Statistical Mixing Guarantees for Contractive Echo State Networks

Pradeep Singh*
Department of Computer Science and Engineering
Indian Institute of Information Technology Surat
Surat, India
pradeep.singh@iiitsurat.ac.in

Balasubramanian Raman
Department of Computer Science and Engineering
Indian Institute of Technology Roorkee
Roorkee, India
bala@cs.iitr.ac.in

Abstract

Reservoir computing turns sequence learning into linear regression on fixed dynamical features, but those features come from a single highly dependent trajectory, so the true amount of usable data is often unclear. We develop a transport-based learning theory for contractive echo state networks driven by stochastic inputs by viewing the reservoir as a Markov process and proving a Wasserstein contraction that guarantees a unique stationary state law and explicit mixing rates. Using contraction-to-concentration tools, we obtain finite-sample deviation bounds for time-averaged features and empirical covariance estimates, and translate them into excess-risk and stability guarantees for ridge-trained linear readouts. The theory yields an explicit effective-sample-size principle and quantifies a certified trade-off: moving the contraction certificate closer to one can permit longer dynamical memory while weakening the guaranteed mixing rate and increasing the sufficient data requirement. Overall, the results complement echo-state and capacity analyses by providing verifiable design rules that link stability parameters to mixing time, generalization, and stationary readout accuracy under i.i.d. inputs, with an augmented-state extension for contractive Markov inputs.

1 Introduction

Reservoir computing is often sold as an almost paradoxical bargain: keep a high-dimensional recurrent network fixed, drive it with an input stream, and learn only a linear readout—yet still obtain strong performance on forecasting and sequence modeling tasks [13, 19, 18]. The bargain hides a subtle statistical question. Training the readout does not see independent samples; it sees a *single trajectory* of reservoir states produced by a dynamical system. Whether that trajectory contains "enough data" depends not only on the dataset length T, but also on how quickly the driven reservoir *mixes*—i.e., how fast it forgets its past in distribution. Without a quantitative mixing theory, even widely used heuristics (e.g., tuning reservoirs near the edge of stability or "criticality") are statistically ambiguous: slower forgetting may improve dynamical memory, but it can also inflate variance by shrinking the effective sample size.

What we know and what is missing. Foundational theory for ESNs largely focuses on well-posedness and representational power. The echo state property (ESP) and fading memory conditions formalize when the state is uniquely determined by the input history and when the influence of the remote past decays [13, 18, 17]. On the expressivity side, ESNs and related reservoirs are known to be universal approximators for broad classes of fading-memory filters [9]. Complementary lines of work

*Corresponding author: pradeep.singh@iiitsurat.ac.in

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

quantify *what* information reservoirs can store and process via memory capacity and information processing capacity measures [29, 5, 2], and relate computational performance to dynamical regimes such as the "edge of chaos" [1]. These results explain stability, approximation, and memory, but they leave open a basic learning-theoretic link: *how does the reservoir's dynamical contraction translate into finite-sample generalization and trainability of the readout when data are temporally dependent?* General statistical learning bounds for reservoir families under weak dependence exist [6], but they do not directly yield an explicit, mechanistic "mixing time $\rightarrow$ effective sample size $\rightarrow$ generalization" pipeline tied to the reservoir's contraction parameters.

Recent related theory. A growing body of recent work has begun to put reservoir computing on a more explicitly *statistical* and *dynamical-systems* footing, complementing classical ESP and universality analyses. On the learning side, finite-sample generalization for reservoir-style regressors has been sharpened in several directions, including tighter bounds that make explicit how dependence and model size interact with the number of observed time steps [10]. In parallel, a dynamical-systems viewpoint has been used to explain why linear readouts trained by Tikhonov (ridge) regression can approximate observables generated by *ergodic* dynamical systems from a single trajectory [11]. Other works develop structural explanations for reservoir efficacy via embedding arguments and network-size reduction principles (e.g. time-delay tradeoffs) [4]. Most closely in spirit to our approach, transport-based analyses have been proposed for driven systems underlying reservoirs: under verifiable contractivity-type conditions one can establish existence and uniqueness of asymptotically invariant state laws and continuity of these laws with respect to changes in the input distribution in Wasserstein distance [20]. Finally, approximation theory for random recurrent architectures and even infinite-dimensional reservoir models has advanced substantially, clarifying when and how random reservoirs can approximate large classes of dynamical input–output maps [7, 8]. These developments motivate a central question addressed here: *when training is done on a single dependent reservoir trajectory, what is the quantitative data requirement implied by the reservoir's stability/memory parameters?*

Our viewpoint: transport contraction as a quantitative mixing certificate. This paper proposes a different organizing principle: study the reservoir not only as a state map, but as a *Markov operator* acting on probability laws of the state. When the reservoir update is contractive in state (the classical ESP-friendly regime), the induced Markov kernel is contractive in Wasserstein distance, a metric from optimal transport [28]. This perspective connects reservoir design to the modern "metric ergodicity" toolkit for Markov processes: contraction in Wasserstein can be interpreted as a form of positive coarse Ricci curvature [22], and it enables nonasymptotic concentration for time averages of Lipschitz observables along a single trajectory [14]. We leverage this chain of implications to derive explicit and verifiable learning guarantees for linear readouts trained on correlated reservoir features.

Mechanism at a glance. Technically, our results can be read as a chain of implications with explicit constants: state contraction $(L) \implies$ Wasserstein contraction $(\kappa \geq 1 - L) \implies$ time-average concentration (rate κT) $\implies$ finite-sample ridge guarantees. The first step is purely dynamical: uniform Lipschitz contraction of the reservoir update in state implies a Wasserstein–1 contraction of the induced Markov operator on state *laws*. The second step is probabilistic: transport contraction (equivalently, positive coarse Ricci curvature) yields bounded-diameter concentration for Lipschitz observables along a *single* Markov trajectory, with an exponent proportional to the curvature gap κ. The final step is statistical: once the empirical Gram matrix and feature–target moments concentrate at rate κT, standard perturbation arguments for ridge regression turn these deviations into parameter and excess-risk bounds. This pipeline turns the informal "dependence reduces data" intuition into a verifiable *effective-sample-size principle* driven by κ.

Contributions. Under uniform contractivity and i.i.d. stochastic inputs, with extensions to contractive augmented Markov models, we provide:

- **Wasserstein mixing for contractive reservoirs.** We prove that the reservoir-induced Markov operator contracts Wasserstein–1 at an explicit rate tied to the reservoir's contraction constant, implying uniqueness of a stationary state distribution and geometric mixing bounds.

- **Concentration and effective sample size.** We obtain finite-T deviation inequalities for time-averaged reservoir features and quadratic forms (empirical Gram matrices), yielding an explicit effective-sample-size scaling when correlations decay geometrically.

- **Finite-sample readout guarantees.** For ridge-trained linear readouts, we convert the above concentration into excess-risk and stability bounds that expose a concrete trade-off between *dynamical memory* and *statistical efficiency*.
- **Design rules.** The final corollaries read as actionable constraints: given a target generalization error and confidence, our bounds specify how long a training sequence is needed as a function of contraction, feature Lipschitz constants, and regularization—turning "tune near criticality" into a quantitatively testable prescription.

2 Setting and Transport Preliminaries

This paper studies *statistical* trainability of contractive reservoirs from a *single dependent trajectory*. While the echo state property ensures that a reservoir state is well-defined and forgets remote history pointwise [13, 18], learning a readout from time samples requires quantitative control of *distributional* forgetting and mixing. Our organizing tool is optimal transport: we show that contractive reservoirs induce Markov operators that contract the Wasserstein–1 distance, yielding explicit mixing rates and, in subsequent sections, concentration and effective-sample-size bounds [28, 22, 14].

2.1 Contractive reservoirs driven by stochastic inputs

Let Σ be a finite input alphabet. We consider reservoirs on a compact state space $X \subset \mathbb{R}^N$ equipped with the metric $d(x, y) = \|x - y\|$ (Euclidean norm), and write $\mathrm{diam}(X) := \sup_{x,y\in X} d(x, y) < \infty$. For each symbol $s \in \Sigma$, let $F_s : X \to X$ be the corresponding state update map. A standard ESN instance is

$$F_s(x) = (1 - \alpha)x + \alpha\, \phi(Wx + Us + b), \tag{1}$$

where ϕ is coordinate-wise 1-Lipschitz, $\alpha \in (0, 1]$, and W, U, b are fixed parameters [13, 18].

We assume the reservoir is driven by a stochastic input process $(U_t)_{t\geq 0}$ taking values in Σ, and evolves as

$$x_{t+1} = F_{U_t}(x_t), \qquad t \geq 0. \tag{2}$$

Throughout the main paper we focus on the simplest and most transparent case.

Assumption 2.1 (Compactness and uniform contractivity)**.**

1. *(**Forward invariance**) $F_s(X) \subseteq X$ for all $s \in \Sigma$.*
2. *(**Uniform contraction**) There exists $L \in (0, 1)$ such that*

$$d(F_s(x), F_s(y)) \leq L\, d(x, y)\ \forall x, y \in X,\ \forall s \in \Sigma. \tag{3}$$

3. *(**i.i.d. inputs**) $(U_t)_{t\geq 0}$ are i.i.d. with law $\nu \in \Delta(\Sigma)$, independent of x_0.*

Assumption 2.1(2) is the classical Lipschitz sufficient condition for the echo state property in ESNs, e.g. for (1) one may take $L \leq (1 - \alpha) + \alpha\|W\|$ when ϕ is 1-Lipschitz [13, 18]. Assumption 2.1(3) makes $(x_t)_{t\geq 0}$ a time-homogeneous Markov chain on X; extensions to weakly dependent inputs are handled by augmenting the state (Appendix C).

2.2 The induced Markov kernel and operator

Under Assumption 2.1(3), the one-step transition kernel P on X is

$$P(x, A) := \mathbb{P}(F_U(x) \in A) = \sum_{s\in\Sigma} \nu(s)\, \mathbf{1}\{F_s(x) \in A\}, \tag{4}$$

where $A \subseteq X$ Borel, $U \sim \nu$. We use the standard dual actions of P: on bounded measurable functions $f : X \to \mathbb{R}$, $(Pf)(x) := \int f(y)\, P(x, dy) = \mathbb{E}\big[f(F_U(x))\big]$; on probability measures $\mu \in \mathcal{P}(X)$, $(\mu P)(A) := \int P(x, A)\, \mu(dx)$. Equivalently, writing $(F_s)_\sharp\mu$ for the pushforward of μ by F_s is

$$\mu P = \sum_{s\in\Sigma} \nu(s)\, (F_s)_\sharp\mu. \tag{5}$$

2.3 Wasserstein–1 distance and its Lipschitz dual

Because X is compact, every $\mu \in \mathcal{P}(X)$ has finite first moment. The Wasserstein–1 distance between $\mu, \eta \in \mathcal{P}(X)$ is

$$W_1(\mu, \eta) := \inf_{\gamma \in \Gamma(\mu,\eta)} \int_{X \times X} d(x, y)\, \gamma(dx, dy), \tag{6}$$

where $\Gamma(\mu, \eta)$ denotes couplings with marginals μ and η. We will repeatedly use the Kantorovich–Rubinstein duality [28]:

$$W_1(\mu, \eta) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left\{ \int f\, d\mu - \int f\, d\eta \right\}, \tag{7}$$

where $\|f\|_{\mathrm{Lip}} := \sup_{x \neq y} \frac{|f(x)-f(y)|}{d(x,y)}$.

