Theory and Computation for Modern Probabilistic Models

Jackson Loper
B.A. Brown University, RI, 2009
M.S, Brown University, RI, 2012

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the Division of Applied Mathematics at Brown University

Providence, Rhode Island
May 2017

This dissertation by Jackson Loper is
accepted in its present form by the Division
of Applied Mathematics as satisfying the
dissertation requirement for the degree of
Doctor of Philosophy.

Date	Stuart Geman, Ph.D., Advisor

Recommended to the Graduate Council

Date	Matthew Harison, Ph.D., Reader
Date	Kavita Ramanan, Ph.D., Reader

Approved by the Graduate Council

Date	Andrew Campbell, Dean of the Graduate School

The Vita of Jackson Loper

Jackson Loper graduated from Tandem Friends School in 2004. He began attending Brown University in 2004. In 2005 he took a leave of absence to work for one year at Ready to Learn Providence, building social networks with providers of early childhood education for low-income individuals ages 3-5. In 2009 he received a Bachelors of Arts in Applied Mathematics from Brown University. After graduation, he spent one year doing research with Stuart Geman and then decided to join the Brown University Division of Applied Mathematics as a graduate student.

Acknowledgments

Thank you Rick Benjamin, for your unfailing commitment to deliberation of the social and ethical implications of all our actions - mathematical, poetical, or otherwise. Thank you Janet Cooper-Nelson for helping me see the historical context of my work - whether in the significance of Charles Brenton Fisk's flight from the Manhattan Project to organ making, in the global role of the apparatus of United States technology, or in the hundred stories you have shared over the years. Thank you Brady Earnhart for your elevating instruction on the topic of concrete nouns - a valuable asset in any abstract endeavor and an invaluable asset in the endeavor of this thesis. Thank you Catherine Sutton for making sure my body didn't fall apart.

Thank you Stuart Geman for your infectious excitement, your advice, and your forbearance throughout the last eight years. Thank you Matthew Harrison for your many insights and for your kindness. Thank you Kavita Ramanan for giving me an exuberant first introduction to rigorous probability theory and a reading of my last work here at Brown.

Thank you 周光耀 for keeping me from completely losing my mind.

Finally, I give thanks for the consistently unbelievable blessing that is my family. You have been the ultimate inspiration in everything that really matters: my brothers in math, my dog in Zen, and my parents in prioritizing love above everything else. Speaking of which: Emma Beth Ellsworth Gross, you're the best.

Contents

1. Finite Approximations for Point Processes

Point processes arise in many different contexts. In neuroscience they may model the firing of neurons (e.g. [25]), in geostatistics they may model the presence of notable aspects of the terrain (e.g. [35]), in computer vision they may model the presence of objects being detected (e.g. [10]), and in general they provide a tractable approach to model random sets.

One very general class of point processes arise from distributions which are absolutely continuous with respect to some general Poisson point process. We will refer to these as Poisson-Absolutely-Continuous (PAC) point processes. Among many others, this class includes the Gibbs processes, most Cox processes, most determinantal processes, and area-interaction-type processes. It also includes many posterior distributions on latent point processes, assuming the prior model is itself a PAC process; that is, if X is a PAC process and $Y|X = x \sim q(\cdot; x)$, then the distribution on $X|Y = y$ is itself often a PAC process.

One difficulty in analyzing such processes is that the the number of points is generally not fixed. This makes it impossible to directly apply common techniques such as Gibbs sampling to analyze the process. However, there are some general-purpose techniques available for analyzing point processes. The Geyer-Moller method is perhaps the oldest general-purpose technique for drawing samples from a point process. This method is a straightforward application of reversible-jump monte-carlo to the problem at hand (cf.

[15]). In some cases, however, this method was found to be somewhat computationally burdensome. Various improvements have been made over time, such as focusing on small neighborhoods to speed up the exploration of the space (e.g. [28]), making clever choices of the birth-of-point and death-of-point proposal distributions (e.g. [10]), developing efficient parallelization (e.g. [1]), and so-forth. However, in order to deal with the fact that the number of points is not fixed, all of these methods are forced to go through all the careful details of a reversible-jump method. These details complicate both one's thinking and one's algorithms. It also generally leads to a focus on Markov-chain monte-carlo methods for obtaining samples, since standard variational Bayesian techniques (e.g. mean field) may be difficult to apply directly.

To simplify thinking and make it easier to apply a wider variety of inference methods, we propose a fairly obvious solution: represent the process with a fixed number of random variables. We will see exactly what we mean by "fixed number of random variables" shortly. It necessarily involves some approximation. However, it certainly doesn't mean that we need to fix the number of points. Moreover, if we allow ourselves *enough* variables, we can ensure that our approximation error is negligible. The benefit of fixing the number of variables is simple: all of the standard techniques for probabilistic inference become applicable without any modification. Here we will present one simple way to find a good approximation to a point process, using only a fixed number of variables.

Before we look at our proposed approximation, we will take a moment to consider what is already known about approximating point processes. There is a wide variety of literature on the subject. However, much of this literature focuses on showing how the various relatively simple distributions emerge from complex situations. For example, there are papers investigating how the stationary distribution of various birth-death processes can approach a Poisson process (e.g. [4]) or compound Poisson processes (e.g. [5]), how so-called "hard-core" processes approach Poisson processes (e.g. [8]), and how the stationary distribution of various birth-death processes can approach distributions for

which the positions of points are independent and identically distributed conditioned on the number of points (e.g. [36]). By contrast, our focus is more computational and more general. We will start with some general PAC point process and endeavor to approximate it with a simple, fixed-variable model.

The basic idea of our finite-variable approximation is pretty straightforward. We first observe that it is easy to sample a Poisson point process if we are allowed an unbounded number of variables. Take $\eta \sim \text{Poisson}(\lambda)$, and $X_1, X_2 \cdots$ an infinite sequence of i.i.d random variables; we can then produce

$$X = \sum_{i=1}^{\eta} \delta_{X_i}$$

where δ_α indicates the delta measure, $\delta_\alpha(A) = \mathbb{I}_{\alpha \in A}$. The variable X is said to be a Poisson point process. We next observe that it is pretty easy to approximate X with only a finite number of variables. Fix some modestly large constant N. We take N independent variables $\xi_1, \xi_2 \cdots \xi_N \in \{0, 1\}$, N independent variables $\tilde{X}_1 \cdots \tilde{X}_n$, and define

$$\tilde{X} = \sum_{i=1}^{N} \xi_i \delta_{\tilde{X}_i}$$

Note that N is not a random variable here. Instead, the number of points in $\tilde{X}$ is controlled by the $\{\xi_i\}$ variables. Also note that $\left\{\tilde{X}_i\right\}$ variables are not necessarily identically distributed, nor are the $\{\xi_i\}$ variables. Despite all this, the distribution of X and the distribution of $\tilde{X}$ may nonetheless be quite similar. The accuracy of the approximation $X \overset{(d)}{\approx} \tilde{X}$ is of course limited if N is small. However, as long as $\mathbb{E}\left[\sum \xi_i\right] = \lambda$ and N is large enough that we can make $\sum \mathbb{E}\left[\xi_i\right]^2 < \epsilon$, Le Cam's theorem immediately tells us that

$$\sum_n \left| \mathbb{P}\left(\tilde{X}(\mathcal{P}) = n\right) - \mathbb{P}\left(X(\mathcal{P}) = n\right) \right| < 2\epsilon$$

Here $X(\mathcal{P})$ and $\tilde{X}(\mathcal{P})$ indicate the total number of points in X and $\tilde{X}$, respectively. If we further require that $\sum_{i=1}^{N} \mathbb{E}[\xi_i] \mathbb{P}\left(\tilde{X}_i \in A\right) = \mathbb{P}(X_i \in A)$ for every measurable $A \subset \mathcal{P}$ and make $\sup_i \mathbb{E}[\xi_i]$ sufficiently small, the total variation between $\tilde{X}$ and X can be made arbitrarily small (we will show this momentarily). This gives us a way to approximate Poisson point processes. What's more, we can use this building block to approximate more sophisticated point processes, by suitably reweighting the joint distribution of $\left\{\left(\xi_i, \tilde{X}_i\right)\right\}$.

Before we continue, let us fix some notation.

- Let $(\mathcal{P}, \mathcal{B}, \lambda)$ denote some finite Radon measure space, i.e. $\mathcal{P}$ is endowed with a locally finite, inner regular Hausdorff topology, and $\mathcal{B}$ are the corresponding borel sets.

- We will be working with random objects whose support lies in the space of random measures, i.e.
$$\mathfrak{N}_n = \left\{\sum_{i=1}^{n} \delta_{\alpha_i}\right\}_{\alpha \in \mathcal{P}^n} \qquad \mathfrak{N} = \bigcup_{n=0}^{\infty} \mathfrak{N}_n$$
We will take the measurable subsets of $\mathfrak{N}$ to be the natural ones inherited from $(\mathcal{P}, \mathcal{B})$, i.e. that is, those generated by
$$\left\{\sum_{i=1}^{n} \delta_{\alpha_i} : \ \alpha_i \in A_i\right\}_{\substack{A_i \in \mathcal{B},\ i \in 1 \cdots n \\ n \in 0,1,2\cdots}}$$

- We say that $X \sim \mathrm{P}oissonProcess(\mathcal{P}, \mathcal{B}, \lambda)$ if X is a Poisson process with intensity λ. That is, for any bounded measurable $\phi : \mathfrak{N} \to \mathbb{R}^+$, we have that
$$\mathbb{E}[\phi(X)] = \sum_{n=0}^{\infty} \frac{e^{-\lambda(\mathcal{P})}}{n!} \int_{\mathcal{P}^n} \phi\left(\sum_{i=1}^{n} \delta_{\alpha_i}\right) \lambda^n(d\alpha_{1\cdots n})$$

- We say that $Y \sim \mathrm{P}ACProcess(\mathcal{P}, \mathcal{B}, \lambda, f)$ if

$$\mathbb{E}\left[\phi(Y)\right] = \mathbb{E}\left[f(X)\phi(X)\right]$$

for any bounded measurable function $\phi : \mathfrak{N} \to \mathbb{R}^+$ and any $X \sim \mathrm{PoissonProcess}(\mathcal{P}, \mathcal{B}, \lambda)$.

With this notation in hand, we can now carefully define the finite-variable approximations which we are proposing.

Definition. Let $M = (g_1 \cdots g_N)$ denote any collection of positive functions on $(\mathcal{P}, \mathcal{B}, \lambda)$ such that $\sum_{i=1}^{N} g_i(\alpha) = 1,\ \forall \alpha \in \mathcal{P}$ and $\int g_i(\alpha)\lambda(d\alpha) < 1,\ \forall i \in 1 \cdots N$.

- We say that $\tilde{X}$ is a **Poisson Localization,** $\tilde{X} \sim \mathrm{P}oissonLocalization(\mathcal{P}, \mathcal{B}, \lambda, M)$ if its distribution can be expressed by the following equations:

$$\begin{aligned}
\lambda_i &\triangleq \int g_i(\alpha)\lambda(d\alpha) \\
\xi_i &\sim \mathrm{Bernoulli}\left(\lambda_i\right) \qquad \text{independently } \forall i \\
\tilde{X}_i &\sim \left[\alpha \mapsto \frac{1}{\lambda_i} g_i(\alpha)\lambda(d\alpha)\right] \qquad \text{independently } \forall i \\
\tilde{X} &\triangleq \sum_{i=1}^{N} \delta_{X_i}\xi_i
\end{aligned}$$

  Two remarks at this time:

  - The expression for the distribution of $\tilde{X}_i$ may be a bit hard on the eyes. The intuition is fairly simple, however: it is the measure λ weighted by a renormalized version of g_i.
    - We emphasize again that N is not random. It is a fixed constant of the localization approximation. The number of points is not controlled by N, but rather by the binary variables $\xi_1, \xi_2 \cdots \xi_N \in \{0, 1\}$.
  - We say that $\tilde{Y}$ is a **PAC Localization,** $\tilde{Y} \sim \mathrm{P}ACLocalization(\mathcal{P}, \mathcal{B}, \lambda, f, M)$

if

$$\mathbb{E}\left[\phi(\tilde{Y})\right] = \frac{\mathbb{E}\left[f(\tilde{X})\phi(\tilde{X})\right]}{\mathbb{E}\left[f(\tilde{X})\right]}$$

for any bounded measurable function $\phi : \mathfrak{N} \to \mathbb{R}^+$ and some $\tilde{X} \sim \mathrm{P}oissonLocalization(\mathcal{P}, \mathcal{B}, \lambda, M$
As with the Poisson localization, every such $\tilde{Y}$ may be written as $\tilde{Y} = \sum \zeta_i \delta_{\tilde{Y}_i}$.
It is worthwhile to now take a moment to consider the following question: what is the correct distribution of the variables $\left\{\left(\zeta_i, \tilde{Y}_i\right)\right\}$ such that $\sum \zeta_i \delta_{\tilde{Y}_i} \sim \mathrm{P}ACLocalization(\mathcal{P}, \mathcal{B}, \lambda, f, M)$? The answer can be readily expressed as a density with respect to $\{0,1\}^N \times (\mathcal{P}, \mathcal{B}, \lambda)^N$, namely

$$p(\zeta_{1\cdots N}, d\tilde{y}_{1\cdots N}) \propto f\left(\sum_{i=1}^{N} \zeta_i \delta_{\tilde{y}_i}\right) \prod_{i=1}^{N} \left((\lambda_i)^{\zeta_i} (1-\lambda_i)^{1-\zeta_i}\right) g_i(\tilde{y}_i) \lambda^N (d\tilde{y}_{1\cdots N})$$

This density may appear daunting, but it is simply the distribution of the Poisson localization reweighted by f. The fact that this density is explicitly available (up to a normalizing constant) makes it very straightforward to implement the usual techniques for Bayesian inference.

By making good choices for the densities $M = g_1 \cdots g_N$, we can produce localizations which make computation easy. In general, the $\left\{\left(\zeta_i, \tilde{Y}_i\right)\right\}$ variables will be neither identical nor independent. However, we may be able to choose M so that the dependencies are computationally manageable. For example, we can choose M so that for each $i \in 1 \cdots N$, g_i is only supported on a small subset of $\mathcal{P}$. Perhaps the support of the M distributions could tile together in an overlapping fashion, like the shingles of a roof. This allows our distribution to break into coherent local pieces, permitting both the "local moves" of the kind advocated by [28] and parallel computational strategies similar to those developed by [1]. These kinds of locality also make it reasonable to adopt mean field approximations for $\left\{\left(\zeta_i, \tilde{Y}_i\right)\right\}$. Since a density for $\left\{\left(\zeta_i, \tilde{Y}_i\right)\right\}$ is explicitly available (up to a normalizing constant), all of these ideas are straightforward to implement.

However, even though $\tilde{Y} \sim \mathrm{P}ACLocalization(\mathcal{P}, \mathcal{B}, \lambda, f, M)$ is easy to work with, it wouldn't be very useful if it is a poor approximation for $Y \sim \mathrm{P}ACProcess(\mathcal{P}, \mathcal{B}, \lambda, f)$. Under what conditions is the localization a good approximation? It is to this question we now turn our attention.

Theorem. *Let $(\mathcal{P}, \mathcal{B})$ some measurable space and λ a finite Poisson intensity measure on $(\mathcal{P}, \mathcal{B})$. Let $M^{(1)}, M^{(2)}, \cdots = \left(g_1^{(1)} \cdots g_{N^{(1)}}^{(1)}\right), \left(g_1^{(2)} \cdots g_{N^{(2)}}^{(2)}\right), \cdots$ denote a sequence of densities on $(\mathcal{P}, \mathcal{B}, \lambda)$ such that $\sum_{i=1}^{N^{(r)}} g_i^{(r)}(\alpha) = 1$ for every $\alpha \in \mathcal{P}$. Let $\lambda_i^{(r)} \triangleq \int g_i^{(r)}(\alpha)\lambda(d\alpha)$. Define*

- $\tilde{X}^{(r)} \sim \mathrm{P}oissonLocalization(\mathcal{P}, \mathcal{B}, \lambda, M^{(r)})$ and $X \sim \mathrm{P}oissonProcess(\mathcal{P}, \mathcal{B}, \lambda)$
  - $\tilde{Y}^{(r)} \sim \mathrm{P}ACLocalization(\mathcal{P}, \mathcal{B}, \lambda, f, M^{(r)})$ and $Y \sim \mathrm{P}ACProcess(\mathcal{P}, \mathcal{B}, \lambda, f)$ for some density f

  If

$$\lim_{r \to \infty} \sup_{i \in 1 \cdots N^{(r)}} \lambda_i^{(r)} = 0$$

  then the distributions of $\tilde{X}^{(r)}$ converge to that of X in total variation, and likewise the distributions of $\tilde{Y}^{(r)}$ converge to that of Y in total variation.

Proof. We will start by looking at $\left\{\tilde{X}^{(r)}\right\}_r$ and X. We can then use our analysis of these random objects to understand $\left\{\tilde{Y}^{(r)}\right\}_r$ and Y.

The distribution of the Poisson localizations $\left\{\tilde{X}^{(r)}\right\}_r$ are themselves absolutely continuous with respect to the distribution of X. Indeed, let

$$h^{(r)}\left(\sum_{i=1}^{n} \delta_{\alpha_i}\right) = \begin{cases} e^{\lambda(\mathcal{P})} \left(\prod_{i=1}^{N^{(r)}} \left(1 - \lambda_i^{(r)}\right)\right) \sum_{\sigma \in \mathrm{Sym}(n, N^{(r)})} \prod_{i=1}^{n} \frac{g_{\sigma(i)}^{(r)}(\alpha_i)}{1 - \lambda_{\sigma(i)}^{(r)}} & \text{if } n \leq N \\ 0 & \text{otherwise} \end{cases}$$

where $\mathrm{Sym}(n, N)$ denotes the set of injections $\sigma : \{1 \cdots n\} \to \{1 \cdots N\}$. Then we may immediately calculate that for any measurable bounded $\phi : \mathfrak{N} \to \mathbb{R}^+$, we have that

$\mathbb{E}\left[\phi\left(\tilde{X}^{(r)}\right)\right] = \mathbb{E}\left[h\left(X\right)\phi(X)\right]$. We can use this h to show that $\tilde{X}^{(r)}$ and X are similar in total variation; this total variation distance can be explicitly written as

$$\delta_{TV}\left(\mathrm{P}oisson Localization(\mathcal{P},\mathcal{B},\lambda,M^{(r)}),\mathrm{P}oisson Process(\mathcal{P},\mathcal{B},\lambda)\right) = \frac{1}{2}\mathbb{E}\left[\left|h^{(r)}(X)-1\right|\right]$$

Let's look at some bounds for h: □

- $h^{(r)}(x)$ is bounded from above by a constant, uniform in r and x. In showing this, there are two cases to consider:
  - If $x = \sum_{i=1}^{n}\delta_{\alpha_i}$ with $n > N^{(r)}$, $h^{(r)}(x) = 0$.
    - If $x = \sum_{i=1}^{n}\delta_{\alpha_i}$ with $n \leq N^{(r)}$, then

$$\begin{aligned} h^{(r)}\left(\sum_{i=1}^{n}\delta_{\alpha_i}\right) &= e^{\lambda(\mathcal{P})}\left(\prod_{i=1}^{N^{(r)}}\left(1-\lambda_i^{(r)}\right)\right)\sum_{\sigma\in\mathrm{Sym}(n,N^{(r)})}\prod_{i=1}^{n}\frac{g_{\sigma(i)}^{(r)}(\alpha_i)}{1-\lambda_{\sigma(i)}^{(r)}} \\ &\leq e^{\lambda(\mathcal{P})}\left(\prod_{i=1}^{N^{(r)}}\left(1-\lambda_i^{(r)}\right)\right)\sum_{k_1=1}^{N^{(r)}}\cdots\sum_{k_n=1}^{N^{(r)}}\prod_{i=1}^{n}\frac{g_{\sigma(i)}^{(r)}(\alpha_i)}{1-\lambda_{\sigma(i)}^{(r)}} \\ &\leq e^{\lambda(\mathcal{P})}\sum_{k_1=1}^{N^{(r)}}\cdots\sum_{k_n=1}^{N^{(r)}}\prod_{i=1}^{n}g_{\sigma(i)}^{(r)}(\alpha_i) \\ &= e^{\lambda(\mathcal{P})} \end{aligned}$$

      Here we have used the facts that $\lambda_i^{(r)} \in [0,1)$ and $\sum_i^{N^{(r)}} g_i^{(r)}(\alpha) = 1$ for every α.

    Putting those two cases together, we see that for every x and every r, we have that

$$h^{(r)}(x) \leq e^{\lambda(\mathcal{P})} \qquad (\dagger)$$

  - For any fixed value of x, we can get an even better upper bound on $h^{(r)}$ in the limit of large r. Since $\log(1-k)/k < -1$ and $\sum_{i=1}^{N}\lambda_i = \lambda(\mathcal{P})$, we obtain

that

$$\begin{aligned}\log\left(e^{\lambda(\mathcal{P})}\prod_{i=1}^{N}\left(1-\lambda_i^{(r)}\right)\right) &= \lambda(\mathcal{P})\left(1+\sum_{i=1}^{N}\frac{\lambda_i^{(r)}}{\lambda(\mathcal{P})}\left(\frac{\log(1-\lambda_i^{(r)})}{\lambda_i^{(r)}}\right)\right)\\ &\leq \lambda(\mathcal{P})\left(1+\sum_{i=1}^{N}\frac{\lambda_i^{(r)}}{\lambda(\mathcal{P})}(-1)\right)\\ &= 0\end{aligned}$$

And so, applying the fact that $\sum_{i=1}^{N} g_i^{(r)}(\alpha) = 1$, we get that

$$h^{(r)}\left(\sum_{i=1}^{n}\delta_{\alpha_i}\right) \leq \sum_{\sigma\in\mathrm{Sym}(n,N^{(r)})}\prod_{i=1}^{n}\frac{g_{\sigma(i)}^{(r)}(\alpha_i)}{1-\lambda_{\sigma(i)}^{(r)}} \leq \left(\frac{1}{1-\sup_i\lambda_i^{(r)}}\right)^n \qquad (*)$$

Obviously when n is large enough, this bound becomes pretty useless. However, for any fixed x we can use this bound to force $h(x) \leq 1+\epsilon$ for any arbitrary ϵ by taking $\sup_i \lambda_i$ sufficiently small.

Proof. It should now begin to be clear why the distributions of $\tilde{X}^{(r)}$ must converge to that of X in total variation. We have that □

- $h^{(r)}(x) \geq 0$ for every x and every r
  - $\mathbb{E}\left[h^{(r)}(X)\right] = 1$ for every r
  - Equation $(*)$ tells us that $\limsup_{r\to\infty} h^{(r)}(x) \leq 1$ for every $x \in \mathfrak{N}$
  - Equation (†) tells us that $h^{(r)}$ is bounded from above, uniformly in r

  As Lemma 13 from Appendix A shows, this is sufficient to prove that

$$\lim_{r\to\infty}\frac{1}{2}\mathbb{E}\left[\left|h^{(r)}(X)-1\right|\right] = 0$$

  and so the distributions of $\tilde{X}^{(r)}$ converge to that of X in total variation.

