The Annals of Probability
2003, Vol. 31, No. 2, 1028–1051

NONDIFFERENTIABILITY OF THE TIME CONSTANTS OF FIRST-PASSAGE PERCOLATION

BY J. MICHAEL STEELE AND YU ZHANG

University of Pennsylvania and Colorado State University

We study the paths of minimal cost for first-passage percolation in two dimensions and obtain an exponential bound on the tail probability of the *ratio* of the lengths of the shortest and longest of these. This inequality permits us to answer a long-standing question of Hammersley and Welsh on the *shift differentiability* of the time constant. Specifically, we show that for subcritical Bernoulli percolation the time constant is not shift differentiable when p is close to one-half.

1. Introduction and main results. As usual in percolation theory, we view the two-dimensional rectangular lattice $\mathbb{Z}^2$ as a graph with an edge e between each pair of vertices u and v in $\mathbb{Z}^2$ for which $\|u - v\| = 1$ where the norm is defined by $\|u - v\| = |u_1 - u_2| + |v_1 - v_2|$. We identify the edge $e = (u, v)$ with the *open* line segment in $\mathbb{R}^2$ from u to v, and to each edge we associate a random variable $x(e)$ that one may view as the amount of time that is needed to go from u to v. In general, the random variables $\{x(e)\}$ are assumed to be independent with a common distribution F that has a finite mean, but, in fact, we are mainly concerned with Bernoulli random variables and shifted Bernoulli variables (i.e., random variables that take the values t and $t + 1$ with probability p and $1 - p$, respectively).

A path γ from the vertex u to the vertex v is understood to be an alternating sequence of *distinct* vertices and edges $\{v_0, e_1, v_1, \ldots, e_n, v_n\}$ such that e_i is the edge between v_{i-1} and v_i, and the terminal vertices are $v_0 = u$ and $v_n = v$; so, for us, a path is always a self-avoiding path. Finally, we use $|\gamma|$ to denote the *length*, or, more precisely, the number of edges in the path γ.

For any path γ, the *passage time* of γ is defined to be the sum

$$\tau(\gamma) = \sum_{e \in \gamma} x(e),$$

and the first-passage time from u to v is defined as the infimum of $\tau(\gamma)$ over all γ paths in $\mathbb{Z}^2$ from u to v. When $u = (m, 0)$ and $v = (n, 0)$, the first-passage time is denoted by $a_{m,n}$, and a key property of the random variables $\{a_{m,n} : 0 \le m < n < \infty\}$ is that they form a subadditive process in the sense of

Received July 2001; revised March 2002.

AMS 2000 subject classifications. Primary 82B43; secondary 60K35.

Key words and phrases. First-passage percolation, Bernoulli percolation, Hammersley, Welsh, differentiability, time constants, shortest path, longest path, surgery.

Kingman, so Kingman's subadditive ergodic theorem then tells us that there exists a finite constant $\mu(F)$ such that

$$\lim_{n\to\infty} a_{0,n}/n = \mu(F) \qquad \text{almost surely and in } L^1. \tag{1.1}$$

The *time constant* $\mu(F)$ defined by the limit (1.1) has been studied extensively, but much remains unknown about its behavior. The main idea pursued here is that useful information about $\mu(F)$ may be obtained by studying the ratio of lengths of the longest path and the shortest path that are contained in a set of minimum cost paths from u to v.

Lengths of shortest routes. The set of all paths from u to v in $\mathbb{Z}^2$ is denoted by $\mathcal{P}[u, v]$, and a path from u to v that attains the minimal cost over all possible paths is called a *route*. Hammersley and Welsh (1965) observed that since $\mathcal{P}[u, v]$ is infinite, the set of such optimal time paths can be empty, but they proved that if the edge times $\{x(e)\}$ are bounded then the set of routes $\mathcal{R}[u, v]$ is nonempty with probability 1. Hammersley and Welsh also conjectured that the boundedness hypothesis could be dropped, and this was later confirmed by results of Smythe and Wierman (1977) and Wierman and Reh (1978).

One of the many useful quantities introduced in Hammersley and Welsh (1965) is the length N_n of the shortest route from $(0, 0)$ to $(n, 0)$, or

$$N_n = \min\{|\gamma| : \gamma \in \mathcal{R}[(0, 0), (n, 0)]\}.$$

The asymptotic behavior of N_n is not as well understood as one might like, but some basic facts are known. In particular, in the supercritical case when $F(0) > 1/2$, Zhang and Zhang (1984) proved that there is a finite constant $\lambda(F)$ such that

$$\lim_{n\to\infty} N_n/n = \lambda(F) \qquad \text{almost surely and in } L^1.$$

It is still not known if N_n/n converges when $F(0) \le 1/2$, but Kesten (1980) proved that for $F(0) < 1/2$ there are constants $h = h(F) < \infty$ and $C = C(F) > 0$ such that

$$P(N_n \ge hn) \le \exp(-Cn), \tag{1.2}$$

so for $F(0) < 1/2$ we at least know

$$1 \le \liminf N_n/n \le \limsup N_n/n \le h \qquad \text{almost surely.} \tag{1.3}$$

Although these results may seem to suggest that one also has a genuine limit for N_n/n in the subcritical case $F(0) < 1/2$, a proof of this conjecture still seems far away. Moreover, the behavior of N_n/n in the critical case $F(0) = 1/2$ suggests that the problem may be quite subtle. For example, Kesten [(1986), page 259] conjectures that the ratio N_n/n should diverge to infinity when $F(0) = 1/2$.

The Hammersley–Welsh differentiation principle. Hammersley and Welsh (1965) also studied the problem of the convergence of N_n/n, and they discovered a remarkable connection between this convergence problem and the smoothness of the function $\mu(F)$ under certain perturbations of F. More precisely, their idea was to relate their original percolation problem with edge weights $\{x(e)\}$ to a new percolation problem with edge weights $x'(e) = x(e) + t$ for some small $t \in \mathbb{R}$.

If we use $F \oplus t = F(x - t)$ to denote the distribution of $x(e) + t$ and write $\mu(F \oplus t)$ to denote the corresponding time constant, then Hammersley and Welsh [(1965), page 101] proved that $\mu(F \oplus t)$ is a concave function of t on any open interval I where $\mu(F \oplus t)$ is finite. As a consequence of this concavity, one sees that the left derivative $D^-\mu(F \oplus t)$ and the right derivative $D^+\mu(F \oplus t)$ both exist for all $t \in I$. More notably, Hammersley and Welsh discovered that in some important cases that the convergence of N_n/n would follow if one could show the smoothness of $\mu(F \oplus t)$ as a function of t. This convergence criterion was subsequently refined by Smythe and Wierman [(1978), pages 129–130] and Kesten (1980) who proved that if $F(0) < \frac{1}{2}$ then with probability one we have

$$D^+\mu(F \oplus t)\big|_{t=0} \le \liminf_{n\to\infty} N_n/n \le \limsup_{n\to\infty} N_n/n \le D^-\mu(F \oplus t)\big|_{t=0}. \tag{1.4}$$

This result naturally implies that N_n/n must converge with probability one if the time constant $\mu(F \oplus t)$ is differentiable at $t = 0$.

For many years, the convergence criterion of Hammersley and Welsh has offered a tantalizing approach to the asymptotic behavior of N_n, but even long-standing prospects may prove illusory. The main result obtained here reveals that the Hammersley–Welsh criterion faces a fundamental limitation. We will show that the function $\phi(t) = \mu(F \oplus t)$ *fails* to be differentiable at zero for the most interesting choices of the edge weight distribution F.

THEOREM 1 (Nondifferentiability of the time constant). *There exist constants $\delta > 0$ and $\rho > 1$ such that for all $\frac{1}{2} - \delta \le p < \frac{1}{2}$, the Bernoulli percolation with $F(0) = p$ satisfies*

$$D^-\mu(F \oplus t)\big|_{t=0} \ge \rho D^+\mu(F \oplus t)\big|_{t=0}. \tag{1.5}$$

Moreover, $D^+\mu(F \oplus t) \ge 1$ for all $t \ge 0$, so the function $\phi(t) = \mu(F \oplus t)$ is not differentiable at zero.

For Bernoulli percolation one often writes $\mu(p)$ in place of $\mu(F)$, and one should take care not to misread Theorem 1 as an assertion about the nondifferentiability of $\mu(p)$ as a function of p. The theorem rather addresses the nondifferentiability of the time constant $\phi(t) = \mu(F \oplus t)$ of the t-shifted Bernoulli distribution with edge probability parameter p. Specifically, it tells us that for subcritical p close to one-half the t-shifted Bernoulli time constants are not differentiable as a function of the shift size t. This is precisely the type of nondifferentiability that one needs in order to show that the Hammersley–Welsh program for proving the convergence of N_n/n cannot be completed.

Organization of the arguments. The proof of Theorem 1 requires the development of several tools that may be useful for other problems of two-dimensional first-passage percolation. Our main technical results, Theorems 2 and 3, tell us that one has considerable flexibility in the choice of a minimal cost path. The first of these shows how one can find a large number of opportunities for "surgeries" that change the length of a path, and the second shows how such surgeries can be used in near-critical, subcritical Bernoulli percolation to find minimum cost paths that have greatly different lengths. Once Theorem 3 is obtained, the nondifferentiability theorem can be proved by easy estimates of the difference quotient inequalities that one finds from an elementary optimality argument.