Two standard consequences (proofs deferred to Appendix A) connect reservoir contractivity to transport contraction:

Lemma 2.2 (Pushforward contracts Wasserstein)**.** *If $T : X \to X$ is K-Lipschitz, then $W_1(T_\sharp \mu, T_\sharp \eta) \le K\, W_1(\mu, \eta)$ for all $\mu, \eta \in \mathcal{P}(X)$.*

Lemma 2.3 (Mixtures preserve contraction)**.** *Let $\{T_i\}_{i \in I}$ be measurable maps and $(p_i)_{i \in I}$ a probability mass function. Then $W_1(\sum_i p_i\, (T_i)_\sharp \mu,\ \sum_i p_i\, (T_i)_\sharp \eta) \le \sum_i p_i\, W_1\big((T_i)_\sharp \mu, (T_i)_\sharp \eta\big)$.*

2.4 Observables and the learning objects

Later sections translate mixing into finite-sample guarantees for linear readouts. In the main results we consider supervised targets that are *state observables*: we fix a measurable function $y : X \to \mathbb{R}$ and define $y_t := y(x_t)$. We will analyze empirical averages such as $G_T := \frac{1}{T} \sum_{t=1}^{T} \phi(x_t)\phi(x_t)^\top$, $b_T := \frac{1}{T} \sum_{t=1}^{T} \phi(x_t)\, y(x_t)$. Extensions with observation noise or extra target variables require a joint Markov state satisfying the contraction and observable assumptions; Appendix C gives the augmented-state contraction criterion. The central question is how the dependence induced by (2) affects the concentration of G_T and b_T around their stationary counterparts. In Section 3 we show that Assumption 2.1 implies geometric mixing in W_1 with explicit rate L, providing the quantitative input to concentration and effective-sample-size results (Section 4).

3 Wasserstein Contraction and Mixing of Contractive Reservoirs

Assumption 2.1 implies more than pointwise forgetting of initial conditions: it yields a *distributional* contraction principle for the Markov operator induced by stochastic driving. This section establishes (i) W_1-contraction of the operator P, (ii) existence and uniqueness of a stationary reservoir law, and (iii) explicit geometric mixing rates and correlation decay for Lipschitz observables. These statements are the probabilistic backbone for our later concentration and learning bounds, and they connect reservoir design (through L) to statistical efficiency.

3.1 One-step Wasserstein contraction

Theorem 3.1 (Wasserstein contraction of the reservoir Markov operator)**.** *Under Assumption 2.1, for all $\mu, \eta \in \mathcal{P}(X)$,*

$$W_1(\mu P, \eta P) \ \le\ L\, W_1(\mu, \eta). \tag{8}$$

Equivalently, for all $x, y \in X$,

$$W_1\big(P(x, \cdot), P(y, \cdot)\big) \ \le\ L\, d(x, y). \tag{9}$$

Curvature interpretation. In the sense of Ollivier's coarse Ricci curvature [22], the Markov chain has curvature bounded below by

$$\kappa \ :=\ 1 - \sup_{x \neq y} \frac{W_1(P(x, \cdot), P(y, \cdot))}{d(x, y)} \ \ge\ 1 - L \ > 0. \tag{10}$$

Positive curvature (here, explicit) is precisely the condition that enables sharp mixing and concentration bounds for Markov chains on metric spaces [22, 14].

3.2 Stationary law and geometric mixing

Corollary 3.2 (Unique stationary reservoir distribution)**.** *There exists a unique invariant probability measure $\pi \in \mathcal{P}(X)$ such that $\pi P = \pi$.*

Corollary 3.3 (Geometric ergodicity in Wasserstein)**.** *For any initial distribution $\mu \in \mathcal{P}(X)$ and any $t \geq 0$,*

$$W_1(\mu P^t, \pi) \;\leq\; L^t\, W_1(\mu, \pi). \tag{11}$$

In particular, for any initial state $x \in X$,

$$W_1(\delta_x P^t, \pi) \;\leq\; L^t \operatorname{diam}(X). \tag{12}$$

Mixing time. For $0 < \varepsilon < \operatorname{diam}(X)$, define $t_{\mathrm{mix}}(\varepsilon) := \inf\{t \in \mathbb{N}_0 : \sup_{x \in X} W_1(\delta_x P^t, \pi) \leq \varepsilon\}$. For $\varepsilon \geq \operatorname{diam}(X)$, $t_{\mathrm{mix}}(\varepsilon) = 0$. From (12),

$$t_{\mathrm{mix}}(\varepsilon) \;\leq\; \left\lceil \frac{\log(\operatorname{diam}(X)/\varepsilon)}{\log(1/L)} \right\rceil. \tag{13}$$

Thus $1/\log(1/L)$ plays the role of a *statistical memory time scale*: as $L \uparrow 1$, the certified mixing-time upper bound deteriorates.

Coupling intuition and a linear sanity check. The Wasserstein contraction in Theorem 3.1 can be understood through a concrete coupling: run two copies of the reservoir with *the same* driving input $U_0, U_1, \ldots$ but different initial states x_0, x_0'. Uniform contractivity gives the pathwise bound $d(x_{t+1}, x'_{t+1}) = d(F_{U_t}(x_t), F_{U_t}(x'_t)) \leq L\, d(x_t, x'_t)$, hence $d(x_t, x'_t) \leq L^t d(x_0, x'_0)$. This synchronous coupling is precisely the mechanism behind W_1-contraction: it furnishes an explicit transport plan whose expected cost decays geometrically. It is also a useful "sanity check" against the classical linear-stable case. For a linear reservoir $x_{t+1} = Wx_t + Uu_t$ with $\|W\| < 1$ and i.i.d. inputs, the stationary solution is the stable linear filter $x_t = \sum_{k \geq 0} W^k U u_{t-1-k}$ and correlations decay at a geometric rate controlled by $\|W\|^t$. Our mixing-time bound (13) recovers the same qualitative dependence: as the reservoir approaches criticality (here $\|W\| \uparrow 1$), dependence persists longer and the distributional forgetting time scale diverges like $1/\log(1/L)$.

3.3 Mixing of expectations and decay of correlations

Wasserstein mixing immediately controls convergence of expectations of Lipschitz observables via the dual form (7).

Proposition 3.4 (Mixing for Lipschitz observables)**.** *Let $f : X \to \mathbb{R}$ be Lipschitz. Then for any initial distribution μ and $t \geq 0$,*

$$\left| \int f\, d(\mu P^t) - \int f\, d\pi \right| \leq \|f\|_{\mathrm{Lip}}\, W_1(\mu P^t, \pi) \leq \|f\|_{\mathrm{Lip}}\, L^t\, W_1(\mu, \pi). \tag{14}$$

In particular, for $\mu = \delta_x$, $|\mathbb{E}[f(x_t) \mid x_0 = x] - \pi(f)| \leq \|f\|_{\mathrm{Lip}}\, L^t \operatorname{diam}(X)$.

We also record a simple correlation bound under stationarity, which will be used conceptually (and, in the appendix, quantitatively) to motivate effective sample size.

Proposition 3.5 (Geometric decay of correlations)**.** *Assume $x_0 \sim \pi$ so that $(x_t)_{t \geq 0}$ is stationary. For Lipschitz $f, g : X \to \mathbb{R}$ and any $t \geq 0$,*

$$\left|\operatorname{Cov}\big(f(x_0), g(x_t)\big)\right| \;\leq\; \|f\|_{\mathrm{Lip}}\, \|g\|_{\mathrm{Lip}}\, \operatorname{diam}(X)^2\, (1-\kappa)^t. \tag{15}$$

Toward effective sample size. Proposition 3.5 bounds the variance of Lipschitz time averages on the scale $\|f\|_{\mathrm{Lip}}^2 \operatorname{diam}(X)^2/(\kappa T)$. This motivates the concentration-based proxy $T_{\mathrm{eff}} = \kappa T \geq (1-L)T$; a variance-normalized autocorrelation time also depends on $\operatorname{Var}_\pi(f)$ (Section 4).

4 Concentration and Effective Sample Size for Reservoir Features

The previous section established that contractive reservoirs driven by i.i.d. inputs generate a Markov chain whose law contracts in Wasserstein–1 at rate L (Theorem 3.1). This distributional contraction

yields geometric mixing and correlation decay (Corollary 3.3 and Proposition 3.5). We now convert these mixing guarantees into *finite-sample* concentration bounds for the empirical quantities used to train linear readouts, such as the empirical feature mean, Gram matrix, and feature–target correlations. The central message is that geometric mixing induces an explicit *effective sample size* governed by the contraction gap $1 - L$.

4.1 A curvature-based concentration inequality for time averages

Let $(x_t)_{t\geq 0}$ be the reservoir Markov chain on (X, d) induced by Assumption 2.1, and let π denote its unique stationary law (Corollary 3.2). Define the *coarse Ricci curvature lower bound*

$$\kappa \;:=\; 1 - \sup_{x\neq y} \frac{W_1(P(x,\cdot), P(y,\cdot))}{d(x,y)} \;\geq\; 1 - L \;>\; 0, \tag{16}$$

and the *state-space diameter*

$$D_X := \operatorname{diam}(X). \tag{17}$$

The case $D_X = 0$ is deterministic and all deviations below vanish. For $D_X > 0$, we use a bounded-diameter consequence of transport contraction, proved directly in Appendix A. This uses a global scale rather than the finer diffusion and granularity parameters of curvature bounds [14].

Theorem 4.1 (Concentration for Lipschitz time averages)**.** *Assume $x_0 \sim \pi$ (stationarity). Let $f : X \to \mathbb{R}$ be Lipschitz with constant $\|f\|_{\mathrm{Lip}}$ and define the empirical average $\overline{f}_T := \frac{1}{T}\sum_{t=1}^{T} f(x_t)$. For $\|f\|_{\mathrm{Lip}} > 0$, there exist absolute constants $c, C > 0$ such that for all $\varepsilon > 0$ (constant observables have zero deviation),*

$$\mathbb{P}\big(\big|\overline{f}_T - \pi(f)\big| \geq \varepsilon\big) \leq 2\exp\left(- c\,\kappa T \;\min\left\{ \frac{\varepsilon^2}{\|f\|_{\mathrm{Lip}}^2\, D_X^2}, \frac{\varepsilon}{\|f\|_{\mathrm{Lip}}\, D_X} \right\}\right). \tag{18}$$

In particular, for deviations in the sub-Gaussian regime $\varepsilon \leq \|f\|_{\mathrm{Lip}} D_X$,

$$\mathbb{P}\big(\big|\overline{f}_T - \pi(f)\big| \geq \varepsilon\big) \;\leq\; 2\exp\left(-\frac{c\,\kappa T}{\|f\|_{\mathrm{Lip}}^2\, D_X^2}\, \varepsilon^2 \right). \tag{19}$$

Effective sample size. Theorem 4.1 gives the concentration-based proxy $T_{\mathrm{eff}} := \kappa T \geq (1 - L)T$ at scale $\|f\|_{\mathrm{Lip}} D_X$. This is a sufficient, observable-scale guarantee, rather than an identity for the variance-normalized effective sample size. Proposition 3.5 also yields a variance upper bound of order $\|f\|_{\mathrm{Lip}}^2 D_X^2/(\kappa T)$.

Interpreting $T_{\mathrm{eff}} = \kappa T$. The variance identity $\mathrm{Var}(\overline{f}_T) = T^{-2}\sum_{s,t=1}^{T} \mathrm{Cov}(f(x_s), f(x_t))$, together with Proposition 3.5, gives $\mathrm{Var}(\overline{f}_T) \leq 2\|f\|_{\mathrm{Lip}}^2 D_X^2/(\kappa T)$. Theorem 4.1 strengthens this to exponential tails. Using only the state contraction certificate gives $T_{\mathrm{eff}} \geq (1 - L)T$. A larger admissible L weakens this certified lower bound; it does not force the actual curvature or statistical efficiency to deteriorate.

4.2 Vector-valued observables and uniform bounds

Readout training involves vector and matrix statistics built from reservoir features. We first record a convenient reduction: concentration of $\mathbb{R}^p$-valued averages can be obtained by applying Theorem 4.1 to one-dimensional projections and then union bounding over a net on the sphere (a standard argument in high-dimensional probability [27]). We state the resulting bound in a form suited to feature moments.

