

Go with the Winners for Graph Bisection*

Tassos Dimitriou[†]

Department of Computer Science
UC San Diego, CA 92093-0114
tassos@cs.ucsd.edu

Russell Impagliazzo

Department of Computer Science
UC San Diego, CA 92093-0114
russell@cs.ucsd.edu

Abstract

We analyze “Go with the winners” for graph bisection. We introduce a weaker version of expansion called “local expansion”. We show that “Go with the winners” works well in any search space whose sub-graphs with solutions at least as good as a certain threshold have local expansion, and where these sub-graphs do not shrink more than by a polynomial factor when the threshold is incremented. We give a general technique for showing that solution spaces for random instances of problems have local expansion. We apply this technique to the minimum bisection problem for random graphs. We conclude that “Go with the winners” approximates the best solution in random graphs of certain densities with planted bisections in polynomial time and finds the optimal solution in quasi-polynomial time. Although other methods also solve this problem for the same densities, the set of tools we develop may be useful in the analysis of similar problems. In particular, our results easily extend to hypergraph bisection, whereas it is not clear whether the other known techniques do.

1 Introduction

Graph Bisection is the following problem: Given an input graph G find a partition of G into two equal pieces such that the number of edges between the two pieces is the smallest among all possible partitions. Such partitions are called *bisections*. The *size* of a bisection is the number of edges crossing the cut and the *bisection width* is the size of the smallest bisection.

While this problem appears in many applications, VLSI design most notably [BL84, AK95], the problem is known to be NP-complete [GJ79], so we cannot expect

to have good performance in the worst case. Instead, heuristics are used to find good bisections on random graphs drawn from specific distributions. Since a random graph G with n nodes and m edges (we say G is drawn uniformly at random from the distribution $G_{n,m}$) is likely to have a minimum bisection of size $m/2$, we will instead assume the existence of a bisection of size b , where b will be smaller than the size of the average bisection, so that we be able to compare b with the solution returned by the heuristic.

Bui *et al.* [BCLM] consider the class $G_{n,d,b}$ of d -regular graphs on n nodes, with a planted bisection of size $b = o(n^{1-1/\lfloor (d+1)/2 \rfloor})$, and they give an algorithm, based on network flows, that finds with high probability the provably minimum bisection. Boppana [Bop87] considers the $G_{n,m,b}$ model, where graphs are chosen uniformly at random from the set of graphs with n nodes, m edges and bisection width b . His algorithm is based on computing eigenvalues on matrices associated with the input graph and returns the minimum bisection provided the planted bisection satisfies $b \leq \frac{1}{2}m - \frac{5}{2}\sqrt{mn \ln n}$. Finally, Jerrum and Sorkin [JS093] analyze the Metropolis algorithm on the model $G_{n,p,q}$ of graphs with n nodes and a planted bisection of density q separating the two equal sized partitions of higher density p (this is basically equivalent to the previous model with $m = \frac{1}{4}(p+q)n^2$ and $b = \frac{1}{4}qn^2$). They show that the Metropolis algorithm (simulated annealing with a fixed temperature) succeeds in finding the minimum bisection provided that $p - q = n^{\Delta-2}$, where $11/6 \leq \Delta < 2$.

Boppana’s algorithm applies to a fairly large class of graphs, including the graphs defined by [BCLM], provided the difference $\frac{1}{2}m - b$ (called the *deficiency* of a graph in [JS093]) is $\Omega(\sqrt{mn \ln n})$, but his results become vacuous for graphs that are relatively sparse, i.e. $m < 25n \log n$, or equivalently $p < 25 \log n/n$. Similarly, the results of [JS093] apply only when Δ is in the range defined above and moreover become vacuous when $p \leq n^{-1/6}$. Jerrum and Sorkin’s result compares unfavorably with Boppana’s result since the deficiency

*Sloan Research Fellowship BR-3311, grant #93025 of the joint US-Czechoslovak Science and Technology Program, and USA-Israel BSF Grant 92-00043.

[†]Current Address: Computer Technology Institute (C.T.I.), Greece. tassos@cti.gr

(typically $O(n^\Delta)$ in the $G_{n,p,q}$ model) is required to be much higher: $O(n^{11/6})$ against $\Omega(n^{3/2}\sqrt{\log n})$ for dense graphs. However, despite the wide range of applicability of Boppana’s algorithm, the simplicity of the Metropolis process makes it an attractive alternative even in this restricted range, since the former is not simple to implement requiring the Ellipsoid method as a subroutine.

In this paper we analyze “Go with the winners” (GWW), introduced in [AV94] as a modification for simulated annealing and in [DI96] as an optimization algorithm, and prove that it succeeds with high probability in finding a good approximation to the minimum bisection for simple graphs as well as for hypergraphs. We do this in two parts: In the first part we introduce “local expansion”, a combinatorial graph property similar to but weaker than expansion. We show that, if each of the sub-graphs of a search space consisting of solutions better than a certain threshold has this property, and the ratios of sizes for the sub-spaces for consecutive thresholds is at most polynomial, then “Go with the winners” finds optimal solutions in polynomial time. In the second part we show, via a technical lemma involving martingales, that the space of bisections has this weak expansion property. We believe these techniques to be of independent interest, suggesting that other combinatorial problems might be amenable to a similar analysis.

The rest of the paper is organized as follows: In the next section we define the notion of local expansion and prove that if a space has this property then GWW finds with high probability the optimal solution. In Section 3 we introduce the basic terminology about the Graph Bisection problem that we will be using throughout the paper. In Section 4 we give a technical lemma about martingales which we use in Section 5 to prove the main result about planted bisections. Finally, in Section 6 we extend these results to random hypergraphs.

2 Local Expansion

In [DI96] a first attempt was made to characterize the properties a solutions space of a problem Π should have, so that GWW can be applied successfully. Here we show how these properties can be weakened further. The algorithm uses B particles that independently search the space of all solutions and the goal is to find a solution of optimal value. For completeness we repeat the algorithm here as would be applied to find the minimum bisection of a graph G . Denote by \mathcal{N} the graph of all possible solutions. Two solutions are connected by an edge in \mathcal{N} if one results from the other by making a local change. In the case of bisections of a graph G , two bisections are neighbors if one results from the other by

swapping two vertices at opposite sides of the bisection.

GWW for Graph Bisection

Denote by \mathcal{N}_i the space of valid solutions during stage i of the algorithm, i.e. the space of bisections of value less than or equal to $|E| - i$.

- Stage 0: Generate B random bisections from \mathcal{N} and place particles on each one of them.
- Stage i : Proceed in three phases.
 1. In the *predistribution* phase, redistribute the particles according to their *down degree*, which is the number of neighbors $y \in \mathcal{N}_{i+1}$.
 2. In the *randomization* phase, for each new particle perform a random walk for $2t + 1$ steps, restricting the moves as follows: in odd steps, go to a random neighbor $y \in \mathcal{N}_{i+1}$, and, in even steps, go to a random neighbor $y \in \mathcal{N}_i$.
 3. In the *postdistribution* phase redistribute the particles inversely proportional to their *up degree*, which is the number of neighboring bisections $y \in \mathcal{N}_i$.

