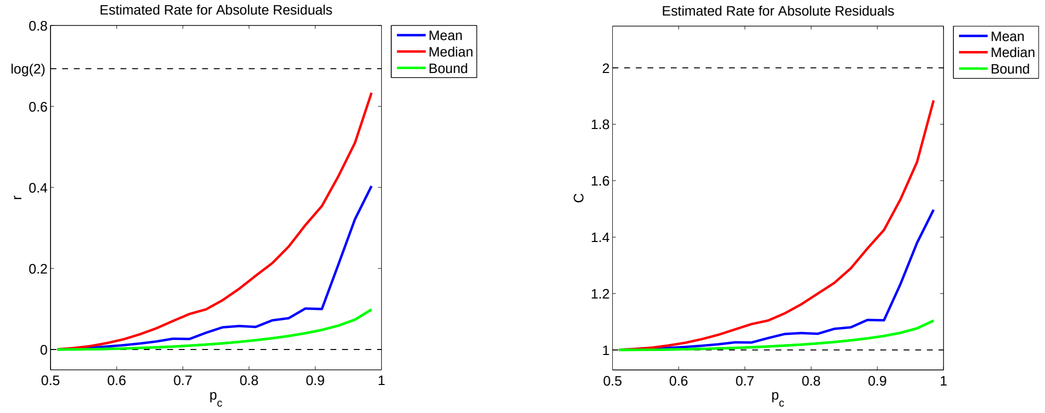

Figure 4.5: *Left Plot:* The rate $r$ of the mean and median of $|X_n - X^*|$ as a function of $p_c$. This assumes that $\mathbb{E}[|X_n - X^*|] \sim e^{-rn}$ and $F^{-0.5}(|X_n - X^*|) \sim e^{-rn}$, where $F$ is the distribution function of $|X_n - X^*|$. The lower bound on $r$ as proven in Theorem 2 (and defined in Lemma 1) is also given. (Here, the notation $g(x) \sim f(x)$ means that $\lim_{x \to x_0} f(x)/g(x) = a$ for some constant $a > 0$.)
*Right Plot:* The rate $C = e^r$ as a function of $p_c$.

### 4.1.2 Confidence Intervals

Let us now focus on the confidence intervals introduced in Sections 3.3–3.4 in macro time. Figure 4.6 shows the upper and lower bound of the 95%-confidence interval $J_n(\alpha)$ (left plot) and $K_n(\alpha)$ (right plot) for $p_c = 0.7$, $X^* = 0.2994$ and $\alpha = 0.05$. Recall that the interval $J_n(\alpha)$ is only a confidence interval for a fixed $n$, but not for the whole sample path, whereas the sequence of intervals $(K_n(\alpha))_n$ is a sequential confidence interval, that is, $x^* \in K_n(\alpha)$ for all $n \in \mathbb{N}$ with high probability. In this sample path, the value $x^*$ never leaves either interval sequence, $(J_n(\alpha))_n$ or $(K_n(\alpha))_n$, but only the sequence $(K_n(\alpha))_n$ provides the probabilistic guarantee for this to happen. By definition $K_n(\alpha) \supseteq K_{n+1}(\alpha)$ for all $n \in \mathbb{N}$. In contrast, this is not necessarily true for the sequence $(J_n(\alpha))_n$ as the length of $J_n(\alpha)$ may increase, as can be observed in Figure 4.6.

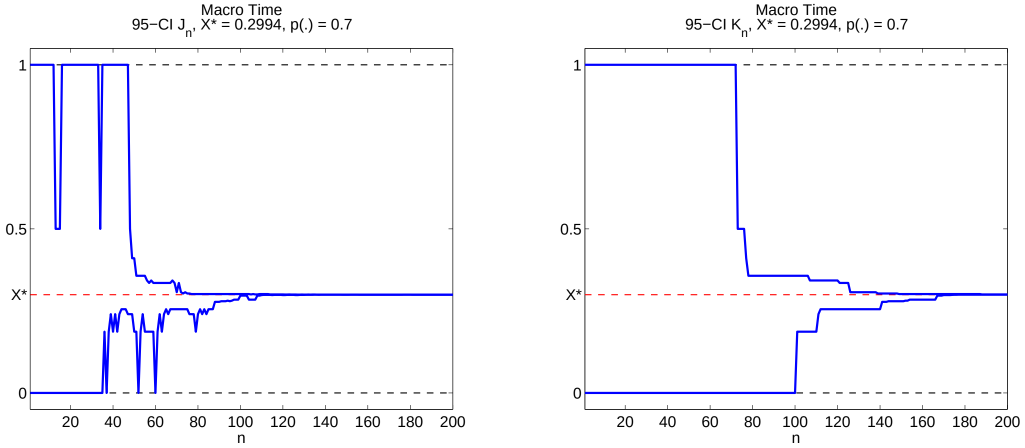

Figure 4.6: A sample path of the confidence interval $(J_n(\alpha))_n$ and the sequential confidence interval $(K_n(\alpha))_n$. Here, $p_c = 0.7, \alpha = 0.05$ and $X^* = 0.2994$.

The sample paths in Figure 4.6 show that after some initial period the lengths of these confidence intervals decrease rapidly. Also, this initial period is significantly longer for the intervals $K_n(\alpha)$. Theorems 5 and 6 prove that for $p_c \geq 0.85$ the asymptotic rate of convergence of the sequence $(|J_n(\alpha)|)_n$ and $(|K_n(\alpha)|)_n$ is geometric. Figures 4.7 and 4.8 show the estimated mean and median paths of $(|J_n(\alpha)|)_n$ for different parameters $p_c$ (the behavior for the sequence $(|K_n(\alpha)|)_n$ is similar, hence omitted). Again the $y$-axis is displayed in log-scale and the linear behavior of the paths indicate a geometric rate of convergence for all tested values $p_c$, giving credence to the conjecture that the geometric rate holds for all parameters $p_c > 1/2$.

The mean and median path of $\big(|J_n(\alpha)|\big)_n$ show an interesting zig-zag behavior. This can be explained: The length of the confidence intervals is determined by the measurement points $(X_n)_n$, which, for a fixed value $p_c$, assume values on a predefined grid that is independent of $X^*$. While the noisy signs $(Z_n(X_n))_n$, which depend on $X^*$, determine the sample path on this grid, the length of the confidence intervals is determined by the fixed grid of possible values that $(X_n)_n$ can attain, depending on $p_c$. This is especially apparent for small $n$, leading to a zig-zag pattern in the sample paths of $\big(|J_n(\alpha)|\big)_n$.

Finally, it is informative to estimate the true cover probabilities of the confidence intervals $(J_n(\alpha))_n$. Figure 4.9, which displays the estimates of the coverage probabilities for $\alpha = 0.05$ and $\alpha = 0.2$, shows that the actual cover probability of the confidence interval is significantly higher than the required level $1 - \alpha$. This difference is mostly due to the use of Hoeffding's bound in the construction of the interval $J_n(\alpha)$, see (3.9).

## 4.2 Convergence Behavior in Wall-Clock Time

### 4.2.1 Absolute Residuals

After investigating the PBA for the setting when $p(\cdot) \equiv p_c$, which corresponds to the performance in macro time, let us now turn our attention to the convergence behavior of the PBA in wall-clock time. We also provide a direct comparison of the PBA to popular SA-type algorithms for a set of functions $g$.

In all examples the stochastic noise is assumed to be standard normal, that is, $\epsilon \sim N(0,1)$, and at each measurement point the PBA performs a test of power one with curved boundary $(k_n)_n$ for the standard normal noise distribution, given by (B.3). Using a test of power one only using the signs $(Z_n(X_n))_n$ instead of the actual noisy function evaluations $(Y_n(X_n))_n$ will lead to very similar convergence behavior in this case, see discussion in Section 1.5.

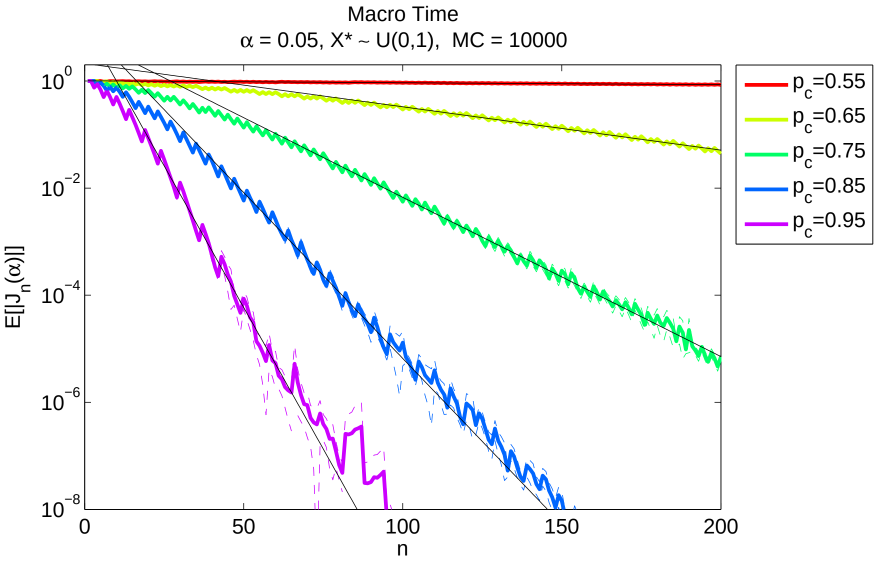

Figure 4.7: Estimated expected lengths $\mathbb{E}[|J_n(\alpha)|]$ for $X^* \sim U(0,1)$, $\alpha = 0.05$, and parameters $p_c = 0.55, 0.65, 0.75, 0.85, 0.95$ (thick lines) with fitted lines (thin black lines). The normal approximation 95%-confidence intervals are also given (dashed). Again, for large large $p_c$ and large $n$ the estimation becomes less reliable.

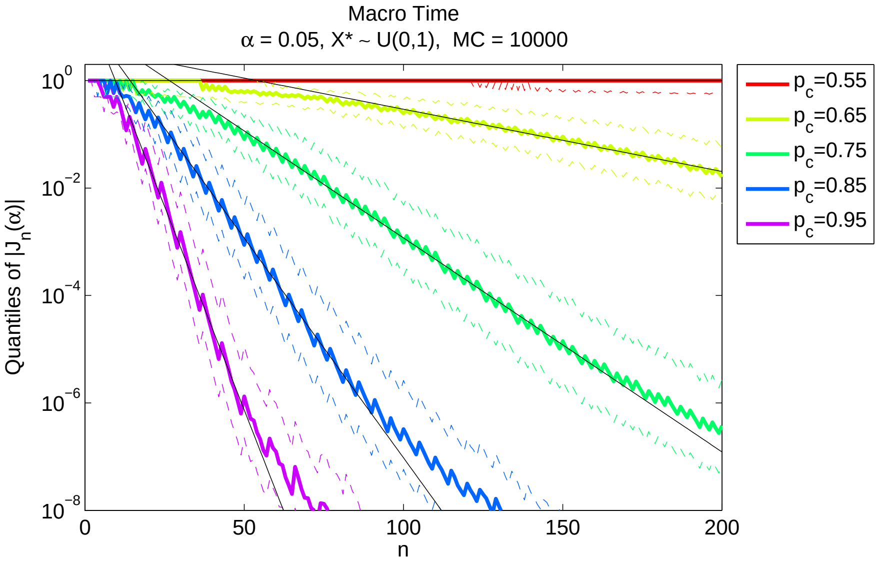

Figure 4.8: Estimated median of $|J_n(\alpha)|$ for $X^* \sim U(0,1)$, $\alpha = 0.05$, and parameters $p_c = 0.55, 0.65, 0.75, 0.85, 0.95$ (thick lines) with a fitted lines (thin black lines). The 20% and 80% quantiles are also given (dashed).