Proposition 4.2 (Concentration for vector averages)**.** *Assume $x_0 \sim \pi$. Let $\psi : X \to \mathbb{R}^p$ be Lipschitz with respect to ℓ_2:*

$$\|\psi(x) - \psi(y)\|_2 \;\leq\; L_\psi\, d(x,y) \qquad \forall x, y \in X, \tag{20}$$

and define $\overline{\psi}_T = \frac{1}{T}\sum_{t=1}^{T} \psi(x_t)$ and $\pi(\psi) = \int \psi\, d\pi$. Then there exist absolute constants $c, C > 0$ such that for all $\delta \in (0, 1)$,

$$\mathbb{P}\left(\|\overline{\psi}_T - \pi(\psi)\|_2 \;>\; C\, L_\psi\, D_X \left(\sqrt{\frac{p + \log(1/\delta)}{\kappa T}} \;\vee\; \frac{p + \log(1/\delta)}{\kappa T} \right) \right) \leq\; \delta. \tag{21}$$

Remark. The $\sqrt{p}$ dependence is the usual price for controlling ℓ_2 deviation in p dimensions; the dependence on κT is the Markov-chain analogue of i.i.d. concentration, with κ playing the role of an inverse correlation time.

4.3 Reservoir features and empirical moments

We now specialize to the feature statistics used in ridge readout training. Fix a feature map $\phi : X \to \mathbb{R}^p$ and a target function $y : X \to \mathbb{R}$. (These results assume state-observable labels; ESP alone does not make an arbitrary target a function of the current state. Additional target variables require a joint contracting Markov model.)

Assumption 4.3 (Bounded Lipschitz features and targets)**.** *There exist constants $B, L_\phi, Y, L_y < \infty$ such that for all $x, x' \in X$: $\|\phi(x)\|_2 \le B, \|\phi(x) - \phi(x')\|_2 \le L_\phi\, d(x, x'), |y(x)| \le Y, |y(x) - y(x')| \le L_y\, d(x, x')$.*

Define the stationary population moments

$$G \;:=\; \mathbb{E}_\pi[\phi(x)\phi(x)^\top], \qquad b \;:=\; \mathbb{E}_\pi[\phi(x)\, y(x)], \tag{22}$$

and their empirical counterparts from a length-T trajectory:

$$G_T \;:=\; \frac{1}{T}\sum_{t=1}^{T} \phi(x_t)\phi(x_t)^\top,\; b_T \;:=\; \frac{1}{T}\sum_{t=1}^{T} \phi(x_t)\, y(x_t). \tag{23}$$

Feature–target moment. The map $\psi_b(x) := \phi(x)y(x)$ is ℓ_2-Lipschitz with constant $L_b := YL_\phi + BL_y$, since $\|\phi(x)y(x) - \phi(x')y(x')\|_2 \le |y(x)|\|\phi(x) - \phi(x')\|_2 + \|\phi(x')\|_2|y(x) - y(x')|$. Applying Proposition 4.2 yields a direct concentration bound for b_T.

Corollary 4.4 (Concentration of b_T)**.** *Under Assumptions 2.1 and 4.3, and assuming $x_0 \sim \pi$, there exist absolute constants $c, C > 0$ such that for all $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\|b_T - b\|_2 > C\,(YL_\phi + BL_y)\, D_X \left(\sqrt{\frac{p + \log(1/\delta)}{\kappa T}} \;\vee\; \frac{p + \log(1/\delta)}{\kappa T} \right) \right) \le \delta. \tag{24}$$

Gram matrix. To control G_T in operator norm, we reduce to quadratic forms. For any unit vector $v \in \mathbb{S}^{p-1}$ define $g_v(x) := (v^\top \phi(x))^2$. By Assumption 4.3, $|v^\top \phi(x)| \le \|\phi(x)\|_2 \le B$, and

$$|g_v(x) - g_v(x')| = |v^\top(\phi(x) - \phi(x'))| \cdot |v^\top(\phi(x) + \phi(x'))| \le (L_\phi d(x, x')) \cdot (2B). \tag{25}$$

so $\|g_v\|_{\mathrm{Lip}} \le 2BL_\phi$. Since $v^\top(G_T - G)v = \frac{1}{T}\sum_{t=1}^{T} g_v(x_t) - \mathbb{E}_\pi[g_v]$, Theorem 4.1 yields sub-Gaussian concentration for each fixed v, and a standard ε-net argument lifts this to operator norm control [27] (details in Appendix A).

Theorem 4.5 (Concentration of G_T in operator norm)**.** *Under Assumptions 2.1 and 4.3, and assuming $x_0 \sim \pi$, there exist absolute constants $c, C > 0$ such that for all $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\|G_T - G\|_{\mathrm{op}} \;>\; C\,(BL_\phi)\, D_X \left(\sqrt{\frac{p + \log(1/\delta)}{\kappa T}} \;\vee\; \frac{p + \log(1/\delta)}{\kappa T} \right) \right) \le\; \delta. \tag{26}$$

Nonstationary initialization and burn-in. For arbitrary $x_0 \sim \mu$, synchronous coupling with a stationary copy bounds the difference of length-T averages after t_0 burn-in steps by $\|f\|_{\mathrm{Lip}} D_X L^{t_0+1}(1 - L^T)/(T(1 - L))$ almost surely. Thus the stationary concentration bound remains valid after adding this deterministic amount to the deviation threshold. Appendix A gives the full statement, also applicable to feature and Gram observables.

5 Finite-Sample Readout Guarantees and Design Tradeoffs

We now translate the concentration bounds of Section 4 into finite-sample guarantees for ridge-trained linear readouts. The results explicitly exhibit how the reservoir contraction gap $\kappa \ge 1 - L$ determines an effective sample size $T_{\mathrm{eff}} = \kappa T$ and therefore the data length required for stable training. This yields a quantitative design tradeoff: a larger contraction certificate L weakens the guaranteed mixing rate and the sufficient training-length bound.

5.1 Ridge readout training and population targets

Given features $\phi(x) \in \mathbb{R}^p$ and target $y(x) \in \mathbb{R}$, consider the regularized population risk $\mathcal{R}_\lambda(w) := \mathbb{E}_\pi\left[(w^\top \phi(x) - y(x))^2\right] + \lambda\|w\|_2^2,\ \lambda > 0$, and the empirical risk based on a length-T trajectory: $\widehat{\mathcal{R}}_{\lambda,T}(w) := \frac{1}{T}\sum_{t=1}^{T}(w^\top \phi(x_t) - y(x_t))^2 + \lambda\|w\|_2^2$. Both objectives are strictly convex quadratics. The associated population and empirical ridge solutions are $w_\lambda^\star := \arg\min_w \mathcal{R}_\lambda(w),\ \widehat{w}_{\lambda,T} := \arg\min_w \widehat{\mathcal{R}}_{\lambda,T}(w)$, with the familiar closed forms [12] $w_\lambda^\star = (G + \lambda I)^{-1} b,\ \widehat{w}_{\lambda,T} = (G_T + \lambda I)^{-1} b_T$, where G, b, G_T, b_T are defined in (22)–(23).

5.2 Parameter error bounds via moment concentration

We first bound $\|\widehat{w}_{\lambda,T} - w_\lambda^\star\|_2$ in terms of deviations of (G_T, b_T) from (G, b).

Proposition 5.1 (Perturbation bound for ridge solutions)**.** *Let $\lambda > 0$ and suppose $\|G_T - G\|_{\mathrm{op}} \le \lambda/2$. Then*

$$\|\widehat{w}_{\lambda,T} - w_\lambda^\star\|_2 \ \le\ \frac{2}{\lambda}\,\|b_T - b\|_2 \ +\ \frac{2}{\lambda}\,\|G_T - G\|_{\mathrm{op}}\,\|w_\lambda^\star\|_2. \tag{27}$$

Theorem 5.2 (Finite-sample ridge estimation under reservoir mixing)**.** *Under Assumptions 2.1 and 4.3, and assuming $x_0 \sim \pi$, there exist absolute constants $c, C > 0$ such that for all $\delta \in (0,1)$, with probability at least $1 - \delta$ the following hold simultaneously. Define the deviation scale $\eta_{T,\delta} := \left(\sqrt{\frac{p+\log(2/\delta)}{\kappa T}} \ \vee\ \frac{p+\log(2/\delta)}{\kappa T}\right)$.*

1. *(**Moment deviations**)*

$$\|b_T - b\|_2 \le C\,(YL_\phi + BL_y)\,D_X\,\eta_{T,\delta}, \|G_T - G\|_{\mathrm{op}} \le C\,(BL_\phi)\,D_X\,\eta_{T,\delta}.$$

2. *(**Ridge parameter error**) If in addition $\|G_T - G\|_{\mathrm{op}} \le \lambda/2$ (which is implied by the previous line for sufficiently large T), then*

$$\|\widehat{w}_{\lambda,T} - w_\lambda^\star\|_2 \le \frac{C}{\lambda}\,(YL_\phi + BL_y)\,D_X\,\eta_{T,\delta} \quad +\ \frac{C}{\lambda^2}\,(B^2 YL_\phi)\,D_X\,\eta_{T,\delta}. \tag{28}$$

Excess risk under the stationary law. Since $\mathcal{R}_\lambda(w)$ is a quadratic with Hessian $2(G + \lambda I)$, the regularized excess risk satisfies

$$\mathcal{R}_\lambda(\widehat{w}_{\lambda,T}) - \mathcal{R}_\lambda(w_\lambda^\star) = \|\widehat{w}_{\lambda,T} - w_\lambda^\star\|_{G+\lambda I}^2 \le (\|G\|_{\mathrm{op}} + \lambda)\,\|\widehat{w}_{\lambda,T} - w_\lambda^\star\|_2^2. \tag{29}$$

and $\|G\|_{\mathrm{op}} \le \mathbb{E}_\pi[\|\phi(x)\|_2^2] \le B^2$ by Assumption 4.3. Plugging (28) into (29) gives a regularized-risk bound of order $K_\lambda \eta_{T,\delta}^2$, where K_λ is defined below. The $O((p + \log(1/\delta))/(\kappa T))$ rate applies when $p + \log(2/\delta) \le \kappa T$, for fixed feature constants and $\lambda > 0$.

5.3 Sample complexity and a design rule

Theorem 5.2 gives an explicit sufficient design rule. Define

$$A_\lambda := D_X\left(\frac{YL_\phi + BL_y}{\lambda} + \frac{B^2 YL_\phi}{\lambda^2}\right), \qquad K_\lambda := (B^2 + \lambda)A_\lambda^2,$$

and put $q = p + \log(2/\delta)$. On the stated moment event, the parameter error is at most $CA_\lambda \eta_{T,\delta}$ and the regularized excess risk is at most $CK_\lambda \eta_{T,\delta}^2$. For target risk $\varepsilon > 0$, a sufficient condition ensuring the Gaussian regime and the Gram event is

$$T\ \ge\ \frac{Cq}{\kappa}\max\left\{1, \left(\frac{BL_\phi D_X}{\lambda}\right)^2, \frac{K_\lambda}{\varepsilon}\right\}. \tag{30}$$

Replacing κ by $1 - L$ gives a conservative sufficient bound. For fixed diameter, feature constants, and regularization, a weaker contraction certificate raises the certified data requirement; these upper bounds do not establish a necessary penalty or a universal memory improvement as L increases.

How to use the bounds as a tuning guide. The sample-complexity scaling (30) suggests a concrete way to "budget" reservoir criticality against dataset length. For a given feature dimension p and training length T, the regime in which ridge readout is well-conditioned is when κT dominates the effective dimension, i.e. when $\kappa T \gg p$ (up to logarithmic confidence factors and Lipschitz/scale constants). Together with the scale and regularization requirements in (30), this makes the Gram deviation small enough for the stated ridge guarantee. Failure of this sufficient condition does not imply instability. Regularization provides a second lever: the condition $\|G_T - G\|_{\mathrm{op}} \leq \lambda/2$ in Proposition 5.1 shows that one can compensate for slower mixing (smaller κ) by increasing λ, at the price of additional bias in the readout. Operationally, this frames hyperparameter selection as a three-way tradeoff between (i) contraction and memory (L), (ii) statistical efficiency ($T_{\mathrm{eff}} = \kappa T$), and (iii) regularization strength (λ). This perspective clarifies why "edge-of-stability" tuning can succeed when abundant data are available, but can become brittle in short-data regimes: near-critical tuning may create richer dynamical features, while its conservative concentration certificate can require more data. Equation (30) makes a sufficient statistical cost explicit. Decreasing a verified L or increasing λ can improve this certificate, provided the diameter and feature constants are also tracked. Spectral-radius scaling alone does not certify the required operator-norm contraction, and larger regularization also changes the population target.

