Probabilistic Bisection Search for Stochastic Root-Finding

Rolf Waeber Peter I. Frazier Shane G. Henderson

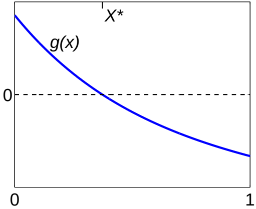
Operations Research & Information Engineering
Cornell University, Ithaca, NY

Sunday October 14, 2012

INFORMS Annual Meeting
Phoenix, AZ

This research is supported by AFOSR YIP FA9550-11-1-0083.

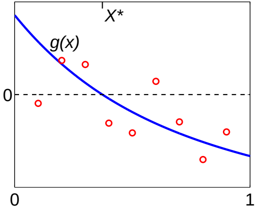
Stochastic Root-Finding Problem



- Consider a function $g : [0,1] \to \mathbb{R}$.
- Assumption: There exists a unique $X^* \in [0,1]$ such that
    - $g(x) > 0$ for $x < X^*$,
    - $g(x) < 0$ for $x > X^*$.

Goal: Find $X^* \in [0,1]$.

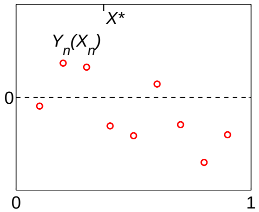
Stochastic Root-Finding Problem



- Consider a function $g : [0,1] \to \mathbb{R}$.
- Assumption: There exists a unique $X^* \in [0,1]$ such that
  - $g(x) > 0$ for $x < X^*$,
  - $g(x) < 0$ for $x > X^*$.

  Goal: Find $X^* \in [0,1]$.
- Can only observe $Y_n(X_n) = g(X_n) + \varepsilon_n(X_n)$, where $\varepsilon_n(X_n)$ is an independent noise with zero mean (median).

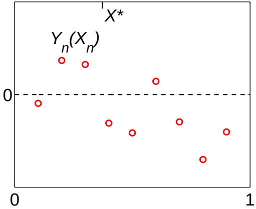
Stochastic Root-Finding Problem



- Consider a function $g : [0, 1] \to \mathbb{R}$.
- Assumption: There exists a unique $X^* \in [0, 1]$ such that
  - $g(x) > 0$ for $x < X^*$,
  - $g(x) < 0$ for $x > X^*$.

  Goal: Find $X^* \in [0, 1]$.
- Can only observe $Y_n(X_n) = g(X_n) + \varepsilon_n(X_n)$, where $\varepsilon_n(X_n)$ is an independent noise with zero mean (median).

Stochastic Root-Finding Problem



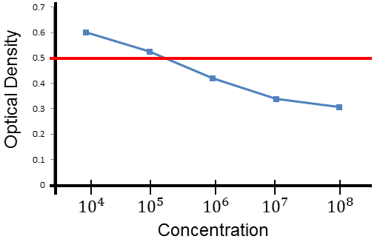
- Consider a function $g : [0, 1] \to \mathbb{R}$.
- Assumption: There exists a unique $X^* \in [0, 1]$ such that
  - $g(x) > 0$ for $x < X^*$,
  - $g(x) < 0$ for $x > X^*$.

  Goal: Find $X^* \in [0, 1]$.
- Can only observe $Y_n(X_n) = g(X_n) + \varepsilon_n(X_n)$, where $\varepsilon_n(X_n)$ is an independent noise with zero mean (median).

  Decisions:
  - Where to place samples X_n for $n = 0, 1, 2, \ldots$
  - How to estimate X^* after n iterations.

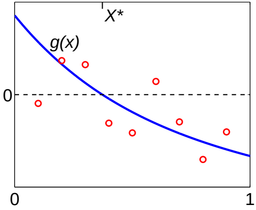
Applications

- Simulation optimization:
  - $g(x)$ as a gradient
- Sequential statistics:



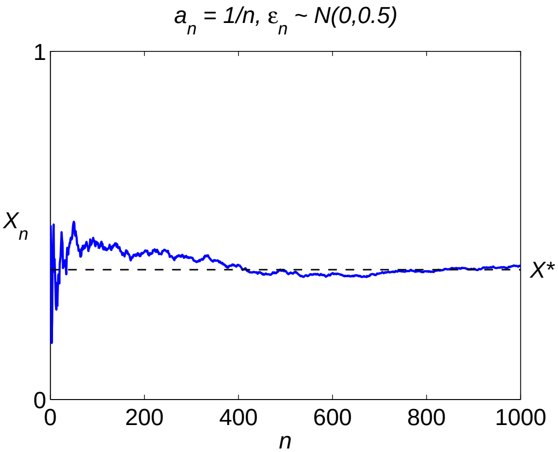
- Finance:
  - Pricing American options
  - Estimating risk measures

Stochastic Approximation (Robbins and Monro, 1951)

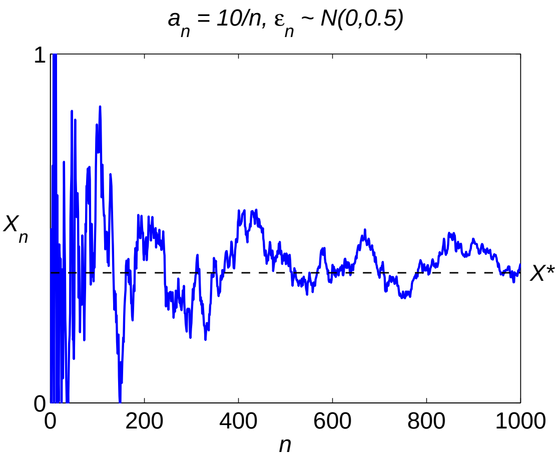


1. Choose an initial estimate $X_0 \in [0,1]$;
2. Determine a tuning sequence $(a_n)_n \geq 0$, $\sum_{n=0}^{\infty} a_n^2 < \infty$, and $\sum_{n=0}^{\infty} a_n = \infty$.
   (Example: $a_n = d/n$ for $d > 0$.)
3. $X_{n+1} = \Pi_{[0,1]}(X_n + a_n Y_n(X_n))$, where $\Pi_{[0,1]}$ is the projection to $[0,1]$.

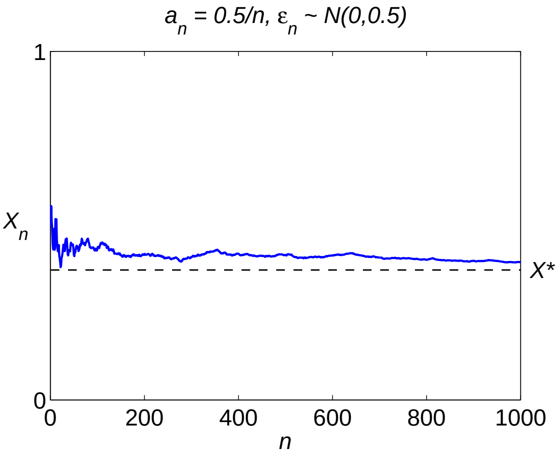
A "Good" Tuning Sequence



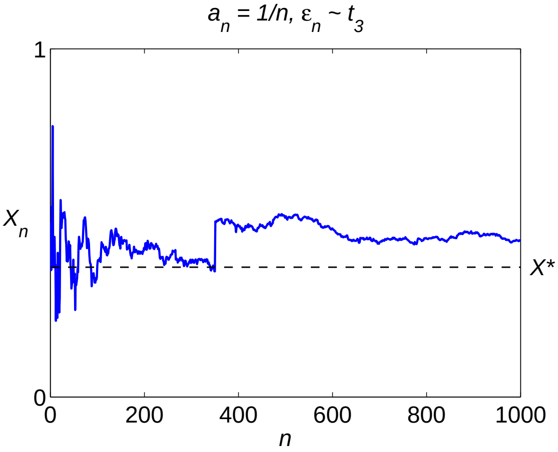
A "Bad" Tuning Sequence – too Large



A "Bad" Tuning Sequence – too Small

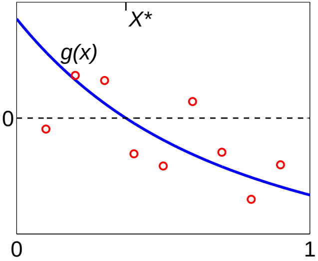


Lack of Robustness – ε_n with Heavy-Tails



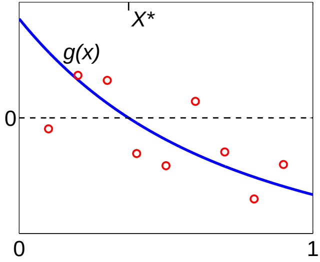
A Different Approach

What about a bisection algorithm?



A Different Approach

What about a bisection algorithm?

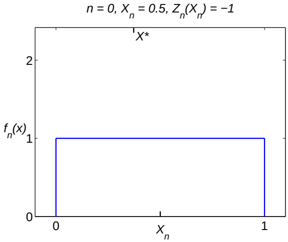


- Deterministic bisection algorithm will fail almost surely.
- Need to account for the noise.

The Probabilistic Bisection Algorithm

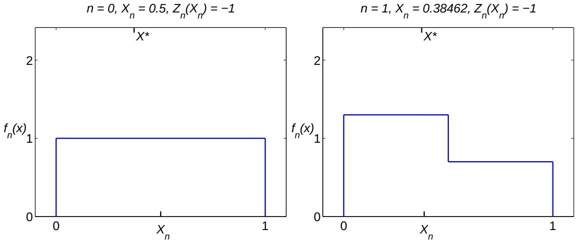
The Probabilistic Bisection Algorithm (Horstein, 1963)

- Input: $Z_n(X_n) := \text{sign}(Y_n(X_n))$.
- Assume a prior density f_0 on $[0, 1]$.



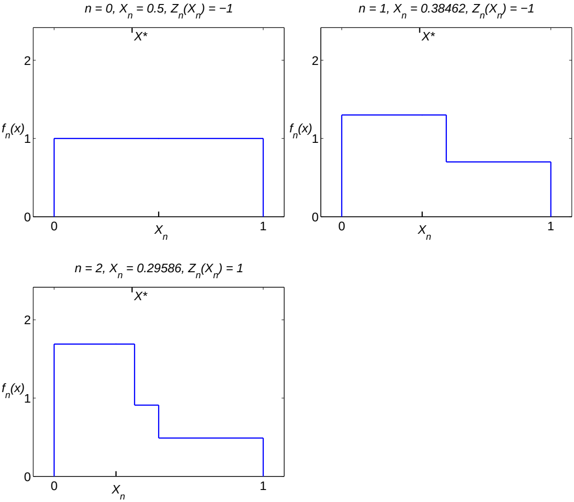
The Probabilistic Bisection Algorithm (Horstein, 1963)

- Input: $Z_n(X_n) := \text{sign}(Y_n(X_n))$.
- Assume a prior density f_0 on $[0, 1]$.



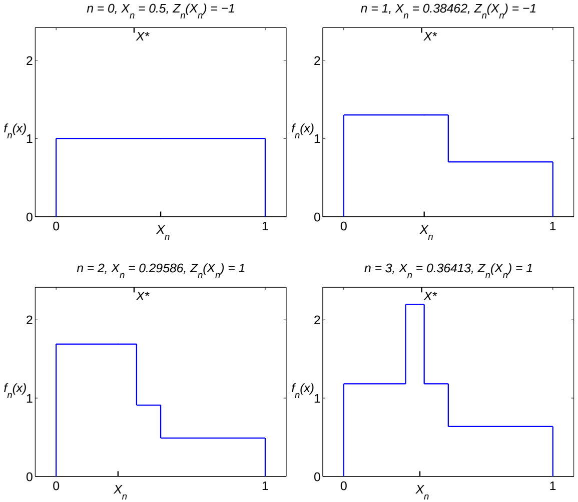
The Probabilistic Bisection Algorithm (Horstein, 1963)

- Input: $Z_n(X_n) := \text{sign}(Y_n(X_n))$.
- Assume a prior density f_0 on $[0, 1]$.

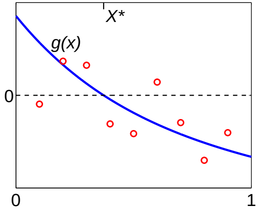


The Probabilistic Bisection Algorithm (Horstein, 1963)

- Input: $Z_n(X_n) := \text{sign}(Y_n(X_n))$.
- Assume a prior density f_0 on $[0, 1]$.