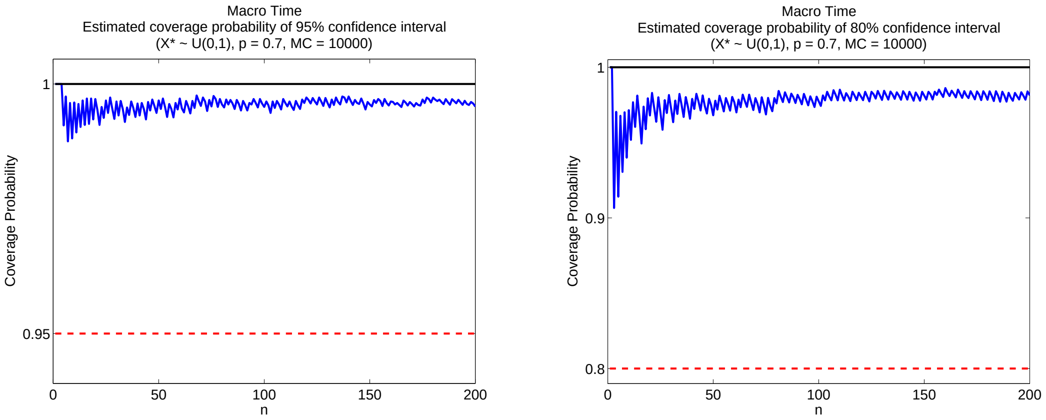

Figure 4.9: Estimated cover probabilities of the $(1-\alpha)$-confidence interval $J_n(\alpha)$ for $\alpha = 0.05$ (left plot) and $\alpha = 0.2$ (right plot).

Let us first compare the sample path behavior of the PBA and the SA. Figure 4.10 shows two sample paths of the PBA (top plots) and two sample paths of the SA algorithm (bottom plot) for the test function $g(x) = X^* - x$. Here, we chose $a_n = 1/n$ and $X_0 = 0$ for the SA, and $p_c = 0.6$ for the PBA as input parameters. In contrast to the SA, which measures at a different point at each iteration, the PBA only changes its current measurement point when "enough" information is collected on the location of $X^*$. In these examples the total number of function evaluations (wall-clock time) is $10^6$, whereas the PBA only uses about 20 macro iterations. If $X_n$ of the PBA is close to $x^*$ the test of power one requires a very long time to decide whether the root is further to the left or right of $X_n$. Such stalling behavior, is observable as long flat lines in the sample paths and can be desirable since in these cases $X_n$ usually corresponds to a good estimate of $X^*$.

Figure 4.11 (top) shows ten sample paths of the PBA and the estimated mean and median path based on 1,000 sample paths; Figure 4.11 (bottom) displays the same for the SA algorithm. For the SA algorithm, it is known that, under some technical assumptions, $(T^{1/2}(X_T - X^*))_T$ is a tight sequence (see Pasupathy and

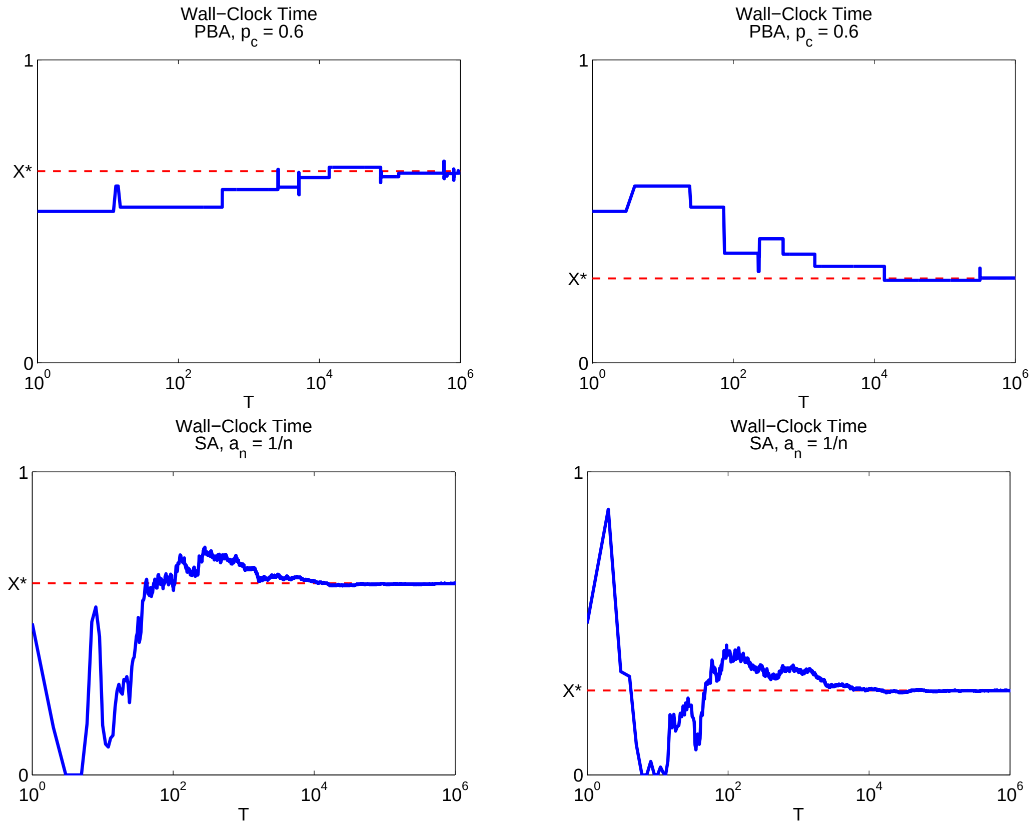

Figure 4.10: Two sample paths of the PBA (top) and two sample paths of the SA (bottom) for two different roots $X^*$ (which are realization of $U(0,1)$ random variables). These plots show the different sampling concepts of the two algorithms: In contrast to the SA algorithm, which changes its measurement location at every iteration, the PBA only changes its measurement location once a strong enough signal regarding the location of $X^*$ has been received. The $x$-axis is shown in log-scale for better visibility.

Kim, 2011; Kushner and Yin, 2003, Chapter 10, for details), whereas for the PBA this does not hold (see Theorem 8). The fact that the PBA converges asymptotically slower than the SA algorithm can already be observed in Figure 4.11, but will become more obvious in Figure 4.14 (below) where the estimated mean paths of the two algorithms are shown on the same graph.

The histograms of the residuals for $T = 100,000$ are given in Figure 4.12. For SA-type algorithms it is known that, under some technical assumptions, the

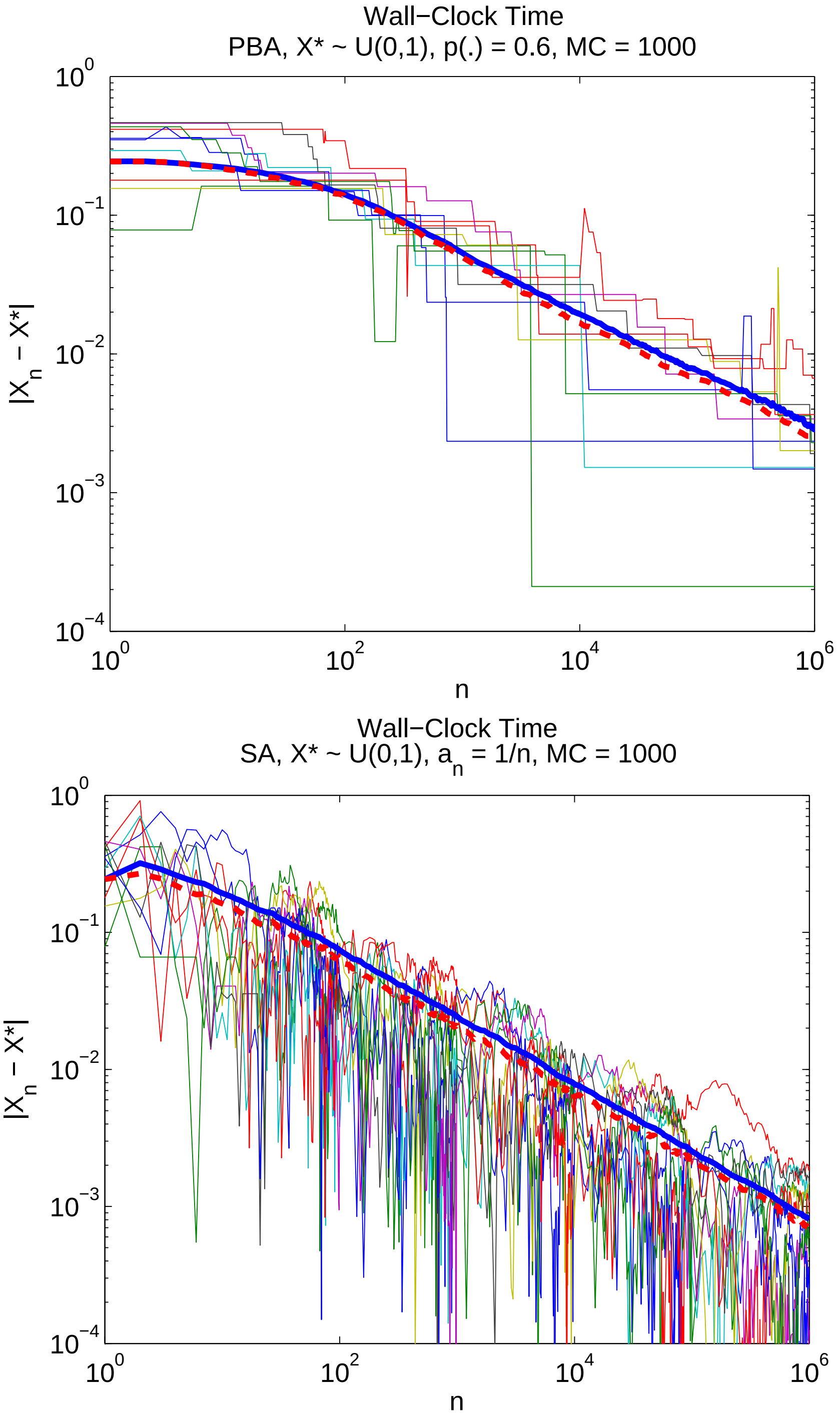

Figure 4.11: The estimated mean and median path of $|X_n - X^*|$ based on 1,000 simulated path for the PBA (top) and SA algorithm (bottom). The first 10 sample paths are also given for each algorithm. The $x$- and $y$-axis are displayed in log-scale.

limiting distribution of $(X_n - X^*)$ is normally distributed (see Pasupathy and Kim, 2011), which can be observed in Figure 4.12 (right plot). The histogram in Figure 4.12 (left plot) suggests that a similar result does not hold for the PBA since the distribution of the residuals seems to have a quite long and flat tail.