The next two sections develop technical results that permit us to restrict our attention to paths that have minimal cost among the paths that are confined to stay within certain rectangles, rather than paths that may wander all over $\mathbb{Z}^2$. Section 3 then develops geometric features of the dual lattice that lead to an essential device for exploiting the independence of the edges above and below a fixed path, while Section 4 shows how a theorem of Turán from graph theory can be used to find a large number of disjoint boxes along any path.

Our main structural arguments are then given in Sections 5 and 6. The first of these helps us see that one can either perform a "surgery-to-lengthen" or a "surgery-to-shorten" many times on almost any path. These arguments are largely combinatorial (or topological). Section 6 then brings probability back into play and assembles all of the pieces that are needed in order for us to complete the proof of Theorem 3 in Section 7.

2. Cylinder variables. For any integers $0 \le m < n$ and any $h \in \mathbb{R}^+ \cup \{\infty\}$, we let $\mathcal{S}(m, n, h)$ denote the set of all paths from the point $(m, 0)$ to the point $(n, 0)$ with edges that are contained in the open real rectangle $(m, n) \times (-h, h)$. If we let

$$t_{m,n}(h) = \inf\{\tau(\gamma) : \gamma \in \mathcal{S}(m, n, h)\},$$

then the process $\{t_{m,n}(h) : 0 \le m < n < \infty\}$ is a natural "rectangular" analog of the point-to-point passage time process $\{a_{n,m}\}$. As before, one can easily check that $\{t_{m,n}(h)\}$ is a subadditive process, and the next lemma confirms that if h goes to infinity linearly with n then the process $\{t_{m,n}(h)\}$ behaves much like $\{a_{n,m}\}$. In this lemma (and subsequently), we write $a_{n,m}(F)$ or $t_{m,n}(h; F)$ whenever there is reason to emphasize the dependence of these processes on the underlying edge weight distribution F.

LEMMA 1. *If the edge weights $\{x(e)\}$ are nonnegative and have a distribution F with a finite mean, then for all $\alpha > 0$ we have*

$$\mu(F) \stackrel{def}{=} \lim_{n\to\infty} a_{0,n}(F)/n = \lim_{n\to\infty} t_{0,n}(\alpha n; F)/n,$$

where the convergence takes place almost surely and in L^1.

PROOF. The proof of the lemma combines a general subadditivity argument with a result of Smythe and Wierman that covers the case of $h = \infty$. Specifically, we will use the fact from Smythe and Wierman [(1978), page 79] that

$$\lim_{n\to\infty} t_{0,n}(\infty)/n = \mu(F) \qquad \text{a.s. and in } L^1. \tag{2.1}$$

First, from the limit (2.1), we see that for any $\varepsilon > 0$ there is an N such that

$$E[t_{0,N}(\infty)] \le N\big(\mu(F) + \varepsilon\big).$$

As $k \to \infty$ the bounded random variables $t_{0,N}(k)$ converge to $t_{0,N}(\infty)$, and they are dominated by the integrable random variable $t_{0,N}(1)$, so we can choose a K such that

$$E[t_{0,N}(K)] \le N\big(\mu(F) + 2\varepsilon\big). \tag{2.2}$$

Now, by subadditivity we have

$$t_{0,n}(K) \le \sum_{1\le j\le \lfloor n/N\rfloor} t_{(j-1)N,jN}(K) + t_{\lfloor n/N\rfloor N,n}(K), \tag{2.3}$$

so the mean bound (2.2), the law of large numbers and the Borel–Cantelli lemma permit us to deduce that

$$\limsup_{n\to\infty} t_{0,n}(K)/n \le \mu(F) + 2\varepsilon \qquad \text{a.s.} \tag{2.4}$$

Also, suboptimality gives us the bounds

$$t_{0,n}(\infty) \le t_{0,n}(\alpha n) \le t_{0,n}(K) \qquad \text{for all } n \ge K/\alpha, \tag{2.5}$$

so, if we use (2.1) to estimate the lower bound and use (2.4) and to estimate the upper bound, we have

$$\mu(F) \le \liminf_{n\to\infty} t_{0,n}(\alpha n)/n \le \limsup_{n\to\infty} t_{0,n}(\alpha n)/n \le \mu(F) + 2\varepsilon \qquad \text{a.s.}$$

Since $\varepsilon > 0$ is arbitrary, the last inequality gives us the required almost sure convergence. Finally, $0 \le t_{0,n}(\alpha n) \le t_{0,n}(1)$ and $t_{0,n}(1)$ is simply a sum of n i.i.d. random variables with finite mean. For such sums, the set of random variables $\{t_{0,n}(1)/n : 1 \le n < \infty\}$ is well known to be uniformly integrable. The collection $\{t_{0,n}(\alpha n)/n : 1 \le n < \infty\}$ is therefore also uniformly integrable, so the L^1 convergence follows from the almost sure convergence. □

The preceding lemma only deals with nonnegative edge weights, but we also need some information on $t_{0,n}(\alpha n; G)$ with a general edge weight distribution G. The hypotheses of Lemma 1 can be relaxed slightly, but such relaxations greatly complicate the proof. Fortunately, we can scrape along with the modest observation that for any G with finite mean and for any positive α we have

$$\mu(G) \le \liminf_{n\to\infty} t_{0,n}(\alpha n; G)/n \qquad \text{a.s.,} \tag{2.6}$$

a fact that follows immediately from the suboptimality bound $a_{0,n} \le t_{0,n}(\alpha n)$.

For Bernoulli percolation, one easily obtains a much more precise understanding of $t_{0,n}(\alpha n)$. As the next lemma shows, a generic subadditivity argument is good enough to give us a useful exponential bound on its upper tail.

LEMMA 2. *For Bernoulli percolation with parameter* $0 < p < 1$ *and for any choice of* $\varepsilon > 0$ *and* $\alpha > 0$, *there exist positive constants* $C_0 = C_0(\varepsilon, p, \alpha) > 0$ *and* $C_1 = C_1(\varepsilon, p, \alpha) > 0$ *such that*

$$P\big(t_{0,n}(\alpha n) \geq n\big(\mu(p) + \varepsilon\big)\big) \leq C_0 \exp(-C_1 n) \qquad \text{for all } n \geq 1. \tag{2.7}$$

PROOF. The subadditivity inequality (2.3) holds for all natural K and N, so, if we let $K = \lfloor \alpha n \rfloor$, we find that for all n and N we have

$$\begin{aligned} t_{0,n}(\alpha n) &\leq \sum_{1 \leq j \leq \lfloor n/N \rfloor} t_{(j-1)N, jN}(\alpha n) + t_{\lfloor n/N \rfloor N, n}(\alpha n) \\ &\leq N + \sum_{1 \leq j \leq m} Z_j, \end{aligned}$$

where $m = \lfloor n/N \rfloor$ and $Z_j = t_{(j-1)N, jN}(\alpha n)$. The $\{Z_j : 1 \leq j \leq m\}$ are independent, and they satisfy $|Z_j - E(Z_j)| \leq N$, so, by the large deviation inequality for bounded random variables [say, as given by Bennett (1962), Equation 8b], one has for all $\lambda > 0$ that

$$P\big(Z_1 + Z_2 + \cdots + Z_m - mE(Z_1) \geq \lambda\big) \leq \exp(-\lambda^2/2mN^2). \tag{2.8}$$

By Lemma 1, we can choose N so that $E(Z_j) = E[t_{0,N}(\alpha N)]$ is bounded above by $N(\mu(p) + \varepsilon/2)$, so for all m such that $m\varepsilon \geq 4$, we have

$$\begin{aligned} P\big(t_{0,n}(\alpha n) \geq n\big(\mu(p) + \varepsilon\big)\big) &\leq P\big(Z_1 + Z_2 + \cdots + Z_m \geq n\big(\mu(p) + \varepsilon\big) - N\big) \\ &\leq P\big(Z_1 + Z_2 + \cdots + Z_m - mE(Z_1) \geq mN\varepsilon/2 - N\big) \\ &\leq \exp\big(-\tfrac{1}{2}m(\varepsilon/4)^2\big). \end{aligned}$$

Since $m \geq n/N - 1$ we have $m\varepsilon > 4$ for large n, and the last inequality gives us our bound (2.7); to cover the smaller values of n, one then just increases C_0. We will not need the explicit values here, but one can check that $C_0 = e$ and $C_1 = \exp(\varepsilon^2/64N)$ will suffice for all $n \geq 1$. □

Tail bounds for the maximum deviation. Another random variable that will help us restrict out attention to well-behaved paths is given by

$$H_n(k) = \max\big\{|y| : (x, y) \text{ is a vertex of } \gamma \in \mathcal{S}(0, n, k) \text{ and } \tau(\gamma) = t_{0,n}(k)\big\},$$

so, $H_n(k)$ is the maximum deviation from the x-axis over all minimum cost paths in $\mathcal{S}(0, n, k)$. To estimate the tail probabilities for $H_n(k)$, first consider the random variable $b'_{0,n}$ that we define to be the infimum of the cost $\tau(\gamma)$ over all paths

in $\mathbb{Z}^2$ from the vertex $(0,0)$ to any point on the line $\{(x,n) \in \mathbb{Z}^2 : x \in \mathbb{Z}\}$. The behavior of this random variable is already well understood, and, after a rotation of coordinates, the results of Grimmett and Kesten [(1984), pages 343–344] tell us that for Bernoulli percolation with parameter p and for any $\varepsilon > 0$ that there exist nonnegative constants $C_0 = C_0(p,\varepsilon)$ and $C_1 = C_1(p,\varepsilon)$ such that

$$P\big(b'_{0,n} \le n\big(\mu(p) - \varepsilon\big)\big) \le C_0 \exp(-C_1 n) \qquad \text{for all } n \ge 1. \tag{2.9}$$

Given this bound, one can get a useful estimate for the tail probabilities of $H_n(k)$ just by looking at the elementary geometry of paths. The next lemma puts this observation into the form that will be used later.