Proof. We now turn our attention to $\tilde{Y}^{(r)}$ and Y. Note that the distribution of $\tilde{Y}^{(r)}$

is absolutely continuous with respect to the distribution of Y. Indeed, let $\mathcal{Z}^{(r)} = \mathbb{E}\left[h^{(r)}(X)\, f(X)\right]$, then we have that

$$\mathbb{E}\left[\phi\left(\tilde{Y}^{(r)}\right)\right] = \mathbb{E}\left[\frac{h^{(r)}(Y)}{\mathcal{Z}^{(r)}}\phi(Y)\right] = \mathbb{E}\left[\frac{h^{(r)}(X)}{\mathcal{Z}^{(r)}}\phi(X)f(X)\right]$$

for any measurable bounded $\phi : \mathfrak{N} \to \mathbb{R}^+$. This makes it pretty easy to analyze the convergence of $\tilde{Y}^{(r)}$. Just as with $\tilde{X}^{(r)}$, we can explicitly calculate the total variation distance:

$$\delta_{TV}\left(\mathrm{P}ACLocalization(\mathcal{P},\mathcal{B},\lambda,f,M^{(r)}), \mathrm{P}ACProcess(\mathcal{P},\mathcal{B},\lambda,f)\right) = \frac{1}{2}\mathbb{E}\left[\left|\frac{h^{(r)}(Y)}{\mathcal{Z}^{(r)}} - 1\right|\right]$$

Now we know that $\mathbb{E}\left[\left|h^{(r)}(X) - 1\right|\right] \to 0$, and the distribution of Y is absolutely continuous with respect to the distribution of X, so it should be no surprise that this total variation also converges to zero in the limit of large r. The only minor hiccup is that we have to be a little bit careful to make sure nothing goes wrong with the normalizing constants, $\mathcal{Z}^{(r)}$. We will split our analysis into three steps: □

- First we will show that $\mathcal{Z}^{(r)} \triangleq \mathbb{E}\left[f(X)h^{(r)}(X)\right]$ behaves as we would hope - that is, it converges towards $\mathbb{E}\left[f(X)\right] = 1$. This is straightforward to show. For any ϵ we have that

$$\begin{aligned}\left|\mathcal{Z}^{(r)} - 1\right| &\leq \mathbb{E}\left[\left|h^{(r)}(Y) - 1\right|\right] \\ &= \mathbb{E}\left[\left|h^{(r)}(Y) - 1\right| \mathbb{I}_{\left|h^{(r)}(Y)-1\right| \leq \epsilon}\right] + \mathbb{E}\left[\left|h^{(r)}(Y) - 1\right| \mathbb{I}_{\left|h^{(r)}(Y)-1\right| > \epsilon}\right] \\ &\leq \epsilon + \left(\sup_y \left|h^{(r)}(y) - 1\right|\right) \mathbb{P}\left(\left|h^{(r)}(Y) - 1\right| > \epsilon\right)\end{aligned}$$

  Remember from equation (†) that $\sup h^{(r)}(y)$ is bounded uniformly in r. Therefore, to show $\lim Z^{(r)} = 1$ it suffices it suffices to show that $\lim_{r\to\infty} \mathbb{P}\left(\left|h^{(r)}(Y) - 1\right| > \epsilon\right) \to 0$ for every ϵ. This is immediate: we already showed that $\lim_{r\to\infty} \mathbb{P}\left(\left|h^{(r)}(X) - 1\right| > \epsilon\right) \to$

0, and the distribution of Y is absolutely continuous with respect to the distribution of X. Thus $\lim \mathcal{Z}^{(r)} = 1$.

– Since we earlier showed that $\lim_r \mathbb{E}\left[\left|h^{(r)}(X) - 1\right|\right] = 0$, it immediately follows that

$$\lim_{r\to\infty} \mathbb{E}\left[\left|\frac{h^{(r)}(X)}{\mathcal{Z}^{(r)}} - 1\right|\right] = 0$$

Since the distribution of Y is absolutely continuous with respect to the distribution of X, this nearly completes the proof. We just have one final step to take care of some details:

– Finally, we can show that

$$\lim_{r\to\infty} \mathbb{E}\left[\left|\frac{h^{(r)}(Y)}{\mathcal{Z}^{(r)}} - 1\right|\right] = 0$$

that is, the distributions of $\tilde{Y}^{(r)}$ converge to that of Y in total variation. To show this, note that for any fixed ϵ we may compute

$$\mathbb{E}\left[\left|\frac{h^{(r)}(Y)}{\mathcal{Z}^{(r)}} - 1\right|\right] \leq \epsilon + \left(\sup_y \left|\frac{h^{(r)}(y)}{\mathcal{Z}^{(r)}} - 1\right|\right) \mathbb{P}\left(\left|\frac{h^{(r)}(Y)}{\mathcal{Z}^{(r)}} - 1\right| > \epsilon\right)$$

Using $\lim_r \mathcal{Z}^{(r)} = 1$ and equation (†), we can always pick r^* large enough so that $\sup_y h^{(r)}(y)/\mathcal{Z}^{(r)}$ is uniformly bounded among all $r > r^*$. Thus to show $\lim_{r\to\infty} \mathbb{E}\left[\left|h^{(r)}(Y)/\mathcal{Z}^{(r)} - 1\right|\right] = 0$ it suffices to show that $\lim_{r\to\infty} \mathbb{P}\left(\left|h^{(r)}(Y)/\mathcal{Z}^{(r)} - 1\right| > \epsilon\right) \to 0$ for every ϵ. As in the first step, this follows immediately from $\lim_{r\to\infty} \mathbb{P}\left(\left|h^{(r)}(X)/\mathcal{Z}^{(r)} - 1\right| > \epsilon\right) 0$ and the absolute continuity of the distribution of Y with respect to the distribution of X. Thus

$$\lim_{r\to\infty} \mathbb{E}\left[\left|\frac{h^{(r)}(Y)}{\mathcal{Z}^{(r)}} - 1\right|\right] = 0$$

as desired.

We will now look at several examples of PAC process distributions, giving explicit formulae for their densities with respect to the distribution of a Poisson process.

Example. *Area-interaction Process.* Let $\mathcal{P} \subset \mathbb{R}^d$, and let μ denote the usual Lebesgue measure. Define

$$U_\epsilon\left(\sum_{i=1}^n \delta_{\alpha_i}\right) = \bigcup_{i=1}^n \{x : \ |x - \alpha_i| < \epsilon\}$$
$$f\left(\sum_{i=1}^n \delta_{\alpha_i}\right) = \beta^n \gamma^{-\mu\left(U_\epsilon\left(\sum_{i=1}^n \delta_{\alpha_i}\right)\right)}$$

Then we say that $\mathrm{A}reaInteractionProcess\left(\mathcal{P}, \beta, r\right) \triangleq \mathrm{P}ACProcess\left(\mathcal{P}, \mathcal{B}, \mu, f\right)$. This process has a certain kind of local quality (we will make this more precise in a moment). Consequently, a localization made up of densities with small support can facilitate efficient parallel computation. Just to look at a toy example, let us consider the case that $\mathcal{P}$ is a circle with circumference 1 - i.e. for every $a \in \mathcal{P}$, we have $a \equiv a + 1$. Let $\epsilon = 1/N$ for some integer N, and consider the localization given by taking

$$g_i(x) = \frac{1}{2}\mathbb{I}_{x \in \left[\frac{i}{N}, \frac{i+2}{N}\right]} \qquad i \in \{0 \cdots N-1\}$$

Taking the distribution of $\left\{\zeta_i, \tilde{Y}_i\right\}_i$ to be given by

$$p(\zeta_{1\cdots N}, d\tilde{y}_{1\cdots N}) \propto f\left(\sum_{i=0}^{N-1} \zeta_i \delta_{\tilde{y}_i}\right) \prod_{i=0}^{N-1} \left(\left(\lambda_i\right)^{\zeta_i}\left(1-\lambda_i\right)^{1-\zeta_i}\right) g_i(\tilde{y}_i) \lambda^N\left(d\tilde{y}_{1\cdots N}\right)$$

and taking $\tilde{Y} = \sum_{i=0}^{N-1} \zeta_i \tilde{Y}_i$. That is, $\tilde{Y} \sim \mathrm{P}ACLocalization(\mathcal{P}, \mathcal{B}, \mu, f, (g_0 \cdots g_{N-1}))$. The distribution on $\zeta_i, \tilde{Y}_i$ has some key conditional independence properties. For example:

$$p\left(\left(\zeta,\tilde{Y}\right)_0, \left(\zeta,\tilde{Y}\right)_3, \left(\zeta,\tilde{Y}\right)_6 \cdots \left(\zeta,\tilde{Y}\right)_{N-2} | \left(\zeta,\tilde{Y}\right)_1, \left(\zeta,\tilde{Y}\right)_2, \left(\zeta,\tilde{Y}\right)_4, \left(\zeta,\tilde{Y}\right)_5 \cdots \left(\zeta,\tilde{Y}\right)_{N-1}\right)$$
$$= \prod_{i=1}^{N/3} p\left(\left(\zeta,\tilde{Y}\right)_{3i} | \left(\zeta,\tilde{Y}\right)_1, \left(\zeta,\tilde{Y}\right)_2, \left(\zeta,\tilde{Y}\right)_4, \left(\zeta,\tilde{Y}\right)_5 \cdots \left(\zeta,\tilde{Y}\right)_{N-1}\right)$$

This is what we mean by saying that the distribution has a "local quality." This independence would allow us perform Gibbs sampling efficiently and in parallel; conditioned on certain variables, we can easily resample large blocks of variables which are all conditionally independent. This is sometimes known as "chromatic" sampling. We could also use this localization for a variational Bayesian approach. Although the variables are not independent, their dependencies are limited: this provides a rich ground for exploring possible variational Bayesian approximations, such as mean field and loopy belief propagation.

Example. *Determinantal Point Processes.* Let $\mathcal{P} \subset \mathbb{R}^d$ and let μ denote the usual Lebesgue measure. Let $C : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ admit the spectral representation

$$C(x, y) = \sum_{k=1}^{\infty} \lambda_k c_k(x) c_k(y)$$

with $\lambda_k < 1$ for all k. Then we say that $Y \sim \mathrm{D}eterminantalPointProcess\,(C)$ if its joint intensities are given by

$$\rho_n\left(x_1, x_2 \cdots x_n\right) = \det \begin{pmatrix} C(\alpha_1, \alpha_1) & \cdots & C(\alpha_1, \alpha_n) \\ \vdots & & \vdots \\ C(\alpha_n, \alpha_1) & \cdots & C(\alpha_n, \alpha_n) \end{pmatrix}$$

This distribution is absolutely continuous with respect to the distribution of a Poisson process with unit intensity. The density is given by the formulae

$$\tilde{C}(x, y) = \sum_{k=1}^{\infty} \frac{\lambda_k}{1 - \lambda_k} c_k(x) c_k(y)$$

$$f\left(\sum_{i=1}^{n} \delta_{\alpha_i}\right) \propto \det \begin{pmatrix} \tilde{C}(\alpha_1, \alpha_1) & \cdots & \tilde{C}(\alpha_1, \alpha_n) \\ \vdots & & \vdots \\ \tilde{C}(\alpha_n, \alpha_1) & \cdots & \tilde{C}(\alpha_n, \alpha_n) \end{pmatrix}$$

according to which $\mathrm{D}eterminantalPointProcess\,(C) = \mathrm{P}ACProcess\,(\mathcal{P}, \mathcal{B}, \mu, f)$ (cf. [27], equation 4.41). If $\tilde{C}$ has compact support, this distribution shares the same local quality as the previous example, making it amenable to the same kinds of localization-based approximate inference techniques mentioned in the previous example.

Example. *Cox Point Process.* Let Z denote some random positive function on $\mathcal{P}$ such that $\mathbb{P}\left(\int Z(\alpha)\lambda(d\alpha) < \infty\right) = 1$. Define the measure λ_Z as

$$\lambda_Z(A) = \int_A Z(\alpha)\lambda(d\alpha)$$

and let $(Y|Z = z) \sim \mathrm{P}oissonProcess(\mathcal{P}, \mathcal{B}, \lambda_Z)$. We then say that $Y \sim \mathrm{C}oxProcess\,(\mathcal{P}, \mathcal{B}, \lambda, Z)$. This distribution is absolutely continuous with respect to the distribution of a Poisson process. Indeed, let

$$f\left(\sum_{i=1}^{n} \delta_{\alpha_i}\right) = \mathbb{E}\left[e^{\lambda(\mathcal{P}) - \int Z(\alpha)\lambda(d\alpha)} \left(\prod_{i}^{n} Z(\alpha_i)\right)\right]$$

Then $\mathrm{C}oxProcess\,(\mathcal{P}, \mathcal{B}, \lambda, Z) = \mathrm{P}ACProcess\,(\mathcal{P}, \mathcal{B}, \lambda, f)$.

One important special case of the Cox point process is the "Cox cluster process." In such processes, the function Z may itself driven by a point process. For example, we may have $W \sim \mathrm{P}oissonProcess(\mathcal{P}, \mathcal{B}, \lambda)$, and

$$Z(\alpha) = \int k(\alpha - w)W(dw)$$

for some positive convolution kernel κ and $Y \sim \mathrm{C}oxProcess(\mathcal{P}, \mathcal{B}, \lambda, Z)$. That is, each point in W increases the intensity for Y in some neighborhood of that point. In this case, the posterior distribution of $W|Y = y$ is itself a PAC process. If κ has compact support, this posterior once again shares the same kind of locality we saw in the previous two examples.

We conclude by considering what happens if we perform the natural generalization of

the Cox cluster process to many levels of hierarchy. That is, consider a process of the form

$$\begin{aligned}
X_1 &\sim \mathrm{P}oissonProcess\left(\mathcal{P}, \mathcal{B}, \lambda\right) \\
X_2|X_1 = x_1 &\sim \mathrm{P}ACProcess\left(\mathcal{P}, \mathcal{B}, \lambda, \pi_1(\cdot; x_1)\right) \\
X_3|X_1 = x_1, X_2 = x_2 &\sim \mathrm{P}ACProcess\left(\mathcal{P}, \mathcal{B}, \lambda, \pi_2(\cdot; x_1, x_2)\right) \\
&\vdots \\
X_\ell|X_1 = x_1 \cdots X_{\ell-1} = x_{\ell-1} &\sim \mathrm{P}ACProcess\left(\mathcal{P}, \mathcal{B}, \lambda, \pi_{\ell-1}(\cdot; x_1 \cdots x_{\ell-1})\right)
\end{aligned}$$

Here π_i indicate conditional densities which are determined by the values of the previous point processes. In the context of computer vision, we take special interest in such distributions as a possible general framework for hierarchical object detection models. We may be therefore interested in some form of posterior inference, such as $X_1|X_\ell = x_\ell$. As with the Cox cluster process, this posterior is a PAC process and can be approximated with a PAC localization. However, unless the forms of $\{\pi_i\}$ are very simple indeed, it will not in general be practical to perform all of the relevant integrals. Instead, we may need approximate Bayesian techniques. For this purpose, one option would be to replace the entire process with a series of localizations:

$$\begin{aligned}
\tilde{X}_1 &\sim \mathrm{P}oissonLocalization\left(\mathcal{P}, \mathcal{B}, \lambda\right) \\
\tilde{X}_2|\tilde{X}_1 = x_1 &\sim \mathrm{P}ACLocalization\left(\mathcal{P}, \mathcal{B}, \lambda, \pi_1(\cdot; x_1)\right) \\
\tilde{X}_3|\tilde{X}_1 = x_1, \tilde{X}_2 = x_2 &\sim \mathrm{P}ACLocalization\left(\mathcal{P}, \mathcal{B}, \lambda, \pi_2(\cdot; x_1, x_2)\right) \\
&\vdots \\
\tilde{X}_\ell|\tilde{X}_1 = x_1 \cdots \tilde{X}_{\ell-1} = x_{\ell-1} &\sim \mathrm{P}ACLocalization\left(\mathcal{P}, \mathcal{B}, \lambda, \pi_{\ell-1}(\cdot; x_1 \cdots x_{\ell-1})\right)
\end{aligned}$$

This naturally raises an important question for future research: how bad is the localization approximation? In particular, how does the approximation error accumulate as we

add more levels of hierarchy? If each localization always has a total variation error of less than ϵ, conditioned on its parameters, what can be said about total variation error of the overall distribution? This problem may be difficult to answer in the completely general case. However, bounds for even very simple examples may lend important insight into whether these kinds of localizations might be safely used as building blocks in a deep learning system.

2. A Particle Method for Mean Field Variational Bayes

Drawing exact samples from complex high-dimensional distributions is often not possible in practice. However, there are a variety of methods available to draw approximate samples. Two popular approximate methods are "Gibbs sampling" and "mean field." In this section, we will show that these algorithms are in fact two extremes on a continuum.

For both Gibbs sampling and mean field, we begin with a distribution p on a product space $\Omega = \prod_i^d \Omega_i$. For simplicity, we will require that $|\Omega_i|$ is finite. We assume that $|\Omega|$ is very large and there is no obvious way to draw exact samples from p without inordinate computation. However, we will assume that it is possible to sample from $p(x_i|x_{\setminus i})$, where $x_{\setminus i}$ denotes $\{x_j\}_{j \neq i}$. Let us see how Gibbs sampling and the mean field approximation approach this situation.

- **Gibbs sampling**. We begin with some $x = (x_1, x_2 \cdots x_d) \in \Omega$, and iteratively update each variable of x according to

$$x_i \leftarrow \text{draw sample from } p(x_i|x_{\setminus i})$$

  This random process is a Markov chain. Subject to mild requirements (e.g. $p(x) > 0$ for all x is more than enough, cf. [14]), it has a unique stationary distribution, namely p. Thus by taking enough iterations, we are guaranteed that the distribution of x will converge to p.

- **Mean field approximation.** Here we try to find the information projection of p to the space of product measure on Ω. That is, we attempt to find

$$q^* = \arg\min_{q \in \mathcal{P}_\Omega} D(q||p)$$

where $D(q||p)$ denotes the KL-divergence, $\sum_{x\in\Omega} q(x) \log \frac{q(x)}{p(x)}$ and $\mathcal{P}_\Omega$ denotes the set of product measure on Ω (i.e. q can be expressed as $q(x_1 \cdots x_d) = \prod_{i=1}^d q_i(x_i)$). In practice, exact computation of q^* is not possible. Instead, it is usually approached via an iterative scheme. Perhaps the simplest such scheme is given by looking at each variable in turn. We initialize $q = \prod_i q_i$ where each q_i is some arbitrary initial distribution on Ω_i. The distribution q is then iteratively updated according to

$$q_i(x_i) \leftarrow \frac{\exp\left(\sum_{x_{\setminus i} \in \prod_{j\neq i}\Omega_i} \prod_{j\neq i} q(x_j) \log p(x_i|x_{\setminus i})\right)}{\sum_{x_i^* \in \Omega_i} \exp\left(\sum_{x_{\setminus i} \in \prod_{j\neq i}\Omega_i} \prod_{j\neq i} q(x_j) \log p(x_i^*|x_{\setminus i})\right)}$$

After sufficient iterations, an approximate sample from p can be drawn by taking $x_i \sim q_i$.

Each method is useful in different contexts. In some cases mean field methods converge much faster than Gibbs sampling methods. On the other hand, mean field approximations have some special challenges:

- The summation over all $x_{\setminus i} \in \prod_{j\neq i} \Omega_i$ may be impossible in practice
- Storing q_i may be difficult. In general, it would require $|\Omega_i| - 1$ values to encode a distribution on Ω_i. When $|\Omega_i|$ is very large, this may be impossible (even though sampling $x_i|x_{\setminus i}$ may be possible, albeit using approximate methods).
- Mean field samples are not asymptotically exact. That is - unless $p \in \mathcal{P}_\Omega$ - there is a lower bound on the difference between q and p, no matter how many iterations

you take.

- Mean field is susceptible to local minima. Although the iterative algorithm generally converges, it often fails to converge to q^*. Thus, the resulting samples are not from p, and, worse, they are not even from q^*. They are merely from a local minima of the KL-divergence.

We here present a new particle approach to these difficulties. Particle methods are not unknown to the mean field community, but this method has novel interest due to its simple form and asymptotic properties. Intuitively speaking, this method consists in replacing each q_i with a collection of samples, $(x_i(1) \cdots x_i(n))$. These samples will be updated randomly in a Markov chain monte-carlo algorithm. By making the process random instead of deterministic, we both ease computational burdens and get stronger guarantees of asymptotic accuracy.

Definition 1. We will now introduce what we call an **n-fold blur chain for p.** Consider a homogeneous Markov chain on $\left(\prod_{i=1}^{d} \Omega_i\right)^n$. Let us denote the transition kernel by

$$M(x, y), \qquad \text{with}$$
$$x = x_1(1) \cdots x_d(1), x_1(2) \cdots x_d(n-1), x_1(n) \cdots x_d(n)$$
$$y = y_1(1) \cdots y_d(1), y_1(2) \cdots y_d(n-1), y_1(n) \cdots y_d(n)$$

The chain is said to be an n-fold blur chain for p if the transition kernel is given by the following stochastic process:

1. Pick $i \sim \text{Uniform}\{1, 2, \cdots d\}$

   a) Sample $y_i(1) \cdots y_i(n)$ independently and identically from the distribution

$$\tilde{p}_i(x_i; x_{\setminus i}(1) \cdots x_{\setminus i}(n)) \triangleq \frac{\exp\left(\frac{1}{n^{d-1}} \sum_{k \in \{1 \cdots n\}^{d-1}} \log p(x_i | X_{\setminus i} = x_{\setminus i}(k))\right)}{\sum_{x_i^* \in \Omega_i} \exp\left(\frac{1}{n^{d-1}} \sum_{k \in \{1 \cdots n\}^{d-1}} \log p(x_i^* | X_{\setminus i} = x_{\setminus i}(k))\right)}$$

b) Let $y_j(1) \cdots y_j(n) = x_j(1) \cdots x_j(n)$ for all $j \neq i$.

