Adv. Appl. Prob. (SGSA) **44,** 330–342 (2012)
Printed in Northern Ireland

CONVEX HULLS OF UNIFORM SAMPLES FROM A CONVEX POLYGON

PIET GROENEBOOM,* *Delft University of Technology*

Abstract

In Groeneboom (1988) a central limit theorem for the number of vertices N_n of the convex hull of a uniform sample from the interior of a convex polygon was derived. To be more precise, it was shown that $\{N_n - \frac{2}{3}r\log n\}/\{\frac{10}{27}r\log n\}^{1/2}$ converges in law to a standard normal distribution, if r is the number of vertices of the convex polygon from which the sample is taken. In the unpublished preprint Nagaev and Khamdamov (1991) a central limit result for the joint distribution of N_n and A_n is given, where A_n is the area of the convex hull, using a coupling of the sample process near the border of the polygon with a Poisson point process as in Groeneboom (1988), and representing the remaining area in the Poisson approximation as a union of a doubly infinite sequence of independent standard exponential random variables. We derive this representation from the representation in Groeneboom (1988) and also prove the central limit result of Nagaev and Khamdamov (1991), using this representation. The relation between the variances of the asymptotic normal distributions of the number of vertices and the area, established in Nagaev and Khamdamov (1991), corresponds to a relation between the actual sample variances of N_n and A_n in Buchta (2005). We show how these asymptotic results all follow from one simple guiding principle. This corrects at the same time the scaling constants in Cabo and Groeneboom (1994) and Nagaev (1995).

Keywords: Convex hull

2010 Mathematics Subject Classification: Primary 60E20
Secondary 49G03; 49F10

1. Introduction

Let N_n be the number of vertices of the convex hull of a sample of size n, drawn uniformly from the interior of a convex polygon with r vertices. It was shown in Groeneboom (1988) that

$$\left\{N_n - \tfrac{2}{3}r\log n\right\}/\left\{\tfrac{10}{27}r\log n\right\}^{1/2} \xrightarrow{\text{D}} \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ denotes the standard normal distribution. This was proved by coupling the sample point process near the boundary of the convex polygon with a Poisson point process, and showing that the relevant part of the sample process could be approximated sufficiently closely by the coupled Poisson point process. The central limit result for N_n was subsequently derived from a corresponding result for the boundary of the convex hull of the approximating Poisson point process. These methods were also applied to the area A_n of the convex hull in Cabo and Groeneboom (1994), but unfortunately the central limit result A_n contained a scaling error (see Remark 3.2).

Received 2 November 2010; revision received 11 November 2011.

* Postal address: DIAM, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.
Email address: p.groeneboom@tudelft.nl

Dedicated to the memories of Alexander Nagaev and Tomasz Schreiber.

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press

Nagaev and Khamdamov (1991), using the coupling of (part of) the sample point process with a Poisson process introduced in Groeneboom (1988), derived the following interesting central limit theorem for the joint distribution of the number of vertices and the area of the convex hull of a uniform sample of n points on the interior of a convex polygon.

Theorem 1.1. (Theorem 1 of Nagaev and Khamdamov (1991).) *Let N_n denote the number of vertices of the convex hull of a uniform sample of size n from the interior of a convex polygon C with $r \geq 3$ vertices and area $A(C)$. Moreover, let A_n denote the area of the convex hull of the sample, and let the scaled 'remaining area' $\bar{A}_n$ be defined by*

$$\bar{A}_n = \frac{n\{A(C) - A_n\}}{A(C)}.$$

Then

$$\big(\tfrac{10}{27}r \log n\big)^{-1/2}\big(N_n - \tfrac{2}{3}r \log n, \bar{A}_n - \tfrac{2}{3}r \log n\big) \xrightarrow{\mathrm{D}} \mathcal{N}(0, \Sigma),$$

where $\mathcal{N}(0, \Sigma)$ denotes the normal distribution with expectation the zero vector and covariance matrix Σ given by

$$\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & \frac{14}{5} \end{pmatrix}.$$

This is an extension of the central limit theorem for the number of vertices N_n in Groeneboom (1988), and one indeed recovers the central limit theorem given there by specializing the above result to the first coordinate. Unfortunately, the preprint Nagaev and Khamdamov (1991), containing this result, was never published. Moreover, it is written in Russian and its length is 50 pages, which might also not have helped its spread in the scientific world.

In a private correspondence Christian Buchta revealed to me that the constant for the central limit theorem for the second component (the remaining area) in Nagaev and Khamdamov (1991) was consistent with a relation he had derived himself between the finite sample variances of N_n and $\bar{A}_n$.

It is the purpose of the present note to give a simple proof of Theorem 1.1, deriving the result from the central limit theorem for N_n in Groeneboom (1988). We think that using the central limit theorem of Groeneboom (1988) considerably simplifies the proof of Theorem 1.1 of Nagaev and Khamdamov (1991) and perhaps more clearly reveals the beauty of their idea. The relation between the variances in Theorem 1.1 can be considered to be a precursor (in an asymptotic sense) of the relation found between the finite sample variances in Buchta (2005).

For recent work on central limit theorems for random polytopes, see, e.g. Bárány and Reitzner (2010a) and Bárány and Reitzner (2010b), where references to earlier work in this area can also be found.