LEMMA 3. *For subcritical Bernoulli percolation with parameter p, there exist positive constants $C_0 = C_0(p)$ and $C_1 = C_1(p)$ such that*

$$P\big(H_n(3n) \ge 2n\big) \le C_0 \exp(-C_1 n) \qquad \textit{for all } n \ge 1.$$

PROOF. The first observation is that for every

$$\omega \in \big\{t_{0,n}(3n) \le n(3\mu(p)/2),\ H_n(3n) \ge 2n\big\}$$

there exists a path with costs bounded by $n(3\mu(p)/2)$ that goes from $(0,0)$ to the line segment $\{(x,2n) : x \in (0,n)\}$ or to the line segment $\{(x,-2n) : x \in (0,n)\}$. Since the quantities $b'_{0,2n}$ and $b'_{0,-2n}$ defined above have the same distribution, we therefore find

$$\begin{aligned} P\big(t_{0,n}(3n) \le n\big(3\mu(p)/2\big), H_n(3n) \ge 2n\big) &\le 2P\big(b'_{0,2n} \le n\big(3\mu(p)/2\big)\big) \\ &= 2P\big(b'_{0,2n} \le 2n\big(\mu(p) - \varepsilon\big)\big), \end{aligned}$$

where $\varepsilon = \mu(p)/4$. The last probability has an exponential bound given by the Grimmett–Kesten inequality (2.9), and by Lemma 2 we also have an exponential bound on $P(t_{0,n}(3n) \ge n(3\mu(p)/2))$. Together these bounds complete the proof of the lemma. $\square$

3. Grounding paths in the dual lattice. We now let $\mathbb{Z}^{*2}$ denote the dual lattice of $\mathbb{Z}^2$ and we view $\mathbb{Z}^{*2}$ as a graph with vertex set

$$V = \big\{v : v = w + (1/2, 1/2) \text{ with } w \in \mathbb{Z}^2\big\}$$

and with edge set consisting of all pairs of vertices (u,v) such that $\|u - v\| = 1$. For any subset A of edges in $\mathbb{Z}^2$, we let A^* denote the subset of edges of $\mathbb{Z}^{*2}$ that meet A, and for any edge e of $\mathbb{Z}^2$ we let e^* denote the unique edge of $\mathbb{Z}^{*2}$ that meets e. Finally, we define the cost $x(e^*)$ of the dual edge e^* by setting $x(e^*) = x(e)$.

Next, we let ∂_n denote the piecewise linear curve that starts at $(0,0)$ and that subsequently visits the points $(0,-3n)$, $(n,-3n)$, and $(n,0)$ in that order; equivalently, ∂_n consist of the boundary of the box $[0,n]\times[0,-3n]$ minus the points on the open segment $(0,n)\times\{0\}$. For each $\gamma\in\mathcal{S}(0,n,3n)$, the set $\partial_n\cup\gamma$ is a simple closed curve of $\mathbb{R}^2$, and we let $\mathrm{int}(\partial_n\cup\gamma)$ denote the open subset of $\mathbb{R}^2$ bounded by $\partial_n\cup\gamma$. For each edge $e\in\gamma$, the dual edge e^* has exactly one vertex in $\mathrm{int}(\partial_n\cup\gamma)$, and we let $v(e^*)$ denote that vertex.

DEFINITION 1 [The event $G(\gamma)$]. For each $\gamma\in\mathcal{S}(0,n,3n)$, we define the event $G(\gamma)$ to be the set of all ω such that for each $e\in\gamma$ there exists a path $\tilde{\gamma}$ in the dual lattice $\mathbb{Z}^{*2}$ with the following four properties:

1. $\tilde{\gamma}$ starts at the vertex $v(e^*)$ in $\mathrm{int}(\partial_n\cup\gamma)$,
2. the last edge of $\tilde{\gamma}$ is an edge of ∂_n^*,
3. every edge of $\tilde{\gamma}$ except its last edge is contained in the open subset of $\mathbb{R}^2$ given by $\mathrm{int}(\partial_n\cup\gamma)$, and
4. one has $x(f)=1$ for all $f\in\tilde{\gamma}$, except possibly for the last edge of $\tilde{\gamma}$.

The $G(\gamma)$ as a "covering partition." The events $G(\gamma)$ with $\gamma\in\mathcal{S}(0,n,3n)$ do not quite form a covering partition of the sample space, but in some ways they come close. Specifically, Lemma 4 tells us that they have a useful coverage property, and Lemma 5 tells us that the $G(\gamma)$ are disjoint subject to a certain natural restriction.

LEMMA 4. *For any integers $0\le i<\infty$ and $1\le n<\infty$, we have*

$$\{t_{0,n}(3n)=i,\,H_n(3n)<2n\}\subset\bigcup_{\gamma\in\mathcal{S}(0,n,2n)}\{\tau(\gamma)=i\}\cap\{t_{0,n}(3n)=i\}\cap G(\gamma).$$

The proof of this lemma requires some understanding of the topology of the *closed* paths in $\mathbb{Z}^2$. In particular, we need a proposition from Kesten (1982) that tells us about the structure of a closed path α that is made up out of four arcs (or subpaths). To be explicit, we recall that an *arc* is like a path in that it consists of an alternating sequence of vertices and connecting edges, but, unlike a path, an arc is not required to begin with a vertex or to end with a vertex. In particular, an arc can be a single vertex, a single edge, or any contiguous part of a path—such as a path minus its two end points.

Now consider a closed path α in $\mathbb{Z}^2$ that is made up of four arcs α_1, α_2, α_3 and α_4 that one meets in clockwise order as one traverses α, and let $\mathrm{int}(\alpha)$ denote the bounded open subset of $\mathbb{R}^2$ that has α as its boundary. Proposition 2.2 of Kesten (1982) tells us that if α_1 and α_3 each contain at least one vertex of $\mathbb{Z}^2$, and, if ω is any configuration of zero–one edge weights, then one of the two following assertions must hold.

Direct assertion. There exists a path β in $\mathbb{Z}^2$ such that each edge of β is in $\text{int}(\alpha)$ and such that (1) γ starts at a vertex of α_1, (2) β ends at a vertex of α_3, and (3) for each edge $e \in \beta$, one has $x(e) = 0$.

Dual assertion. There exists a path β in $\mathbb{Z}^{*2}$ such that each edge of β except the first and the last is in $\text{int}(\alpha)$ and such that (1) β starts with an edge of α_2^*, (2) β ends with an edge of α_4^*, and (3) for each edge $f \in \beta$ (except possibly for the first edge or the last edge of β), one has $x(f) = 1$.

To prove Lemma 4, we just need to show for any $\omega \in \{t_{0,n}(3n) = i, H_n(3n) < 2n\}$ we can find a $\gamma \in \mathcal{S}(0, n, 2n)$ such that $\tau(\gamma) = i$ and $\omega \in G(\gamma)$. We will produce such a γ by an algorithm that creates a finite sequence $\gamma_0, \gamma_1, \ldots, \gamma_k$ of candidates. To begin the algorithm, we note that the argument of Lemma 3 tells us that for any $\omega \in \{t_{0,n}(3n) = i, H_n(3n) < 2n\}$ there is a $\gamma_0 \in \mathcal{S}(0, n, 2n)$ such that

$$\tau(\gamma_0) = i \quad \text{and} \quad \gamma_0 \in \mathcal{S}(0, n, 2n).$$

If $\omega \in G(\gamma_0)$, then γ_0 gives us the γ that we need in order to show that ω is an element of the union in Lemma 4. In this lucky case, the proof of the lemma is therefore complete.

On the other hand, if $\omega \notin G(\gamma_0)$, we need a more substantial argument. In this case, the definition of $G(\gamma_0)$ tells us that there exists an edge $e \in \gamma_0$ such that every path $\tilde{\gamma}$ in $\text{int}(\partial_n \cup \gamma_0)$ from the starting vertex $v(e^*)$ to ∂_n^* must have some edge $f \in \tilde{\gamma}$ other than its last edge for which $x(f) = 0$. Moreover, we can also assume without loss of generality that the candidate edge $e \in \gamma_0$ is not the first edge or the last edge of γ_0, since each of these special edges has a trivial length-two dual path to ∂_n^* that satisfies properties (1)–(4) required by the definition of $G(\gamma_0)$.

