VARIATIONAL ANALYSIS OF INFERENCE FROM DYNAMICAL SYSTEMS

KEVIN MCGOFF AND ANDREW NOBEL

Abstract. We introduce and study a variational framework for the analysis of statistical inference from dynamical systems and ergodic processes. The framework applies to a broad family of inference procedures in which (i) a trajectory from an unknown observed system is fit by a trajectory from a known reference system by minimizing empirical risk, and (ii) a parameter estimate is obtained from the initial state of the best fit reference trajectory. We establish that the empirical risk of the best fit trajectory converges almost surely to a constant that can be expressed, in a variational form, as the minimum expected loss over dynamically invariant couplings (joinings) of the observed and reference systems. Moreover, we establish that the the family of joinings minimizing the expected loss fully characterizes the asymptotic behavior of the estimated parameters. We show that variational analysis can be applied to the well studied problems of maximum likelihood estimation and non-linear regression, as well as to the new problem of estimating a transformation from quantized trajectories subject to noise.

1. Introduction

Although independence assumptions are common in the statistics and machine learning literature, there has been long-standing interest, both theoretical and applied, in the analysis of observations exhibiting long-range dependence. Representative recent work can be found in [1, 2, 13, 24, 34, 48]. A number of recent papers have considered inference from dynamical systems that evolve deterministically over time, and ergodic processes, e.g., [14, 15, 25, 26, 23, 29, 43]. Inference procedures for observations with long range dependence have application in a variety of application areas, including ecology [28, 44], geophysical modeling [3, 17], and data assimilation [27].

In this paper we introduce and study a variational framework for the analysis of estimation schemes for dynamical systems and ergodic processes. The framework applies to a broad family of inference procedures that can be decomposed into two stages: a tracking stage in which the trajectory of a known reference system is fit to the trajectory of an
observed system by minimizing empirical risk; and a translation stage in which a parameter estimate is obtained by applying a continuous invariant map to the initial state of the best-fit reference trajectory.

Our principal results concerning the two-stage inference procedure are the following. First, the empirical risk of the optimal reference trajectory in the tracking stage converges almost surely to a constant that is equal to the minimum expected loss over joinings (dynamically invariant couplings) of the observed and reference systems. Second, the family of optimal joinings, namely those that achieve the minimum expected loss, is non-empty, convex, and compact in the weak topology. Third, the family of optimal joinings characterizes the limiting behavior of the estimates derived in the translation stage. Together, these results constitute the basis for the variational analysis of the two-step inference procedure. In particular, the limiting behavior of two-stage estimates can be studied through the set of optimal joinings of the observed and reference dynamical systems.

Variational analysis has a number of desirable properties. It requires relatively mild assumptions (spelled out in detail below). It readily accommodates model misspecification, as the observed and reference systems need not be related to one another. It addresses the problem of identifiability in a direct way, by characterizing of the limiting parameter set, and provides a systematic means of studying the limiting behavior and consistency of the two-stage parameter estimates. Lastly, variational analysis facilitates the application of powerful results and constructions from the theory of joinings to interesting problems in statistics and machine learning.

As described above, inference based on tracking and translation concerns the (deterministic) trajectories of dynamical systems. However, using standard shift constructions and appropriate loss functions, tracking-translation encompasses a wide variety of inference problems in which empirical risk minimization is applied to stationary ergodic observations. As such, variational analysis can provide insights into the asymptotic behavior of estimates in these problems. To illustrate, we show how variational analysis can be applied to the well studied problems of maximum likelihood estimation and non-linear regression. We then consider in detail the new, more challenging problem of estimating a transformation from quantized trajectories with noise, analysis of which requires results and constructions from the theory of joinings. Further applications of variational analysis to the fitting of low-complexity dynamical models are considered in [31].
1.1. Tracking. In this subsection and the next we describe and analyze tracking-translation inference for dynamical systems. The tracking problem has three basic components: an observed ergodic system, a reference topological system, and an integrable loss function.