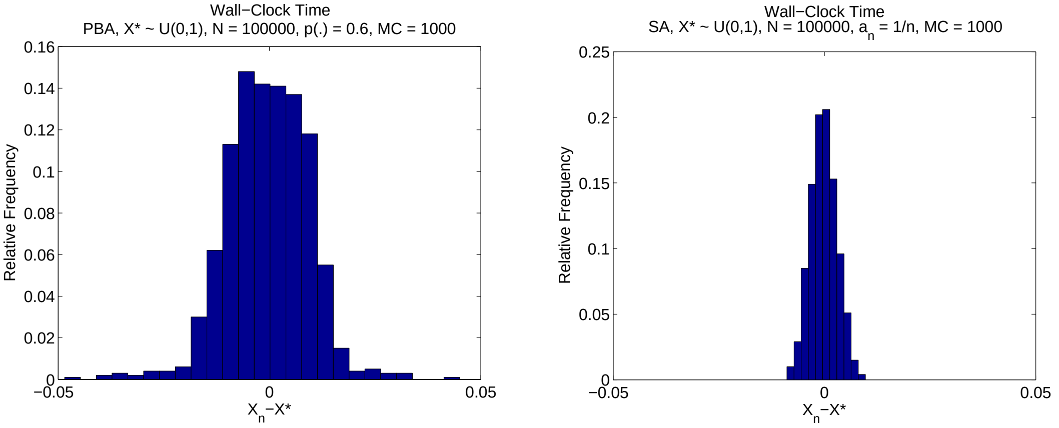

Figure 4.12: The histograms of the residuals $|X_n - X^*|$ for the PBA and SA algorithm after $T = 100,000$ function evaluations.

We now use the performance measure $\mathbb{E}[|\hat{X}_T - X^*|]$ to compare five stochastic root-finding algorithms, each of which defines a sequence of estimators $\hat{X}_T$, on four different test functions. The algorithms are:

1. PBA: After $T$ function evaluations the best estimate of $X^*$ is given by $X_T$, which corresponds to the current measurement point. Here, $p_c$ is a tuning parameter. We chose $p_c = 0.95$.

2. PBA-averaging: After $T$ function evaluations the best estimate of $X^*$ is given by the time-weighted average of the medians generated by the PBA, that is,

$$\hat{X}_T = \frac{1}{T}\left(\sum_{i=0}^{n-1} N_i X_i + (T - \sum_{i=0}^{n-1} N_i) X_n\right),$$

where $(X_n)_n$ is the sequence of medians generated by the PBA. Here, $p_c$ is a tuning parameter. We chose $p_c = 0.95$.

3. PBA-$\varepsilon$-averaging: After $T$ function evaluations the best estimate of $X^*$ is given by the transformed time-weigted average as given by (3.33) in Theorem 9, that is,

$$\hat{X}_T = \frac{1}{\overline{T}} \left( \sum_{i=0}^{n-1} N_i^{\frac{1}{2}+\varepsilon} X_i + (T - \sum_{i=0}^{n-1} N_i)^{\frac{1}{2}+\varepsilon} X_n \right),$$

where the denominator is

$$\overline{T} = \sum_{i=0}^{n-1} N_i^{\frac{1}{2}+\varepsilon} + \left( T - \sum_{i=0}^{n-1} N_i \right)^{\frac{1}{2}+\varepsilon}.$$

The sequence $(X_n)_n$ corresponds to the medians generated by the PBA. Here, $p_c$ and $\varepsilon > 0$ are tuning parameters. We chose $p_c = 0.95$ and $\varepsilon = 10^{-3}$.

4. SA: After $T$ function evaluations the best estimate of $X^*$ is given by $X_T$, where $(X_i)_i$ is the sequence of measurement points generated via

$$X_{i+1} = \Gamma_{[0,1]}(X_i + a_i Y_i(X_i)), \tag{4.1}$$

where $\Gamma_{[0,1]}$ is the projection on the interval $[0, 1]$ and $a_i = c/i$; see Section 1.3 for details. Here, $c$ and $X_0$ are tuning parameters. We chose $c = 1$ and $X_0 = 0.5$.

5. Polyak-Ruppert averaging: After $T$ function evaluations the best estimate of $x^*$ is

$$\hat{X}_T = \frac{1}{T} \sum_{i=1}^{T} X_i,$$

where $(X_i)_i$ is a sequence generated by (4.1), where now $a_i = c/i^\delta$ and $\delta \in (1/2, 1)$. Here, $c$, $\delta$ and $X_0$ are tuning parameters. We chose $c = 1$, $\delta = 0.8$ and $X_0 = 0.5$.

All the above input parameters were chosen based on a sensitivity analysis, with details provided in the next section.

The four test functions are:

1. The piecewise constant function:

$$g_1(x) = \ 0.3\mathbb{1}\{x \le X^* - 0.2\} + 0.2\mathbb{1}\{x \le X^*\}$$
$$- 0.1\mathbb{1}\{x > X^*\} - 0.3\mathbb{1}\{x > X^* + 0.2\}; \tag{4.2}$$

see Figure 4.13. This test function is motivated by simulation-optimization problems on a discrete domain, which can be solved by linear interpolation of the objective function between feasible points; see Lim (2011).

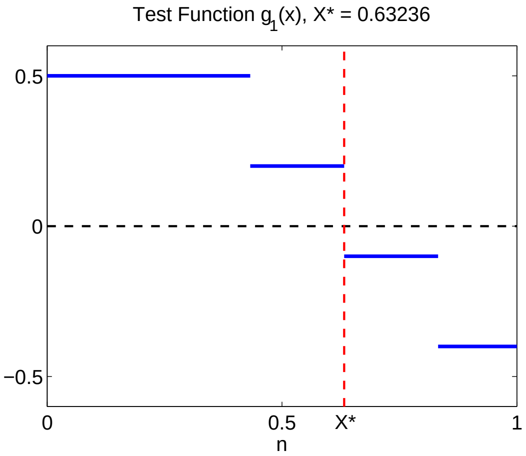

Figure 4.13: The test function $g_1$ as given in (4.2).

2. The linear function:

$$g_2(x) = X^* - x. \tag{4.3}$$

This test function corresponds to a "simple" function on which any stochastic root-finding algorithm should perform reasonably well.

3. The exponential function:

$$g_3(x) = \exp\left(2(X^* - x)\right) - 1. \tag{4.4}$$

Due to the curvature this function appears to be more difficult for root-finding than the linear function (4.3). Such curvature usually introduces a bias in the finite-time estimation of $x^*$, since a measurement to the left of $X^*$ causes a faster move towards the right, compared to a movement induced by a measurement to the right of $X^*$. Therefore, the algorithms (SA-type as well as PBA-type algorithms) tend to spend more time to the right of $X^*$.

4. The cubic function:

$$g_4(x) = (X^* - x)^3, \tag{4.5}$$

this function is considered very difficult for stochastic root-finding algorithms since $g'(X^*) = 0$. Although PBA and SA are still able to produce consistent estimators of $X^*$ in this case, the rate of convergence results do usually not hold anymore. Specifically, Assumption (i) for SA-type algorithms in Section 1.3 and Assumption (c) in Theorem 9 for the PBA are violated. Although in most applications we would not encounter $g'(X^*) = 0$, it is nevertheless informative to test stochastic root-finding algorithms on such difficult functions for which the convergence behavior is not yet well-understood.

Figure 4.14 provides a direct comparison of the five considered stochastic root-finding algorithms for each test function separately.

The comparison of the algorithms on this set of test functions leads to some interesting observations, such as:

- When the function $g$ indeed has a discontinuity at $X^*$, the PBA (without averaging) significantly outperforms SA-type algorithms as well as PBA-type algorithms that use averaging.
- When the function $g$ is continuous, then SA-type algorithms seem to provide

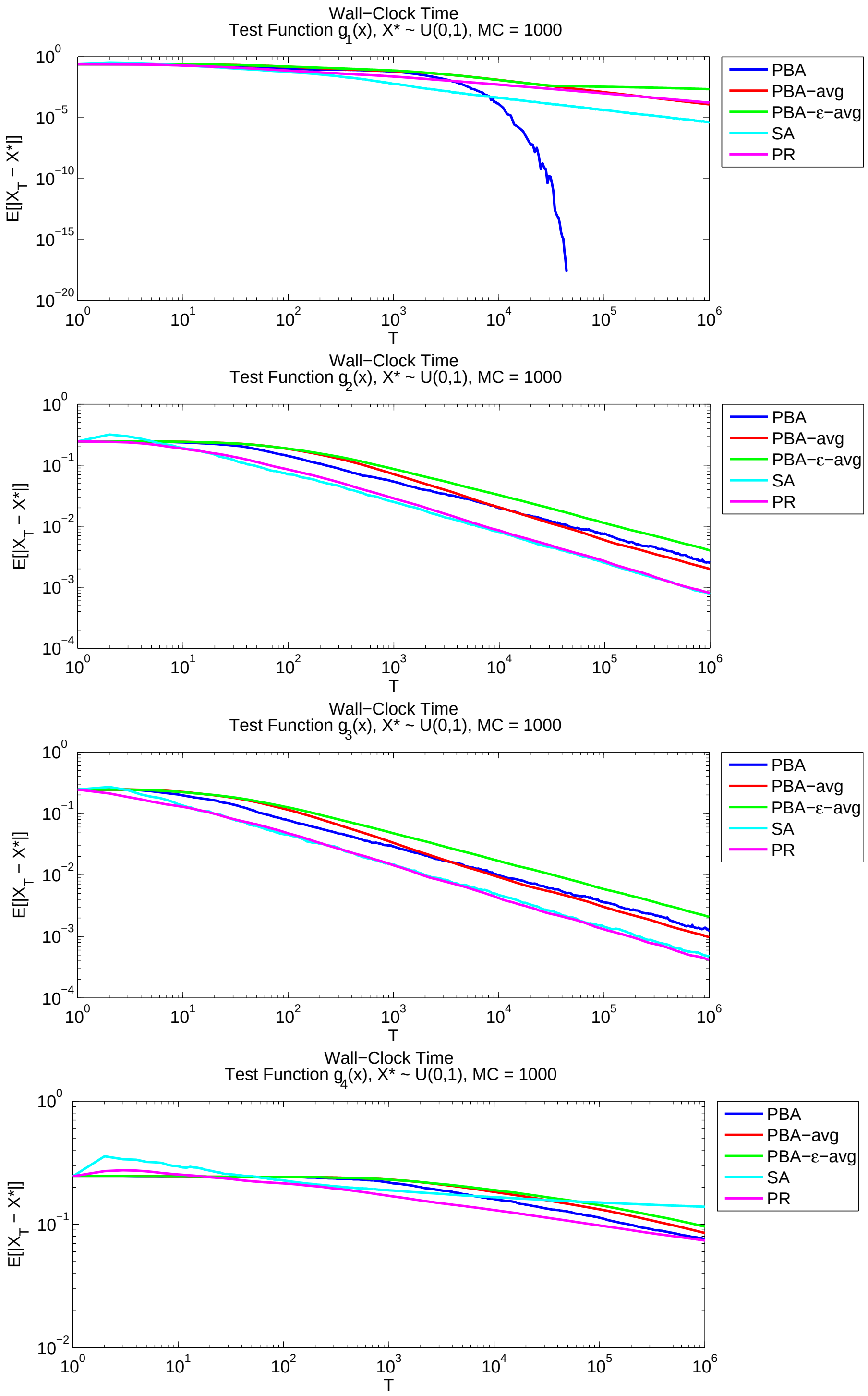

Figure 4.14: The estimated path $\mathbb{E}[|\hat{X}_T - X^*|]$ for test functions $g_1, \ldots, g_4$ and for the five described stochastic root-finding algorithms. Here, $\epsilon \sim N(0,1)$. For better visibility the $x$- and $y$-axes are displayed in log-scale. We use 1,000 independent sample runs to estimate these paths, and the relative error is always smaller than 5%.

slightly better performance in terms of $\mathbb{E}[|\hat{X}_n - X^*|]$ than PBA-type algorithms. This is not surprising since SA-type algorithms tend to achieve the optimal rate, given the tuning sequences was chosen appropriately.