2. Representation of the remaining area by independent and identically distributed exponentials

We consider the Poisson point process $\mathcal{P}$ of intensity 1 in $\mathbb{R}_+^2$, and its left-lower convex hull, as in Groeneboom (1988). To make the connection with Groeneboom (1988), we first restate the definition of the process of vertices $\{W(a) \colon a \in \mathbb{R}_+\}$ consisting of the vertices of the (left-lower) convex hull of a Poisson process $\mathcal{P}$ with intensity 1 in $\mathbb{R}_+^2$.

Definition 2.1. For each $a > 0$, $W(a) = (U(a), V(a))$ is the point of the realization of the Poisson process $\mathcal{P}$ on $\mathbb{R}_+^2$ such that all points of the realization of $\mathcal{P}$ lie to the right of the line $x + ay = c$ which passes through $W(a)$. If there are several such points (which happens

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press

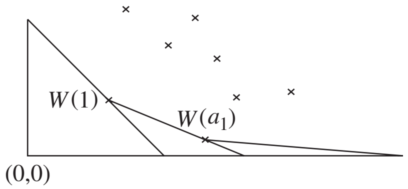

FIGURE 1: The $W(a)$ process.

with probability 0 for fixed a), we define $U(a)$ ($V(a)$) as the supremum (infimum) of the x-coordinates (y-coordinates) of points of this type.

Figure 1 gives a picture of the jump process W, where a_1 is the first jump time after 1. We now have the following result (see also Theorem 2.1 of Nagaev (1995) for a result of this type).

Theorem 2.1. *Let $a_0 = 1$, let $a_1, a_2, \ldots$ be the jump times of the process $\{W(a)\colon a \ge 1\}$, and let D_0 be the area of the isosceles triangle T_0 with a basis, running through $W(1)$, and two equal sides along the x- and y-axes, meeting at the top at the origin. Moreover, let D_i, $i \ge 1$, be the area of the triangle T_i, with top at $W(a_{i-1})$, basis along the x-axis, and sides along the lines $x + a_{i-1}y = U(a_i) + a_i V(a_i)$ and $x + a_i y = U(a_i) + a_i V(a_i)$, where $W(a_i)$, $U(a_i)$, and $V(a_i)$ are defined as in Definition 2.1. Then the following statements hold.*

(i) *The areas $D_0, D_1, \ldots$ form an independent and identically distributed (i.i.d.) sequence of standard exponential random variables.*

(ii) *Let S_i be the length of the line segment connecting $W(a_{i-1})$ and $W(a_i)$, and let L_i be the length of the segment obtained by extending the line segment from $W(a_{i-1})$ to $W(a_i)$ until it crosses the x-axis. Then the random variables S_i^2/L_i^2, $i = 1, 2, \ldots$, form an i.i.d. sequence of* Uniform$(0, 1)$ *random variables, independent of $W(1)$. Moreover, the S_i^2/L_i^2 are independent of the sequence $D_0, D_1, \ldots$.*

Proof. (i) By Lemma 2.4(i) of Groeneboom (1988) we have, for $z \ge 0$,

$$\mathrm{P}\{D_0 > z\} = \mathrm{P}\left\{\frac{1}{2}\{U(1) + V(1)\}^2 > z\right\} = \int_{\{(x,y)\colon (x+y)^2/2 > z\}} \mathrm{e}^{-(x+y)^2/2}\,\mathrm{d}x\,\mathrm{d}y = \mathrm{e}^{-z},$$

showing that D_0 has a standard exponential distribution. Let $\mathcal{F}_a$ denote the σ-algebra generated by the points $\{W(b),\ 1 \le b \le a\}$. Then, as shown in Groeneboom (1988), the process of points $\{W(a),\ a \ge 1\}$ is a Markov process with respect to the filtration $\{\mathcal{F}_a,\ a \ge 1\}$. Now note that, if $i \ge 1$, $D_i > z$ exactly when there are no points in the triangle of area z, with top at $W(a_i)$, basis along the x-axis, and sides along the lines $x + a_{i-1}y = U(a_i) + a_i V(a_i)$ and $x + a_i y = U(a_i) + a_i V(a_i)$. Since this event is independent of the location of the points $W(a_0), \ldots, W(a_{i-1})$, by the Poisson property of the point process in $\mathbb{R}_+^2$, we obtain

$$\mathrm{P}\{D_i > z\} = \mathrm{e}^{-z}, \qquad z \ge 0,$$

where the event $D_i > z$ is independent of $D_0, \ldots, D_{i-1}$ (note that we can use the strong Markov property here).

(ii) The jump measure $M(a, w; \cdot)$ of the process $\{W(a)\colon a > 0\}$ is given by

$$M(a, w; B) = \int_0^y u\,\mathbf{1}_B(au, -u)\,\mathrm{d}u;$$

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press

see Equation (2.22) of Groeneboom (1988). Hence, conditioning on $W(a) = W(a_{i-1}) = (x, y)$ and the event that there is a jump at time a, the location of the next vertex has a density proportional to u (representing the distance of $W(a)$ to the next vertex). So we obtain, for $z \in (0, 1)$,

$$\begin{aligned} \mathrm{P}\left\{\frac{S_i^2}{L_i^2} < z \,\middle|\, W(a) > W(a-) = (x, y)\right\} \\ &= \mathrm{P}\{S_i < L_i\sqrt{z} \mid W(a) > W(a-) = (x, y)\} \\ &= \mathrm{P}\{S_i < y\sqrt{z(1+a^2)} \mid W(a) > W(a-) = (x, y)\} \\ &= \frac{2}{y^2\{1+a^2\}} \int_0^{y\sqrt{z(1+a^2)}} u \,\mathrm{d}u \\ &= z, \end{aligned}$$

where we have used the fact that $\frac{1}{2}y^2\{1+a^2\}$ is the total measure of the jump measure on the line segment of length $y\sqrt{1+a^2}$, connecting (x, y) and $(x + ay, 0)$. This implies that S_i^2/L_i^2 has a uniform distribution, in accordance with Theorem 2.1 of Nagaev (1995). Moreover, since the distribution neither involves the value of $a = a_i$ nor that of $W(a_{i-1})$, the sequence of variables S_i^2/L_i^2 is i.i.d. For the same reason, the variable S_i^2/L_i^2 is independent of D_j, $j \leq i$. It is also seen that S_i^2/L_i^2 is independent of D_j, $j > i$, since the conditional distribution of D_{i+1}, given $W(a_i)$, is standard exponential, independently of the value of $W(a_i)$.

