

Assessing significance in a Markov chain without mixing

Maria Chikina

*Department of Computational and Systems Biology
University of Pittsburgh
3078 Biomedical Science Tower 3
Pittsburgh, PA 15213
U.S.A.**

Alan Frieze[†] and Wesley Pegden[‡]
*Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA 15213
U.S.A.*

(Dated: January 13, 2017)

We present a new statistical test to detect that a presented state of a reversible Markov chain was not chosen from a stationary distribution. In particular, given a value function for the states of the Markov chain, we would like to demonstrate rigorously that the presented state is an outlier with respect to the values, by establishing a *p*-value under the null hypothesis that it was chosen from a stationary distribution of the chain.

A simple heuristic used in practice is to sample ranks of states from long random trajectories on the Markov chain, and compare these to the rank of the presented state; if the presented state is a 0.1%-outlier compared to the sampled ranks (its rank is in the bottom 0.1% of sampled ranks) then this should correspond to a *p*-value of 0.001. This is not rigorous, however, without good bounds on the mixing time of the Markov chain.

Our test is the following: given the presented state in the Markov chain, take a random walk *from the presented state* for any number of steps. We prove that observing that the presented state is an ε -outlier on the walk is significant at $p = \sqrt{2\varepsilon}$, under the null hypothesis that the state was chosen from a stationary distribution. We assume nothing about the Markov chain beyond reversibility, and show that significance at $p \approx \sqrt{\varepsilon}$ is essentially best possible in general. We illustrate the use of our test with a potential application to the rigorous detection of gerrymandering in Congressional districtings.

I. INTRODUCTION

The essential problem in statistics is to bound the probability of a surprising observation, under a *null hypothesis* that observations are being drawn from some unbiased probability distribution. This calculation can fail to be straightforward for a number of reasons. On the one hand, defining the way in which the outcome is surprising requires care; for example, intricate techniques have been developed to allow sophisticated analysis of cases where multiple hypotheses are being tested. On the other hand, the correct choice of the unbiased distribution implied by the null hypothesis is often not immediately clear; classical tools like the *t*-test are often applied by making sim-

plifying assumptions about the distribution in such cases. If the distribution is well-defined, but not be amenable to mathematical analysis, a *p*-value can still be calculated using bootstrapping, if test samples can be drawn from the distribution.

A third way for *p*-value calculations to be nontrivial occurs when the observation is surprising in a simple way, the null hypothesis distribution is known, but where there is no simple algorithm to draw samples from this distribution. In these cases, the best candidate method to sample from the null hypothesis is often through a *Markov chain*, which essentially takes a long random walk on the possible values of the distribution. Under suitable conditions, theorems are available which guarantee that the chain converges to its *stationary distribution*, allowing a random sample to be drawn from a distribution quantifiably close to the target distribution. This principle has given rise to diverse applications of Markov chains, including to simulations of chemical reactions, to Markov chain Monte Carlo statistical methods, to protein folding, and to statistical physics models.

A persistent problem in applications of Markov chains is the often unknown *rate* at which the chain converges to

*email: mchikina@pitt.edu; Research supported in part by NIH grants 1R03MH10900901A1 and U545U54HG00854003.

[†]email: alan@random.math.cmu.edu; Research supported in part by NSF Grants DMS1362785, CCF1522984 and a grant(333329) from the Simons Foundation.

[‡]email: wes@math.cmu.edu; Research supported in part by NSF grant DMS-1363136 and the Sloan foundation.

Author list is alphabetical.

the stationary distribution[1, 2]. It is rare to have rigorous results on the mixing time of a real-world Markov chain, which means that in practice, sampling is performed by running a Markov chain for a “long time”, and hoping that sufficient mixing has occurred. In some applications, such as in simulations of the Potts model from statistical physics, practitioners have developed modified Markov chains in the hopes of achieving faster convergence [3], but such algorithms have still been demonstrated to have exponential mixing times in many settings [4–6].

In this paper, we are concerned with the problem of assessing statistical significance in a Markov chain without requiring results on the mixing time of the chain, or, indeed, any special structure at all in the chain beyond reversibility. Formally, we consider a reversible Markov chain \mathcal{M} on a state space Σ , which has an associated label function $\omega : \Sigma \rightarrow \mathbb{R}$. (The definition of Markov chain is recalled at the end of this section.) The labels constitute auxiliary information, and are not assumed to have any relationship to the transition probabilities of \mathcal{M} . We would like to demonstrate that a presented state σ_0 is unusual for states drawn from a stationary distribution π . If we have good bounds on the mixing time of \mathcal{M} , then we can simply sample from a distribution of $\omega(\pi)$, and use bootstrapping to obtain a rigorous p -value for the significance of the smallness of the label of σ_0 . Such bounds are rarely available, however.

We propose the following simple and rigorous test to detect that σ_0 is unusual relative to states chosen randomly according to π , which does not require bounds on the mixing rate of \mathcal{M} :

The $\sqrt{\varepsilon}$ test: Observe a trajectory $\sigma_0, \sigma_1, \sigma_2, \dots, \sigma_k$ from the state σ_0 , for any fixed k . The event that $\omega(\sigma_0)$ is an ε -outlier among $\omega(\sigma_0), \dots, \omega(\sigma_k)$ is significant at $p = \sqrt{2\varepsilon}$, under the null-hypothesis that $\sigma_0 \sim \pi$.

Here, we say that a real number α_0 is an ε -outlier among $\alpha_0, \alpha_1, \dots, \alpha_k$ if there are at most $\varepsilon(k+1)$ indices i for which $\alpha_i \leq \alpha_0$. In particular, note for the $\sqrt{\varepsilon}$ test, the only relevant feature of the label function is the ranking it imposes on the elements of Σ . In the Supplement, we consider the statistical power of the test, and show that the relationship $p \approx \sqrt{\varepsilon}$ is best possible. We leave as an open question whether the constant $\sqrt{2}$ can be improved.

Roughly speaking, this kind of test is possible because a reversible Markov chain cannot have many *local outliers* (Figure 1). Rigorously, the validity of the test is a consequence of the following theorem.

Theorem I.1. *Let $\mathcal{M} = X_0, X_1, \dots$ be a reversible Markov chain with a stationary distribution π , and suppose the states of \mathcal{M} have real-valued labels. If $X_0 \sim \pi$,*

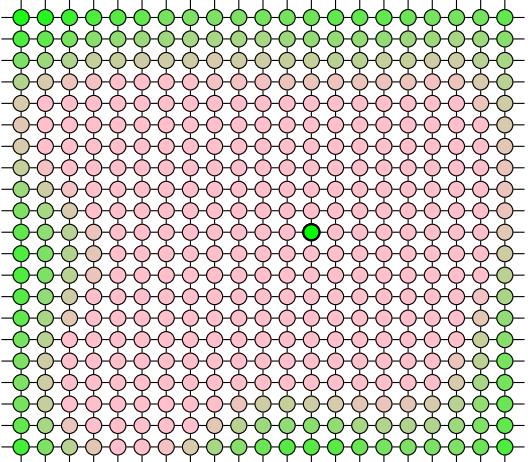


FIG. 1: This schematic illustrates a region of a potentially much larger Markov chain with a very simple structure; from each state seen here, a jump is made with equal probabilities to each of the 4 neighboring states. Colors from green to pink represent labels from small to large. It is impossible to know from this local region alone whether the highlighted green state has unusually small label in this chain overall. But to an unusual degree, this state is a *local outlier*. The $\sqrt{\varepsilon}$ test is based on the fact that *no* reversible Markov chain can have too many local outliers.

then for any fixed k , the probability that the label of X_0 is an ε -outlier from among the list of labels observed in the trajectory $X_0, X_1, X_2, \dots, X_k$ is at most $\sqrt{2\varepsilon}$.