- Of the three tested PBA-type algorithms, the performance of PBA-averaging seems to be the best when $g$ is continuous at $X^*$ and even seems to recover the optimal rate of convergence $O(T^{-1/2})$. This can be observed as the slope of its performance path seems to be very similar to the slope of the performance path of SA-type algorithms.

### 4.2.2 Sensitivity Analysis

In this section we test the sensitivity of the performance measure $\mathbb{E}[|\hat{X}_n - X^*|]$ with respect to the input parameters of the different stochastic root-finding algorithms. We will not provide details for all test functions and parameters, but instead just highlight some general behavior.

We consider the linear test function $g_2(x) = X^* - x$. Figure 4.15 shows the effect of the input parameter $p_c$ on the performance of the three PBA-type algorithms, whereas Figure 4.16 shows the effect of the input parameter $c$ on the performance of the two SA-type algorithms. Here, we fix the other input parameters as $\varepsilon = 10^{-3}$, $\delta = 0.8$ and $X_0 = 0.5$.

These figures indicate that the performance of the PBA is robust towards the choice of the tuning parameter $p_c$, as long as it is reasonably large, say larger than $p_c \geq 0.65$. The performance of the SA algorithm, on the other hand, is more affected by the choice of the tuning sequence, as its performance may become inferior when the tuning sequence is chosen small, for example, $c = 0.1$. In this case

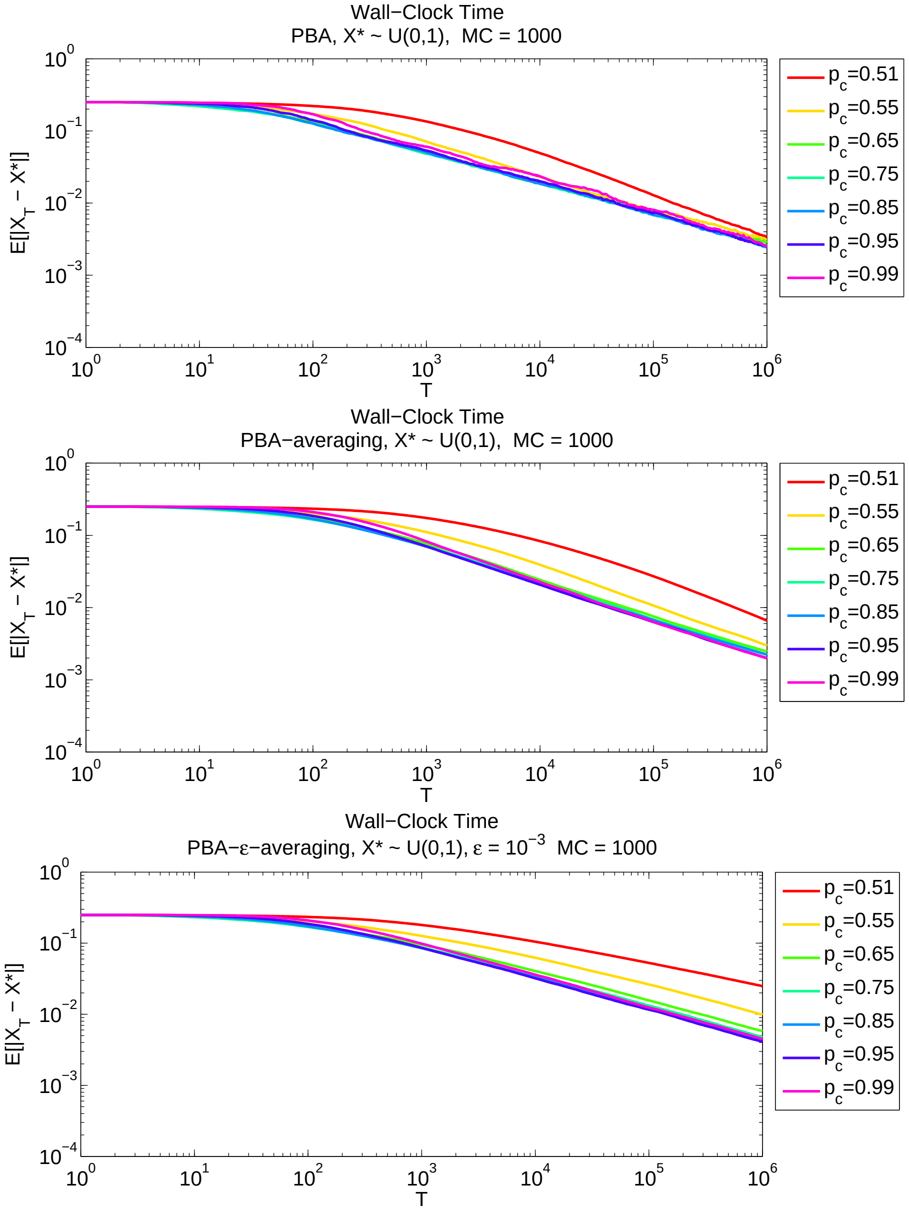

Figure 4.15: The performance of the PBA, PBA-averaging and PBA-$\varepsilon$-averaging for a set of different input parameters $p_c$. The test function is $g_2(x) = X^* - x$. Here, we chose $\varepsilon = 10^{-3}$ for the PBA-$\varepsilon$-averaging method. We see that the performance is stable with respect to the input parameter $p_c$, as long as it is not too close to 1/2. We use 1,000 independent sample runs to estimate these paths, and the relative error is always smaller than 5%.

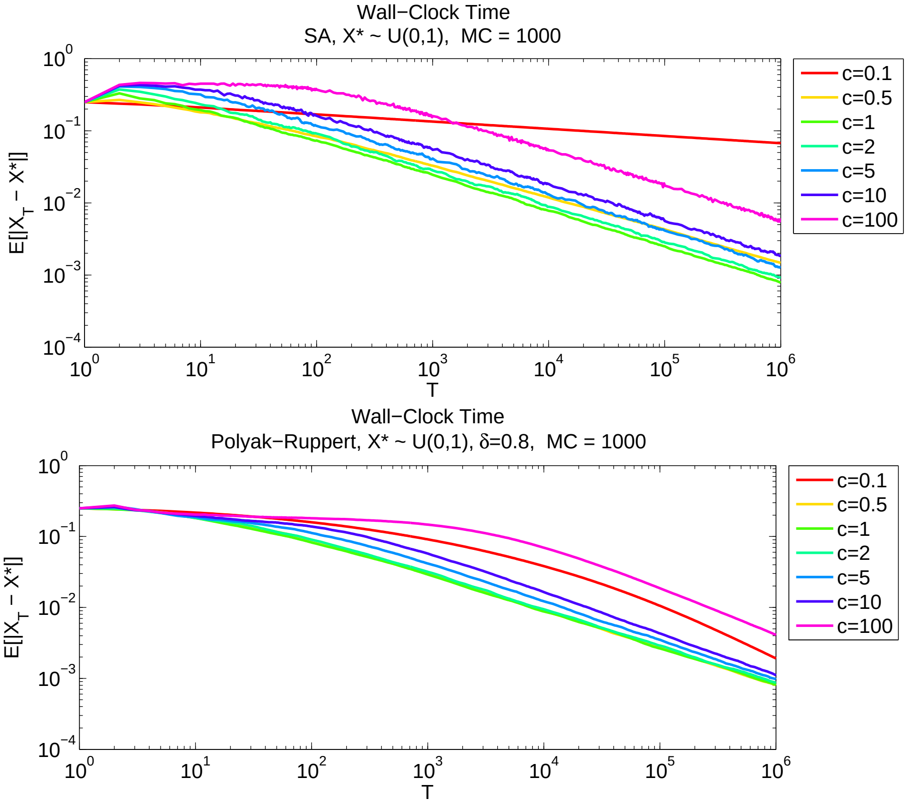

Figure 4.16: The performance of the SA and Polyak-Ruppert algorithms for a set of different input parameters $c$. The test function is $g_2(x) = X^* - x$. Here, we chose $\delta = 0.8$ for the Polyak-Ruppert algorithm and $X_0 = 0.5$ as the starting point for all tested algorithms. If $c$ is chosen too small for the SA algorithm, then the optimal rate of convergence might not be achieved. Also, the normalizing constant of the limiting behavior depends on the chosen tuning sequence. We use 1,000 independent sample runs to estimate these paths, and the relative error is always smaller than 5%.

Assumption (ii) given in Section 1.3 is violated, that is, $c < -1/(2g'(X^*))$. As previously discussed in Section 1.3, the dependency on this assumption is removed by the Polyak-Ruppert averaging method. Even though the optimal rate is achieved, when the tuning sequence that is too large, SA-type algorithms show suboptimal convergence behavior in terms of the limiting constant.

### 4.2.3 Confidence Intervals

As a last empirical experiment we analyze the performance of the confidence interval $J_n(\alpha)$ introduced in Section 3.3 in wall-clock time, and compare its performance to the performance of approximate confidence intervals generated by SA-type algorithms.

Hsieh and Glynn (2002) suggest restarting the SA algorithm several times to construct at least approximate confidence intervals of the root $X^*$. Here, we use 100 independent runs of the SA algorithm ($a_n = 1/n$ and $X_0 = 0.5$) and then, for any fixed $n$, use these 100 estimates to generate an approximate confidence interval of $X^*$. We assume a total simulation budget of $10^7$ samples, therefore each individual run of the SA algorithm consists of $10^5$ function evaluations. We have also used the same approach to construct confidence intervals for the Polyak-Ruppert averaging estimator (using $\delta = 0.8$).

The PBA, on the other hand, does not require repeatedly restarting and hence can use all $10^7$ function evaluations of one single sample path to construct a true confidence interval. Figure 4.17 shows the average width of the confidence intervals based on the PBA ($p_c = 0.95$), as well as the average width of the approximate confidence intervals based on the SA-type algorithms. Here, we use 250 independent samples to estimate the average widths.