Now, with our candidate edge e in hand, we are ready to construct our new candidate path γ_1. First, we consider four arcs that are defined as follows:

- α_1 is the arc that follows γ_0, beginning with the first edge of γ_0 and ending with the first vertex of e,
- α_2 is the arc that consists of just the single edge e,
- α_3 is the arc given by the subpath of γ_0 that begins with the second vertex of e and ends with the edge from $(n-1, 0)$ to $(n, 0)$, and
- α_4 is the arc that begins with the vertex $(n, 0)$, follows the arc defined by ∂_n and ends with the vertex $(0, 0)$.

Kesten's proposition now tells us that there exists a path β with all of its edges in $\text{int}(\partial_n \cup \gamma)$ that begins with a vertex u of α_1 and ends with a vertex v of α_3 for which one has $x(f) = 0$ for every edge f in β. Now we define a new path $\gamma_1 \in \mathcal{S}(0, n, 3n)$ by following γ_0 from $(0, 0)$ to the vertex u, following β from u to v, and following γ_0 from v to $(n, 0)$. Since $x(f) = 0$ for each $f \in \beta$, our new path satisfies $\tau(\gamma_1) = \tau(\gamma_0) = i$, and we see that γ_1 is again a path of minimal cost.

Also, since we have $\omega \in \{H_n(3n) < 2n\}$, the argument of Lemma 3 again tells us that γ_1 is in fact an element of $\mathcal{S}(0, n, 2n)$.

If $\omega \in G(\gamma_1)$, then we can take $\gamma = \gamma_1$ to complete our construction, but if $\omega \notin G(\gamma_1)$, we need to repeat the preceding process to define another path γ_2. In general, for $i = 0, 1, \ldots$ so long as $\omega \notin G(\gamma_i)$, we simply repeat our process to define a subsequent path γ_{i+1}. Since the number of edges enclosed by $\partial_n \cup \gamma_{i+1}$ is strictly less than the number of edges enclosed by $\partial_n \cup \gamma_i$, we see that after a finite number of steps we must arrive at a k such that $\omega \in G(\gamma_k)$. Finally, once such a k is found, we can set $\gamma = \gamma_k$ to complete the proof of Lemma 4 just as we have done twice before.

The $G(\gamma)$ and independence. For any fixed $\gamma \in \mathcal{S}(0, n, 2n)$, we let $\text{ABOVE}(\gamma)$ denote the set of edges of the $\mathbb{Z}^2$ lattice that are in the interior of the real cylinder $[0, n] \times \mathbb{R}$ that are strictly above γ and we let $\text{BELOW}(\gamma)$ denote those that are strictly below γ. The key to the constructions that we will use later is the fact that for a *fixed* γ the event $G(\gamma)$ is measurable with respect to the σ-field $\sigma\{x(e) : e \in \text{BELOW}(\gamma)\}$. Consequently, the event $G(\gamma)$ is independent of the cost $\tau(\gamma)$ and independent of any event that is measurable with respect to the σ-field $\sigma\{x(e) : e \in \text{ABOVE}(\gamma)\}$.

The $G(\gamma)$ and restricted disjointness.

LEMMA 5. *The collection of events $A(\gamma) = G(\gamma) \cap \{\omega : \tau(\gamma) = 0\}$ indexed by $\gamma \in \mathcal{S}(0, n, 2n)$ is a collection of disjoint events, and, consequently, we have*

$$\sum_{\gamma \in \mathcal{S}(0,n,2n)} P\big(\tau(\gamma) = 0, G(\gamma)\big) \leq 1.$$

PROOF. For any pair of unequal paths γ and γ' in $\mathcal{S}(0, n, 2n)$, one either has an $e \in \gamma$ and $e \in \text{ABOVE}(\gamma')$, or one has an $e \in \gamma'$ and $e \in \text{ABOVE}(\gamma)$. There is no loss of generality if we assume that the first case holds.

Now, if $\omega \in A(\gamma)$, the definition of $G(\gamma)$ tells us there is a path β in $\mathbb{Z}^{2*}$ from e^* to ∂_n^* such that $x(f) = 1$ for all $f \in \beta$, except possibly for the edge of β that is in ∂_n^*. Also, since we assume that $e \in \text{ABOVE}(\gamma')$, the Jordan curve theorem tells us that the path β must cross γ' someplace, and, since γ' does not meet any of the dual edges in ∂_n^*, we see that γ' must meet β at an edge of f that has cost one. This tells us γ' contains an edge with cost one, and since $\tau(\gamma') \neq 0$ implies $\omega \notin A(\gamma')$, the proof of the lemma is complete. □

Here we should note that an alternative proof of the Lemma 5 can be based on Proposition 2.3 of Kesten (1982) which establishes the uniqueness of certain paths called "lowest zero-crossings." There is no need to give the details of this alternative proof, but we should note that Lemma 5 is brought into range of

Kesten's Proposition 2.3 by padding the outside of ∂_n with edges that have cost one and by exploiting the uniqueness of the lowest zero-crossing between the outside vertices $(-1, 0)$ and $(n + 1, 0)$.

4. Disjoint boxes on a path. In our main argument, we will need to show that one can often perform a large number of local "surgeries" on the paths of $\mathcal{S}(0, n, 2n)$, and for these arguments to be effective we need to know that for any path γ one can find large number of disjoint boxes with an edge of γ at its "center." To make this notion precise, we need to distinguish between horizontal and vertical edges; specifically, given any horizontal edge $e = \{(x, y), (x + 1, y)\}$, we let

$$T(e) = [x - 2, x + 3] \times [y - 2, y + 2],$$

and for any vertical edge $e = \{(x, y), (x, y + 1)\}$ we take

$$T(e) = [x - 2, x + 2] \times [y - 2, y + 3].$$

In either case, we call $T(e)$ the *box centered at* e, and we regard $T(e)$ as a closed subset of $\mathbb{R}^2$. The specific dimensions of these boxes have been chosen to fit a design that will be described shortly. The problem now is to show that one can always find a set of disjoint centered boxes on γ that has cardinality of order $|\gamma|$.

One nice systematic way to show how one can find a large number of disjoint boxes on a path is to appeal to a bit of graph theory. For any graph (V, E), a set of vertices $A \subset V$ is said to be an *independent* set (in the graph theoretical sense) provided that there is no edge of (V, E) that joints two elements of A. The *independence number* $\alpha(V, E)$ of the graph (V, E) is then defined to be the maximum cardinality of any such independent set. One of the most basic facts about $\alpha(V, E)$ is Turán theorem, which says that if the graph has maximal degree Δ then

$$\alpha(V, E) \geq \frac{|V|}{\Delta + 1}. \tag{4.1}$$

This result and many other versions of Turán theorem are covered in the instructive survey of Aigner (1995). Turán's theorem gives us a tidy way to find a large number of disjoint boxes on a path γ; we just need to define the right graph.

LEMMA 6 (The α_0 lemma). *There are constants $\alpha_0 > 0$ and $N_0 < \infty$ such that for all $n \geq N_0$ and all $\gamma \in \mathcal{S}(0, n, 2n)$ there exists a set of k edges $e_1, e_2, \ldots, e_k$ of γ with $k \geq \alpha_0|\gamma|$ such that*

$$T(e_i) \cap T(e_j) = \varnothing \qquad \text{for all } 1 \leq i < j \leq k$$

and

$$T(e_i) \subset [1, n - 1] \times [-2n - 3, 2n + 3].$$

PROOF. To set up Turán's theorem, we take the vertex set V to be the set

$$V = \{T(e) : e \in \gamma, T(e) \subset [1, n-1] \times [-2n-3, 2n+3]\}.$$

The cardinality of this vertex set satisfies $|V| \geq n - 6$ simply because γ goes from $(0, 0)$ to $(n, 0)$. Next we note that V also satisfies $|V| \geq |\gamma| - 100n$, since there are fewer than $100n$ horizontal and vertical edges of the $\mathbb{Z}^2$ lattice in the sets $[0, 2] \times [-2n, 2n]$ and $[n-2, n] \times [-2n, 2n]$. These two bounds and the constraint $n \geq 3000$ imply that

$$|V| \geq |\gamma|/200. \tag{4.2}$$

Next, we consider the graph with vertex set V and with edge E that we define to be the set of all the (unordered) pairs of elements of V such that $T(e_j) \cap T(e_k) \neq \varnothing$. From the fact that each of the $T(e)$ contains exactly 30 lattice points and each of these can be a "corner point" of at most four of the neighbors of $T(e)$ in the graph (V, E), so the maximal degree of (V, E) is certainly not greater than $4 \cdot 30$. We can then conclude by Turán's theorem that there exists an independent set of elements of V with cardinality that is at least $|V|/(4 \cdot 30 + 1)$, so by our bound (4.2) on $|V|$, we see that the proof of the lemma is complete and that we have $N_0 \leq 3000$ and $\alpha_0 \geq 1/200 \cdot (4 \cdot 30 + 1)$. □

5. Surgeries that lengthen or shorten paths. Our argument pivots on the possibility of performing a large number of surgeries that alter the length of a path and leave the cost of the path unchanged. To make this notion precise, we need formal definitions of a *surgery-to-lengthen* and a *surgery-to-shorten*. We begin with the most natural of these, the surgery-to-lengthen.