Definition 2. That is,

$$M(x,y) = \frac{1}{d} \sum_{i=1}^{d} \left(\prod_{j \neq i} \prod_{k=1}^{n} \mathbb{I}_{x_j(k)=y_j(k)} \right) \prod_{k=1}^{n} \tilde{p}_i(y_{\setminus i}(k); x_{\setminus i}(1) \cdots x_{\setminus i}(n))$$

As long as $p(x) > 0$ for every $x \in \Omega$, n-fold blur chains are guaranteed to have some stationary distribution. The most interesting aspect of this method is how this stationary distribution varies with n. In particular,

- If $n = 1$, it reduces to ordinary Gibbs sampling, so the stationary distribution is exactly p. However, the Markov chain may take quite a long time to converge to that stationary distribution.
- As $n \to \infty$, the stationary distribution approaches the *global* minima of the KL-divergence, q^*, the true mean field approximation. In particular, the marginal distribution of $x_i(k)$ converges to that of q_i^* for each i, k.

The first point is self-evident, but the second point requires a bit of analysis to prove:

Theorem. *If $p(x) > 0$ for every $x \in \Omega$, then the n-fold blur chain for p has a unique stationary distribution, which we will denote β_n. For each n, let*

$$X^n = (X_1^n(1) \cdots X_d^n(1), X_1^n(2) \cdots X_d^n(n-1), X_1^n(n) \cdots X_d^n(n))$$

be distributed according to this stationary distribution. We observe that $X_i^n(1), X_i^n(2) \cdots X_i^n(n)$ are all identically distributed for each i. If q^* is unique, then the distribution of $X_i^n(1)$ converges to q_i^* for each $i \in 1 \cdots d$ as $n \to \infty$.

Proof. First observe that the n-fold blur Markov chain is certainly aperiodic and recurrent, since $M(x,x) > 0$ for every $x \in \Omega^n$, and $\left(M^d\right)(x,y) > 0$ for every $x, y \in \Omega^n$ (cf. [14] for a more detailed proof of why this is sufficient).

Thus we know by ergodicity that the n-fold blur chain has a unique stationary distribution. By direct inspection, we see that this stationary distribution is given by

$$\beta_n(x(1)\cdots x(n)) = \frac{1}{\mathcal{Z}_n} e^{\frac{1}{n^{d-1}}\sum_{i_1=1}^n\cdots\sum_{i_d=1}^n \log p(x_1(i_1)\cdots x_d(i_d))}$$

Indeed, M satisfies detailed balance for this distribution. Pick any $x, y \in \Omega^n$ so that there is some i so that $x_j(k) = y_j(k)\ \forall k, \forall j \neq i$ (otherwise $M(x,y) = 0$, so detailed balance is automatically satisfied). Then

$$\begin{aligned}\frac{M(x,y)}{M(y,x)} &= \frac{\prod_{k=1}^n \tilde{p}_i(y_{\setminus i}(k); x_{\setminus i}(1)\cdots x_{\setminus i}(n))}{\prod_{k=1}^n \tilde{p}_i(x_{\setminus i}(k); y_{\setminus i}(1)\cdots y_{\setminus i}(n))} = \frac{\prod_{k=1}^n \tilde{p}_i(y_{\setminus i}(k); x_{\setminus i}(1)\cdots x_{\setminus i}(n))}{\prod_{k=1}^n \tilde{p}_i(x_{\setminus i}(k); x_{\setminus i}(1)\cdots x_{\setminus i}(n))} \\ &= \exp\left(\frac{1}{n^{d-1}}\sum_{\ell=1}^n \sum_{k\in\{1\cdots n\}^{d-1}} \log\frac{p(y_i(\ell), X_{\setminus i} = x_{\setminus i}(k))}{p(x_i(\ell), X_{\setminus i} = x_{\setminus i}(k))}\right) = \frac{\beta_n(y)}{\beta_n(x)}\end{aligned}$$

So M satisfies detailed balance.

It now remains to show that the distribution of $X_i^n(1)$ converges to q_i^*. The key is to look at the empirical distributions of x, instead of x itself. That is, define $\mu_i(w_i; x) \triangleq \frac{1}{n}\sum_{k=1}^n \mathbb{I}_{x_i(k)=w_i}$, and observe that we may write β_n in terms of $\mu_i(\cdot; x)$ alone. Indeed,

$$\beta_n(x) = \frac{1}{\mathcal{Z}_n} e^{n\sum_{y\in\Omega}\left(\prod_{i=1}^d \mu_i(y_i;x)\right)\log p(y_1\cdots y_d)}$$

To make best use of this representation, we will need a well-known bound for distributions on finite state-spaces. The bound describes the number of ways of achieving a given empirical distribution, ν, from a sequence of n samples. The bound is given in terms of the entropy of that distribution, $\mathcal{H}(\nu)$. For any distribution ν on Ω_i, the bound reads that

$$\frac{1}{(n+1)^{|\Omega_i|}} e^{n\mathcal{H}(\nu)} \leq \#\left\{x_i(1)\cdots x_i(n) :\ \nu = \mu(\cdot; x_i)\right\} \leq e^{n\mathcal{H}(\nu)}$$

Applying this identity, we obtain that

$$\frac{1}{\mathcal{Z}_n\,(n+1)^{|\Omega|}}e^{-nD(\prod\nu_i||p)} \leq \sum_{x:\ \mu_i(\cdot;x)=\nu_i} \beta_n(x) \leq \frac{1}{\mathcal{Z}_n}e^{-nD(\prod\nu_i||p)}$$

What does this signify, intuitively speaking? If $X^n \sim \beta_n$, then the random variables $\{\mu_i(\cdot;X^n)\}_i$ are distributed by something primarily governed by a desire to minimize $D\left(\prod\mu_i(\cdot;X^n||p\right)$. In some sense, this completes the proof. As $n\to\infty$, β_n will place ever-more concentration on empirical distributions that make $D(\prod\mu_i(\cdot;x)||p)$ as small as possible. Asymptotically, the distribution of this random empirical distribution will converge to a dirac delta on the empirical distribution with the smallest possible KL-divergence, i.e. $\lim_{n\to\infty}\prod\mu_i(w_i;X_n) = \prod q_i^*(w_i)$. Finally, note that the distribution of $X_i^n(1)|\mu_i(\cdot;X^n;p)$ is simply μ_i, so it follows that the distribution of $X_i^n(1)$ converges to q_i^*.

To be more precise: □

1. $D(\prod\mu_i(\cdot;X^n)||p)$ **converges in probability to** $D(q^*||p)$**.** To prove this, we will show that when n is large,

$$\mathbb{P}\left(D(\prod\mu_i(\cdot;X^n)||p) > D(q^*||p)+2\epsilon\right)$$

is very small, and

$$\mathbb{P}\left(D(\prod\mu_i(\cdot;X^n)||p) < D(q^*||p)+\epsilon\right)$$

is very large. This will be sufficient to prove convergence in probability, since $\mathbb{P}\left(D(\prod\mu_i(\cdot;X^n)||p) < D(q^*||p)\right) = 0$, by the definition of q^*. Define $A_{n,\epsilon}$ as the set of d-tuples of empirical distributions $\nu = (\nu_1\cdots\nu_d)$ such that $D(q^*||p) \leq D(\prod\nu_i||p) \leq D(q^*||p)+\epsilon$. That is, $A_{n,\epsilon}$ denotes sequences whose empirical distri-

butions are pretty close to q^*. Then we have that

$$\begin{aligned}\mathbb{P}(X^n \in A_{n,2\epsilon}) = \sum_{x:\ \mu(\cdot;x)\in A^c_{n,2\epsilon}} \beta_n(x) &= \sum_{\nu\in A_{n,2\epsilon}} \sum_{x:\ \mu_i(\cdot;x)=\nu_i} \beta_n(x) \\ &\leq \frac{1}{\mathcal{Z}_n} \sum_{\nu\in A^c_{n,2\epsilon}} e^{-nD(\prod \nu_i||p)} \\ &\leq \frac{1}{\mathcal{Z}_n} \left|A_{n,2\epsilon}\right| e^{-n(D(q^*||p)+2\epsilon)}\end{aligned}$$

Of course, $|A_{n,2\epsilon}| < (n+1)^{|\Omega|}$, so we obtain that in fact

$$\mathbb{P}(X^n \in A_{n,2\epsilon}) \leq \frac{(n+1)^{|\Omega|}}{\mathcal{Z}_n} e^{-n(D(q^*||p)+2\epsilon)}$$

That is, β_n places rather little mass on $A^c_{n,2\epsilon}$. On the other hand, for sufficiently large n, $A_{n,\epsilon}$ is not empty. Therefore

$$\mathbb{P}(X^n \in A_{n,2\epsilon}) \geq \sum_{x:\ \mu(\cdot;x)\in A_{n,\epsilon}} \beta_n(x) \geq \frac{1}{\mathcal{Z}_n\,(n+1)^{|\Omega|}} e^{-n(D(\prod \nu_i||p)+\epsilon)}$$

And so β_n does place rather a lot of mass on $A_{n,2\epsilon}$. We can put these two bounds together to see that

$$\frac{\mathbb{P}(X^n \in A^c_{n,2\epsilon})}{\mathbb{P}(X^n \in A_{n,2\epsilon})} \leq (n+1)^{2|\Omega|}\, e^{-n\epsilon}$$

For sufficiently large n, this vanishes. Thus, for any fixed ϵ,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|D(\prod \mu_i(\cdot;X^n)||p) - D(q^*||p)\right| > 2\epsilon\right) = \lim_{n\to\infty} \frac{\mathbb{P}\left(X^n \in A^c_{n,2\epsilon}\right)}{\mathbb{P}\left(X^n \in A_{n,2\epsilon}\right) + \mathbb{P}\left(X^n \in A_{n,2\epsilon}\right)} = 0$$

as desired.

a) **Ergo $\mu_i(\cdot;X^n)$ converges in probability to q_i^* for each i.** This is a simple consequence of the above, together with three facts:

   i. the space of probability distributions on finite sets is compact

ii. the KL divergence is continuous

iii. q^* is unique

b) **Ergo $X_i^n(k)$ converges to q_i^* for each i, k.** This follows from the fact that the distribution on $X_i^n(k)$ is sampled directly from $\mu(\cdot; X^n)$.

This theorem admits some relaxations. For example:

- If q^* comprises a finite set of minimizing distributions, then the distribution of $X_i^n(k)$ will go to some convex combination of those distributions which minimize the KL divergence.
- If $p(x) = 0$ for some x, the theorem may still hold, as long as the resulting Markov chain is still aperiodic and recurrent.

Beyond the interesting fact that Gibbs sampling and Mean field approximations lie on a common continuum, we now turn our attention to the question of whether this theorem has any direct application. There are two considerations in this regard:

1. Is it computationally feasible to sample from an n-fold blur chain? When $d = 2$ or $d = 3$, the answer is generally a straight "yes." The key difficulty is in computing $\tilde{p}_i(x_i; x_{\setminus i}(1) \cdots x_{\setminus i}(n))$, which - from a naive point of view - requires $O(|\Omega_i| \, n^{d-1})$ operations for each update. Obviously when $d \gg 1$, this becomes intractable. However, there are situations when d is large and computation is still quite possible. For example, let us assume p represents a Gibbs random field whose maximal clique has ζ nodes and whose maximal degree is ξ. That is, the distribution of p may be factored as $p(x) = \prod_{C \in \mathcal{C}} f(x_C)$ where $\mathcal{C}$ is a collection of "clique" sets, each set of size ξ or less. Then computation requires $O(\xi \, |\Omega_i| \, n^{\zeta - 1})$ operations per update. For example, the Ising model on a lattice $\mathbb{Z}^k$ is a very complex object, but its cliques contain only 2 nodes and its maximal degree is $2k$. Thus computation of the n-fold blur sample for this object requires only $O(4nk)$ operations. Thus

sampling is indeed feasible in many cases. Furthermore, this sampling process is highly amenable to parallel computational strategies (for example, it is easy to allocate a separate process for each of the n particles). Finally, we note that blur chains may be much easier to compute than the corresponding mean field solutions, since there is no necessity to compute sums over $x_{\setminus i} \in \prod_{j \neq i} \Omega_i$ or explicitly store distributions on Ω_i.

2. Does the n-fold blur chain provide any advantages over Gibbs sampling or mean field approximations? For a preliminary exploration of this question, we looked at the toy model of sampling from the 1-dimensional Ising model. This model has the advantage that exact ground truth is easy to compute. Here, quite arbitrarily, we picked the case of a 10-node Ising model with periodic boundary conditions, a temperature of 6, and an external force field given by the values

$$(0.344, 2.207, -1.716, -0.155, 0.008, 0.292, 1.162, 0.228, 1.316, 0.224)$$

This was simply the first model we tried; other examples yielded similar results. In terms of this model, we then looked at $\mathbb{E}[X_1]$, four different ways.

a) Estimated via Markov chain monte-carlo with Gibbs sampling

b) Estimated via Markov chain monte-carlo with a 2-fold blur chain

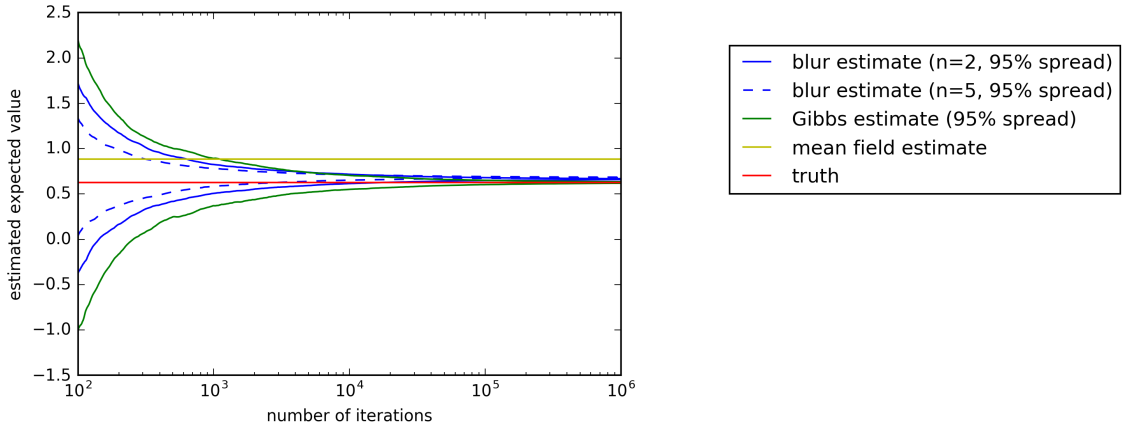
c) Estimated via Markov chain monte-carlo with a 5-fold blur chain

d) Estimated via mean field (we were only able to find one local minima of $D(q||p)$)

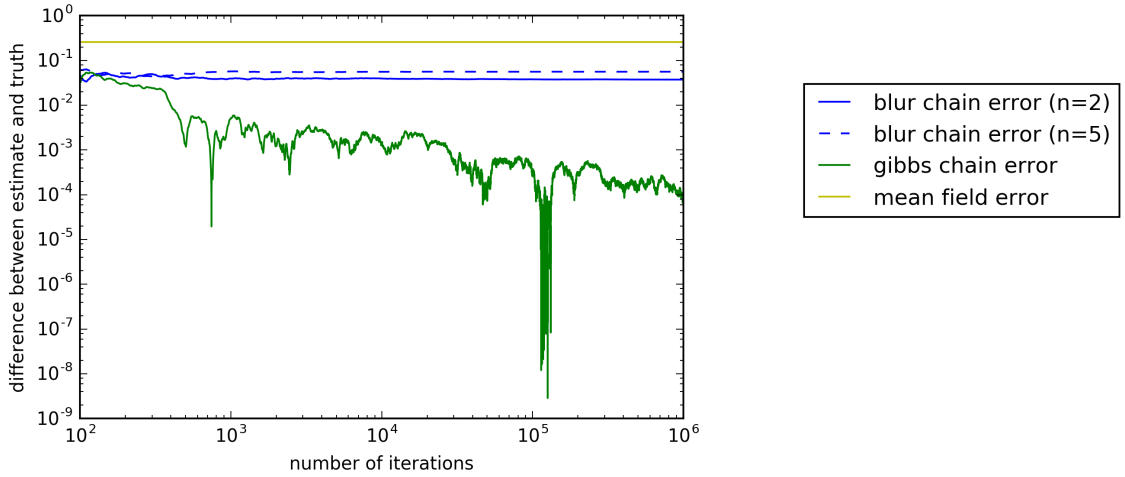
e) Calculated directly via brute force

Using random initial conditions, we ran 500 trials with 10^6 iterations of method (a), (b), and (c). For each trial at each iteration under each method, we produced an estimate for $\mathbb{E}[X_1]$ based on the empirical average of X_1 over all the iterations

up until that point for that trial, excluding the first 100 iterations (we exclude the first iterations to follow a common practice known as "burn-in"). This gave us $500 \times 10^6 \times 3$ estimates for $\mathbb{E}[X_1]$: one for each triple of (trial, iteration, method). For each (iteration, method)-pair, we then calculated the central region which contained the estimates for 95% of the trials. These 95% spreads are plotted below on a log plot, together with the mean field estimate and the true answer for comparison.



Let us also look at another plot, showing the absolute difference between the center of the 95% confidence intervals and the truth, as a function of the number of iterations. This captures the bias of each estimator. We also include the error for the mean field estimate, for comparison.



We highlight two trends in the graphs above:

- The Gibbs sampling estimate converges closer and closer to the exact answer, as we knew it must. On the other hand, the 2-fold blur chain has an asymptotic error of about .005. The 5-fold blur chain has an asymptotic error of about .008. The mean field solution differs from the truth by about .132. Thus we see that the Gibbs method is the most accurate, followed by 2-fold blur chain, followed by the 5-fold blur chain, followed by the mean field solution.
- The mean field solution can be computed almost instantly, and has no variance. The width of the 95% confidence interval of the 5-fold blur estimator tightens fairly quickly as the number of iterations increases. The width of the interval for the 2-fold blur estimator tightens a bit slower. The width of the interval for the Gibbs estimator tightens slowest of all. Thus we see that the Gibbs method converges slowest, followed by the 2-fold blur chain, followed by the 5-fold blur chain, followed by the mean field solution.

Thus, on the whole, we see that the blurred chains fulfill their promise of being an intermediary between Gibbs sampling and mean field. A higher value of n causes chains to converge faster than Gibbs, at the expense of some accuracy. In the limit as $n \to \infty$, we know that the convergence will be achieved in very few steps, but the error will be that of the mean field solution.

We conclude by noting that the blur chain is almost certainly not the only algorithm that lies on the continuum between Gibbs and mean field. It would be interesting to explore other intermediaries, in search of the most effective ones.

For example, many mean field algorithms do not use the simple per-variable iteration updates that I described above; instead, $D(q||p)$ is optimized directly using gradient methods. However, these gradient methods also involve computing expectations, and it would be interesting to try to understand how replacing those expectations with particle

approximations would effect their result. In this case, the random variables of the Markov chain would be the parameters of q, instead of particles. The particles would merely serve as the mechanism of the transition kernel of the Markov chain. As a result, the stationary distribution of such an algorithm would not be so simple as it is for the blur chain. However, further investigation might be able to at least show that the stationary distribution converges towards the mean field solution in the limit as the number of particles used goes to infinity. In general, the difficulty is that even the slightest error due to particle approximations may accumulate through the Markov chain. Does the effect of this error blow up exponentially? Or can we show that the error is controlled?

In the case of the blur chain, at least, we have herein shown that the error is tamed. It remains to be seen whether the same can be said for other particle-based methods in the continuum between Gibbs sampling and mean field.

3. Brownian Motion Hitting Probabilities for Small Sets

The computer simulation of long biological polymers continues to provide a great challenge. One of the difficulties is that such polymers spend an inordinately long time wiggling somewhat purposelessly around in space. It is only very rarely that pieces of polymers come close enough to each other that significant energetic interactions kick in. At such key moments, the behavior of the molecule may be very intricate and it may move very swiftly. At other times, molecules may be seem to be doing barely anything at all.

This makes direct simulation of the molecular dynamics quite challenging. If a system is accurate enough to handle the fast-timescale behavior, it is too slow to handle the long-timescale behavior. If it is fast enough to handle the long-timescale behavior, it may not be accurate enough to handle the short-timescale behavior. One obvious solution is to try to divide our computation into a short-timescale analysis and a long-timescale analysis. In theory, we should be able to put these two together to produce a good solution.

One way to bridge the gap between short- and long-timescales is by dividing the space of possible polymer poses into a finite number of regions and building a discrete Markov chain that moves between those regions. Here by "polymer pose" we simply mean the collective positions of all the parts of the polymers. Each state in the state-space of the Markov chain corresponds to one of the regions of pose space. Ideally, each region will be relatively small, and we can therefore use simulations to try to understand what

the transition probabilities should be (for example, cf [9]). To simulate long-timescale behavior, we can simply run the Markov chain, which is much more computationally efficient than simulating the whole dynamics exactly.

Unfortunately, it is not possible to divide the space of all possible poses into a small number of small regions. In practice, some regions will simply need to be huge. Here "huge" is defined quite practically. In particular, consider any given initial pose inside a pose region. We need to be able to simulate the polymer as it moves from that initial pose across the boundary into another pose region. If the amount of time required for this transition is longer than we can tractably simulate using modern computers, we consider the pose region to be "huge." If we have "huge" sets then we will not be able to actually compute the transition probabilities which we need.

For example, consider the set of all poses of an RNA molecule in which none of the nucleotides are bound to any other nucleotides, or even close to binding to any other nucleotides; that is, RNA molecules that aren't close to forming any so-called "secondary structure" (the secondary structure of an RNA molecule indicates the bonds between non-adjacent nucleotides in the RNA sequence). This set of poses is enormous - there are quite a lot of ways that RNA molecules can twist around without forming a secondary structure. Simulating a polymer moving through this space would be very computationally burdensome. This is especially frustrating, because in a certain sense this set is very "boring." In particular, there are no complicated energetic interactions happening as long as there is no secondary structure.

This raises a very interesting question: is there an easier way to understand the behavior of polymers inside very large, "boring" regions? We will focus on a toy example. We will consider the case that we have a large boring region, $N \subset \mathbb{R}^n$, with two very small important regions tucked inside it, A and B. Thus we have one "huge" set and two tractable "small" sets. We will assume that the poses of a molecule lie N, and that in practice these poses always lie inside N; for simplicity, we will assume that this is

because of very high energetic barriers at ∂N, effectively producing a reflecting boundary. The small interesting regions A, B may indicate places where the energy landscape becomes complex. Outside the interesting regions, we assume that the energy landscape governing the molecule is flat. Given that a molecule starts somewhere in this boring region, our task is to understand which interesting region the molecule is likely to hit first. More generally, we want to know the probability with which it hits each important region inside of it. If we could perform these kinds of calculations, we would be able to compute the transition probabilities we need to make a Markov-chain-based simulation truly practical - at least for this toy example.

It turns out that that we can understand the dynamics of such a system quite exactly in the asymptotic limit as the diameter of A and B becomes small. What's more, asymptotically speaking, the dynamics do not actually depend on where the polymer lies inside the region. Thus it may indeed be quite reasonable to consider a very large "boring" region as just a single state in a Markov chain simulation technique. If we could leverage the techniques to analyze this toy example to more sophisticated examples, we may be able to make significant progress on efficient simulation of polymers.

3.1. Setup and statement of the theorem

Let $\{X_t\}$ denote a Brownian motion contained inside a connected, open, bounded set $N \subset \mathbb{R}^D$ by a reflecting boundary at ∂N. Now fix two subsets $A, B \subset N$, and a starting point $x_0 \in N$. If $X_0 = x_0$, what is the probability that $\{X_t\}$ hits A before it hits B? That is: if $\tau \triangleq \inf\{t : \ X_t \in \partial A \cup \partial B\}$ then what is $\mathbb{P}^{x_0}(X_\tau \in \partial A)$?

The answer depends upon the shapes and sizes of A, B, along with the placements of A, B, x_0 inside the set N. Different scenarios may call for different methods of approximating $\mathbb{P}^{x_0}(X_\tau \in \partial A)$. It turns out that we can construct a pretty good approximation for scenarios in which the following parameters can be controlled:

- **Circumscribing spheres should be small.** Let x_A, x_B, r_A, r_B such that $A \subset \mathscr{B}(x_A, r_A)$ and $B \subset \mathscr{B}(x_B, r_B)$, where $\mathscr{B}(x_0, r) \triangleq \{x : |x - x_0| < r\}$. We want r_A, r_B small.

- A, B **should be far from each other and** ∂N**.** Let $d_A > r_A, d_B > r_B$ such that

  - $\mathscr{B}(x_A, d_A) \cap (\mathscr{B}(x_B, d_B) \cup \partial N) = \emptyset$
  - $\mathscr{B}(x_B, d_B) \cap (\mathscr{B}(x_A, d_A) \cup \partial N) = \emptyset$

  We want d_A, d_B large.

- **Brownian motion should be able to mix in** N**.** Let c_N denote the Poincare-Wirtinger constant for N, i.e. $\|f\|^2_{\mathscr{L}^2(N)} \leq c_N \||\nabla f|\|^2_{\mathscr{L}^2(N)}$ for all weakly differentiable f with vanishing mean. Among other things, this captures how long it takes Brownian motion on N to mix. We do not want c_N to be too large.

- **The dimension must be at least 3.**

- **The sets must be regular**. We require that A, B, N are open and bounded with C^3 boundary.

- **The reflection must be normal.** We assume $\{X_t\}$ is given by the standard Skorokhod construction of normally reflecting brownian motion inside N.

For the reader's convenience, we review some standard definitions of smooth boundaries and normally reflecting brownian motion in Appendix B.1.

When the above parameters can be properly controlled, we will see that $\mathbb{P}^{x_0}(X_\tau \in \partial A)$ does not significantly depend on the placement of A, B, x_0 inside of N. Instead, the probability is largely governed by the "harmonic capacity" of the sets A and B:

Definition. Let $C \subset \mathbb{R}^D$ any open, bounded set with C^3 boundaries, $D \geq 3$. Take $\{W_t\}$ to denote a Brownian motion on $\mathbb{R}^D$, and define $w(x) = \mathbb{P}^x(\inf\{t : W_t \in C\} < \infty)$. Then we shall define the **harmonic capacity** of C is as $cap(C) \triangleq \int_{\mathbb{R}^D \setminus C} |\nabla w(x)|^2 \, dx$.

We shall call the function w the **harmonic capacity function**. It is immediate that the capacity scales with the size of C like the diameter of the set to the power of $D-2$. That is, just as $\text{Vol}(rC)/\text{Vol}(C) = r^D$, the harmonic capacity satisfies $cap\left(rC\right)/cap\left(C\right) = r^{D-2}$. We refer the reader to [2] for a general exposition of the capacity of sets.

We will show that the likelihood ratio $\mathbb{P}^{x_0}\left(X_\tau \in \partial A\right)/\mathbb{P}^{x_0}\left(X_\tau \in \partial B\right)$ is roughly given by the ratio of the harmonic capacities of the two sets, $cap\left(A\right)/cap\left(B\right)$. In the limit as A, B grow small or $D \to \infty$, this approximation becomes exact.

Theorem. *Fix d_A, d_B, c_N. Define $h^{**} = cap\left(A\right)/\left(cap\left(A\right) + cap\left(B\right)\right)$.*

- For any $D, \gamma > 0, \epsilon > 0$ we can pick $c = c(d_A, d_B, c_N, D, \gamma, \epsilon) > 0$ so that $|\mathbb{P}^x(X_\tau \in \partial A) - h^{**}| < \epsilon$ for all x outside of $\mathscr{B}\left(x_A, \gamma\right), \mathscr{B}\left(x_B, \gamma\right)$ for all scenarios with $r_A, r_B < c$.
  - For any $r_A, r_B, \gamma > 1, \epsilon > 0$ we can pick $c = c(d_A, d_B, c_N, r_A, r_B, \gamma, \epsilon)$ so that $|\mathbb{P}^x(X_\tau \in \partial A) - h^{**}| < \epsilon$ for all x outside of $\mathscr{B}\left(x_A, \gamma r_A\right), \mathscr{B}\left(x_B, \gamma r_B\right)$ for all scenarios with $D > c$.

Proof. Deferred to Section 3.4. □

When the conditions of this theorem are met for small ϵ, we observe that the function $u(x) = \mathbb{P}^x(X_\tau \in \partial A)$ must be quite "flat." In particular, the theorem claims that $\sup_x |u(x) - h^{**}| < \epsilon$ on a region at modest distance from A, B. The main difficulty in proving the theorem will be in showing that u is really quite flat.

In Section 3.2 we will show that u is flat in the sense that $\inf_{\bar{u}} \int \left(u(x) - \bar{u}\right)^2 dx$ is small for some scalar $\bar{u}$. We will use this in Section 3.3 to show that $\sup_x |u(x) - \bar{u}|$ is also small on a region at modest distance from A, B. Having shown that u is uniformly close to *some* constant, it will be straightforward to show that this constant must be very close to h^{**}; we use this to prove the theorem in Section 3.4.

3.2. The variance of u is small

Let $u(x) = \mathbb{P}^x(X_\tau \in \partial A)$. It is well-known that u satisfies three boundary constraints (cf. Appendix B.1):

$$\begin{aligned} u(\partial A) &= 1 \\ u(\partial B) &= 0 \\ \langle \nabla u\,(\partial N)\,, \mathbf{n} \rangle &= 0 \end{aligned}$$

Here $\mathbf{n}$ indicates the surface normal of ∂N.

The function u has two important characterizations, which we will use throughout:

1. **The harmonic characterization.** u is the unique function which is harmonic on $\Omega = N \backslash A \backslash B$ and satisfies the boundary constraints.

2. **The variational characterization.** u is the unique solution to $\min \int |\nabla w|^2$, among all functions $w \in H^1(\Omega)$ satisfying the boundary constraints. Here $H^1 = W^{1,2}$ indicates the usual Sobolev space.

A more detailed account of these well-known characterizations is given in Appendix B.1. The harmonic characterization follows from the application of Corollary 18 with $g = \mathbb{I}_A$ and $\mathcal{L} = \Delta/2$; the variational characterization then follows from Theorem 21.

The variational characterization is the first hint that u might be inclined to be somewhat flat - indeed, the integral of its squared absolute gradient should be smaller than that of any other function satisfying the boundary constraints. To make use of this variational characterization, we will construct an approximation, $\hat{u}$, satisfying the boundary constraints, for which we can directly calculate $\int |\nabla \hat{u}|^2$. This will give us an upper bound on $\int |\nabla u|^2$, giving us our first numerical measure of just how flat u is.

The approximation $\hat{u}$ will be produced by breaking up Ω into three separate regions, and trying to treat each of them somewhat separately. Let $\beta = \frac{d_A}{r_A} \wedge \frac{d_B}{r_B}$ and $\tilde{\beta} = \frac{\beta+1}{2}$.

Notice that, by assumption, $\mathscr{B}(x_A, \tilde{\beta} r_A) \supset A$ is guaranteed to be far away from $B, \partial N$ and $\mathscr{B}(x_B, \tilde{\beta} r_B) \supset B$ is guaranteed to be far away from $A, \partial N$. On the other hand, if d_A, d_B are large, then the boundary of these balls is also guaranteed to far away from the sets which they enclose. With this in mind, we will observe that Ω can be partitioned into three regions: $\mathscr{B}(x_A, \tilde{\beta} r_A) \backslash A$, $\mathscr{B}(x_B, \tilde{\beta} r_B) \backslash B$, and $N \backslash \Big(\mathscr{B}(x_A, \tilde{\beta} r_A) \cup \mathscr{B}(x_B, \tilde{\beta} r_B) \Big)$. Now, for each $h \in (0,1)$, we will define our approximation $\hat{u}_h$ as the unique continuous function on $N \backslash (A \cup B)$ such that

- $\hat{u}_h$ satisfies the boundary constraints
- $\hat{u}_h$ is harmonic inside $\mathscr{B}(x_A, \tilde{\beta} r_A) \backslash A$ and $\mathscr{B}(x_B, \tilde{\beta} r_B) \backslash B$
- $\hat{u} = h$, a constant, when outside of $\mathscr{B}(x_A, \tilde{\beta} r_A)$ and $\mathscr{B}(x_B, \tilde{\beta} r_B)$

To pick h, we will solve

$$\min_{h \in (0,1)} \int_{\Omega} |\nabla \hat{u}_h|^2$$

We can get the *exact* solution to this problem in terms of the something called the "condenser capacities," $\kappa_A = cap\Big(A \subset \mathscr{B}(x_A, \tilde{\beta} r_A)\Big)$ and $\kappa_B = cap\Big(B \subset \mathscr{B}(x_B, \tilde{\beta} r_A)\Big)$.

Definition 3. Let $C \subset C' \subset \mathbb{R}^D$. Take $\{W_t\}$ to denote a Brownian motion on $C \backslash C'$, and $w(x) = \mathbb{P}^x \left(X_{\inf\{t:\ W_t \in \partial C \cup \partial C'\}} \in \partial C \right)$. Then the **condenser capacity of** $C \subset C'$ is defined as $cap\,(C \subset C') \triangleq \int_{C' \backslash C} |\nabla w|^2$. The function w is called the **condenser capacity function**. Condenser capacity functions are the unique harmonic functions subject to their boundary conditions, $w(\partial C) = 1$ and $w(\partial C') = 0$ (cf. Appendix B.1; apply Corollary 18 with $g = \mathbb{I}_{\partial C}$ and $\mathcal{L} = \Delta/2$). Note that the harmonic capacity is the limit of the condenser capacity as C' grows large. Indeed - as we we will later show in Lemma 10 - $cap\,(C \subset C') \approx cap\,(C)$ when C is very small or C' is very large.

The condenser capacity functions - suitably rescaled - are exactly what we need for the harmonic portion of our approximation, $\hat{u}_h$. Let w_A, w_B denote the condenser capacity

functions for $A \subset \mathscr{B}(x_A, \tilde{\beta} r_A)$ and $B \subset \mathscr{B}(x_B, \tilde{\beta} r_B)$, respectively, then it turns out we can write $\hat{u}_h$ explicitly as

$$\hat{u}_h(x) = \begin{cases} (1-h)\, w_A(x) + h & \text{if } x \in \mathscr{B}(x_A, \tilde{\beta} r_A) \\ h w_B(x) & \text{if } x \in \mathscr{B}(x_B, \tilde{\beta} r_B) \\ h & \text{else} \end{cases}$$

It follows directly from the definition of the condenser capacities κ_A, κ_B that $\int_\Omega |\nabla \tilde{u}_h|^2 = \kappa_A (1-h)^2 + \kappa_B h^2$. Taking derivatives and setting equal to zero, we see that the optimal solution h^* is given by the equation

$$h^* = \frac{\kappa_A}{\kappa_A + \kappa_B}$$

And so

$$\int_N |\nabla \hat{u}_{h^*}|^2 = \kappa_A (1-h^*)^2 + \kappa_B (h^*)^2 = \frac{\kappa_A \kappa_B}{\kappa_A + \kappa_B}$$

When r_A^{D-2}, r_B^{D-2} are small, κ_A, κ_B become small. This forces $\int_N |\nabla \hat{u}|^2$ to be small, which then forces $\int_N |\nabla u|^2$ to be small. This is our first hint that u should be pretty flat. Together with a Poincare inequality, this tells us that the overall variance of u must also be small. We summarize these results in the lemma below:

Lemma 4. *Extend u to all of N by taking $u(A) = 1$ and $u(B) = 0$. Let $\bar{u} = \frac{1}{|N|} \int_N u$. Then $\int |\nabla u|^2 < \kappa_A \kappa_B / (\kappa_A + \kappa_B)$, and the variance of u is bounded by*

$$\int_N (u - \bar{u})^2 \leq c_N \frac{\kappa_A \kappa_B}{\kappa_A + \kappa_B}$$

Moreover,

$$\frac{\kappa_A \kappa_B}{\kappa_A + \kappa_B} \leq \left(\frac{2\pi^{\frac{D}{2}} (D-2)}{\Gamma\left(\frac{D}{2}\right) \left(1 - \tilde{\beta}^{2-D}\right)} \right) \frac{r_A^{D-2} r_B^{D-2}}{r_A^{D-2} + r_B^{D-2}}$$

Thus the variance vanishes as $r_A^{D-2} r_B^{D-2}$ vanish.

Proof. First notice that $\int |\nabla u|^2 \leq \int_N |\nabla \hat{u}_{h^*}|^2 = \kappa_A \kappa_B / (\kappa_A + \kappa_B)$, because u is the minimizer of the integrated squared norm of the gradient among all functions satisfying the boundary conditions. The variance bound follows directly from applying the Poincare inequality, $\left\| f - \bar{f} \right\|_{\mathscr{L}^2(N)} \leq c_N \left\| |\nabla f| \right\|_{\mathscr{L}^2(N)}$.

It only remains is to put an upper bound on the $\kappa_A \kappa_B / (\kappa_A + \kappa_B)$. For this, we note that if we replaced A with $A' = \mathscr{B}(x_A, r_A)$ and B with $B' = \mathscr{B}(x_B, r_B)$, the value of $\int |\nabla u_h|^2$ would only get worse (this replacement corresponds to adding an additional constraint to $\hat{u}$). Thus we can upper bound $\int |\nabla u_{h^*}|^2$ using $cap\left(A' \subset \mathscr{B}(x_A, \tilde{\beta} r_A)\right)$ and $cap\left(B' \subset \mathscr{B}(x_B, \tilde{\beta} r_B)\right)$ in place of κ_A, κ_B. Since spheres are very easy to work with, it is straightforward to solve the partial differential equation $\Delta w = 0$ with the appropriate boundary conditions to compute the corresponding capacity function for this condenser. Integrating the squared norm of the gradient, we can then readily compute that that $cap\left(A' \subset \mathscr{B}(x_A, \tilde{\beta} r_A)\right) = r_A^{D-2} 2\pi^{\frac{D}{2}} (D-2) / \Gamma\left(\frac{D}{2}\right) \left(1 - \tilde{\beta}^{2-D}\right)$ and likewise for B; the bound follows immediately. □

3.3. We can control $|u - \bar{u}|$ uniformly on a region away from A, B

We now know that $\int (u - \bar{u})^2$ is small. We can use this to show that $|u(x) - \bar{u}|$ can actually be uniformly controlled. Our argument will follow two steps:

1. Starting from a point far away from $A \cup B$, it takes a long time to hit either A or B.

2. If it takes long enough, then the process will forget it's initial conditions. Thus the probability that it hits ∂A before ∂B will not much depend on the initial condition, x. Thus $u(x)$ will be close to some constant for all x away from $A \cup B$. This constant

will have to be close to $\bar{u}$.

3.3.1. It takes a long time to hit A or B

Regardless of whether we hit A or B first, how long will the whole process take? Getting an exact handle on this time is challenging. However, we can get some lower bounds. We'll start by pretending B doesn't exist, and just focus on how long it would take to hit A.

Let $\lambda_A = \inf\{t : X_t \in \partial A\}$ - the time to hit ∂A, regardless of whether we passed through B en route. While simpler than τ, getting the distribution of λ_A is still difficult, because $\partial A, \partial N$ could be very tricky objects indeed. So we will approximate this problem using another process. We replace both ∂A and ∂N with spheres.

Let $\tilde{X}_t$ denote a different Brownian motion, with

- **Reflecting boundary** $\partial\mathscr{B}(x_A, \beta r_A)$
- **Absorbing time** $\tilde{\lambda}_A = \inf\left\{t : \tilde{X}_t \in \partial\mathscr{B}(x_A, r_A)\right\}$

Getting the distribution of $\tilde{\lambda}_A$ is possible, because the boundaries are so well understood. It's also useful, because $\tilde{\lambda}_A$ is a lower bound on λ_A. Indeed, we can explicitly construct a coupling of $\left(\tilde{\lambda}_A, \lambda_A\right)$ so that $\tilde{\lambda}_A \leq \lambda_A$ almost surely, namely $\tilde{\lambda}_A = \int_0^{\{\inf t:\ X_t \in \partial\mathscr{B}(x_A, r_A)\}} \mathbb{I}_{X_t \in \mathscr{B}(x_A, \beta r_A)}$.

To get the distribution of $\tilde{\lambda}_A$, we will analyze the function

$$x \mapsto \mathbb{E}^x\left[e^{-\frac{1}{2}\xi^2 \tilde{\lambda}_A}\right]$$

For any fixed ξ, this function can be characterized with a partial differential equation, just like u. However, unlike u, this differential equation can be solved exactly. We can use this solution - together with a Chernoff bound - to show that $\tilde{\lambda}$ is small with small probability.

Lemma. *For any* $|\tilde{x}_0| \geq \frac{\beta+1}{2} r_A$*, we have*

$$\mathbb{P}^{\tilde{x}_0}\left(\tilde{\lambda} < (\beta r_A)^2 \frac{\beta^{\frac{D-2}{2}}}{D(D-2)}\right) \leq 6\beta^{\frac{2-D}{2}} e^{3(D-2)\beta^{-1+\frac{2-D}{4}}}$$

Proof. Straightforward but rather technical. Deferred to Appendix B.2, Lemma 23. □

The choice to bound $\mathbb{P}^{\tilde{x}_0}\left(\tilde{\lambda} < (\beta r_A)^2 \frac{\beta^{\frac{D-2}{2}}}{D(D-2)}\right)$ may seem a bit complicated and odd. However, we will be wanting to consider the case that $\beta r_A \approx d_A$ is fixed, and we are sending either $r_A \to 0$ or $D \to \infty$. In either case, $\beta^{D-2/2}$ will go to infinity and the bound will go to zero, thus showing that $\tilde{\lambda}$ is getting large.

All that now remains is to apply this lemma twice to get a bound on τ:

Corollary 5. *Assume* x_0 *is outside of* $\mathscr{B}\left(x_A, \tilde{\beta} r_A\right) \cup \mathscr{B}\left(x_B, \tilde{\beta} r_B\right)$*. Then*

$$\mathbb{P}^{x_0}\left(\tau \leq (\beta r_A \wedge \beta r_B)^2 \frac{\beta^{\frac{D-2}{2}}}{D(D-2)}\right) \leq 12\beta^{\frac{2-D}{2}} e^{3(D-2)\beta^{-1+\frac{2-D}{4}}}$$

Proof. Apply the coupling in which $\tilde{\lambda}_A \leq \lambda_A$ almost surely and likewise $\tilde{\lambda}_B \leq \lambda_B$ for the corresponding times. Then apply a union bound. □

3.3.2. If we don't hit A or B for a long time, then $u(x_0) \approx \bar{u}$

If $\{X_t\}$ has a long time to wander around before it hits A or B, the chance of hitting A before B doesn't really depend on where you start. In fact:

1. If t small enough so $\mathbb{P}^{x_0}(\tau < t) \approx 0$, then $u(x_0)$ is approximately given by $\mathbb{E}^{x_0}[u(X_t)]$
2. If t is large enough so that $\{X_t\}$ has had time to mix, then $\mathbb{E}^{x_0}[u(X_t)] \approx \bar{u}$.

Assuming we can find a t which is small enough to satisfy (1) and large enough to satisfy (2), we must have that $u(x_0) \approx \bar{u}$. This argument is made rigorous in the following lemma.

Lemma 6. *For any $t > 0$, and any x_0 outside of $\mathscr{B}(x_A, \tilde{\beta} r_A), \mathscr{B}(x_B, \tilde{\beta} r_B)$, we have*

$$|u(x_0) - \bar{u}| \leq \mathbb{P}(\tau \leq t) + \sqrt{c_N e^{-t/c_N} D \left(\frac{\beta - 1}{2\beta}\right)^{-D} (\beta r_A \wedge \beta r_B)^{-D} \left(\frac{D-2}{1-\tilde{\beta}^{2-D}}\right) \frac{r_A^{D-2} r_B^{D-2}}{r_A^{D-2} + r_B^{D-2}}}$$

Proof. Pick any $x_0 \in \partial\mathscr{B}(x_A, \tilde{\beta} r_A) \cup \partial\mathscr{B}(x_B, \tilde{\beta} r_B)$. We are interested in $u(x_0) = \mathbb{P}^{x_0}(X_\tau \in \partial A)$. We want to show that if the Markov process $\{X_t\}$ has time to mix, then we will obtain some nice properties. Unfortunately, the mixing bounds we use do not interact well with the degenerate initial condition $X_0 \sim \delta_{x_0}$. Therefore, we find it convenient to instead consider a nondegenerate initial distribution, q, given by:

$$\begin{aligned} S &\triangleq \mathscr{B}\left(x_0, \frac{\beta - 1}{2\beta}(\beta r_A \wedge \beta r_B)\right) \\ q(x) &= \text{Uniform}(x; S) \end{aligned}$$

Using the initial condition q instead of δ_{x_0} doesn't actually change anything of importance to us. This is because u is guaranteed to be harmonic on all of S; the averaging property of harmonic functions thus gives that $u(x_0) = \int q(x) u(x) dx$. Combining this with the Dynkin formula, $u(x) = \mathbb{E}^x[u(X_\tau)]$, we obtain that

$$u(x_0) = \int q(x) \mathbb{E}^x[u(X_\tau)] = \mathbb{E}^q[u(X_\tau)]$$

So instead of dealing with $u(x_0)$ directly, we can deal with $\mathbb{E}^q[u(X_\tau)]$, which is less degenerate.

Now, two observations:

□

1. $u(x_0) = \mathbb{E}^q[u(X_\tau)] \approx \mathbb{E}^q[u(X_t)]$, assuming $\mathbb{P}(\tau < t)$ is small. Indeed,

$$|\mathbb{E}^q[u(X_\tau) - u(X_t)]| = |\mathbb{E}^q[(u(X_\tau) - u(X_t))\mathbb{I}_{\tau<t}] + \mathbb{E}^q[(u(X_\tau) - u(X_t))\mathbb{I}_{\tau \geq t}]|$$

Notice that $0 \leq u \leq 1$ (for the first term) and $\mathbb{E}^q\left[(u(X_\tau) - u(X_t))\,\mathbb{I}_{\tau \geq t}\right] = 0$ (for the second term). Thus

$$\left|\mathbb{E}^q\left[u(X_\tau) - u(X_t)\right]\right| \leq \mathbb{P}(\tau < t)$$

a) $\mathbb{E}^q\left[u(X_t)\right] \approx \bar{u}$. Let $p^q(x,t)$ indicates the marginal distribution on X_t conditioned on $X_0 \sim q$. We apply Cauchy-Schwartz:

$$\left|\mathbb{E}^q\left[u(X_t)\right] - \bar{u}\right| = \int_N (u(x) - \bar{u})\left(p^q(x,t) - \frac{1}{|N|}\right) \leq \sqrt{\int_N (u(x) - \bar{u})^2 \int_N \left(p^q(x,t) - \frac{1}{|N|}\right)^2}$$

We see that we have two forms of error to control:

i. Lemma 4 controls the variance term:

$$\int_N (u(x) - \bar{u})^2 \leq c_N \left(\frac{2\pi^{\frac{D}{2}}(D-2)}{\Gamma\left(\frac{D}{2}\right)\left(1 - \tilde{\beta}^{2-D}\right)}\right) \frac{r_A^{D-2} r_B^{D-2}}{r_A^{D-2} + r_B^{D-2}}$$

ii. The second term is controlled by applying Gronwall's inequality applied with the Poincare inequality to the Fokker-Planck equation $p_t(x,t) = \Delta_x p(x,t)/2$ (cf. Appendix B.1). We obtain that

$$\int_N \left(p^q(x,t) - \frac{1}{|N|}\right)^2 \leq e^{-t/c_N} \int_N \left(p^q(x,0) - \frac{1}{|N|}\right)^2$$

This is where the non-degeneracy of q is so important. Noting that $\left\|q - |N|^{-1}\right\|^2_{\mathscr{L}^2(N)} < |S|^{-1}$ and $|S|^{-1} = \Gamma\left(D/2+1\right)\pi^{-D/2}\left(\frac{\beta-1}{2\beta}\left(\beta r_A \wedge \beta r_B\right)\right)^{-D}$, we translate this into the bound

$$\int_N \left(p^q(x,t) - \frac{1}{|N|}\right)^2 \leq e^{-t/c_N}\Gamma\left(\frac{D}{2}+1\right)\pi^{-D/2}\left(\frac{\beta-1}{2\beta}\left(\beta r_A \wedge \beta r_B\right)\right)^{-D}$$

All together, cancelling the Γ functions and the $\pi^{D/2}$ terms,

$$|\mathbb{E}^q[u(X_t)] - \bar{u}|^2 \leq e^{-t/c_N} D \left(\frac{\beta-1}{2\beta}\right)^{-D} (\beta r_A \wedge \beta r_B)^{-D} \left(\frac{D-2}{1-\tilde{\beta}^{2-D}}\right) \frac{r_A^{D-2} r_B^{D-2}}{r_A^{D-2} + r_B^{D-2}}$$

Proof. Putting (1) and (2) together completes the proof for the case $x_0 \in \partial\mathscr{B}(x_A, \tilde{\beta} r_A) \cup \partial\mathscr{B}(x_B, \tilde{\beta} r_B)$. However, since the set $N \backslash \mathscr{B}(x_A, \tilde{\beta} r_B) \backslash \mathscr{B}(x_A, \tilde{\beta} r_B)$ is connected (and $u - \bar{u}$ is harmonic with vanishing normal derivative on ∂N), the maximum principle gives that this bound actually holds everywhere away from $\mathscr{B}(x_A, \tilde{\beta} r_B), \mathscr{B}(x_A, \tilde{\beta} r_B)$. □

One difficulty in applying this lemma is in choosing the correct value of t. However, Corollary 5 gives us a fairly good value of t to work with. The result is a bound on $|u - \bar{u}|$ which can be made uniform over regions at some distance from A, B.

Corollary 7. *Fix d_A, d_B, and c_N.*

- For all fixed $\gamma > 0, \epsilon > 0, D \geq 3$ one can pick $c > 0$ small enough so that $|u(x_0) - \bar{u}| < \epsilon$ for all x_0 outside of $\mathscr{B}(x_A, \gamma), \mathscr{B}(x_B, \gamma)$ whenever $r_A, r_B < c$.
  - For all $\gamma > 1, \epsilon > 0, r_A, r_B$ one can pick D large enough so that $|u(x_0) - \bar{u}| < \epsilon$ for all x_0 outside of $\mathscr{B}(x_A, \gamma r_A), \mathscr{B}(x_B, \gamma r_B)$.

Proof. Applying the lemma above with with Corollary 5, we get

$$\begin{aligned} |u(x_0) - \bar{u}| \quad \leq \quad & 12\beta^{\frac{2-D}{2}} e^{3(D-2)\beta^{-1+\frac{2-D}{4}}} \\ & + \sqrt{c_N e^{-(\beta r_A \wedge \beta r_B)^2 \frac{\beta^{\frac{D-2}{2}}}{D(D-2)}/c_N} D \left(\frac{\beta-1}{2\beta}(\beta r_A \wedge \beta r_B)\right)^{-D} \left(\frac{D-2}{1-\tilde{\beta}^{2-D}}\right) \frac{r_A^{D-2} r_B^{D-2}}{r_A^{D-2} + r_B^{D-2}}} \end{aligned}$$

for all $x_0 \in \partial\mathscr{B}(x_A, \tilde{\beta} r_A) \cup \partial\mathscr{B}(x_B, \tilde{\beta} r_B)$. Now

□

- Fix $\gamma > 0, D \geq 3$. Note that if we could set $\beta r_A = \beta r_B = \gamma$ steady and send r_A, r_B to zero, the bound would become arbitrarily small. Unfortunately, our assumptions do not allow us control over the ratio r_A/r_B; if this ratio is very bad, the power of our our bound could run into some artifacts, due to the fact that we used the same coefficient β for our spheres which contain A and B. Fortunately, for any $k > r_A$, $A \subset \mathscr{B}(x_A, r_A) \implies A \subset \mathscr{B}(x_A, k)$; we can use this to adjust r_A, r_B to control any artifacts.
  - Fix $\gamma > 1, r_A, r_B$. Pick $\beta = \gamma$ and send $D \to \infty$; note that the bound becomes arbitrarily small. This case is slightly less obvious because of terms such as $\left(\frac{\beta-1}{2\beta}\right)^{-D}$, which is generally exploding with D. However, all of these terms are completely outclassed by $\exp\left(-\left(\beta r_A \wedge \beta r_B\right)^2 \frac{\beta^{\frac{D-2}{2}}}{D(D-2)}/c_N\right)$.

It is worth noticing that if d_A, d_B shrink along with r_A, r_B, we do *not* get any such guarantees. Indeed, if A, B are arbitrarily small and yet their environment is *also* correspondingly small, we really cannot say anything. If space is cramped and tiny, it becomes impossible to show that $\mathbb{P}(\tau \leq t)$ is small. Thus, our results will only hold if A, B have significant breathing room around them, i.e. a ball around them is completely empty.

3.4. Main theorem

We have now shown that $|u - \bar{u}| < \epsilon$ on a large portion of its domain; that is, u is really very flat. All that remains is discover the actual value of $\bar{u}$.

We will show that $\bar{u} \approx cap\left(A\right)/\left(cap\left(A\right) + cap\left(B\right)\right)$, in two steps:

1. $\bar{u} \approx \kappa_A/\left(\kappa_A + \kappa_B\right)$ - here the reader may recall κ_A, κ_B are certain condenser capacities from Section 3.2

2. $\kappa_A/\left(\kappa_A + \kappa_B\right) \approx cap\left(A\right)/\left(cap\left(A\right) + cap\left(B\right)\right)$

This will let us prove our main theorem.

3.4.1. The average value of u is roughly determined by κ_A, κ_B

We now return to the variational variational characterization of u, i.e. $\int |\nabla u|^2 \leq \min \int |\nabla w|^2$ among all w satisfying the same boundary conditions.

Having demonstrated a uniform bound on $|\bar{u} - u|$ in a region at modest distance from A, B, we now consider the relationship between $\bar{u}$ and $\int |\nabla u|^2$. Note that if $\bar{u}$ is too high, the drop from $\bar{u}$ to $u(\partial B) = 0$ will create a large integrated gradient. If $\bar{u}$ is too small, the climb from $\bar{u}$ to $u(\partial A) = 1$ will also create a very large integrated gradient. To make this idea rigorous, we need to show a lower bound on $\int |\nabla f|^2$ for any function which climbs from a low value to a high value:

Lemma 8. *Let $C \subset C' \subset \mathbb{R}^D$, open bounded sets with thrice continuously differentiable boundaries. Then, for any $f \in H^1(C' \backslash C)$ with $f(\partial C) \leq a \leq b \leq f(\partial C')$,*

$$\int_{C' \backslash C} |\nabla f|^2 \geq (b-a)^2 \, cap\left(C \subset C'\right)$$

Proof. Note that $\int |\nabla f|^2$ scales quadratically with multiplication. Thus, WLOG, we will assume that $a = 0$ and $b = 1$. We will also assume that $f = 0$ on C and $f = 1$ on $\mathbb{R}^D \backslash C'$ - this doesn't change $\int_{C' \backslash C} |\nabla f|^2$.

Define $S_0 = \{x : \ f(x) \leq 0\}$ and $S_1 = \{x : \ f(x) \leq 1\}$ we obtain $C \subset S_0 \subset S_1 \subset C'$. Take

$$\tilde{f} = \begin{cases} 0 & x \in S_0 \\ 1 & x \in C' \backslash S_1 \\ f(x) & \text{else} \end{cases}$$

Then observe that:

- $\int_{C' \backslash C} |\nabla f|^2 = \int_{S_0} |\nabla f|^2 + \int_{C' \backslash S_1} \int |\nabla f|^2 + \int \left|\nabla \tilde{f}\right|^2 \geq \int \left|\nabla \tilde{f}\right|^2$
  - $\tilde{f}$ satisfies $\tilde{f}(\partial C) = 0$ and $\tilde{f}(\partial C') = 1$, so $\int \left|\nabla \tilde{f}\right|^2 \geq cap\left(C \subset C'\right)$

□

– $\tilde{f} \in H^1$

So $\int |\nabla f|^2 \geq \int \left|\nabla \tilde{f}\right| \geq cap\left(C \subset C'\right)$, which concludes the proof.

Our result now follows from combining Lemma 8 and Lemma 4:

Lemma 9. *Let $|u(x_0) - \bar{u}| < \epsilon$ for all x_0 outside of $\mathscr{B}\left(x_A, \tilde{\beta} r_A\right), \mathscr{B}\left(x_B, \tilde{\beta} r_B\right)$. Then*

$$\left|\bar{u} - \frac{\kappa_A}{\kappa_A + \kappa_B}\right| \leq 2\sqrt{\epsilon}$$

Proof. By assumption, $u(x_0) < \bar{u} + \epsilon$ for $x_0 \in \partial\mathscr{B}\left(x_A, \tilde{\beta} r_A\right)$, and $u(\partial A) = 1$. That is, u must climb up from $\bar{u}+\epsilon$ to 1. It follows from Lemma 8 that the integrated squared norm of the gradient must be large in that climb, i.e. $\int_{\mathscr{B}\left(x_A, \tilde{\beta} r_A\right)} |\nabla u|^2 \geq (1 - \bar{u} - \epsilon)^2 \kappa_A$. Using the fact that $u(x_0) > \bar{u} - \epsilon$ for $x_0 \in \partial\mathscr{B}\left(x_B, \tilde{\beta} r_B\right)$ and $u(\partial B) = 0$, the same argument yields that $\int_{\mathscr{B}\left(x_B, \tilde{\beta} r_B\right)} |\nabla u|^2 \geq (\bar{u} - \epsilon)^2 \kappa_B$. Putting those together, we have a *lower* bound on $\int |\nabla u|^2$.

$$\int |\nabla u|^2 \geq (1 - \bar{u} - \epsilon)^2 \kappa_A + (\bar{u} - \epsilon)^2 \kappa_B$$
$$= (1 - \bar{u})^2 \kappa_A + \bar{u}^2 \kappa_B + 2\epsilon\left(\epsilon - (1 - \bar{u})\kappa_A - \epsilon\bar{u}\kappa_B\right)$$

On the other hand, we already know $\int |\nabla u|^2$ can't be too big; as stated in Lemma 4,

$$\int |\nabla u|^2 \leq \frac{\kappa_A \kappa_B}{\kappa_A + \kappa_B} = \min_h \left((1 - h)^2 \kappa_A + h^2 \kappa_B\right)$$

Together, these bounds create a quadratic inequality on $\bar{u}$:

$$(1 - \bar{u})^2 \kappa_A + (\bar{u})^2 \kappa_B + 2\epsilon\left(\epsilon - (1 - \bar{u})\kappa_A - \epsilon\bar{u}\kappa_B\right) \leq \min_h \left((1 - h)^2 \kappa_A + h^2 \kappa_B\right)$$

Noting the similarity of the LHS and RHS when ϵ vanishes, it should not be hard to believe that this inequality will force $\bar{u} \approx \arg\min_h (1 - h)^2 \kappa_A + h^2 \kappa_B$. Indeed, it is

straightforward to compute the result of this quadratic inequality on $\bar{u}$. The roots are given by

$$\frac{\kappa_A - \epsilon(\kappa_A + \kappa_B) \pm 2\sqrt{(1-\epsilon)\epsilon\kappa_A\kappa_B}}{\kappa_A + \kappa_B}$$

Applying the inequality $\frac{\sqrt{ab}}{a+b} \leq \frac{1}{2}$, and $(1-\epsilon) < 1$ we have

$$\left| \bar{u} - \frac{\kappa_A}{\kappa_A + \kappa_B} \right| \leq \epsilon + \sqrt{\epsilon} \leq \epsilon + \sqrt{\epsilon}$$

Noting that $|u(x_0) - \bar{u}| \leq 1$ trivially, we have $\epsilon < 1$, so $\epsilon < \sqrt{\epsilon}$. Our result follows. □

We have now shown both that that $|u - \bar{u}| < \epsilon$ on a large portion of its domain and that that $\bar{u}$ is close to $\frac{\kappa_A}{\kappa_A + \kappa_B}$. Our result is now close at hand.

3.4.2. The condenser capacity and the harmonic capacity are not so different

We now show that

$$h^* = \frac{\kappa_A}{\kappa_A + \kappa_B} \approx \frac{cap(A)}{cap(A) + cap(B)} = h^{**}$$

First we put bounds on the individual capacities:

Lemma 10. *Let $A \subset \mathscr{B}(x_0, r_1)$. Then for any $r_2 > r_1$,*

$$cap(A) \leq cap(A, \mathscr{B}(x_0, r_2)) \leq cap(A) / \left(1 - \left(\frac{r_1}{r_2}\right)^{D-2}\right)^2$$

Proof. Let w, v denote the capacity functions for $cap(A, \mathscr{B}(x_0, r_2)), cap(A)$, respectively.

The lower bound is easy to see. We know w is the unique minimizer of $\int |\nabla g|^2$ subject to $g(\partial A) = 1$ and $g(\partial\mathscr{B}(x_0, r_2)) = 0$, whereas v is the minimizer of the same objective, subject only to $g(\partial A) = 1$ and $g(\infty) = 0$. Thus, since v is less constrained, $cap(A) = \int |\nabla v|^2$ will certainly achieve a lower value than $cap(A, \mathscr{B}(x_0, r_2)) = \int |\nabla w|^2$. This

proves the lower bound.

To prove the upper bound, let us consider the value of v when $x \in \partial\mathscr{B}(x_0, r_2)$. This value indicates the probability that a Brownian motion started at x will ever hit A. This probability is certainly less than the chance that a Brownian motion started at x will ever hit $\partial\mathscr{B}(x_0, r_1)$, which is given explicitly by $\frac{|x-x_0|^{2-D}}{r_1^{2-D}}$. Thus $x \in \partial\mathscr{B}(x_0, r_2) \implies v(x) \leq \frac{r_2^{2-D}}{r_1^{2-D}}$. It follows that

$$\tilde{v}(x) = \frac{v(x) - \frac{r_2^{2-D}}{r_1^{2-D}}}{1 - \frac{r_2^{2-D}}{r_1^{2-D}}}$$

satisfies both $\tilde{v}(\partial A) = 1$ and $\tilde{v}(\partial\mathscr{B}(x_0, r_2)) \leq 0$. Lemma 8 therefore tells us that $\frac{cap(A)}{\left(1-\frac{r_2^{2-D}}{r_1^{2-D}}\right)^2} = \int |\nabla\tilde{v}|^2 \geq \int |\nabla w|^2 = cap\,(A, \mathscr{B}(x_0, r_2))$. □

Corollary 11. $\left|\frac{\kappa_A}{\kappa_A+\kappa_B} - \frac{cap(A)}{cap(A)+cap(B)}\right| \leq \left(\frac{\tilde{\beta}^{2-D}}{1-\tilde{\beta}^{2-D}}\right)$

Apply Lemma 10 to argue that

$$\begin{aligned} cap\,(A) &\leq \kappa_A \leq cap\,(A) \,/\, \left(1-\tilde{\beta}^{2-D}\right)^2 \\ cap\,(B) &\leq \kappa_B \leq cap\,(B) \,/\, \left(1-\tilde{\beta}^{2-D}\right)^2 \end{aligned}$$

It is straightforward to calculate that therefore

$$\begin{aligned} \frac{\kappa_A}{\kappa_A+\kappa_B} - \frac{cap\,(A)}{cap\,(A)+cap\,(B)} &\leq \frac{cap\,(A)}{cap\,(A)+cap\,(B)} \times \tilde{\beta}^{2-D}\left(\frac{1}{1-\tilde{\beta}^{2-D}}\right) \\ \frac{\kappa_A}{\kappa_A+\kappa_B} - \frac{cap\,(A)}{cap\,(A)+cap\,(B)} &\geq \frac{cap\,(A)}{cap\,(A)+cap\,(B)} \times \tilde{\beta}^{2-D}\left(1-\tilde{\beta}^{2-D}\right) \end{aligned}$$

Note that that $1/\left(1-\tilde{\beta}^{2-D}\right) > \left(1-\tilde{\beta}^{2-D}\right)$. So, choosing this larger bound on either

side,

$$\begin{aligned}\left|\frac{\kappa_A}{\kappa_A+\kappa_B}-\frac{cap\left(A\right)}{cap\left(A\right)+cap\left(B\right)}\right| &\leq \frac{cap\left(A\right)}{cap\left(A\right)+cap\left(B\right)}\left(\frac{\tilde{\beta}^{2-D}}{1-\tilde{\beta}^{2-D}}\right)\\ &\leq \left(\frac{\tilde{\beta}^{2-D}}{1-\tilde{\beta}^{2-D}}\right)\end{aligned}$$

3.4.3. Final statement

Our theorem now follows immediately:

Theorem. *Fix d_A, d_B, c_N. Define $h^{**} = cap\left(A\right) / \left(cap\left(A\right) + cap\left(B\right)\right)$.*

- For any $D, \gamma > 0, \epsilon > 0$ we can pick $c > 0$ so that $|\mathbb{P}^x(X_\tau \in \partial A) - h^{**}| < \epsilon$ for all x outside of $\mathscr{B}\left(x_A, \gamma\right), \mathscr{B}\left(x_B, \gamma\right)$ for all scenarios with $r_A, r_B < c$.
  - For any $r_A, r_B, \gamma > 1, \epsilon > 0$ we can pick c so that $|\mathbb{P}^x(X_\tau \in \partial A) - h^{**}| < \epsilon$ for all x outside of $\mathscr{B}\left(x_A, \gamma r_A\right), \mathscr{B}\left(x_B, \gamma r_B\right)$ for all scenarios with $D > c$.

Proof. Without loss of generality, assume $\epsilon < 1$. Apply Corollary 7 to make $|u(x_0) - \bar{u}| \leq \epsilon^2/9$. We can then apply Lemma 9 to show $|\bar{u} - \kappa_A/\left(\kappa_A + \kappa_B\right)| \leq \epsilon/3$. Finally, taking a more extreme r_A, r_B or D if necessary, ensure that $\tilde{\beta}^{2-D}/\left(1-\tilde{\beta}^{2-D}\right) \leq \frac{\epsilon}{3}$. Corollary 11 and the triangle inequality then give the desired result, since $\epsilon^2/9 + \epsilon/3 + \epsilon/3 < \epsilon$. □

3.5. Empirical results

The theoretical results above are only asymptotic, and we were interested to know whether the results of the theorem were relevant in regimes which one might be interested in.

To this end, we ran some empirical simulations in which

- $N = \mathscr{B}(0, 1) \subset \mathbb{R}^5$
- $A = \mathscr{B}((-.8, 0, 0, 0, 0), .1)$

- $B = \mathscr{B}((.5, .6, 0, 0, 0), .05)$

Starting at various values of $x_0 \notin \mathscr{B}(x_A, .3) \cup \mathscr{B}(x_B, .2)$, we then asked how often we hit A versus B. Our theorem would argue that the probability of hitting A versus B would be relatively independent of x_0. This probability should be roughly

$$\mathbb{P}^x(X_\tau \in \partial A) \approx \frac{cap\,(A)}{cap\,(A) + cap\,(B)} = \frac{.1^3}{.05^3 + .1^3} = \frac{8}{9} = 0.\bar{8}$$

However, we wanted to know what actually happens. Using Monte-Carlo simulations, we were able to obtain fairly accurate estimates based on different starting conditions. Here are some examples:

- $x_0 = (.21, -.59, .003, .67, -.179)$ - $\mathbb{P}^{x_0}(X_\tau \in \partial A) = .893$
- $x_0 = (-.22, .19, -.029, -.075, -.29)$ - $\mathbb{P}^{x_0}(X_\tau \in \partial A) = .883$
- $x_0 = (0.62, -0.33, 0.072, 0.066, -0.61)$ - $\mathbb{P}^{x_0}(X_\tau \in \partial A) = .905$
- $x_0 = (.537, -0.051, .0996, .70913, -0.313)$ - $\mathbb{P}^{x_0}(X_\tau \in \partial A) = .885$
- $x_0 = (-0.26, 0.52, 0.61, 0.09, -0.268)$ - $\mathbb{P}^{x_0}(X_\tau \in \partial A) = .908$

These numbers are all pretty close to 8/9. Other points yielded similar results. This suggests that the theorem is at least somewhat in effect even in this regime. In regimes where the sets were larger or the initial position x_0 was closer to A, B, the performance began to degrade. However, it was not completely obvious how the accuracy of the theorem's approximation depended upon interaction between the various parameters. Since there are so many parameters involved, it is a difficult question to try to attack empirically.