From Figure 4.17 it becomes obvious that the width of the confidence intervals based on the PBA cannot compete with the width of the approximate confidence intervals based on SA-type algorithms. The reason being that the PBA only allows a few macro iterations and then stalls at an estimate very close to $X^*$. While such a close point estimate of $X^*$ is desirable, it prevents the algorithm from

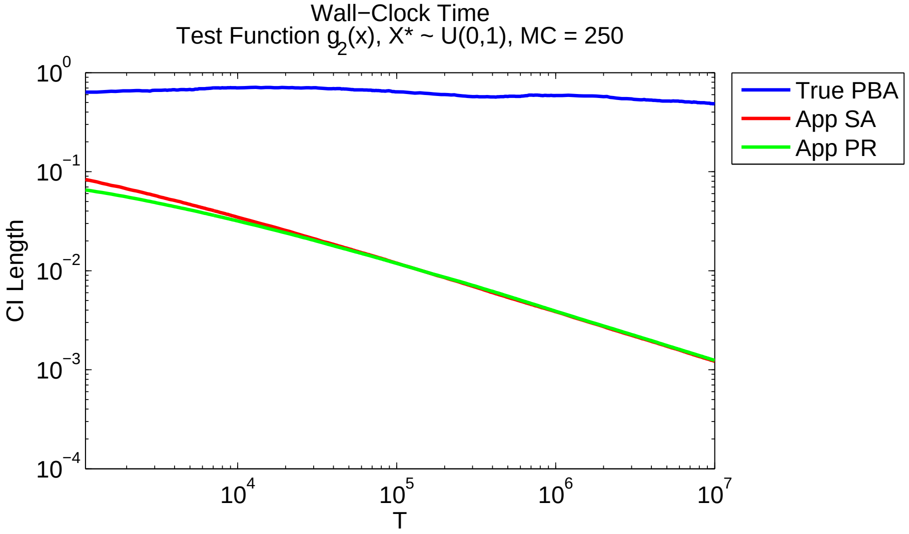

Figure 4.17: The estimated width of the true 95%-confidence intervals based on the PBA, as well as the estimated width of the approximate 95%-confidence intervals based on the SA algorithm and the Polyak-Ruppert averaging algorithm. For better visibility the $x$- and $y$-axes are shown in log-scale. Here, we use 250 independent samples to estimate the widths of the confidence intervals and the relative error is always smaller than 5%.

evaluating further measurement points and the confidence interval will not decrease any further. We have tested several other input parameters $p_c$ and confidence levels $\alpha$, but the general behavior remains the same. In order to achieve more macro iterations, and with this improve the quality of the confidence intervals, one could measure at different points than the median of the density $f_n$, such as quantiles, or realizations of a random variable with density $f_n$. Further investigation of this approach is left for future research.

Finally, it is informative to consider the estimated coverage probabilities of the confidence intervals; see Figure 4.18. In all 250 sample runs the point $X^*$ never leaves the confidence intervals defined by the PBA, whereas the coverage probability of the SA-type algorithms is alarming bad for small $T$, but slowly converges to the required confidence level $1 - \alpha$. This comparison shows the main difference between the two approaches of confidence intervals: the PBA generates a true (but conservative) confidence interval for all $T \in \mathbb{N}$, whereas the SA-type algorithms provide approximate confidence intervals, which will contain the root $X^*$ only as $T \to \infty$.

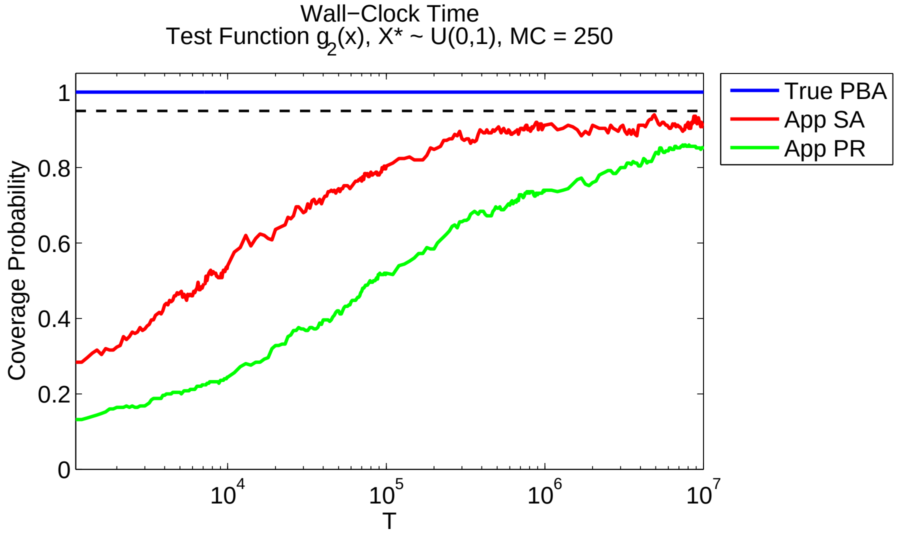

Figure 4.18: The estimated coverage probabilities of the 95%-confidence intervals based on the PBA as well as for the approximate 95%-confidence intervals based on SA-type algorithms. We used 250 independent sample runs to estimate these quantities. The PBA produces conservative confidence intervals that provide a statistical guarantee for every $T \in \mathbb{N}$, whereas the SA-type algorithms produces confidence intervals that are only valid as $T \to \infty$. For better visibility the $x$-axis is displayed in log-scale.

CHAPTER 5

# CONCLUSIONS AND FUTURE RESEARCH

In this thesis we have provided a thorough analysis and discussion of the Probabilistic Bisection Algorithm (PBA) for the one-dimensional stochastic root-finding problem. Since this algorithm is based on a bisection approach, it is conceptually very different from existing stochastic root-finding algorithms that mimic steepest-descent algorithms. The PBA was introduced in Horstein (1963), but very little has been known about its theoretical properties. We have shown several key results for the PBA, such as consistency and rate of convergence for different stochastic root-finding settings. In summary, if the function $g$ (for which one wants to find a root) has a discontinuity at the root $x^*$ then the sequence of estimators generated by the PBA converges to $x^*$ at a geometric rate, and therefore significantly outperforms existing stochastic root-finding algorithms. If $g$ is continuous at $x^*$, then, assuming a certain conjecture holds, an "averaged" estimator based on the PBA converges to $x^*$ at a near-optimal rate. In addition to these asymptotic properties, we have shown that the PBA also provides the simulation analyst with finite-time guarantees on the location of $x^*$, such as a confidence interval. To the best of our knowledge, this is the first stochastic root-finding method that provides such a guarantee.

The presented results directly pose two important questions regarding the PBA that are left for future research:

1. Does the Conjecture 1 or the weaker Conjecture 2 indeed hold? While these conjectures seem very reasonable based on intuitive arguments and numerical examples, formal proofs of them are still missing. The correctness of the second conjecture is needed for the proof of Theorem 9 to hold, which in turn

shows that an averaging-scheme attains a near-optimal rate of convergence for general stochastic root-finding problems.

2. Does the time-weighted average of the medians, that is, $\hat{X}_T = \frac{1}{T+1}\sum_{i=0}^{T} X_i$, where $(X_i)_i$ are the measurement points of the PBA in wall-clock time, indeed converge to $x^*$ at the known optimal rate $O(T^{-1/2})$ as the empirical examples suggest? If the answer is positive, the asymptotic performance of the PBA can directly compete with the asymptotic performance of popular SA-type algorithms, and one would want to develop a finer understanding of the rate of convergence of the algorithm through the multiplicative constant in $O(T^{-1/2})$.

From a larger perspective there exist several important future research directions regarding the PBA, including:

1. Analyzing different sampling schemes based on the PBA updating. The original PBA states that at each step the function $g$ should be evaluated at the median of the density $f_n$, and that the median in turn also provides a good approximation of $x^*$. Therefore, the PBA always tries to measure as close as possible to $x^*$. Such a sampling scheme is reasonable if the probability of a correct sign $p(\cdot)$ is constant, which is the setting the PBA originally was designed for. But when $p(\cdot)$ varies with $x$—especially when $\lim_{x\to x^*} p(x) = 1/2$—then the test of power one is likely to require many function evaluations the closer the current measurement points is to $x^*$. In this case it might be beneficial to measure at a different point than the median, such as at a quantile of the density $f_n$ or at the realization of a random variable with probability density $f_n$.

2. Investigating the robustness of the PBA. One of the drawbacks of SA-type

algorithms is that they may lack robustness, for example, when the noise distribution has heavy tails. The PBA, which maintains a posterior density on the location of the root $x^*$, is more robust with respect to extreme noise observations. This advantage of the PBA over SA-type algorithms seems to merit further investigation.

3. Analyzing the rate of convergence with respect to overall time. Our measure of wall-clock time keeps track of the number of function evaluations. While this is a reasonable measure for many applications it would be also informative to account for the overall time of the algorithm on a human time scale. For example, updating the posterior density of the PBA requires significantly more work than an updating step of SA-type algorithms. In contrast, there is also a cost associated with switching the measurement point $x$ (as discussed, for example, in Hong and Nelson, 2005) and SA-type algorithms will be slowed by their frequent switching, whereas the PBA only switches sporadically. Furthermore, the evaluation of the function $g$ might require a varying amount of work depending on the prescribed point $x$. All the above considerations affect the rate of convergence of the algorithm on a human time scale, and it is important to compare stochastic root-finding algorithms under this perspective as well.

4. Extending the PBA to higher dimensions. The method of centers of gravity, developed independently in Levin (1965) and Newman (1965), generalizes noise-free bisection (in the optimization setting) to higher dimensions. Nemirovski and Yudin (1983) provide a discussion of complexity and efficiency results of the method of centers of gravity and the subsequent ellipsoid method for deterministic optimization problems. A similar multivariate extension of the PBA seems plausible. Major challenges are proper updating,

the tracking of the posterior density, as well as the introduction of multidimensional tests of power one. A multivariate extension, however, would be very useful for many applications, including simulation optimization.

APPENDIX A

# ADDITIONAL RESULTS AND PROOFS

The following lemma provides the Bayesian updating for the case when $p(\cdot)$ is constant and known, that is, $p(\cdot) \equiv p$ for some constant $p \in (1/2, 1)$. For this, let $\mathcal{G}_n = \sigma\left(X_m, Z_m(X_m) : 0 \leq m \leq n\right)$ be the $\sigma$-algebra generated by the measurement points $(X_i)_{i=0}^n$ and signs $(Z_i(X_i)_{i=0}^n$ and $\mathcal{G}_{-1}$ be the trivial $\sigma$-algebra.

**Lemma 8.** *The domain of the prior density function $f_0$ is $[0, 1]$. Assume that $p(\cdot) \equiv p$, for some constant $p \in (1/2, 1)$ and that $q = 1 - p$. The sequence of posterior densities $(f_n)_n$ is given by the following iterative process, where $x$ is a point in the interior of $f_n$ at which the function $g$ is called at step $n$.*

$$\text{If } Z_n(x) = +1, \text{ then } f_{n+1}(y) = \begin{cases} \eta(x)^{-1} p f_n(y), & \text{if } y \geq x, \\ \eta(x)^{-1}(1-p) f_n(y), & \text{if } y < x, \end{cases} \tag{A.1}$$

$$\text{if } Z_n(x) = -1, \text{ then } f_{n+1}(y) = \begin{cases} (1-\eta(x))^{-1}(1-p) f_n(y), & \text{if } y \geq x, \\ (1-\eta(x))^{-1} p f_n(y), & \text{if } y < x, \end{cases} \tag{A.2}$$

*where $\eta(x) = \mathbb{P}(Z_n(x) = +1 | \mathcal{G}_{n-1}) = (1 - F_n(x))p + F_n(x)(1-p)$ and $F_n$ denotes the cdf of the density $f_n$.*

*Proof.* Conditional on $X^*$ and $\mathcal{G}_{n-1}$, the random variable $Z_n(x)$ assumes the value $+1$ with the following probabilities:

$$\mathbb{P}(Z_n(x) = +1 | X^* \geq x, \mathcal{G}_{n-1}) = p,$$
$$\mathbb{P}(Z_n(x) = +1 | X^* < x, \mathcal{G}_{n-1}) = 1 - p.$$

The conditional distribution of the event $\{Z_n(x) = +1\}$ given $\mathcal{G}_{n-1}$ is then computed as

$$\begin{aligned}
&\mathbb{P}(Z_n(x) = +1|\mathcal{G}_{n-1}) \\
&\qquad\qquad\qquad = \mathbb{P}(X^* \geq x|\mathcal{G}_{n-1})\mathbb{P}(Z_n(x) = +1|X^* \geq x, \mathcal{G}_{n-1}) \\
&\qquad\qquad\qquad\quad + \mathbb{P}(X^* < x|\mathcal{G}_{n-1})\mathbb{P}(Z_n(x) = +1|X^* < x, \mathcal{G}_{n-1}) \\
&\qquad\qquad\qquad = (1 - F_n(x))p + F_n(x)(1-p) = \eta(x),
\end{aligned} \tag{A.3}$$

where the first equation follows from the law of total probability. The result now follows from Bayes' rule. That is, on the event $\{Z_n(x) = +1\}$ we have

$$\begin{aligned}
f_{n+1}(y) &= \frac{\mathbb{P}(Z_n(x) = +1|\mathcal{G}_{n-1}, X^* = y)f_n(y)}{\mathbb{P}(Z_n(x) = +1|\mathcal{G}_{n-1})} \\
&= \begin{cases} \eta(x)^{-1} p f_n(y), & \text{if } y \geq x, \\ \eta(x)^{-1}(1-p) f_n(y), & \text{if } y < x. \end{cases}
\end{aligned}$$

The expression (A.2) for $f_{n+1}(y)$ on the event $\{Z_n(x) = -1\}$ is derived similarly.