DEFINITION 2 (Surgery-to-lengthen). If $e \in \gamma$ and $\gamma \in \mathcal{S}(0, n, 2n)$, we say that there exists a *surgery-to-lengthen* γ in $T(e)$ if there exists a set A of edges $e'_1, e'_2, \ldots, e'_s$ in the set $\text{ABOVE}(\gamma) \cap T(e)$ and there exists a set D of edges $e_1, e_2, \ldots, e_s$ in $\gamma \cap T(e)$ such that if one deletes the edges of D from γ and adjoins the edges of A one obtains a path $\gamma' \in \mathcal{S}(0, n, 3n)$ with $|\gamma'| > |\gamma|$.

The definition of a *surgery-to-shorten* is similar, but there is a small difference that has some important consequences. For a surgery-to-shorten we no longer require the set of deleted edges to be contained in $T(e)$. This means that a surgery-to-shorten is not local like a surgery-to-lengthen; the passage from γ to γ' in a *surgery-to-shorten* may rip out parts of γ that appear almost anywhere in the rectangle $[0, n] \times (-2n, 2n)$. The formal definition of a surgery-to-shorten almost looks redundant, but, given the physical gap between the two types of surgery, it seems prudent to be explicit.

DEFINITION 3 (Surgery-to-shorten). If $e \in \gamma$ and $\gamma \in \mathcal{S}(0, n, 2n)$, we say that there exists a *surgery-to-shorten* γ in $T(e)$ if there exists a set A of edges $e'_1, e'_2,$

$\ldots, e'_s$ in the set $\text{ABOVE}(\gamma) \cap T(e)$ and there exist some set D of edges $e_1, e_2, \ldots, e_s$ in γ such that if one deletes the edges of D from γ and adjoins the edges of A one obtains a $\gamma' \in \mathcal{S}(0, n, 3n)$ with $|\gamma'| < |\gamma|$.

The most important feature of this pair of surgical operations is that either one or the other is (almost) always available to us.

THEOREM 2 (Surgery theorem). *For any path $\gamma \in \mathcal{S}(0, n, 2n)$ and any $e \in \gamma$ such that $T(e) \subset [1, n-1] \times (-2n+3, 2n-3)$, the path γ either has a surgery-to-lengthen in $T(e)$ or has a surgery-to-shorten in $T(e)$.*

The proof of Theorem 2 depends on the examination on number of cases, but the following simple lemma gives us a way to deal quickly with many of these.

LEMMA 7 (Short paths from close-by points). *Suppose the vertices v and w are both on the path γ and $\|v - w\| = 1$, but the edge (v, w) is not in γ. If $(v, w) \in \text{ABOVE}(\gamma) \cap T(e)$, then there is a surgery-to-shorten γ in $T(e)$.*

PROOF. Since v and w are on γ but (v, w) is not and edge of γ, the number of edges on γ from v to w is at least three. Thus, the path γ' defined by following γ to v, taking the new edge (v, w) and then following γ from w to $(0, n)$ has at least two fewer edges than γ. Since we assume that $(v, w) \in \text{ABOVE}(\gamma) \cap T(e)$, we therefore meet the definition of a surgery-to-shorten. □

Boxes and cases of boxes: *Proof of Theorem* 2. First consider a horizontal edge e and its associated box $T(e)$. If $e \in \gamma$ and $\gamma \in S(0, n, 2n)$, we view $e = (u, v)$ as an ordered pair where u is the first vertex on γ as one goes from $(0, 0)$ to $(n, 0)$. When $T(e) \subset [1, n-1] \times [-2n-2, 2n+2]$, there must be edges e_- and e_+ that precede and succeed e on γ, and each of these edges can have three possible orientations. Thus, if we fix e to be a horizontal edge, there are nine cases that we need to consider. One then needs another set of nine cases to cover the situation when the center edge is vertical edges, but, by symmetry, we will only need to consider the case of horizontal edges.

The left-hand column of Figure 1 gives four of the nine cases where the center edge is horizontal, and the left-hand column of Figure 2 gives two more cases. By the left-right asymmetry of cases 4–6 given in Figures 1 and 2, we see that the proof of Theorem 2 will be complete if we show that in each of the six listed cases we always have either a surgery-to-shorten or a surgery-to-lengthen.

In Figures 1 and 2, the center edge e is labeled e (logically enough), and when the endpoints of e are needed they are labeled u and v. The edge of γ preceding e

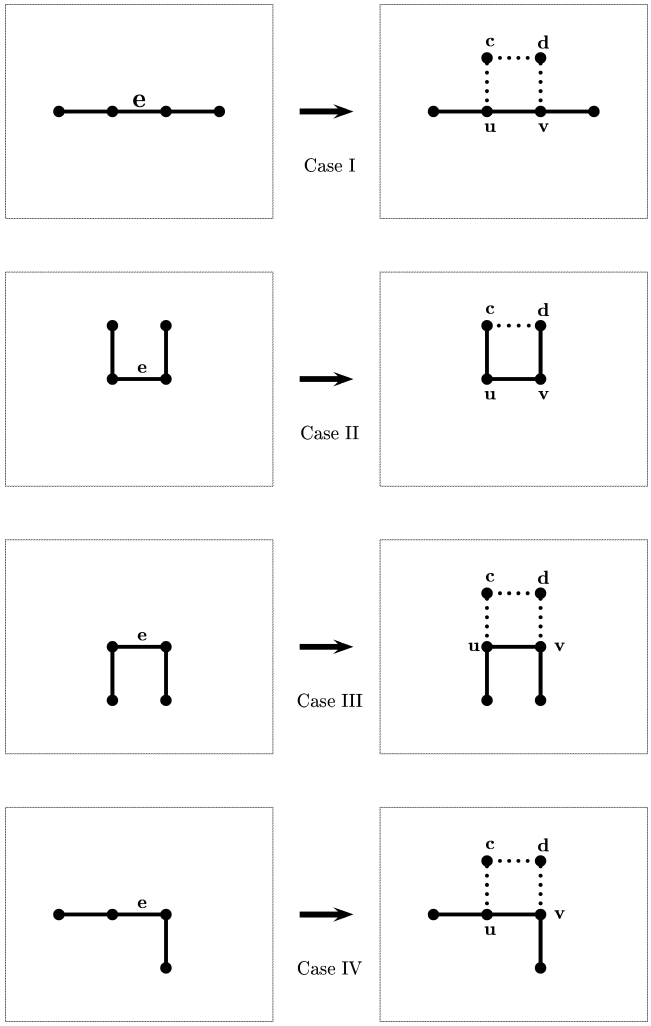

FIG. 1. *The first four cases.*

is denoted by (a, u) and the one following e is denoted by (v, b). Only these three edges of γ are drawn in Figures 1 and 2, but there will be other edges of γ at other places in the box that are not drawn. The role of the undrawn edges will be made explicit as they are met in the context of our argument.

The interpretation of the dotted lines and other labeled points of Figures 1 and 2 will also become evident as we investigate the individual cases, but we should comment on the location of the *edge symbol* e. We know that the edge $e = (u, v)$ has the set $\text{ABOVE}(\gamma)$ on one side and the set $\text{BELOW}(\gamma)$ on the other. We use the graphical convention of placing the symbol e on the side of the edge

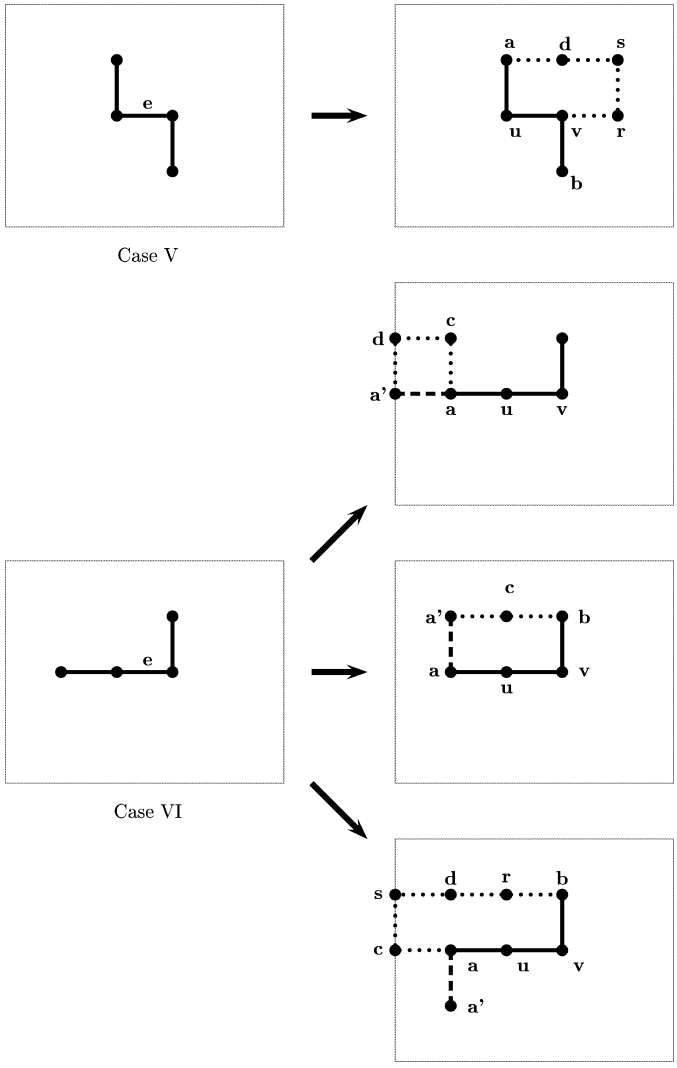

FIG. 2. *Cases five and six.*

$e = (u, v)$ where one finds ABOVE(γ), and this convention is important for the enumeration of the cases that we must consider. Our enumeration is designed to permit us to draw the symbol e directly above the edge $e = (u, v)$, and the difference between cases 2 and 3 depends precisely on the distinction enforced by this design. To minimize clutter, we only draw the edge symbol e in the first columns of Figures 1 and 2.

One useful way to think about the placement of the symbol e is to note that if there is a continuous path in $\mathbb{R}^2$ from the symbol e to the point z such that the path does not go through γ or ∂_n, then the point z is in ABOVE(γ). This intuitive test

for membership in $\text{ABOVE}(\gamma)$ may sound casual, but it is completely rigorous; it is nothing more or less than the Jordan curve theorem.

Case 1. In this case, the edges (a, u), (u, v) and (v, b) are all horizontal. To begin, we consider the vertex c directly above u. We know that the edge (u, c) is not in γ since γ cannot have a vertex with degree exceeding two. Also, the edge (u, c) is contained in $\text{ABOVE}(\gamma)$ because of the orientation indicated by the label e.

Now, if we have $c \in \gamma$, then Lemma 7 tell us that we have a surgery-to-shorten in $T(e)$. Thus, we may assume that $c \notin \gamma$, and, by the same argument, we may assume that the vertex d that is directly above v is also not an element of γ. These observations tell us that without loss of generality we may assume that the set of edges $A = \{(u, c), (c, d), (d, u)\}$ is contained $\text{ABOVE}(\gamma)$ and does not meet γ except at u and v. This makes a surgery-to-lengthen obvious. We simply delete e from γ and add the edges of A to find a new path that meets all the requirements of a surgery-to-lengthen in $T(e)$.

As a cautionary point, one should note here that even when we start out with three horizontal edges in $T(e)$ we have no guarantee that $T(e)$ permits a surgery-to-lengthen. Here, and in all subsequent cases, we simply argue that we can *either* find a surgery-to-lengthen *or* find a surgery-to-shorten. In every case, Lemma 7 is used to argue that if there is not a surgery-to-shorten, then we have enough room to make a surgery-to-lengthen.

Case 2. This is the only completely trivial case. The vertices c and d are on γ; they satisfy $\|c - d\| = 1$, and (u, v) is in $\text{ABOVE}(\gamma)$, so Lemma 7 gives us a surgery-to-shorten.

Case 3. As we mentioned earlier, this case serves to remind us that the location of the *symbol* e in the figure tells us the side of the *edge* $e = (u, v)$ where one finds $\text{ABOVE}(\gamma)$. Our enumeration lists the cases to be considered in a such way that we can always place the symbol e above the edge $e = (u, v)$. This symbol thus becomes part of the definition of the case, so, for example, if one were to draw the symbol e below the edge $e = (u, v)$ in the diagram for Case 2, one would find a diagram that is in fact equivalent to Case 3.

To deal with Case 3, we first note (as in the Case 1) that we may assume that c and d are not on γ, since otherwise Lemma 7 would provide a surgery-to-shorten. Also, there is a trivial path in $\mathbb{R}^2$ from the edge symbol e to the elements of $A = \{(u, c), (c, d), (d, u)\}$ so all of these edges are elements of $\text{ABOVE}(\gamma)$. By adding the edges of A and deleting the edge e, we find a surgery-to-lengthen.

Case 4. This case is safely skipped since it is completely parallel to Cases 1 and 3.

Case 5. This is an interesting case. We can argue as before that we may assume that the vertices d and r are in $\text{ABOVE}(\gamma)$ and not on γ, yet we need a new argument to show that we may assume that the indicated vertex s is likewise in $\text{ABOVE}(\gamma)$ and not on γ.

First, we note that the edge (v, d) is not in γ since v cannot have degree 3 in γ. We then see that there is a continuous path from the symbol e to the point s that does not meet γ or ∂_n, so by our earlier discussion of the symbol e, we see that $s \in \text{ABOVE}(\gamma)$, unless it happens that s is an element of γ.

Suppose to the contrary, that $s \in \gamma$. If s follows b as one traces γ from $(0, 0)$ to $(n, 0)$, we simply replace the path segment along γ from a to s with the edges (a, d) and (d, s). At least 6 edges have been cut out and only two added, so we have a surgery-to-shorten. On the other hand, if s precedes a on γ, then the path along γ from s to a must have length at least 4, and we may this path segment with (s, d) and (d, a) to get a surgery-to-shorten.

The only remaining possibility is that s is not on γ. In this case, we replace (a, u) and (u, v) with the path $a \to d \to s \to r \to v$, and we have a surgery-to-lengthen.

Case 6. As one would expect, the last case is the hardest. We can no longer drive the argument by just the orientation of three edges (a, u), (u, v) and (v, b). We also need to consider the orientation of the edge (a', a) that precedes (a, u) on γ. The three possibilities are broken out in Figure 2, and the three possible orientations of (a', a) are indicated by *dashed* edges.

In the first possibility, the edges $(a', a), (a, u)$ and (u, v) are all horizontal. This is exactly the same circumstance that we studied in Case 1. Here one should confirm that the construction used before can be used again without stepping out of the box, but this is an easy check.

In the second possibility, (a', a) goes north from a. Here we note as usual that we may assume that c is not on γ, or else we would have a surgery-to-shorten. As a consequence, we can take the edges (a, c) and (c, b) and delete the segment of γ given by $a' \to a \to u \to v \to b$. This gives us a surgery-to-shorten.

In the third possibility, (a', a) goes south from a as indicated by the dashed line at the bottom of Figure 2. We first observe that arguments we have given twice before permit us to assume that r and d are not on γ, or else Lemma 7 would provide a surgery-to-shorten. Now, since $(a, d) \notin \gamma$ there is a path in $\mathbb{R}^2$ from the edge symbol e to c that does not cross γ, so we see that $c \in \text{ABOVE}(\gamma)$. Lemma 7 therefore tells us that we may assume that $c \notin \gamma$.

The whole focus is now on s. Since $(a, d) \notin \gamma$, there is a path in $\mathbb{R}^2$ from the edge symbol e to s that does not cross γ, so we see that $s \in \text{ABOVE}(\gamma)$, unless we happen to have $s \in \gamma$. Assume for the moment that $s \in \gamma$. In this case, we need to ask if s comes before a' or after b on γ. If s comes before a', then we can remove the segment of γ from s to a and replace it with the edges (s, c) and (c, a). At least three edges are removed and only two are added, so we find a surgery-to-shorten.

On the other hand, if s comes after b on γ, then we replace the segment of γ from b to s with the edges (b, r), (r, d), and (d, s). At least five edges are dropped and only three are added, so again we find a surgery-to-shorten.

At last, we may assume that $s \notin \gamma$. In this case, we may replace the path segment $a \to u \to v \to b$ with the new segment $a \to c \to s \to d \to r \to b$. This alteration gives us a surgery-to-lengthen, and thus completes the analysis of Case 6 and the proof of Theorem 2.

6. Geometry of the long and short routes. The random variables at the heart of our analysis add two small twists to the minimum length variable N_n introduced by Hammersley and Welsh (1965). First, for technical reasons, we need to consider cylinder variables. Second, and more important, we consider both the longest *and* the shortest routes; specifically, we introduce

$$N_n^+ = \max\{|\gamma| : \tau(\gamma) = t_{0,n}(3n) \text{ for } \gamma \in \mathcal{S}(0, n, 3n)\}$$

and

$$N_n^- = \min\{|\gamma| : \tau(\gamma) = t_{0,n}(3n) \text{ for } \gamma \in \mathcal{S}(0, n, 3n)\}.$$

For us, the most important feature of these variables rests in the fact that we can prove that for certain subcritical values of p, the random variable N_n^+ is virtually guaranteed to be larger than N_n^- by a factor of at least $(1 + \varepsilon)$ for a fixed $\varepsilon > 0$. The next theorem makes this principle precise.

THEOREM 3 (Long and short routes). *There exists a constant $\delta > 0$ such that for Bernoulli percolation with $1/2 - \delta \le p < 1/2$, we have three constants $C_0 = C_0(p) > 0$, $C_1 = C_1(p) > 0$, and $\rho = \rho(p) > 1$ such that*

$$P(N_n^+ \le \rho N_n^-) \le C_0 \exp(-C_1 n) \qquad \textit{for all } n \ge 1.$$

We have already developed most of the facts needed to prove Theorem 3. What remains is a pleasant calculation that breaks naturally into four steps. The first step introduces a decomposition that sets up the exploitation of the disjointness property of the $G(\gamma)$.