Go to the next stage.

The intuition of the above algorithm is that in the i -th stage and beyond we are implicitly deleting all bisections of size larger than $|E| - i$ from the space of all possible bisections. This divides \mathcal{N} into components. The redistribution phase will make sure that particles in locally optimal bisections will be distributed among non-local solutions, while the randomization phase will ensure that particles remain uniformly distributed as they advance to deeper and deeper levels of \mathcal{N} (here t is the time necessary for the random walk to converge to the stationary distribution).

Denote by T the (implicit) tree that results from the above decomposition of the space \mathcal{N} , where each node at level i of T represents a component of bisections of size $|E| - i$ or smaller. It was shown in [DI96] that if T has the following properties:

Properties

1. There are polynomially many components at each level of T and their sizes are within some polynomial $p(n)$ of each other,
2. Each component has good expansion and
3. $\frac{|\mathcal{N}_i|}{|\mathcal{N}_{i+1}|} < q(n)$, for some polynomial $q(n)$.

then particles remain uniformly distributed among the various components and GWW can find the optimal solution with high probability.

In this section we will deal with state spaces that do not possess the above properties in such clear cut form. In particular, we will consider the case where the state space is not entirely disconnected, as suggested by Property 1, but consists of *loosely connected* parts. Property 2 in such a case may not even be true. So instead, we will introduce the notion of *local expansion* and prove from first principles that after the randomization phase of the algorithm particles remain uniformly distributed. This will have the effect of collapsing Properties 1 and 2 into a new property which characterizes the *entire* search graph and it is easier to work with (in Section 5 we will see that if we take a random walk in \mathcal{N}_i such that the probability of never visiting a solution outside \mathcal{N}_i stays polynomially small, then \mathcal{N}_i expands locally). Property 3 will stay as it is however, since otherwise the space of all solutions would have many local optima which would probably be bad for any search algorithm.

The proof is a combination of two different parts: In the first part, we define this weak expansion property and use it to argue that a random walk on the space of solutions diffuses rapidly, i.e. never restricts itself to a small portion of the state space. More precisely, Lemma 1 states that this property is sufficient to guarantee a decrease in the collision probability by a constant factor at each step of the random walk. In the second part, we use the above result to show that particles remain well distributed in every stage of the algorithm.

In the discussion that follows, by “random walk” we mean the random walk each particle has to perform during the randomization phase of stage i of the algorithm. Also denote by \mathcal{N}_i the graph of valid solutions during the same stage. We remind again here the definition of expansion:

Definition 1 *The expansion of a graph \mathcal{N} is:*

$$v = \min_{|S| \leq \frac{|\mathcal{N}|}{2}} \left\{ \frac{|N(S)|}{|S|} \right\},$$

where $N(S)$ denotes the neighbors of S in $\mathcal{N} - S$.

Let now $p(n)$ be some polynomial.

Definition 2 *A graph \mathcal{N} has local expansion v for some polynomial $p(n)$ if and only if all subsets S of size less than $|\mathcal{N}|/p(n)$ have at least $v|S|$ neighbors outside S .*

Observe how this “implies” Property 1 above. If \mathcal{N}_i locally expands then the only way for a component not to have expansion v is if its size is larger than a fraction

Assumptions

Denote by \mathcal{N}_i the space of valid solutions during stage i of the algorithm.

1. \mathcal{N}_i locally expands.
2. $\frac{|\mathcal{N}_i|}{|\mathcal{N}_{i+1}|} < q(n)$, for some polynomial $q(n)$.

Figure 1: Yet an even weaker set of assumptions with “combinatorial flavor”.

$1/p(n)$ of \mathcal{N}_i . But in this case there cannot be more than $p(n)$ such components. Similarly, a component cannot be exponentially small (less than a polynomially small fraction of \mathcal{N}_i) as this would contradict the definition of local expansion. The new set of properties is shown in Figure 1.

We now want to prove what Property 2 was doing for us, namely that a random walk on the nodes of \mathcal{N}_i converges rapidly to the stationary distribution. However, the usual arguments involving the expanding properties of \mathcal{N}_i do not seem to carry over here. What we will do instead is first show that the collision probability decreases by a constant factor at each step of the random walk and then show that this guarantees the uniform distribution of particles.

Definition 3 *Let f be a probability distribution on some set X . The probability of picking the same element twice from X is called the collision probability of X and is given by $\|f\| = \sum_{i \in X} f_i^2$.*

The following lemma attempts to capture the decrease in collision probability of a random walk on the nodes of an arbitrary graph \mathcal{N} with weak expansion properties.

Lemma 1 *Let \mathcal{N} be a graph with local expansion v for some polynomial $p(n)$. Let also f be a distribution on the vertices of \mathcal{N} and P be the probability matrix of the random walk on \mathcal{N} . If $\|f\| \geq 4p(n)/|\mathcal{N}|$ then $\|fP\| \leq (1 - \frac{v^2}{16d^2})\|f\|$, where d is the maximum degree of \mathcal{N} .*

Proof: We follow the argument of Mihail[M89] to bound the rate at which the collision probability decreases in terms of the expansion of small sets. The convergence rate of a random walk can be viewed as the diffusion of a “charge”, essentially the discrepancy vector that expresses the distance from stationarity, along the edges of the expander. In our case however, this

argument has to be modified a little bit for two reasons. First, because only small sets have good expansion, and second because the graph may be disconnected and the stationary distribution not well defined.

In what follows let $m = \lfloor |\mathcal{N}|/p(n) \rfloor$ be the value for which sets of size smaller than m have expansion v . Order the vertices of \mathcal{N} so that $f_1 \geq f_2 \geq \dots \geq f_{|\mathcal{N}|}$, let A_k be the set of the first k vertices and (A_k, \bar{A}_k) be the set of edges between A_k and its complement.

Claim 1 *The difference $\|f\| - \|fP\|$ satisfies:*

$$\|f\| - \|fP\| \geq \frac{1}{4d^2} \frac{\left[\sum_{k=1}^{|\mathcal{N}|-1} (f_k^2 - f_{k+1}^2) |(A_k, \bar{A}_k)| \right]^2}{\|f\|},$$

where d is the maximum degree of the graph \mathcal{N} .