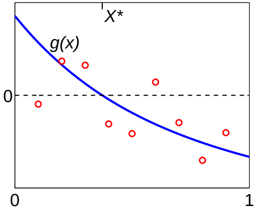


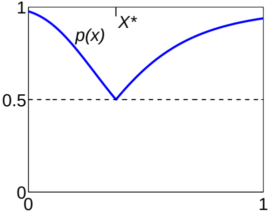
Stochastic Root-Finding Revisited



$$Z_n(X_n) = \begin{cases} \text{sign}\left(g(X_n)\right) & \text{with probability } p(X_n), \\ -\text{sign}\left(g(X_n)\right) & \text{with probability } 1 - p(X_n). \end{cases}$$

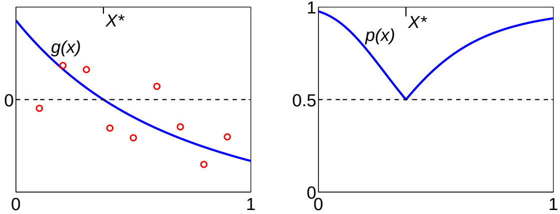
Stochastic Root-Finding Revisited





$$Z_n(X_n) = \begin{cases} \text{sign}\,(g(X_n)) & \text{with probability } p(X_n), \\ -\text{sign}\,(g(X_n)) & \text{with probability } 1 - p(X_n). \end{cases}$$

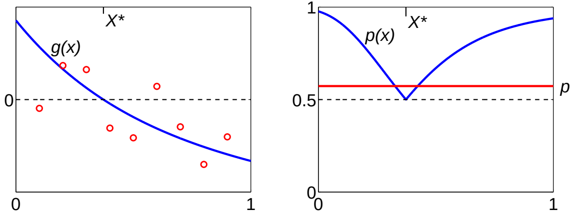
Stochastic Root-Finding Revisited



$$Z_n(X_n) = \begin{cases} \text{sign}\,(g(X_n)) & \text{with probability } p(X_n), \\ -\text{sign}\,(g(X_n)) & \text{with probability } 1 - p(X_n). \end{cases}$$

- The probability of a correct sign $p(\cdot)$ depends on $g(\cdot)$ and the noise $(\varepsilon_n)_n$.

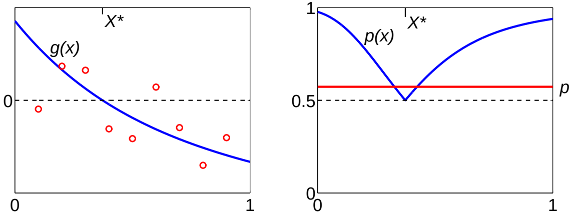
Stochastic Root-Finding Revisited



$$Z_n(X_n) = \begin{cases} \text{sign}\left(g(X_n)\right) & \text{with probability } p(X_n), \\ -\text{sign}\left(g(X_n)\right) & \text{with probability } 1 - p(X_n). \end{cases}$$

- The probability of a correct sign $p(\cdot)$ depends on $g(\cdot)$ and the noise $(\varepsilon_n)_n$.
- Stylized Setting:
  - $p(\cdot)$ is constant.

Stochastic Root-Finding Revisited



$$Z_n(X_n) = \begin{cases} \text{sign}\left(g(X_n)\right) & \text{with probability } p(X_n), \\ -\text{sign}\left(g(X_n)\right) & \text{with probability } 1 - p(X_n). \end{cases}$$

- The probability of a correct sign $p(\cdot)$ depends on $g(\cdot)$ and the noise $(\varepsilon_n)_n$.
- Stylized Setting:
  - $p(\cdot)$ is constant.
  - $p(\cdot)$ is known.

Stylized Setting

- $g(x)$ is a step function with a jump at X^*, for example, in edge detection applications (Castro and Nowak, 2008).

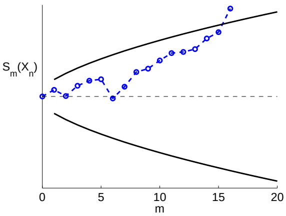
Stylized Setting

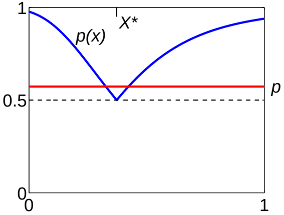
- $g(x)$ is a step function with a jump at X^*, for example, in edge detection applications (Castro and Nowak, 2008).
- Sample sequentially at point X_n and use $S_m(X_n) = \sum_{i=1}^m Y_{n,i}(X_n)$ to construct an α-*level test of power 1* (Siegmund, 1985):

$$N_n = \inf\left\{ m : |S_m| \geq [(m+1)(\log(m+1) + 2\log(1/\alpha))]^{1/2} \right\}.$$

Then $\mathbb{P}_{X_n = X^*}\{N_n < \infty\} \leq \alpha$, $\mathbb{P}_{X_n \neq X^*}\{N_n < \infty\} = 1$, and

$$\mathbb{P}_{X_n < X^*}\{S_{N_n}(X_n) > 0\} \geq 1 - \alpha/2 = p_c,$$
$$\mathbb{P}_{X_n > X^*}\{S_{N_n}(X_n) < 0\} \geq 1 - \alpha/2 = p_c.$$





The Probabilistic Bisection Algorithm (Horstein, 1963)

Notation: $p(\cdot) = p_c \in (1/2, 1]$ and $q_c = 1 - p_c$.

The Probabilistic Bisection Algorithm (Horstein, 1963)

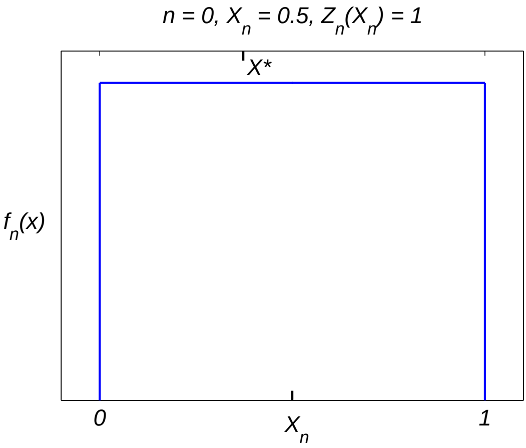
Notation: $p(\cdot) = p_c \in (1/2, 1]$ and $q_c = 1 - p_c$.

1. Place a prior density f_0 on the root X^*, f_0 has domain $[0, 1]$.
   Example: $U(0, 1)$.
2. For n=0,1,2, ...
   (a) Measure at the median $X_n := F_n^{-1}(1/2)$.
   (b) Update the posterior density:

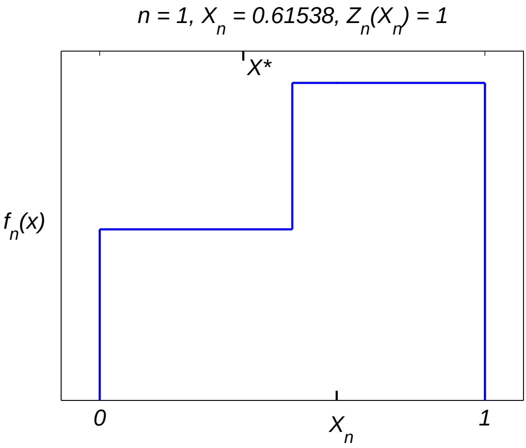
$$\text{if } Z_n(X_n) = +1, \qquad f_{n+1}(x) = \begin{cases} 2p_c \cdot f_n(x), & \text{if } x > X_n, \\ 2q_c \cdot f_n(x), & \text{if } x \leq X_n, \end{cases}$$

$$\text{if } Z_n(X_n) = -1, \qquad f_{n+1}(x) = \begin{cases} 2q_c \cdot f_n(x), & \text{if } x > X_n, \\ 2p_c \cdot f_n(x), & \text{if } x \leq X_n. \end{cases}$$