□

*Proof of Proposition 1.* The definition of entropy and the tower property of conditional expectation imply

$$\mathbb{E}\big[H(f_N)\,\big|\,X_{N-1} = x, f_{N-1}\big] = \mathbb{E}\big[-\log_2 f_N(X^*)\,\big|\,X_{N-1} = x, f_{N-1}\big].$$

Using the updating equations described in Lemma 8 for the query $X_{N-1} = x$ we can decompose the random variable $-\log_2 f_N(X^*)|X_{N-1} = x, f_{N-1}$ into a sum of three terms for both possible outcomes of $Z_{N-1}(X_{N-1})$:

$$\begin{aligned}
&\text{If } Z_{N-1}(X_{N-1}) = +1, \text{ then} \\
&\quad -\log_2 f_N(X^*)|X_{N-1} = x, f_{N-1} \\
&\qquad\qquad = -\log_2 f_{N-1}(X^*) - \log_2 \eta(x)^{-1} - \begin{cases} \log_2 p, & \text{if } X^* \geq x, \\ \log_2(1-p), & \text{if } X^* < x, \end{cases} \\
&\text{if } Z_{N-1}(X_{N-1}) = -1, \text{ then} \\
&\quad -\log_2 f_N(X^*)|X_{N-1} = x, f_{N-1} \\
&\qquad\qquad = -\log_2 f_{N-1}(X^*) - \log_2 (1-\eta(x))^{-1} - \begin{cases} \log_2(1-p), & \text{if } X^* \geq x, \\ \log_2 p, & \text{if } X^* < x. \end{cases}
\end{aligned}$$

By the linearity of the expectation operator we can calculate the expected value of each of the three terms separately. The first term is independent of $Z_{N-1}(X_{N-1})$ and simply recovers the entropy at time $N-1$,

$$\mathbb{E}\big[-\log_2 f_{N-1}(X^*)\,\big|\,X_{N-1} = x, f_{N-1}\big] = H(f_{N-1}).$$

To evaluate the second term we use the fact that $\mathbb{P}(Z_{N-1}(x) = +1|f_{N-1}) = (1-F_{N-1}(x))p + F_{N-1}(x)(1-p) = \eta(x)$ as was shown in (A.3). And the expectation of the second term is

$$\begin{aligned}
&\mathbb{E}\big[\mathbb{1}\{Z_{N-1}(x) = +1\} \log_2 \eta(x) + \mathbb{1}\{Z_{N-1}(x) = -1\} \log_2(1-\eta(x))\,\big|\,f_{N-1}\big] \\
&\quad = \eta(x) \log_2 \eta(x) + (1-\eta(x)) \log_2(1-\eta(x)).
\end{aligned}$$

The third term is equal to $\log_2 p$ when the sign $Z_{N-1}(X_{N-1})$ is correct and $\log_2(1-p)$ otherwise. This is independent of the measurement point $x$. Hence the expectation of the third term equals $-p\log_2 p - (1-p)\log_2(1-p)$. Combining these three terms together and noting that the first and third terms do not depend on the measurement location $x$ yields

$$\inf_{x\in[0,1]} \mathbb{E}[H(f_N)|X_{N-1} = x, f_{N-1}] = H(f_{N-1}) - p\log_2 p - (1-p)\log_2(1-p) \\ + \inf_{x\in[0,1]} \left[\eta(x)\log_2 \eta(x) + (1-\eta(x))\log_2(1-\eta(x))\right].$$

The inner expression over which we take the infimum depends on $x$ only through $\eta(x)$, the probability of observing $Z_{N-1}(x) = +1$, which can take values in $[0,1]$. Consider the function $\eta\log_2\eta + (1-\eta)\log(1-\eta)$, which is strictly convex and has a global minimum at $\eta = 1/2$. Further $\eta(x) = 1/2$ when $F_{N-1}(x) = 1/2$, which shows that the optimal choice of $x$ is the median of the pdf $f_{N-1}$. Finally, combining all three terms yields

$$\mathbb{E}[H(f_N)|F_{N-1}(X_{N-1}) = 1/2, f_{N-1}] = \\ H(f_{N-1}) - p\log_2 p - (1-p)\log_2(1-p) - 1,$$

and this finishes the proof. □

*Proof of Theorem 1.* We show for each $n = 0, 1, \ldots, N$ that the value function is as claimed in (2.3), and that the median achieves the minimum in Bellman's recursion (2.2). This is sufficient to show the claim.

We proceed by backward induction on $n$. The value function clearly has the claimed form at the final time, $n = N$. Now, fix any $n < N$ and assume that the value function is of the form claimed for $n+1$. Then Bellman's recursion and the induction hypothesis show that

$$\begin{aligned}
&V_n(f_n) \\
&= \inf_{x\in[0,1]} \mathbb{E}\big[V_{n+1}(f_{n+1}) \,\big|\, X_n = x, f_n\big] \\
&= \inf_{x\in[0,1]} \mathbb{E}\big[H(f_{n+1}) - (N-n-1)(1 + p\log_2 p + (1-p)\log_2(1-p)) \,\big|\, X_n = x, f_n\big] \\
&= \inf_{x\in[0,1]} \mathbb{E}\big[H(f_{n+1}) \,\big|\, X_n = x, f_n\big] - (N-n-1)(1 + p\log_2 p + (1-p)\log_2(1-p)).
\end{aligned}$$

Finally, Proposition 1 shows that the infimum is achieved at the median $\inf\{x : F_n(x) \geq 1/2\}$, and that the resulting value is

$$V_n(f_n) = H(f_n) - (N-n)(1 + p\log_2 p + (1-p)\log_2(1-p)),$$

as stated in the theorem. □

*Proof of Lemma 1.* Let us first focus on the definition of $C$. The reason for defining $C$ in this way will become clear towards the end of the proof. Consider the function

$$U(u) = \left(\frac{u + D}{\log(2p) - \log(2q)}\right)^2,$$

and note that

1. $U$ is convex and non-negative;
2. $U(|D|) = 0$, because $D < 0$.

These two properties imply that there exists a unique $\tilde{u} \in (0, |D|)$ such that $U(\tilde{u}) = \tilde{u}$. Then define $C = e^{\tilde{u}}$ and consequently $1 < C < e^{|D|}$.

Now we return to the random walk $(R_n)_n$. For any $n \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{P}\left(e^{R_n} > C^{-n}/2\right) &= \mathbb{P}\left(R_n > \log(C^{-n}/2)\right) \\
&= \mathbb{P}\left(R_0 + \sum_{j=1}^{n} \psi_j > \log(2^{-1}C^{-n})\right) \\
&\leq \mathbb{P}\left(\log(1/2) + \sum_{j=1}^{n} \psi_j > \log(1/2) - n\log C\right) \\
&= \mathbb{P}\left(\sum_{j=1}^{n} \psi_j > -n\log C\right) \\
&= \mathbb{P}\left(\sum_{j=1}^{n} \psi_j - nD > -n\log C - nD\right) \\
&= \mathbb{P}\left(\overline{\psi}_n - D > -\log C - D\right),
\end{aligned}$$

where $\overline{\psi}_n = n^{-1}\sum_{j=1}^{n} \psi_j$ and $\mathbb{E}[\psi_j] = (\log(2p) + \log(2q))/2 = D$. The increments $\psi_j$ are iid and bounded, and $C < e^{|D|}$ which implies that $-\log C - D > 0$, so we can apply Hoeffding's bound[1]:

$$\mathbb{P}\left(e^{R_n} > C^{-n}/2\right) \leq \exp\left(-2\left(\frac{\log C + D}{\log(2p) - \log(2q)}\right)^2 n\right).$$

Now by definition of $C$

$$\left(\frac{\log C + D}{\log(2p) - \log(2q)}\right)^2 = \log C,$$

and hence $\mathbb{P}(e^{R_n} > C^{-n}/2) \leq C^{-2n}$, which holds for any chosen $n \in \mathbb{N}$. □

*Proof of Theorem 3.* Consider arbitrary $\varepsilon > 0$. Proposition 4 shows that $\mathbb{P}\big(c^n \mathbb{E}_n[|X_n - X^*|] > \varepsilon\big) \leq C^{-n}$ for $n > \hat{N} = 0 \vee \widetilde{N}(\varepsilon, c, C)$, then

$$\sum_{n=0}^{\infty} \mathbb{P}\big(c^n \mathbb{E}_n[|X_n - X^*|] > \varepsilon\big) \leq \hat{N} + \frac{C^{-\hat{N}+1}}{C-1} < \infty.$$

[1] Let $X_1, \ldots, X_n$ be iid bounded random variables, that is, $\mathbb{P}(X_i \in [a,b]) = 1$. Then for the empirical mean $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ the inequality $\mathbb{P}\big(\overline{X} - \mathbb{E}[\overline{X}] \geq t\big) \leq \exp\big(-2t^2 n(b-a)^{-2}\big)$ holds when $t \geq 0$. See Hoeffding (1963).