Corollary 2.1. *Let the sequences $a_0, a_1, \ldots$ and $V(a_0), V(a_1), \ldots$ be defined as in Theorem 2.1, and let $\tau_i = V(a_i)/V(a_{i-1})$, $i = 1, 2, \ldots$. Then the sequence of random variables $\tau_1, \tau_2, \ldots$ is i.i.d. and*

$$(1-\tau_i)^2 \sim \mathrm{Uniform}(0, 1).$$

Moreover, the random variables τ_i are independent of $V(a_0) = V(1)$ and the areas D_i, where D_i is defined as in Theorem 2.1.

Proof. This follows from part (ii) of Theorem 2.1 since

$$1 - \tau_i = 1 - \frac{V(a_i)}{V(a_{i-1})} = \frac{V(a_{i-1}) - V(a_i)}{V(a_{i-1})} = \frac{S_i}{L_i}, \qquad i = 1, \ldots,$$

where the last equality is the proportionality relation, well known from elementary geometry.

The following result is the key to Theorem 1.1.

Corollary 2.2. *For $m = 2, 3 \ldots$, let $N(1, m)$ be the number of jumps of the process $\{W(a) : a \in [1, m]\}$, and let $[\mathrm{E}\,N(1, m)]$ be the largest integer smaller than or equal to $\mathrm{E}\,N(1, m)$. Then the following statements hold.*

(i) $\mathrm{E}\,N(1, m) = \frac{1}{3}\log m$.

(ii) *As $m \to \infty$, the bivariate random variable*

$$\left(\frac{N(1, m) - \mathrm{E}\,N(1, m)}{\sqrt{5\log m/27}}, \sum_{i=1}^{[\mathrm{E}\,N(1,m)]} \frac{D_i - 1}{\sqrt{\mathrm{E}\,N(1, m)}}\right)$$

converges in distribution to a bivariate normal distribution with expectation 0 *and covariance matrix equal to the identity matrix I.*

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press

Proof. (i) This is Theorem 2.4(i) of Groeneboom (1988), which is a simple consequence of the fact that the expected jump rate of the process $\{W(a)\colon a \geq 1\}$ is given by $1/(3a)$.

(ii) The area D_i of the triangle T_i, as defined in Theorem 2.1, is given by

$$D_i = \tfrac{1}{2}V(a_{i-1})(V(a_{i-1}) + a_i V(a_{i-1}) - V(a_{i-1}) - a_{i-1}V(a_{i-1})) = \tfrac{1}{2}V(a_{i-1})^2(a_i - a_{i-1}). \tag{2.1}$$

Define

$$U_i = U(a_i), \qquad V_i = V(a_i), \quad \text{and} \quad W_i = (U_i, V_i), \qquad i = 0, 1, \dots .$$

It is clear that (2.1) gives a tridiagonal system for solving a_i in terms of the D_i and V_i. We obtain, using $a_0 = 1$,

$$a_n = 1 + 2\sum_{i=1}^{n} \frac{D_i}{V_{i-1}^2}, \qquad n \geq 1.$$

We now define, for $n \geq 1$,

$$Y_n = V_{n-1}^2\left\{1 + 2\sum_{i=1}^{n}\frac{D_i}{V_{i-1}^2}\right\} = V_{n-1}^2 a_n.$$

Thus,

$$\log a_n = -2\log V_{n-1} + \log Y_n, \tag{2.2}$$

and, hence, we get the 'switching relation':

$$N(1, m) \geq n \iff a_n \leq m \iff -2\log V_{n-1} + \log Y_n \leq \log m. \tag{2.3}$$

By Corollary 2.1,

$$\mathrm{E}\,V_n^2 = \mathrm{E}\,V_0^2\prod_{i=1}^{n}\tau_i^2 = 6^{-n}\,\mathrm{E}\,V_0^2, \qquad \mathrm{E}\left(\frac{V_n^2}{V_k^2}\right) = \prod_{i=k+1}^{n}\mathrm{E}\,\tau_i^2 = 6^{-(n-k)}, \quad n > k \geq 0.$$

Since, by Theorem 2.1, the τ_i are also independent of the D_i, we obtain, for all $k \geq 1$,

$$\begin{aligned}\mathrm{E}\,Y_n &= 6^{-(n-1)}\,\mathrm{E}\,V_0^2 + 2\sum_{j=1}^{n}\mathrm{E}\left(\frac{V_{n-1}^2}{V_{j-1}^2}\right)\\ &= 6^{-(n-1)}\,\mathrm{E}\,V_0^2 + 2\sum_{j=1}^{n-1}6^{-j}\\ &\leq 6^{-(n-1)}\,\mathrm{E}\,V_0^2 + 2\sum_{j=1}^{\infty}6^{-j}.\end{aligned}$$