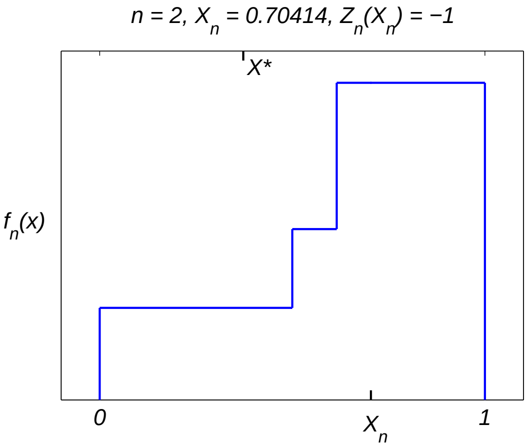
Sample Path of Posterior Distributions



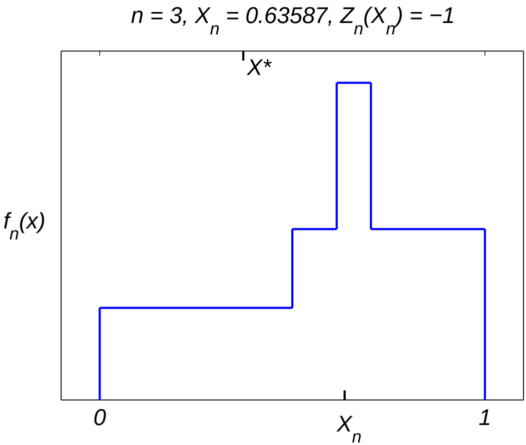
Sample Path of Posterior Distributions



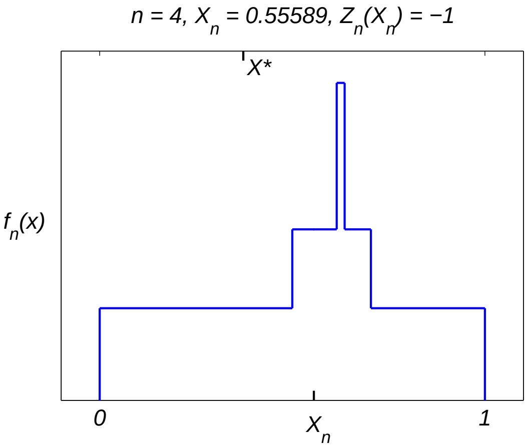
Sample Path of Posterior Distributions



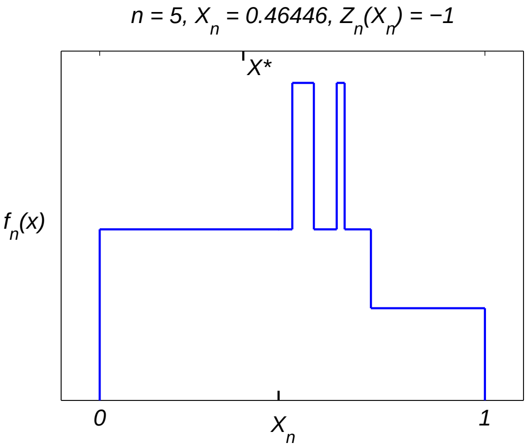
Sample Path of Posterior Distributions



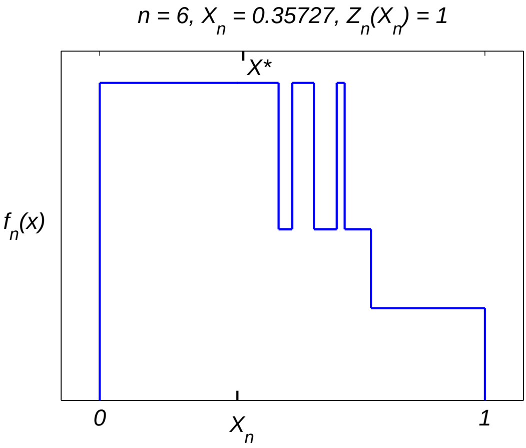
Sample Path of Posterior Distributions



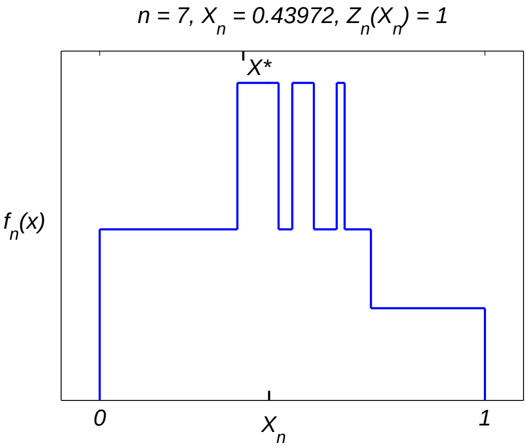
Sample Path of Posterior Distributions



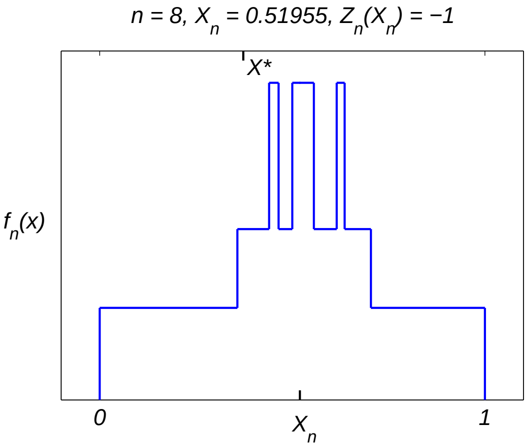
Sample Path of Posterior Distributions



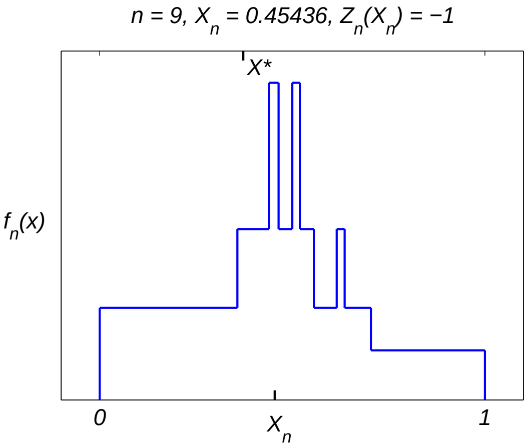
Sample Path of Posterior Distributions



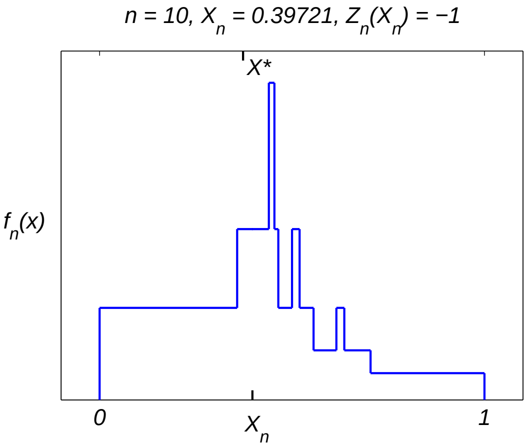
Sample Path of Posterior Distributions



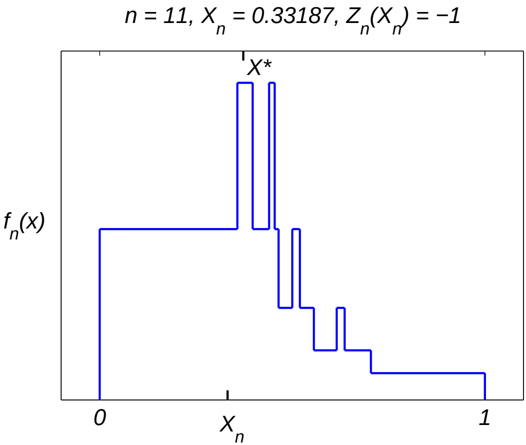
Sample Path of Posterior Distributions



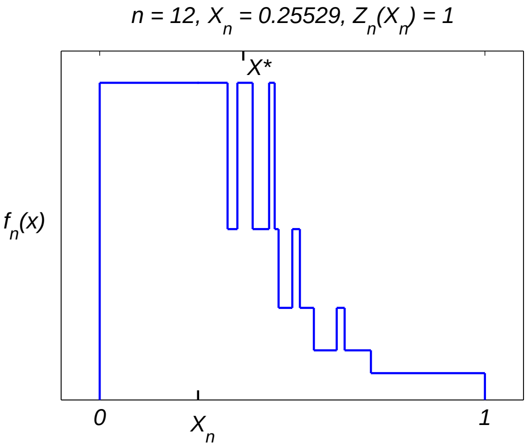
Sample Path of Posterior Distributions



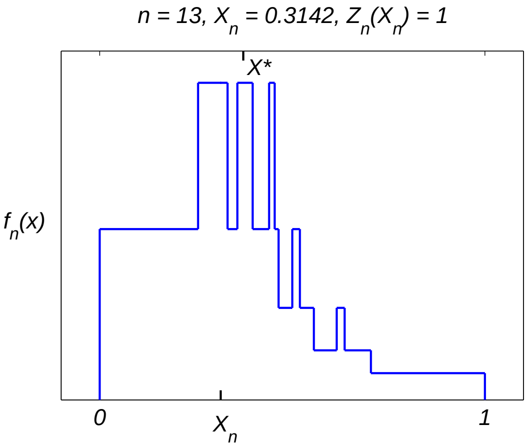
Sample Path of Posterior Distributions



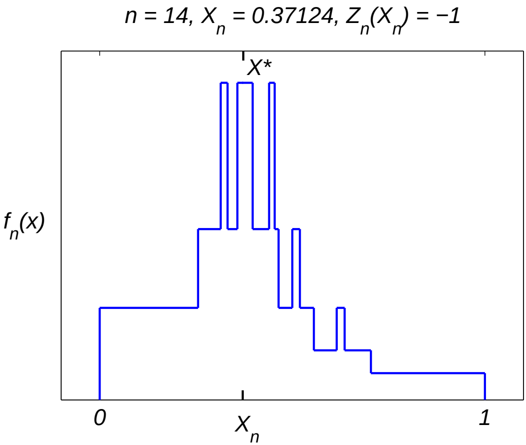
Sample Path of Posterior Distributions



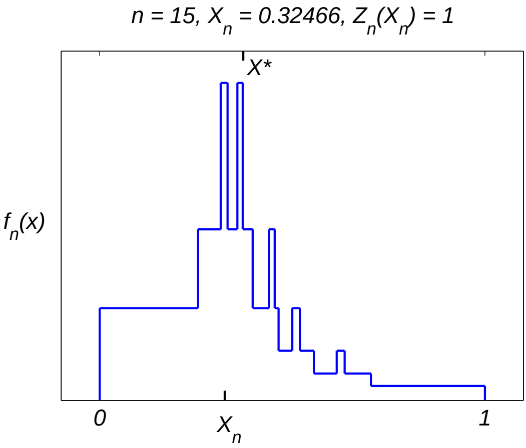
Sample Path of Posterior Distributions



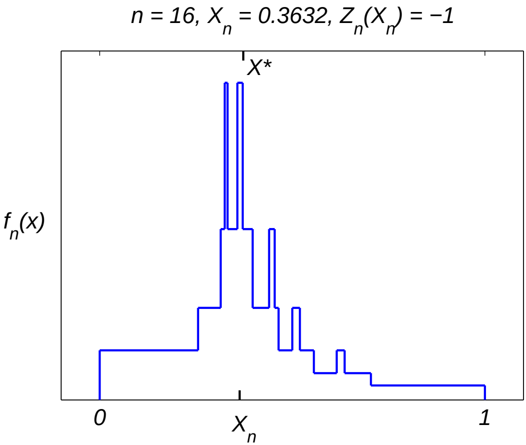
Sample Path of Posterior Distributions



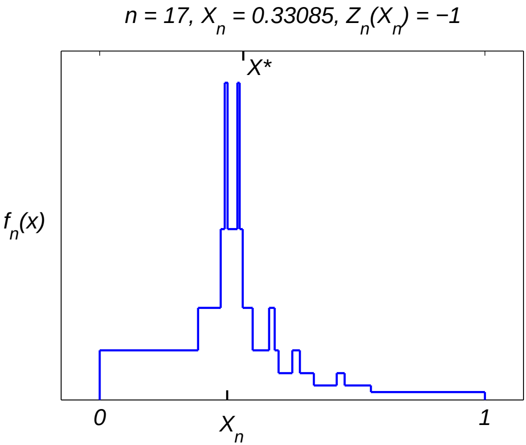
Sample Path of Posterior Distributions



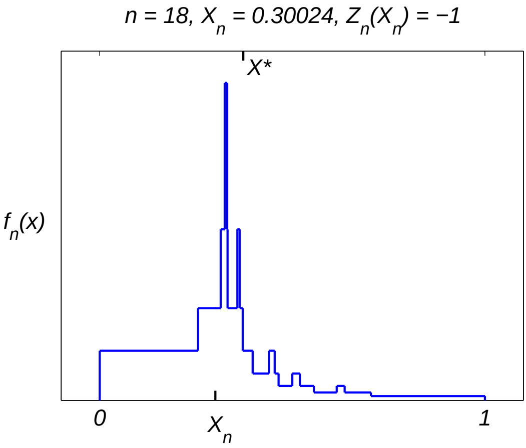
Sample Path of Posterior Distributions



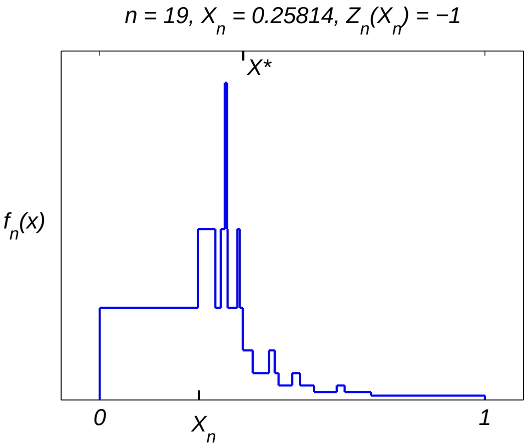
Sample Path of Posterior Distributions



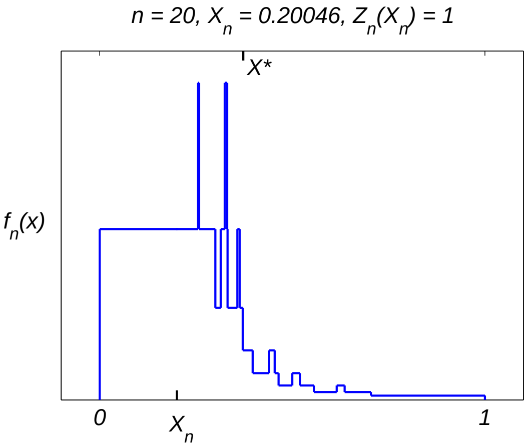
Sample Path of Posterior Distributions



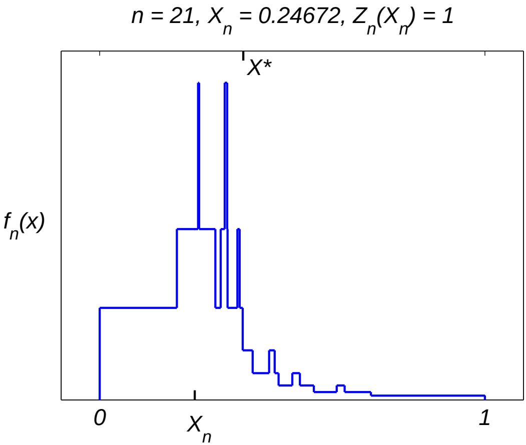
Sample Path of Posterior Distributions



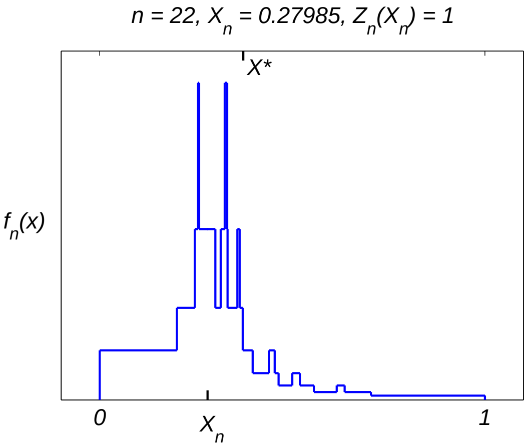
Sample Path of Posterior Distributions



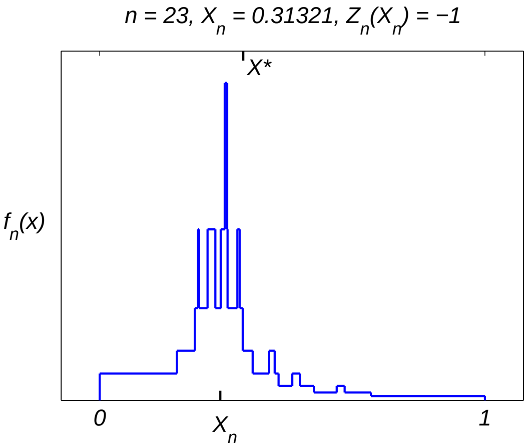
Sample Path of Posterior Distributions



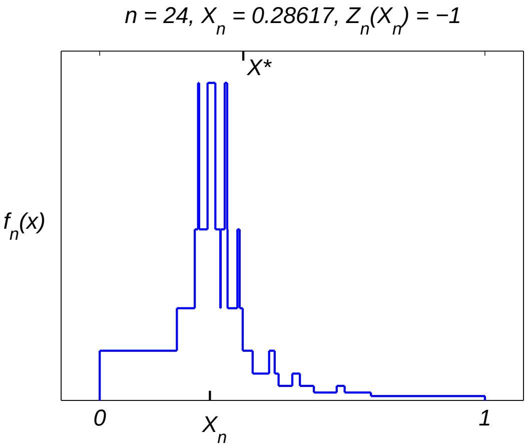
Sample Path of Posterior Distributions



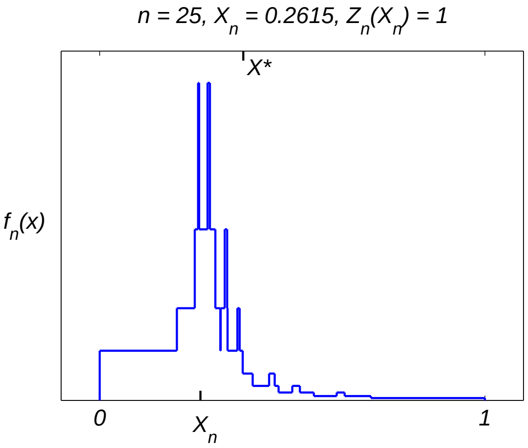
Sample Path of Posterior Distributions



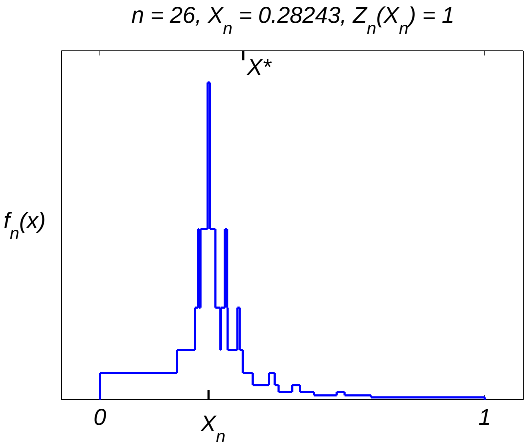
Sample Path of Posterior Distributions



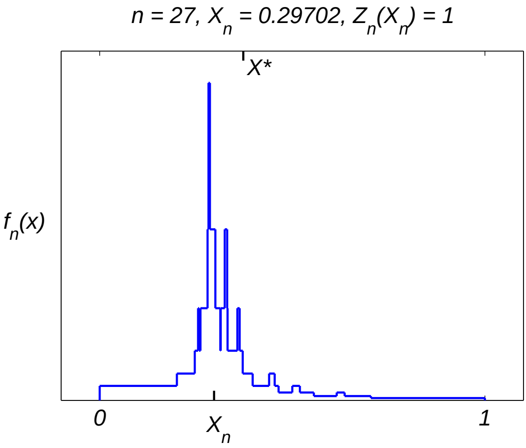
Sample Path of Posterior Distributions