Proof: As in [M89]. \square

Using the definition of expansion v we know that $|(A_k, \bar{A}_k)| \geq v|A_k|$ for $k \leq m$, hence

$$\begin{aligned} \sum_{k=1}^{|\mathcal{N}|-1} (f_k^2 - f_{k+1}^2) |(A_k, \bar{A}_k)| &\geq v \sum_{k=1}^m (f_k^2 - f_{k+1}^2) k \\ &\geq v \left(\sum_{k=1}^m f_k^2 - m f_{m+1}^2 \right) \\ &\geq \frac{v}{2} \|f\| \end{aligned} \quad (1)$$

where we used the bound $f_k \leq 1/k$ and the hypothesis $\|f\| \geq 4/m$ (a few steps were omitted here). Combining Claim 1 with (1), we conclude that

$$\|fP\| \leq \left(1 - \frac{v^2}{16d^2}\right) \|f\|.$$

\square

Corollary 1 *For any initial distribution f and any random walk on \mathcal{N} , the number of steps required for the collision probability to drop below $4p(n)/|\mathcal{N}|$ is $t = \frac{16d^2}{v^2} \ln \frac{|\mathcal{N}|}{4p(n)}$.*

We are now going to use the above results to argue that particles remain almost uniformly distributed, during stage i of the algorithm, among the solutions of \mathcal{N}_i . Let P^t be probability matrix of t steps of the random walk, where t is the time constant found in Corollary 1. This is an $N_i \times N_i$ stochastic matrix so both the columns and the rows sum to 1, where $N_i = |\mathcal{N}_i|$. In particular, the *average* entry of this matrix has value $1/N_i$.

Consider now the B particles at the beginning of the random walk. Each one corresponds to some row (not necessarily distinct) of the matrix P^t . We would like to show that for each such row, each of the N_i entries has value close to $1/N_i$. This may not be the case however, as the collision probability can be as large as $4p(n)/|\mathcal{N}_i|$. What we will be able to argue though is that each solution x has the right average probability of being visited, over *all* B particles. Or in other words, prove that the weighted average for each column x of the matrix is about B/N_i .

In the analysis that follows we will show that p_x , the probability of visiting solution x , is very close to its expected value B/N_i . Denote by κ the cluster size (accumulation of particles at the same solution) at the beginning of the random walk¹, and by σ_x^2 the variance of each entry of column x . It is not difficult to see that the variance is bounded by the collision probability divided by N_i , which is in turn bounded by $4p(n)/N_i^2$. Using Chernoff bounds we have:

$$\begin{aligned} \Pr\left[|p_x - \frac{B}{N_i}| > \alpha \frac{B}{N_i}\right] &< e^{-(\alpha B/N_i \kappa)^2 / B \sigma_x^2} \\ &< e^{-\alpha^2 B / 4 \kappa^2 p(n)} \\ &< 1/N_i^2 \end{aligned}$$

provided $B = O(\alpha^{-2} \kappa^2 p(n) \ln N_i)$. Since N_i is at most exponentially large, the number of particles is polynomially bounded. We conclude that with high probability and for all solutions x the probability p_x of visiting x is given by $p_x = (1 \pm \alpha)B/N_i$.

How much error is accumulated after h stages of the algorithm? Let D_i be the distribution of a random particle p at the beginning of the randomization phase of stage i . We will show that for all stages i and for all solutions x

$$\frac{(1 - \alpha_i)}{N_i} \leq \Pr_{p \in D_i} [p = x] \leq \frac{(1 + \alpha_i)}{N_i} \quad (2)$$

where the value of α_i 's will be determined shortly.

Clearly, (2) is true for stage 1: After the initialization stage and because of the good expansion properties of the *whole* space \mathcal{N} , all solutions are equally likely, so α_1 can be taken to be any polynomially small quantity. Assume inductively that (2) is true for stage i . The analysis above shows that after the randomization

¹Intuitively, clusters will form during the re-distribution stages, when certain nodes will spawn several probes in the next generation, and others will die out. The expected size of a cluster at a node x with down-degree d_x and up-degree u_x will be $d_x / (u_x \text{Avg}_{y \in B} (d_y) * \text{Avg}_{y \in B} (1/u_y)) \leq d_x q(n) d \leq q(n) d^2$. So it will turn out that we can use $\kappa = O(q(n) d^2)$ as a good estimate.

phase, (2) will be true for stage $i + 1$ for α_{i+1} satisfying $(1 + \alpha_{i+1}) = (1 + \alpha_i)(1 + \alpha)$. Setting, for all i , $\alpha_i = \alpha = 1/h^2$, we see that after h stages the resulting distribution is almost uniform (each solution gets visited with probability at least $(1 - \alpha)^h/N_i$ and at most $(1 + \alpha)^h/N_i$).

Using the value of α found above and bounding the cluster size by $\kappa = O(q(n)\hat{d}^2)$ we see that the number of particles is given by $B = O(\hat{d}^2 h^6 p(n) q^2(n) \ln|\mathcal{N}|)$.

Theorem 1 *Let \mathcal{N} be a search graph for a given problem Π satisfying the assumptions of Figure 1. Then “Go with the winners” succeeds, with high probability and in time $O(Bht)$, in finding an optimal solution for Π using $B = O(\hat{d}^2 h^6 p(n) q^2(n) \ln|\mathcal{N}|)$ particles, where \hat{d} is the maximum degree of \mathcal{N} and t is the time constant found in Corollary 1.*

3 Notation for Graph Bisection

The $G_{n,p,q}$ model consists of all graphs G with $n/2$ white vertices and $n/2$ black vertices, with edge probability p between vertices of the same color and edge probability q between vertices of different color. It can be shown[BCLM] that with high probability the partition with the set of white vertices in one size is the minimum one, so the bisection width of $G \in G_{n,p,q}$ will be taken to be $\frac{1}{4}qn^2$. In the rest of the paper, when no confusion arises from the context, bisection will mean either a partition of the graph into two equal size pieces or the size of that partition. We will also denote by L, R the left and right sides of the partition and by W, B the set of white and black vertices.

Without loss of generality we will assume that the graphs have an even number of vertices, i.e. drawn from $G_{2n,p,q}$, so that they can be bisected exactly. Borrowing some notation from [JS03], we define the *imbalance* of a partition π to be the value $k \geq n/2$ for which there are k white vertices in one side of the partition and $n - k$ in the other. From the discussion above, the minimum partition is likely to have imbalance n . We also define the *deficiency* of the graph G to be the difference between a random bisection and the planted one. In the $G_{n,m,b}$ model the deficiency is $\frac{1}{2}m - b$. In the $G_{2n,p,q}$ model $m = n^2(p + q)$ and $b = n^2q$. Thus the deficiency is about $\frac{1}{2}n^2(p - q)$.

Given a graph $G = (V, E)$ from $G_{2n,p,q}$ we will apply GWW to it in order to find a partition with imbalance close to n . Our goal in the rest of the paper is to show that for almost all graphs $G \in G_{2n,p,q}$ a large fraction of the space \mathcal{N} has the local expansion property, thus obtaining an approximation algorithm for the minimum

bisection. Key to the analysis will be a generalization of the following puzzle:

Puzzle: Consider a particle that starts at the origin and at each time step makes a move in the positive or negative direction with probability $1/2$. What is the probability that after $2n$ steps the particle always stayed below the x -axis? If we introduce a positive bias ϵ for the upward moves, what is the largest value of ϵ so that the new probability is within some polynomial of the unbiased case?