The observed dynamical system is a triple \((\mathcal{Y}, T, \nu)\) consisting of a non-empty Polish space \(\mathcal{Y}\), a Borel measurable map \(T : \mathcal{Y} \to \mathcal{Y}\), and a Borel probability measure \(\nu\) on \(\mathcal{Y}\) that is invariant and ergodic under \(T\). Recall that \(\nu\) is invariant under \(T\) if \(\nu(T^{-1}A) = \nu(A)\) for each Borel set \(A \subseteq \mathcal{Y}\), and furthermore \(\nu\) is ergodic under \(T\) if \(T^{-1}A = A\) implies \(\nu(A) \in \{0, 1\}\). The reference dynamical system is a pair \((\mathcal{X}, S)\) consisting of a non-empty, compact metric space \(\mathcal{X}\) and a continuous map \(S : \mathcal{X} \to \mathcal{X}\). The loss \(\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\) is a lower semicontinuous function such that \(\sup_{x \in \mathcal{X}} |\ell(x, y)| \leq \ell^*(y)\) for some \(\ell^* \in L^1(\nu)\). In what follows, the conditions above will be referred to as the standard assumptions.

In the tracking problem we have access to a single trajectory \(y, Ty, T^2y, \ldots\) of the observed system \((\mathcal{Y}, T, \nu)\) with initial state \(y\) drawn according to \(\nu\). Here \(T^k\) denotes the k-fold composition of \(T\) with itself; by convention, \(T^0\) is the identity. At time \(n\) we observe the initial segment of the trajectory \(y, Ty, \ldots, T^{n-1}y\) and identify a corresponding initial state \(x\) of the reference system \((\mathcal{X}, S)\) that minimizes the empirical risk \(n^{-1} \sum_{i=0}^{n-1} \ell(S^ix, T^iy)\). Our assumptions on \(\ell\) and \((\mathcal{X}, S)\) ensure that a minimizing initial condition exists.

There is an evident asymmetry in the specification of the observed and reference systems: the observed system is equipped with an invariant ergodic measure, while the reference system is specified without reference to an invariant measure. By the Krylov-Bogoliubov Theorem [22], the family \(\mathcal{M}(\mathcal{X}, S)\) of \(S\)-invariant Borel probability measures on \(\mathcal{X}\) is non-empty. These measures provide a link between the observed and reference systems through the notion of joinings. Recall that the product transformation \(S \times T\) is defined on \(\mathcal{X} \times \mathcal{Y}\) by \((S \times T)(x, y) = (Sx, Ty)\).

**Definition 1.1.** Let \(\mu \in \mathcal{M}(\mathcal{X}, S)\). A Borel probability measure \(\lambda\) on \(\mathcal{X} \times \mathcal{Y}\) is said to be a joining of the systems \((\mathcal{X}, S, \mu)\) and \((\mathcal{Y}, T, \nu)\) if \(\lambda\) is invariant under \(S \times T\) and the marginals of \(\lambda\) on \(\mathcal{X}\) and \(\mathcal{Y}\) are \(\mu\) and \(\nu\), respectively.

A joining is a coupling of the measures \(\mu\) and \(\nu\) with the additional property that the coupling is invariant (stationary) under the product transformation \(S \times T\). Joinings were introduced and first studied by Furstenberg [9] and have played an important role in ergodic theory.
since then, see \cite{6, 10}. For each $\mu \in \mathcal{M}(X, S)$ let $\mathcal{J}(\mu, \nu)$ denote the family of all joinings of $(X, S, \mu)$ and $(Y, T, \nu)$. Note that $\mathcal{J}(\mu, \nu)$ contains the product measure $\mu \otimes \nu$ and is therefore non-empty. Define $\mathcal{J}(S : \nu) = \bigcup_{\mu \in \mathcal{M}(X, S)} \mathcal{J}(\mu, \nu)$. It is easy to see that $\mathcal{J}(S : \nu)$ is just the set of Borel probability measures $\lambda$ on $X \times Y$ such that $\lambda$ is invariant under $S \times T$ and the marginal of $\lambda$ on $Y$ is $\nu$.