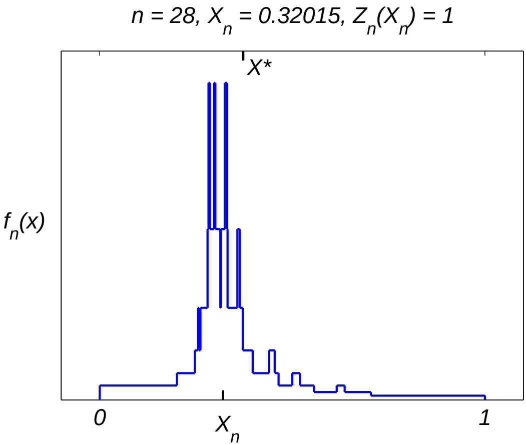
Sample Path of Posterior Distributions



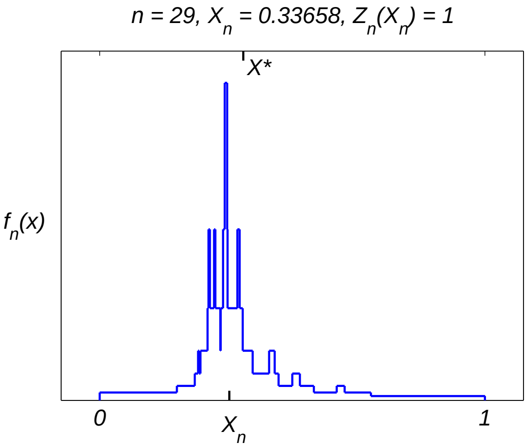
Sample Path of Posterior Distributions



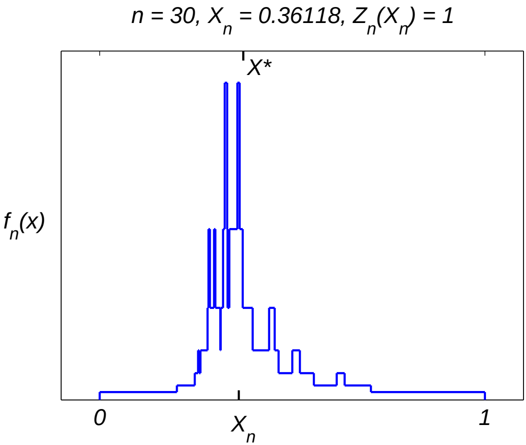
Sample Path of Posterior Distributions



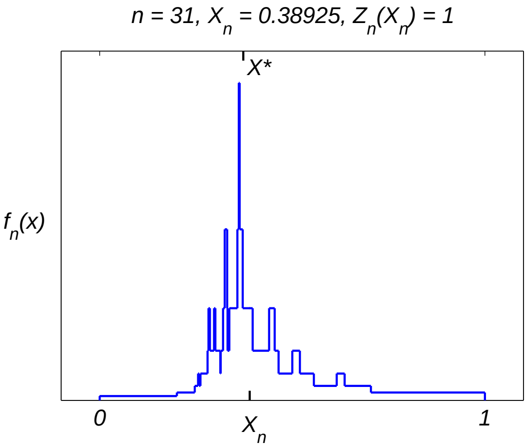
Sample Path of Posterior Distributions



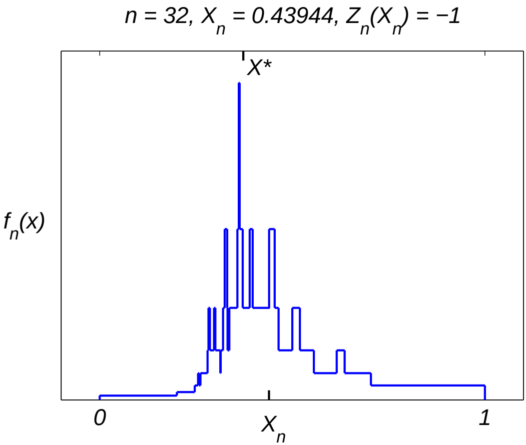
Sample Path of Posterior Distributions



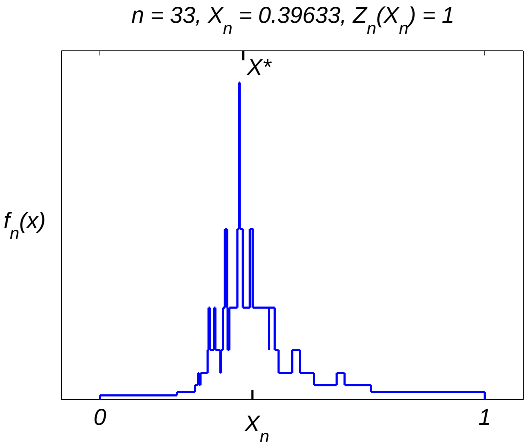
Sample Path of Posterior Distributions



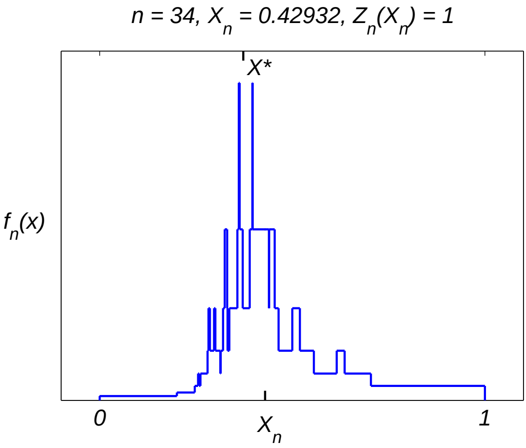
Sample Path of Posterior Distributions



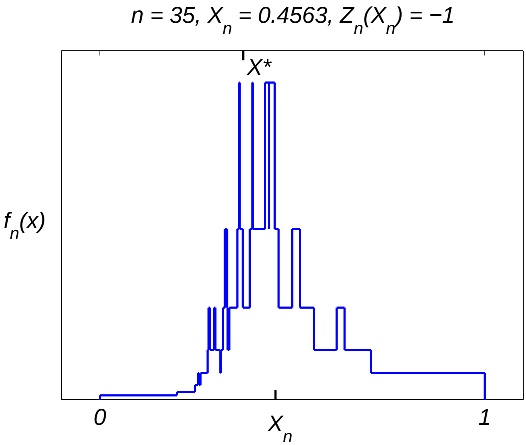
Sample Path of Posterior Distributions



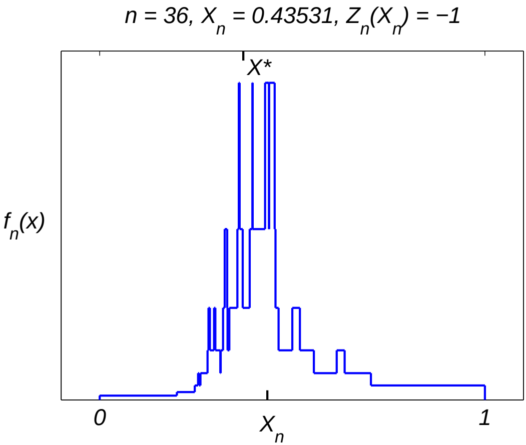
Sample Path of Posterior Distributions



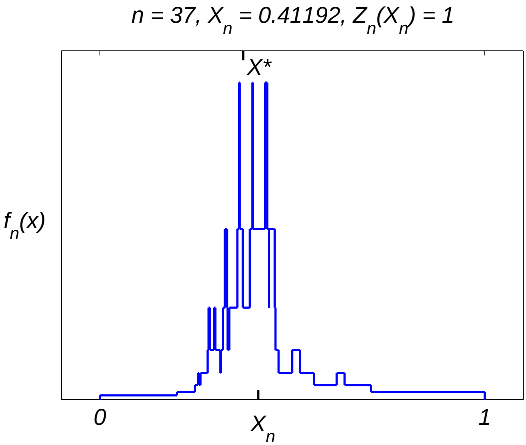
Sample Path of Posterior Distributions



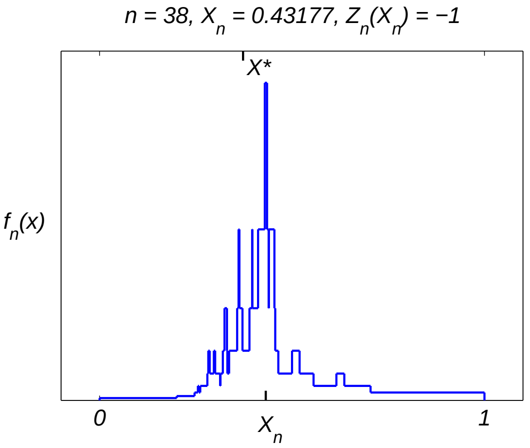
Sample Path of Posterior Distributions



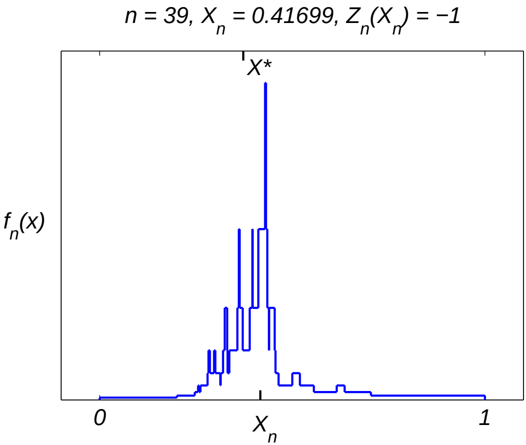
Sample Path of Posterior Distributions



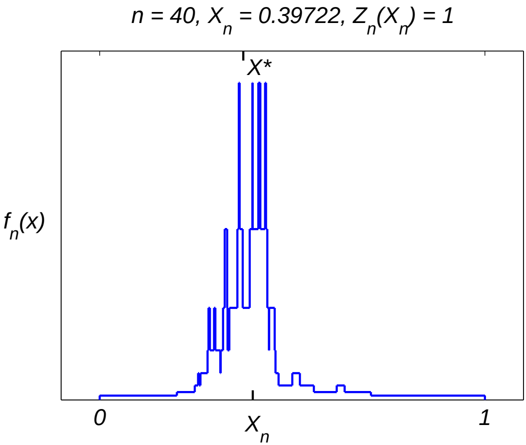
Sample Path of Posterior Distributions



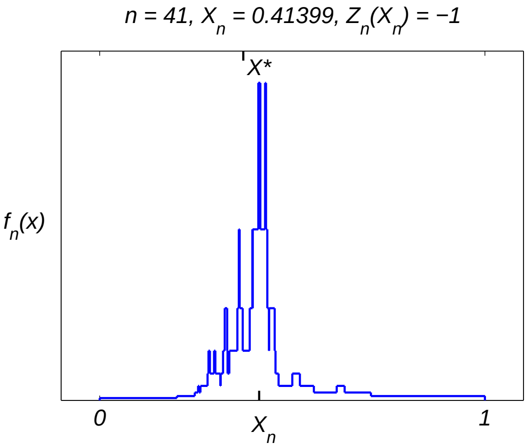
Sample Path of Posterior Distributions



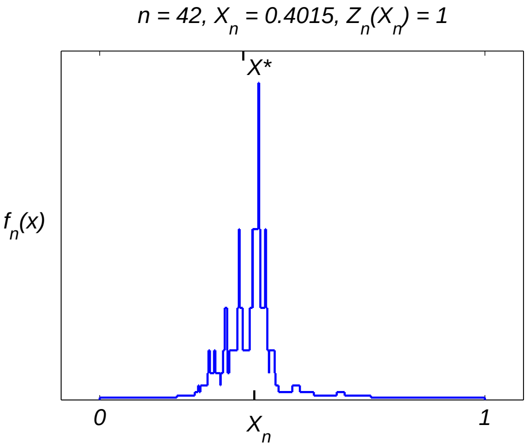
Sample Path of Posterior Distributions



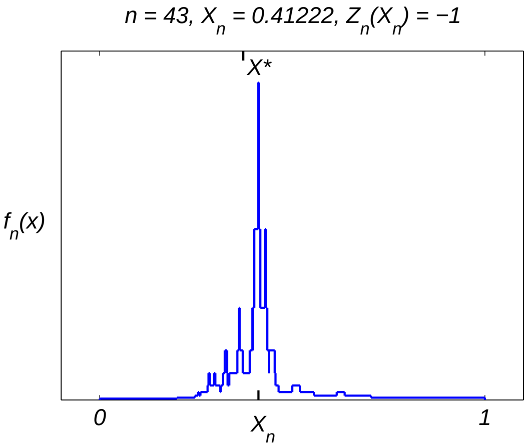
Sample Path of Posterior Distributions



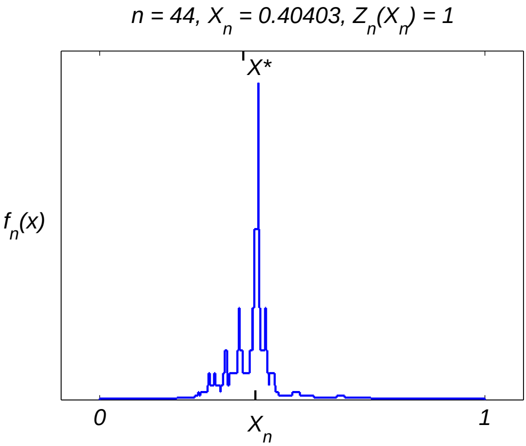
Sample Path of Posterior Distributions



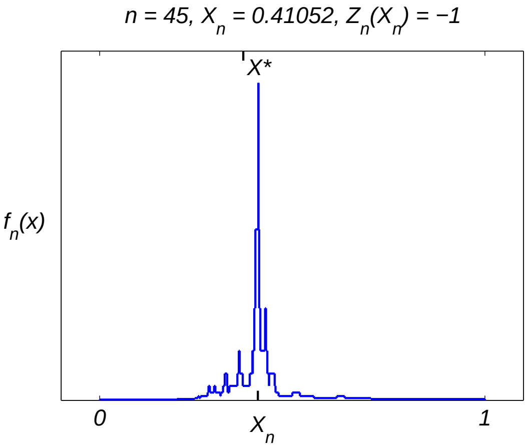
Sample Path of Posterior Distributions



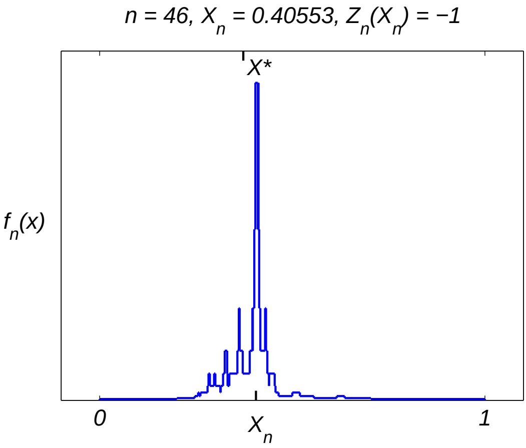
Sample Path of Posterior Distributions



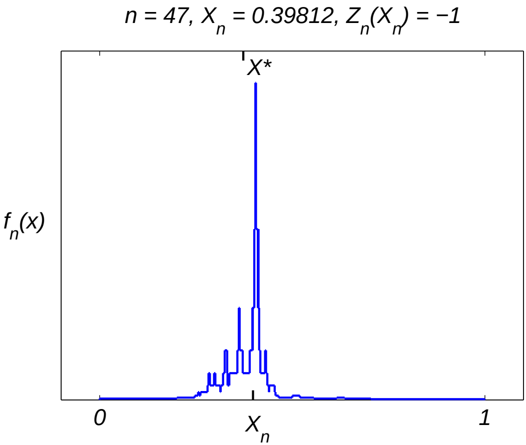
Sample Path of Posterior Distributions



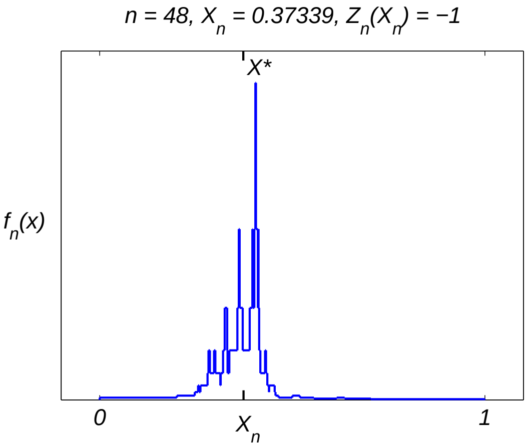
Sample Path of Posterior Distributions



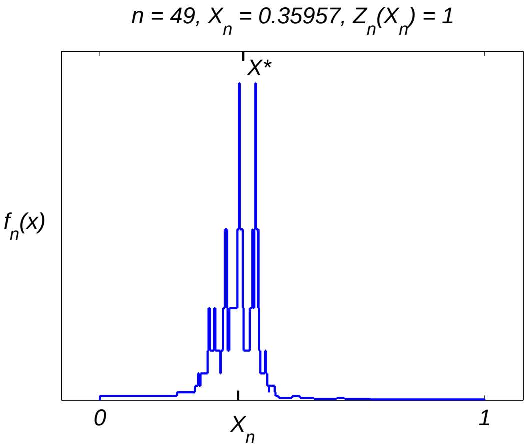
Sample Path of Posterior Distributions



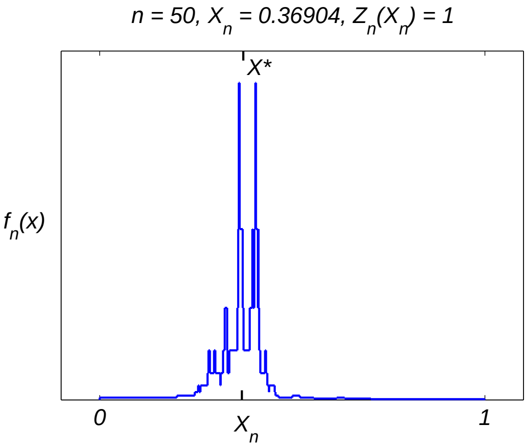
Sample Path of Posterior Distributions



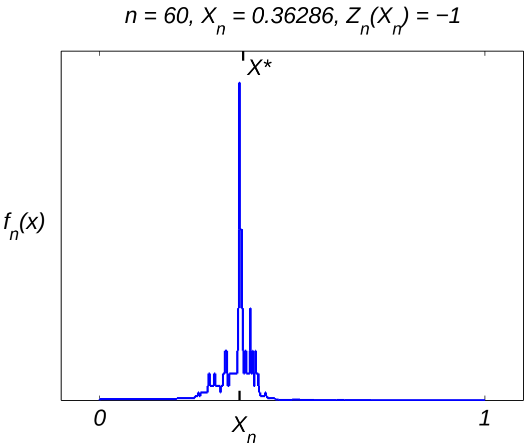
Sample Path of Posterior Distributions



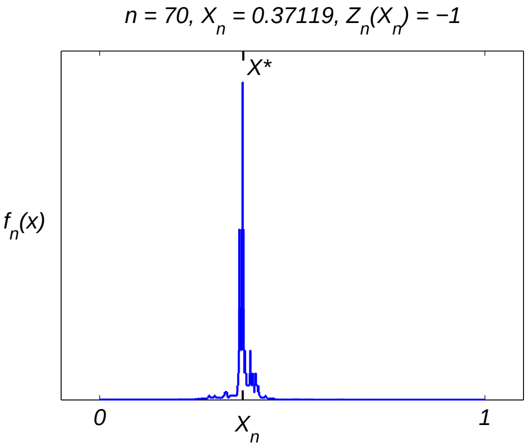
Sample Path of Posterior Distributions



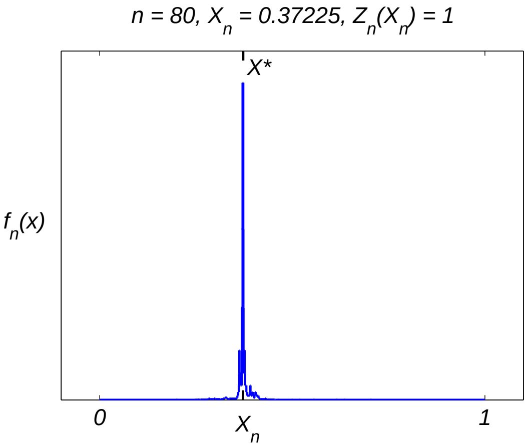
Sample Path of Posterior Distributions