Answer: The number of paths that stay below the x -axis and take the particle from the point $(0, 0)$ to the point $(2n, -2y)$, where $0 \leq y \leq n$, is

$$\sum_{y=0}^n \frac{2y+1}{2n+1} \binom{2n+1}{n+y+1} = \binom{2n}{n} > 2^{2n}/(n+1).$$

Thus the probability that the particle stays negative is at least $1/(n+1)$, and not exponentially small as one might have expected. To answer the second question, we consider only the paths that end up in $(2n, 0)$. Their number is $\frac{1}{n+1} \binom{2n}{n} > 2^{2n}/(n+1)^2$ and moreover the number of upward moves is equal to the number of downward moves. The probability that the biased particle stays below the x -axis in this case is therefore at least

$$\begin{aligned} \left(\frac{1}{2} - \epsilon\right)^n \left(\frac{1}{2} + \epsilon\right)^n \frac{1}{n+1} \binom{2n}{n} &> (1 - 4\epsilon^2)^n \frac{1}{(n+1)^2} \\ &> \frac{1}{(n+1)^3}, \end{aligned}$$

provided that $\epsilon < \sqrt{\ln n/8n}$. In the next section we will give a generalization of the above argument to martingales that will be used in Section 5 to show that the space \mathcal{N} has the local expansion property.

4 A Lemma about Martingales

We start by giving the definition of a martingale.

Definition 4 *Let x_i , $i \geq 0$ be independent random variables. The sequence $X = x_0, x_1, \dots$ is called a [perfect] martingale if for all $i \geq 0$*

$$E[x_{i+1} - x_i \mid x_i, \dots, x_0] = 0.$$

Similarly, an ϵ -martingale is a sequence x_0, x_1, \dots of independent random variables so that for all $i \geq 0$

$$E[x_{i+1} - x_i \mid x_i, \dots, x_0] \leq \epsilon.$$

The main lemma whose proof is given at the appendix is analogous to the second part of the puzzle. That is, assume our martingale is biased towards the positive direction by an ϵ amount. What is the largest value of ϵ so that for all $0 \leq i \leq t$, $x_i \leq x_0$?

Lemma 2 *Let $X = x_0, x_1, \dots$ be an ϵ -martingale with values in $[-M, M]$. Denote by D_+ (resp. D_-) the distribution on the positive (resp. negative) values of $(x_i - x_{i-1})$, given x_{i-1}, \dots, x_0 , and by E_+, E_- the corresponding expectations. Let also p_+ (resp. p_-) be the probability that $(x_i - x_{i-1})$ is greater (resp. smaller) than zero. Choose c_p, c_E so that $c_p \geq \max(p_+^{-1}, p_-^{-1})$ and $c_E \leq \min(E_+ + E_-)$. Then*

$$\Pr[\forall i, i \leq t, x_i \leq x_0] \geq \left(\frac{1}{M+1} - e^{-4t\epsilon^2/c_E^2} \right) e^{-4tc_p\epsilon^2/c_E^2}.$$

In particular, if $\epsilon = \frac{c_E}{2} \sqrt{\frac{\ln 1/\sigma}{t}}$ then $\Pr[\forall i, i \leq t, x_i \leq x_0] \geq \left(\frac{1}{M+1} - \sigma \right) \sigma^{c_p}$. Choosing $\sigma = 1/2M$, this probability becomes polynomially small, provided c_p is $O(1)$.

5 Properties of the Space of Bisections

In this section we prove that GWW can approximate the minimum bisection. Let $m = n^2(p+q)$, the total number of edges, and $b = n^2q$, the size of the planted bisection in the $G_{2n,p,q}$ model.

Theorem 2 *For any graph $G \in G_{2n,p,q}$ such that $\frac{1}{2}m - b > \sqrt{mn^2/\ln n}$ or $\frac{1}{2}m - b < \sqrt{m \ln n}$ “Go with the winners” succeeds in finding a bisection that is within $(1+\theta)$ of the optimum bisection using a number of particles B that is polynomial in $n^{O(\ln \frac{1}{\theta})}$.*

Our goal will be to prove that a large part of the space \mathcal{N} has the local expansion property so that GWW can be applied successfully. Since the algorithm is guaranteed to find the optimum bisection for that part of the space of solutions, by relating the size of that bisection to the optimum bisection the theorem will follow.

The proof consists of two parts: In the first part, we show that if we are at a bisection u , then almost all bisections of size less than some threshold T that depends only on the current stage can be reached from u . We do this by proving that a random walk starting at u is unlikely to visit any bisection of value greater than T . In other words, if our walk is represented by a particle starting at u , then with high probability the particle will always be below T . The connection with the results of the previous section is clearer now.

In the second part, we use this property to argue that all sets of size up to a polynomially small fraction of \mathcal{N}_i expand well. Combining the two parts together with the result of Section 2 the theorem will follow.

5.1 Good Sideways Movement for Bisections

Suppose we are at a bisection u of value less than T and imbalance k . Repeat the following $t = O((n-k) \ln \frac{n}{n-k})$ times (this value of t will be justified later): Pick at random two vertices l, r at opposite sides of the partition that have the same color and interchange them. What is the probability that we never visit a bisection of value greater than T ? Observe that if this probability is polynomially small and the resulting bisection is a random one, then there can be no more than a polynomial number of components. This however, doesn’t exclude the possibility of these “components” being loosely connected and that’s why we will only try to prove that the local expansion property holds. Bear also in mind that this random walk is *not* part of the algorithm. The algorithm has no way of knowing how the vertices are colored or what is the imbalance of a bisection.

Let $x_0 = |u|$ be the cost of bisection u , and y_i be the change in cost resulting by interchanging two vertices in the i -th step of the walk. The distribution of y_i can be expressed as the sum and difference of 8 sets of edges according to the color and the location of their endpoints in the partition. To fix a notation let E_{LWLW} be the set of edges among the k white vertices on the left side, E_{LWRW} be the set of edges connecting the k white vertices on the left with the $n-k$ white vertices on the right and similarly define all other cases.

Given such a set, denote by N the total number of potential edges in the set. Each edge in the set appears with probability P equal to either p or q , and the contribution V of each edge to the expected change is the probability that the edge is connected to either l or r . The possible combinations are shown in Figure 2. Then y_i is given by the sum NPV over all possible sets of edges which is not difficult to see that if the sets $E_{LWLW}, \dots, E_{RWRB}$ didn’t deviate from their expected values shown in Figure 2, $x_i = x_{i-1} + y_i$ would be a perfect martingale. Since this is not necessarily the case, we have to find the values of imbalance k for which the sequence x_0, x_1, \dots, x_t remains an ϵ -martingale.