Our first principal result shows that the minimal empirical risk in the tracking problem has a limit, and that the limit has a simple variational form involving the loss function $\ell$ and the family $\mathcal{J}(S : \nu)$. The proof of the theorem is given in Section 3.

**Theorem 1.2** (Variational expression for limiting empirical risk). Under the standard assumptions, for $\nu$-almost every initial state $y \in Y$,

$$
\liminf_{n} \frac{1}{n} \sum_{k=0}^{n-1} \ell(S^k x, T^k y) = \inf_{\lambda \in \mathcal{J}(S : \nu)} \int \ell d\lambda := L(S : \nu),
$$

and the second infimum is attained by some joining $\lambda$ in $\mathcal{J}(S : \nu)$.

The value of $L(S : \nu)$ captures the “closeness” of the systems $(X, S)$ and $(Y, T, \nu)$ with respect to the loss function $\ell$. It follows from the integrability assumptions on $\ell$ that $L(S : \nu)$ is finite (see Lemma 3.1). We next investigate the joinings in $\mathcal{J}(S : \nu)$ that achieve the minimum expected loss $L(S : \nu)$. Recall that $\mu$ is said to be an extreme point of a convex family of probability measures $\mathcal{M}$ if $\mu = t \mu_1 + (1 - t) \mu_2$ with $t \in (0, 1)$ and $\mu_1, \mu_2 \in \mathcal{M}$ implies $\mu_1 = \mu_2 = \mu$. The proof of the following theorem can be found in Appendix A.

**Theorem 1.3** (Structural of optimal joinings). Under the standard assumptions, the set of optimal joinings

$$
\mathcal{J}_{\text{min}}(S : \nu) = \left\{ \lambda \in \mathcal{J}(S : \nu) : \int \ell d\lambda = L(S : \nu) \right\}
$$

is non-empty, convex, and compact in the weak topology. Furthermore, a joining $\lambda$ is an extreme point of $\mathcal{J}_{\text{min}}(S : \nu)$ if and only if it is ergodic under $S \times T$.

**Remark 1.4.** It follows from Theorem 1.3 and the Krein-Milman theorem that there exists an ergodic optimal joining. Further, by considering the $X$-marginal of an ergodic optimal joining, we see that there exists an ergodic measure $\mu$ for the system $(X, S)$ that can be optimally joined with $\nu$.

1.2. **Translation.** For a fixed sample size $n$ the tracking problem is a special case of empirical risk minimization in which the initial segment of trajectory $y, Ty, \ldots, T^{n-1} y$ is fit using a family of sequences.
(x, Sx, ..., Sn−1x) indexed by initial states x ∈ X. In the translation stage of the inference procedure, a parameter estimate is obtained from the initial state of a trajectory that minimizes, or nearly minimizes, empirical risk.

Let Θ be a compact metrizable parameter space. A parameter map is a continuous function ϕ : X → Θ that is invariant under the dynamics of the reference system in the sense that ϕ ◦ S = ϕ. (In what follows these conditions on Θ and ϕ are included in the standard assumptions.) Invariance ensures that the value θ = ϕ(x) is constant on the trajectory x, Sx, S2x, ... and may therefore be viewed as a property of the entire trajectory of x under S. The two-stage inference procedure is formalized as follows.