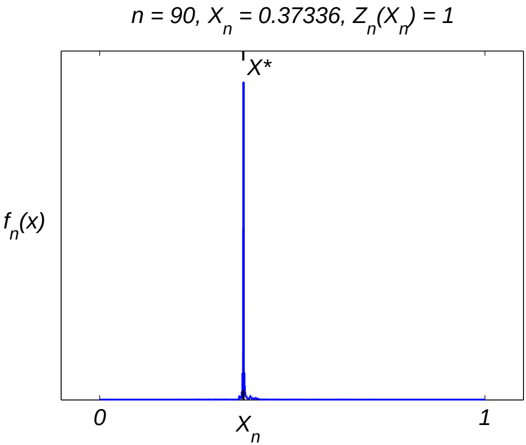
Sample Path of Posterior Distributions



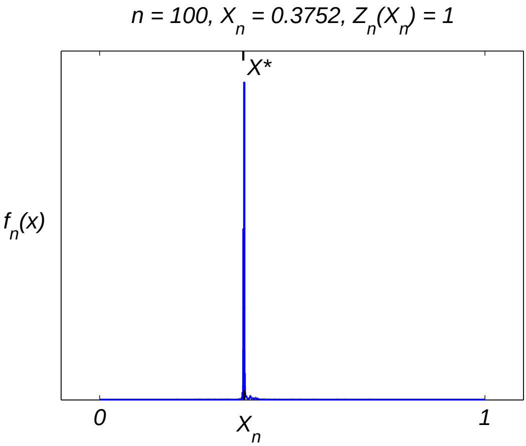
Sample Path of Posterior Distributions



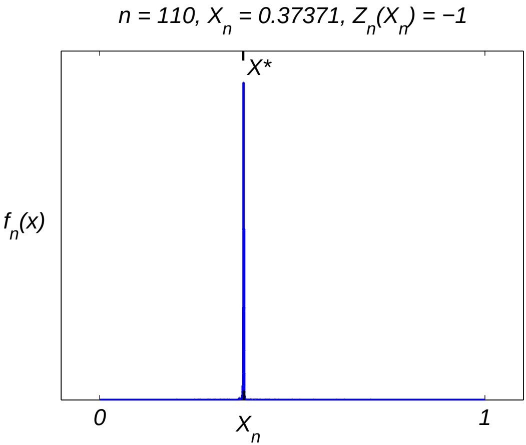
Sample Path of Posterior Distributions



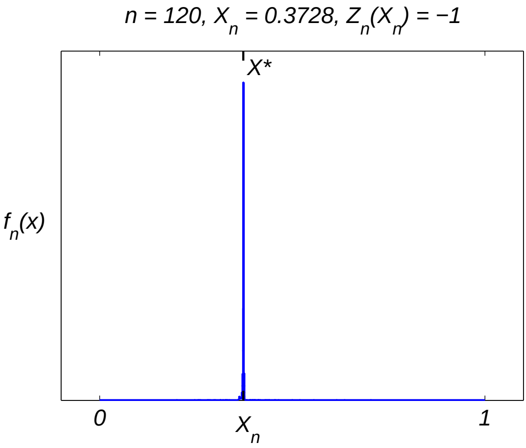
Sample Path of Posterior Distributions



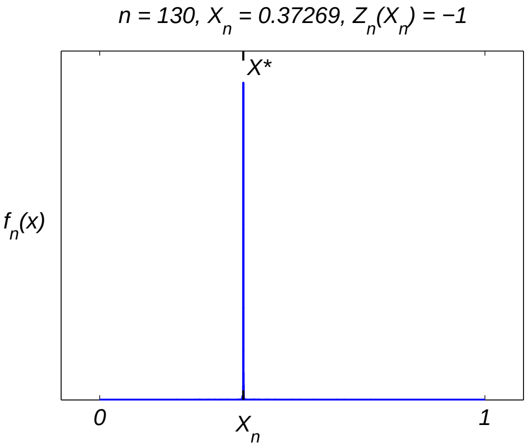
Sample Path of Posterior Distributions



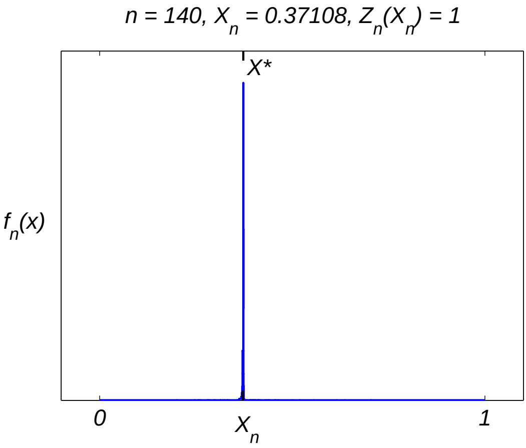
Sample Path of Posterior Distributions



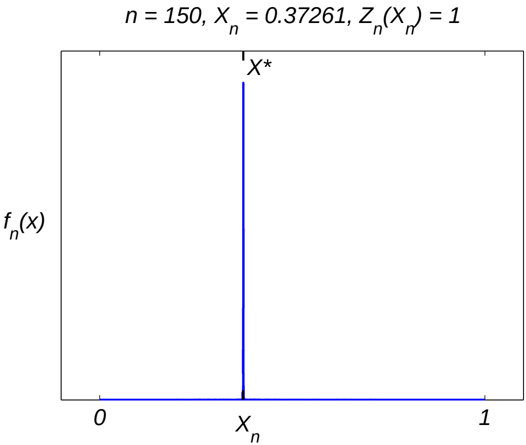
Sample Path of Posterior Distributions



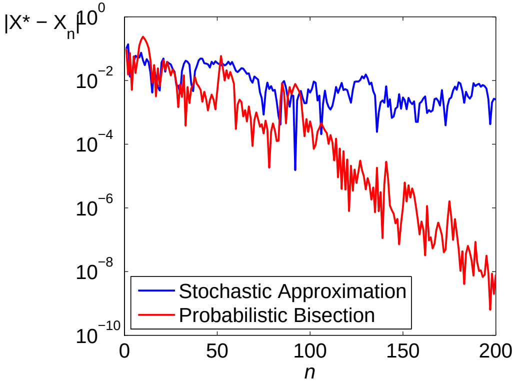
Sample Path of Posterior Distributions

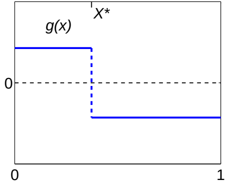


Sample Path of Posterior Distributions



Comparison to Stochastic Approximation





Literature Review: Probabilistic Bisection Algorithm

- First introduced in Horstein (1963).
- Discretized version: Burnashev and Zigangirov (1974).
- Feige et al. (1997), Karp and Kleinberg (2007), Ben-Or and Hassidim (2008), Nowak (2008), Nowak (2009), ...
- Survey paper: Castro and Nowak (2008)

Literature Review: Probabilistic Bisection Algorithm

- First introduced in Horstein (1963).
- Discretized version: Burnashev and Zigangirov (1974).
- Feige et al. (1997), Karp and Kleinberg (2007), Ben-Or and Hassidim (2008), Nowak (2008), Nowak (2009), ...
- Survey paper: Castro and Nowak (2008)

> *"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."*

Algorithm Analysis

Consistency

Setting for probabilistic bisection with power 1 tests:

- $X^* \in [0, 1]$ fixed and unknown.
- $X_n \neq X^*$ for any finite $n \in \mathbb{N}$.
- $p(X_n) \geq p_c$ for all $n \in \mathbb{N}$.
- $p_c \in (1/2, 1)$ is an input parameter.

Consistency

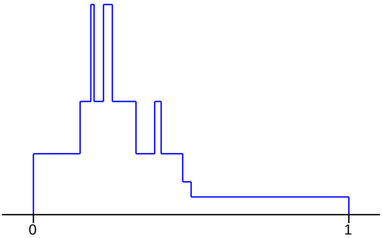
Setting for probabilistic bisection with power 1 tests:

- $X^* \in [0,1]$ fixed and unknown.
- $X_n \neq X^*$ for any finite $n \in \mathbb{N}$.
- $p(X_n) \geq p_c$ for all $n \in \mathbb{N}$.
- $p_c \in (1/2, 1)$ is an input parameter.

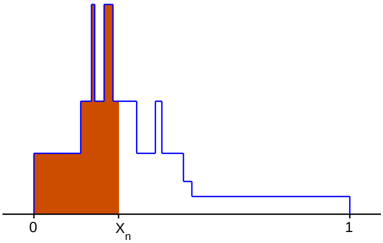
Theorem

$X_n \to X^*$ almost surely as $n \to \infty$.

Analysis of Posterior Density