6 Conclusion, Limitations, and Outlook

We presented a transport-based statistical theory for contractive echo state networks. The core observation is that uniform state contraction induces contraction of the associated Markov operator in Wasserstein–1 distance, yielding a positive curvature gap $\kappa \geq 1 - L$ and therefore explicit geometric mixing of the reservoir state distribution. Using a direct bounded-diameter curvature argument, motivated by transport concentration [14], we derived finite-T deviation bounds for time-averaged reservoir features and for the empirical moments that govern ridge readout training. This leads to an explicit *effective sample size* $T_{\mathrm{eff}} = \kappa T$, and to finite-sample readout guarantees and sample-complexity laws that quantify a principled design tradeoff: near-critical tuning may increase dynamical memory while weakening the certified mixing rate and increasing the sufficient training length.

A complementary metric to capacity and memory. Reservoir practice often reports memory capacity or task-specific accuracy as a proxy for "how good" a reservoir is, while criticality heuristics are justified informally by increased dynamical memory. Our results suggest reporting a second, conceptually orthogonal quantity alongside such metrics: a *mixing certificate* (e.g. $\kappa \geq 1 - L$ or a derived $t_{\mathrm{mix}}(\varepsilon)$), which controls a sufficient concentration-based proxy for the information in the trajectory. From this angle, a reservoir design is not only a feature generator but also a *dependent-data sampler*; stability parameters simultaneously govern expressiveness and statistical efficiency. Making this dual role explicit can help practitioners separate "better features" from "noisier estimation," and it provides a principled way to decide when near-critical tuning is beneficial versus when it is data-limited.

Limitations and Outlook. Our results are deliberately scoped to a clean regime to obtain transparent rates. First, we assume *uniform contractivity* in the reservoir state. Every verified $L < 1$ is covered, although the bounds may deteriorate as $L \uparrow 1$; regimes without such a certificate require different tools (e.g., nonuniform hyperbolicity or random dynamical systems methods). Second, we treat i.i.d. finite-alphabet driving; Appendix C proves an augmented-state extension for contractive Markov inputs and a weaker Lipschitz dependence bound for general mixing inputs; continuous i.i.d. inputs admit the same integral-kernel argument under uniform contraction and compact forward invariance. Third, our concentration bounds are stated for bounded Lipschitz observables, which is natural in the transport framework but may be conservative for heavy-tailed targets or non-Lipschitz features. Several directions appear promising. (i) *Dependent and structured inputs*: extending the effective-sample-size law to Markov, mixing, or adversarial-but-stable input processes via augmented-state contraction. (ii) *Beyond linear readouts*: combining transport mixing with complexity measures for nonlinear readouts or kernelized ridge to obtain analogous trainability laws. (iii) *Design at fixed data budget*: using the explicit dependence on $(1 - L)$ and λ to formalize optimal operating points that balance dynamical memory against statistical efficiency for a given training length. (iv) *Nonuniform contraction*: replacing a uniform L by pathwise or average contraction rates to capture

heterogeneous "forgetting speeds" across input regimes while preserving quantitative mixing and learning guarantees.

Impact Statement. These guarantees may support more reliable and data-efficient reservoir readouts by making dependence and sufficient training lengths explicit. Applying them beyond verified contraction, target, or input assumptions could create unjustified confidence in predictions. Reporting those assumptions and the diameter, feature, and regularization constants helps limit this risk.

Acknowledgments and Disclosure of Funding

Funding. This research received no specific funding from any public, commercial, or not-for-profit funding agency.

Competing interests. The authors declare that they have no competing financial interests related to this work.

References

[1] N. Bertschinger and T. Natschläger. Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, 16(7):1413–1436, 2004.

[2] J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar. Information processing capacity of dynamical systems. *Scientific Reports*, 2:514, 2012.

[3] R. L. Dobrushin. Central limit theorem for nonstationary markov chains. i. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

[4] X.-Y. Duan, X. Ying, S.-Y. Leng, J. Kurths, W. Lin, and H.-F. Ma. Embedding theory of reservoir computing and reducing reservoir network using time delays. *Phys. Rev. Res.*, 5:L022041, May 2023.

[5] S. Ganguli, D. Huh, and H. Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105(48):18970–18975, 2008.

[6] L. Gonon, L. Grigoryeva, and J.-P. Ortega. Risk bounds for reservoir computing. *Journal of Machine Learning Research*, 21(240):1–61, 2020.

[7] L. Gonon, L. Grigoryeva, and J.-P. Ortega. Approximation bounds for random neural networks and reservoir systems. *The Annals of Applied Probability*, 33(1):28–69, 2023.

[8] L. Gonon, L. Grigoryeva, and J.-P. Ortega. Infinite-dimensional reservoir computing. *Neural Networks*, 179:106486, 2024.

[9] L. Grigoryeva and J.-P. Ortega. Echo state networks are universal. *Neural Networks*, 108:495–508, 2018.

[10] X. Han, Y. Zhao, and M. Small. A tighter generalization bound for reservoir computing. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(4):043115, 04 2022.

[11] A. G. Hart, J. L. Hook, and J. H. Dawes. Echo state networks trained by tikhonov least squares are l2(μ) approximators of ergodic dynamical systems. *Physica D: Nonlinear Phenomena*, 421:132882, 2021.

[12] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[13] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks. *GMD Technical Report*, 148:1–47, 2001.

[14] A. Joulin and Y. Ollivier. Curvature, concentration and error estimates for markov chain monte carlo. *The Annals of Probability*, 38(6):2418–2442, 2010.

[15] A. Kontorovich and K. Ramanan. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, Nov. 2008.

[16] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

[17] M. Lukoševičius. A practical guide to applying echo state networks. In *Neural Networks: Tricks of the Trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 659–686. Springer, 2012.

[18] M. Lukoševičius and H. Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

[19] W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560, 2002.

[20] G. Manjunath and J.-P. Ortega. Transport in reservoir computing. *Physica D: Nonlinear Phenomena*, 449:133744, 2023.

[21] K. Marton. A measure concentration inequality for contracting markov chains. *Geometric and Functional Analysis*, 6(3):556–571, 1996.

[22] Y. Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

[23] D. Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20(79):1–32, 2015.

[24] P.-M. Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *The Annals of Probability*, 28(1):416 – 461, 2000.

[25] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[26] J. A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1–2):1–230, May 2015.

[27] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[28] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

[29] O. L. White, D. D. Lee, and H. Sompolinsky. Short-term memory in orthogonal neural networks. *Physical Review Letters*, 92(14):148102, 2004.

A Complete Proofs

This appendix collects full proofs of all statements in the main paper. Throughout, $X \subset \mathbb{R}^N$ is compact with metric $d(x, y) = \|x-y\|$, Σ is finite, and the reservoir update maps $\{F_s : X \to X\}_{s\in\Sigma}$ satisfy Assumption 2.1. The input symbols $(U_t)_{t\geq 0}$ are i.i.d. with law ν, and the induced Markov kernel P is given by (4). We write $\mathcal{P}(X)$ for Borel probability measures on X and W_1 for the Wasserstein–1 distance (6).

A.1 Transport preliminaries

We repeatedly use Kantorovich–Rubinstein duality [28]:

$$W_1(\mu, \eta) = \sup_{\|f\|_{\mathrm{Lip}}\leq 1} \left\{ \int f \, d\mu - \int f \, d\eta \right\}. \tag{31}$$

Proof of Lemma 2.2 (pushforward contracts Wasserstein). Let $T : X \to X$ be K-Lipschitz. Fix $\mu, \eta \in \mathcal{P}(X)$ and let $\gamma \in \Gamma(\mu, \eta)$ be any coupling. Define the pushforward coupling $\widetilde{\gamma} := (T, T)_\sharp \gamma$, i.e. the law of $(T(X), T(Y))$ when $(X, Y) \sim \gamma$. Then $\widetilde{\gamma} \in \Gamma(T_\sharp\mu, T_\sharp\eta)$ and

$$\int d(x', y') \, \widetilde{\gamma}(dx', dy') = \int d(T(x), T(y)) \, \gamma(dx, dy) \leq K \int d(x, y) \, \gamma(dx, dy).$$

Taking the infimum over $\gamma \in \Gamma(\mu, \eta)$ yields $W_1(T_\sharp\mu, T_\sharp\eta) \leq K \, W_1(\mu, \eta)$. □

Proof of Lemma 2.3 (mixtures preserve contraction). Let $\{T_i\}_{i\in I}$ be measurable maps and $(p_i)_{i\in I}$ a probability mass function. Fix $\mu, \eta \in \mathcal{P}(X)$. For each i, choose a coupling $\gamma_i \in \Gamma((T_i)_\sharp\mu, (T_i)_\sharp\eta)$ (e.g. an optimal one). Define the mixture coupling $\gamma := \sum_{i\in I} p_i \, \gamma_i$ (finite or countable sum). Then γ is a probability measure on $X \times X$, and its marginals are

$$\gamma(\cdot \times X) = \sum_i p_i (T_i)_\sharp\mu, \qquad \gamma(X \times \cdot) = \sum_i p_i (T_i)_\sharp\eta,$$

so $\gamma \in \Gamma(\sum_i p_i (T_i)_\sharp\mu, \ \sum_i p_i (T_i)_\sharp\eta)$. Therefore,

$$W_1\left(\sum_i p_i (T_i)_\sharp\mu, \ \sum_i p_i (T_i)_\sharp\eta \right) \leq \int d(x, y) \, \gamma(dx, dy) = \sum_i p_i \int d(x, y) \, \gamma_i(dx, dy).$$

Taking γ_i optimal (or taking infimum over γ_i) yields the claim. □

A.2 Proofs for Section 3

Proof of Theorem 3.1. Recall from (5) that

$$\mu P = \sum_{s\in\Sigma} \nu(s) \, (F_s)_\sharp\mu, \qquad \eta P = \sum_{s\in\Sigma} \nu(s) \, (F_s)_\sharp\eta.$$

By Lemma 2.3,

$$W_1(\mu P, \eta P) \leq \sum_{s\in\Sigma} \nu(s) \, W_1((F_s)_\sharp\mu, (F_s)_\sharp\eta).$$

By Lemma 2.2 and Assumption 2.1(2), $W_1((F_s)_\sharp\mu, (F_s)_\sharp\eta) \leq \mathrm{Lip}(F_s) \, W_1(\mu, \eta) \leq L \, W_1(\mu, \eta)$. Hence

$$W_1(\mu P, \eta P) \leq \sum_{s\in\Sigma} \nu(s) \, L \, W_1(\mu, \eta) = L \, W_1(\mu, \eta),$$

which is (8). For (9), apply (8) with $\mu = \delta_x$ and $\eta = \delta_y$. □

Lemma A.1 (Completeness of $(\mathcal{P}(X), W_1)$ on compact X)**.** *If (X, d) is compact metric, then $(\mathcal{P}(X), W_1)$ is complete and W_1 metrizes weak convergence on $\mathcal{P}(X)$.*

Proof. This is standard; see [28]. For completeness: if (μ_n) is Cauchy in W_1, then it is tight (trivial on compact X). By Prokhorov, extract a weakly convergent subsequence $\mu_{n_k} \Rightarrow \mu$. On compact spaces, weak convergence implies convergence in W_1 because the cost is bounded and continuous; hence $W_1(\mu_{n_k}, \mu) \to 0$. Since (μ_n) is Cauchy, the full sequence converges to μ in W_1. □