We emphasize that Theorem I.1 makes no assumptions on the structure of the Markov chain beyond reversibility. In particular, it applies even if the chain is not irreducible (in other words, even if the state space is not connected) even though in this case the chain will never mix.

In Section III we apply the test to Markov chains generating random political districtings, for which no results on rapid mixing exist. In particular, we show that for various simple choices of constraints on what constitutes a “valid” Congressional districting (e.g., that the districts are contiguous, and satisfy certain geometric constraints), the current Congressional districting of Pennsylvania is significantly biased, under the null hypothesis of a districting chosen at random from the set of valid districtings. (We obtain p -values between $\approx 2.5 \cdot 10^{-4}$ and $\approx 8.1 \cdot 10^{-7}$ for the constraints we considered.)

One hypothetical application of the $\sqrt{\varepsilon}$ test is the possibility of rigorously demonstrating that a chain is not mixed. In particular, suppose that Research Group 1 has run a reversible Markov chain for n_1 steps, and believes that this was sufficient to mix the chain. Research Group 2 runs the chain for a further n_2 steps, producing a trajectory of total length $n_1 + n_2$, and notices that a property of interest changes in these n_2 further steps. Heuristically, this suggests that n_1 steps was not suffi-

cient to mix the chain, and the $\sqrt{\varepsilon}$ test quantifies this reasoning rigorously. For this application, however, we must allow X_0 to be distributed not exactly as the stationary distribution π , but as some distribution π' whose total variation distance to π is small, as this is the scenario for a “mixed” Markov chain. In the Supplement, we give a version of Theorem I.1 which applies in this scenario.

One area of research related to the present manuscript concerns methods for *perfect sampling* from Markov chains. Beginning with the Coupling From The Past algorithm of Propp and Wilson[7, 8] and several extensions[9, 10], these techniques are designed to allow sampling of states *exactly* from the stationary distribution π , without having rigorous bounds on the mixing time of the chain. Compared with $\sqrt{\varepsilon}$ test, perfect sampling techniques have the disadvantage that they require the Markov chain to possess certain structure for the method to be implementable, and that the time it takes to generate each perfect sample is unbounded. Moreover, although perfect sampling methods do not require rigorous bounds on mixing times to work, they will not run efficiently on a slowly mixing chain. The point is that for a chain which has the right structure, and which actually mixes quickly (in spite of an absence of a rigorous bound on the mixing time), algorithms like CFTP can be used to rigorously generate perfect samples. On the other hand, the $\sqrt{\varepsilon}$ test applies to *any* reversible Markov chain, regardless of the structure, and has running time k chosen by the user. Importantly, it is quite possible that the test can detect bias in a sample even when k is much smaller than the mixing time of the chain, as seems to be the case in the districting example discussed in Section III. Of course, unlike perfect sampling methods, the $\sqrt{\varepsilon}$ test can only be used to demonstrate a given sample is not chosen from π ; it does not give a way for generating samples from π .

II. DEFINITIONS

We remind the reader that a Markov chain is a discrete time random process; at each step, the chain jumps to a new state, which only depends on the previous state. Formally, a Markov chain \mathcal{M} on a state space Σ is a sequence $\mathcal{M} = X_0, X_1, X_2, \dots$ of random variables taking values in Σ (which correspond to states which may be occupied at each step) such that for any $\sigma, \sigma_0, \dots, \sigma_{n-1} \in \Sigma$,

$$\begin{aligned} \Pr(X_n = \sigma | X_0 = \sigma_0, X_1 = \sigma_1, \dots, X_{n-1} = \sigma_{n-1}) \\ = \Pr(X_1 = \sigma | X_0 = \sigma_0). \end{aligned}$$

Note that a Markov chain is completely described by the distribution of X_0 and the transition probabilities $\Pr(X_1 = \sigma_1 | X_0 = \sigma_0)$ for all pairs $\sigma_0, \sigma_1 \in \Sigma$. Terminology is often abused, so that the *Markov chain* refers only

to the ensemble of transition probabilities, regardless of the choice of distribution for X_0 .

With this abuse of terminology, a *stationary distribution* for the Markov chain is a distribution π such that $X_0 \sim \pi$ implies that $X_1 \sim \pi$, and therefore that $X_i \sim \pi$ for all i . When the distribution of X_0 is a stationary distribution, the Markov chain X_0, X_1, \dots is said to be *stationary*. A stationary chain is said to be *reversible* if for all i, k , the sequence of random variables $(X_i, X_{i+1}, \dots, X_{i+k})$ is identical in distribution to the sequence $(X_{i+k}, X_{i+k-1}, \dots, X_i)$. Finally a chain is *reducible* if there is a pair of states σ_0, σ_1 such that σ_1 is inaccessible from σ_0 via legal transitions, and *irreducible* otherwise.

A simple example of a Markov chain is a random walk on a directed graph, beginning from an initial vertex X_0 chosen from some distribution. Here Σ is the vertex-set of the directed graph. If we are allowed to label the directed edges with positive reals and the probability of traveling along an arc is proportional to the label of the arc (among those leaving the present vertex), then any Markov chain has such a representation, as the transition probability $\Pr(X_1 = \sigma_1 | X_0 = \sigma_0)$ can be taken as the label of the arc from σ_0 to σ_1 . Finally, if the graph is undirected, the corresponding Markov chain is reversible.

III. DETECTING BIAS IN POLITICAL DISTRICTING

A central feature of American democracy is the selection of Congressional districts in which local elections are held to directly elect national representatives. Since a separate election is held in each district, the proportions of party affiliations of the slate of representatives elected in a state does not always match the proportions of statewide votes cast for each party. In practice, large deviations from this seemingly desirable target do occur.

Various tests have been proposed to detect *gerrymandering* of districtings, in which a districting is drawn in such a way as to bias the resulting slate of representatives towards one party; this can be accomplished by concentrating voters of the unfavored party in a few districts. One class of methods to detect gerrymandering concerns heuristic ‘smell tests’ which judge whether a districting seems generally reasonable in its statistical properties (see, e.g., [11, 12]). For example, such tests may frown upon districtings in which difference between the mean and median vote on district-by-district basis is unusually large [13].

The simplest statistical smell test, of course, is whether the party affiliation of the elected slate of representatives

is close in proportion to the party affiliations of votes for representatives. Many states have failed this simple test spectacularly, such as in Pennsylvania, where in 2012, 48.77% of votes were cast for Republican representatives and 50.20% for Democrat representatives, in an election which resulted in a slate of 13 Republican representatives and 5 Democrat representatives.