By the lemma of Borel-Cantelli it follows that $\mathbb{P}\big(c^n\mathbb{E}_n[|X_n - X^*|] > \varepsilon \text{ i.o.}\big) = 0$[2]. Since this holds for any $\varepsilon > 0$ it follows that $c^n\mathbb{E}_n[|X_n - X^*|] \to 0$, and hence $\mathbb{E}_n[|X_n - X^*|] \to 0$, almost surely as $n \to \infty$. □

*Proof of Corollary 1.* Let $E$ be the set of probability one where the convergence $\mathbb{E}_n[|X_n - X^*|] \to 0$ holds. For a sample path $\omega \in E$ it holds that the probability measure of $|X_n - X^*|$ converges weakly to a point mass at 0 (since $L^1$-convergence implies convergence in distribution), which is equivalent of $\mathbb{P}_n(\cdot)(\omega)$ converging weakly to a point mass at $X^*(\omega)$. □

*Proof of Corollary 2.* Consider a sample path $\omega \in E$, where $E$ is the set of probability one on which $\lim_{n\to\infty} F_n(x) = \mathbb{1}\{x \geq X^*\}$ holds. Then for any $\varepsilon > 0$ it holds that $\lim_{n\to\infty} F_n(X^* - \varepsilon) = 0$ and $\lim_{n\to\infty}(1 - F_n(X^* + \varepsilon)) = 0$, and hence there exists an $N(\varepsilon) \in \mathbb{N}$ such that $F_n(X^* + \varepsilon) - F_n(X^* - \varepsilon) > 0.5$ for all $n > N(\varepsilon)(\omega)$. By definition of the median it follows that $X_n \in (X^* - \varepsilon, X^* + \varepsilon)$ for all $n > N(\varepsilon)$. Since this holds for any $\varepsilon > 0$ it follows that $X_n \to X^*$ as $n \to \infty$ on this sample path, and hence $X_n \to X^*$ almost surely as $n \to \infty$. □

*Proof of Lemma 5.* As shown in the proof of Proposition 8, $d - q\beta > 0$. Now consider arbitrary $r \in (0, d - q\beta)$, $n \geq N_1$ and $\alpha \in (0, 1)$. Then

$$\mathbb{P}\left(W_n \geq \frac{b_n - rn}{\beta}\right) = \mathbb{P}\left(\frac{1}{n}W_n - q \geq \frac{b_n/n - r}{\beta} - q\right) = \mathbb{P}\left(\frac{1}{n}W_n - q \geq t\right), \tag{A.4}$$

where we defined

$$t = \frac{b_n/n - r - q\beta}{\beta}.$$

[2] i.o. stands for infinitely often, that is, $\{c^n\mathbb{E}_n[|X_n - X^*|] > \varepsilon \text{ i.o.}\} = \bigcap_{n=0}^{\infty}\bigcup_{j=n}^{\infty}\{c^j\mathbb{E}_j[|X_j - X^*|] > \varepsilon\}$.

In order to use Hoeffding's bound we need that $t > 0$, indeed

$$\begin{aligned}
t &= \frac{d - n^{-1/2}\left(-\frac{1}{2}\log(\alpha/2)\right)^{1/2}\beta - r - q\beta}{\beta} \\
&= \frac{d - r - q\beta}{\beta} - n^{-1/2}\left(-\frac{1}{2}\log(\alpha/2)\right)^{1/2} \\
&\geq \frac{d - r - q\beta}{\beta} - \frac{d - r - q\beta}{(2\log(2/\alpha))^{1/2}}\left(-\frac{1}{2}\log(\alpha/2)\right)^{1/2} \\
&= \frac{d - r - q\beta}{\beta} - \frac{d - r - q\beta}{2\beta} > 0,
\end{aligned}$$

where the second last inequality follows since $n \geq N_1$. Applying Hoeffding's inequality[3] to (A.4) shows that

$$\mathbb{P}\left(W_n \geq \frac{b_n - rn}{\beta}\right) \leq \exp(-2t^2 n).$$

It remains to show that $\exp(-2t^2 n) \leq \alpha/2$, that is, $tn^{1/2} \geq ((1/2)\log(2/\alpha))^2$, indeed,

$$\begin{aligned}
tn^{1/2} &= \frac{b_n/n - r - q\beta}{\beta} n^{1/2} \\
&= \frac{d - n^{-1/2}\left(-(1/2)\log(\alpha/2)\right)^{1/2}\beta - r - q\beta}{\beta} n^{1/2} \\
&= \left(\frac{d - r - q\beta}{\beta}\right) n^{1/2} - \left(-\frac{1}{2}\log(\alpha/2)\right)^{1/2} \\
&\geq \left(\frac{d - r - q\beta}{\beta}\right)(2\log(2/\alpha))^{1/2}\left(\frac{\beta}{d - r - q\beta}\right) - \left(-\frac{1}{2}\log(\alpha/2)\right)^{1/2} \\
&= (2\log(2/\alpha))^{1/2} - \left(\frac{1}{2}\log(2/\alpha)\right)^{1/2} \\
&= \left(\frac{1}{2}\log(2/\alpha)\right)^{1/2},
\end{aligned}$$

where the inequality follows since $n \geq N_1$. □

*Proof of Lemma 6.* Let $(k_n)_n$ be the curved boundary for the test of power one for Bernoulli random variables as defined in (B.5) where $\gamma$ is replaced with $\alpha/2$.

[3] Let $X_1, \ldots, X_n$ be iid bounded random variables, that is, $\mathbb{P}(X_i \in [a, b]) = 1$. Then for the empirical mean $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ the inequality $\mathbb{P}\big(\overline{X} - \mathbb{E}[\overline{X}] \geq t\big) \leq \exp\big(-2t^2 n(b-a)^{-2}\big)$ holds when $t \geq 0$. See Hoeffding (1963).

Then, by construction of this test,

$$\mathbb{P}\left(|W_n - nq| \geq k_n \text{ for some } n \geq 1\right) \leq \alpha/2,$$

and

$$\begin{aligned}\{|W_n - nq| \geq k_n \text{ for some } n \geq 1\} &\supseteq \{W_n - nq \geq k_n \text{ for some } n \geq 1\} \\ &= \{W_n \geq k_n + nq \text{ for some } n \geq 1\}.\end{aligned}$$

To finish the proof it remains to show that there exists a constant $N_2$ such that

$$k_n + nq \leq \frac{a_n - rn}{\beta}, \tag{A.5}$$

for all $n \geq N_2$. Using the definitions of $(a_n)_n$ given in (3.17) (use $p_c = p$) and $(k_n)_n$ given in (B.5) (use $\gamma = \alpha/2$), the condition (A.5) is equivalent to

$$\left(\frac{\beta q - d + r}{\beta}\right) n + 2\left[-\frac{1}{2}\log\left(\frac{\alpha}{n+1}\right)\right]^{1/2} n^{1/2} \leq 0,$$

for all $n \geq N_2$. Since $r < d - q\beta$ such a constant $N_2$ always exists. □

APPENDIX B

# TESTS OF POWER ONE

In this appendix we provide additional details from statistical tests of power one, also referred to as tests with curved boundaries. See, Robbins (1970) and Siegmund (1985), Chapter 4, for in-depth discussions of such tests.

Let $(\xi_n(\theta))_n$ be an iid sequence of random variables with mean $\mathbb{E}[\xi_i(\theta)] = \theta$. The goal of a test of power one is to decide whether the hypothesis $\theta < \theta_0$ or the alternative $\theta > \theta_0$ holds. Such a test observes the random walk $S_n(\theta) = \sum_{i=1}^{n}(\xi_i(\theta) - \theta_0)$ until

$$N(\theta) = \inf\{n \geq 1 : S_n(\theta) \geq |k_n|\}, \tag{B.1}$$

where $(k_n)_n$ is an increasing positive sequence, also referred to as a curved boundary. When $N(\theta) < \infty$, the test decides that $\theta < \theta_0$ if $S_{N(\theta)} < 0$ and decides $\theta > \theta_0$ if $S_{N(\theta)} > 0$; and when $N(\theta) = \infty$ it does not provide a decision. Such a test needs to satisfy the following properties:

$$\begin{aligned} \mathbb{P}\left(N(\theta_0) < \infty\right) &\leq \gamma, \qquad \text{and} \\ \mathbb{P}\left(N(\theta) < \infty\right) &= 1, \qquad \text{for all } \theta \neq \theta_0, \end{aligned}$$

where $\gamma \in (0, 1)$ is a confidence parameter. Often, the second property follows immediately by the law of large numbers, whereas the first property needs more justification. An elegant method of verifying the first property is by means of a likelihood ratio argument, initially introduced in Ville (1939). (See also Wald, 1947 and Robbins, 1970.)

**Proposition 13** (Ville, 1939). *Suppose that under $\mathbb{P}$ for each $n \geq 1$ the random variables $\xi_1, \ldots, \xi_n$ have a pdf $\rho_n$ with respect to a $\sigma$-finite measure $\lambda^{(n)}$ on the Borel sets of an $n$-dimensional metric space, and that $\mathbb{P}'$ is any other joint probability* *distribution of the sequence* $(\xi_n)_n$ *such that* $\xi_1, \ldots, \xi_n$ *have a pdf* $\rho'_n$ *with respect to the same measure* $\lambda^{(n)}$*. Define the likelihood ratio* $L_n = \rho'_n/\rho_n$ *when* $g_n > 0$*. Then, for any* $\gamma \in (0,1)$*,*

$$\mathbb{P}\left(L_n \geq 1/\gamma \text{ for some } n \geq 1\right) \leq \gamma. \tag{B.2}$$

*Proof (Ville, 1939; Robbins, 1970).* Let $N = \inf\{n \geq 1 : \rho'_n \geq \rho_n/\gamma\}$, then $\mathbb{P}\left(\rho_n = 0 \text{ for some } n \geq 1\right)$, and

$$\begin{aligned}\mathbb{P}\left(L_n \geq 1/\gamma \text{ for some } n \geq 1\right) &= \mathbb{P}\left(N < \infty\right)\\ &= \sum_{i=1}^{\infty} \int_{(N=n)} \rho_n d\lambda^{(n)}\\ &\leq \gamma \sum_{i=1}^{\infty} \int_{(N=n)} \rho'_n d\lambda^{(n)} = \gamma \cdot \mathbb{P}'(N < \infty) \leq \gamma. \qquad \square\end{aligned}$$

Let us list the explicit constructions of the boundary $(k_n)_n$ for three different distributions of $(\xi_n(\theta))_n$ that are used in this thesis.

1. *Normal Distribution (Robbins, 1970; Siegmund, 1985):* Assume that $(\xi(\theta)_n)_n$ is a sequence of iid $N(\theta, 1)$ random variables and denote with $\mathbb{P}_\theta(\cdot)$ the corresponding probability measure. Here, we test the hypothesis $\theta < 0$ versus $\theta > 0$. Consider the normal mixture distribution of the form $\mathbb{P}'(\cdot) = \int_{-\infty}^{\infty} P_\theta(\cdot)\phi(x)dx$, where $\phi(x)$ is the standard normal pdf. Then

$$L_n = \int_{-\infty}^{\infty} \exp(xS_n - \frac{1}{2}nx^2)\phi(x)dx,$$

   and the stopping time defined in Proposition 13 is equal to (B.1), where

$$k_n = \left((n+1)[\log(n+1) - 2\log\gamma]\right)^{1/2}. \tag{B.3}$$

2. *Bernoulli Distribution (Robbins, 1970):* Assume that $(\xi(\theta)_n)_n$ is a sequence of iid Bernoulli($\theta$) random variables and denote with $\mathbb{P}_\theta(\cdot)$ the corresponding

probability measure. Here, we test the hypothesis $\theta < \theta_0$ versus $\theta > \theta_0$. Consider the uniform mixture of Bernoulli distributions with parameter $0 < x < 1$, that is, $\mathbb{P}'(x_1, \ldots, x_n) = \int_0^1 x^{\sum_{i=1}^n x_i}(1-x)^{n-\sum_{i=1}^n x_i}dx$. Using the construction from Proposition 13 a test of power one is defined by the stopping rule

$$N(\theta) = \inf\left\{n \geq 1 : \binom{n}{B_n}\theta_0^{B_n}(1-\theta_0)^{n-B_n} \leq \frac{\gamma}{n+1}\right\}, \tag{B.4}$$

where $B_n = \sum_{i=1}^n \xi_i(\theta)$. Applying Hoeffding's bound to $\binom{n}{B_n}\theta_0^{B_n}(1-\theta_0)^{n-B_n}$ defines a test of power with stopping rule as given in (B.1), where

$$k_n = \left(n\left(\log(n+1) - \log 2 - \log\gamma\right)/2\right)^{1/2}. \tag{B.5}$$

3. *Simple Random Walk:* Assume that $\left(\xi_n(\theta)\right)_n$ is a sequence of iid random variables with distribution $\mathbb{P}\left(\xi_i = +1\right) = (1+\theta)/2$ and $\mathbb{P}\left(\xi_i = -1\right) = (1-\theta)/2$, for $\theta \in [-1, 1]$. We use the test designed for Bernoulli random variables to test the hypothesis $\theta < \theta_0$ against the alternative $\theta > \theta_0$. In this case $S_n(\theta) = 2\widetilde{S}_n(\theta) - n$, where $\widetilde{S}_n(\theta)$ is a random walk with Bernoulli increments, and hence the stopping rule (B.1) where

$$k_n = \left(2n\left(\log(n+1) - \log 2 - \log\gamma\right)\right)^{1/2}, \tag{B.6}$$

defines a test of power one for the simple random walk with increments $\xi \in \{-1, +1\}$.