A failed first attempt: One approach to do this is by bounding the deviation of each such set from its expected value by an ϵ amount. Doing so, the change y_i , which is the sum NPV for each such set, will be

Set of edges	# of edges N	Prob. P	Value V
$\overline{E_{LWLW}}$	$\binom{k}{2}$	p	$\frac{2}{k}$
$\overline{E_{LWRW}}$	$k(n-k)$	p	$-\frac{n-2}{k(n-k)}$
$\overline{E_{LWRB}}$	k^2	q	$-\frac{1}{k}$
$\overline{E_{LWLB}}$	$k(n-k)$	q	$\frac{1}{k}$
$\overline{E_{LBRW}}$	$(n-k)^2$	q	$-\frac{1}{n-k}$
$\overline{E_{LBRB}}$	$k(n-k)$	p	0
$\overline{E_{RWRW}}$	$\binom{n-k}{2}$	p	$\frac{2}{n-k}$
$\overline{E_{RWRB}}$	$k(n-k)$	q	$\frac{1}{n-k}$

Figure 2: Different sets of edges contributing to the expected change y_i .

bounded by $O(\epsilon)$ and the resulting sequence will be an ϵ martingale. In fact we will consider only the second set of edges shown in Figure 2 since its contribution is the dominating one. Using Chernoff bounds we get:

$$\begin{aligned} \Pr[|E_{LWRW} - NPV| > \epsilon] &< e^{-(\epsilon/V)^2/NP} \\ &= e^{-\epsilon^2/NPV^2} \quad (3) \\ &\leq \frac{1}{n^{O(1)}} \end{aligned}$$

provided $\epsilon = O(\sqrt{NPV^2 \ln n})$. Substituting the values of N, P and V we find that $\epsilon = O(\sqrt{\frac{np \ln n}{n-k}})$. What is now the probability q_{valid} that in t steps we never visit a bisection of size greater than that of u ? Plugging in the value of ϵ found above to the probability in Lemma 2 and estimating M, c_p and c_E by $n^2, O(1)$ and $O(\sqrt{np})$ respectively, we see that q_{valid} is given by:

$$\begin{aligned} q_{valid} &= e^{-O(\frac{tnp \ln n}{(n-k)np})} \\ &= n^{-O(\frac{t}{n-k})} \\ &= n^{-O(\ln \frac{n}{n-k})} \end{aligned}$$

since $t = O((n-k) \ln \frac{n}{n-k})$. Observe that this probability remains polynomially small so long $n-k = O(n)$. This is almost all we had to prove as this probability must also be greater than the probability q_ϵ shown in (3). The reason being that q_ϵ is the probability that the deviation fails to be within the required bounds, so we want to make sure that the total failure probability $q_\epsilon + (1 - q_{valid})$ is less than one, or $q_\epsilon < q_{valid}$. This is not the case however, as one can show by substitution that $q_{valid} = q_\epsilon^{O(\ln \frac{n}{n-k})}$. This approach was bounded to fail because two contradictory situations arose. Trying to stay below u meant exactly entering the region where the deviation becomes big. But we can stay below u only if the deviation is small.

A better approach: We failed in our previous approach because we restricted the random walk to stay below u and not below the threshold T , as was originally our goal. The smaller is u from its expected value, the larger the deviation becomes but also the larger the distance from T becomes. Hence we might be able to correct the problem by taking the value of u into account.

Let v be the neighbor of u . Then the expected change $|v| - |u|$ satisfies (to simplify the notation, hereby u will also denote the size of the bisection)

$$\begin{aligned} \text{Exp}[v - u] &\leq \left(\frac{1}{k} + \frac{1}{n-k}\right)(E_k - u) \\ &< \frac{2}{n-k}(E_k - u) \\ &= \epsilon_0 + \frac{2}{n-k}(T - u) \quad (4) \end{aligned}$$

where $E_k = 2k(n-k)(p-q) + n^2q$ is the expected value of a bisection of imbalance k and $\epsilon_0 = \frac{2}{n-k}(E_k - T)$. Define now the extended value of u to be

$$P(u, t_I) = u + \frac{2}{n-k}t_I(T - u)$$

where t_I is the length of a subinterval of the random walk so that $\frac{2}{n-k}t_I < 1$. Since $t = O((n-k) \ln \frac{n}{n-k})$ there can be at most $O(\ln \frac{n}{n-k})$ such subintervals. Our goal now is to show that for two neighboring bisections u, v the expected change $P(v, t_I - 1) - P(u, t_I)$ is at most ϵ_0 , thus making the whole process an ϵ_0 -martingale. By direct substitution we have:

$$\begin{aligned} P(v, t_I - 1) - P(u, t_I) &= \\ &= \left(1 - \frac{2}{n-k}t_I\right)(v - u) - \frac{2}{n-k}(T - v) \end{aligned}$$

Taking expectations of both sides and using (4) we obtain

$$\text{Exp}[P(v, t_I - 1) - P(u, t_I)] \leq \epsilon_0$$

Thus the whole process is an ϵ_0 -martingale for each subinterval of length t_I . What is the probability that we stay below the threshold T ? Using the results of Lemma 2, for each subinterval of the random walk this probability is

$$\begin{aligned} q_I &= e^{-O(t_I c_p \epsilon_0^2 / c_E^2)} \\ &= e^{-O((n-k) \epsilon_0^2 / np)} \end{aligned}$$

Thus the overall probability is simply q_I raised to the number of such intervals, i.e.

$$\begin{aligned} q_{valid} &= q_I^{O(\ln \frac{n}{n-k})} \\ &= e^{-O(\frac{(n-k) \epsilon_0^2}{np} \ln \frac{n}{n-k})} \quad (5) \end{aligned}$$

We need now estimate the maximum allowable value of ϵ_0 , so that q_{valid} remains polynomially small. Since ϵ_0 is proportional to the difference $E_k - T$, this will also give us the minimum value of thresholds T that the algorithm can be applied successfully.