This implies, by Markov's inequality,

$$Y_n = O_p(1) \quad \text{as } n \to \infty.$$

Since we also have $Y_n \geq 2D_n$ for all $n \geq 1$, where D_n has a standard exponential distribution, from this we obtain

$$|\log Y_n| = O_p(1) \quad \text{as } n \to \infty.$$

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press

We now obtain, from (2.2),

$$\frac{\log a_n - 3n}{\sqrt{5n}} = \frac{-2\log V_{n-1} + \log Y_n - 3n}{\sqrt{5n}} = \frac{-2\log V_{n-1} - 3n}{\sqrt{5n}} + O_p(n^{-1/2})$$

as $n \to \infty$. Moreover, since

$$-2\log V_{n-1} = -2\sum_{i=1}^{n-1} \log\left(\frac{V_i}{V_{i-1}}\right) - 2\log V_0 = -2\sum_{i=1}^{n-1} \log \tau_i - 2\log V_0,$$

we obtain, by the central limit theorem,

$$\frac{\log a_n - 3n}{\sqrt{5n}} = \frac{-2\sum_{i=1}^{n-1}\log\tau_i - 3n}{\sqrt{5n}} + o_p(1) \xrightarrow{\mathrm{D}} \mathcal{N}(0,1) \quad \text{as } n \to \infty, \tag{2.4}$$

where $\mathcal{N}(0,1)$ denotes the standard normal distribution.

Let

$$B_1(m) = \sum_{i=1}^{[\mathrm{E}\,N(1,m)]} \frac{\mathcal{D}_i - 1}{\sqrt{\mathrm{E}\,N(1,m)}}$$

and

$$B_2(m) = \frac{N(1,m) - \mathrm{E}\,N(1,m)}{\sqrt{5\log m/27}},$$

and let, for fixed $y \in \mathbb{R}$, $n = n_{m,y} \in \mathbb{N}$ be defined by

$$n = \left[\mathrm{E}\,N(1,m) + y\sqrt{\tfrac{5}{27}\log m}\right] \quad \text{as } m \to \infty. \tag{2.5}$$

Then we find, using (2.3) and (2.4), that, as $m \to \infty$,

$$\begin{aligned}
&\mathrm{P}\{B_1(m) \geq x,\ B_2(m) \geq y\} \\
&\quad = \mathrm{P}\left\{B_1(m) \geq x,\ N(1,m) \geq \mathrm{E}\,N(1,m) + y\sqrt{\tfrac{5}{27}\log m}\right\} \\
&\quad \sim \mathrm{P}\{B_1(m) \geq x,\ N(1,m) \geq n\} \\
&\quad = \mathrm{P}\{B_1(m) \geq x,\ \log a_n \leq \log m\} \\
&\quad = \mathrm{P}\left\{B_1(m) \geq x,\ \frac{\log a_n - 3n}{\sqrt{5n}} \leq \frac{\log m - 3n}{\sqrt{5n}}\right\} \\
&\quad \sim \mathrm{P}\left\{B_1(m) \geq x,\ \frac{-2\sum_{i=1}^{n-1}\log\tau_i - 3n}{\sqrt{5n}} \leq \frac{\log m - 3\,\mathrm{E}\,N(1,m) - y\sqrt{5\log m/3}}{\sqrt{5n}}\right\} \\
&\quad \sim \mathrm{P}\{B_1(m) \geq x\}\,\mathrm{P}\left\{\frac{-2\sum_{i=1}^{n-1}\log\tau_i - 3n}{\sqrt{5n}} \leq -\frac{y\sqrt{5\log m/3}}{\sqrt{5\log m/3}}\right\} \\
&\quad = \mathrm{P}\{B_1(m) \geq x\}\,\mathrm{P}\left\{\frac{-2\sum_{i=1}^{n-1}\log\tau_i - 3n}{\sqrt{5n}} \leq -y\right\},
\end{aligned}$$

where we have used part (i), (2.5), and Corollary 2.1 (independence of the τ_i and the D_i) in the second to last line. Since, by (2.4),

$$\mathrm{P}\left\{\frac{-2\sum_{i=1}^{n-1}\log\tau_i - 3n}{\sqrt{5n}} \leq -y\right\} \to \Phi(-y) = 1 - \Phi(y),$$

where Φ is the standard normal distribution function, the result now follows.

3. The central limit theorem

In this section we prove a two-dimensional central limit theorem, by combining the results of the preceding section with the results in Groeneboom (1988).

Theorem 3.1. *Let $N(a,b)$ be the number of jumps in the interval $[a,b]$ of the process W, as defined in Definition 2.1, and let $D(a,b)$ be the area of the union of the triangles T_i, corresponding to points of jump $a_i \in [a,b]$, as defined in Theorem 2.1. Then*

$$\left(\frac{5}{27}\log\left(\frac{b}{a}\right)\right)^{-1/2}\left(N(a,b)-\frac{1}{3}\log\left(\frac{b}{a}\right), D(a,b)-\frac{1}{3}\log\left(\frac{b}{a}\right)\right) \xrightarrow{\mathrm{D}} N(0,\Sigma)$$

as $b/a \to \infty$, where $N(0,\Sigma)$ is a bivariate normal distribution with expectation 0 *and covariance matrix defined by*

$$\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & \frac{14}{5} \end{pmatrix}.$$

Proof. As shown by the transformation to a stationary process (see Equation (2.27) of Groeneboom (1988)), the distribution of $N(a,b)$ depends only on the ratio b/a. The same construction shows that the distribution of $D(a,b)$ depends only on the ratio b/a. So we only have to prove the result for $a=1$ and $b>1$.