Heuristic statistical tests such as these all suffer from lack of rigor, however, due to the fact that the statistical properties of ‘typical’ districtings are not rigorously characterized. For example, it has been shown [14] that Democrats may be at a natural disadvantage when drawing electoral maps even when no bias is at play, because Democrat voters are often highly geographically concentrated in urban areas. Particularly problematic is that the degree of geographic clustering of partisans is highly variable from state to state: what looks like a gerrymandered districting in one state may be a natural consequence of geography in another.

Some work has been done in which the properties of a “valid” districting are defined (which may be required to have roughly equal populations among districts, have districts with reasonable boundaries, etc.) so that the characteristics of a given districting can be compared with what would be “typical” for a valid districting of the state in question, by using computers to generate random districtings [15, 16]; see also [13] for discussion. However, much of this work has relied on heuristic sampling procedures which do not have the property of selecting districtings with equal probability (and, more generally, whose distributions are not well-characterized), undermining rigorous statistical claims about the properties of typical districts.

In an attempt to establish a rigorous framework for this kind of approach, several groups [17–19] have used Markov chains to sample random valid districtings for the purpose of such comparisons. Like many other applications of real-world Markov chains, however, these methods suffer from the completely unknown mixing time of the chains in question. Indeed, no work has even established that the Markov chains are irreducible (in the case of districtings, this means that any valid districting can be reached from any other by a legal sequence of steps), even if valid districtings were only required to consist of contiguous districts of roughly equal populations. And, indeed, for very restrictive notions of what constitutes a valid districting, irreducibility certainly fails.

As a straightforward application of the $\sqrt{\varepsilon}$ test, we can achieve rigorous p -values in Markov models of political districtings in spite of the lack of bounds on mixing times of the chains. In particular, for all choices of the constraints on valid districtings we tested, the $\sqrt{\varepsilon}$ test showed that the current Congressional districting of Pennsylvania is an outlier at significance thresholds rang-

ing from $p \approx 2.5 \cdot 10^{-4}$ and $p \approx 8.1 \cdot 10^{-7}$. Detailed results of these runs are in the Supplement.

A key advantage of the Markov chain approach to gerrymandering is that it rests on a rigorous framework; namely, comparing the actual districting of a state with typical (i.e., random) districtings from a well-defined set of valid districtings. The rigor of the approach thus depends on the availability of a precise definition of what constitutes a valid districting; in principle and in practice, this is a a thorny legal question. While some work on Markov chains for redistricting (in particular, [19]) has aimed to account for complex constraints on valid districtings, our main goal in the present manuscript is to illustrate the application of the $\sqrt{\varepsilon}$ test. In particular, we have erred on the side of using relatively simple sets of constraints on valid districtings in our Markov chains, while checking that our significance results are not highly sensitive to the parameters that we use. On the other hand, our test immediately gives a way of putting the work such that in [19] on a rigorous statistical footing.

The full description of the Markov chain we use in the present work is given in the supplement, but its basic structure is as follows: Pennsylvania is divided into roughly 9000 Census blocks. (These blocks can be seen upon close inspection of Figure 2.) We define a division of these blocks into 18 districts to be a valid districting of Pennsylvania if districts differ in population by less than 2%, are contiguous, are simply connected (districts do not contain holes) and are “compact” in ways we discuss in the supplement; roughly, this final condition prohibits districts with extremely contorted structure. The state space of the Markov chain is the set of valid districtings of the state, and one step of the Markov chain consists of randomly swapping a precinct on the boundary of a district to a neighboring district, if the result is still a valid districting. As we discuss in the supplement, the chain is adjusted slightly to ensure that the uniform distribution on valid districtings is indeed a stationary distribution for the chain. Observe that this Markov chain has a potentially huge state space; if the only constraint on valid districtings was that the districts have roughly equal population, there would be 10^{10000} or so valid districtings. Although contiguity and especially compactness are severe restrictions which will decrease this number substantially, it seems difficult to compute effective upper bounds on the number of resulting valid districtings, and certainly, it is still enormous. Impressively, these considerations are all immaterial to our very general method.

Applying the $\sqrt{\varepsilon}$ test involves the choice of a label function $\omega(\sigma)$, which assigns a real-number to each districting. We have conducted runs using two label functions: ω_{var} is the (negative) variance of the proportion of Democrats in each district of the districting (as measured by 2012 presidential votes), and ω_{MM} is the differ-

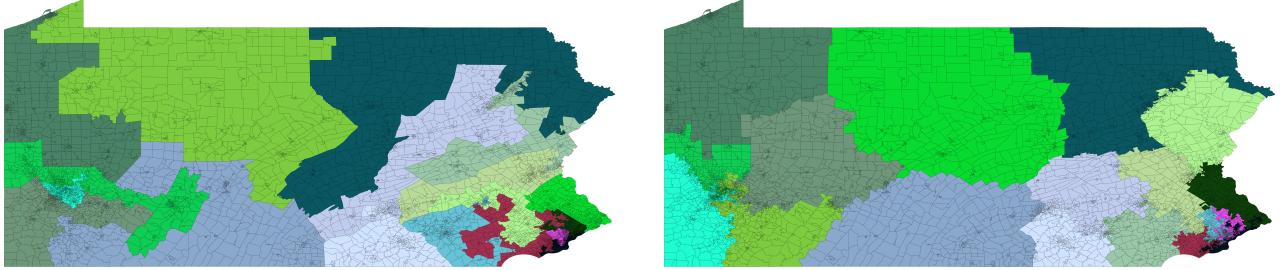


FIG. 2: **Left:** The current districting of Pennsylvania. **Right:** A districting produced by the Markov chain after 2^{40} steps. (Detailed parameters for this run are given in the supplement.)

ence between the median and mean of the proportions of Democrats in each district. ω_{MM} is motivated by the fact that this metric has a long history of use in gerrymandering, and is directly tied to the goals of gerrymandering, while the use of the variance is motivated by the fact that it can change quickly with small changes in districtings. These two choices are discussed further in the Supplement, but an important point is that our use of these label functions **is not** based on an assumption that small values of ω_{var} or of ω_{MM} directly imply gerrymandering. Instead, as Theorem I.1 is valid for any fixed label function, these labels are tools used to demonstrate significance, which are chosen because they are simple and natural functions on vectors which can be quickly computed, seem likely to be different for typical versus gerrymandered districtings, and have the potential to change relatively quickly with small changes in districtings. For the various notions of valid districtings we considered, the $\sqrt{\varepsilon}$ test demonstrated significance at p -values in the range 10^{-4} to 10^{-5} for the ω_{MM} label function, and in the range 10^{-4} and 10^{-7} for the ω_{var} label function.

As noted earlier, the $\sqrt{\varepsilon}$ test can easily be used with more complicated Markov chains which capture more intricate definitions of the set of valid districtings. For example, the current districting of Pennsylvania splits fewer rural counties than the districting on the right in Figure 2, and the number of county splits is one of many metrics for valid districtings considered by the Markov chains developed in [19]. Indeed, our test will be of particular value in cases where complex notions of what constitute a valid districting slow the chain to make the heuristic mixing assumption particularly questionable. Regarding mixing time: even our chain with relatively weak constraints on the districtings (and very fast running time in implementation) appears to mix too slowly to sample π , even heuristically; in Figure 2, we see that several districts seem still to have not left their general position from the initial districting, even after 2^{40} steps.