This suggests a natural direction for future research. The asymptotic bounds that we obtained in the general case were necessarily quite poor. However, in the special case that A, B, N are all spheres, it should be possible to obtain a much more accurate estimate of

the error. Although this simple setup may not directly apply to any practical problems, it might be very useful for giving at least some indication of how the error would behave in more complex settings.

3.6. Conclusions

There are several assumptions in this work that make it difficult to apply directly to the problem of biological polymers.

- We have assumed that the energy landscape is actually flat inside of N, instead of merely being mostly flat. It does not seem that there is anything essential about our arguments that depend upon this fact. A more careful analysis is necessary to determine if the gist of our idea carries over to that case.

- We have assumed that A, B do not intersect the boundary of N. This is perhaps implausible in many cases. A next step would be to consider the case that A, B are actually *subsets* of the boundary of N. More generally, if A, B are arbitrary sets which intersect the boundary of N, we may still be able to say some things. For example, we may still be able to show that the probability of hitting A or B does not depend strongly upon the initial condition, X_0.

If these difficulties could be overcome, techniques in the style of those presented could potentially allow us to gain much greater understanding of the large, boring regions of space which are so very difficult to simulate in practice. The very "boringness" of the regions gives us optimism that some efficient analysis should be possible.

4. Medical Treatments and the "Worst Probability of Harm"

It is universally acknowledged that the general public finds medical research confusing. From the New York Times to the Public Library of Science, confusion reigns over the host of contradictory results from the science of medicine (cf. [23, 20]). Many theories have been put forward to explain this bewildering state of affairs, including p-hacking, the undue influence of business priorities on the scientific process, and exaggerated representations of science by the media. However, there are other causes which are much simpler and much easier to treat.

The disregard for effect size is one clear culprit. Medical practitioners are well aware that scientific studies with small effect sizes should not be taken too seriously (cf [33]). One simple reason is that a small effect may not be worth the time it takes to think about it. What's worse, however, is that a small effect may be devastatingly sensitive to seemingly trivial variations in the experimental setup. These sensitivity issues are already at the forefront of genomic science, where so-called "batch effects" play havoc with statistics (cf. [17]). The problem is just as palpable in medical science, where two experiments that seem almost identical can nonetheless yield two contradictory results, both of which are statistically significant at the .001 level. A study may find a result which is statistically significant, but it in no way follows that the result is scientifically significant. As long as one gathers enough data, almost every experiment is doomed to find a statistically significant result - whether or not that result says anything about how

life works or how people should behave.

Fortunately, there is a growing attention to effect size and the many different methods that people use to quantify it. One common difficulty is that any measure of effect size must be at least somewhat intuitive. If a quantity is too abstract, it may be ignored by practitioners. Here we will try to develop a measure of effect size that has a clear, intuitive meaning, and makes as few assumptions as possible. Our method originated from something called the Number Needed to Treat (NNT), used for experiments with binary outcomes (e.g. life or death). We develop this intuitive idea into a new measure, called the "worst probability of harm." Our goals for this measure are twofold:

- To clarify certain assumptions around the NNT
- To generalize the essence of the NNT to the case when the outcomes are not binary

The NNT actually does an excellent job of measuring effect size, and our main focus is on the second point. However, a proper understanding of the assumptions of the NNT lead naturally to the generalization we will propose. For this reason we will take some time exploring those assumptions. We proceed as follows:

1. *Binary outcomes.* We will try to think clearly about the assumptions implicit in the NNT measure of effect size. A new measure of effect size will naturally present itself: "the worst probability of harm." This section provides context, but it is not essential for understanding the remainder of the article.

2. *Non-binary outcomes.* We will see that the worst probability of harm is also well-defined for treatments with non-binary outcomes.

3. *Basic properties.* In the general non-binary case, the best way to calculate the worst probability of harm is not at all obvious. However, it turns out to be straightforward. Formulae are presented.

4. *Examples.* We will show that the worst probability of harm yields some rather startling results when applied to a variety of articles in medical science.

5. *Statistical considerations.* We will conclude by considering some nonpararmetric methods of estimating the worst probability of harm from actual data.

4.1. Binary outcomes

Let us consider the set of patients with a given condition, such as cancer or heart disease. Let us additionally consider two mutually exclusive treatment options which are available to the patients. We will call them treatment A and B. Perhaps treatment B is actually some form of placebo, or no treatment at all. In evaluating these treatments, we often look at some indication of patient success. For example, we might look at whether the patient is still alive three years after treatment.

This sort of situation is a good place to begin our exploration of effect size. After performing a Randomized Controlled Trial (RCT), we obtain two numbers:

- q_A - the proportion of people who live when given treatment A
- q_B - the proportion of people who live when given treatment B

Based on just these two numbers, a practitioner needs to have some way of understanding the medical importance of a given treatment.

We'll start by looking at the Number Needed to Treat (NNT), a common way to measure effect size for this situation. It carries a powerful intuition, making it a great practical tool to help practitioners understand what matters and what doesn't. However, as we shall see, it also carries a hidden assumption. Taking a hard look at that assumption will naturally suggest a new measure of effect size, which is similar to the NNT but doesn't make any assumptions.

Conceptually, the NNT is based on the idea that the population of patients may be divided into four groups:

	Live	Die
A	$q_A = 80\%$	$1 - q_A = 20\%$
B	$q_B = 30\%$	$1 - q_B = 70\%$

		A	
		Live	Die
B	Live	p_W =?	p_B =?
	Die	p_A =?	p_L =?

$$p_W + p_B = q_B = 30\%$$
$$p_A + p_L = 1 - q_B = 70\%$$
$$p_W + p_A = q_A = 80\%$$
$$p_B + p_L = 1 - q_B = 20\%$$

Table 4.1.: An example of a Randomized Controlled Trial (RCT). The result of the experiment are shown in the left table. These results tell us two things. First, if you pick a random person and give them treatment A, they will live with 80% probability. Second, if you pick a random person and give them treatment B, they will live with 30% probability. However, this left table doesn't tell us everything we might want to know. For example: how many people would live *regardless* of which treatment they were given? That information cannot be deduced from q_A, q_B. This proportion corresponds to to the variable p_W, which is represented as one of the cells in the table on the right. All we can say about the table on the right is that we know the *row and column sums.* These row and column sums correspond to the four equations shown. Note that the fourth equation is actually redundant to the first three, so in effect we have four unknowns and three equations. This leaves a degree of uncertainty about the values of p_W, p_B, p_A, p_L.

- The winners - let p_W denote the proportion of people who will survive no matter whether they get treatment A or B
- The losers - let p_L denote the proportion of people who will die no matter whether they get treatment A or B
- The treatment-A population - let p_A denote the proportion of people who will only live if given treatment A
- The treatment-B population - let p_B denote the proportion of people who will only live if given treatment B

It would be nice if we could determine the percentage of people in each of these groups.

Unfortunately, an RCT *cannot* directly provide this information. Instead, an RCT can only tell us $q_A = p_W + p_A$ and $q_B = p_W + p_B$. This is due to the fact that the treatment options are mutually exclusive. We cannot give a patient both treatments (or, put another way, giving them both treatments would really be yet a third kind of treatment). Therefore, once we have decided to give a patient treatment A, there is simply no way to determine how they might have fared if we had given them treatment B.

However, we can nonetheless say something about those four values, p_B, p_W, p_A, p_L. To see how it works, we'll take an example - say $q_A = 80\%$ and $q_B = 30\%$. That is, on average, patients survive more often when given the treatment A. This is not enough information to determine p_B, p_W, p_A, p_L, but we can make a *guess* at those values. For example, we might hazard the guess that treatment never hurts - that is, $p_B = 0$. If we assume this is true and then apply our knowledge of q_A, q_B, all four quantities can be deduced:

$$\begin{aligned} p_B &= 0\% \\ p_W &= 30\% = q_B \\ p_A &= 50\% = q_A - q_B \\ p_L &= 20\% = 1 - q_A \end{aligned}$$

Under this assumption, the number needed to treat is calculated as

$$\mathrm{N}NT \triangleq 1/\left(p_A\right) = 1/\left(q_A - q_B\right)$$

In our case, $\mathrm{N}NT = 1/\left(.5\right) = 2$. Small values of the NNT suggest that p_A is large, which signifies that there is a large proportion of people who will only live if given treatment A. If $\mathrm{N}NT = 2$, that suggests a very optimistic picture for a practitioner: for every 10 patients they give treatment A, the practitioner will save 5 lives.

However, there is a very important assumption that has been made here. It is assumed that nobody's primary outcome is actually made *worse* by the adoption of treatment A. An article may describe certain negative secondary outcomes or "side-effects" of the treatment, but these are assumed to be the only harm that the treatment causes. However, it is possible that the treatment actually causes harm to the primary outcome for some patients. Here is another set of values which is *equally consistent* with the experimental data, $q_A = 80\%$ and $q_B = 30\%$:

$$p_B = 20\%$$
$$p_W = 10\%$$
$$p_A = 70\%$$
$$p_L = 0\%$$

This presents a rather different picture. For every 10 people a practitioner gives treatment A, they will actually save 7 lives - but they will also kill 2 people. What's more, since $p_L = 0$, we could have saved *every patient* if we only knew which treatment to give them. If these numbers were true, it would be imperative to try to figure out why some people only live under treatment A and some people only live under treatment B. This could be accomplished by analyzing how covariates influence treatment outcomes (e.g. [19], [13]). For example, perhaps younger patients do better with treatment B.

To properly highlight the fact that we don't know exactly what kind of picture we are actually dealing with, we propose a new measure of effect size.

Definition. Given q_A, q_B, the **worst probability of harm** for adopting treatment A over treatment B indicates the highest possible probability that a patient will actually do worse taking treatment A. That is, the largest possible value of p_B, subject to the requirement that $p_W + p_A = q_A$ and $p_W + p_B = q_B$. The **worst probability of ineffectiveness** indicates the highest possible probability that treatment A will not

improve a patients outcome. That is, the largest possible value of $1 - p_A$, subject to the same requirements.

In the case of the example above, with $q_A = 80\%$ and $q_B = 20\%$, we may calculate that the largest possible value of q_B is 20% (if it were any higher, we could not have $p_L + p_B = 1 - q_A = 20\%$). Thus the worst probability of harm is 20%. This quantity may give additional context to the significance of the NNT as it is usually calculated. It has a very clear message to the practitioner: on average, for every 10 patients they treat, they may be killing as many as 2 although they save as many as 7. It is important to note that we cannot be sure that the treatment kills 20% of patients; another possibility, equally consistent with the data, is that the treatment doesn't kill anyone. Perhaps it only saves 50% of patients and has no effect on the other 50%. However, it seems important that the practitioner should understand all the possibilities which are fully consistent with the experimental evidence.

Assuming this ambiguity is clear in the mind of the practitioner, we feel that the NNT remains perhaps the overall best tool to express the effect size measured by a binary-outcome experiment. Unfortunately, for experiments with non-binary outcomes, there is no obvious generalization of this concept. Fortunately, the worst probability of harm (which we hope captures something of the essence of the NNT) does generalize to non-binary outcomes.

4.2. Non-binary outcomes

Let us now assume that treatment outcome can be measured as a value in $\Omega \subset \mathbb{R}$, with more positive numbers indicating a better outcome. In this case, we may be interested

in the following kinds of proportions:

$$F_{X,Y}(x,y) \triangleq \text{proportion of people for whom}$$
$$\text{their outcome is} \leq x \qquad \text{if given treatment } A$$
$$\text{their outcome is} \leq y \qquad \text{if given treatment } B$$

In other words, let

- X denote the outcome of a patient if they were given treatment A
- Y denote the outcome of a patient if they were given treatment B

then $F_{X,Y}(x,y)$ is the joint cumulative distribution function of X, Y for a randomly selected patient. Given enough subjects, an RCT can readily estimate F_X and F_Y, the marginal distribution of X and Y. As before, an RCT cannot directly estimate the joint. However, if we know the marginal distributions, we can determine the worst probability of harm:

Definition. Let F_X, F_Y denote two cumulative distribution functions corresponding to the outcomes of patients under treatments A, B respectively. We define the **worst probability of harm** for (F_X, F_Y) as the highest possible probability that adopting treatment A over treatment B will result in an inferior outcome, i.e. the quantity $\sup_Q Q(X < Y)$ among all distributions Q for which $X \sim F_X$ and $Y \sim F_Y$. Correspondingly, we define the **worst probability of ineffectiveness** as $\sup_Q Q(X \leq Y)$.

Let us contrast the worst probability of harm with measures of central tendency, such as mean or median. Medical practitioners often rely heavily on measures of central tendency in making their decisions; if $\mathbb{E}[X] > \mathbb{E}[Y]$ then treatment A may be considered preferable. Yet even though $\mathbb{E}[X] > \mathbb{E}[Y]$ there are cases in the literature in which the worst probability of harm is 80% or more. That is, it may be that the treatment actually harms *more than 80% of patients to whom it is applied.* In general, a naive consideration

of mean or median may be quite misleading. This is especially troubling since the NNT is not available in this continuous case. In this case, the worst probability of harm can offer insight into how a treatment may affect patients.

Of course, to make use of these concepts, we need to be able to compute them. Although it is not obvious, this turns out to be relatively straightforward, as we shall see in the next section.

4.3. Basic properties and computation

Theorem. *Let F_X, F_Y denote two cumulative distribution functions. Let $\tilde{\lambda}$ denote the worst probability of ineffectiveness. Let $\tilde{\gamma}(k) = 1 - F_Y(k) + F_X(k)$ and $\tilde{\gamma}^* \triangleq 1 \wedge \inf_k \tilde{\gamma}(k)$. Then $\tilde{\lambda} = \tilde{\gamma}^*$ and there exists a distribution Q which achieves $Q(X \leq Y) = \tilde{\lambda}$.*

Proof. Cf. [11]. □

4.4. Examples

To show what kind of insights we might hope to obtain from calculating the worst probability of harm, we looked at six papers from the New England Journal of Medicine. We picked the papers purely on the basis that they

- had some continuous-valued outcome,
- provided enough information for us to get a sense of the distributions of those outcomes under two treatments (e.g. standard errors and number of subjects, interquartile ranges, etc)
- were randomized and controlled, and
- had a statistically significant result at the .05 level or below.

Somewhat astonishingly, the worst probability of harm was greater than 75% for all but one of the papers. That is, it may well be that the treatments advocated by five of these papers actually could harm 75% of the patients to whom they are applied. Here are the results.

1. Worst probability of harm: 86%. A study on the efficacy of estradiol in improving the vascular health for postmenopausal patients [18]. The primary outcome was the rate of change in carotid-artery intramedia thickness (CIMT). By randomly assigning treatment to 643 patients, the study found that the following two treatment conditions were associated with roughly the following two distributions of outcomes:

   - CIMT of patients given estradiol - $\mathcal{N}\left(.0044, .00996698^2\right)$
   - CIMT of patients given placebo - $\mathcal{N}(.0078, .010184^2)$

   This suggests that, on average, estradiol slows the rate at which arteries thicken - a good thing. However, we ask whether it is possible that estradiol may in some cases actually increase the rate at which arteries thicken. To answer that question, let us calculate the worst probability of harm. Since good outcomes are associated with lower CIMT scores, we take $X \sim \mathcal{N}\left(-.0044, .00996698\right)$ and $Y \sim \mathcal{N}(-.0078, .010184^2)$ and calculate the worst probability of harm as $\sup_Q Q(X < Y)$ over all Q with the appropriate marginal distributions on X, Y. Notice that the variances for the distributions of X, Y here are *population* variances, rather than the standard errors of the mean estimators - this difference comes out to a crucial factor of $\sqrt{n}$, where n is the number of patients. We emphasize this point here because the variance plays a very important role in determining the worst probability of harm.

   If we take it as exact truth that the distributions under the two outcomes are given by the Gaussians described above, we can compare the cumulative distri-

bution functions for these two distributions and then numerically calculate that $\sup_Q Q(X < Y) = .86574\cdots$. That is, it may be that estradiol actually *worsens* the situation for as much as 86% of the population. In retrospect, this should not be terribly shocking. If we look at the scale of the population variances (which are quite large), the difference between the population means is almost negligible. Due to the large sample size, that difference is indeed statistically significant, but the effect size is simply negligible. This manifests in the fact that we cannot be sure that the treatment is not actually harming most patients.

2. Worst probability of harm: 80%. A study on the effects of spinal fusion for lumbar spondylolisthesis in the context of laminectomy. The primary outcome was the change in a physical health summary score after two years. Here are the distributions:[16]

   - Health score change for patients with fusion - $\mathcal{N}\left(15.2, 11.61^2\right)$
   - Heath score change for patients without fusion - $\mathcal{N}\left(9.5, 11.81^2\right)$

   Since $15.2 > 9.5$, the authors conclude that fusion is generally good thing for patients. We caution, however, that if we take these distributions as truth then it may be that as many as 80.1% of patients would do better with the nonfusion treatment than with the fusion treatment. We again emphasize that this large probability in no way contradicts the fact that the average outcome in the fusion treatment is better than the average outcome in the nonfusion treatment.

3. Worst probability of harm: 90%. A study on the effects of testosterone treatment on the results of the Psychosexual Daily Questionnaire [32]. They find the following distributions:

   - Score for patients given testosterone treatment - $\mathcal{N}\left(.2, 1.6^2\right)$
   - Score for patients given placebo - $\mathcal{N}\left(-.1, 1.4^2\right)$

Taking these distributions as truth, we calculate that it may be that fully 90% of patients would do better with a placebo treatment than the testosterone treatment.

4. Worst probability of harm: 77%. A study on the ability of Caplacizumab to rapidly normalize certain platelet counts [30]. The primary outcome was the number of days before the platelet count was back within a normal range. They find the following characteristics of the distribution:

   - Days until platelet count is normalized for patients with Caplacizumab - mean is 3.0, variance is 1.07^2
   - Days until platelet count is normalized for patients given placebo - mean is 4.9, variance is 5.09^2

   Since the number of days is obviously always positive, a normal distribution seemed too unrealistic. We therefore fit these statistics to a gamma distributions. Taking those distributions as truth, we can obtain that the worst probability of harm is 77%.

5. Worst probability of harm: 84%. A study on the effects of metformin on maternal weight gain [34]. We look at what they found for the maternal weight gain in kilograms, low weight being preferred. We found that

   - Weight gain for patients given metformin - $\mathcal{N}(4.6, 4.37^2)$
   - Weight gain for patients given placebo - $\mathcal{N}\left(6.3, 4.67^2\right)$

   Taking these distributions as truth, the worst probability of harm may be calculated as 84.9%

6. Worst probability of harm: 4% - a different story. On the sixth study we looked at, we finally came upon a substantial, undeniable effect [31]. This study looks at how a drug called Andexanet reverses the anticoagulant activity of another drug, called apixaban. That is, Andexanet has anti-anti-coagulant properties. These properties

were measured by the percent change of a chromogenic assay of certain enzymatic activity. They find roughly these distributions:

- Change given Andexanet - $\mathcal{N}\left(94\%, 9.8^2\right)$
- Change given placebo - $\mathcal{N}\left(21\%, 27^2\right)$

Taking these distributions as truth, we find that the worst probability of harm is only 4%. That is, even in the worst case scenario, we have that the situation for at least 96% of people are actually improved instead of worsened by the application of Andexanet. This worst probability of harm is so low because the difference between the means is very significant relative to the variances.

There are two significant limitations to the picture we have presented in the examples above. We will now turn the the remedy for both of these limitations.

The first issue pertains to our definition of "harm." Let us imagine that we have women who seek to experience less maternal weight gain. Let us further imagine that there exists a patient who would gain 5.001 kilograms under a treatment and 4.9999 kilograms under a placebo. It is ridiculous to say that applying this treatment to this patient actually caused harm (except, perhaps, insofar as someone had to pay for the treatment). Instead, is important for medical practitioners to define what sort of differences are actually clinically significant. This suggests a new definition.

Definition 12. Let $\xi(y)$ denote some monotone function on the space of patient outcomes, such that $x < \xi(y)$ if and only if the difference between x and y is clinically significant. For example, we might have $\xi(y) = y - \epsilon$, where ϵ was some threshold. The **worst probability of clinically significant harm** for (F_X, F_Y) is then defined as $\sup_Q Q(X < \xi(Y))$ among all Q with the corresponding marginal cumulative distribution functions. From a mathematical point of view, the analysis of this object is essentially the same as the analysis for the worst probability of harm. From a practical point of view, however, the worst probability of clinically significant harm may be considerably

more important than the worst probability of harm.

The second issue pertains to estimation. In the examples above, we chose parametric families somewhat arbitrarily, estimated parameters, and used those parameters to calculate the worst probability of harm. This chain of assumptions and estimates certainly could introduce significant errors. For this reason, the calculations from all of the examples above should in no way be considered for scientific or medical purposes. In actual practice, we must either use nonparametric methods or use model families that we have good reason to trust. We must also calculate some measure of uncertainty, for example by looking at confidence intervals. It is to these issues we now turn our attention.

4.5. Statistical considerations

Perhaps the simplest approach to estimation would be to take an entirely nonparametric approach. For example, we can readily apply the Dvoretesky-Kiefer-Wolfowitz (DKW) inequality to obtain a valid confidence interval on the worst probability of harm - without making any assumptions whatsoever on the form of F_X, F_Y.

Theorem. *Let $X(1) \cdots X(n_Y) \sim F_X$ and $Y(1) \cdots Y(n_Y) \sim F_Y$. Let $\hat{F}_X$ and $\hat{F}_Y$ denote the empirical cumulative distribution functions (e.g. $\hat{F}_X(x) \triangleq \frac{1}{n_X} \# \{i : X(i) \leq x\}$). Pick any ϵ. Let $\tilde{\lambda}$ denote the worst probability of ineffectiveness for (F_X, F_Y) and let $\hat{\tilde{\lambda}}$ denote the worst probability of ineffectiveness for $\left(\hat{F}_X, \hat{F}_Y\right)$. Then*

$$\mathbb{P}\left(\left|\hat{\tilde{\lambda}} - \tilde{\lambda}\right| > \epsilon\right) < 2e^{-n_X \epsilon^2/2} + 2e^{-n_Y \epsilon^2/2}$$
$$\mathbb{P}\left(\left|\hat{\lambda} - \lambda\right| > \epsilon\right) < 2e^{-n_X \epsilon^2/2} + 2e^{-n_Y \epsilon^2/2}$$

For example, in the special case that $n_X = n_Y$, we obtain that our estimates $\pm\sqrt{\frac{1}{n} \log \frac{4}{\alpha}}$ yield confidence intervals with coverage of at least $(1 - \alpha)$.