While the method of likelihood ratios as given in Proposition 13 provides tests of power one, these are not necessarily optimal tests. The performances of different tests of power are usually compared by the expected stopping time as a function of $\theta$, especially as $\theta \to \theta_0$. To this end, Robbins and Siegmund (1974) and Lai (1977) (Theorem 1, Example 1) show that the expected hitting time of a test with curved boundary of the form $O\big((n\log n)^{1/2}\big)$ (which includes tests defined by (B.3), (B.5) and (B.6)) satisfies

$$\mathbb{E}[N(\theta)] \sim (\theta-\theta_0)^{-2}\log(|\theta-\theta_0|^{-1}), \quad \text{as } \theta \to {\theta_0}^{1}. \tag{B.7}$$

This result is intuitive, since one would expect that a test of power one on average requires more samples than a fixed-sized hypothesis test of $\theta = \theta_0$ versus $\theta = \theta_1$, which requires, under mild assumptions on the noise distribution, $O(|\theta_1 - \theta_0|^{-2})$ samples. The above tests are not asymptotically optimal since

$$\mathbb{E}[N(\theta)] \sim (\theta-\theta_0)^{-2}\,|\log|\log|\theta-\theta_0|||, \quad \text{as } \theta \to \theta_0, \tag{B.8}$$

is an optimal rate for the expected hitting time of a test of power one when the increment distribution belongs to the exponential family (Farrell, 1964). For some noise distribution there exist explicit curved boundaries $(k_n)_n$ that achieve this optimal rate, which are of the form

$$k_n \sim \left(2n(\log\log(n+e) + c\log\log\log(n+e^e))\right)^{1/2}, \quad \text{as } n \to \infty, \tag{B.9}$$

for some $c > 3/2$. Such optimal tests, however, are difficult to calibrate to a required confidence level $\gamma$ and only hold for large $n$, that is, the stopping time assumes the form $N = \inf\{n \geq n_1 : |S_n| \geq k_n\}$ for some large constant $n_1 \in \mathbb{N}$. See Farrell (1964), Robbins and Siegmund (1974) and Lai (1977) for proofs and discussions of the above results.

Although optimal tests of power one may perform slightly better asymptotically, the tests defined by the simpler boundaries (B.3), (B.5) and (B.6) are sufficient for the presented results in this thesis. The potential improvement of using optimal tests, as well as the investigation of nonasymptotic behavior of different tests of power one, are both promising topics for future research.

[1] Here, the notation $g(x) \sim f(x)$ means that $\lim_{x\to x_0} f(x)/g(x) = a$ for some constant $a > 0$.

# BIBLIOGRAPHY

Ben-Or, M. and Hassidim, A. (2008). The Bayesian learner is optimal for noisy binary search (and pretty good for quantum as well). In *49th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 221–230. IEEE.

Billingsley, P. (1999). *Convergence of Probability Measures, 2nd edition.* Wiley-Interscience.

Bottou, L. (2004). Stochastic learning. In *Advanced Lectures on Machine Learning.* Bousquet, O., von Luxburg, U., and Rätsch, G., Eds., volume 3176 of *Lectures Notes in Computer Science*, pp. 146–168. Springer.

Broadie, M., Cicek, D., and Zeevi, A. (2011). General bounds and finite-time improvement for the Kiefer-Wolfowitz stochastic approximation algorithm. *Oper. Res.*, 59(5):1211–1224.

Burnashev, M. V. and Zigangirov, K. S. (1974). An interval estimation problem for controlled observations. *Problemy Peredachi Informatsii*, 10(3):51–61.

Casella, G. and Berger, R. L. (2002). *Statistical Inference, 2nd edition.* Duxbury.

Castro, R. M. and Nowak, R. D. (2008a). Active learning and sampling. In *Foundations and Applications of Sensor Management.* Hero, A. O., Castañón, D. A., Cochran, D., and Kastella, K., Eds., pp. 177–200. Springer.

Castro, R. M. and Nowak, R. D. (2008b). Minimax bounds for active learning. *IEEE Trans. Inform. Theory*, 54(5):2339–2353.

Cont, R. and Kukanov, A. (2012). Optimal order placement in limit order markets. *Available at SSRN: http://ssrn.com/abstract=2155218.*

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* John Wiley & Sons, New York, N.Y.

Dunkel, J. and Weber, S. (2010). Stochastic root finding and efficient estimation of convex risk measures. *Oper. Res.*, 58(5):1505–1521.

Durrett, R. (2005). *Probability: Theory and Examples, 3rd edition.* Duxbury Advanced Series.

Ehrlichman, S. M. T. and Henderson, S. G. (2007). Finite-sample performance guarantees for one-dimensional stochastic root finding. In *Proceedings of the Winter Simulation Conference.* Henderson, S. G., Biller, B., Hsieh, M.-H., Shortle, J., Tew, J. D., and Barton, R. R., Eds., pp. 313–321, Piscataway, N.J. IEEE.

Farrell, R. H. (1964). Asymptotic behavior of expected sample size in certain one sided tests. *Ann. Math. Statist.*, 35(1):36–72.

Feige, U., Raghavan, P., Peleg, D., and Upfal, E. (1994). Computing with noisy information. *SIAM J. Comput.*, 23(5):1001–1018.

Frazier, P. I., Powell, W. B., and Dayanik, S. (2008). A knowledge-gradient policy for sequential information collection. *SIAM J. Control Optim.*, 47(5):2410–2439.

Frees, E. W. and Ruppert, D. (1990). Estimation following a sequentially designed experiment. *J. Amer. Statist. Assoc.*, 85(412):1123–1129.

Fu, M. C., Glover, F. W., and April, J. (2005). Simulation optimization: a review, new developments, and applications. In *Proceedings of the Winter Simulation Conference.* Kuhl, M. E., Steiger, N. M., Armstrong, F. B., and Joines, J. A., Eds., pp. 83–95, Piscataway, N.J. IEEE.

Gut, A. (2009). *Stopped Random Walks: Limit Theorems and Applications.* Springer.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58(301):13–30.

Hong, J. L. and Nelson, B. L. (2005). The tradeoff between sampling and switching: New sequential procedures for indifference-zone selection. *IIE Transactions*, 37(7):623–634.

Horstein, M. (1963). Sequential transmission using noiseless feedback. *IEEE Trans. Inform. Theory*, 9(3):136–143.

Hsieh, M.-H. and Glynn, P. W. (2002). Confidence regions for stochastic approximation algorithms. In *Proceedings of the Winter Simulation Conference.* Yücesan, E., Chen, C.-H., Snowdon, J. L., and Charnes, J. M., Eds., pp. 370–376, Piscataway, N.J. IEEE.

Jedynak, B., Frazier, P. I., and Sznitman, R. (2012). Twenty questions with noise: Bayes optimal policies for entropy loss. *J. Appl. Prob.*, 49(1):114–136.

Karp, R. M. and Kleinberg, R. (2007). Noisy binary search and its applications. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 881–890. SIAM.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, 23(3):462–466.

Kushner, H. J. and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications.* Springer.

Lai, T. L. (1977). Power-one tests based on sample sums. *Ann. Statist.*, 5(5):866–880.

Lai, T. L. (2003). Stochastic approximation. *Ann. Statist.*, 31(2):391–406.

Levin, A. I. (1965). On an algorithm for the minimization of convex functions. *Dokl. Akad. Nauk*, 160:1244–1247. (translated in *Dokl. Math.* 6, 286–290, 1965).

Lim, E. (2011). On the convergence rate for stochastic approximation in the nonsmooth setting. *Math. Oper. Res.*, 36(3):527–537.

Nemirovski, A. S. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York.

Newman, D. J. (1965). Location of the maximum on unimodal surfaces. *J. ACM*, 12(3):395–398.

Nowak, R. D. (2008). Generalized binary search. In *46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 568–574.

Nowak, R. D. (2009). Noisy generalized binary search. In *Adv. Neural Inf. Process. Syst. 22*. Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., Eds., pp. 1366–1374.

Pasupathy, R. and Kim, S. (2011). The stochastic root-finding problem: overview, solutions, and open questions. *ACM Trans. Model. Comput. Simul.*, 21(3):19.

Pelc, A. (1989). Searching with known error probability. *Theoret. Comput. Sci.*, 63(2):185–202.

Polyak, B. T. (1990). New method of stochastic approximation type. *Autom. Remote Control*, 51:937–946.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855.

Rivest, R. L., Meyer, A. R., Kleitman, D. J., Winklmann, K., and Spencer, J. (1980). Coping with errors in binary search procedures. *J. Comput. System Sci.*, 20(3):396–404.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Ann. Math. Statist.*, 41(5):1397–1409.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407.

Robbins, H. and Siegmund, D. (1974). The expected sample size of some tests of power one. *Ann. Statist.*, 2(3):415–436.

Ruppert, D. (1991). Stochastic approximation. In *Handbook of Sequential Analysis*. Gosh, B. K. and Sen, P. K., Eds. Marcel-Dekker.

Siegmund, D. (1985). *Sequential Analysis: tests and confidence intervals*. Springer.

Ville, J. (1939). Etude critique de la notion de collectif. *Monographies des Probabilites*, 3:144.

Waeber, R., Frazier, P. I., and Henderson, S. G. (2010). Performance measures for ranking and selection procedures. In *Proceedings of the Winter Simulation Conference*. Johansson, B., Jain, S., Montoya-Torres, J., Hugan, J., and Yücesan, E., Eds., pp. 1235–1245, Piscataway, N.J. IEEE.

Waeber, R., Frazier, P. I., and Henderson, S. G. (2011). A Bayesian approach to stochastic root finding. In *Proceedings of the Winter Simulation Conference*.

Jain, S., Creasey, R. R., Himmelspach, J., White, K. P., and Fu, M., Eds., pp. 4033–4045, Piscataway, N.J. IEEE.

Waeber, R., Frazier, P. I., and Henderson, S. G. (2012a). A framework for selecting a selection procedure. *ACM Trans. Model. Comput. Simul.*, 22(3):16.

Waeber, R., Frazier, P. I., and Henderson, S. G. (2012b). Bisection search with noisy responses. *in review at SIAM J. Control Optim.*

Wald, A. (1947). *Sequential Analysis.* Wiley, New York.

Zangenehpour, S. (1993). Method and apparatus for positioning head of disk drive using zone-bit-recording. US Patent 5,257,143.