We know, from Theorem 2.4 of Groeneboom (1988), that $\mathrm{E}\,N(1,a)=\frac{1}{3}\log a$ and $\mathrm{var}(N(1,a)) \sim (\frac{5}{27})\log a$ as $a\to\infty$. Moreover,

$$D(1,a) = \sum_{a_i\in[1,a]} D_i = \sum_{a_i\in[1,a]} \mathrm{area}(T_i),$$

where the T_i are the triangles of Theorem 2.1. So we can consider $D(1,a)$ as a random sum of standard exponential random variables, where the number of terms in the sum is equal to the random variable $N(a,b)$. Reasoning heuristically, as in the case of a compound Poisson distribution, we would obtain

$$\mathrm{E}(D(1,a)) = \mathrm{E}\,N(1,a) = \tfrac{1}{3}\log a$$

and

$$\mathrm{var}(D(1,a)) = \mathrm{E}\,N(1,a) + \mathrm{var}(N(1,a)) \sim \tfrac{1}{3}\log a + \tfrac{5}{27}\log a = \tfrac{14}{27}\log a.$$

We now show that we can prove the result by using this heuristic idea.

We write $D(1,a)-\frac{1}{3}\log a$ as the sum of the terms $A_1(a)$ and $A_2(a)$, where

$$A_1(a) = \sum_{i=1}^{[\mathrm{E}\,N(1,a)]} D_i - \frac{1}{3}\log a,$$

defining $[\mathrm{E}\,N(1,a)]$ as the largest integer not exceeding $\mathrm{E}\,N(1,a)=\frac{1}{3}\log a$, and

$$A_2(a) = \begin{cases} \displaystyle\sum_{i=[\mathrm{E}\,N(1,a)]+1}^{N(1,a)} D_i & \text{if } N(1,a) > [\mathrm{E}\,N(1,a)], \\ \displaystyle -\sum_{i=N(1,a)+1}^{[\mathrm{E}\,N(1,a)]} D_i & \text{if } N(1,a) \le [\mathrm{E}\,N(1,a)]. \end{cases}$$

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press

We now have, if $N(1,a) > [\mathrm{E}\,N(1,a)]$,

$$\sum_{i=[\mathrm{E}\,N(1,a)]+1}^{N(1,a)} D_i = \sum_{i=[\mathrm{E}\,N(1,a)]+1}^{N(1,a)} (D_i - 1) + N(1,a) - [\mathrm{E}\,N(1,a)],$$

and, similarly, if $N(1,a) \le [\mathrm{E}\,N(1,a)]$,

$$-\sum_{i=N(1,a)+1}^{[\mathrm{E}\,N(1,a)]} D_i = -\sum_{i=N(1,a)+1}^{[\mathrm{E}\,N(1,a)]} (D_i - 1) + N(1,a) - [\mathrm{E}\,N(1,a)],$$

where both sides are 0 if $N(1,a) = [\mathrm{E}\,N(1,a)]$. Hence, we can write

$$D(1,a) - \tfrac{1}{3}\log a = A_1(a) + N(1,a) - [\mathrm{E}\,N(1,a)] + R(a),$$

where

$$R(a) = \begin{cases} \displaystyle\sum_{i=[\mathrm{E}\,N(1,a)]+1}^{N(1,a)} (D_i - 1) & \text{if } N(1,a) > [\mathrm{E}\,N(1,a)], \\ \displaystyle -\sum_{i=N(1,a)+1}^{[\mathrm{E}\,N(1,a)]} (D_i - 1) & \text{if } N(1,a) \le [\mathrm{E}\,N(1,a)]. \end{cases}$$

Fix $\varepsilon > 0$. By Theorem 2.4 of Groeneboom (1988), there exist an $M = M(\varepsilon) > 0$ and an $a_0 = a_0(M)$ so that

$$\mathrm{P}\left\{\left|\frac{N(1,a) - [\mathrm{E}\,N(1,a)]}{\sqrt{\log a}}\right| > M\right\} < \varepsilon, \qquad a \ge a_0.$$

Define

$$n_-(a) = [\mathrm{E}\,N(1,a)] - M\sqrt{\log a}, \qquad n_+(a) = [\mathrm{E}\,N(1,a)] + M\sqrt{\log a}.$$

Then, by Doob's inequality,

$$\begin{aligned} &\mathrm{P}\left\{\max_{m\in[[\mathrm{E}\,N(1,a)]+1,n_+(a)]}\left|\sum_{i=[\mathrm{E}\,N(1,a)]}^{m}(D_i-1)\right| > \varepsilon\sqrt{\log a}\right\} \\ &+\mathrm{P}\left\{\max_{m\in[n_-(a),[\mathrm{E}\,N(1,a)]]}\left|\sum_{i=m}^{[\mathrm{E}\,N(1,a)]}(D_i-1)\right| > \varepsilon\sqrt{\log a}\right\} \\ &\quad\le \frac{n_+(a) - n_-(a) + 1}{\varepsilon^2(\log a)} \\ &\quad\sim \frac{2M}{\varepsilon^2\sqrt{\log a}} \\ &\quad\to 0 \quad \text{as } a \to \infty. \end{aligned}$$

These relations imply that $R(a)/\sqrt{\log a} = o_p(1)$ and $a \to \infty$, and, hence,

$$\frac{D(1,a) - \mathrm{E}\,N(1,a)}{\sqrt{\log a}} = \frac{\sum_{i=1}^{[\mathrm{E}\,N(1,a)]}(D_i - 1)}{\sqrt{\log a}} + \frac{N(1,a) - [\mathrm{E}\,N(1,a)]}{\sqrt{\log a}} + o_p(1). \tag{3.1}$$

The result now follows from Corollary 2.2 and Theorem 2.4 of Groeneboom (1988).