Proof. The DKW inequality, together with a union bound, tells us that we may assume

that

$$\sup \left| \hat{F}_X(x) - F_X(x) \right| \leq \epsilon/2$$
$$\sup \left| \hat{F}_Y(y) - F_Y(y) \right| \leq \epsilon/2$$

with at least probability $1 - \left(2e^{-n_X \epsilon^2/2} + 2e^{-n_Y \epsilon^2/2}\right)$. Taking these two inequalities as fact, we see immediately that $\left|\hat{\lambda} - \lambda\right| < \epsilon$ and $\left|\hat{\tilde{\lambda}} - \tilde{\lambda}\right| < \epsilon$. After all, since $F_X(k) \leq \hat{F}_X(k)+\epsilon$, it follows that $\lim_{\ell \uparrow k} F_X(\ell) \leq \lim_{\ell \uparrow k} \hat{F}_X(\ell)+\epsilon$. The same goes for the inequality in the other direction. From there, the computations are completely straightforward. $\square$

The theorem above provides a straightforward way to get rigorous statistical bounds on the worst probability of harm/ineffectiveness for any pair of distributions, without making any assumptions on the parametric form of ones distribution. However, the coverage of a DKW-based interval generally seems to be too broad.

For example, let us consider the case that $X \sim \mathcal{N}(1, 2^2)$ and $Y \sim \mathcal{N}(0, 1^2)$. The worst probability of harm for (X, Y) is about 65.486%. If we are given n random samples of X, n random samples of Y, and use the theorem above to produce a confidence interval with 95% nominal coverage, then how often does the true probability of harm fall inside the interval? That is, what is the true coverage of our interval? For each $n \in \{5, 55, 105 \cdots 955\}$, we look at the answer to that question when n is the number of samples. The results are to be found in Figure 4.1. We estimate the true coverage by using 5000 randomly generated datasets for each n.

It is clear that the interval is much more conservative than it needs to be. There are at least three reasons why this happens:

1. The DKW inequality constructs an envelope around the empirical cumulative distribution with constant width for all of $\mathbb{R}$. However, in the computation of $\lambda, \tilde{\lambda}$, it may be that we do not significantly care about fluctuations in F_X, F_Y on all of $\mathbb{R}$: in fact, we only care about estimating F_X, F_Y close to the point where $F_Y - F_X$

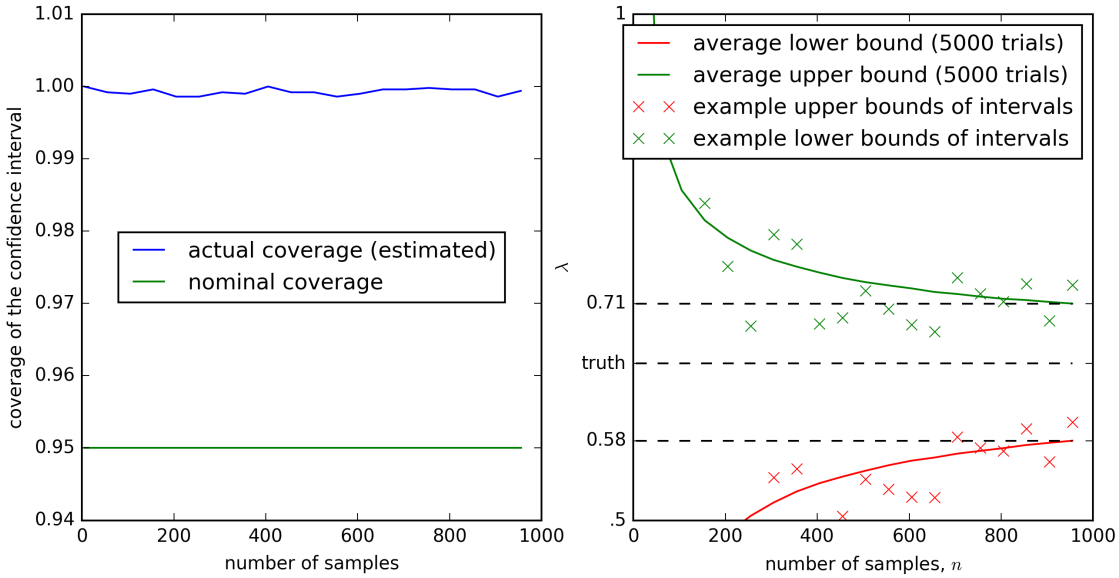


Figure 4.1.: The DKW-based confidence interval is almost comically conservative, but it is at least always valid. Here we assume $X \sim \mathcal{N}(1, 2^2)$ and $Y \sim \mathcal{N}(0, 1^2)$. In this figure we look at how the confidence interval changes as the number of samples, n, increases. We focus on $n \in \{5, 55, 105 \cdots 955\}$. On the left we compare the nominal coverage to an estimate of the true coverage. On the right, we try to show how the confidence intervals converge on the true parameter value. The solid lines indicate the average upper and lower bounds of confidence intervals, averaged over 5000 trials for each value of n. Each "x" marker indicates a bound for a confidence interval yielded from just one of those trials. We show 50 examples in all. The dotted lines at .71 and .58 indicate the average upper and lower bounds of a confidence interval with $n = 1000$.

reaches its peak. The natural solution to this problem would be to use alternative cumulative distribution estimators which place a varying-width envelope around the estimates of F_X, F_Y; this would allow us to be more confident where we need to be confident. This naturally raises the question of how best to select the right envelope for our problem. One direction would be to first estimate a rough sketch of F_X, F_Y with a subset of the data, use that sketch to determine where accuracy is needed, and then use the rest of the data to construct confidence envelopes which are carefully tuned so that our intervals around $\lambda, \tilde{\lambda}$ obtain give good coverage.

2. Our task is to estimate the difference between F_X and F_Y, but the DKW inequality gives an envelope on F_X and an envelope on F_Y. Applying a union bound, this certainly gives us a way to put an envelope on $(F_Y - F_X)$ - but it is unlikely to be the best way. This problem appears even in the simple case that the outcomes are binary, i.e. $X, Y \in \{0, 1\}$. In this case we can use samples to produce two estimates, $\hat{p}_X \approx \mathbb{P}(X = 1)$ and $\hat{p}_Y \approx \mathbb{P}(Y = 1)$. If we put a confidence interval on p_X and a confidence interval on p_Y, a union bound can then give us a confidence interval on $p_X - p_Y$ - but it is worse than a confidence interval taken directly on $p_X - p_Y$.

3. The DKW-inequality does not immediately lend itself to one-sided confidence intervals in both directions. Medical professionals may be only interested in showing that the worst probability of harm is low, but a naive use of the inequality does not allow them to leverage this to obtain increased statistical power for what they are interested in.

We here present a very simple scheme that begins to address these three concerns:

1. Use 20% of your data to estimate $k = \arg\max(F_Y - F_X)$. This identifies a point where we need to accurately model F_X, F_Y.

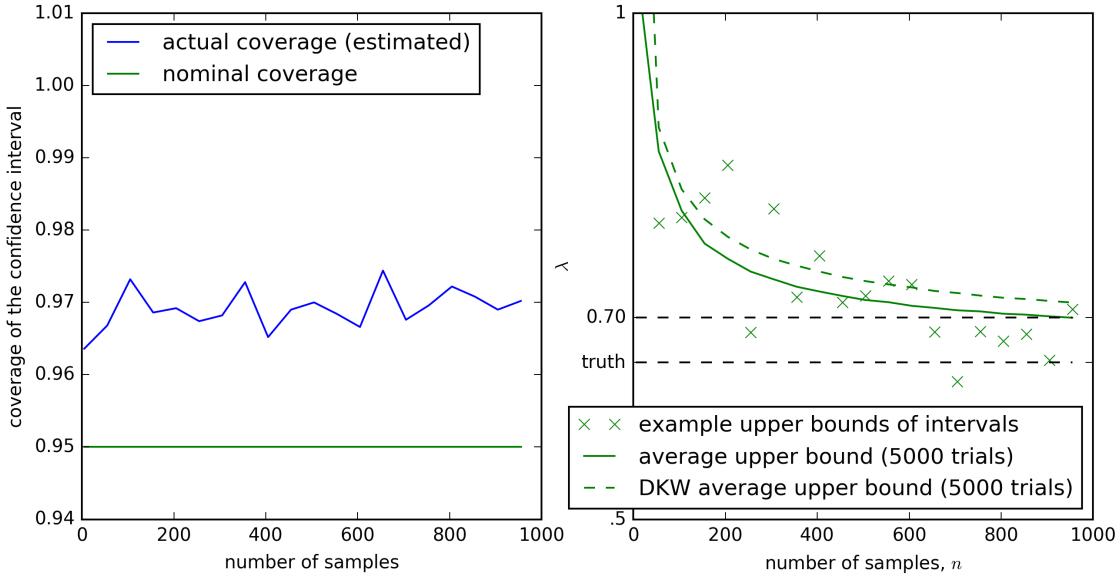


Figure 4.2.: A confidence interval based on simple difference of proportions is rather less conservative, and also allows us to take the one-sided intervals that we may be most interested in. As in the previous figure, we look at how the confidence intervals respond to changes in the number of samples, $n \in \{5, 55, \cdots 955\}$. The left side compares nominal and actual (estimated) coverage. The right figure shows typical upper bounds that you would obtain from this method given various amounts of data. The solid line gives the average upper bound over 5000 trials for each n, and each "x" is an example of an upper bound yielded from one of those 5000 trials. The dotted line gives the average DKW-based upper bound for 5000 trials, for comparison.

2. Use the rest of your data to obtain a one-sided 95% confidence interval of the form $\lim_{\ell \uparrow k} F_X(\ell) - F_Y(k) \leq \beta$. This can be much more effective than combining two confidence intervals, one for $P(X < k) = \lim_{\ell \uparrow k} F_X(\ell)$ and one for $F_Y(k)$.

3. Recall that for *any* value of k, we have $\lambda \leq 1 - F_Y(k) - \lim_{\ell \uparrow k} F_X(\ell)$. Thus, as long as we are only looking for a one-sided interval, $\lambda \leq 1 - \beta$ does the job.

The results of this scheme are displayed in Figure 4.2, again using 5000 randomly generated datasets for each value of the number of samples, n. As the figure suggests, this scheme is better than the DKW-based scheme, but definitely still conservative. We at-

tribute this to the fact that $\lambda < 1 - F_Y(k) - \lim_{\ell \uparrow k} F_X(\ell)$ for most values of k. This strict inequality makes our method more conservative than it needs to be. It is very likely that simple schemes such as this one are strictly suboptimal, and it would be interesting to try to understand how much they could be improved. For example, perhaps it would be possible to obtain clear results in certain well-constrained settings - perhaps if F_X, F_Y are known to be Gaussian. Even in this well-constrained case, it is far from obvious how one should construct an optimal confidence interval for λ. Four separate simultaneous confidence intervals for the means and variances of the Gaussians would certainly give you a way to compute bounds on λ, but it is very unlikely to be the best way. More generally, is it typically the case that knowledge of the parametric form of F_X, F_Y would enable us to significantly outperform non-parametric methods? If so, is there any general method by which such knowledge could be usefully applied?

Bibliography

[1] Christophe Avenel, Pierre Fortin, and Dominique Béréziat. Parallel birth and death process for cell nuclei extraction in histopathology images. In *2013 42nd International Conference on Parallel Processing*, pages 429–438. IEEE, 2013.

[2] S Axler, P Bourdon, and W. Ramey. *Harmonic Function Theory*. Springer-Verlag, 2001.

[3] A Azzam. Smoothness properties of solutions of mixed boundary value problems for elliptic equations in sectionally smooth n-dimensional domains. In *Annales Polonici Mathematici*, volume 1, pages 81–93, 1981.

[4] Andrew D Barbour and Timothy C Brown. Stein's method and point process approximation. *Stochastic Processes and their Applications*, 43(1):9–31, 1992.

[5] Andrew D Barbour and Marianne Månsson. Compound poisson process approximation. *Annals of probability*, pages 1492–1537, 2002.

[6] Richard F Bass and Pei Hsu. Some potential theory for reflecting brownian motion in holder and lipschitz domains. *The Annals of Probability*, pages 486–508, 1991.

[7] Krzysztof Burdzy, Zhen-Qing Chen, and Donald E Marshall. Traps for reflected brownian motion. *Mathematische Zeitschrift*, 252(1):103–132, 2006.

[8] Louis HY Chen and Aihua Xia. Stein's method, palm theory and poisson process approximation. *Annals of probability*, pages 2545–2569, 2004.

[9] John D Chodera, William C Swope, Jed W Pitera, and Ken A Dill. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Modeling & Simulation*, 5(4):1214–1226, 2006.

[10] Xavier Descombes, Robert Minlos, and Elena Zhizhina. Object extraction using a stochastic birth-and-death dynamics in continuum. *Journal of Mathematical Imaging and Vision*, 33(3):347–359, 2009.

[11] Yanqin Fan and Sang Soo Park. Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(03):931–951, 2010.

[12] PJ Fitzsimmons and Jim Pitman. Kac's moment formula and the Feynman–Kac formula for additive functionals of a Markov process. *Stochastic processes and their applications*, 79(1):117–134, 1999.

[13] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[14] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[15] Charles J Geyer and Jesper Møller. Simulation procedures and likelihood inference for spatial point processes. *Scandinavian journal of statistics*, pages 359–373, 1994.

[16] Zoher Ghogawala, James Dziura, William E Butler, Feng Dai, Norma Terrin, Subu N Magge, Jean-Valery CE Coumans, J Fred Harrington, Sepideh Amin-Hanjani, J Sanford Schwartz, et al. Laminectomy plus fusion versus laminectomy alone for lumbar spondylolisthesis. *New England Journal of Medicine*, 374(15):1424–1434, 2016.

[17] Stephanie C Hicks, Mingxiang Teng, and Rafael A Irizarry. On the widespread and critical impact of systematic bias and batch effects in single-cell rna-seq data. *bioRxiv*, page 025528, 2015.

[18] Howard N Hodis, Wendy J Mack, Victor W Henderson, Donna Shoupe, Matthew J Budoff, Juliana Hwang-Levine, Yanjie Li, Mei Feng, Laurie Dustin, Naoko Kono, et al. Vascular effects of early versus late postmenopausal treatment with estradiol. *New England Journal of Medicine*, 374(13):1221–1231, 2016.

[19] Guido W Imbens and Donald B Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, pages 305–327, 1997.

[20] John PA Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005.

[21] Mourad EH Ismail. Complete monotonicity of modified Bessel functions. *Proceedings of the American Mathematical Society*, 108(2):353–361, 1990.

[22] Weining Kang and Kavita Ramanan. On the submartingale problem for reflected diffusions in domains with piecewise smooth boundaries. *arXiv preprint arXiv:1412.0729*, 2014.

[23] Gina Kolata. We're so confused: The problems with food and exercise studies. *New York Times*, August 2016.

[24] Andrea Laforgia and Pierpaolo Natalini. Some inequalities for modified Bessel functions. *Journal of Inequalities and Applications*, 2010(1):1, 2010.

[25] Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349(6244):184–187, 2015.

[26] Pierre-Louis Lions and Alain-Sol Sznitman. Stochastic differential equations with reflecting boundary conditions. *Communications on Pure and Applied Mathematics*, 37(4):511–537, 1984.

[27] Odile Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, pages 83–122, 1975.

[28] Mathias Ortner, Xavier Descombes, and Josiane Zerubia. *A Reversible Jump MCMC Sampler for Object Detection in Image Processing*, pages 389–401. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[29] RB Paris. An inequality for the Bessel function j. *SIAM journal on mathematical analysis*, 15(1):203–205, 1984.

[30] Flora Peyvandi, Marie Scully, Johanna A Kremer Hovinga, Spero Cataland, Paul Knöbl, Haifeng Wu, Andrea Artoni, John-Paul Westwood, Magnus Mansouri Taleghani, Bernd Jilma, et al. Caplacizumab for acquired thrombotic thrombocytopenic purpura. *New England Journal of Medicine*, 374(6):511–522, 2016.

[31] Deborah M Siegal, John T Curnutte, Stuart J Connolly, Genmin Lu, Pamela B Conley, Brian L Wiens, Vandana S Mathur, Janice Castillo, Michele D Bronson, Janet M Leeds, et al. Andexanet alfa for the reversal of factor xa inhibitor activity. *New England Journal of Medicine*, 373(25):2413–2424, 2015.

[32] Peter J Snyder, Shalender Bhasin, Glenn R Cunningham, Alvin M Matsumoto, Alisa J Stephens-Shields, Jane A Cauley, Thomas M Gill, Elizabeth Barrett-Connor, Ronald S Swerdloff, Christina Wang, et al. Effects of testosterone treatment in older men. *New England Journal of Medicine*, 374(7):611–624, 2016.

[33] Gail M Sullivan and Richard Feinn. Using effect size-or why the p value is not enough. *Journal of graduate medical education*, 4(3):279–282, 2012.

[34] Argyro Syngelaki, Kypros H Nicolaides, Jyoti Balani, Steve Hyer, Ranjit Akolekar, Reena Kotecha, Alice Pastides, and Hassan Shehata. Metformin versus placebo in obese pregnant women without diabetes mellitus. *New England Journal of Medicine*, 374(5):434–443, 2016.

[35] Rasmus Waagepetersen, Yongtao Guan, Abdollah Jalilian, and Jorge Mateu. Analysis of multispecies point patterns by using multivariate log-gaussian cox processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(1):77–96, 2016.

[36] Aihua Xia and Fuxi Zhang. On the asymptotics of locally dependent point processes. *Stochastic Processes and their Applications*, 122(9):3033–3065, 2012.

A. A simple convergence lemma

Lemma 13. *Let $(\Omega, \mathcal{F}, \mu)$ denote a probability space. Let the random object X carry the distribution μ. Let $X^{(1)}, X^{(2)} \cdots$ denote a sequence of random objects with support on Ω, whose distributions are all absolutely continuous with respect to μ. Let $f^{(1)}, f^{(2)} \cdots$ denote the corresponding densities of these random variables. If*

- $\limsup_{r\to\infty} f^{(r)}(x) \leq 1$, almost everywhere w.r.t. μ
  - $f^{(r)}(x) \leq M$ for some constant $M > 1$, uniformly in r

  then the distributions of $X^{(r)}$ converge to that of X in total variation.

Proof. Fix ϵ. Applying Egoroff's theorem, find a region $S \subset \Omega$ with $\mu(S) < \frac{\epsilon}{M-1}$ and

$$\limsup_{r\to\infty} \left(\sup_{x\notin S} f^{(r)}(x) \right) \leq 1$$

Using this limit, we can choose r^* so that

$$\sup_{x\notin S} f^{(r)}(x) \leq 1 + \epsilon$$

for every $r > r^*$. Then for any measurable $A \subset \Omega$, we have that

$$\begin{aligned}
\mathbb{P}\left(X^{(r)} \in A\right) - \mathbb{P}\left(X \in A\right) &= \mathbb{E}\left[\mathbb{I}_{X \in A}\left(f_r(X) - 1\right)\right] \\
&= \mathbb{E}\left[\mathbb{I}_{X \in A \cap S}\left(f_r(X) - 1\right)\right] + \mathbb{E}\left[\mathbb{I}_{X \in A \backslash S}\left(f_r(X) - 1\right)\right] \\
&\leq \mu(S)\left(M - 1\right) + \mu(A \backslash S)\left(1 + \epsilon - 1\right) \\
&\leq 2\epsilon
\end{aligned}$$

Applying the same argument, we have that

$$\mathbb{P}\left(X \in A\right) - \mathbb{P}\left(X^{(r)} \in A\right) = \mathbb{P}\left(X^{(r)} \in \Omega \backslash A\right) - \mathbb{P}\left(X \in \Omega \backslash A\right) \leq 2\epsilon$$

Since A was arbitrary, we have that

$$\sup_A \left|\mathbb{P}\left(X^{(r)} \in A\right) - \mathbb{P}\left(X \in A\right)\right| \leq 2\epsilon$$

for all $r \geq r^*$.

Thus

$$\lim_{r \to \infty} \sup_A \left|\mathbb{P}\left(X^{(r)} \in A\right) - \mathbb{P}\left(X \in A\right)\right| = 0$$

as desired. □

B. Brownian Motion

B.1. Known results

The gist and main components of the following theorems are well-known, but we could not find any place which combined these pieces in quite the way we needed.

Definition 14. A set $A \subset \mathbb{R}^D$ is said to have C^3 **boundary** if for every $x_0 \in \partial A$, we can find $\epsilon > 0$ so that A is locally isometric to the hypograph of some function in $C^3(\mathbb{R}^{D-1})$.

Definition 15. Let $\{B_t\}$ denote a brownian motion on $\mathbb{R}^D$. The **SDE with Normal Reflection (SDE-NR)** problem is a tuple (N, μ, σ, q), where

- $N \subset \mathbb{R}^D$ denotes an open set with C^1 boundary.
  - $\mathbf{n}(x)$ denotes the outward-facing surface normal of ∂N.
  - $\sigma : \bar{N} \to \mathbb{R}^{D \times D}$ and $\mu : \bar{N} \to \mathbb{R}^D$ are diffusion and drift coefficients.
  - q is a probability measure on $\bar{N}$.

  A semimartingale $\{Y_t\}$ is said to be a solution to the SDE-NR if we can find some

$\{K_t\}$ adapted to $\{B_t\}$ so that

$$\begin{aligned} Y_0 &\sim q \\ Y_t &\in \bar{N} \\ Y_t &= Y_0 + \int_0^t \sigma(Y_s) dB_s + \int_0^t \mu(Y_s) ds + K_t \\ |K|_t &= \int_0^t \mathbb{I}_{Y_s \in \partial C} d\,|K|_s < \infty \\ K_t &= -\int_0^t \mathbf{n}(X_s) d\,|K|_s \end{aligned}$$

almost surely, where $|K|_t$ is the total variation of K on $[0,t]$. It is said to be **Normally Reflecting Brownian Motion** if $\sigma = I$ and $\mu = 0$.

Theorem 16. *(Existence/uniqueness of SDE-NR solutions) If N is bounded with a C^3 boundary and $\{\sigma_{ij}\}, \{\mu_i\}$ are Lipschitz on $\bar{N}$, then the SDE-NR problem (N, μ, σ, q) has an almost-surely unique solution, $\{Y_t\}$, satisfying the strong Markov property.*

Proof. Cf [26]. □

Theorem 17. *(Martingale property of SDE-NR solutions) Let (N, μ, σ, q) denote an SDE-NR problem satisfying the conditions of Theorem 16. Define the operator $\mathcal{L}$ by*

$$(\mathcal{L}f)(y) \triangleq \sum_{j=1}^{D} \mu_j(y) \left(\nabla_{y_j} f\right)(y) + \frac{1}{2} \sum_{i,j}^{D} \left(\sigma(y)\sigma^T(y)\right)_{i,j} \left(\nabla_{y_i} \nabla_{y_j} f\right)(y)$$

Then for any $f \in C^2(\bar{N})$ with $\langle f, \mathbf{n} \rangle = 0$ on ∂N, the process

$$Z_t = f(Y_t) - \int_0^t (\mathcal{L}f)(Y_s) ds$$

is a martingale.