Analysis of Posterior Density



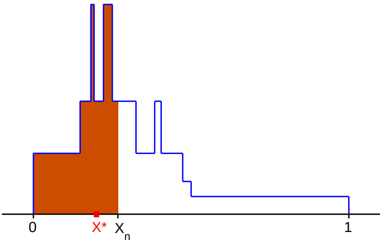
- If $Z_n = +1$:

$$f_{n+1}(x) = 2q_c \cdot f_n(x), \quad x < X_n,$$
$$f_{n+1}(x) = 2p_c \cdot f_n(x), \quad x \geq X_n,$$

- If $Z_n = -1$:

$$f_{n+1}(x) = 2p_c \cdot f_n(x), \quad x < X_n,$$
$$f_{n+1}(x) = 2q_c \cdot f_n(x), \quad x \geq X_n.$$

Analysis of Posterior Density



Case I: If $X^* < X_n : \mathbb{P}(Z_n = +1) = p(X_n) \geq p_c$

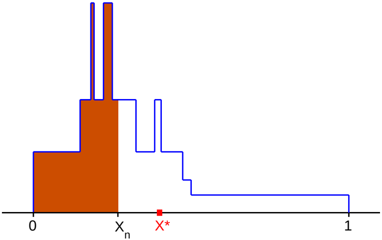
- If $Z_n = +1$:

$$f_{n+1}(x) = 2q_c \cdot f_n(x), \quad x < X_n,$$
$$f_{n+1}(x) = 2p_c \cdot f_n(x), \quad x \geq X_n,$$

- If $Z_n = -1$:

$$f_{n+1}(x) = 2p_c \cdot f_n(x), \quad x < X_n,$$
$$f_{n+1}(x) = 2q_c \cdot f_n(x), \quad x \geq X_n.$$

Analysis of Posterior Density



Case II: If $X^* > X_n : \mathbb{P}(Z_n = +1) = 1 - p(X_n) \leq p_c$

- If $Z_n = +1$:

$$f_{n+1}(x) = 2q_c \cdot f_n(x), \quad x < X_n,$$
$$f_{n+1}(x) = 2p_c \cdot f_n(x), \quad x \geq X_n,$$

- If $Z_n = -1$:

$$f_{n+1}(x) = 2p_c \cdot f_n(x), \quad x < X_n,$$
$$f_{n+1}(x) = 2q_c \cdot f_n(x), \quad x \geq X_n.$$

Analysis of Posterior Density cont.

- The dynamics of $f_n(x)$ are very complicated for almost all $x \in [0,1]$.

Analysis of Posterior Density cont.

- The dynamics of $f_n(x)$ are very complicated for almost all $x \in [0,1]$. HOWEVER, the dynamics of $f_n(X^*)$ are rather simple:

$$f_{n+1}(X^*) = \begin{cases} 2p_c \cdot f_n(X^*), & \text{with probability } p(X_n) \geq p_c, \\ 2q_c \cdot f_n(X^*), & \text{with probability } q(X_n) \leq q_c. \end{cases}$$

Analysis of Posterior Density cont.

- The dynamics of $f_n(x)$ are very complicated for almost all $x \in [0,1]$. HOWEVER, the dynamics of $f_n(X^*)$ are rather simple:

$$f_{n+1}(X^*) = \begin{cases} 2p_c \cdot f_n(X^*), & \text{with probability } p(X_n) \geq p_c, \\ 2q_c \cdot f_n(X^*), & \text{with probability } q(X_n) \leq q_c. \end{cases}$$

- A sample path of $f_n(X^*)$ dominates a sample path of a coupled geometric random walk $(W_n)_n$ with dynamics

$$W_{n+1} = \begin{cases} 2p_c \cdot W_n, & \text{with probability } p_c, \\ 2q_c \cdot W_n, & \text{with probability } q_c. \end{cases}$$

Analysis of Posterior Density cont.