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press

Using the methods from Groeneboom (1988) in going from the Poisson approximation to the sample process, we can now easily deduce the central limit result Theorem 1.1 from Theorem 3.1. The latter method is also used in Nagaev and Khamdamov (1991).

Remark 3.1. Instead of working directly with relation (2.1), expressing the differences between successive slopes of the convex hull in terms of the area of the corresponding rectangle and the y-coordinate of the vertex at the intersection of the line segments with these slopes, Nagaev and Khamdamov (1991) first wrote this relation in the form

$$D_i = \frac{1}{2}V(a_{i-1})^2\left(\frac{U(a_i)-U(a_{i-1})}{V(a_{i-1})-V(a_i)} - \frac{U(a_{i-1})-U(a_{i-2})}{V(a_{i-2})-V(a_{i-1})}\right),$$

and then deduced a recursive relation for the $U(a_i)$ in terms of the $V(a_i)$ and D_i from this. They then defined the random time

$$\theta_T = \inf\{i : U(a_i) \ge T\},$$

and considered sums of the form $\sum_{i=1}^{\theta_T} D_i$. This seems to lead to more complicated proofs.

Remark 3.2. The scaling constants for the central limit theorem for the area in Cabo and Groeneboom (1994) are not correct, although a correct application of the methods used in that paper would lead to the central limit theorem for the area, which is part of Theorem 1.1. We here tried to present the results of the unpublished preprint Nagaev and Khamdamov (1991) in an easily understandable way, where the presentation is considerably simplified by the use of martingales, Doob's inequality, and the results from Groeneboom (1988). In view of this simpler approach, and also the fact that Theorem 1.1 is in fact a stronger (two-dimensional) result, this approach seems preferable to the approach in Cabo and Groeneboom (1994). On the other hand, the computations along the lines of Cabo and Groeneboom (1994) give precise information on the first and second moments, as shown below in Section 4.

Although Nagaev (1995) hinted at the proof of Theorem 1.1, there are many important missing steps, which can only be filled in by referring to the unpublished preprint Nagaev and Khamdamov (1991). It seems fair to say that, without knowledge of this preprint, deducing the result from Nagaev (1995) is pretty hard. Moreover, the crucial relation (3.7) of Nagaev (1995) contains an incorrect scaling constant (the constant $\frac{5}{4}$ there should be $\frac{20}{27}$), which further complicates the derivation of Theorem 1.1. For this reason, we gave a simplified and self-contained treatment above.

Remark 3.3. Buchta (2005) gave the following relation between the sample variances of N_n and $\bar{A}_n$ (using the notation of Theorem 1.1):

$$\frac{(n+1)(n+2)\operatorname{var}(\bar{A}_n)}{n^2} = \operatorname{var}(N_n) + d_{n+2}.$$

Here

$$d_n = (\mathrm{E}\,N_n)^2 - \frac{n(\mathrm{E}\,N_{n-1})^2}{n-1} - (2n-1)\,\mathrm{E}\,N_n + 2n\,\mathrm{E}\,N_{n-1} \sim \mathrm{E}\,N_n \sim \frac{9}{5}\operatorname{var}(N_n)$$

as $n \to \infty$. Hence,

$$\operatorname{var}(\bar{A}_n) \sim \tfrac{14}{5}\operatorname{var}(N_n) \quad \text{as } n \to \infty,$$

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press

in accordance with the covariance matrix Σ in Theorem 1 of Nagaev and Khamdamov (1991) (Theorem 1.1 above). Note that the decomposition of the variance of $\bar{A}_n$ corresponds to the decomposition (3.1), where d_{n+2} corresponds to the variance of the exponentials ξ_i in (3.1) and $\operatorname{var}(N_n)$ corresponds to the variance of the second term on the right-hand side of (3.1).

From Theorem 2 of Buchta (2003), for the number of vertices N_n of the convex hull of the points $(0,1)$, $(1,0)$, and $P_1, \dots, P_n$, where $P_1, \dots, P_n$ is a uniform sample from the interior of the triangle with vertices $(0,0)$, $(0,1)$, and $(1,0)$,

$$\mathrm{E}\, N_n = \frac{1}{3}\left\{2\sum_{i=1}^{n}\frac{1}{i} + 1\right\}$$

and

$$\operatorname{var}(N_n) = \frac{1}{27}\left\{10\sum_{i=1}^{n}\frac{1}{i} + 12\sum_{i=1}^{n}\frac{1}{i^2} - 28 + \frac{12}{n+1}\right\}.$$

This gives

$$\mathrm{E}\, N_n \sim \tfrac{2}{3}\log n, \qquad \operatorname{var}(N_n) \sim \tfrac{10}{27}\log n, \quad \text{as } n \to \infty, \tag{3.2}$$

which corresponds to the distribution results derived in Groeneboom (1988), as is also noted in Buchta (2003).

The results in Groeneboom (1988) and Nagaev and Khamdamov (1991) imply only that a normal limit distribution for the number of vertices of the convex hull of a uniform sample is obtained from the interior of a convex polygon with r vertices by centering with $\frac{2}{3}r\log n$ and dividing by $(\frac{10}{27}r\log n)^{1/2}$. It is not proved there that the variance of the number of vertices itself is also of the order $\frac{10}{27}r\log n$. In principle, it is possible to have a central limit theorem where the scaling needed to obtain the central limit result is different from that obtained from the actual variance. However, the only remaining consideration to go from (3.2) to the result that the variance itself is also of the order $\frac{10}{27}r\log n$ seems to be the appropriate use of the independence of what happens in the corners of the polygons, so that we can conclude that the variance is the sum of the variances of the number of vertices in these corners. Moreover, we have to go from what happens in the triangle to what happens in the corners of the polygon. This is the subject of the current research by Buchta. Results for higher moments of the convex hull of a uniform sample from a triangle with vertices $(0,0)$, $(0,1)$, and $(1,0)$ are given in Buchta (2012).

4. Simulations

Let $N(a,b)$ and $D(a,b)$ be defined as in Theorem 3.1. The distribution of these random variables depends only on the ratio b/a, and in this section we present some simulation results for these random variables, taking $a=1$ and replacing b by a.

The algorithm, given in Section 4 of Nagaev (1995), was used to simulate part of the boundary of the convex hull of a Poisson process with intensity 1 in the first quadrant. The starting triangle is bounded by the x-axis, y-axis, and a line of the form $x+y=c$, where $c>0$. Its area D_0 has a standard exponential distribution and the point $W(1)$ is uniformly distributed on the line segment which is the hypotenuse of this triangle.

With the algorithm of Nagaev (1995) we can now generate the points $W(a)$, $a \ge 1$, and simulate in this way the distributions of $N(1,a)$ and $D(1,a)$. We start with $N(1,a)$ and recall

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press

TABLE 1: Comparison of $\mathrm{E}\,N(1,a)$ and $\mathrm{var}(N(1,a))$ with simulated and asymptotic values.

$\log a$	$\mathrm{E}\,N(1,a)$		$\mathrm{var}(N(1,a))$		
	Simulated	Exact	Simulated	Exact	Asymptotic
10	3.3519	3.3333	2.1193	2.0596	1.8519
50	16.6668	16.6667	9.5908	9.4670	9.2593
100	33.4259	33.3333	18.7039	18.7263	18.5185

TABLE 2: Comparison of $\mathrm{E}\,D(1,a)$ and $\mathrm{var}(D(1,a))$ with simulated and asymptotic values.

$\log a$	$\mathrm{E}\,D(1,a)$		$\mathrm{var}(D(1,a))$		
	Simulated	Exact	Simulated	Exact	Asymptotic
10	3.3664	3.3333	5.4089	5.3040	5.1852
50	16.6576	16.6667	26.1452	26.0448	25.9259
100	33.4933	33.3333	52.3304	51.9707	51.8519

the exact expressions for the expectation $\mathrm{E}\,N(1,a)$ and $\mathrm{var}(N(1,a))$ from Groeneboom (1988, Theorem 2.4):

$$\mathrm{E}\,N(1,a) = \tfrac{1}{3}\log a,$$

$$\mathrm{var}(N(1,a)) = \frac{5}{27}\log a + \frac{4}{9}(\tan^{-1}(\sqrt{a-1}))^2 + \frac{8}{9}\left(\frac{\tan^{-1}(\sqrt{a-1})}{\sqrt{a-1}} - 1\right). \tag{4.1}$$

As noted at the top of page 34 in Cabo and Groeneboom (1994), the formula for the variance of $N(1,a)$, given in Theorem 2.1 of Groeneboom (1988), contained a typo (the argument of the first $\tan^{-1}$ above was a instead of $\sqrt{a-1}$), and the correct formula is in fact given on page 365 of Groeneboom (1988) (which we use here). Note that these are exact expressions for $\mathrm{E}\,N(1,a)$ and $\mathrm{var}(N(1,a))$ and not asymptotic ones.

Table 1 shows the means and variances for 10 000 simulations for $\log a = 10$, 50, and 100. The exact values are given to four decimal places.

It is seen from Table 1 that $\mathrm{E}\,N(1,a)$ and $\mathrm{var}(N(1,a))$ are quite close to the simulated values and that, not unexpectedly, for $a = 10$, the exact expression for the variance of $N(1,a)$, given by (4.1), is closer to the simulated value than the asymptotic value.

Similarly, we performed 10 000 simulations for $\log a = 10$, 50, and 100 to simulate the behavior of $D(1,a)$. Using the (corrected) methods of computation of Cabo and Groeneboom (1994) (details are given in Groeneboom (2011b)), it can be shown that

$$\mathrm{E}\,D(1,a) = \tfrac{1}{3}\log a$$

and, defining $\alpha = a - 1$, that

$$\begin{aligned}\mathrm{var}(D(1,a)) = \frac{14}{27}\log a + \frac{2}{3\alpha^2} + \frac{4}{9\alpha} - \frac{44}{45} - \frac{2\{3+\alpha(3-4\alpha)\}\tan^{-1}(\sqrt{\alpha})}{9\alpha^{5/2}}\\ + \frac{4}{9}(\tan^{-1}(\sqrt{\alpha}))^2.\end{aligned}$$

These are again exact expressions for $\mathrm{E}\,D(1,a)$ and $\mathrm{var}(D(1,a))$ and not asymptotic ones. The results are given in Table 2.

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press

TABLE 3: Comparison of $\mathrm{E}\,\nu_t$ and $\mathrm{var}(\nu_t)$ with simulated and asymptotic values.

$\log t$	$\mathrm{E}\,\nu_t$		$\mathrm{var}(\nu_t)$		
	Simulated	Exact	Simulated	$\frac{20}{27}\log t$	$\frac{5}{4}\log t$ (Nagaev (1995))
10	13.0778	13.3333	7.2630	7.40741	12.5
50	66.4792	66.6667	37.6192	37.0370	62.5
100	133.1330	133.3333	74.542	74.0741	125

We finally turn our attention to Relation (3.7) of Nagaev (1995). This relation gives asymptotic expressions for the expectation and variance of the number, ν_t, of vertices falling in a disk S_t with radius t and center $(0, 0)$. On the basis of the results in Groeneboom (1988), it is to be expected that

$$\mathrm{E}\,\nu_t \sim \tfrac{4}{3}\log t, \qquad \mathrm{var}(\nu_t) \sim \tfrac{20}{27}\log t, \quad \text{as } t \to \infty, \tag{4.2}$$

whereas Relation (3.7) of Nagaev (1995) gives the above relation for $\mathrm{E}\,\nu_t$, but $\frac{5}{4}\log t$ as the asymptotic expression for $\mathrm{var}(\nu_t)$. The argument for (4.2) is that, first of all, ν_t can be expected to behave asymptotically as the number of vertices with coordinates $x > y$ such that $x < t$ plus the number of vertices with coordinates $y \geq x$ such that $y < t$, since vertices with large x-coordinates will, with high probability, be very close to the x-axis and vertices with large y-coordinates will, with high probability, be very close to the y-axis. Secondly, again by Groeneboom (1988), the number of vertices with coordinates $x > y$ such that $x < t$ will behave asymptotically as $N(1, t^2)$, and, similarly, the number of vertices with coordinates $y \geq x$ such that $y < t$ will behave asymptotically as $N(1/t^2, 1)$.

By the construction of the algorithm in Nagaev (1995), we can simulate the number of vertices $W(a)$, $a \geq 1$, satisfying $U(a)^2 + V(a)^2 < t^2$, by running the algorithm till we obtain a vertex $W(a)$ such that

$$U(a)^2 + V(a)^2 \geq t^2.$$

The resulting asymptotic behaviors of $\mathrm{E}\,\nu_t$ and $\mathrm{var}(\nu_t)$ are obtained from this by multiplying the results by the factor 2. In Table 3 we present the results for 10 000 simulations for $\log t = 10$, 50, and 100.

Table 3 clearly suggests that the factor $\frac{5}{4}$ is much too large and that the correct approximation is indeed given by (4.2).

5. Concluding remarks

There is a remarkable analogy between the behavior of the left-lower convex hull of the Poisson point process, discussed above, and the least concave majorant of (one-sided) Brownian motion without drift, as analyzed in Groeneboom (1983). In the same way there is an analogy between the behavior of the lower convex hull of the Poisson point process inside a parabola, as analyzed in Groeneboom (1988) and Nagaev (1995), and the least concave majorant of Brownian motion with a parabolic drift, as studied in Groeneboom (1989) and Groeneboom (2011a). Why this is the case is still somewhat of a mystery and deserves (in my view) further investigation.

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press

Acknowledgements

I would like to posthumously thank Tomasz Schreiber for sending me the unpublished preprint Nagaev and Khamdamov (1991), and Christian Buchta for making me aware of Buchta (2003) and sending me the preprint Buchta (2012).

References

Bárány, I. and Reitzner, M. (2010a). Poisson polytopes. *Ann. Prob.* **38,** 1507–1531.

Bárány, I. and Reitzner, M. (2010b). On the variance of random polytopes. *Adv. Math.* **225,** 1986–2001.

Buchta, C. (2003). On the distribution of the number of vertices of a random polygon. *Anz. Österreich. Akad. Wiss. Math. Natur. Kl.* **139,** 17–19.

Buchta, C. (2005). An identity relating moments of functionals of convex hulls. *Discrete Comput. Geom.* **33,** 125–142.

Buchta, C. (2012). On the boundary structure of the convex hull of random points. *Adv. Geom.* **12,** 179–190.

Cabo, A. J. and Groeneboom, P. (1994). Limit theorems for functionals of convex hulls. *Prob. Theory Relat. Fields* **100,** 31–55.

Groeneboom, P. (1983). The concave majorant of Brownian motion. *Ann. Prob.* **11,** 1016–1027.

Groeneboom, P. (1988). Limit theorems for convex hulls. *Prob. Theory Relat. Fields* **79,** 327–368.

Groeneboom, P. (1989). Brownian motion with a parabolic drift and Airy functions. *Prob. Theory Relat. Fields* **81,** 79–109.

Groeneboom, P. (2011a). Vertices of the least concave majorant of Brownian motion with parabolic drift. *Electron. J. Prob.* **16,** 2234–2258.

Groeneboom, P. (2011b). The remaining area of the convex hull of a Poisson process. Preprint. Available at http://arxiv.org/abs/1111.2504v2.

Nagaev, A. V. and Khamdamov, I. M. (1991). Limit theorems for functionals of random convex hulls. Preprint. Institute of Mathematics, Academy of Sciences of Uzbekistan.

Nagaev, A. V. (1995). Some properties of convex hulls generated by homogeneous Poisson point processes in an unbounded convex domain. *Ann. Inst. Statist. Math.* **47,** 21–29.

https://doi.org/10.1239/aap/1339878714 Published online by Cambridge University Press