**Definition 1.5.** A sequence of measurable functions θn : Yn → Θ, n ≥ 1, is an optimal ϕ-estimation scheme if θn = ϕ ◦ xn where the functions xn : Yn → X are such that for ν-almost every y in Y,

\[
\lim_n \frac{1}{n} \sum_{k=0}^{n-1} \ell(S^k \hat{x}_n, T^k y) = \lim \inf_n \frac{1}{n} \sum_{k=0}^{n-1} \ell(S^k x, T^k y),
\]

with \( \hat{x}_n = x_n(y, ..., T^{n-1} y) \). Thus the estimate \( \hat{\theta}_n = \theta_n(y, ..., T^{n-1} y) \) is obtained by applying the parameter map ϕ to an initial state \( \hat{x}_n \) of the reference system obtained by (asymptotically) minimizing the average loss with the observed trajectory y, ..., \( T^{n-1} y \).

We now address the limiting behavior of the estimates \( \hat{\theta}_n \). For each \( \theta \in \Theta \) let \( X_{\theta} = \varphi^{-1}\{\theta\} \) be the set of states in X that are mapped to \( \theta \), and let \( S_{\theta} \) be the restriction of \( S \) to \( X_{\theta} \). It is easy to see that \( X_{\theta} \) is a compact subset of X that is invariant under \( S \), and therefore \( (X_{\theta}, S_{\theta}) \) is a topological dynamical system. Thus the parameter map \( \varphi \) gives rise to a family of topological systems, indexed by the parameters \( \theta \in \Theta \), each of which can act as a reference system for tracking the observed ergodic system \( (Y, T, \nu) \).

It follows from Theorem 1.2 that the limiting average loss of tracking \( (Y, T, \nu) \) using the reference system \( (X_{\theta}, S_{\theta}) \) is equal to \( L(S_{\theta} : \nu) \). It is shown in Lemma 4.2 below that \( L(S : \nu) = \inf_{\theta} L(S_{\theta} : \nu) \), and we therefore study the family

\[
\Theta_{\text{min}} = \arg\min_{\theta \in \Theta} L(S_{\theta} : \nu)
\]

of parameters with minimal limiting loss. It is not difficult to show that

\[
\Theta_{\text{min}} = \{ \theta \in \Theta : \exists \lambda \in J_{\text{min}}(S : \nu) \text{ s.t. } \lambda(\varphi^{-1}\{\theta\} \times Y) = 1 \}.
\]
Thus $\Theta_{\text{min}}$ is the set of parameters whose associated states in $X$ support an optimal joining with the observed system $(Y, T, \nu)$. In this sense $\Theta_{\text{min}}$ is the set of parameters in $\Theta$ most compatible with the observed system $(Y, T, \nu)$, and it is a natural limit set for optimal $\varphi$-estimation schemes. The next theorem is our principal inference result. Its proof appears in Section 4.

**Theorem 1.6** (Convergence of estimators). Under the standard assumptions $\Theta_{\text{min}}$ is non-empty and compact. Moreover, if $(\theta_n)_{n \geq 1}$ is an optimal $\varphi$-estimation scheme, then $\hat{\theta}_n = \theta_n(y, \ldots, T^{n-1}y)$ converges to $\Theta_{\text{min}}$ for $\nu$-almost every $y \in Y$. Conversely, for every $\theta_0 \in \Theta_{\text{min}}$ there exists an optimal $\varphi$-estimation scheme that converges almost surely to $\theta_0$.

Theorem 1.6 fully characterizes the limiting behavior of optimal $\varphi$-estimation schemes. In particular, it reduces questions about identifiability and consistency to the analysis of the set $\Theta_{\text{min}}$, and the family $J_{\text{min}}(S : \nu)$ of optimal joinings. In this way the theorem facilitates the application of joining constructions and properties to the analysis of statistical problems, e.g., in the analysis of quantized data below, where joinings play a critical role in the analysis. The theorem places no restrictions on the relation between the observed and reference systems, which need not be the same.