STEP 1 (A decomposition). We first note that we have the trivial inclusion $\{N_n^+ \le \rho N_n^-\} \subset A_n \cup B_n \cup C_n$ where we take

$$A_n = \{N_n^+ \le \rho N_n^-, t_{0,n}(3n) \le 2\mu(p)n, H_n(3n) < 2n\}$$

and where we take

$$B_n = \{H_n(3n) \ge 2n\} \quad \text{and} \quad C_n = \{t_{0,n}(3n) \ge 2\mu(p)n\}.$$

Lemmas 3 and 2 provide exponential bounds for the events B_n and C_n, so, to prove Theorem 3, we only need to obtain an exponential upper bound on the event A_n. By Lemma 4 we also have the decomposition

$$\begin{aligned} &A_n \cap \{t_{0,n}(3n) = k\} \\ &\quad \subset \bigcup_{\gamma \in \mathcal{S}(0,n,2n)} \{N_n^+ \le \rho N_n^-, t_{0,n}(3n) = k, \tau(\gamma) = k\} \cap G(\gamma). \end{aligned} \tag{6.1}$$

STEP 2 (A surgery count). Our main task now is to estimate the probability of the union (6.1), and this is done most easily by introducing three new random variables $\nu_n^+(\gamma)$, $\nu_n^-(\gamma)$, and $\nu_n(\gamma)$. The random variable $\nu_n^+(\gamma)$ is defined to be the maximum value of k such that there exist k edges $e_1, e_2, \ldots, e_k$ of γ such that for each $1 \le i \le k$ one has the four properties:

1. $T(e_i) \cap T(e_j) = \varnothing$ if $i \ne j$,
2. $T(e_i) \subset [1, n-1] \times [-2n-2, 2n+2]$,
3. all edges in $T(e_i) \cap \text{ABOVE}(\gamma)$ have cost 0, and
4. γ has a surgery-to-lengthen in $T(e_i)$.

The random variable $\nu_n^-(\gamma)$ is then given by the same recipe as $\nu_n^+(\gamma)$, except that "surgery-to-lengthen" is replaced by "surgery-to-shorten," and finally, ν_n is simply taken to be the random variable one gets by requiring just the first three conditions (without any surgery requirements).

The surgery theorem tells us that every box that satisfies the first two conditions either has a surgery-to-lengthen or a surgery-to-shorten, so we have

$$\max\bigl(\nu_n^+(\gamma), \nu_n^-(\gamma)\bigr) \ge \nu_n(\gamma)/2. \tag{6.2}$$

One of the benefits of the random variable $\nu_n(\gamma)$ is that it is easily estimated from below in terms of a binomial random variable. We first just note that are at least $\alpha_0|\gamma|$ disjoint boxes on γ and that are contained $[1, n-1] \times [-2n-2, 2n+2]$ and for any such box there are certainly not more than 49 edges in $T(e) \cap \text{ABOVE}(\gamma)$. This says that $\nu_n(\gamma)$ may be stochastically bounded below by the sum of $\alpha_0|\gamma|$ independent Bernoulli random variables with parameter p^{49}. The large deviation bound for Bernoulli sums [say, as given by Hoeffding (1963), Theorem 2] then gives us

$$\delta \le \tfrac{1}{2} p^{49} \quad \Longrightarrow \quad P\bigl(\nu_n(\gamma) \le \delta \alpha_0 |\gamma|\bigr) \le \exp(-\alpha_0 p^{98} |\gamma|/8). \tag{6.3}$$

In our application of this bound, we will need to exploit the uniformity that holds when the range of p is restricted; the uniform estimate that we use is summarized in the next lemma.

LEMMA 8 (Good boxes on a path). *There are positive constants C_0 and C_1 that do not depend on either on $\varepsilon > 0$ or p such that for all n and all $\gamma \in \mathcal{S}(0, n, 2n)$, we have for all $n \geq 1$, all $1/4 \leq p \leq 1/2$ and all $0 < \varepsilon < p^{50}$ that*

$$P\big(\nu_n(\gamma) \leq \varepsilon|\gamma|\big) \leq C_0 \exp(-C_1|\gamma|).$$

STEP 3 (Estimation of the gap $N_n^+ - N_n^-$). We now know that we have many surgical opportunities on any path, but we still need to show that this leads us to an effective lower bound on the difference between N_n^+ and N_n^-.

LEMMA 9. *If we set $\varepsilon = \rho - 1 > 0$, then for any $\gamma \in \mathcal{S}(0, n, 3n)$ we have*

$$\begin{aligned}\big\{\omega : N_n^+ < \rho N_n^-, t_{0,n}(3n) = k, \tau(\gamma) = k\big\}& \\ \cap G(\gamma) \subset \big\{\omega : t_{0,n}(3n) = k, \tau(\gamma) = k, \nu_n(\gamma) \leq 2\varepsilon|\gamma|\big\} \cap G(\gamma)&.\end{aligned}$$

PROOF. For any ω in the first set, we have

$$N_n^+(\omega) - N_n^-(\omega) \leq \varepsilon N_n^-(\omega),$$

and we first claim that for all $\gamma \in \mathcal{S}(0, n, 3n)$ with $\tau(\gamma) = t_{0,n}(3n)$, we also have the bound

$$N_n^+(\omega) - N_n^-(\omega) \geq \max\big(\nu_n^+(\gamma), \nu_n^-(\gamma)\big) \geq \tfrac{1}{2}\nu_n(\gamma)(\omega). \tag{6.4}$$

To see why this is so, we first note that each surgery-to-lengthen takes place entirely with in a box and always moves us from a minimal cost path to another minimal cost path, so we have

$$N_n^+ \geq |\gamma| + \nu_n^+(\gamma) \geq N_n^- + \nu_n^+(\gamma). \tag{6.5}$$

The corresponding bound with $\nu_n^-(\gamma)$ is a bit more subtle since the surgery-to-shorten is no longer local. We do have the parallel inequality,

$$N_n^- \leq |\gamma| - \nu_n^-(\gamma) \leq N_n^+ - \nu_n^-(\gamma), \tag{6.6}$$

but the proof is more algorithmic. We know that there are $\nu_n^-(\gamma) = k$ disjoint boxes $T(e_1), T(e_2), \ldots, T(e_k)$ on γ that permit a surgery-to-shorten, so we take the first of these and perform the surgery. This surgery will shorten the path and move from a minimal cost path to a minimal cost path, but the surgery may also cut out a an undetermined number of the other boxes $T(e_2), T(e_3), \ldots, T(e_k)$ that are on the path γ.

What one needs to notice is that if m boxes are cut out by the first surgery, then our new path is shorter than γ by at least $m + 1$ edges (actually many more). If $m = k$, then our construction is complete, and otherwise the path γ' created by the first surgery has $k - m > 0$ boxes remaining that permit a surgery-to-shorten. In the second case, we go to the first of the remaining boxes perform another surgery-to-shorten. If one continues in this way, one obtains a path that has cost no greater

than the cost of γ and which as at least k fewer edges. This completes the proof of the first inequality of (6.6), and the second inequality is obvious.

Finally, when we put the bound (6.4) together with the hypothesis $N_n^+ < \rho N_n^-$, we see that

$$\tfrac{1}{2}\nu_n(\gamma)(\omega) \le \varepsilon N_n^-(\omega) \le \varepsilon|\gamma|,$$

and this is precisely the estimate one needs in order to say that ω is in the second set of the lemma. $\square$

STEP 4 (A final calculation). All of the elements are in place for the proof of Theorem 3. We first apply the decomposition (6.1), then use Lemma 9, Boole's inequality and independence to find

$$\begin{aligned}
P(A_n) &\le \sum_{k=0}^{2\mu(p)n} P\Bigg(\bigcup_{\gamma\in\mathcal{S}(0,n,2n)} \{N_n^+ \le \rho N_n^-,\ t_{0,n}(3n)=k, \tau(\gamma)=k,\ G(\gamma)\}\Bigg)\\
&\le \sum_{k=0}^{2\mu(p)n} P\Bigg(\bigcup_{\gamma\in\mathcal{S}(0,n,2n)} \{\nu_n(\gamma) \le 2\varepsilon|\gamma|,\ t_{0,n}(3n)=k, \tau(\gamma)=k,\ G(\gamma)\}\Bigg)\\
&\le \sum_{k=0}^{2\mu(p)n} \sum_{\gamma\in\mathcal{S}(0,n,2n)} P\big(\nu_n(\gamma) \le 2\varepsilon|\gamma|,\ \tau(\gamma)=k,\ G(\gamma)\big)\\
&= \sum_{k=0}^{2\mu(p)n} \sum_{\gamma\in\mathcal{S}(0,n,2n)} P\big(\nu_n(\gamma) \le 2\varepsilon|\gamma|\big)P\big(\tau(\gamma)=k\big)P(G(\gamma)).
\end{aligned}$$

Now, for $0 < p \le \frac{1}{2}$, we then have

$$\frac{P(\tau(\gamma)=k)}{P(\tau(\gamma)=0)} = \binom{|\gamma|}{k}\Big(\frac{p}{1-p}\Big)^k \le \binom{|\gamma|}{k},$$

and by the entropy bound [say, as given in Engel (1997), Corollary 2.6.2], we have

$$\binom{|\gamma|}{k} \le \exp\big(|\gamma|\, H(k/|\gamma|)\big)$$

where $H(x) = -x\log(x) - (1-x)\log(1-x)$ for $0 < x < 1$ and $H(0) = H(1) = 0$; so for $k \le 2\mu(p)n$ and $2\mu(p) \le 1/2$, we also have

$$H(k/|\gamma|) \le H\big(2n\mu(p)/|\gamma|\big) \le H(2\mu(p)).$$

To make the most of this bound, we recall that the map $p \mapsto \mu(p)$ is continuous and $\mu(1/2) = 0$ by Theorem 6.1, Remark 6.2 and Theorem 6.9 of Kesten (1986), so we can choose $\delta > 0$ such that

$$p \in \big(\tfrac{1}{2} - \delta, \tfrac{1}{2}\big] \quad\Longrightarrow\quad H(2\mu(p)) \le C_1/2,$$

where C_1 is the constant of Lemma 8. Here we should underscore that C_1 does not depend on p, except that we require $p \geq 1/4$.