On the same note, it should also be kept in mind that while our result gives a method to rigorously disprove that a given districting is unbiased—e.g., to show that the districting is unusual among districtings X_0 distributed according to the stationary distribution π —it

does so *without* giving a method to sample from the stationary distribution. In particular, our method can not answer the question of how many seats Republicans and Democrats should have in a typical districting of Pennsylvania, because we are still not mixing the chain. Instead, Theorem I.1 has given us a way to disprove $X_0 \sim \pi$ without sampling π .

IV. PROOF OF THEOREM I.1

We let π denote any stationary distribution for \mathcal{M} , and suppose that the initial state X_0 is distributed as $X_0 \sim \pi$, so that in fact $X_i \sim \pi$ for all i . We say σ_j is ℓ -small among $\sigma_0, \dots, \sigma_k$ if there are at most ℓ indices $i \neq j$ among $0, \dots, k$ such that the label of σ_i is at most the label of σ_j . In particular, σ_j is 0-small among $\sigma_0, \sigma_1, \dots, \sigma_k$ when its label is the unique minimum label, and we encourage readers to focus on this $\ell = 0$ case in their first reading of the proof.

For $0 \leq j \leq k$, we define

$$\begin{aligned} \rho_{j,\ell}^k &:= \Pr(X_j \text{ is } \ell\text{-small among } X_0, \dots, X_k) \\ \rho_{j,\ell}^k(\sigma) &:= \Pr(X_j \text{ is } \ell\text{-small among } X_0, \dots, X_k \mid X_j = \sigma) \end{aligned}$$

Observe that since $X_s \sim \pi$ for all s , we also have that

$$(1) \quad \begin{aligned} \rho_{j,\ell}^k(\sigma) &= \\ \Pr(X_{s+j} \text{ is } \ell\text{-small among } X_s, \dots, X_{s+k} \mid X_{s+j} = \sigma) \end{aligned}$$

We begin by noting two easy facts.

Observation IV.1. $\rho_{j,\ell}^k(\sigma) = \rho_{k-j,\ell}^k(\sigma)$.

Proof. Since $\mathcal{M} = X_0, X_1, \dots$ is stationary and reversible, the probability that $(X_0, \dots, X_k) = (\sigma_0, \dots, \sigma_k)$ is equal to the probability that $(X_0, \dots, X_k) = (\sigma_k, \dots, \sigma_0)$ for any fixed sequence $(\sigma_0, \dots, \sigma_k)$. Thus, any sequence $(\sigma_0, \dots, \sigma_k)$ for which $\sigma_j = \sigma$ and σ_j is a ℓ -small corresponds to an equiprobable sequence $(\sigma_k, \dots, \sigma_0)$ for which $\sigma_{k-j} = \sigma$ and σ_{k-j} is ℓ -small. \square

Observation IV.2. $\rho_{j,2\ell}^k(\sigma) \geq \rho_{j,\ell}^j(\sigma) \cdot \rho_{0,\ell}^{k-j}(\sigma)$.

Proof. Consider the events that X_j is a ℓ -small among X_0, \dots, X_j and among X_j, \dots, X_k , respectively. These events are conditionally independent, when conditioning on the value of $X_j = \sigma$, and $\rho_{j,\ell}^j(\sigma)$ gives the probability of the first of these events, while applying equation (1) with $s = j$ gives that $\rho_{0,\ell}^{k-j}(\sigma)$ gives the probability of the second event.

Finally, when both of these events happen, we have that X_j is 2ℓ -small among X_0, \dots, X_k . \square

We can now deduce that

$$(2) \quad \rho_{j,2\ell}^k(\sigma) \geq \rho_{j,\ell}^j(\sigma) \cdot \rho_{0,\ell}^{k-j}(\sigma) = \rho_{0,\ell}^j(\sigma) \cdot \rho_{0,\ell}^{k-j}(\sigma) \geq (\rho_{0,\ell}^k(\sigma))^2.$$

Indeed, the first inequality follows from Observation IV.2, the equality follows from Observation IV.1, and the final inequality follows from the fact that $\rho_{j,\ell}^k(\sigma)$ is monotone nonincreasing in k for fixed j, ℓ, σ .

Observe now that $\rho_{j,\ell}^k = \mathbf{E} \rho_{j,\ell}^k(X_j)$, where the expectation is taken over the random choice of $X_j \sim \pi$.

Thus taking expectations in (2) we find that

$$(3) \quad \begin{aligned} \rho_{j,2\ell}^k &= \mathbf{E} \rho_{j,2\ell}^k(\sigma) \geq \mathbf{E} \left((\rho_{0,\ell}^k(\sigma))^2 \right) \\ &\geq (\mathbf{E} \rho_{0,\ell}^k(\sigma))^2 = (\rho_{0,\ell}^k)^2. \end{aligned}$$

where the second of the two inequalities is the Cauchy-Schwartz inequality.

For the final step in the proof, we sum the left and right-hand sides (3) to obtain

$$\sum_{j=0}^k \rho_{j,2\ell}^k \geq (k+1)(\rho_{0,\ell}^k)^2$$

If we let ξ_j ($0 \leq j \leq k$) be the indicator variable which is 1 whenever X_j is 2ℓ -small among X_0, \dots, X_k , then $\sum_{j=0}^k \xi_j$ is the number of 2ℓ -small terms, which is always at most $2\ell+1$, so that linearity of expectation gives that

$$(4) \quad 2\ell+1 \geq (k+1)(\rho_{0,\ell}^k)^2,$$

giving that

$$(5) \quad \rho_{0,\ell}^k \leq \sqrt{\frac{2\ell+1}{k+1}}.$$

This proves Theorem I.1, as if X_i is an ε -outlier among X_0, \dots, X_k , then X_i is necessarily ℓ -small among X_0, \dots, X_k for $\ell = \lfloor \varepsilon(k+1)-1 \rfloor \leq \varepsilon(k+1)-1$, and then we have $2\ell+1 \leq 2\varepsilon(k+1)-1 \leq 2\varepsilon(k+1)$. \square

Acknowledgment

We are grateful for helpful conversations with John Nangle, Danny Sleator, and Dan Zuckerman.

-
- [1] A. Gelman and D. B. Rubin, Statistical science pp. 457–472 (1992).
 - [2] A. Gelman and D. R. Rubin, in *Bayesian Statistics* (Oxford University Press, 1992).
 - [3] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **58**, 86 (1987), URL <http://link.aps.org/doi/10.1103/PhysRevLett.58.86>.
 - [4] C. Borgs, J. Chayes, and P. Tetali, Probability Theory and Related Fields **152**, 509 (2012).
 - [5] C. Cooper and A. Frieze, Random Structures and Algorithms **15**, 242 (1999).
 - [6] V. K. Gore and M. R. Jerrum, Journal of Statistical Physics **97**, 67 (1999).
 - [7] J. G. Propp and D. B. Wilson, Random structures and Algorithms **9**, 223 (1996).
 - [8] J. Propp and D. Wilson, Microsurveys in Discrete Probability **41**, 181 (1998).
 - [9] J. A. Fill, in *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing* (ACM, 1997), pp. 688–695.
 - [10] M. Huber, Annals of Applied Probability pp. 734–753 (2004).
 - [11] S. S.-H. Wang, Available at SSRN 2671607 (2015).
 - [12] J. F. Nagle, Election Law Journal **14**, 346 (2015).
 - [13] M. D. McDonald and R. E. Best, Election Law Journal **14**, 312 (2015).
 - [14] J. Chen and J. Rodden, Quarterly Journal of Political Science **8**, 239 (2013).
 - [15] C. Cirincione, T. A. Darling, and T. G. O'Rourke, Political Geography **19**, 189 (2000).
 - [16] P. A. Rogerson and Z. Yang, Social Science Computer Review **17**, 27 (1999).
 - [17] B. Fifield, M. Higgins, K. Imai, and A. Tarr, Tech. Rep., Working Paper. Available at <http://imai.princeton.edu/research/files/redist.pdf> (2015).
 - [18] L. Chenyun Wu, J. Xiaotian Dou, A. Frieze, D. Miller, and D. Sleator, arXiv preprint arXiv:1510.03247 (2015).
 - [19] C. Vaughn, S. Bangia, B. Dou, S. Guo, and J. Mattingly (2016).
 - [20] S. Ansolabehere, M. Palmer, and A. Lee, *Precinct-level election data, Harvard dataverse, v1*, <http://hdl.handle.net/1902.1/21919>.
 - [21] F. Y. Edgeworth, Journal of the Royal Statistical Society **60**, 681 (1897).
 - [22] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times* (American Mathematical Soc., 2009).
 - [23] D. Aldous and J. Fill, *Reversible markov chains and random walks on graphs*.
 - [24] A. Frieze and M. Karoński, *Introduction to random graphs* (Cambridge University Press, 2015).

Supplement: Assessing significance without mixing

S1. PRECINCT DATA

Precinct level voting data and associated shape files were obtained from the Harvard Election Data Archive (<http://projects.iq.harvard.edu/eda/home>) [20]. The data for Pennsylvania contains 9256 precincts. The data was altered in two ways: 258 precincts that were contained within another precinct were merged and 79 precincts that were not contiguous were split into continuous areas, with voting and population data distributed proportional to the area. The result was a set of 9060 precincts. All geometry calculations and manipulations were accomplished in R with “maptools”, “rgeos”, and “BARD” R packages. The final input to the Markov chain is a set of precincts with corresponding areas, neighbor precincts, the length of the perimeter shared with each neighbor, voting data from 2012, and the current Congressional district the precinct belongs to.

S2. VALID DISTRICTINGS

We restrict our attention to districtings satisfying 4 restrictions, each of which we describe here.

A. Contiguous districts

A valid districting must have the property that each of its districts is contiguous. In particular, two precincts are considered adjacent if the length of their shared perimeter is nonzero (in particular, precincts meeting only at a point are *not* adjacent), and a district is contiguous if any pair of precincts is joined by a sequence of consecutively adjacent pairs.

B. Simply connected districts

A valid districting must have the property that each of its districts is simply connected. Roughly speaking, this means the district cannot have a “hole”. Precisely, a district is simply connected if for any circuit of precincts in the district, all precincts in the region bounded by the circuit also lie in the district.

Apart from aesthetic reasons for insisting that districtings satisfy this condition, there is also a practical reason: it is easier to have a fast local check for contiguity when one is also maintaining that districtings are simply connected.

C. Small population difference

According to the “one person, one vote” doctrine, Congressional districts for a state are required to be roughly equal in population. In the current districting of Pennsylvania, for example, the maximum difference in district population from the average population is roughly 1%. Our chain can use different tolerances for population difference between districts and the average, and the tolerances used in the runs below are indicated.

D. Compactness

If districtings were drawn randomly with only the first three requirements, the result would be districtings in which districts have very complicated, fractal-like structure (since most districtings have this property). The final requirement on valid districtings prevents this, by ensuring that the districts in the districting have a reasonably nice shape. This requirement on district shape is commonly termed “compactness”, is explicitly required of Congressional districts by the Pennsylvania constitution.

Although compactness of districts does not have a precise legal definition, various precise metrics have been proposed to quantify the compactness of a given district mathematically. One of the simplest and most commonly used metrics is the Polsby-Popper metric, which defines the compactness of a district as

$$C_D = \frac{4\pi A_D}{P_D^2},$$

where A_D and P_D are the area and perimeter of the district D . Note that the maximum value of this measure is 1, which is achieved only by the disc, as a result of the isoperimetric inequality. All other shapes have compactness between 0 and 1, and smaller values indicate more “contorted” shapes.

Perhaps the most straightforward use of this compactness measure is to enforce some threshold on compactness, and require valid districtings to have districts whose compactness is above that lower bound. (For consistency with our other metrics, we actually impose an upper bound on the reciprocal $1/C_D$ of the Polsby-Popper compactness C_D of each district D .) In the table of runs given in Section S5, this is the L^∞ compactness metric.

One drawback of using this method is that the current districting of Pennsylvania has a few districts which have very low compactness values (they are much stranger looking than the other districts). Applying this restriction will allow all 18 districts to be as bad as the threshold chosen, so that, in particular, we will be sampling districtings from space in which all 18 districts may be as bad as the worst district in the current plan. In fact, because there are more noncompact regions than compact ones, one expects that in a typical such districting, all 18 districts would be nearly as bad as the currently worst example.

To address this issue, and also to demonstrate the robustness of our finding for the districting question, we also consider some alternate restrictions on the districting, which measure how reasonable the districting as a whole is with regard to compactness. For example, one simple measure of this is to have a threshold for the maximum allowable sum

$$\frac{1}{C_1} + \cdots + \frac{1}{C_{18}}$$

of the inverse compactness values of the 18 districts. This is the L^1 metric in the table in Section S5. Similarly, we could have a threshold for the maximum allowable sum of squares

$$\frac{1}{C_1^2} + \cdots + \frac{1}{C_{18}^2}.$$

This is the L^2 metric in the table. Finally, we can have a simple condition which simply ensures that the total perimeter

$$P_1 + \cdots + P_{18}$$

is less than some threshold.

E. Other possible constraints

It is possible to imagine many other reasonable constraints on valid districtings. For example, the PA constitution currently requires of districtings for the Pennsylvania Senate and Pennsylvania House of Representatives that:

Unless absolutely necessary no county, city, incorporated town, borough, township or ward shall be divided in forming either a senatorial or representative district.

There is no similar requirement for U.S. Congressional districts in Pennsylvania, which is what we consider here, but it is still a reasonable constraint to consider.

There are also interesting legal-questions about the extent to which Majority-Minority districts (in which an ethnic minority is an electoral majority) are either required to be intentionally created, or forbidden to be intentionally created. On the one hand, the U.S. Supreme Court ruled in *Thornburg v. Gingles* (1986) that in certain cases, a geographically concentrated minority population is entitled to a Congressional district in which it constitutes a majority. On the other hand, several U.S. Supreme Court cases (*Shaw v. Reno* (1993), *Miller v. Johnson* (1995), and *Bush v. Vera* (1996)) Congressional districtings were thrown out because they contained intentionally-drawn Majority-Minority districts which were deemed to be a “racial gerrymander”. In any case, we have not attempted to answer the question of whether or how the existence of Majority-Minority districts should be quantified. (We suspect that the unbiased procedure of drawing a random districting is probably acceptable under current Majority-Minority district requirements, but in any case, our main intent is to demonstrate the application of the $\sqrt{\varepsilon}$ test.)

Importantly, we emphasize that any constraint on districtings which can be precisely defined (i.e., by giving an algorithm which can identify whether a districting satisfies the constraint) can be used in the Markov Chain setting in principle.

S3. THE MARKOV CHAIN

The Markov chain \mathcal{M} we use has as its state space Σ the space of all valid districtings (with 18 districts) of Pennsylvania. Note that there is no simple way to enumerate these, and there is an enormous number of them.

A simple way to define a Markov chain on this state space is to transition as follows:

1. From the current state, determine the set S of all pairs (ρ, D) , where ρ is a precinct in some district D_ρ , and $D \neq D_\rho$ is a district which is adjacent to ρ . Let N_S denote the size of this set.
2. From S , choose one pair (ρ_0, D_0) uniformly at random.
3. Change the district membership of ρ_0 from D_{ρ_0} to D_0 , **if** the resulting district is still valid.

Let the Markov chain with these transition rules be denoted by \mathcal{M}_0 . This is a perfectly fine reversible Markov Chain to which our theorem applies, but the uniform distribution on valid districtings is not stationary for \mathcal{M}_0 , and so we cannot use \mathcal{M}_0 to make comparisons between a presented districting and a uniformly random valid districting.

A simple way to make the uniform distribution stationary is to “regularize” the chain, that is, to modify the chain so that the number of legal steps from any state is equal. (This is not the case for \mathcal{M}_0 , as the number of precincts on the boundaries of districts will vary from districting to districting.) We do this by adding loops to each possible state. In particular, using a theoretical maximum N_{\max} on the possible size of N_S for any district, we modify the transition rules as follows:

1. From the current state, determine the set S of all pairs (ρ, D) , where ρ is a precinct in some district D_ρ , and $D \neq D_\rho$ is a district which is adjacent to ρ . Let N_S denote the size of this set.
2. With probability $1 - \frac{N_S}{N_{\max}}$, remain in the current state for this step. With probability $\frac{N_S}{N_{\max}}$, continue as follows:
3. From S , choose one pair (ρ_0, D_0) uniformly at random.
4. Change the district membership of ρ_0 from D_{ρ_0} to D_0 , **if** the resulting district is still valid. If it is not, remain in the current district for this set.

In particular, with this modification, each state has exactly N_{\max} possible transitions, which are each equally likely; many of these transitions are loops back to the same state. (Some of these loops arise from Step 2, but some also arise when the **if** condition in Step 4 fails.)

S4. THE LABEL FUNCTION

In principle, any label function ω could be used in the application of the $\sqrt{\varepsilon}$ test; note that Theorem I.1 places no restrictions on ω . Thus when we choose which label function to use, we are making a choice based on what is likely to achieve good significance, rather than what is valid statistical reasoning (subject to the caveat discussed in the last paragraph of this section). To choose a label function which was likely to allow good statistical power, we want to have a function which:

1. is likely very different for a gerrymandered districting compared with a typical districting, and
2. is sensitive enough that small changes in the districting might be detected in the label function.

While the role of the first condition in achieving outlier status is immediately obvious, the second property is also crucial to detecting significance with our test, which makes use of trajectories which may be quite small compared with the mixing time of the chain. For the $\sqrt{\varepsilon}$ test to succeed at demonstrating significance, it is not enough for the presented state σ_0 to actually be an outlier against π with respect to ω ; this must also be detectable on trajectories of the fixed length k , which may well be too small to mix the chain. This second property discourages the use of “coarse grained” label functions such as the number of seats out of 18 the Democrats would hold with the districting in question, since many swaps would be needed to shift a representative from one party to another.

We considered two label functions for our experiments (each selected with the above desired properties in mind) to demonstrate the robustness of our framework. The first label function ω_{var} we used is simply the negative of the variance in the proportions of Democrat voters among the districts. Thus, given a districting σ , $\omega_{\text{var}}(\sigma)$ is calculated as

$$\omega_{\text{var}}(\sigma) = - \left(\frac{\delta_1^2 + \delta_2^2 + \cdots + \delta_{18}^2}{18} - \left(\frac{\delta_1 + \delta_2 + \cdots + \delta_{18}}{18} \right)^2 \right)$$

where for each $i = 1, \dots, 18$, δ_i is the fraction of voters in that district which voted for the Democrat presidential candidate in 2012. We suspect that the variance is a good label function from the standpoint of the first characteristic listed above, but a great label function from the standpoint of the second characteristic. Recall that in practice, gerrymandering is accomplished by packing the voters of one party into a few districts, in which they make up an overwhelming majority. This, naturally, results in a high-variance vector of party proportions in the districts. However, high-variance districtings can exist which do not favor either party (note, for example, that the variance is symmetric with respect to Democrats and Republicans, ignoring third-party affiliations). This means that for a districting which is biased against π due to a partisan gerrymander to “stand out” as an outlier, it must have especially high variance. In particular, statistical significance will be weaker than it might be for a label function which is more strongly correlated with partisan gerrymandering. On the other hand, ω_{var} can detect very small changes in the districting, since essentially every swap will either increase or decrease the variance. Indeed, for the run of the chain corresponding to the L^∞ constraint (Section S5), $\omega_{\text{var}}(X_0)$ was strictly greater than $\omega_{\text{var}}(X_i)$ for the entire trajectory ($1 \leq i \leq 2^{40}$). That is, for the L^∞ constraint, the current districting of Pennsylvania was the absolute worst districting seen according to ω_{var} among the more than trillion districtings on the trajectory.

The second label function we considered is calculated simply as the difference between the median and the mean of the ratios $\delta_1, \dots, \delta_{18}$. This simple metric, called the “Symmetry Vote Bias” by McDonald and Best [13] and denoted as ω_{MM} by us, is closely tied to the goal of partisan gerrymandering. As a simple illustration of the connection, we consider the case where the median of the ratios $\delta_1, \dots, \delta_{18}$ is close to $\frac{1}{2}$. In this case, the mean of the δ_i 's tracks the fraction of votes the reference party wins in order to win half the seats. Thus a positive Symmetry Vote Bias corresponds to an advantage for the reference party, while a negative Symmetry Vote Bias corresponds to an disadvantage. The use of the Symmetry Vote Bias in evaluating districtings dates at least to the 19th century

work of Edgeworth [21]. These features make it an excellent candidate from the standpoint of our first criterion: gerrymandering is very likely to be reflected in outlier-values of ω_{MM} .