Definition 5 *Let u be a random bisection of imbalance k . We call a threshold T good for k if and only if*

$$\Pr_u[u \leq T] \geq \frac{1}{n^{O(1)}}$$

It is not difficult to see that if u is a random k -bisection, then the probability that the value of u is less than T is about $p_{k,T} = e^{-(E_k - T)^2 / E_k}$. Thus T is good for k if $E_k - T \leq \sqrt{E_k \ln n}$ or $T \geq E_k - \sqrt{E_k \ln n}$. Since E_k is $O(n^2 p)$, the difference $E_k - T$ can be at most $O(\sqrt{n^2 p \ln n})$ and thus $\epsilon_0 = O(\frac{1}{n-k} \sqrt{n^2 p \ln n})$. Plugging this value into (5), we finally have

$$q_{valid} = n^{-O(\frac{n}{n-k} \ln \frac{n}{n-k})} \quad (6)$$

By being more careful we can get rid of the extra $n/(n-k)$ factor in the exponent. We now have to find the smallest value of threshold T that the algorithm can be applied successfully. More precisely, we need to find the T so that the k -bisections for which T is bad, constitute only a polynomially small fraction of the space of bisections that have value less than T . By doing so, we will ultimately find the range of distributions $G_{2n,p,q}$ for which the algorithm is effective. Denote by N_k , $n/2 \leq k \leq n$, the number of k -bisections, i.e. $N_k = \binom{n}{k}^2$. Our goal is to find the conditions under which

$$\frac{\sum_{k \text{ is bad}} N_k p_{k,T}}{\sum_k N_k p_{k,T}} \leq \frac{1}{n^{O(1)}} \quad (7)$$

Let r_k be equal to the ratio $N_k p_{k,T} / N_{k+1} p_{k+1,T}$. If we can show that for all $k < k_0$, where k_0 is the largest k which is bad for T , r_k is less than one, then (7) will be satisfied since the numerator will be at most $(k_0 - n/2) N_{k_0} p_{k_0,T}$ while the denominator will be at least $N_{k_0+1} p_{k_0+1,T}$. This quantity however is polynomially small as the ratio N_{k_0} / N_{k_0+1} is bounded by a constant while the ratio $p_{k_0,T} / p_{k_0+1,T}$ is polynomially small since k_0 is bad for T .

Expanding r_k we have

$$\begin{aligned} r_k &= \left(\frac{k+1}{n-k} \right)^2 e^{-\frac{(E_k - T)^2}{E_k}} e^{\frac{(E_{k+1} - T)^2}{E_{k+1}}} \\ &< \left(\frac{k+1}{n-k} \right)^2 e^{-(E_k - E_{k+1}) \frac{(E_k - T)}{E_k}} \\ &< \left(\frac{k+1}{n-k} \right)^2 e^{-2(2k-n+1)(p-q) \frac{\sqrt{E_k \ln n}}{E_k}} \\ &= e^{-2(2k-n+1) \left(\frac{p-q}{n} \sqrt{\frac{\ln n}{p} - \frac{2}{n}} \right)} \end{aligned} \quad (8)$$

where the last inequality follows after estimating $(k+1)/(n-k)$ as $e^{2(2k-n+1)/n}$ when both k and $n-k$ are $O(n)$, and overestimating E_k by $n^2 p$. It thus follows that so long $(p-q) > 2\sqrt{p/\ln n}$, the number of k -bisections for which k is bad for T is always a polynomially small fraction of the state space, no matter how small T becomes. Translating this condition on $(p-q)$ back to the deficiency of the graph, we find that the difference $\frac{1}{2}m - b$ must satisfy

$$\frac{1}{2}m - b > \sqrt{mn^2 / \ln n} \quad (9)$$

Similarly, we can show that GWW can be effective when

$$\frac{1}{2}m - b \leq \sqrt{m \ln n} \quad (10)$$

Observe that (9) applies only for very dense graphs ($m = \Omega(n^2 / \ln n)$). In the next section however we will see that this situation improves a lot if one considers hypergraphs.

It remains now to see why after $t = O((n-k) \ln \frac{n}{n-k})$ steps the resulting bisection is a random one. Consider the set of white vertices in u . Initially, we have k vertices on the left side of the partition and $n-k$ on the right. At the end of the walk we want to argue that the resulting partition v is a random one. What would be the properties of such v ? We know by standard probabilistic arguments that v would have $(n-k)^2/n$ white vertices in common with the $n-k$ vertices of u . Thus, our goal is to perform the walk so that $(n-k) - (n-k)^2/n$ random vertices get removed from the right and be replaced with random vertices from the left. With one extra detail, however. When we remove a vertex that originally belonged to the right side and put it to the left, we are not allowed to move it back at a later step. In other words, we mark such a vertex once it is moved to the left, and we cannot switch it back again. There is no limitation however with vertices that originated from the left side, as they can be switched back and forth.

This walk is reminiscent of sampling with replacement from a population of N distinct elements until r different elements are obtained. In our case, $N = n-k$ and $r = (n-k) - (n-k)^2/n$, the vertices that have to be removed. The expected sample size (length of random walk) in such a situation is less than $N \ln \frac{N}{N-r}$ (see also [Fel68, Chapter IX.3]). Replacing the values of N and r we find that $t = O((n-k) \ln \frac{n}{n-k})$.

It is not difficult to see that the resulting bisection v is uniform among all bisections that have $(n-k)^2/n$ white vertices in common with u . To make it random among all bisections, we simply choose a value I for the size

of intersection with the right probability (the expected size is $(n - k)^2/n$), and we perform the random walk until $n - k - I$ vertices get removed.

5.2 Sideways Movement Implies Local Expansion

Our goal in this part is to prove that all sets S of size less than a polynomially small fraction of \mathcal{N}_i expand well. If we compute the number of neighbors $N(S)$, then the expansion will be at least $|N(S)|/|S|$, since S is arbitrary. The way to do this is by counting the number of edges leaving S . The number of neighbors will simply be the number of edges divided by the maximum degree, which is less than n^2 in our case.

A standard technique [JS88] of counting the number of edges leaving S is first to define a set of canonical paths among the vertices of \mathcal{N}_i and then establish that no edge is used by too many of these paths. If the congestion of each edge is bounded by $b|\mathcal{N}_i|$, the expansion will be at least $1/2bd$, where d is the maximum degree of the graph. In our case however, it seems difficult to get an easy bound on the congestion of each edge due to the special nature of the random walk. What we will do instead is argue that two valid paths starting in S are unlikely to have any nodes in common. Since these paths exit S using different edges, by estimating the number of such disjoint paths we will be able to get a lower bound on the number of neighbors of S .

Let \mathcal{P}_S be the set of paths leaving S , i.e. $|\mathcal{P}_S| = |S||\bar{S}|$, and \mathcal{V}_S be the set of paths that do not use any invalid bisections, i.e. $|\mathcal{V}_S| = q_{valid}|\mathcal{P}_S|$.

Lemma 3 $\Pr_{P_1, P_2 \in \mathcal{P}_S}[P_1 \cap P_2 \neq \emptyset] \leq n^2/|S|$.

Proof: Consider two random paths in S . The probability that these paths start from the same node is equal to the collision probability of the uniform distribution in S , which is equal to $1/|S|$. Due to the good expansion properties of the *whole* space of bisections and in light of Lemma 1, the collision probability at each step of the walk is always bounded by $1/|S|$. What is the probability that two specific nodes, one from each path, are the same? This probability is bounded by the maximum of the collision probabilities of these nodes, which is always less than $1/|S|$. Summing over all t^2 possible pairs, where $t < n$ is the length of the random walk, the lemma follows. \square

It follows that two paths in \mathcal{V}_S have a common node with probability at most $q_c = n^2/q_{valid}^2|S|$. Define now a graph H where each node represents one of the $|S||\bar{S}|$ possible paths leaving S . Two such nodes are connected

by an edge if and only if the corresponding paths are not disjoint, something that happens with probability q_c . Our goal is to find an independent set in H , i.e. a set of disjoint paths. Applying Turan's theorem in H we see that there exist an independent set of size at least $q_{valid}^2|S|/2n^2$. Since these paths exit S using different edges, we conclude that the number of these edges is at least $q_{valid}^2|S|/2n^2$ and therefore the expansion is at least $q_{valid}^2/4n^4$, since S is arbitrary.