- The dynamics of $f_n(x)$ are very complicated for almost all $x \in [0,1]$. HOWEVER, the dynamics of $f_n(X^*)$ are rather simple:

$$f_{n+1}(X^*) = \begin{cases} 2p_c \cdot f_n(X^*), & \text{with probability } p(X_n) \geq p_c, \\ 2q_c \cdot f_n(X^*), & \text{with probability } q(X_n) \leq q_c. \end{cases}$$

- A sample path of $f_n(X^*)$ dominates a sample path of a coupled geometric random walk $(W_n)_n$ with dynamics

$$W_{n+1} = \begin{cases} 2p_c \cdot W_n, & \text{with probability } p_c, \\ 2q_c \cdot W_n, & \text{with probability } q_c. \end{cases}$$

- The process $f_n(X^*)$ behaves almost like a geometric random walk **independently of** $(X_n)_n$. The goal is then to locate this geometric random walk efficiently!

Confidence Intervals for X^*

- Notation: $\mu = p_c \ln 2p_c + q_c \ln 2q_c$.
- For $\alpha \in (0,1)$, define

$$b_n = n\mu - n^{1/2}(-0.5 \ln \alpha)^{1/2}(\ln 2p_c - \ln 2q_c).$$

- Define

$$J_n = conv(x \in [0,1] : f_n(x) \geq e^{b_n}).$$

Confidence Intervals for X^*

- Notation: $\mu = p_c \ln 2p_c + q_c \ln 2q_c$.
- For $\alpha \in (0,1)$, define

$$b_n = n\mu - n^{1/2}(-0.5 \ln \alpha)^{1/2}(\ln 2p_c - \ln 2q_c).$$

- Define

$$J_n = conv(x \in [0,1] : f_n(x) \geq e^{b_n}).$$

Theorem

For $\alpha \in (0,1)$,

$$\mathbb{P}(X^* \in J_n) \geq 1 - \alpha,$$

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

Confidence Intervals for X^*

- Notation: $\mu = p_c \ln 2p_c + q_c \ln 2q_c$.
- For $\alpha \in (0,1)$, define

$$b_n = n\mu - n^{1/2}(-0.5 \ln \alpha)^{1/2}(\ln 2p_c - \ln 2q_c).$$

- Define

$$J_n = conv(x \in [0,1] : f_n(x) \geq e^{b_n}).$$

Theorem

For $\alpha \in (0,1)$,

$$\mathbb{P}(X^* \in J_n) \geq 1 - \alpha,$$

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

Proof:

Application of Hoeffding's inequality.

□

Size of Confidence Interval

Theorem

Choose $p_c \geq 0.85$, $\alpha \in (0,1)$. *For* $0 < r < \mu - q_c \ln 2p_c$ *there exists a* $N(p_c, r, \alpha) \in \mathbb{N}$, *such that*

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

for all $n \geq N(p_c, r, \alpha)$.

Size of Confidence Interval

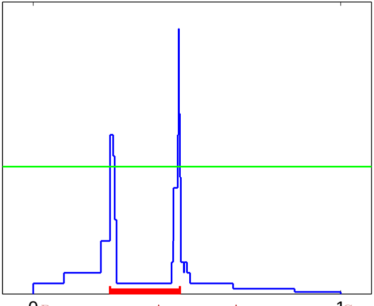
Theorem

Choose $p_c \geq 0.85$, $\alpha \in (0,1)$. *For* $0 < r < \mu - q_c \ln 2p_c$ *there exists a* $N(p_c, r, \alpha) \in \mathbb{N}$, *such that*

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

for all $n \geq N(p_c, r, \alpha)$.

Proof Idea:



Rate of Convergence

Theorem

Define $\hat{X}_n$ to be any point in J_n, then there exists $r > 0$ such that

$$\mathbb{E}[|X^* - \hat{X}_n|] = O(e^{-rn}).$$

Rate of Convergence

Theorem

Define $\hat{X}_n$ to be any point in J_n, then there exists $r > 0$ such that

$$\mathbb{E}[|X^* - \hat{X}_n|] = O(e^{-rn}).$$

- This is extremely fast compared to stochastic approximation:

$$O(e^{-rn}) \text{ vs. } O(n^{-1/2}).$$

Rate of Convergence

Theorem

Define $\hat{X}_n$ to be any point in J_n, then there exists $r > 0$ such that

$$\mathbb{E}[|X^* - \hat{X}_n|] = O(e^{-rn}).$$

- This is extremely fast compared to stochastic approximation:

$$O(e^{-rn}) \text{ vs. } O(n^{-1/2}).$$

- And we have true confidence intervals for X^*.

Rate of Convergence

Theorem

Define $\hat{X}_n$ to be any point in J_n, then there exists $r > 0$ such that

$$\mathbb{E}[|X^* - \hat{X}_n|] = O(e^{-rn}).$$

- This is extremely fast compared to stochastic approximation:

$$O(e^{-rn}) \text{ vs. } O(n^{-1/2}).$$

- And we have true confidence intervals for X^*.
- But n is the number of measurement points, what about total wall-clock time?

Wall-Clock Time

At each iteration of the Probabilistic Bisection Algorithm:

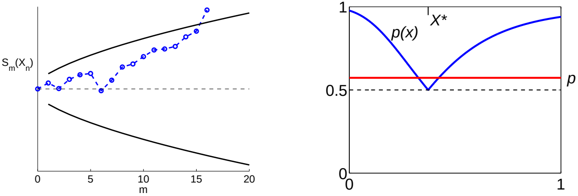
- Sample sequentially at point X_n and observe $S_m(X_n) = \sum_{i=1}^m Y_{n,i}(X_n)$, until

$$N_n = \inf\left\{m : |S_m| \geq [(m+1)(\log(m+1) + 2\log(1/\alpha))]^{1/2}\right\},$$

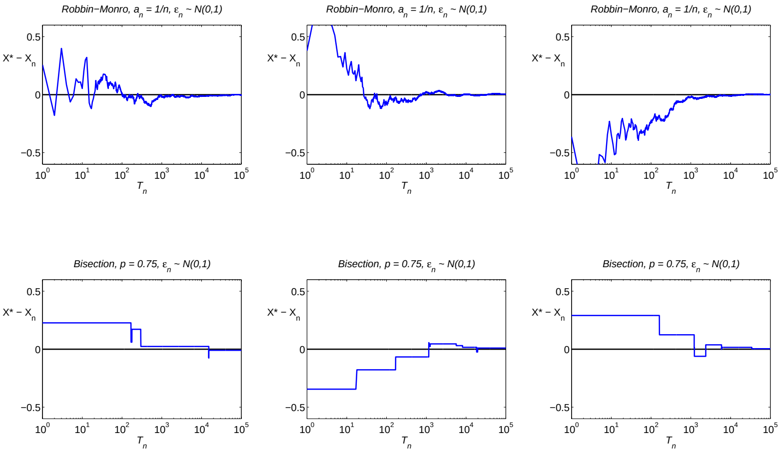
then $\mathbb{P}_{X_n = X^*}\{N_n < \infty\} \leq \alpha$, $\mathbb{P}_{X_n \neq X^*}\{N_n < \infty\} = 1$, and

$$\mathbb{P}_{X_n < X^*}\{S_{N_n}(X_n) > 0\} \geq 1 - \alpha/2 = p_c,$$
$$\mathbb{P}_{X_n > X^*}\{S_{N_n}(X_n) < 0\} \geq 1 - \alpha/2 = p_c.$$

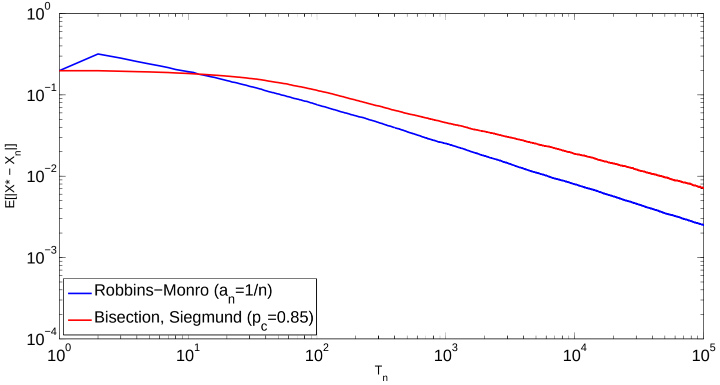
- Wall-clock time: $T_n = \sum_{i=1}^n N_n$.



Sample Paths



Numerical Comparison



Rate of Convergence in Wall-Clock Time?

- Farrell (1964):

$$\mathbb{E}_{g(x)}[N] \sim (1/g(x))^2 \log\log(1/|g(x)|) \text{ as } g(x) \to 0,$$

and for all tests of power one, if $\mathbb{P}_0(N = \infty) > 0$, then

$$\lim_{g(x)\to 0} g(x)^2 \mathbb{E}_{g(x)}[N] = \infty.$$

Rate of Convergence in Wall-Clock Time?

- Farrell (1964):

$$\mathbb{E}_{g(x)}[N] \sim (1/g(x))^2 \log\log(1/|g(x)|) \text{ as } g(x) \to 0,$$

and for all tests of power one, if $\mathbb{P}_0(N = \infty) > 0$, then

$$\lim_{g(x)\to 0} g(x)^2 \mathbb{E}_{g(x)}[N] = \infty.$$

Theorem

1. $\lim_{n\to\infty} \mathbb{E}[|X^* - X_n|(T_n)^{1/2}] = \infty$.
2. $(|X^* - X_n|(T_n)^{1/2})_n$ *is not tight.*

Rate of Convergence in Wall-Clock Time?

- Farrell (1964):

$$\mathbb{E}_{g(x)}[N] \sim (1/g(x))^2 \log\log(1/|g(x)|) \text{ as } g(x) \to 0,$$

and for all tests of power one, if $\mathbb{P}_0(N = \infty) > 0$, then

$$\lim_{g(x)\to 0} g(x)^2 \mathbb{E}_{g(x)}[N] = \infty.$$

Theorem

1. $\lim_{n\to\infty} \mathbb{E}[|X^* - X_n|(T_n)^{1/2}] = \infty$.
2. $(|X^* - X_n|(T_n)^{1/2})_n$ *is not tight.*

- If

$$g(x) \to 0 \text{ as } x \to X^*,$$

and if we use X_n as the best estimate of X^* then the Probabilistic Bisection Algorithm with power one tests is **asymptotically slower** than Stochastic Approximation.

Conjecture

- X_n might not be the best estimate for X^* when we use power one tests.
- Intuitively, observations where we spend more time should also be closer to X^*, hence an estimator of the form

$$\tilde{X}_n = \frac{1}{T_n} \sum_{i=1}^{n} N_i X_i$$

should perform better.

Conjecture

- X_n might not be the best estimate for X^* when we use power one tests.
- Intuitively, observations where we spend more time should also be closer to X^*, hence an estimator of the form

$$\tilde{X}_n = \frac{1}{T_n} \sum_{i=1}^{n} N_i X_i$$

  should perform better.
- **Conjecture:** For any $\beta > 0$ it holds that

$$\mathbb{E}[|\tilde{X}_n - X^*|] = O(T_n^{-\frac{1}{2(1+\beta)}}),$$

  (if g satisfies some growth conditions).

Conjecture

- X_n might not be the best estimate for X^* when we use power one tests.
- Intuitively, observations where we spend more time should also be closer to X^*, hence an estimator of the form

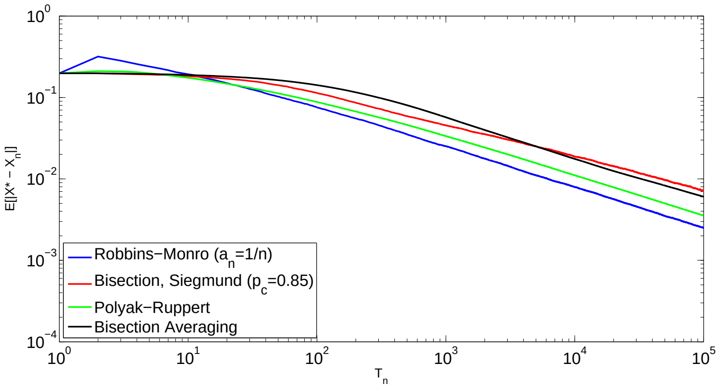
$$\tilde{X}_n = \frac{1}{T_n} \sum_{i=1}^{n} N_i X_i$$

  should perform better.
- **Conjecture:** For any $\beta > 0$ it holds that

$$\mathbb{E}[|\tilde{X}_n - X^*|] = O(T_n^{-\frac{1}{2(1+\beta)}}),$$

  (if g satisfies some growth conditions).
- **Sufficient Condition:** $|X_n - X^*| = O(e^{-rn})$ for some $r > 0$.