On the other hand, ω_{MM} is a rather slow-changing function, compared with ω_{var} . To see this, observe that in the calculation

$$\text{Symmetry Vote Bias} = \text{median} - \text{mean},$$

the mean is essentially fixed, so that changes in ω_{MM} depend on changes in the median. In initial changes to the districting, only changes to the two districtings giving rise to the median (two since 18 is even) can have a significant impact on ω_{MM} . (On the other hand, changes to any district directly affect ω_{var} .)

It is likely possible to make better choices for the label function ω to achieve better significance. For example, the metric B_G described by Nagle [12] seems likely to be excellent from the standpoints of conditions 1 and 2 simultaneously. However, we have restricted ourselves to the simple choices of ω_{var} and ω_{MM} to clearly demonstrate our method, and to make it obvious that we have not tried many labeling functions before finding some that worked (in which case, a multiple hypothesis test would be required).

One point to keep in mind is that often when applying the $\sqrt{\varepsilon}$ test—including in the present application to gerrymandering—we will wish to apply the test, and thus need to define a label function, after the presented state σ_0 is already known. In these cases, care must be taken to choose a “canonical” label function ω , so that there is no concern that ω was carefully crafted in response to σ_0 (in this case, a multiple hypothesis correction would be required, for the various possible ω 's which could have been crafted, depended on σ_0). ω_{var} and ω_{MM} are excellent choices from this perspective: the variance is a common and natural function on vectors, and the Symmetry Vote Bias has an established history in the evaluation of gerrymandering (and in particular, a history which predates the present districting of Pennsylvania).

S5. RUNS OF THE CHAIN

In Table I we give the results of the 8 runs of the chain under various conditions. Each run was for $k = 2^{40}$ steps. Code and input data for our Markov chain are available at the corresponding author's website (<http://math.cmu.edu/~wes>).

Generally, after an initial “burn-in” period, we expect the chain to (almost) never again see states as unusual as the current districting of Pennsylvania, which means that we expect the test to demonstrate significance proportional to the inverse of the square root of the number of steps (i.e., the p -value at 2^{42} steps should be half the p -value at 2^{40} steps). In particular, for the L^1 , L^2 , and L^∞ constraints, these runs never revisited states as bad as the initial state after 2^{21} steps for the ω_{MM} label, and after 2^6 steps for the ω_{var} label. Note that this agrees with our guess that ω_{var} had the potential to change more quickly than ω_{MM} . The perimeter constraint did revisit enough states as bad as the the given state with respect to the ω_{var} label to adversely affect its p -value with respect to the ω_{var} label.

population threshold	compactness measure	compactness threshold	initial value	(steps) $k =$	label function	ε -outlier at $\varepsilon =$	significant at $p =$
2%	perimeter	≤ 125	121.2...	2^{40}	ω_{var}	$3.0974 \cdot 10^{-8}$	$2.4889 \cdot 10^{-4}$
					ω_{MM}	$5.7448 \cdot 10^{-10}$	$3.3896 \cdot 10^{-5}$
2%	L^1	≤ 160	156.4...	2^{40}	ω_{var}	$5.0123 \cdot 10^{-11}$	$1.0012 \cdot 10^{-5}$
					ω_{MM}	$5.6936 \cdot 10^{-10}$	$3.3745 \cdot 10^{-5}$
2%	L^2	≤ 44	43.06...	2^{40}	ω_{var}	$8.2249 \cdot 10^{-11}$	$1.2826 \cdot 10^{-5}$
					ω_{MM}	$6.8038 \cdot 10^{-10}$	$3.6888 \cdot 10^{-5}$
2%	L^∞	≤ 25	24.73...	2^{40}	ω_{var}	$3.3188 \cdot 10^{-13}$	$8.1472 \cdot 10^{-7}$
					ω_{MM}	$6.9485 \cdot 10^{-8}$	$3.7279 \cdot 10^{-4}$

TABLE I: Runs of the redistricting Markov Chain, with results of the $\sqrt{\varepsilon}$ test.

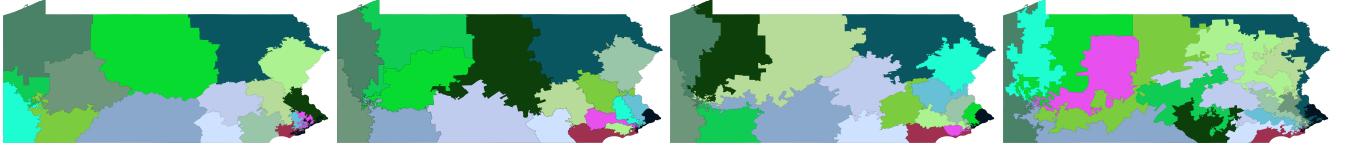


FIG. S1: The last state from each of the above runs of the chain (perimeter, L^1 , L^2 , and L^∞ , respectively). Note that the L^∞ districting is quite ugly; with this notion of validity, every district among the 18 is allowed to be as noncompact as the worst district in the current Pennsylvania districting. The perimeter constraint produces a districting which appears clean at a large scale, but allows rather messy city districts, since they contribute only moderately to the perimeter anyways. The L^1 and L^2 constraints are more balanced. Note that none of these districtings should be expected to be geometrically “nicer” than the current districting of Pennsylvania. Indeed, the point of our Markov Chain framework is to compare the present districting of Pennsylvania with other “just as bad” districtings, to observe that even among this set, the present districting is atypical.

This may reflect our guess that the ω_{var} label is worse than the ω_{MM} label in terms of how easily it can distinguish gerrymandered districtings from random ones.

The parameters for the first row were used for Fig 2 of the paper.

One quick point: although we have experimented here with different compactness measures, there is no problem of multiple hypothesis correction to worry about, as *every* run we encountered produces strong significance for the bias of the initial districting. The point of experimenting with the notion of compactness is to demonstrate that this a robust framework, and that the finding is unlikely to be sensitive to minor disagreements over the proper definition of the set of valid districtings.

S6. AN EXAMPLE WHERE $p \approx \sqrt{\varepsilon}$ IS BEST POSSIBLE

It might be natural to suspect that observing that a presented state σ is an ε -outlier on a random trajectory σ is significant something like $p \approx \varepsilon$ rather than the $p \approx \sqrt{\varepsilon}$ established by Theorem I.1. However, since Theorem I.1 places no demand on the mixing rate of \mathcal{M} , it should instead seem remarkable that any significance can be shown in general, and indeed, we show by example in this section that significance at $p \approx \sqrt{\varepsilon}$ is essentially best possible.

Let N be some large integer. We let \mathcal{M} be the Markov chain where X_0 is distributed uniformly in $[0, 1, 2, \dots, N - 1]$, and, for any $i \geq 1$, X_i is equal to $X_{i-1} + \zeta_i$ computed modulo N , where ζ_i is 1 or -1 with probability $\frac{1}{2}$. Note that the chain is stationary and reversible.

If N is chosen large relative to k , then with probability arbitrarily close to 1, the value of X_0 is at distance greater than k from 0 (in both directions). Conditioning on this event, we have that X_0 is minimum among X_0, \dots, X_k if and only if all the partial sums $\sum_{i=1}^j \zeta_i$ are positive. This is just the probability that a k -step 1-dimensional random walk from the origin takes a first step to the right and does not return to the origin. This is a classical problem in random walks, which can be solved using the reflection principle.