When we apply the last estimate in our upper bound on $P(A_n)$, we find

$$P(A_n) \leq \sum_{k=0}^{2\mu(p)n} \sum_{\gamma \in \mathcal{S}(0,n,2n)} C_0 \exp(-C_1|\gamma|) \exp(C_1|\gamma|/2)\, P\big(G(\gamma) \cap \{\tau(\gamma) = 0\}\big)$$
$$\leq 2\mu(p)n \exp(-C_1|\gamma|/2),$$

where in the second inequality we took advantage of Lemma 5. Finally, when we combine this bound with our exponential bounds on $P(B_n)$ and $P(C_n)$, we then complete the proof of Theorem 3. □

7. The nondifferentiability of $\mu(F \oplus t)$. The nondifferentiability of the mapping $t \mapsto \mu(F \oplus t)$ asserted by Theorem 1 is actually a consequence of two inequalities for the difference quotients that one obtains from optimality considerations. The key observation is that if we take $r > 0$ and look at the passage times $t_{0,n}(3n) = t_{0,n}(3n, F)$ and $t_{0,n}(3n, F \oplus r)$ associated with the Bernoulli edge times $\{x(e)\}$ and their shifted cousins $\{x(e) + r\}$, then we have

$$t_{0,n}(3n, F \oplus r) \leq t_{0,n}(3n, F) + rN_n^-, \tag{7.1}$$

since a path $\gamma \in \mathcal{S}(0, n, 3n)$ that achieves the optimal time $t_{0,n}(3n, F)$ and the shortest length N_n^- for the Bernoulli edge times $\{x(e)\}$ will realize a passage time under the edge times $\{x(e) + r\}$ that is given by the right-hand side of inequality (7.1). The required bound then follows from the optimality of the cost $t_{0,n}(3n, F \oplus r)$ for the passage problem with the shifted edge times $\{x(e) + r\}$.

In exactly the same way, one finds that for any $s < 0$, we have

$$t_{0,n}(3n, F \oplus s) \leq t_{0,n}(3n, F) + sN_n^+, \tag{7.2}$$

and as a consequence we have the two bounds on the difference quotients

$$\frac{t_{0,n}(3n, F \oplus r) - t_{0,n}(3n)}{nr} \leq \frac{N_n^-}{n} \quad \text{and} \quad \frac{t_{0,n}(3n, F \oplus s) - t_{0,n}(3n)}{ns} \geq \frac{N_n^+}{n}.$$

Now, if we let $A(n) = \{\omega : N_n^+ \geq \rho N_n^-\}$ where $\rho > 1$ is chosen as in Theorem 3, we then find

$$\frac{t_{0,n}(3n, F \oplus s) - t_{0,n}(3n, F)}{ns}$$
$$\geq \frac{N_n^+}{n} \geq \rho \mathbb{1}_{A(n)} \frac{N_n^-}{n} \geq \rho \mathbb{1}_{A(n)} \frac{t_{0,n}(3n, F \oplus r) - t_{0,n}(3n, F)}{nr}.$$

If we now let $n \to \infty$, then Lemma 1 and equation (2.6) on the convergence to the time constant will team up with the almost sure convergence of $\mathbb{1}_{A(n)}$ to 1 given by Theorem 3 to tell us that for all $s < 0 < r$ we have the difference quotient bound

$$\frac{\mu(F \oplus s) - \mu(F)}{s} \geq \rho \frac{\mu(F \oplus r) - \mu(F)}{r}. \tag{7.3}$$

Here we should note that $\mu(F \oplus s)$ can be equal to minus infinity for some values of s, but this possibility does not interfere with the truth of inequality (7.3). Moreover, for $p < 1/2$ there is an open interval $I = I(p)$ that depends on p and contains zero such that $\mu(F \oplus s)$ is a finite concave function on $I = I(p)$. Thus, the left and right derivatives of $\phi(t)$ both exist at $t = 0$, and the inequality (7.3) for the difference quotients is more than we need to assert the analogous inequality of the one-sided derivatives,

$$D^{-}\phi(t)\big|_{t=0} \geq \rho D^{+}\phi(t)\big|_{t=0}.$$

This inequality almost completes the proof of the nondifferentiability of $\phi(\cdot)$, but we still need to check that $D^{+}\phi(t)|_{t=0} \neq 0$. Actually, we will show the stronger fact that

$$D^{+}\phi(t)\big|_{t=0} \geq 1.$$

One way to check this assertion is to consider any path γ with $\tau(\gamma) = t_{0,n}(3n, F \oplus r)$ and to note that γ is a suboptimal path for the unshifted Bernoulli percolation. This observation tells us that for $r > 0$ we have

$$t_{0,n}(3n, F) \leq \tau(\gamma) - r|\gamma|.$$

By definition, γ is $F \oplus r$ optimal, so the last inequality implies

$$\frac{t_{0,n}(3n, F \oplus r) - t_{0,n}(3n, F)}{nr} \geq \frac{|\gamma|}{n} \geq 1.$$

Now, when we let $n \to \infty$, we find

$$\frac{\mu(F \oplus r) - \mu(F)}{r} \geq 1.$$

This is more than we need to deduce $D^{+}\phi(t)|_{t=0} \geq 1$, so the proof of Theorem 1 is complete. $\square$

8. Concluding remarks. The nondifferentiability theorem proved here puts to rest a long-standing question. It was widely believed that the time constants $\phi(t) = \mu(F \oplus t)$ of shifted subcritical Bernoulli percolation must be differentiable at $t = 0$, but now we know that is simply not the case. This closes the door on an important approach to the convergence problem for N_n/n. Nevertheless, it is still seems likely that N_n/n converges for subcritical Bernoulli percolation, although now we suspect that the proof of this natural conjecture may be more subtle that might have been imagined earlier.

From the monotonicity of $\phi(t) = \mu(F \oplus t)$, we know that $\phi(t)$ is differentiable for almost all t in the set $D_p = \{t : \phi(t) > -\infty\}$. We know now that $t = 0$ is in the exceptional set for certain subcritical values of p, but we also believe that $t = 0$ is the only exceptional point. Specifically, we conjecture that $\phi(t)$ is differentiable for all $t \in D_p = \{t : \phi(t) > -\infty\}$ provided that $t \neq 0$ and $p \neq 1/2$. The result of

Zhang and Zhang (1984) adds credibility to this conjecture for $p > 1/2$, but the fact that N_n/n is likely to diverge for $p = 1/2$ also suggests that the analysis of the subcritical and the supercritical cases may be quite different.

The simplest problem suggested by our conjecture is that $\phi(t)$ is differentiable for all $p > 1/2$ and all sufficiently large values of t. This specific problem may not be difficult, yet it seems to offer a logical focus for the attack on a larger set of stubborn analytical questions.

Acknowledgment. The authors are pleased to thank a careful referee who helped us to improve the statement of Lemma 3 and who made other useful remarks that improved our exposition.

REFERENCES

AIGNER, M. (1995). Turán's graph theorem. *Amer. Math. Monthly* **102** 808–816.

BENNETT, G. (1962). Probability inequalities for sums of independent random variables. *J. Amer. Statist. Assoc.* **57** 33–45.

ENGLE, E. (1997). *Sperner Theory*. Cambridge Univ. Press.

GRIMMETT, G. and KESTEN, H. (1984). First-passage percolation, network flows, and electrical resistances. *Z. Wahrsch. Verw. Gebiete* **66** 335–366.

HAMMERSLEY, J. M. and WELSH, D. J. A. (1965). First-passage percolation, subadditive processes, stochastic networks, and generalized renewal theory. In *Bernoulli, Bayse, Laplace Anniversary Volume* (J. Neyman and L. Le Cam, eds.) 61–110. Springer, Berlin.

HOEFFING, W. (1963). Probability inqualities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **26** 13–30.

KESTEN, H. (1980). On the time constant and path length of first-passage percolation. *Adv. in Appl. Probab.* **12** 848–863.

KESTEN, H. (1982). *Percolation for Mathematicians*. Birkhäuser, Boston.

KESTEN, H. (1986). *Aspects of First Passage Percolation. Lecture Notes in Math.* **1180**. Springer, New York.

SMYTHE, R. T. and WIERMAN, J. C. (1977). First passage percolation on the square lattice. I. *Adv. in Appl. Probab.* **9** 38–54.

SMYTHE, R. T. and WIERMAN, J. C. (1978). *First Passage Percolation on the Square Lattice. Lecture Notes in Math.* **671**. Springer, New York.

WIERMAN, J. and REH, W. (1978). On conjectures in first-passage percolation theory. *Ann. Probab.* **6** 388–397.

ZHANG, Y. and ZHANG, Y. C. (1984). A limit theorem for N_{0n}/n in first-passage percolation. *Ann. Probab.* **12** 1068–1076.

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104–6340
E-MAIL: steele@wharton.upenn.edu

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF COLORADO
COLORADO SPRINGS, COLORADO 80933–7150
E-MAIL: yzhang@vision.uccs.edu