Numerical Comparison Cont.



Conclusions

Conclusions

Positive:

- Provides true confidence interval of the root X^*.
- Works extremely well if there is a jump at $g(X^*)$ (geometric rate of convergence).
- Only one tuning parameter.

Drawbacks:

- Seems to be asymptotically slower than Stochastic Approximation (but by not much).
- Higher computational cost.

Future Research:

- Robustness of algorithm.
- Use parallel computing (very little switching of $(X_n)_n$).
- Extension to higher dimensions.

THANK YOU!

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

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

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

R. H. Farrell (1964): Asymptotic behavior of expected sample size in certain one sided tests. *The Annals of Mathematical Statistics* 35:36–72.

U. Feige, P. Raghavan, D. Peleg and E. Upfal (1997): Computing with noisy information. *SIAM Journal of Computing* :1001–1018.

M. Horstein (1963): Sequential transmission using noiseless feedback. *IEEE Transactions on Information Theory* 9:136–143.

R. Karp and R. Kleinberg (2007): Noisy binary search and its applications. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, pp. 881–890.

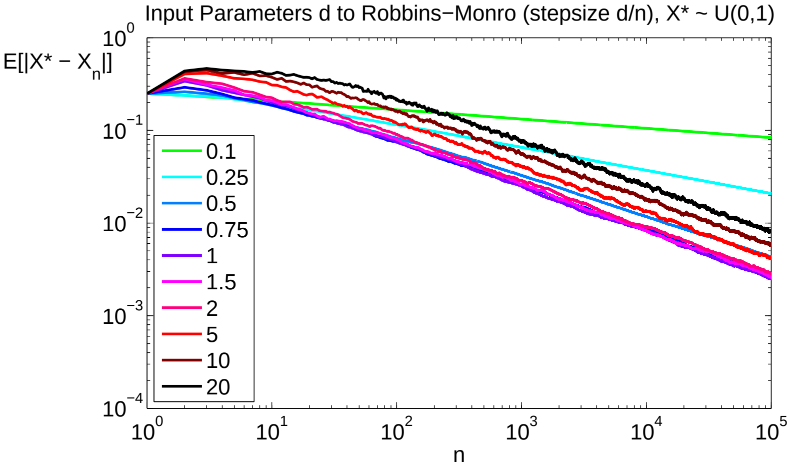
R. Nowak (2008): Generalized binary search. In *46th Annual Allerton Conference on Communication, Control, and Computing*. pp. 568–574.

R. Nowak (2009): Noisy generalized binary search. In *Advances in neural information processing systems*, volume 22. pp. 1366–1374.

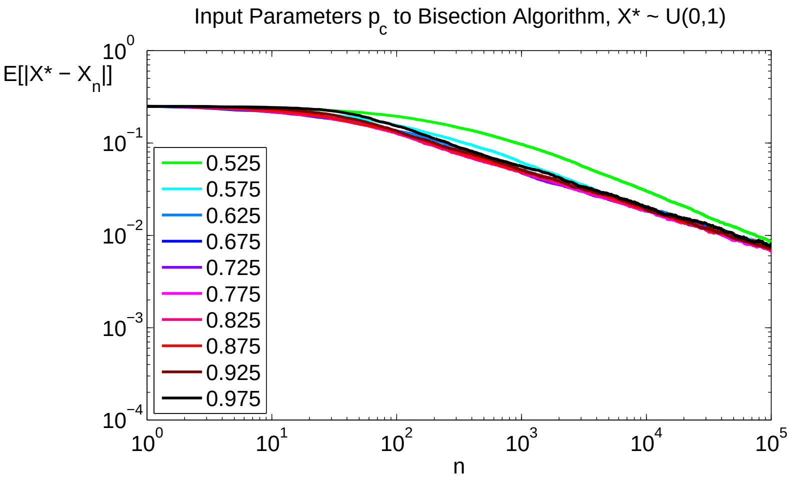
H. Robbins and S. Monro (1951): A stochastic approximation method. *The Annals of Mathematical Statistics* 22:400–407.

D. Siegmund (1985): *Sequential Analysis: tests and confidence intervals*. Springer.

Tuning Parameters

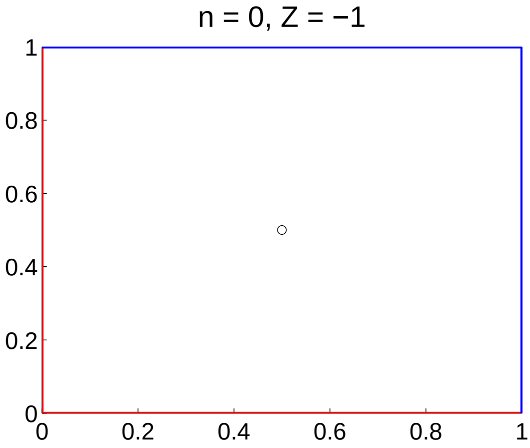


Tuning Parameters



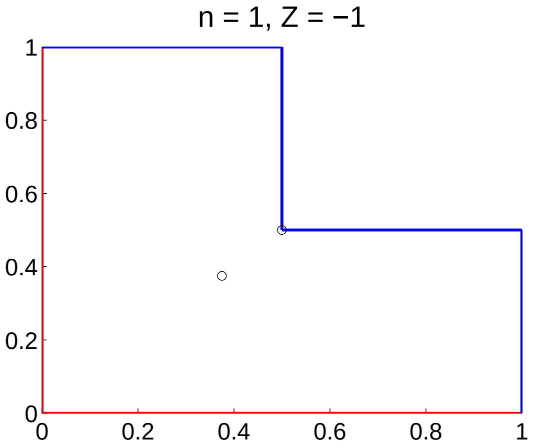
Higher Dimensions

Boundary detection:



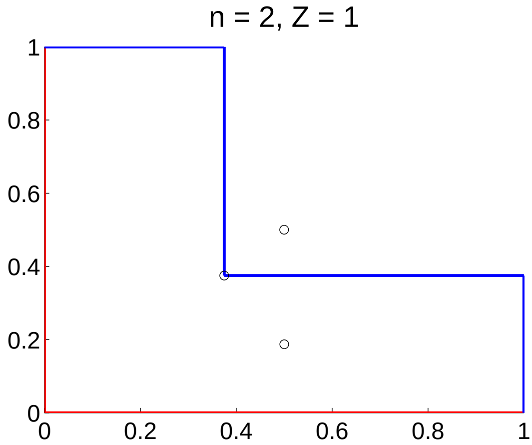
Higher Dimensions

Boundary detection:



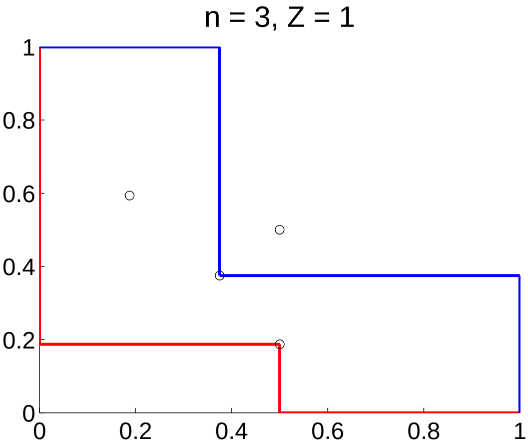
Higher Dimensions

Boundary detection:



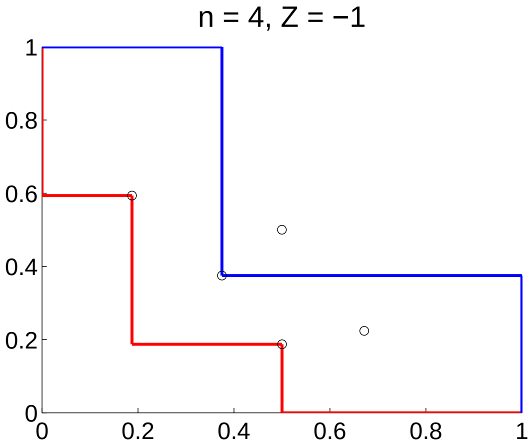
Higher Dimensions

Boundary detection:



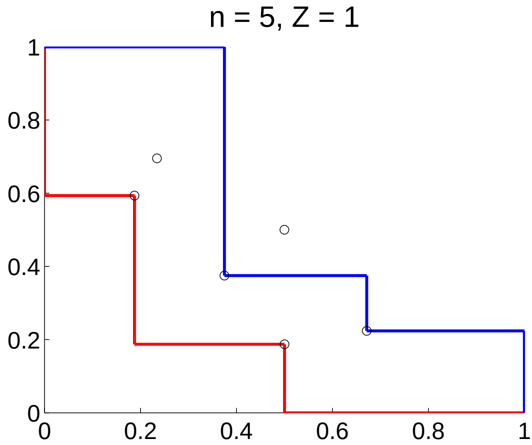
Higher Dimensions

Boundary detection:



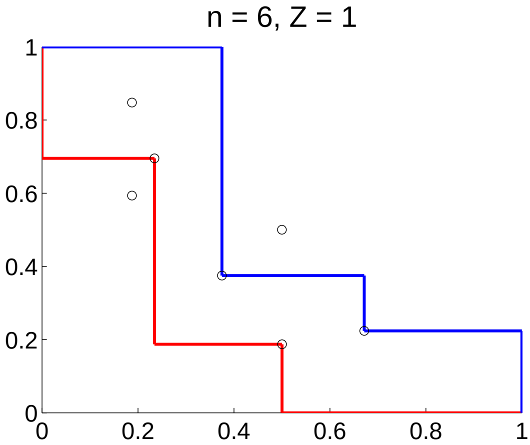
Higher Dimensions

Boundary detection:



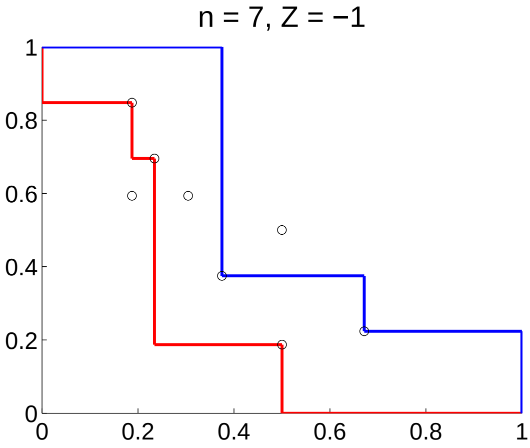
Higher Dimensions

Boundary detection:



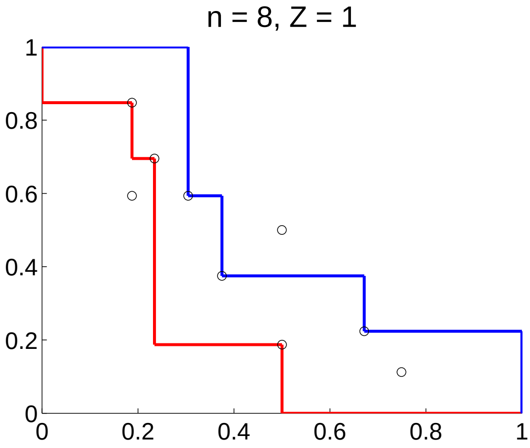
Higher Dimensions

Boundary detection:



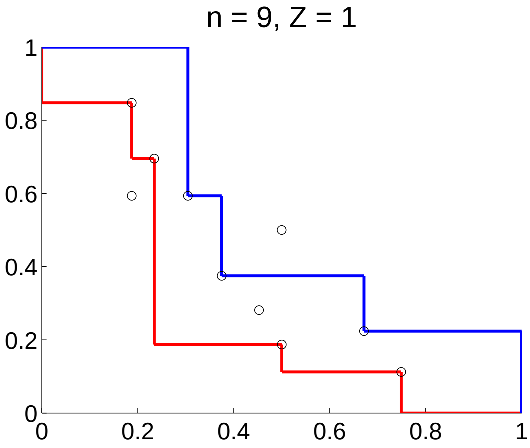
Higher Dimensions

Boundary detection:



Higher Dimensions

Boundary detection:



Higher Dimensions

Boundary detection: