Exponential Suppression of Transient-Basin Contributions in Trajectory-Weighted Markov Chain Measures

Bryan Ehrlich

March 16, 2026

Abstract

We prove an exponential suppression bound for occupation-time ratios in finite-state reversible Markov chains with Metropolis dynamics. Given two metastable basins with Freidlin–Wentzell communication heights $\Delta_s > \Delta_b$, and a bounded positive trajectory functional ρ, the ratio of ρ-weighted occupation in the shallow basin to that in the deep basin satisfies $\mu_{\mathrm{BB}}/\mu_{\mathrm{stable}} \leq C \exp(-(\Delta_s - \Delta_b - \alpha)/\varepsilon)$, where α parameterizes the observation horizon and C is an $O(1)$ prefactor. The proof assembles seven constituent lemmas drawing on Freidlin–Wentzell cycle decomposition, BEGK capacity asymptotics, quasi-stationary distribution convergence, and Donsker–Varadhan large deviations, with explicit error scaling $\gamma = \min(\alpha/2, \Delta_s - \alpha)$ at every composition step. A three-state chain validation confirms the bound across 9 parameter combinations, with exact prefactor agreement at $p = 0.5$. The result is motivated by the Boltzmann brain problem in cosmology: within this idealized finite-state framework, the bound shows that any trajectory-weighted measure assigning positive weight to a deep metastable basin will exponentially suppress contributions from shallower transient basins, regardless of the specific form of ρ.

1 Introduction

1.1 The Boltzmann brain problem

Statistical mechanics predicts that thermal fluctuations in a system at equilibrium will eventually produce any configuration, including one that instantaneously mimics a conscious observer – a "Boltzmann brain." In cosmological settings with eternal de Sitter phases, the expected number of Boltzmann brains can vastly exceed the number of evolved observers, leading to the paradox that a randomly sampled "typical" observer should be a Boltzmann brain rather than a product of Darwinian evolution [1, 8, 4, 9].

The standard formulation counts observers by their instantaneous existence: if a configuration at time t qualifies as an observer, it receives equal weight regardless of its causal history, dynamical stability, or capacity for sustained information processing. This counting prescription lies at the heart of the problem – it treats a thermal fluctuation persisting for 10^{-43} seconds identically to a biological brain with decades of coherent experience.

1.2 Existing approaches to BB suppression

Several strategies have been proposed to resolve the Boltzmann brain problem:

1. **Geometric cutoffs.** The causal diamond measure [14] and the scale-factor cutoff measure [10] regulate the divergent spacetime volume by restricting the counting region, suppressing late-time fluctuation observers relative to early-time evolved observers.

2. **Cognitive instability.** Carroll [11] argues that a cosmological theory predicting Boltzmann brain domination is "cognitively unstable": an observer who accepts the theory should conclude they are probably a Boltzmann brain, thereby undermining the empirical evidence for the theory.
3. **Quantum dissolution.** Boddy, Carroll, and Pollack [12] argue that in a truly quantum treatment of de Sitter space, thermal fluctuations do not produce Boltzmann brains because the quantum state is stationary and unitarily evolving.
4. **Xerographic typicality.** Hartle and Srednicki [13] question whether "typicality" reasoning (we should expect to be typical observers) is the correct inference framework, proposing instead that we condition on our specific data.
5. **Information-theoretic approaches.** Tegmark [15] proposes "perceptronium" – the most general substance capable of subjective experience – characterized by information integration properties.

The present work takes a different approach: rather than modifying the cosmological measure or the inference framework, we study a specific trajectory-weighted functional (the experiential density ρ, defined below) and prove that *any* bounded positive functional with a positive lower bound on the dominant basin's quasi-stationary distribution produces exponential suppression of transient-basin contributions. The setting is a finite-state Markov chain – a toy model, not a cosmological calculation – but the mathematical mechanism (metastability gap $\Rightarrow$ exponential suppression) is cleanly isolated.

1.3 The experiential measure framework

The experiential measure framework replaces instantaneous observer counting with a trajectory-weighted functional that privileges sustained, self-modeling systems. The key quantity is the *experiential density*:

$$\rho(p) = I(B; M)\left(1 - \frac{I(B; M)}{H(B)}\right), \tag{1}$$

where B and M are the brain and model subsystems of a composite state space $\Omega = B \times M$, $I(B; M)$ is their mutual information, $H(B)$ is the entropy of the brain subsystem, and all information-theoretic quantities are computed in nats (ln). The factor $I(B; M)$ rewards states where the model M carries information about the brain B (self-modeling), while the factor $(1 - I(B; M)/H(B))$ penalizes the degenerate case $I = H(B)$ where the brain is a deterministic function of the model (no genuine experience).

The functional ρ belongs to the family of statistical complexity measures introduced by Lopez-Ruiz, Mancini, and Calbet [17], which quantify the interplay between order and disorder in a probability distribution. It is structurally related to Tononi's integrated information Φ [16], which also measures how much a system's parts inform each other beyond what is available from the parts individually. We emphasize that the theorem proved below does *not* depend on the specific form of ρ: the exponential suppression holds for any bounded positive functional f satisfying $f(\nu_{\mathrm{QSD}}^{B_{\mathrm{stable}}}) \geq c > 0$ and $f(p) \leq f_{\max} < \infty$. The experiential density is one natural choice from the LMC family, but it is not unique.

The experiential measure over a trajectory $\{p_t\}_{t\in[0,T]}$ is the time integral

$$\mu([0, T]) = \int_0^T \rho(p_t)\, dt \qquad \text{(units: nat-seconds).} \tag{2}$$

This functional automatically assigns more weight to metastable states – those that persist long enough to accumulate experiential measure – and less weight to transient fluctuations that are reabsorbed before ρ can integrate to a significant value.

1.4 Theorem A and the contribution of this paper

The central claim, Theorem A, is that within this framework the Boltzmann brain experiential measure is *exponentially suppressed* relative to the stable-observer measure:

$$\frac{\mu_{\mathrm{BB}}}{\mu_{\mathrm{stable}}} \le C \cdot \exp\left(-\frac{\Delta_s - \Delta_b - \alpha}{\varepsilon}\right),$$

where Δ_s and Δ_b are the Freidlin–Wentzell communication heights of the stable and Boltzmann brain basins, α is a parameter controlling the observation horizon, ε is the noise strength (temperature), and C is a finite constant independent of ε.

This paper provides the first complete, self-contained assembly of the proof. The argument composes seven lemmas, each grounded in established results from metastability theory:

(i) **Basin Partition** (Freidlin–Wentzell [7]): exact cycle decomposition of the energy landscape.

(ii) **Residence Time** (BEGK [2]): sharp asymptotics for mean exit times via capacities.

(iii) **QSD Convergence** (Champagnat–Villemonais [5]): exponential convergence to the quasi-stationary distribution within a metastable basin.

(iv) **Stable Measure Lower Bound**: integration of ρ along the QSD trajectory.

(v) **BB Occupation Upper Bound**: excursion analysis bounding time spent in the Boltzmann brain basin.

(vi) **Concentration**: conversion of expected-value bounds to high-probability statements via BEGK exponential laws.

(vii) **Ratio Assembly**: division of upper and lower bounds with explicit error composition.

Every composition step carries an explicit error term, and we track the error rates through the full dependency chain, showing that the composite error decays at rate $\gamma = \min(\alpha/2,\, \Delta_s - \alpha) > 0$. For the typical regime where α is well away from both endpoints of the interval $(0, \Delta_s - \Delta_b)$, $\gamma = \alpha/2$. We validate the bound on a three-state chain where exact analytical formulas are available, confirming agreement across 9 parameter combinations.

2 Setup and Notation

Let $(\Omega, Q_\varepsilon)_{\varepsilon>0}$ be a family of finite irreducible continuous-time Markov chains with state space $\Omega = B \times M$ ($|B|, |M| < \infty$) and generators of Metropolis type:

$$Q_\varepsilon(x, y) = r(x, y) \exp\bigl(-[E(y) - E(x)]^+/\varepsilon\bigr), \qquad x \neq y, \tag{3}$$

where $r(x, y) > 0$ for all pairs (x, y) with $x \neq y$ (irreducibility), $r(x, y) = r(y, x)$ (reversibility with respect to the Gibbs measure $\pi_\varepsilon(x) \propto \exp(-E(x)/\varepsilon)$), and $[a]^+ = \max(a, 0)$. Row sums are zero: $Q_\varepsilon(x, x) = -\sum_{y\neq x} Q_\varepsilon(x, y)$. We adopt the convention of row-vector left multiplication: $dp/dt = pQ_\varepsilon$.

The experiential density functional is $\rho(p) = I(B;M) \cdot (1 - I(B;M)/H(B))$, with entropy and mutual information computed in nats. The trajectory functional is $\mu([0,T]) = \int_0^T \rho(p_t)\, dt$.

Standing assumptions:

(A1) The Freidlin–Wentzell cycle hierarchy decomposes Ω into metastable sets including B_{stable} with communication height Δ_s and B_{BB} with communication height Δ_b, satisfying $\Delta_s > \Delta_b$.

(A2) **Density bounds:** $\rho(\nu_{\text{QSD}}^{B_{\text{stable}}}) \geq c > 0$ (the quasi-stationary distribution on B_{stable} has nontrivial self-modeling), and $\rho(p) \leq \rho_{\max} = H(B)/4$ for all distributions p on Ω.

(A3) The observation horizon is $T_\varepsilon = \exp((\Delta_s - \alpha)/\varepsilon)$ for a fixed $\alpha \in (0, \Delta_s - \Delta_b)$.

(A4) The initial distribution p_0 is concentrated on B_{stable}.

3 Statement of Theorem A

Theorem 1 (Boltzmann Brain Negligibility). *Under assumptions (A1)–(A4), define the experiential measures over* $[0, T_\varepsilon]$*:*

$$\mu_{\text{stable}} = \int_0^{\min(\tau_{B^c_{\text{stable}}}, T_\varepsilon)} \rho(p_t)\, dt, \tag{4}$$

$$\mu_{B_{\text{BB}}} = \int_0^{T_\varepsilon} \rho(p_t) \cdot \mathbf{1}_{X_t \in B_{\text{BB}}}\, dt. \tag{5}$$

Then, as $\varepsilon \to 0$*, with probability at least* $1 - O(\exp(-\alpha/(2\varepsilon)))$*:*

$$\frac{\mu_{B_{\text{BB}}}}{\mu_{\text{stable}}} \leq C \cdot \exp\left(-\frac{\Delta_s - \Delta_b - \alpha}{\varepsilon}\right) \cdot \big(1 + \delta(\varepsilon)\big), \tag{6}$$

where

$$C = \frac{\rho_{\max}}{c} \cdot \frac{K_b}{K_s^2} \tag{7}$$

is a finite constant independent of ε *(*K_s, K_b *are the BEGK capacity prefactors for* B_{stable} *and* B_{BB}*), and*

$$|\delta(\varepsilon)| \leq C' \exp(-\gamma/\varepsilon), \qquad \gamma = \min(\alpha/2,\, \Delta_s - \alpha) > 0. \tag{8}$$

In particular, $\mu_{B_{\text{BB}}}/\mu_{\text{stable}} \to 0$ *exponentially fast as* $\varepsilon \to 0$*.*

Remark 2 (Constraint on α). *The parameter* α *satisfies* $0 < \alpha < \Delta_s - \Delta_b$*. At* $\alpha = 0$*, the observation horizon equals the mean exit time from* B_{stable} *and the exit probability is* $O(1)$*. At* $\alpha = \Delta_s - \Delta_b$*, the exponent vanishes. For* $0 < \alpha < \Delta_s - \Delta_b$*, the bound gives genuine exponential suppression, with* α *controlling the trade-off between observation time and suppression strength.*

Remark 3 (Routing probabilities and the prefactor). *The prefactor* $C = (\rho_{\max}/c)(K_b/K_s^2)$ *uses the bound* $q \leq 1$ *on the routing probability* $q = \mathbb{P}(\textit{reach } B_{\text{BB}} \mid \textit{exit } B_{\text{stable}})$*. For systems where* q *is known, the prefactor sharpens to* $C_q = q \cdot (\rho_{\max}/c)(K_b/K_s^2)$*. For the three-state chain with branching parameter* p *at the saddle,* $q = 1 - p$*, and* $C = 0.25$ *is tight when* $p = 0.5$*.*

Remark 4 (Error rate γ). *The error rate* $\gamma = \min(\alpha/2,\, \Delta_s - \alpha)$ *arises from the minimum over all error rates in the composition chain (see* (22)*). When* $\alpha \leq 2\Delta_s/3$*, the binding constraint is* $\alpha/2$*; when* $\alpha > 2\Delta_s/3$*, it is* $\Delta_s - \alpha$*. For the typical regime where* α *is well away from both* 0 *and* $\Delta_s - \Delta_b$*,* $\gamma = \alpha/2$*. Choosing* $\eta = \exp(-\beta/\varepsilon)$ *for any* $\beta \in (0, \alpha)$ *gives* $\gamma = \min(\beta, \Delta_s - \alpha)$*, but the constraint from* $\Delta_s - \alpha$ *cannot be removed. The rate* γ *does not affect the leading-order suppression* $\exp(-(\Delta_s - \Delta_b - \alpha)/\varepsilon)$*.*

4 Constituent Lemmas

Lemma 5 (Basin Partition)**.** (Freidlin–Wentzell [7], Chapter 6, Theorem 6.3.1.) *For the family* $(Q_\varepsilon)_{\varepsilon>0}$ *on the finite state space* Ω *with energy function* E*, the Freidlin–Wentzell cycle hierarchy defines a partition of* Ω *into metastable sets* $\{B_1, B_2, \dots, B_k\}$ *at each hierarchy level, with communication heights*

$$\Delta_i = V(B_i, B_i^c) - \min_{x \in B_i} E(x),$$

where $V(A, A^c) = \min_{x\in A, y\notin A} V(x, y)$ *and*

$$V(x, y) = \min_{\gamma:x\to y} \max_{i=0,\dots,|\gamma|-1} E(\gamma_i) - E(x)$$

is the minimal communication height (minimum over all paths of the maximal energy barrier).

Identify B_{stable} *and* B_{BB} *as two cycles with* $\Delta_s > \Delta_b > 0$*.*

Error term: *None. This is an exact graph-theoretic construction, independent of* ε*.*

Output: *Partition* $(B_{\text{stable}}, B_{\text{BB}}, \Omega \setminus (B_{\text{stable}} \cup B_{\text{BB}}))$ *with communication heights* $\Delta_s, \Delta_b \in \mathbb{R}_{>0}$*.*

Remark 6. *For finite state spaces with Metropolis rates, the communication height coincides with the energy barrier:* $V(x, y) = \max_\gamma E(\gamma_*) - E(x)$ *along the optimal path. The partition is determined entirely by the energy landscape* E *and the adjacency structure.*

Lemma 7 (Residence Time Lower Bound)**.** (BEGK [2], Theorems 1.2 and 1.4; [3], Theorem 7.8.) *Let* $\tau_{B^c_{\text{stable}}} = \inf\{t > 0 : X_t \notin B_{\text{stable}}\}$*. For any starting state* $x \in B_{\text{stable}}$*:*

$$\mathbb{E}_x[\tau_{B^c_{\text{stable}}}] = K_s \cdot \exp(\Delta_s/\varepsilon) \cdot (1 + \delta_2), \tag{9}$$

where $K_s > 0$ *is a prefactor from the BEGK capacity formula (a rational function of rate prefactors and energies, hence* $O(1)$ *in* ε*), and*

$$|\delta_2| \le C_2 \exp(-c_2/\varepsilon), \qquad c_2 \ge \Delta_s - \Delta_b > 0. \tag{10}$$

The rate constant c_2 *is determined by the sub-leading communication height in the cycle hierarchy.*

Additionally, the rescaled exit time converges in distribution:

$$\frac{\tau_{B^c_{\text{stable}}}}{\mathbb{E}_x[\tau_{B^c_{\text{stable}}}]} \xrightarrow{d} \text{Exp}(1) \qquad \text{as } \varepsilon \to 0, \tag{11}$$

providing concentration: for any $\eta > 0$ *and* ε *sufficiently small,*

$$\mathbb{P}_x\left(\left|\frac{\tau_{B^c_{\text{stable}}}}{\mathbb{E}_x[\tau_{B^c_{\text{stable}}}]} - 1\right| > \eta\right) \le 2\exp(-\eta/2). \tag{12}$$

Output: *Lower bound on* $\mathbb{E}_x[\tau_{B^c_{\text{stable}}}]$ *with multiplicative error* δ_2*, plus the exponential law* (11) *providing distributional convergence.*

Lemma 8 (QSD Convergence)**.** (Champagnat–Villemonais [5], Theorem 2.1.) *The killed chain on* B_{stable} *(with absorption upon hitting* B^c_{stable}*) has a unique quasi-stationary distribution* ν_{QSD} *satisfying*

$$\mathbb{P}_{\nu_{\text{QSD}}}(X_t \in \cdot \mid t < \tau_{B^c_{\text{stable}}}) = \nu_{\text{QSD}}(\cdot) \qquad \text{for all } t \ge 0.$$

For any initial distribution p_0 supported on B_{stable}:

$$\left\|\mathbb{P}_{p_0}(X_t \in \cdot \mid t < \tau_{B^c_{\text{stable}}}) - \nu_{\text{QSD}}\right\|_{\text{TV}} \leq C_3 \exp(-\gamma_D \cdot t), \tag{13}$$

where $\gamma_D > 0$ is the spectral gap of the killed generator $Q_\varepsilon^{B_{\text{stable}}}$.

The spectral gap $\gamma_D = O(1)$ as $\varepsilon \to 0$: within-basin transition rates are $O(1)$ (internal barriers $< \Delta_s$), and the exponentially slow inter-basin rates are removed by the killing. This follows from the Cheeger inequality for the killed chain ([3], Section 7.2).

Since the residence time $\tau_{B^c_{\text{stable}}} \sim \exp(\Delta_s/\varepsilon) \gg 1/\gamma_D$ for small ε, the chain spends all but an $O(1)$-duration transient at the QSD.

Output: *QSD ν_{QSD} on B_{stable} with $\rho(\nu_{\text{QSD}}) \geq c > 0$ and convergence rate $\gamma_D = O(1)$.*

Lemma 9 (Stable Measure Lower Bound)**.** (Combination of Lemmas 7 and 8.) *The experiential measure accumulated during residence in B_{stable} satisfies*

$$\mu_{\text{stable}} = \int_0^{\min(\tau_{B^c_{\text{stable}}}, T_\varepsilon)} \rho(p_t)\, dt \geq c \cdot T_{\min} \cdot (1 - \delta_4), \tag{14}$$

where $T_{\min} = \min(\tau_{B^c_{\text{stable}}}, T_\varepsilon)$ and $\delta_4 = O(1/(\gamma_D \cdot T_{\min}))$.

Proof. Decompose $[0, T_{\min}]$ into transient $[0, t_0]$ and QSD phase $[t_0, T_{\min}]$ with $t_0 = O(1/\gamma_D) = O(1)$ chosen so that $\|p_{t_0}^{\text{killed}} - \nu_{\text{QSD}}\|_{\text{TV}} < \delta$.

During the QSD phase, by Lipschitz continuity of ρ (with constant $L_\rho = O(H(B))$) and (13):

$$|\rho(p_t) - \rho(\nu_{\text{QSD}})| \leq L_\rho C_3 e^{-\gamma_D t}.$$

Using $\rho \geq 0$ on $[0, t_0]$:

$$\begin{aligned}\mu_{\text{stable}} &\geq \int_{t_0}^{T_{\min}} [\rho(\nu_{\text{QSD}}) - L_\rho C_3 e^{-\gamma_D t}]\, dt \\ &\geq c \cdot (T_{\min} - t_0) - L_\rho C_3/\gamma_D \\ &= c \cdot T_{\min} \left(1 - \frac{t_0}{T_{\min}} - \frac{L_\rho C_3}{c\gamma_D T_{\min}}\right).\end{aligned}$$

Since $t_0 = O(1)$ and $T_{\min}$ is exponentially large in $1/\varepsilon$, both correction terms give $\delta_4 = O(1/T_{\min})$. □

Remark 10. *The lower bound is expressed in terms of the random variable $T_{\min}$ rather than the deterministic $K_s \exp(\Delta_s/\varepsilon)$. This is essential: when $\tau_{B^c_{\text{stable}}} > T_\varepsilon$ (the dominant case for $\alpha > 0$), $T_{\min} = T_\varepsilon < K_s \exp(\Delta_s/\varepsilon)$, and the old bound $cK_s \exp(\Delta_s/\varepsilon)(1 - \delta_4)$ would exceed the actual $\mu_{\text{stable}} \leq \rho_{\max} T_\varepsilon$.*

Lemma 11 (BB Occupation Upper Bound)**.** (Renewal theory; BEGK [2], Theorem 1.2 for exit times from B_{BB}; [3], Chapter 7 for the excursion structure.) *During the observation horizon T_ε, the expected BB experiential measure satisfies*

$$\mathbb{E}[\mu_{B_{\text{BB}}}] \leq \rho_{\max} \cdot \frac{K_b}{K_s} \cdot \exp((\Delta_b - \alpha)/\varepsilon) \cdot (1 + \delta_5), \tag{15}$$

where $|\delta_5| \leq C_5 \exp(-c_5/\varepsilon)$ with $c_5 \geq \Delta_s - \Delta_b > 0$.

Proof. The argument proceeds via excursion analysis.

Excursion probability. The chain starts in B_{stable} and must exit before visiting B_{BB}. By (9) and (11):

$$\mathbb{P}(\tau_{B^c_{\text{stable}}} < T_\varepsilon) \leq \frac{T_\varepsilon}{\mathbb{E}[\tau_{B^c_{\text{stable}}}]}(1 + o(1)) = \frac{\exp(-\alpha/\varepsilon)}{K_s(1 + \delta_2)} = O(\exp(-\alpha/\varepsilon)). \tag{16}$$

Conditional BB occupation. Upon exiting B_{stable}, the chain reaches B_{BB} with routing probability $q \leq 1$. By BEGK Theorem 1.2 applied to B_{BB}, the mean residence time during one visit is $\mathbb{E}[\tau_{B_{\text{BB}} \to B^c_{\text{BB}}}] = K_b \exp(\Delta_b/\varepsilon)(1 + \delta'_5)$.

Combining. Since $\rho(p) \leq \rho_{\max}$ for all p:

$$\begin{aligned} \mathbb{E}[\mu_{B_{\text{BB}}}] &\leq \rho_{\max} \cdot \frac{\exp(-\alpha/\varepsilon)}{K_s} \cdot K_b \exp(\Delta_b/\varepsilon) \cdot (1 + O(\exp(-c_5/\varepsilon))) \\ &= \rho_{\max} \cdot \frac{K_b}{K_s} \cdot \exp((\Delta_b - \alpha)/\varepsilon) \cdot (1 + \delta_5). \end{aligned}$$

□

Lemma 12 (Concentration)**.** (BEGK [2], Theorem 1.4; optionally Donsker–Varadhan [6].) *By the BEGK exponential law, the following hold simultaneously with probability at least* $1 - 2e^{-\eta/2} - o(1)$*:*

$$\mu_{\text{stable}} \geq c \cdot K_s \cdot (1 - \eta) \exp(\Delta_s/\varepsilon) \cdot (1 - \delta_4), \tag{17}$$

$$\mu_{B_{\text{BB}}} \leq \rho_{\max} \cdot \frac{K_b}{K_s} \cdot \exp((\Delta_b - \alpha)/\varepsilon) \cdot (1 + \delta_5 + \delta_6), \tag{18}$$

where $\delta_6 = O(\exp(-\alpha/\varepsilon))$ *accounts for the BB occupation deviation.*

For the BB occupation: the excursion probability is $O(\exp(-\alpha/\varepsilon)) \to 0$*, so* $\tau_{B_{\text{BB}}}(T_\varepsilon) = 0$ *with probability* $1 - O(\exp(-\alpha/\varepsilon))$*. Conditional on a visit, the single-visit duration concentrates via the exponential law applied to* B_{BB} *exit.*

For generality, one may invoke the Donsker–Varadhan large deviation principle for the time-averaged empirical measure $L_T = (1/T)\int_0^T \mathbf{1}_{X_t \in A}\, dt$*, giving* $\mathbb{P}(|L_T - \pi(A)| > \eta) \leq \exp(-T \cdot I_{\text{DV}}(\eta))$ *with* $I_{\text{DV}}(\eta) > 0$*. Since* $T = T_\varepsilon$ *is exponentially large, this gives super-exponentially tight concentration. For the present application, the simpler BEGK route suffices.*

Output: *High-probability bounds on* μ_{stable} *(lower) and* $\mu_{B_{\text{BB}}}$ *(upper).*

Lemma 13 (Ratio Assembly)**.** (Direct combination of Lemmas 9–12.) *Under assumptions (A1)–(A4), as* $\varepsilon \to 0$*:*

$$\frac{\mu_{B_{\text{BB}}}}{\mu_{\text{stable}}} \leq C \cdot \exp\left(-\frac{\Delta_s - \Delta_b - \alpha}{\varepsilon}\right) \cdot (1 + \delta_7), \tag{19}$$

where $C = (\rho_{\max}/c)(K_b/K_s^2)$ *and* $|\delta_7| \leq C_7 \exp(-c_7/\varepsilon)$ *with* $c_7 = \min(\Delta_s - \Delta_b, \alpha) > 0$*.*

5 Proof of Theorem A

The proof composes the seven lemmas of Section 4 through the dependency chain $L1 \to (L2, L3) \to (L4, L5) \to L6 \to L7$.

The key structural feature is a **case analysis** on whether the chain exits B_{stable} during $[0, T_\varepsilon]$:

- **Case 1 (no exit):** $\tau_{B^c_{\text{stable}}} > T_\varepsilon$. This is the dominant case.
- **Case 2 (early exit):** $\tau_{B^c_{\text{stable}}} \leq T_\varepsilon$. This has probability $O(\exp(-\alpha/\varepsilon))$, absorbed into the failure budget.

Step 1: Basin partition (Lemma 5)

By Lemma 5, the cycle hierarchy defines metastable basins B_{stable} and B_{BB} with $\Delta_s > \Delta_b > 0$. This is exact – no error term, no approximation.

Output: Partition $(B_{\text{stable}}, B_{\text{BB}}, \Omega \setminus (B_{\text{stable}} \cup B_{\text{BB}}))$ with heights Δ_s, Δ_b.

Step 2: Mean exit time (Lemma 7)

By Lemma 7, $\mathbb{E}_x[\tau_{B^c_{\text{stable}}}] = K_s \exp(\Delta_s/\varepsilon)(1+\delta_2)$ with $|\delta_2| \le C_2 \exp(-c_2/\varepsilon)$, $c_2 \ge \Delta_s - \Delta_b$. The exponential law gives $\tau_{B^c_{\text{stable}}}/\mathbb{E}[\tau_{B^c_{\text{stable}}}] \xrightarrow{d} \text{Exp}(1)$.

Input: Basin B_{stable} with height Δ_s from Step 1.

Output: Mean exit time with multiplicative error δ_2 and the exponential law providing concentration.

Step 3: QSD convergence (Lemma 8)

By Lemma 8, the killed chain converges to ν_{QSD} with rate $\gamma_D = O(1)$ and $\rho(\nu_{\text{QSD}}) \ge c > 0$.

Input: Basin B_{stable} from Step 1.

Output: QSD with density bound and convergence rate.

Step 4: Stable measure lower bound (Lemma 9)

By Lemma 9, $\mu_{\text{stable}} \ge c \cdot T_{\min} \cdot (1-\delta_4)$ where $T_{\min} = \min(\tau_{B^c_{\text{stable}}}, T_\varepsilon)$ and $\delta_4 = O(1/T_{\min})$.

On event $A_1 = \{\tau_{B^c_{\text{stable}}} > T_\varepsilon\}$: $T_{\min} = T_\varepsilon = \exp((\Delta_s - \alpha)/\varepsilon)$, giving $\delta_4 = O(\exp(-(\Delta_s-\alpha)/\varepsilon))$.

Inputs: Exit time from Step 2; QSD from Step 3.

Output: Lower bound on μ_{stable} in terms of $T_{\min}$, valid for any realization.

Step 5: BB measure upper bound (Lemma 11)

By Lemma 11, $\mathbb{E}[\mu_{B_{\text{BB}}}] \le \rho_{\max}(K_b/K_s)\exp((\Delta_b - \alpha)/\varepsilon)(1+\delta_5)$ with $c_5 \ge \Delta_s - \Delta_b$.

Input: Exit time from Step 2; basin structure from Step 1.

Output: Expected-value upper bound on $\mu_{B_{\text{BB}}}$ (converted to high-probability in Step 6).

Step 6: Case analysis and concentration (Lemma 12)

Define events:

$$\begin{aligned} A_1 &= \{\tau_{B^c_{\text{stable}}} > T_\varepsilon\} && \text{(no exit)},\\ A_2 &= \{\tau_{B^c_{\text{stable}}} \le T_\varepsilon\} && \text{(early exit)}. \end{aligned}$$

Probability of A_2: By the exponential law with $T_\varepsilon/\mathbb{E}[\tau_{B^c_{\text{stable}}}] = \exp(-\alpha/\varepsilon)/(K_s(1+\delta_2))$:

$$\mathbb{P}(A_2) \le \frac{\exp(-\alpha/\varepsilon)}{K_s}(1+o(1)) = O(\exp(-\alpha/\varepsilon)). \tag{20}$$

Case 1 – on A_1: The chain remains in B_{stable} throughout $[0, T_\varepsilon]$, so $\mu_{B_{\text{BB}}} = 0$ and

$$\frac{\mu_{B_{\text{BB}}}}{\mu_{\text{stable}}} = 0 \le C \cdot \exp\left(-\frac{\Delta_s - \Delta_b - \alpha}{\varepsilon}\right).$$

The bound (6) is trivially satisfied.

Case 2 – on A_2: The chain exits B_{stable} and may visit B_{BB}. We do not need to bound the ratio on this event because A_2 is absorbed into the failure probability:

$$\mathbb{P}(A_2) = O(\exp(-\alpha/\varepsilon)) \ll O(\exp(-\alpha/(2\varepsilon))),$$

since $\alpha/\varepsilon > \alpha/(2\varepsilon)$ for all $\varepsilon > 0$.

Combined:

$$\mathbb{P}\left(\frac{\mu_{B_{\text{BB}}}}{\mu_{\text{stable}}} > C \cdot \exp\left(-\frac{\Delta_s - \Delta_b - \alpha}{\varepsilon}\right) \cdot (1 + \delta(\varepsilon))\right) \leq \mathbb{P}(A_2) = O(\exp(-\alpha/(2\varepsilon))). \tag{21}$$

Although the bound is trivially satisfied on A_1 (where the ratio is zero), the right-hand side $C\exp(-(\Delta_s - \Delta_b - \alpha)/\varepsilon)$ captures the correct scaling: as $\alpha \to 0$, the observation time approaches the mean exit time, $\mathbb{P}(A_2) \to O(1)$, and the bound approaches the ergodic ratio $C\exp(-(\Delta_s - \Delta_b)/\varepsilon)$.

Step 7: Ratio assembly and error composition (Lemma 13)

It remains to verify the error term $\delta(\varepsilon)$ and the prefactor C.

Ergodic-limit consistency. As $\alpha \to 0$, using Steps 4–5:

$$\begin{aligned}\frac{\mathbb{E}[\mu_{B_{\text{BB}}}]}{\mathbb{E}[\mu_{\text{stable}}]} &\leq \frac{\rho_{\max}(K_b/K_s)\exp((\Delta_b - \alpha)/\varepsilon)(1 + \delta_5)}{c \cdot K_s \exp(\Delta_s/\varepsilon)(1 - \delta_2)} \\ &= \frac{\rho_{\max} K_b}{c K_s^2 (1 - \delta_2)} \cdot \exp\left(\frac{\Delta_b - \alpha - \Delta_s}{\varepsilon}\right) \cdot (1 + \delta_5),\end{aligned}$$

confirming $C = (\rho_{\max}/c)(K_b/K_s^2)$.

Error collection. The correction factor from all composition steps is:

$$\frac{(1 + \delta_5)}{(1 - \delta_2)(1 - \delta_4')(1 - \eta)} = 1 + O(\exp(-\gamma/\varepsilon)),$$

where $\eta = \exp(-\alpha/(2\varepsilon))$ and

$$\gamma = \min(c_2, \Delta_s - \alpha, c_5, \alpha/2) = \min(\Delta_s - \Delta_b, \Delta_s - \alpha, \alpha/2) = \min(\alpha/2,\, \Delta_s - \alpha), \tag{22}$$

since $\alpha/2 < \alpha < \Delta_s - \Delta_b$ and $\Delta_s - \alpha > 0$ (both follow from $0 < \alpha < \Delta_s - \Delta_b < \Delta_s$). When $\alpha \leq \Delta_b$, the ordering $\alpha/2 < \Delta_s - \alpha$ holds and $\gamma = \alpha/2$. When $\alpha > \Delta_b$, one can have $\Delta_s - \alpha < \alpha/2$ (specifically when $\alpha > 2\Delta_s/3$), so the general expression retains both terms. For the typical regime where α is well away from the endpoints of $(0, \Delta_s - \Delta_b)$, $\gamma = \alpha/2$.

Prefactor verification. C is $O(1)$ in ε: $\rho_{\max} = (\ln|B|)/4$ depends only on $|B|$; $c = \rho(\nu_{\text{QSD}}) > 0$ is bounded away from zero by (A2); K_s, K_b are BEGK capacity prefactors, rational functions of rates and energies on a finite state space ([3], Theorem 7.8). □

6 Error Composition Verification

Proposition 14 (Error product preserves exponential form)**.** *The product of all correction factors arising in Steps 2–7 is* $1 + O(\exp(-\gamma/\varepsilon))$ *with* $\gamma = \min(\alpha/2,\, \Delta_s - \alpha) > 0$.

Proof. The individual error terms and their rates are:

Source	**Error term**	**Rate**	**Origin**
L2	$\delta_2 = O(e^{-c_2/\varepsilon})$	$c_2 \geq \Delta_s - \Delta_b$	BEGK Thm 1.2
L3	$\delta_3(t) = C_3 e^{-\gamma_D t}$	$\gamma_D = O(1)$	CV16 Thm 2.1
L4	$\delta_4' = O(1/T_\varepsilon)$	$\Delta_s - \alpha$	L2+L3, on A_1
L5	$\delta_5 = O(e^{-c_5/\varepsilon})$	$c_5 \geq \Delta_s - \Delta_b$	BEGK for B_{BB}
$(1-\eta)^{-1}$	$O(e^{-\alpha/(2\varepsilon)})$	$\alpha/2$	choice of η

All rates are strictly positive:

$$\begin{aligned} c_2 &\geq \Delta_s - \Delta_b > 0, \\ \Delta_s - \alpha &> \Delta_b > 0, \\ c_5 &\geq \Delta_s - \Delta_b > 0, \\ \alpha/2 &> 0. \end{aligned}$$

The combined correction is $1 + O(\exp(-\gamma/\varepsilon))$ with

$$\gamma = \min(c_2, \Delta_s - \alpha, c_5, \alpha/2) = \min(\Delta_s - \Delta_b, \Delta_s - \alpha, \alpha/2) = \min(\alpha/2,\ \Delta_s - \alpha),$$

since $\alpha/2 < \Delta_s - \Delta_b$ (from $\alpha < \Delta_s - \Delta_b$) and $\Delta_s - \alpha > 0$ (from $\alpha < \Delta_s$).

No polynomial-in-$1/\varepsilon$ prefactors arise: each C_i depends only on the problem parameters $(|B|, |M|, r, E)$, and the product of finitely many such terms introduces at most a constant factor. □

7 Dependency Graph

The seven lemmas form a directed acyclic graph:

Topological order: $L1 \to (L2, L3) \to (L4, L5) \to L6 \to L7$.

Source	**Target**	**Object transferred**	**Type match**
L1	L2	Basin B_{stable} with height Δ_s (exact)	Exact → exact
L1	L3	Basin B_{stable} as subset of Ω (exact)	Exact → exact
L2	L4	$\mathbb{E}[\tau_{B^c_{\text{stable}}}]$ (expectation + exp. law)	Expectation → high-prob. via BEGK Thm 1.4
L3	L4	ν_{QSD} with rate γ_D (TV bound)	Distribution + rate → transient error
L2	L5	Exit time giving excursion probability	Expectation → expected occupation
L4	L7	$\mu_{\text{stable}} \geq c\,T_{\min}(1-\delta_4)$	Lower bound (high-prob. via L6)
L5	L7	$\mathbb{E}[\mu_{B_{\text{BB}}}]$ upper bound	Expected value → high-prob. via L6
L6	L7	High-prob. bounds on both measures	High-prob. → ratio bound

The two potential type mismatches (expectations flowing where high-probability bounds are needed) occur at L2→L4 and L5→L7. Both are resolved by the BEGK exponential law (Lemma 7, eq. (11)) and the concentration argument (Lemma 12), respectively.

8 Numerical Validation: Three-State Chain

We validate Theorem A on the three-state chain $s \leftrightarrow u \leftrightarrow b$ with parameters:

- $B_{\text{stable}} = \{s\}$ (stable state), $B_{\text{BB}} = \{b\}$ (Boltzmann brain state), $u =$ saddle.
- Communication heights: $\Delta_s = 3.0$, $\Delta_b = 1.0$.
- Experiential densities: $\rho_s = 0.8$, $\rho_b = 0.2$.
- Branching probability p at the saddle: probability of returning to s from u.

For this system, the parameters specialize as:

- $K_s = K_b = 1$ (single-state basins with exit rate $\exp(-\Delta/\varepsilon)$).
- $c = \rho_s = 0.8$ (QSD on $\{s\}$ is δ_s).
- $\rho_{\max} = \rho_b = 0.2$ for the BB contribution.
- Routing probability: $q = 1 - p$.

8.1 Theorem A prefactor

The general bound gives:

$$C = \frac{\rho_{\max}}{c} \cdot \frac{K_b}{K_s^2} = \frac{0.2}{0.8} \cdot \frac{1}{1} = 0.25.$$

The exact ergodic ratio for this chain (from detailed balance) is:

$$\frac{\mu_{B_{\mathrm{BB}}}}{\mu_{\mathrm{stable}}} = \frac{\rho_b}{\rho_s} \cdot \frac{(1-p)}{p} \cdot \exp\left(-\frac{\Delta_s - \Delta_b}{\varepsilon}\right).$$

The exact ergodic prefactor $C_{\mathrm{exact}} = (\rho_b/\rho_s)(1-p)/p$ equals $C = 0.25$ when $p = 0.5$ (exact match).

8.2 Test results

We tested 9 combinations of (p, α) with $p \in \{0.3, 0.5, 0.7\}$ and $\alpha \in \{0.5, 1.0, 1.5\}$ using the verification code `three_state_chain.py`. All 9 test cases satisfy the Theorem A bound, with exact prefactor agreement at $p = 0.5$.

p	C_{exact}	C_{bound}	Tight at $p = 0.5$?	Sub-ergodic valid?
0.3	0.583	0.25	No	Yes (ratio = 0 on A_1)
0.5	0.250	0.25	Exact match	Yes
0.7	0.107	0.25	Conservative	Yes

For $p < 0.5$, the bound $C = 0.25$ is smaller than C_{exact}, so it is not a valid upper bound on the exact ergodic ratio. However, Theorem A with $\alpha > 0$ remains valid because on event A_1 (probability $\geq 1 - O(\exp(-\alpha/\varepsilon))$), the ratio is exactly zero. The discrepancy at $p \neq 0.5$ arises from the renewal structure (factor $1/p$ in the ergodic limit), which is beyond the scope of the sub-ergodic theorem.

8.3 Key numerical findings

1. **All 9 test cases pass.** For each (p, α) combination, the computed ratio satisfies the Theorem A bound, confirming the exponential scaling and case-analysis structure.
2. $C = 0.25$ **for** $p = 0.5$ is an exact match between the general bound and the specific three-state chain prefactor.
3. **Error rate** $\gamma = \min(\alpha/2,\ \Delta_s - \alpha)$**.** The observed error decay is consistent with the theoretical rate.
4. **Case analysis structure confirmed.** The exit-vs-no-exit decomposition is numerically validated: on A_1 the ratio is exactly zero, and $\mathbb{P}(A_2) = O(\exp(-\alpha/\varepsilon))$ as predicted.

9 Discussion

9.1 Summary of results

We have proven that in finite-state reversible Markov chains with Metropolis dynamics, the trajectory-weighted occupation of a shallow metastable basin is exponentially suppressed relative to that of a deeper basin, with explicit control over all error terms. Specifically:

- The ratio $\mu_{B_{\mathrm{BB}}}/\mu_{\mathrm{stable}}$ decays as $\exp(-(\Delta_s - \Delta_b - \alpha)/\varepsilon)$ with probability at least $1 - O(\exp(-\alpha/(2\varepsilon)))$.
- The composite error term preserves exponential form with rate $\gamma = \min(\alpha/2, \Delta_s - \alpha) > 0$, verified through explicit tracking of all seven lemma error terms.
- The prefactor $C = (\rho_{\max}/c)(K_b/K_s^2)$ is finite and independent of ε, with each component verified to be $O(1)$.
- The proof structure – a case analysis on whether the chain exits B_{stable} during the observation window – provides the correct framework for the sub-ergodic regime ($\alpha > 0$) while maintaining consistency with the ergodic limit ($\alpha \to 0$).

9.2 Trivial satisfaction and the role of α

For $\alpha > 0$, the dominant event $A_1 = \{\tau_{B^c_{\mathrm{stable}}} > T_\varepsilon\}$ has probability $1 - O(\exp(-\alpha/\varepsilon))$, and on A_1 the chain never exits B_{stable}, so $\mu_{B_{\mathrm{BB}}} = 0$ and the bound is trivially satisfied. The suppression mechanism in this regime is *basin trapping*: the observation horizon expires before the chain can escape the deep basin, and neither the form of ρ nor any information-theoretic content plays a role. This is structurally parallel to geometric cutoff measures in cosmology (causal diamond [14], scale-factor cutoff [10]), which suppress Boltzmann brains by truncating the time window before late-time fluctuations can dominate.

The non-trivial regime is the ergodic limit $\alpha \to 0$, where the observation horizon approaches the mean exit time, $\mathbb{P}(A_2) \to O(1)$, and the chain makes $O(1)$ excursions to B_{BB}. In this regime the prefactor $C = (\rho_{\max}/c)(K_b/K_s^2)$ – and hence the choice of ρ – determines the numerical value of the suppression ratio, even though the exponential rate $\Delta_s - \Delta_b$ is independent of ρ. For $\alpha > 0$, the experiential measure adds nothing beyond what a simple basin-exit analysis provides.

Remark 15 (Role of ρ versus metastability)**.** *The exponential suppression factor* $\exp(-(\Delta_s - \Delta_b - \alpha)/\varepsilon)$ *comes entirely from the metastability structure (the communication height gap* $\Delta_s - \Delta_b$*). The experiential density* ρ *enters only in the* $O(1)$ *prefactor* $C = (\rho_{\max}/c)(K_b/K_s^2)$*. Any positive bounded functional* f *with* $f(\nu_{\mathrm{QSD}}^{B_{\mathrm{stable}}}) \geq c > 0$ *would produce the same exponential suppression. In particular, a plain occupation-time ratio (*$f \equiv 1$*) gives* $C = K_b/K_s^2$ *– the same exponential decay with no information-theoretic content. The specific choice of* ρ *determines which states are "weighted" but not the rate of suppression.*

9.3 Limitations

Several limitations of the current result should be noted:

1. **Finite state spaces.** The proof relies on finiteness of Ω in several places: the Freidlin–Wentzell cycle decomposition is a graph-theoretic construction on finite graphs; the BEGK capacity formula applies to finite reversible chains; and $\rho_{\max} = H(B)/4 < \infty$ requires $|B| < \infty$. Extension to infinite (e.g., continuous) state spaces would require replacing these tools with their continuum counterparts (potential-theoretic capacities, Witten Laplacian estimates).
2. **Reversible dynamics.** The Metropolis-type generator ensures detailed balance, which underlies both the BEGK capacity asymptotics and the Freidlin–Wentzell cycle hierarchy. Non-reversible dynamics (e.g., systems with persistent currents) would require different metastability tools.

3. **Routing probability.** The bound uses $q \leq 1$ for the probability of reaching B_{BB} upon exiting B_{stable}. For systems where $q \ll 1$ (e.g., when the saddle preferentially routes back to B_{stable}), the bound is conservative. Sharpening the prefactor to include routing probabilities is straightforward but requires system-specific information.

4. **Single-excursion regime.** The proof bounds at most one excursion to B_{BB} during $[0, T_\varepsilon]$. The ergodic-limit analysis (multiple excursions over exponentially long times) introduces a renewal factor $1/p$ that is not captured by the current argument. This is reflected in the $p < 0.5$ discrepancy for the three-state chain.

9.4 Connection to the broader program

Theorem A is the first component of a program to formalize properties of the experiential measure framework. The broader program includes:

- **Lipschitz stability** of the experiential density ρ under perturbations of the distribution p, ensuring that the density functional is well-behaved under the continuous-time dynamics [18].
- **Born–Fisher test**: numerical testing of whether the experiential measure framework reproduces Born rule statistics via a variational principle. The conjecture was falsified for exchange-plus-dephasing Lindblad models [19], establishing a boundary on the framework's quantum applicability.
- **Extension to continuous state spaces** and non-reversible dynamics, which would make the framework applicable to realistic physical systems beyond the toy models considered here.

The three-state chain serves as a minimal but complete test case where all components of the proof can be verified analytically and numerically, providing confidence in the machinery before tackling more complex systems.

References

[1] A. Albrecht and L. Sorbo, *Can the universe afford inflation?*, Phys. Rev. D **70** (2004) 063528.

[2] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein, *Metastability and low lying spectra in reversible Markov chains*, Comm. Math. Phys. **228** (2002) 219–255.

[3] A. Bovier and F. den Hollander, *Metastability: A Potential-Theoretic Approach*, Grundlehren der mathematischen Wissenschaften **351**, Springer (2015).

[4] R. Bousso and B. Freivogel, *A paradox in the global description of the multiverse*, J. High Energy Phys. **2007**(06) (2007) 018.

[5] N. Champagnat and D. Villemonais, *Exponential convergence to quasi-stationary distribution and Q-process*, Probab. Theory Relat. Fields **164** (2016) 243–283.

[6] M. D. Donsker and S. R. S. Varadhan, *Asymptotic evaluation of certain Markov process expectations for large time*, Comm. Pure Appl. Math. I: **28** (1975) 1–47; II: **28** (1975) 279–301; III: **29** (1976) 389–461; IV: **36** (1983) 183–212.

[7] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, 3rd ed., Grundlehren der mathematischen Wissenschaften **260**, Springer (2012).

[8] D. N. Page, *Is our universe likely to decay within 20 billion years?*, Phys. Rev. D **78** (2008) 063535; hep-th/0610079 (2006).

[9] L. Dyson, M. Kleban, and L. Susskind, *Disturbing implications of a cosmological constant*, J. High Energy Phys. **2002**(10) (2002) 011.

[10] A. De Simone, A. H. Guth, A. Linde, M. Noorbala, M. P. Salem, and A. Vilenkin, *Boltzmann brains and the scale-factor cutoff measure of the multiverse*, Phys. Rev. D **82** (2010) 063520.

[11] S. M. Carroll, *Why Boltzmann Brains Are Bad*, arXiv:1702.00850 (2017).

[12] K. K. Boddy, S. M. Carroll, and J. Pollack, *De Sitter space without dynamical quantum fluctuations*, Found. Phys. **46** (2016) 702–735; arXiv:1405.0298 (2014).

[13] J. B. Hartle and M. Srednicki, *Are we typical?*, Phys. Rev. D **75** (2007) 123523.

[14] R. Bousso, *Holographic probabilities in eternal inflation*, Phys. Rev. Lett. **97** (2006) 191302.

[15] M. Tegmark, *Consciousness as a state of matter*, Chaos, Solitons & Fractals **76** (2015) 238–270; arXiv:1401.1219 (2014).

[16] G. Tononi, *An information integration theory of consciousness*, BMC Neuroscience **5** (2004) 42.

[17] R. Lopez-Ruiz, H. L. Mancini, and X. Calbet, *A statistical measure of complexity*, Phys. Lett. A **209** (1995) 321–326.

[18] B. Ehrlich, *Lipschitz Stability of the Experiential Density Functional*, preprint (2026).

[19] B. Ehrlich, *Falsification of the Born–Fisher–Experiential Conjecture in a Qubit Toy Model*, preprint (2026).