What is the maximum size of S ? S has to be such so that a random walk starting in S should not end up in S . This will happen if the probability $|S|/|\mathcal{N}_i|$ of returning in S is smaller than q_{valid} . In other words $|S| \ll q_{valid}|\mathcal{N}_i|$. We conclude this section by proving Theorem 2.

Proof of Theorem 2: Set $p(n) = 1/q_{valid} = O(n^{\ln \frac{2}{n-2k}})$. From the discussion above it follows that all sets of size less than $|\mathcal{N}_i|/p(n)$ have expansion at least $1/p^2(n)n^4$. Thus the space \mathcal{N}_i satisfies the notion of local expansion and the results of Section 2 apply here. Since $p(n)$ remains a polynomial as long as $n - k = O(n)$, we can apply GWW to find the minimum bisection for that part of the space of solutions. But how good is this bisection? Let u be the k -bisection returned by the algorithm. The expected size of u is $E_k = 2k(n - k)(p - q) + n^2q$. Set now $k = \delta n$ and $p - q = \gamma q$. Let $b = n^2q$ be the size of the planted bisection. We can write the expected size of u as $E_k = (1 + 2\delta(1 - \delta)\gamma)b < (1 + 2(1 - \delta)\gamma)b = (1 + \theta)b$. Thus the algorithm succeeds in finding a bisection that is within $(1 + \theta)$ of the optimum using a number of particles that is a polynomial on $p(n) = n^{O(\ln \frac{2}{\theta})}$. Since the interesting case is when $\gamma = \Omega(1)$, the algorithm uses polynomially many particles when θ remains a constant. Better approximations can be found at the expense of the particles used and hence the running time of algorithm. The theorem follows since equations (9) and (10) give the range of distributions for which the algorithm is effective. \square

6 Extension to Hypergraphs

Graph bisection is a problem that appears frequently in VLSI because of the hierarchical approach often used in system design. A very useful representation of such a system is by a hypergraph $H(V, E)$, where the node set $V = \{1, 2, \dots, n\}$ represents the set of different modules of the system and the hyperedges $E = \{1, 2, \dots, m\}$ represent the input-output relationships among the various modules. For example edge $e = \{v_1, v_4, v_5\}$ might

represent a signal connecting modules v_1 , v_4 and v_5 . In general, an edge $e \in E$ can be any subset of V such that $|e| \geq 2$. We call a hypergraph $H(V, E)$ l -uniform if each of the edges contains exactly l vertices of V .

In the previous sections we studied the behavior of GWW for simple graphs (or 2-uniform hypergraphs). We saw in Theorem 2 that the algorithm can approximate the size of the planted bisection provided the graph is quite dense. In what follows we will show that GWW can do much better in bisecting random l -uniform hypergraphs, where $l \geq 3$. We start our discussion by considering random 3-uniform hypergraphs.

Our model consists of all graphs $H_{2n,p,q}$ with n white and n black vertices. An edge of such a graph is called *monochromatic* if it connects vertices of the same color. Monochromatic edges in H appear with probability p , while edges containing different colored vertices appear with probability q . A bisection of H is a partition of the graph into two equal sized pieces and the size of the bisection is the number of hyperedges with at least two vertices in different sides of the partition. If $q < p$ then with high probability the partition with the set of white vertices in one side is the minimum one, so the bisection width of $H \in H_{2n,p,q}$ will be taken to be $2\binom{n}{1}\binom{n}{2}q$.

Working as in Section 3, the deficiency is about $\frac{1}{4}n^3(p-q)$. We will be able to show that GWW can approximate the minimum bisection for graphs H whose deficiency satisfies $\frac{3}{4}m - b \geq \frac{1}{2}\sqrt{mn^2/\ln n}$. This is already an improvement over the results of Section 5 as this implies $p \geq 1/n \ln n$. Observe that a dense 3-uniform hypergraph has $p = O(1)$ while a sparse one has $p = O(1/n^2)$. The details are omitted from this paper.

7 Future Research

We gave a new property that characterizes the entire space of solutions, namely local expansion, and a general technique for showing that solutions spaces for random instances of problems have this property. We applied this technique to the minimum bisection problem and showed how GWW can approximate the minimum bisection for random graphs as well as hypergraphs. We believe the set of tools we developed may be found helpful in the analysis of similar optimization problems such as SAT, cliques (see also [Jer]), and so on.

8 Acknowledgements

We would like to thank David Zuckerman for pointing out and explaining the significance of reference [M89] to

us, and for other helpful conversations.

References

- [AV94] D. Aldous and U. Vazirani. "Go with the winners" Algorithms. In *Proc. 35th FOCS*, pages 492–501, 1994.
- [AK95] C. J. Alpert and A. B. Kahng. Recent Directions in Netlist Partitioning: A Survey. *Integration: The VLSI Journal*, 19:1–81, 1995.
- [BL84] S. Bhatt and F. T. Leighton. A framework for solving VLSI graph layout problems. *Journal of Computer and System Sciences*, 28:300–343, 1984.
- [Bop87] R. B. Boppana. Eigenvalues and graph bisection: An average case analysis. In *Proc. 28th FOCS*, pages 280–285, 1987.
- [BCLM] T. Bui, S. Chaudhuri, T. Leighton and M. Sipser. Graph bisection algorithms with good average case behavior. In *Proc. 25th FOCS*, pages 181–192, 1984.
- [DI96] T. Dimitriou and R. Impagliazzo. Towards an analysis of local optimization algorithms. In *Proc. 28th STOC*, 1996.
- [Fel68] William Feller. *An Introduction to Probability Theory and its Applications*. Vol. 1, 3rd Edition, 1968, John Wiley & Sons.
- [GJ79] M. R. Garey and D. S. Johnson. *Computers and Intractability*. Freeman, San Francisco, CA, 1979.
- [JS88] M. R. Jerrum and A. Sinclair. Conductance and the rapid mixing property of Markov chains: The approximation of the permanent resolved. In *Proc. 20th STOC*, pages 235–244, 1988.
- [Jer] M. R. Jerrum. Large cliques elude the Metropolis process. In *Random Structures and Algorithms*, 3(4), 347–359, 1992.
- [JSo93] M. R. Jerrum and G. Sorkin. Simulating annealing for graph bisection. In *Proc. 34th FOCS*, pages 94–103, 1993.
- [M89] M. Mihail. Conductance and Convergence of Markov Chains: A Combinatorial Treatment of Expanders. In *Proc. 30th FOCS*, pages 526–531, 1989.

Appendix

Here, we complete the proof of Lemma 2. Our first result is analogous to the first part of the puzzle and estimates the probability of a martingale “staying below” the x -axis which is now translated to x_0 . In particular it answers the following question: What is the probability that in t steps the values of the variables x_1, x_2, \dots, x_t are all less than $x_0 + 1$? (If the martingale takes on only integer values, this is the same as being $\leq x_0$).

Lemma 4 *Let X be a perfect martingale in $[-M, M]$. Then $\Pr[\forall i, 0 \leq i \leq t, x_i < x_0 + 1] \geq 1/(M + 1)$.*

Proof: Let p be the probability that in t steps the particle stays below x_0 . We will dominate X with another martingale X' such that $x'_i = x_i$ when $x_i \leq x_0$ but $x'_i = x_{i_0}$, for $i > i_0$ when $x_{i_0} > x_0$. Clearly, if X is a perfect martingale then X' is also perfect. The expected difference $x'_i - x'_0$, $0 \leq i \leq t$, will be at most $p(-M) + (1-p)(+1)$ which is bounded from above by $E[x'_i - x'_0] = E[x_i - x_0] = 0$. Solving for p , we obtain $p \geq 1/(M + 1)$. \square

We now turn our attention to the second part of the puzzle. Assume our martingale is biased towards the positive direction by an ϵ amount. What is the largest value of ϵ so that for all $0 \leq i \leq t, x_i \leq x_0$?

Proof of Lemma 2: The idea is to reduce the biased case to the unbiased one by showing that the number of biased sequences x_1, \dots, x_t is polynomially related to the number of unbiased sequences $\tilde{x}_1, \dots, \tilde{x}_t$ that stay below x_0 . Then, using Lemma 4 the result will follow.

Let $\tilde{x}_1, \dots, \tilde{x}_t$ be chosen as follows. Let D be the conditional distribution on x_i given that $x_1 = \tilde{x}_1, \dots, x_{i-1} = \tilde{x}_{i-1}$, and let $D_+, D_-, E_+, E_-, p_+, p_-$ be as in the statement of the lemma. Since X is an ϵ -martingale, these will satisfy

$$p_+ E_+ - p_- E_- \leq \epsilon.$$

“Adjust” now p_+, p_- by setting $q_+ = p_+ - \delta$ and $q_- = p_- + \delta$, so that $q_+ E_+ - q_- E_- = 0$, by picking $\delta = (p_+ E_+ - p_- E_-)/(E_+ + E_-)$. Clearly, $\delta \leq \epsilon/(E_+ + E_-) \leq \epsilon/c_E$. Sample now \tilde{x}_i from D_+ with probability q_+ and from D_- with probability q_- . It is not difficult to see that $\tilde{X} = \tilde{x}_1, \dots, \tilde{x}_t$ is a perfect martingale. Thus the set

$$G = \{\alpha_1, \dots, \alpha_t \mid \alpha_i \leq x_0\},$$

of all sequences $\alpha_1, \dots, \alpha_t$ that stay below x_0 has according to Lemma 4 measure at least $1/(M + 1)$ or in other words $\Pr[\tilde{X} \in G] \geq 1/(M + 1)$. Define now the set B of bad sequences $\vec{\alpha} = (\alpha_1, \dots, \alpha_t)$ so that

$$B = \{\vec{\alpha} \mid \Pr[X = \vec{\alpha}] < \sigma^{c_p} \Pr[\tilde{X} = \vec{\alpha}]\},$$

where $\sigma < 1$ and $c_p = \max(p_+^{-1}, p_-^{-1})$.

Claim 2 $\Pr[\tilde{X} \in B] < \sigma$.

Using the claim we know that $\Pr[\tilde{X} \in G - B] \geq (\frac{1}{M+1} - \sigma)$. Applying the definition of B , this means that $\Pr[X \in G - B] \geq (\frac{1}{M+1} - \sigma)\sigma^{c_p}$ and the lemma follows. Thus in order to complete the proof we only have to show that the claim is true.

Proof of Claim 2: $\vec{\alpha} = (\alpha_1, \dots, \alpha_t) \in B$ if and only if the following is true:

$$\Pr[X = \vec{\alpha}] < \sigma^{c_p} \Pr[\tilde{X} = \vec{\alpha}] \quad \text{or,}$$

$$\ln \frac{\Pr[\tilde{X} = \vec{\alpha}]}{\Pr[X = \vec{\alpha}]} > c_p \ln \frac{1}{\sigma} \quad \text{or,}$$

$$\sum_{i=1}^t \ln \frac{\Pr[\tilde{x}_i = \alpha_i \mid \alpha_{i-1}, \dots, \alpha_1]}{\Pr[x_i = \alpha_i \mid \alpha_{i-1}, \dots, \alpha_1]} > c_p \ln \frac{1}{\sigma} \quad \text{or,}$$

$$\sum_{i=1}^t v_i > c_p \ln \frac{1}{\sigma},$$

where we set $v_i = \ln \frac{\Pr[\tilde{x}_i = \alpha_i \mid \alpha_{i-1}, \dots, \alpha_1]}{\Pr[x_i = \alpha_i \mid \alpha_{i-1}, \dots, \alpha_1]}$. We will now use Hoeffding’s lemma to argue that this happens with probability at most σ . The expectation E_i of each v_i is given by

$$\begin{aligned} E_i &= q_+ \ln \frac{q_+}{p_+} + q_- \ln \frac{q_-}{p_-} \\ &= q_+ \ln(1 - \frac{\delta}{p_+}) + q_- \ln(1 + \frac{\delta}{p_-}) \\ &< -q_+ \frac{\delta}{p_+} + q_- \frac{\delta}{p_-} \\ &= \delta^2 \left(\frac{1}{p_+} + \frac{1}{p_-} \right) \\ &< 2c_p \delta^2 \end{aligned}$$

Thus, the total expected change $\sum_i E_i$ is bounded by $2tc_p \delta^2$. Similarly, the absolute value of each v_i is bounded by $c_p \delta$. Applying Hoeffding’s inequality we get

$$\begin{aligned} \Pr\left[\sum_i v_i - \sum_i E_i > \beta\right] &< e^{-(\beta/|v_i|)^2/t} \\ &< e^{-\beta^2/tc_p^2 \delta^2} \end{aligned}$$

Choosing $\beta = \sum_i E_i$ and setting the above probability equal to σ we find that

$$\Pr\left[\sum_i v_i > 4tc_p \delta^2\right] < e^{-4t\delta^2} \quad \text{or,}$$

$$\Pr\left[\sum_i v_i > c_p \ln \frac{1}{\sigma}\right] < \sigma \quad \text{for} \quad \delta = \frac{1}{2} \sqrt{\frac{\ln 1/\sigma}{t}}.$$

The lemma follows since $\epsilon \geq c_E \delta$. \square