In particular, for k even, the probability is given by

$$\frac{1}{2^{k+1}} \binom{k}{k/2} \sim \frac{1}{\sqrt{2\pi k}}.$$

Since being the minimum out of X_0, \dots, X_k corresponds to being an ε -outlier for $\varepsilon = \frac{1}{k+1}$, this example shows that the probability of being an ε -outlier can be as high as $\sqrt{\varepsilon}/2\pi$.

The best possible value of the constant in the $\sqrt{\varepsilon}$ test appears to be an interesting problem for future work.

S7. NOTES ON STATISTICAL POWER

The effectiveness of the $\sqrt{\varepsilon}$ test depends on the availability of a good choice for ω , and the ability to run the test for long enough (in other words, choose k large enough) to detect that the presented state is a local outlier.

It is possible, however, to make a general statement about the power of the test when k is chosen large relative to the actual mixing time of the chain. Recall that one potential application of the test is in situations where the mixing time of the chain is actually accessible through reasonable computational resources, even though this can't be proved rigorously, because theoretical bounds on the mixing time are not available. In particular, we do know that the test is very likely to succeed when k is sufficiently large, and $\omega(\sigma_0)$ is atypical.

Theorem S7.1. *Let \mathcal{M} be a reversible Markov Chain on Σ , and let $\omega : \Sigma \rightarrow \mathbb{R}$ be arbitrary. Given σ_0 , suppose that for a random state $\sigma \sim \pi$, $\Pr(\omega(\sigma) \leq \omega(\sigma_0)) \leq \varepsilon$. Then with probability at least*

$$\rho \geq 1 - \left(1 + \frac{\varepsilon k}{10\tau_2}\right) \frac{1}{\sqrt{\pi_{\min}}} \exp\left(\frac{-\varepsilon^2 k}{20\tau_2}\right)$$

$\omega(\sigma)$ is an 2ε -outlier among $\omega(\sigma_0), \omega(\sigma_1), \dots, \omega(\sigma_k)$, where $\sigma_0, \sigma_1, \dots$ is a random trajectory starting from σ_0 .

Here τ_2 is the *relaxation time* for \mathcal{M} , defined as $1/(1 - l_2)$, where l_2 is the second eigenvalue of \mathcal{M} . τ_2 is thus the inverse of the spectral gap for \mathcal{M} , and is intimately related to the mixing time of \mathcal{M} [22–24]. The probability ρ in Theorem S7.1 converges exponentially quickly to 1, and, moreover, is very close to 1 once k is large relative to τ_2 . In particular, Theorem S7.1 shows that the $\sqrt{\varepsilon}$ test *will work* when the test is run for long enough. Of course, one strength of the $\sqrt{\varepsilon}$ test is that it can sometimes demonstrate bias even when k far too small to mix the chain, as is almost certainly the case for our application to gerrymandering. When these short- k runs are successful at detecting bias is of course dependent on the relationship of the presented state σ_0 and its local neighborhood in the chain.

Theorem S7.1 is an application of the following theorem of Gillman:

Theorem S7.2. *Let $\mathcal{M} = X_0, X_1, \dots$ be a reversible Markov Chain on Σ , let $A \subseteq \Sigma$, and let $N_k(A)$ denote the number of visits to A among X_0, \dots, X_k . Then for any $\gamma > 0$,*

$$\Pr(N_k(A)/n - \pi(A) > \gamma) \leq \left(1 + \frac{\gamma n}{10\tau_2}\right) \sqrt{\sum_{\sigma} \frac{\Pr(X_0 = \sigma)^2}{\pi(\sigma)}} \exp\left(\frac{-\gamma^2 n}{20\tau_2}\right).$$

Proof of Theorem S7.1. We apply Theorem S7.2, with A as the set of states $\sigma \in \Sigma$ such that $\omega(\sigma) \leq \omega(\sigma_0)$, with $X_0 = \sigma_0$, and with $\gamma = \varepsilon$. By assumption, $\pi(A) \leq \varepsilon$, and Theorem S7.2 gives that

$$\Pr(N_k(A)/k > 2\varepsilon) \leq \left(1 + \frac{\varepsilon k}{10\tau_2}\right) \sqrt{\frac{1}{\pi_{\min}}} \exp\left(\frac{-\varepsilon^2 k}{20\tau_2}\right).$$

□

S8. A RESULT FOR SMALL VARIATION DISTANCE

In this section, we give a corollary of Theorem I.1 which applies to the setting where X_0 is distributed not as a stationary distribution π , but instead with small total variation distance to π .

The *total variation distance* $\|\rho_1 - \rho_2\|_{\text{TV}}$ between probability distributions ρ_1, ρ_2 on a probability space Ω is defined to be

$$(S1) \quad \|\rho_1 - \rho_2\|_{\text{TV}} := \sup_{E \subseteq \Omega} |\rho_1(E) - \rho_2(E)|.$$

Corollary S8.1. *Let $\mathcal{M} = X_0, X_1, \dots$ be a reversible Markov chain with a stationary distribution π , and suppose the states of \mathcal{M} have real-valued labels. If $\|X_0 - \pi\|_{\text{TV}} \leq \varepsilon_1$, then for any fixed k , the probability that the label of X_0 is an ε -outlier from among the list of labels observed in the trajectory $X_0, X_1, X_2, \dots, X_k$ is at most $\sqrt{2\varepsilon} + \varepsilon_1$.*

The theorem is intuitively clear; we provide a formal proof below for completeness.

Proof. If ρ_1, ρ_2 , and τ are probability distributions, then we have that the product distributions (ρ_1, τ) and (ρ_2, τ) satisfy

$$(S2) \quad \|(\rho_1, \tau) - (\rho_2, \tau)\|_{\text{TV}} = \|\rho_1 - \rho_2\|_{\text{TV}}.$$

Our plan now is to split the randomness in the trajectory X_0, \dots, X_k of the Markov Chain into two independent sources: the initial distribution is $X_0 \sim \rho$, and τ is the uniform distribution on sequences of length k of real numbers r_1, r_2, \dots, r_k in $[0, 1]$. We can view the distribution of the trajectory X_0, X_1, \dots, X_k as the product (ρ, τ) by using sequences of reals r_1, \dots, r_k to choose transitions in the chain; from $X_i = \sigma_i$, if there are L transitions possible, with probabilities p_1, \dots, p_L , then we make the t th possible transition if $r_i \in [p_1 + \dots + p_{t-1}, p_1 + \dots + p_{t-1} + p_t]$.

Now we have from (S2) that if $\|\rho - \pi\|_{\text{TV}} \leq \varepsilon_1$, then $\|(\rho, \tau) - (\pi, \tau)\|_{\text{TV}} \leq \varepsilon_1$. Therefore, any event which would happen with probability at most p for the sequence X_0, \dots, X_k when $X_0 \sim \pi$ must happen with probability at most $p + \varepsilon_1$ when $X_0 \sim \rho$ where $\|\rho - \pi\|_{\text{TV}} \leq \varepsilon_1$. This gives the Corollary. \square