Proof. Follows directly from Ito's formula applied to $f(Y_t)$. A completely rigorous argument can be found in Theorem 2 of [22]. □

Corollary 18. *Let N be connected and let $A \subset N$ denote an open set with C^3 boundary. Let $\tau = \inf\{t : Y_t \in A\}$ such that $\mathbb{E}[\tau] < \infty$. Let $f, g \in C^2$. The following functions can be characterized in terms of partial differential equations:*

Function	*Differential Equation*	*∂A Boundary*	*∂N Boundary*
$u_0(x) = \mathbb{E}^x[g(X_\tau)]$	$0 = \mathcal{L}u_0$	$g = u_0$	$0 = \langle \nabla u_0, \mathbf{n} \rangle$
$u_1(x) = \mathbb{E}^x\left[g(X_\tau)\int_0^\tau f(X_s)ds\right]$	$0 = \mathcal{L}u_1 + fu_0$	$0 = u_1$	$0 = \langle \nabla u_0, \mathbf{n} \rangle$
$u_n(x) = \mathbb{E}^x\left[g(X_\tau)\left(\int_0^\tau f(X_s)ds\right)^n\right]$	$0 = \mathcal{L}u_n + nfu_{n-1}$	$0 = u_n$	$0 = \langle \nabla u_0, \mathbf{n} \rangle$
$u_e = \mathbb{E}^x\left[g(X_\tau)e^{\int_0^\tau f(X_s)ds}\right]$	$0 = \mathcal{L}u_e + fu_e$	$g = u_e$	$0 = \langle \nabla u_0, \mathbf{n} \rangle$

Proof. First note that all the differential equations have unique solutions in $C^2(\bar{N}\backslash A)$, per [3]. The equation for u_0 follows directly from Theorem 17 with the optional stopping theorem, and likewise the equation for u_1 when $g(x) = 1$. There is some subtlety in the general case for $u_1, u_2, \cdots u_e$ - but it is easily resolved by applying $\mathbb{E}[\tau] < \infty$ and the compactness of $\bar{N}$ to show that $\mathbb{E}^y\left[|g(X_\tau)|\int_0^\tau |nf(Y_s)u_{n-1}(Y_s)ds|\right] < \infty$. The existence of this integral allows the classic results of Kac to be applied (cf. [12] for a good summary).

□

Corollary 19. *(Fokker-Planck and Kolmogorov Equations).*

- Let $f \in C^2(\bar{N})$ with vanishing derivative on ∂N; then $h(t,x) = \mathbb{E}^x[f(X_t)] \implies h_t(t,x) = (\mathcal{L}h(t,\cdot))(x)$.
  - Let $q \in C^2$ denote a continuous density on $\bar{N}$ with vanishing derivative on ∂N; then $X_t \sim p(t,x) \implies p_t(t,x) = (\mathcal{L}^* p(t,\cdot))(x)$, where $\mathcal{L}^*$ denotes the adjoint of $\mathcal{L}$.

Proof. The smoothness allows us to interchange integrals and derivatives from Theorem 17 to deduce that $(\mathcal{L}g)(x) = \lim_{r\downarrow 0}\frac{1}{r}(\mathbb{E}^x[g(X_r)] - g(x))$ for any $g \in C^2(\bar{N})$; the formula for h_t follows from an application of this to $g = h(t,\cdot)$. Similarly, interchanging integrals and derivatives yields that $\int p_t(y,t)f(y) = \int p(y,t)(\mathcal{L}f)(y)$ for any $f \in C^2$; the formula

for p_t then follows from the definition of the adjoint and the density of C^2 functions in $\mathscr{L}^2(\bar{N})$. □

Lemma 20. *(Expected hitting time is finite). Let N be connected and bounded. Let $A \subset N$ denote an open set with C^3 boundary. Let $\tau = \inf\{t : Y_t \in A\}$, and let $\sigma = I, \mu = 0$, so $\{Y_t\}$ is a reflecting brownian motion. Then $\mathbb{E}[\tau] < \infty$, so 18 may be applied.*

Proof. It is sufficient to show that the marginal density of X_t converges uniformly exponentially fast, for every initial condition $X_0 = x_0$ (cf. [7]). This is shown in [6]. □

Theorem 21. *(Harmonic functions minimize the integrated squared norm of the gradient). Let A, N denote two open bounded sets with C^3 boundary, $A \subset N$, and take $f \in C^2$. Let u denote the unique solution to the boundary problem $\Delta u = 0$, $u = f$ on ∂A, $\langle \nabla u, \mathbf{n} \rangle = 0$ on∂N. Then for any $v \in H^1(\bar{N})$ satisfying the same boundary conditions, $\int |\nabla v|^2 \geq \int |\nabla u|^2$.*

Proof. Let $w = u - v$. Then $\int |\nabla v|^2 = \int |\nabla u + \nabla w|^2 = \int |\nabla u|^2 + \int |\nabla w|^2 + 2\int \langle \nabla u, \nabla w \rangle$. Since $u \in C^2(\bar{N} \backslash A)$ and $w \in H^1(\bar{N})$, we can use integration by parts and the fact that u is harmonic to obtain $\int \langle \nabla u, \nabla w \rangle = \int_{\partial A \cup \partial N} w(x) \langle \nabla u, \mathbf{n} \rangle - \int w(x) \Delta u$ - but $\Delta u = 0$, $w(x)$ vanishes on ∂A, and $\langle \nabla u, \mathbf{n} \rangle$ vanishes on ∂N, so we obtain zero overall. Thus $\int |\nabla v|^2 = \int |\nabla u|^2 + \int |\nabla w|^2 \geq \int |\nabla u|^2$. □

B.2. Timing estimates

Recall that $\tilde{X}_t$ indicates Brownian motion on $\mathbb{R}^D$ in the annulus between a reflecting sphere of radius $r = \beta r_A$ and an absorbing sphere of radius $r = r_A$, where both spheres are centered at x_A. Our task in this section is to understand the distribution of $\tilde{\lambda} = \inf\left\{t : \tilde{R}_t = r_A\right\}$, where $\tilde{R}_t = \left|\tilde{X}_t - x_A\right|$.

We will have special interest in the asymptotics as $r_A \to 0$ for some fixed value of βr_A. Therefore, without loss of generality, we will mostly assume that $r_A \beta = 1$. When $r_A \beta \neq$

1, we can use the space/time correspondence of brownian motion - $\left(\tilde{\lambda}/k^2; \beta r_A = k, \beta = c\right) \overset{(d)}{=} \left(\tilde{\lambda}; \beta r_A = 1, \beta = c\right)$ - to make the appropriate calculations.

Our first step will be to get the Laplace transform of the density of $\tilde{\lambda}$.

Lemma 22. *Assume* $\beta r_A = 1$. *Let* $\tilde{\lambda} = \inf\left\{t : \tilde{R}_t = r_A\right\}$, $h^\xi(r) = \mathbb{E}^r\left[e^{-\frac{1}{2}\xi^2\tilde{\lambda}}\right]$ *for* $r_A \leq r \leq 1$. *Then*

$$h^\xi(r) = \left(\frac{r_A}{r}\right)^\nu \frac{I_{\nu+1}(\xi)K_\nu(r\xi) + K_{\nu+1}(\xi)I_\nu(r\xi)}{I_{\nu+1}(\xi)K_\nu(r_A\xi) + K_{\nu+1}(\xi)I_\nu(r_A\xi)}$$

where $\nu = \frac{D-2}{2}$ *and* $I_\nu \triangleq \sum_{m=0}^\infty \frac{(x/2)^{2m+\alpha}}{m!\Gamma(m+\alpha+1)}$, $K_\nu = \lim_{\alpha\to\nu} \frac{\pi(I_{-\alpha}-I_\alpha)}{2\sin(\alpha\pi)}$ *are the modified Bessel functions of first and second kind.*

Proof. Applying Ito's lemma to $\tilde{R}_t = \left|\tilde{X}_t - x_A\right|$, we deduce that $\tilde{R}_t$ is itself a Markov process, governed by $d\tilde{R}_t = dW_t + \frac{D-1}{2\tilde{R}_t}dt$, where W_t is a 1-dimensional Brownian motion. This process is sometimes called the "Bessel process."

Thus $\tilde{R}_t$ satisfies a stochastic differential equation with smooth coefficients and boundaries, so we can apply Corollary 18 to argue that □

1. h^ξ must satisfy the differential equation $\mathcal{L}h^\xi = \frac{1}{2}\xi^2 h^\xi$ - here the generator $\mathcal{L}$ is given by $(\mathcal{L}f)(r) \triangleq \frac{1}{2}f''(r) + \frac{D-1}{2r}$
   a) $h^\xi(r_A) = 1$
   b) $h^\xi_r(1) = 0$

Proof. We will now show that the formula for h^ξ stated in this theorem is the unique solution to these requirements.:

□

1. **The differential equation**: we will first show that our formula satisfies $\mathcal{L}h^\xi = \frac{1}{2}\xi^2 h^\xi$. Since $\tilde{R}_t$ is called the Bessel process, it should perhaps not be surprising

that our key tool will be something called the "modified Bessel equation," $0 = x^2y''(x) + xy'(x) - \left(x^2 + \nu^2\right) y(x)$.

Let y be any solution to the modified Bessel equation, and define $w(r) = r^{-\nu}y(\xi r)$. Then it is straightforward to calculate that $(\mathcal{L}w)(r) = \frac{1}{2}r^{-\nu-2}\left[r^2\xi^2y''(\xi r) + r\xi y'(\xi r) - \nu^2 y(\xi r)\right]$. Now we have assumed y is a solution to the modified Bessel equation, so the term in brackets can be rewritten, yielding $(\mathcal{L}w)(r) = \frac{1}{2}r^{-\nu-2}\left[(\xi r)^2 y(\xi r)\right] = \frac{1}{2}\xi^2 w$. That is, w satisfies the differential equation we're interested in. Thus for any solution to a modified Bessel equation, y, we can construct a solution to our differential equation, w.

To apply this idea to our case, recall the well-known fact that the modified Bessel functions I_ν, K_ν are both solutions to the modified Bessel equation, and so, in particular,

$$y^*(x) = r_A^\nu \frac{I_{\nu+1}(\xi)K_\nu(x) + K_{\nu+1}(\xi)I_\nu(x)}{I_{\nu+1}(\xi)K_\nu(r_A\xi) + K_{\nu+1}(\xi)I_\nu(r_A\xi)}$$

is also a solution to the modified Bessel equation. Therefore we compute that our formula indeed satisfies $\mathcal{L}h^\xi = \frac{1}{2}\xi^2 h^\xi$.

a) **The absorbing boundary condition:** we will show that our formula satisfies $h^\xi(r_A) = 1$. This is immediate; the numerators and the denominators coincide when $r = r_A$.

b) **The reflecting boundary condition:** we will show that our formula satisfies $h_r^\xi(1) = 0$. This can be directly verified using the well-known identities $I_\nu' = \frac{1}{2}(I_{\nu+1} + I_{\nu-1})$, $K_\nu' = -\frac{1}{2}(K_{\nu+1} + K_{\nu-1})$, $I_{\nu-1}(x) = I_{\nu+1}(x) + \frac{2\nu}{x}I_\nu(x)$, and $K_{\nu-1}(x) = K_{\nu+1}(x) - \frac{2\nu}{x}K_\nu(x)$.

We can use this Laplace transform to show that $\tilde{\lambda}$ is large with high probability. We will be particularly interested in the case where our initial position is given halfway in-between the inner and outer shells.

Lemma 23. *For any $r^* \geq \frac{\beta+1}{2} r_A$, we have*

$$\mathbb{P}^{r^*}\left(\tilde{\lambda} < (\beta r_A)^2 \frac{\beta^{\frac{D-2}{2}}}{D\left(D-2\right)}\right) \leq 6\beta^{\frac{2-D}{2}} e^{3(D-2)\beta^{1+\frac{D-2}{4}}}$$

Proof. Note again that it is sufficient to show that

$$\mathbb{P}^{r^*}\left(\tilde{\lambda} < \frac{\beta^{\frac{D-2}{2}}}{D\left(D-2\right)}\right) \leq 6\beta^{\frac{2-D}{2}} e^{3(D-2)\beta^{1+\frac{D-2}{4}}}$$

when $\beta r_A = 1$. After all, for any fixed β we have that $\mathbb{P}^{r^*}\left(\tilde{\lambda}/k^2 < \frac{\beta^{\frac{D-2}{2}}}{D(D-2)}; \beta r_A = k\right) = \mathbb{P}^{r^*}\left(\tilde{\lambda} < \frac{\beta^{\frac{D-2}{2}}}{D(D-2)}; \beta r_A = 1\right)$. Therefore, fix $r_A\beta = 1$. Our argument proceeds in four steps: show that $\mathbb{P}^r(\tilde{\lambda} < t) \approx \mathbb{P}^1(\tilde{\lambda} < t)$, get a bound on $\mathbb{E}^1\left[e^{\frac{1}{2}\xi^2\tilde{\lambda}}\right]$, apply a Chernoff bound, and put it all together.

□

1. **Show** $\mathbb{P}^r\left(\tilde{\lambda} < t\right) \approx \mathbb{P}^1\left(\tilde{\lambda} < t\right)$. Let $T = \inf\left\{t : \tilde{R}_t \in \{r_A, 1\}\right\} \leq \tilde{\lambda}$, the first time we hit *either* the inner or outer boundary. Using the Markov property, we have that

$$\begin{aligned}\mathbb{P}^r\left(\tilde{\lambda} < t\right) &= \mathbb{P}^r(\tilde{R}_T = r_A)\mathbb{P}^r\left(\tilde{\lambda} < t|\tilde{R}_T = r_A\right) + \mathbb{P}^r(\tilde{R}_T = 1)\mathbb{P}^r\left(\tilde{\lambda} < t|\tilde{R}_T = 1\right) \\ &\leq \mathbb{P}^r(\tilde{R}_T = r_A) + \mathbb{P}^1\left(\tilde{\lambda} < t\right)\end{aligned}$$

Now, apply Corollary 18 and solve the corresponding partial differential equation to deduce that $\mathbb{P}^r(\tilde{R}_T = r_A) = \left(r^{-2\nu} - 1\right) / \left(r_A^{-2\nu} - 1\right)$. We are interested in the case $r^* \geq (r_A + 1)/2$, yielding

$$\mathbb{P}^{r^*}\left(\tilde{R}_T = r_A\right) \leq \frac{\left(\frac{r_A+1}{2}\right)^{-2\nu} - 1}{r_A^{-2\nu} - 1} = r_A^{\nu} \frac{\left(\left(\frac{r_A+1}{2}\right)^2\right)^{-\nu} - 1}{r_A^{-\nu} - r_A^{\nu}}$$

Applying the inequalities $r_A^{\nu} < 1$ and $1 + x^2 \geq 2x$, we can simplify this to obtain

$\mathbb{P}^{r^*}\left(\tilde{R}_T = r_A\right) \leq r_A^\nu$, so

$$\mathbb{P}^r\left(\tilde{\lambda} < t\right) \leq r_A^\nu + \mathbb{P}^1\left(\tilde{\lambda} < t\right)$$

a) **Estimate $h^\xi(1)$.** Applying Lemma 22 with the well-known Wronskian identity $I_{\nu+1}(x)K_\nu(x) + K_{\nu+1}(x)I_\nu(x) = \frac{1}{x}$, we get

$$h^\xi(1) = \frac{r_A^\nu \xi^{-1}}{I_{\nu+1}(\xi)K_\nu\left(r_A\xi\right) + K_{\nu+1}\left(\xi\right)I_\nu\left(r_A\xi\right)}$$

Let's take a closer look at

$$\begin{aligned} I_{\nu+1}(\xi)K_\nu\left(r_A\xi\right) + K_{\nu+1}\left(\xi\right)I_\nu\left(r_A\xi\right) &\geq I_{\nu+1}(\xi)K_\nu\left(r_A\xi\right) \\ &= \frac{I_{\nu+1}(\xi)}{I_{\nu+1}(r_A\xi)}\frac{I_{\nu+1}(r_A\xi)}{I_\nu(r_A\xi)}I_\nu(r_A\xi)K_\nu\left(r_A\xi\right) \end{aligned}$$

Note that since $\tilde{d}_A > r_A$ we can applying a bound of [29], namely $\frac{I_\nu(x)}{I_\nu(y)} \geq \left(\frac{x}{y}\right)^\nu$ for $x \geq y$. We obtain

$$\geq r_A^{-\nu-1}\frac{I_{\nu+1}(r_A\xi)}{I_\nu(r_A\xi)}I_\nu(r_A\xi)K_\nu\left(r_A\xi\right)$$

Applying a bound of [24], $\frac{I_{\nu+1}(x)}{I_\nu(x)} \geq \frac{x}{(\nu+1)+\sqrt{(\nu+1)^2+x^2}}$, we obtain

$$\begin{aligned} &\geq r_A^{-\nu-1}\frac{r_A\xi}{(\nu+1)+\sqrt{(\nu+1)^2+\left(r_A\xi\right)^2}}I_\nu(r_A\xi)K_\nu\left(r_A\xi\right) \\ &\geq r_A^{-\nu-1}\frac{r_A\xi}{2\left(\nu+1\right)+r_A\xi}I_\nu(r_A\xi)K_\nu\left(r_A\xi\right) \end{aligned}$$

Applying $I_\nu \triangleq \sum_{m=0}^\infty \frac{(x/2)^{2m+\nu}}{m!\Gamma(m+\nu+1)} \geq \frac{(x/2)^\nu}{\Gamma(\nu+1)}$, we get

$$\geq r_A^{-\nu-1}\frac{r_A\xi\left(r_A\xi/2\right)^\nu}{\Gamma\left(\nu+1\right)\left(2\left(\nu+1\right)+r_A\xi\right)}K_\nu\left(r_A\xi\right)$$

Since $\nu \geq \frac{1}{2}$, we may apply Theorem 1 of [21], $K_\nu(x) \geq 2^{\nu-1}\Gamma(\nu)e^{-x}x^{-\nu}$ to get

$$\begin{aligned} &\geq r_A^{-\nu-1}\frac{r_A\xi\left(r_A\xi/2\right)^\nu}{\Gamma\left(\nu+1\right)\left(2\left(\nu+1\right)+r_A\xi\right)}2^{\nu-1}\Gamma(\nu)e^{-r_A\xi}\left(r_A\xi\right)^{-\nu} \\ &= \frac{r_A^{-\nu}\xi e^{-r_A\xi}}{2\nu\left(2\left(\nu+1\right)+r_A\xi\right)} \end{aligned}$$

So, in summary,

$$\begin{aligned} h^\xi &= \frac{r_A^\nu\xi^{-1}}{I_{\nu+1}(\xi)K_\nu\left(r_A\xi\right)+K_{\nu+1}\left(\xi\right)I_\nu\left(r_A\xi\right)} \\ &\leq \frac{r_A^\nu\xi^{-1}2\nu\left(2\left(\nu+1\right)+r_A\xi\right)}{r_A^{-\nu}\xi e^{-r_A\xi}} \\ &= r_A^{2\nu}\xi^{-2}2\nu\left(2\nu+2+r_A\xi\right)e^{r_A\xi} \end{aligned}$$

b) **Chernoff bound.** We will now apply a Chernoff bound:

$$\begin{aligned} \mathbb{P}^1\left(\tilde{\lambda}<\ell\right) &\leq e^{\frac{1}{2}\xi^2\ell}h^\xi(\tilde{d}_A) \\ &\leq e^{\frac{1}{2}\xi^2\ell}r_A^{2\nu}\xi^{-2}2\nu\left(2\nu+2+r_A\xi\right)e^{r_A\xi} \end{aligned}$$

Using Corollary 18 to calculate $\mathbb{E}\left[\tilde{\lambda}\right]$, it can be estimated that $\tilde{\lambda}$ grows roughly like $r_A^{-2\nu}/4\nu\left(\nu+1\right)$. We'll therefore take ℓ to be a modest $r_A^{-\nu}/4\nu\left(\nu+1\right)$. We need to cancel the $e^{\frac{1}{2}\xi\ell}$ term, so we take $\xi=\sqrt{8\nu(\nu+1)}r_A^{\nu/2}=\sqrt{2/\ell}$. The result is

$$\begin{aligned} \mathbb{P}^1\left(\tilde{\lambda}<\frac{r_A^{-\nu}}{4\nu\left(\nu+1\right)}\right) &\leq e^1r_A^{2\nu}\frac{r_A^{-\nu}}{8\nu\left(\nu+1\right)}2\nu\left(2\nu+2+\sqrt{8\nu(\nu+1)}r_A^{1+\nu/2}\right)e^{\sqrt{8\nu(\nu+1)}r_A^{1+\nu/2}} \\ &= r_A^\nu\left(\frac{1}{2}+\frac{\nu r_A^{1+\nu/2}}{\sqrt{2\nu\left(\nu+1\right)}}\right)e^{1+\sqrt{8\nu(\nu+1)}r_A^{1+\nu/2}} \end{aligned}$$

Since $r_A < 1$ and $\frac{x}{\sqrt{2x(x+1)}} < 1$ and $e^1 < 3$, we get

$$\leq \frac{3}{2} r_A^{\nu}(3)\, e^{\sqrt{8\nu(\nu+1)} r_A^{1+\nu/2}}$$

Writing $8\nu(\nu+1) = 2(2\nu)^2 + 4(2\nu)$ and noting that since $\nu \geq \frac{1}{2}$ we have $(2\nu) \leq (2\nu)^2$ and so $8\nu(\nu+1) \leq 24\nu^2 \leq (5\nu)^2$. Also $\frac{9}{2} \leq 5$. All told:

$$\leq 5 r_A^{\nu} e^{5\nu r_A^{1+\nu/2}}$$

c) **Putting it together.** Combining (1) and (3), we have that

$$\begin{aligned} \mathbb{P}^{r^*}\left(\tilde{\lambda} < \frac{r_A^{-\nu}}{4\nu(\nu+1)}\right) &\leq r_A^{\nu} + 5 r_A^{\nu} e^{5\nu r_A^{1+\nu/2}} \\ &\leq 6 r_A^{\nu} e^{5\nu r_A^{1+\nu/2}} \end{aligned}$$

Converting to the language of D and β, we get

$$\begin{aligned} \mathbb{P}^{r^*}\left(\tilde{\lambda} < \frac{\beta^{\frac{D-2}{2}}}{D(D-2)}\right) &\leq 6 r_A^{\frac{D-2}{2}} e^{5(\frac{D-2}{2}) r_A^{1+\frac{D-2}{4}}} \\ &\leq 6\beta^{\frac{2-D}{2}} e^{3(D-2)\beta^{-1+\frac{2-D}{4}}} \end{aligned}$$