Proof of Corollary 3.2. Define the operator $\mathcal{T} : \mathcal{P}(X) \to \mathcal{P}(X)$ by $\mathcal{T}(\mu) = \mu P$. By Theorem 3.1, $\mathcal{T}$ is a strict contraction on $(\mathcal{P}(X), W_1)$ with contraction constant $L \in (0, 1)$. By Lemma A.1, $(\mathcal{P}(X), W_1)$ is complete, so Banach's fixed-point theorem implies that $\mathcal{T}$ has a unique fixed point $\pi \in \mathcal{P}(X)$ satisfying $\pi P = \pi$. □

Proof of Corollary 3.3. Iterating Theorem 3.1 gives, for any $t \geq 0$,

$$W_1(\mu P^t, \eta P^t) \leq L^t \, W_1(\mu, \eta).$$

Taking $\eta = \pi$ and using $\pi P^t = \pi$ yields (11). For (12), note that for any $x \in X$,

$$W_1(\delta_x, \pi) = \inf_{\gamma \in \Gamma(\delta_x, \pi)} \int d(x, y) \, \gamma(dx, dy) = \int d(x, y) \, \pi(dy) \leq \operatorname{diam}(X),$$

so $W_1(\delta_x P^t, \pi) \leq L^t \operatorname{diam}(X)$. □

Proof of Proposition 3.4. Let $f : X \to \mathbb{R}$ be Lipschitz. If $\|f\|_{\mathrm{Lip}} = 0$ the statement is trivial, so assume $\|f\|_{\mathrm{Lip}} > 0$ and define $\widetilde{f} = f / \|f\|_{\mathrm{Lip}}$, so that $\|\widetilde{f}\|_{\mathrm{Lip}} \leq 1$. Then by Kantorovich–Rubinstein duality (31),

$$\left| \int f \, d(\mu P^t) - \int f \, d\pi \right| = \|f\|_{\mathrm{Lip}} \left| \int \widetilde{f} \, d(\mu P^t) - \int \widetilde{f} \, d\pi \right| \leq \|f\|_{\mathrm{Lip}} \, W_1(\mu P^t, \pi).$$

Applying Corollary 3.3 yields (14). The specialization $\mu = \delta_x$ follows from (12). □

Proof of Proposition 3.5. Assume stationarity: $x_0 \sim \pi$. Write $\bar{f} = f - \pi(f)$ and $\bar{g} = g - \pi(g)$. By the Markov property and stationarity,

$$\operatorname{Cov}(f(x_0), g(x_t)) = \mathbb{E}_\pi\big[\bar{f}(x_0) \, \bar{g}(x_t)\big] = \mathbb{E}_\pi\big[\bar{f}(x_0) \, (P^t \bar{g})(x_0)\big] .$$

Hence

$$|\operatorname{Cov}(f(x_0), g(x_t))| \leq \|\bar{f}\|_\infty \, \|P^t \bar{g}\|_\infty.$$

We bound each factor using Lipschitzness on compact X:

$$\|\bar{f}\|_\infty = \sup_x |f(x) - \pi(f)| \leq \sup_{x,y} |f(x) - f(y)| \leq \|f\|_{\mathrm{Lip}} \operatorname{diam}(X).$$

For the second factor, note that for any $x \in X$,

$$|(P^t \bar{g})(x)| = \left| \int g \, d(\delta_x P^t) - \int g \, d\pi \right| \leq \|g\|_{\mathrm{Lip}} \, W_1(\delta_x P^t, \pi) \leq \|g\|_{\mathrm{Lip}} \, (1 - \kappa)^t \operatorname{diam}(X),$$

where the kernel contraction coefficient is $1 - \kappa$ by definition, so its t-step coefficient is at most $(1 - \kappa)^t \leq L^t$. Combining the two bounds yields (15). □

A.3 Proofs for Section 4

A.3.1 A transport-curvature concentration inequality

We use the following bounded-diameter consequence of Wasserstein contraction. Its direct proof keeps the curvature dependence explicit; the finer diffusion-scale inequalities in [14] require different scale factors.

Lemma A.2 (Bounded-diameter curvature concentration)**.** *Let (x_t) be a stationary Markov chain on a compact metric space of diameter $D_X > 0$, with stationary law π, and suppose*

$$W_1(P(x, \cdot), P(y, \cdot)) \leq (1 - \kappa) d(x, y), \qquad 0 < \kappa \leq 1.$$

For every Lipschitz f with $K := \|f\|_{\mathrm{Lip}} > 0$, every integer $T \geq 1$, and every $\varepsilon > 0$,

$$\mathbb{P}\left(\left| T^{-1} \sum_{t=1}^{T} f(x_t) - \pi(f) \right| \geq \varepsilon \right) \leq 2 \exp\left(-c_0 \kappa T \frac{\varepsilon^2}{K^2 D_X^2} \right), \qquad c_0 = (1 - e^{-1})^2.$$

Constant observables have zero deviation.

Proof. Set $m = \lceil 1/\kappa \rceil \le 2/\kappa$, and $a = (1-\kappa)^m \le e^{-1}$. The skeleton kernel P^m contracts Lipschitz seminorms by a. For any length-n stationary skeleton $(Y_i)_{i=1}^n$, let $S = \sum_{i=1}^n (f(Y_i) - \pi(f))$. In its Doob martingale, the conditional future-sum function at time i is $H_i = \sum_{j=0}^{n-i} (P^m)^j f$, whose Lipschitz constant is at most $K/(1-a)$. Each martingale difference therefore has conditional range of length at most $KD_X/(1-a)$, including the first difference. Conditional Hoeffding's lemma and iteration give, for every real θ,

$$\mathbb{E}e^{\theta S} \le \exp\left(\frac{n\theta^2 K^2 D_X^2}{8(1-a)^2}\right).$$

Partition $\{1, \dots, T\}$ into its m residue classes. Write S_r for the centered sum in class r and n_r for its size; empty classes contribute zero. Hölder's inequality, which requires no independence between classes, yields

$$\begin{aligned}\mathbb{E}\exp\left(\frac{\theta}{T}\sum_{r=1}^m S_r\right) &\le \prod_{r=1}^m \left(\mathbb{E}e^{m\theta S_r/T}\right)^{1/m} \\ &\le \exp\left(\frac{m\theta^2 K^2 D_X^2}{8T(1-a)^2}\right) \le \exp\left(\frac{\theta^2 K^2 D_X^2}{4\kappa T(1-e^{-1})^2}\right),\end{aligned}$$

where $\sum_r n_r = T$. Chernoff optimization for the two tails proves the claim. □

Proof of Theorem 4.1. Theorem 3.1 and the definition of κ give kernel contraction by $1 - \kappa \le L$. Apply Lemma A.2. Its all-deviation quadratic bound implies (18), since $\min\{u^2, u\} \le u^2$ for $u \ge 0$, and also implies (19). □

A.3.2 Vector concentration

We will use a standard ε-net argument for suprema over the unit sphere [27].

Lemma A.3 (ε-net reduction for the Euclidean norm)**.** *Let $z \in \mathbb{R}^p$, $0 < \varepsilon < 1$, and let $\mathcal{N}_\varepsilon \subset \mathbb{S}^{p-1}$ be an ε-net of the unit sphere in ℓ_2, i.e. for every $v \in \mathbb{S}^{p-1}$ there exists $u \in \mathcal{N}_\varepsilon$ with $\|v - u\|_2 \le \varepsilon$. Then*

$$\|z\|_2 = \sup_{v \in \mathbb{S}^{p-1}} v^\top z \;\le\; \frac{1}{1-\varepsilon} \sup_{u \in \mathcal{N}_\varepsilon} u^\top z.$$

Moreover, there exists an ε-net with cardinality $|\mathcal{N}_\varepsilon| \le (1 + 2/\varepsilon)^p$.

Proof. Fix z. Let $v^\star = z/\|z\|_2$ (if $z = 0$ the claim is trivial). Choose $u \in \mathcal{N}_\varepsilon$ with $\|u - v^\star\|_2 \le \varepsilon$. Then

$$u^\top z = u^\top(\|z\|_2 v^\star) = \|z\|_2\, u^\top v^\star \ge \|z\|_2\,(1 - \|u - v^\star\|_2) \ge (1-\varepsilon)\|z\|_2,$$

so $\|z\|_2 \le (1-\varepsilon)^{-1} \sup_{u \in \mathcal{N}_\varepsilon} u^\top z$. The existence of an ε-net with $|\mathcal{N}_\varepsilon| \le (1 + 2/\varepsilon)^p$ is standard [27]. □

Proof of Proposition 4.2. Let $z := \overline{\psi}_T - \pi(\psi) \in \mathbb{R}^p$. Fix $\varepsilon = 1/2$ and let $\mathcal{N} := \mathcal{N}_{1/2}$ be a $1/2$-net of $\mathbb{S}^{p-1}$ with $|\mathcal{N}| \le 5^p$ (Lemma A.3). Then

$$\|z\|_2 \le 2 \sup_{v \in \mathcal{N}} v^\top z. \tag{32}$$

For any fixed $v \in \mathbb{S}^{p-1}$, define the scalar observable $f_v(x) := v^\top \psi(x)$. By (20),

$$|f_v(x) - f_v(y)| \le \|v\|_2\, \|\psi(x) - \psi(y)\|_2 \le L_\psi\, d(x, y),$$

so $\|f_v\|_{\mathrm{Lip}} \le L_\psi$. Applying Theorem 4.1 (or Lemma A.2) to f_v yields, for any $t > 0$,

$$\mathbb{P}\big(|v^\top z| \ge t\big) \le 2\exp\left(-c\,\kappa T\, \min\left\{\frac{t^2}{L_\psi^2\, D_X^2},\; \frac{t}{L_\psi\, D_X}\right\}\right).$$

Using the all-deviation quadratic bound of Lemma A.2, a union bound over $\mathcal{N}$ gives

$$\mathbb{P}\left(\sup_{v \in \mathcal{N}} |v^\top z| \ge t\right) \le 2|\mathcal{N}| \exp\left(-c\,\kappa T\, \frac{t^2}{L_\psi^2 D_X^2}\right) \le 2\exp\left(p \log 5 - c\,\kappa T\, \frac{t^2}{L_\psi^2 D_X^2}\right).$$

Choose

$$t := C\, L_\psi\, D_X \left(\sqrt{\frac{p + \log(1/\delta)}{\kappa T}} \;\vee\; \frac{p + \log(1/\delta)}{\kappa T} \right)$$

with C large enough so that the union bound gives probability at most δ. Combining with (32) yields (21) (absorbing factors into constants). □

Proof of Corollary 4.4. Define $\psi_b(x) := \phi(x)\, y(x) \in \mathbb{R}^p$. Under Assumption 4.3,

$$\begin{aligned} \|\psi_b(x) - \psi_b(x')\|_2 &= \|\phi(x)y(x) - \phi(x')y(x')\|_2 \\ &\le |y(x)|\|\phi(x) - \phi(x')\|_2 + \|\phi(x')\|_2 |y(x) - y(x')| \\ &\le (Y L_\phi + B L_y) d(x, x'). \end{aligned}$$

Hence ψ_b is Lipschitz with $L_{\psi_b} = Y L_\phi + B L_y$ and bounded by $\|\psi_b(x)\|_2 \le BY$. Applying Proposition 4.2 with $\psi = \psi_b$ gives (24). □

A.3.3 Operator-norm concentration for the Gram matrix

We use a standard net argument for symmetric matrices [27].

Lemma A.4 (ε-net reduction for the operator norm)**.** *Let $M \in \mathbb{R}^{p\times p}$ be symmetric and let $\mathcal{N}_\varepsilon$ be an ε-net of $\mathbb{S}^{p-1}$. Then*

$$\begin{aligned} \|M\|_{\mathrm{op}} &= \sup_{v \in \mathbb{S}^{p-1}} |v^\top M v| \\ &\le \frac{1}{1 - 2\varepsilon} \sup_{u \in \mathcal{N}_\varepsilon} |u^\top M u|, \qquad \varepsilon \in (0, 1/2). \end{aligned}$$

Proof. See [27]. For completeness: for any $v \in \mathbb{S}^{p-1}$ pick $u \in \mathcal{N}_\varepsilon$ with $\|u - v\|_2 \le \varepsilon$. Then using symmetry and Cauchy–Schwarz,

$$\begin{aligned} |v^\top M v - u^\top M u| &= |(v-u)^\top M v + u^\top M (v-u)| \\ &\le \big(\|v-u\|_2 \|v\|_2 + \|u\|_2 \|v-u\|_2\big) \|M\|_{\mathrm{op}} \\ &\le 2\varepsilon \|M\|_{\mathrm{op}}. \end{aligned}$$

Thus $|v^\top M v| \le |u^\top M u| + 2\varepsilon \|M\|_{\mathrm{op}}$. Taking sup over v and rearranging yields the claim. □

Proof of Theorem 4.5. Let $M := G_T - G$. Then M is symmetric and

$$\begin{aligned} \|M\|_{\mathrm{op}} &= \sup_{v \in \mathbb{S}^{p-1}} |v^\top M v| \\ &= \sup_{v \in \mathbb{S}^{p-1}} \left| \frac{1}{T} \sum_{t=1}^{T} (v^\top \phi(x_t))^2 - \mathbb{E}_\pi[(v^\top \phi(x))^2] \right|. \end{aligned}$$

Fix $v \in \mathbb{S}^{p-1}$ and define $g_v(x) := (v^\top \phi(x))^2$. By Assumption 4.3, $|v^\top \phi(x)| \le \|\phi(x)\|_2 \le B$, and for $x, x' \in X$,

$$\begin{aligned} |g_v(x) - g_v(x')| &= |v^\top(\phi(x) - \phi(x'))|\, |v^\top(\phi(x) + \phi(x'))| \\ &\le \|\phi(x) - \phi(x')\|_2 \big(\|\phi(x)\|_2 + \|\phi(x')\|_2\big) \\ &\le (L_\phi d(x, x'))(2B). \end{aligned}$$

so $\|g_v\|_{\mathrm{Lip}} \le 2BL_\phi$. Applying Theorem 4.1 to g_v gives, for any $t > 0$,

$$\mathbb{P}\left(\left| \frac{1}{T} \sum_{t=1}^{T} g_v(x_t) - \pi(g_v) \right| \ge t \right) \le 2 \exp\left(-c\, \kappa T\, \min\left\{ \frac{t^2}{(2BL_\phi)^2 D_X^2},\; \frac{t}{(2BL_\phi) D_X} \right\} \right).$$

Now choose $\varepsilon = 1/4$ and let $\mathcal{N}$ be a $1/4$-net of $\mathbb{S}^{p-1}$ with $|\mathcal{N}| \le 9^p$. By Lemma A.4,

$$\|M\|_{\mathrm{op}} \le 2 \sup_{u \in \mathcal{N}} |u^\top M u|.$$

The all-deviation quadratic bound of Lemma A.2 and a union bound over $\mathcal{N}$ yield

$$\mathbb{P}\left(\sup_{u\in\mathcal{N}} |u^\top M u| \geq t\right) \leq 2|\mathcal{N}| \exp\left(-c\kappa T \frac{t^2}{(2BL_\phi)^2 D_X^2}\right) \leq 2\exp\left(p\log 9 - c\kappa T \frac{t^2}{(2BL_\phi)^2 D_X^2}\right).$$

Set

$$t := C\,(BL_\phi)\,D_X\left(\sqrt{\frac{p+\log(1/\delta)}{\kappa T}} \vee \frac{p+\log(1/\delta)}{\kappa T}\right)$$

with C large enough so that the union bound gives probability at most δ. Then $\sup_{u\in\mathcal{N}} |u^\top M u| \leq t$ with prob. $\geq 1-\delta$, hence $\|M\|_{\mathrm{op}} \leq 2t$, which is (26) after absorbing constants. □

A.3.4 Nonstationary initialization and burn-in (details)

Proposition A.5 (Burn-in bias and concentration)**.** *Let $x_0 \sim \mu$ be arbitrary and let f be Lipschitz. The expectation bound remains*

$$|\mathbb{E}_\mu f(x_{t_0}) - \pi(f)| \leq \|f\|_{\mathrm{Lip}} L^{t_0} W_1(\mu,\pi).$$

For $\overline{f}_{T,t_0} = T^{-1}\sum_{k=1}^{T} f(x_{t_0+k})$, put

$$a_{T,t_0} = \frac{\|f\|_{\mathrm{Lip}} D_X L^{t_0+1}(1-L^T)}{T(1-L)}.$$

Then the right-hand side of (18) *bounds $\mathbb{P}_\mu(|\overline{f}_{T,t_0} - \pi(f)| \geq \varepsilon + a_{T,t_0})$ for every $\varepsilon > 0$.*

Proof. The expectation bound follows from Proposition 3.4. Couple $x_0 \sim \mu$ with $x_0' \sim \pi$ and drive both reservoirs by the same i.i.d. inputs, independent of the initial pair. Uniform contraction gives $d(x_j, x_j') \leq D_X L^j$ almost surely. Summing the resulting observable differences proves $|\overline{f}_{T,t_0} - \overline{f}'_{T,t_0}| \leq a_{T,t_0}$. The second average is stationary, so Theorem 4.1 and the triangle inequality prove the probability statement. The same coupling applies to vector averages with Lipschitz constant L_ψ, and to Gram quadratic forms with constant $2BL_\phi$ before the net argument. □

A.4 Proofs for Section 5

Proof of Proposition 5.1. Write $A := G_T + \lambda I$ and $B := G + \lambda I$, so $\widehat{w} = A^{-1}b_T$ and $w^\star = B^{-1}b$. Then

$$\widehat{w} - w^\star = A^{-1}(b_T - b) + (A^{-1} - B^{-1})b.$$

Using the resolvent identity $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1} = A^{-1}(G - G_T)B^{-1}$ gives

$$\widehat{w} - w^\star = A^{-1}(b_T - b) + A^{-1}(G - G_T)w^\star.$$

Taking norms,

$$\|\widehat{w} - w^\star\|_2 \leq \|A^{-1}\|_{\mathrm{op}}\,\|b_T - b\|_2 + \|A^{-1}\|_{\mathrm{op}}\,\|G_T - G\|_{\mathrm{op}}\,\|w^\star\|_2.$$

If $\|G_T - G\|_{\mathrm{op}} \leq \lambda/2$, then

$$A = (G+\lambda I) + (G_T - G) \succeq \lambda I - \|G_T - G\|_{\mathrm{op}} I \succeq (\lambda/2) I,$$

so $\|A^{-1}\|_{\mathrm{op}} \leq 2/\lambda$. Plugging this yields (27). □

Lemma A.6 (Bound on $\|w_\lambda^\star\|_2$)**.** *Under Assumption 4.3, $\|w_\lambda^\star\|_2 \leq BY/\lambda$.*

Proof. We have $w_\lambda^\star = (G+\lambda I)^{-1}b$. Since $G \succeq 0$, $\|(G+\lambda I)^{-1}\|_{\mathrm{op}} \leq 1/\lambda$. Moreover, $\|b\|_2 = \|\mathbb{E}_\pi[\phi(x)y(x)]\|_2 \leq \mathbb{E}_\pi[\|\phi(x)\|_2 |y(x)|] \leq BY$. Therefore $\|w_\lambda^\star\|_2 \leq (1/\lambda)\,BY$. □

Proof of Theorem 5.2. Let $\delta \in (0,1)$ and define the events

$$\mathcal{E}_b := \{\|b_T - b\|_2 \le C_b\, D_X\, \eta_{T,\delta}\}, \qquad \mathcal{E}_G := \{\|G_T - G\|_{\mathrm{op}} \le C_G\, D_X\, \eta_{T,\delta}\},$$

with $C_b := C(YL_\phi + BL_y)$ from Corollary 4.4 and $C_G := C(BL_\phi)$ from Theorem 4.5. By these results (with confidence parameters $\delta/2$ each),

$$\mathbb{P}(\mathcal{E}_b) \ge 1 - \delta/2, \qquad \mathbb{P}(\mathcal{E}_G) \ge 1 - \delta/2,$$

so by a union bound $\mathbb{P}(\mathcal{E}_b \cap \mathcal{E}_G) \ge 1 - \delta$. This proves Item (1).

For Item (2), work on the event $\mathcal{E}_b \cap \mathcal{E}_G$. If additionally $\|G_T - G\|_{\mathrm{op}} \le \lambda/2$, Proposition 5.1 gives

$$\|\widehat{w}_{\lambda,T} - w^\star_\lambda\|_2 \le \frac{2}{\lambda}\|b_T - b\|_2 + \frac{2}{\lambda}\|G_T - G\|_{\mathrm{op}}\, \|w^\star_\lambda\|_2.$$

Using Lemma A.6 yields

$$\|\widehat{w}_{\lambda,T} - w^\star_\lambda\|_2 \le \frac{2}{\lambda}\|b_T - b\|_2 + \frac{2BY}{\lambda^2}\|G_T - G\|_{\mathrm{op}}.$$

Substituting the bounds on $\|b_T - b\|_2$ and $\|G_T - G\|_{\mathrm{op}}$ from Item (1) gives (28) (absorbing factors into the constant C).

Finally, the condition $\|G_T - G\|_{\mathrm{op}} \le \lambda/2$ holds whenever

$$C_G D_X \eta_{T,\delta} \le \frac{\lambda}{2} \quad \Longleftrightarrow \quad T \ge \frac{p + \log(2/\delta)}{\kappa} \max\left\{\frac{4C_G^2 D_X^2}{\lambda^2}, \frac{2C_G D_X}{\lambda}\right\},$$

which makes the dependence on $T_{\mathrm{eff}} = \kappa T$ explicit. □

B Technical Background on Wasserstein Contraction and Curvature

This appendix section summarizes the transport and "curvature" notions used in the main paper, and records standard implications that connect Wasserstein contraction to ergodicity and concentration. Our goal is twofold: (i) to make the meaning of the contraction coefficient and curvature gap κ precise, and (ii) to clarify why contractive reservoirs fall into a regime where sharp nonasymptotic concentration is available.

B.1 Wasserstein distances and basic inequalities

Let (X, d) be a compact metric space. Denote by $\mathcal{P}(X)$ the set of Borel probability measures on X and by $\Gamma(\mu, \eta)$ the set of couplings of $\mu, \eta \in \mathcal{P}(X)$. For $p \ge 1$, the p-Wasserstein distance is

$$W_p(\mu, \eta) := \left(\inf_{\gamma \in \Gamma(\mu,\eta)} \int d(x,y)^p\, \gamma(dx, dy)\right)^{1/p}. \tag{33}$$

In this paper we specialize to $p = 1$, which has a particularly convenient dual representation.

Theorem B.1 (Kantorovich–Rubinstein duality)**.** *For $\mu, \eta \in \mathcal{P}(X)$,*

$$W_1(\mu, \eta) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left\{\int f\, d\mu - \int f\, d\eta\right\}. \tag{34}$$

This is classical; see [28]. A frequent corollary is the Lipschitz test-function bound: for any Lipschitz f,

$$\left|\int f\, d\mu - \int f\, d\eta\right| \le \|f\|_{\mathrm{Lip}}\, W_1(\mu, \eta). \tag{35}$$

B.2 Markov kernels, Dobrushin coefficients, and Wasserstein contraction

A Markov kernel P on X acts on measures by $(\mu P)(A) = \int P(x, A)\, \mu(dx)$, and on functions by $(Pf)(x) = \int f(y) P(x, dy)$. A central quantity is the Wasserstein Dobrushin coefficient:

$$\alpha_{W_1}(P) := \sup_{x \ne y} \frac{W_1(P(x,\cdot), P(y,\cdot))}{d(x,y)} \in [0, \infty). \tag{36}$$

This coefficient controls one-step contractivity of P on $\big(\mathcal{P}(X), W_1\big)$:

Lemma B.2 (Kernel contraction ⇒ operator contraction)**.** *For any $\mu, \eta \in \mathcal{P}(X)$,*

$$W_1(\mu P, \eta P) \le \alpha_{W_1}(P)\, W_1(\mu, \eta).$$

Proof. Fix μ, η and let $\gamma \in \Gamma(\mu, \eta)$. Construct a coupling of μP and ηP by first sampling $(X, Y) \sim \gamma$ and then, conditionally on (X, Y), sampling (X', Y') from an optimal coupling of $P(X, \cdot)$ and $P(Y, \cdot)$. Taking expectations yields

$$W_1(\mu P, \eta P) \le \int W_1(P(x, \cdot), P(y, \cdot))\, \gamma(dx, dy) \le \alpha_{W_1}(P) \int d(x, y)\, \gamma(dx, dy).$$

Infimizing over γ proves the claim. □

In the special case where X is equipped with the discrete metric $d_\Sigma(s, s') = \mathbf{1}\{s \neq s'\}$, W_1 coincides with total variation:

$$W_1(\mu, \eta) = \|\mu - \eta\|_{\mathrm{TV}} \qquad \text{when } d = d_\Sigma. \tag{37}$$

Hence $\alpha_{W_1}(P)$ reduces to the classical Dobrushin contraction coefficient in total variation [3, 16].

B.3 Coarse Ricci curvature and the curvature gap

Ollivier's coarse Ricci curvature [22] packages the contraction coefficient into a geometric form. For $x \neq y$, define

$$\kappa(x, y) := 1 - \frac{W_1(P(x, \cdot), P(y, \cdot))}{d(x, y)}. \tag{38}$$

A uniform curvature lower bound $\kappa(x, y) \ge \kappa_0 > 0$ is equivalent to $\alpha_{W_1}(P) \le 1 - \kappa_0 < 1$. In the main paper we use the global curvature gap

$$\kappa_0 := \inf_{x \neq y} \kappa(x, y) = 1 - \alpha_{W_1}(P). \tag{39}$$

Positive curvature $\kappa_0 > 0$ implies geometric mixing in W_1 and, more strongly, enables concentration for empirical means along the chain [22, 14].

Reservoir specialization. In our contractive reservoir setting with i.i.d. inputs,

$$P(x, \cdot) = \sum_{s \in \Sigma} \nu(s)\, \delta_{F_s(x)}(\cdot),$$

and uniform state contraction $d(F_s(x), F_s(y)) \le L d(x, y)$ implies $W_1(P(x, \cdot), P(y, \cdot)) \le L d(x, y)$ (Theorem 3.1), so

$$\kappa_0 \ge 1 - L. \tag{40}$$

Thus the reservoir Markov chain has explicit positive curvature when $L < 1$.

B.4 Ergodicity and concentration: alternative routes

The manuscript uses curvature-based concentration [14]. For completeness we note two complementary approaches commonly used in Markov chain concentration:

- **Marton couplings and transportation inequalities.** Contractivity can be combined with coupling arguments to prove T_1 (transport-entropy) inequalities and concentration (e.g. [21, 23]).
- **Spectral methods.** For reversible chains, Poincaré/log-Sobolev inequalities yield concentration; on finite spaces these relate to spectral gaps [16]. Wasserstein contraction provides a nonreversible, metric alternative that is especially natural for contractive dynamical systems.

These viewpoints are consistent: positive curvature implies a form of transport inequality and concentration under suitable regularity, while in finite settings one can often compare the resulting rates to spectral-gap bounds [14, 23].

C Extensions: Dependent Inputs and Augmented-State Formulations

The manuscript assumes i.i.d. input symbols (U_t) so that the reservoir state (x_t) is Markov. In many applications, inputs are temporally correlated. This section explains how to extend the theory to weakly dependent inputs by embedding the system into an *augmented Markov chain* and re-establishing Wasserstein contraction.

C.1 Markov inputs: augmented chain on $\Sigma \times X$

Assume now (U_t) is a time-homogeneous Markov chain on the finite alphabet Σ with transition matrix Q and stationary distribution ν. The driven reservoir evolves as

$$x_{t+1} = F_{U_t}(x_t), \qquad U_{t+1} \sim Q(U_t, \cdot).$$

Then (x_t) is generally not Markov, but the augmented process

$$z_t := (U_t, x_t) \in Z := \Sigma \times X \tag{41}$$

is Markov with kernel $\widetilde{P}$ defined by

$$\widetilde{P}\big((s,x), A \times B\big) = \sum_{s' \in A} Q(s,s')\, \mathbf{1}\{F_s(x) \in B\}. \tag{42}$$

To quantify contraction, we must specify a metric on Z. Let $d_\Sigma(s,s') := \mathbf{1}\{s \neq s'\}$ be the discrete metric on Σ. For $\rho > 0$ define the weighted product metric

$$d_\rho\big((s,x),(s',y)\big) := d_\Sigma(s,s') + \rho\, d(x,y). \tag{43}$$

The next theorem shows that if the input chain has a one-step contraction in total variation (equivalently, in W_1 under d_Σ), then the augmented chain contracts in W_1 under d_ρ.

Input-chain contraction coefficient. Define the Dobrushin coefficient of Q as

$$\alpha(Q) := \sup_{s \neq s'} \|Q(s,\cdot) - Q(s',\cdot)\|_{\mathrm{TV}} \;=\; \sup_{s \neq s'} W_1\big(Q(s,\cdot), Q(s',\cdot)\big) \quad (d_\Sigma), \tag{44}$$

so that $\alpha(Q) < 1$ is a Doeblin-type one-step contraction condition [3, 16].

Symbol-sensitivity of the reservoir. Define the worst-case change in the next state when the input symbol changes:

$$M := \max_{s \neq s'} \sup_{x \in X} d\big(F_s(x), F_{s'}(x)\big) \;<\; \infty, \tag{45}$$

which is finite by compactness of X and continuity of F_s (in typical ESN instances F_s is continuous).

Theorem C.1 (Wasserstein contraction for the augmented Markov chain)**.** *Assume uniform state contraction $d(F_s(x), F_s(y)) \le L d(x,y)$ with $L \in (0,1)$, and assume $\alpha(Q) < 1$. Fix $\rho > 0$ such that $\alpha(Q) + \rho M < 1$ (any $\rho > 0$ is admissible when $M = 0$), and define d_ρ by* (43)*. Then the augmented kernel $\widetilde{P}$ contracts W_1 on (Z, d_ρ):*

$$W_1^{(\rho)}(\widetilde{\mu}\widetilde{P}, \widetilde{\eta}\widetilde{P}) \;\le\; \widetilde{L}\, W_1^{(\rho)}(\widetilde{\mu}, \widetilde{\eta}), \qquad \widetilde{L} := \max\{L,\ \alpha(Q) + \rho M\} \;<\; 1, \tag{46}$$

where $W_1^{(\rho)}$ denotes Wasserstein–1 computed with cost d_ρ. Consequently, $\widetilde{P}$ has a unique stationary law $\widetilde{\pi}$ on $\Sigma \times X$ and mixes geometrically in $W_1^{(\rho)}$ at rate $\widetilde{L}$.

Proof. It suffices to bound the kernel contraction coefficient in the metric d_ρ. Fix $(s,x),(s',y) \in Z$. Consider a coupling (S_1, S_1') of $Q(s,\cdot)$ and $Q(s',\cdot)$ achieving total variation distance:

$$\mathbb{P}(S_1 \neq S_1') = \|Q(s,\cdot) - Q(s',\cdot)\|_{\mathrm{TV}} \le \alpha(Q)\, d_\Sigma(s,s').$$

(When $s = s'$, take $S_1 = S_1'$ a.s.) Set deterministically

$$X_1 := F_s(x), \qquad Y_1 := F_{s'}(y).$$

This defines a coupling of $\widetilde{P}((s,x),\cdot)$ and $\widetilde{P}((s',y),\cdot)$ because (S_1, X_1) has law $\widetilde{P}((s,x),\cdot)$ and similarly for (S_1', Y_1).

Now

$$\begin{aligned}\mathbb{E}\big[d_\rho((S_1,X_1),(S_1',Y_1))\big] &= \mathbb{E}[d_\Sigma(S_1,S_1')] + \rho\, d(F_s(x),F_{s'}(y)) \\ &\le \alpha(Q)\, d_\Sigma(s,s') + \rho\Big(d(F_s(x),F_s(y)) + d(F_s(y),F_{s'}(y))\Big) \\ &\le \alpha(Q)\, d_\Sigma(s,s') + \rho\Big(Ld(x,y) + M\, d_\Sigma(s,s')\Big) \\ &= \big(\alpha(Q)+\rho M\big)\, d_\Sigma(s,s') + L\,(\rho d(x,y)).\end{aligned}$$

Taking the maximum of the two coefficients yields

$$\mathbb{E}\big[d_\rho((S_1,X_1),(S_1',Y_1))\big] \le \max\{L,\alpha(Q)+\rho M\}\,\big(d_\Sigma(s,s')+\rho d(x,y)\big) = \widetilde{L}\, d_\rho((s,x),(s',y)).$$

Since Wasserstein–1 is the infimum over couplings, this shows $W_1^{(\rho)}(\widetilde{P}((s,x),\cdot),\widetilde{P}((s',y),\cdot)) \le \widetilde{L}\, d_\rho((s,x),(s',y))$. Applying Lemma B.2 (with metric d_ρ) gives (46). The fixed-point and mixing consequences follow exactly as in Corollaries 3.2–3.3. □

If $\alpha(Q)=1$. For irreducible aperiodic finite Q, a coalescent input coupling satisfies $\mathbb{P}(U_j \ne U_j') \le A_0 r^j \mathbf{1}\{U_0 \ne U_0'\}$ for constants $A_0<\infty$ and $r<1$ [16]. Iterating the reservoir inequality under this coupling gives

$$\mathbb{E}d_\rho(z_t,z_t') \le \rho L^t d(x,y) + A_0\left(r^t + \rho M \sum_{j=0}^{t-1} L^{t-1-j} r^j\right) d_\Sigma(s,s').$$

Both coefficients tend to zero, so some augmented m-step kernel contracts $W_1^{(\rho)}$ strictly. Banach's theorem and the finitely many intervening steps then give uniqueness of the stationary law and geometric Wasserstein convergence. This analyzes the full augmented kernel; substituting Q^m into the one-step formula would omit the intermediate inputs.

C.2 General weak dependence: mixing inputs and fading-memory maps

Beyond Markov inputs, uniform contractivity makes the stationary echo state a fading-memory functional of the input past. This controls dependence tested by Lipschitz observables; it does not imply total-variation or β-mixing of the state process.

Proposition C.2 (Lipschitz dependence inherited from mixing inputs)**.** *Let $(U_t)_{t\in\mathbb{Z}}$ be stationary with absolute-regularity coefficients $\beta_U(k)$, and let (x_t) be its stationary causal echo-state solution. For a random variable H measurable with respect to $\sigma(x_s : s\le 0)$, with $|H|\le 1$, and a Lipschitz $f: X\to\mathbb{R}$, every integer $t\ge 1$ and $0\le m<t$ satisfy*

$$|\mathrm{Cov}(H, f(x_t))| \le 2\|f\|_{\mathrm{Lip}} D_X\big(\beta_U(t-m) + L^m\big). \tag{47}$$

Proof. Fix $x_*\in X$ and replace x_{t-m} by x_* before applying inputs $U_{t-m},\dots,U_{t-1}$, obtaining $x_t^{(m)}$. Then $d(x_t,x_t^{(m)}) \le D_X L^m$, so replacement changes the covariance by at most $2\|f\|_{\mathrm{Lip}} D_X L^m$. Causality makes H measurable with respect to the input past up to time -1, whereas $x_t^{(m)}$ depends only on the indicated future block. By the total-variation definition of absolute regularity, their covariance, after subtracting $f(x_*)$, is at most $2\|f\|_{\mathrm{Lip}} D_X \beta_U(t-m)$. Combining these bounds proves (47). □

For example, on $X=[0,1]$, let $x_{t+1}=(x_t+U_t)/2$ with independent $U_t\sim\mathrm{Bernoulli}(1/2)$ and stationary $x_0\sim\mathrm{Unif}[0,1]$. Then the fractional part of $2^t x_t$ equals x_0 almost surely. The event $\{x_0<1/2\}$ is therefore also measurable from x_t at every lag and has covariance $1/4$ with its recovered indicator. Hence the state process is not β-mixing, despite contraction. General-input concentration requires additional assumptions [24, 15]; the main exponential bounds use the Markov or contracting augmented-Markov route.

D Matrix Concentration Details for Gram Matrices

This section complements Theorem 4.5 by (i) spelling out the net-based operator-norm argument in a modular way, and (ii) recording alternative concentration routes (entrywise bounds, and matrix Bernstein via blocking) that may be useful in variants.

D.1 From scalar concentration to operator norm via nets

Let $M \in \mathbb{R}^{p\times p}$ be symmetric. The operator norm satisfies

$$\|M\|_{\mathrm{op}} = \sup_{v\in\mathbb{S}^{p-1}} |v^\top M v|. \tag{48}$$

Lemma A.4 shows that for an ε-net $\mathcal{N}_\varepsilon$ of $\mathbb{S}^{p-1}$ with $\varepsilon \in (0, 1/2)$,

$$\|M\|_{\mathrm{op}} \le \frac{1}{1-2\varepsilon} \sup_{u\in\mathcal{N}_\varepsilon} |u^\top M u|. \tag{49}$$

For $\varepsilon = 1/4$ one has $1/(1-2\varepsilon) = 2$, and the net size bound is $|\mathcal{N}_{1/4}| \le 9^p$ [27].

In our Gram matrix setting,

$$M := G_T - G = \frac{1}{T}\sum_{t=1}^{T} \Big(\phi(x_t)\phi(x_t)^\top - \mathbb{E}_\pi[\phi(x)\phi(x)^\top]\Big),$$

and for each $v \in \mathbb{S}^{p-1}$ we define the scalar observable

$$g_v(x) := (v^\top \phi(x))^2. \tag{50}$$

Then $v^\top M v = \frac{1}{T}\sum_{t=1}^{T} g_v(x_t) - \pi(g_v)$.

Uniform Lipschitz control in v. If ϕ is L_ϕ-Lipschitz and $\|\phi\|_2 \le B$ (Assumption 4.3), then for all $v \in \mathbb{S}^{p-1}$,

$$\|g_v\|_{\mathrm{Lip}} \le 2BL_\phi, \qquad \|g_v\|_\infty \le B^2. \tag{51}$$

Hence Theorem 4.1 applies uniformly in v with the same Lipschitz constant. A union bound over $\mathcal{N}_{1/4}$ then yields Theorem 4.5.

D.2 Entrywise bounds (simple but dimension-worse)

An alternative is to bound each entry $(G_T - G)_{ij}$ and union bound over i, j. Define $h_{ij}(x) := \phi_i(x)\phi_j(x)$. If each coordinate ϕ_i is Lipschitz with constant $\le L_\phi$ and bounded by B, then

$$\|h_{ij}\|_{\mathrm{Lip}} \le 2BL_\phi, \qquad \|h_{ij}\|_\infty \le B^2.$$

Applying Theorem 4.1 and union bounding over p^2 entries yields

$$\|G_T - G\|_{\max} \lesssim BL_\phi\, D_X \sqrt{\frac{\log(p/\delta)}{\kappa T}},$$

and therefore

$$\|G_T - G\|_{\mathrm{op}} \le p\, \|G_T - G\|_{\max},$$

which is typically looser by a factor $\sqrt{p}$ compared to the net-based bound that scales like $\sqrt{p}$ rather than p. The net argument supplies a sufficient $\sqrt{p}$ dependence here.

D.3 Matrix Bernstein via blocking (optional sharper variants)

If one prefers matrix-concentration tools directly, one can combine: (i) matrix Bernstein for independent self-adjoint matrix sums [25, 26], and (ii) a blocking argument for geometrically mixing sequences (a standard approach in dependent-data concentration). For Markov chains, such results are developed systematically in [23] (scalar concentration) and can be adapted to matrix-valued observables by blocking + Tropp's inequality.

Such a blocking argument requires an additional absolute-regularity assumption on the matrix sequence. The block length must be chosen to control a coupling error, and the resulting bound includes the block variance, block length, and residual dependence terms. Wasserstein contraction alone does not supply this assumption, as the example in Appendix C shows. Consequently we do not assert a matrix-Bernstein bound under the main assumptions; the proved net-based bound is Theorem 4.5.

E Additional Corollaries, Variants, and Discussion

This section collects corollaries and variants that follow from the main results with minimal additional work, and highlights how the transport-mixing perspective complements existing reservoir theory.

E.1 Variance and integrated autocorrelation time bounds

A convenient summary of dependence is the integrated autocorrelation time. Under stationarity and geometric correlation decay (Proposition 3.5), one obtains a variance bound for time averages.

Corollary E.1 (Variance bound for Lipschitz time averages)**.** *Assume $x_0 \sim \pi$ and let $f : X \to \mathbb{R}$ be Lipschitz. Then*

$$\mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} f(x_t)\right) \;\le\; \frac{C}{\kappa T}\,\|f\|_{\mathrm{Lip}}^2\,\mathrm{diam}(X)^2, \tag{52}$$

for a universal constant $C > 0$, with $\kappa \ge 1 - L$.

Proof. Write $\overline{f}_T = \frac{1}{T}\sum_{t=1}^{T} f(x_t)$ and expand $\mathrm{Var}(\overline{f}_T) = \frac{1}{T^2}\sum_{s,t}\mathrm{Cov}(f(x_s), f(x_t))$. By stationarity and Proposition 3.5, $|\mathrm{Cov}(f(x_0), f(x_t))| \le \|f\|_{\mathrm{Lip}}^2 \mathrm{diam}(X)^2(1-\kappa)^t$. Summing the geometric series gives $(1 + 2\sum_{t\ge1}(1-\kappa)^t)/T = (2/\kappa - 1)/T \le 2/(\kappa T)$, proving the claim with $C = 2$. □

This supports the concentration-based proxy $T_{\mathrm{eff}} = \kappa T$ at the stated Lipschitz-diameter scale; it does not identify a variance-normalized effective sample size.

E.2 Unregularized least squares when G is well-conditioned

If the population Gram matrix G is strictly positive definite, one can obtain guarantees for ordinary least squares (OLS) by perturbation, similarly to ridge. Let $\lambda_{\min}(G) = \gamma > 0$. If $\|G_T - G\|_{\mathrm{op}} \le \gamma/2$, then G_T is invertible and

$$\|\widehat{w}_{0,T} - w_0^\star\|_2 \le \frac{2}{\gamma}\|b_T - b\|_2 + \frac{2}{\gamma}\|G_T - G\|_{\mathrm{op}}\|w_0^\star\|_2,$$

with $w_0^\star = G^{-1}b$. The same concentration bounds yield an OLS sample-complexity requirement of order $T \gtrsim \frac{p+\log(1/\delta)}{\kappa}\max\{1, (BL_\phi D_X/\gamma)^2\}$ to ensure invertibility, with stability then given by the displayed perturbation bound.

E.3 Metric choice and norm tuning

Our contraction constant L depends on the choice of metric on X. In Euclidean spaces, choosing an equivalent norm (or weighted Euclidean metric) can strictly reduce the Lipschitz constants of the maps $\{F_s\}$. In ESNs, this corresponds to choosing a matrix norm induced by an inner product, a classical technique for tightening contractivity conditions (e.g., comparing $\|W\|_2$ and $\|W\|_P$ for a positive definite P). Metric tuning can improve the certificate, but the diameter and feature Lipschitz constants must be recomputed in that metric; reducing L alone need not improve the full statistical bound.

E.4 Continuous inputs and quantization error

A finite alphabet simplifies notation. Continuous i.i.d. inputs also give the Markov kernel $P(x, A) = \int \mathbf{1}\{F_u(x) \in A\}\nu(du)$, and the same proofs apply under measurability, uniform contraction, and compact forward invariance. Quantization $Q : \mathbb{R}^d \to \Sigma$ is an optional approximation. If the reservoir map is Lipschitz in the input (as in (1) with bounded U), then quantization induces a controlled perturbation: $d\big(F_u(x), F_{Q(u)}(x)\big) \le C_u\|u - Q(u)\|$, and standard stability arguments for contractive systems imply the resulting state perturbation is bounded by a geometric series in L. This yields an additional approximation term in the concentration and learning bounds that is proportional to the quantization resolution (and amplified by $(1 - L)^{-1}$), making explicit a tradeoff between continuous-input fidelity and statistical tractability.

E.5 Relationship to existing reservoir theory

The transport-mixing viewpoint complements two dominant theoretical lenses in reservoir computing:

- **ESP and fading memory.** ESP-type conditions ensure well-posedness and pointwise forgetting [13, 18]. Our results quantify *distributional* forgetting and its statistical consequences for learning from dependent trajectories.
- **Capacity and approximation.** Memory capacity and information-processing capacity quantify representational aspects [2], and universality results characterize approximation power [9]. Our bounds address a different question: *given a finite trajectory length, when does training reliably identify a good readout?* The answer depends on mixing and yields an explicit T_{eff} law.

We view these as complementary: a reservoir may have high dynamical capacity but low statistical efficiency if L is too close to 1 for the available data budget.

E.6 Beyond uniform contraction

Uniform contraction is a strong but clean assumption. Two natural generalizations are:

- **Average contraction / random contraction.** Replace a uniform $L < 1$ by a condition on Lyapunov exponents along the driven dynamics. One may still obtain pullback attractors and, under additional regularity, mixing in suitable Wasserstein-type distances, but the effective rate becomes input-dependent.
- **Piecewise contractive reservoirs.** If the maps are contractive on regions and the chain spends most time in those regions, one may derive drift-and-minorization (Harris) conditions and obtain geometric ergodicity. This connects to classical Markov stability theory and suggests extensions beyond the strict ESP regime.

Developing these extensions while retaining explicit design rules remains an interesting direction.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction summarize the main theoretical contributions (Wasserstein contraction, mixing, concentration, ridge readout guarantees) and explicitly scope the results to the contractive regime and stochastic inputs; see Abstract, Section 1, and Sections 3–5.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper includes a dedicated Limitations paragraph in Section 6, discussing uniform contractivity, i.i.d. finite-alphabet inputs in the main text, and bounded-Lipschitz observables, and points to extensions in Appendix C.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: All assumptions are stated explicitly (e.g., Assumptions 2.1 and 4.3), results are numbered and cross-referenced, and complete proofs are provided in Appendix A with supporting technical background in Appendix B.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A].

   Justification: The submission is purely theoretical and does not report experimental results supporting the main claims; the claims are supported by theorem statements and proofs (Appendix A).

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A].

   Justification: No experiments are included and no datasets or code assets are required to validate the main theoretical results.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A].

   Justification: The paper contains no experimental evaluation; the main results concern finite-sample theory under stated assumptions (Sections 2–5 and Appendix A).

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A].

   Justification: The paper does not include experiments; theoretical statements provide nonasymptotic probability bounds directly (Sections 4–5).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A].

   Justification: The paper contains no experiments and therefore does not require reporting compute resources.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work is theoretical, uses no human subjects, no sensitive data, and poses no special ethical risks; it is presented with explicit assumptions, limitations, and proofs.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The Impact Statement discusses reliability and data efficiency, the risk of overconfidence when assumptions fail, and mitigation through reporting validity conditions and constants.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release models or datasets with dual-use risk; it provides theoretical analysis and proofs only.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A].

    Justification: The paper does not use or distribute third-party code, datasets, or pretrained models as assets; it relies on standard bibliographic citations for prior literature.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not introduce or release new datasets, codebases, or pretrained models; the contributions are theoretical results.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human subjects research.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The paper does not involve human subjects and therefore does not require IRB approval.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [N/A].

Justification: LLMs are not part of the core methodology or theoretical development; the paper's contributions are mathematical results and proofs about contractive ESNs.