Inference through tracking and translation involves deterministic observations and fitting, both in the absence of noise. Nevertheless, by appropriate choice of the observed system $(Y, T, \nu)$, the reference system $(X, S)$, and the loss function $\ell(x, y)$, Theorem 1.6 may be applied to a variety of statistical problems involving stationary ergodic observations. In the next three sections, Theorem 1.6 is used to establish the consistency of empirical risk minimization procedures for maximum likelihood estimation, nonlinear regression, and estimation of a transformation from quantized observations. The latter problem has not received much attention in the literature, but we believe it to be of independent interest.

1.3. Maximum likelihood estimation under ergodic sampling.

We show here how Theorem 1.6 can be used to derive classical results on the consistency of maximum likelihood estimation. For a thorough discussion of this topic, see van der Vaart [45].

Let $U$ be a Polish space, and let $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ be a family of probability densities $f_\theta : U \to [0, \infty)$ with respect to a fixed Borel measure $Q$ on $U$. Assume that $\Theta$ is a compact metric space and that $(\theta, u) \mapsto f_\theta(u)$ is an upper semi-continuous map from $\Theta \times U$ to $\mathbb{R}$.
Suppose that we observe the values of a stationary ergodic process \( U_0, U_1, \ldots \in \mathcal{U} \) and wish to identify a density \( f_\theta \in \mathcal{P} \) that best approximates the marginal distribution of the observed process in the sense that

\[
\mathbb{E} \log f_\theta(U) = \max_{\theta' \in \Theta} \mathbb{E} \log f_{\theta'}(U),
\]

where \( U \) has the same distribution as \( U_1 \). Let \( \theta_n : \mathcal{U}^n \to \Theta, n \geq 1, \) be measurable estimators that maximize the marginal log-likelihood in the sense that

\[
\lim_{n} \frac{1}{n} \sum_{i=0}^{n-1} \log f_{\hat{\theta}_n}(U_i) = \sup_{\theta \in \Theta} \lim_{n} \frac{1}{n} \sum_{i=0}^{n-1} \log f_{\theta}(U_i) \quad \text{wp1}
\]

where \( \hat{\theta}_n = \theta_n(U_0, \ldots U_{n-1}) \). The existence of measurable estimators satisfying (1.6) follows from standard arguments, see Lemma 4.1. Note that the marginal distribution of the observations \( U_i \) need not have a density in \( \mathcal{P} \) and need not be absolutely continuous with respect to the reference measure \( Q \).

The problem described above can be expressed as a two-stage inference procedure in the following way. To begin, we represent the observed process \( \{U_i\}_{i \geq 0} \) as a measure preserving system \((\mathcal{Y}, T, \nu)\), where \( \mathcal{Y} \) is the sequence space \( \mathcal{U}^\mathbb{N} \), \( T \) is the left-shift on \( \mathcal{Y} \), and \( \nu \) is the measure on \( \mathcal{U}^\mathbb{N} \) induced by \( \{U_i\} \). Let the state space \( \mathcal{X} \) of the reference system be equal to the parameter space \( \Theta \) and, as the inference task involves no dynamics beyond those of the observations \( U_i \), let \( S \) be the identity map on \( \Theta \). Finally, let \( \ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \) be defined by \( \ell(\theta, (u_i)_{i \geq 0}) = -\log f_\theta(u_0) \), and let \( \varphi : \mathcal{X} \to \Theta \) be the identity map. These correspondences are detailed in Table 1. A direct application of Theorem 1.7 yields the following classical result, which is similar to Theorem 5.14 of [45].

**Theorem 1.7.** If \( \mathbb{E} \sup_{\theta \in \Theta} |\log f_\theta(U)| \) is finite, then \( \hat{\theta}_n \) converges almost surely to the set \( \Theta_0 = \arg\max_{\theta \in \Theta} \mathbb{E} \log f_\theta(U) \).

The theorem shows that, even in the misspecified setting, the empirical maximum likelihood estimators converge to the set of optimal parameters, i.e., those that best approximate the observed process in the sense of (1.7). If the supremum in the theorem fails to be measurable, then one may replace the expectation there by an outer expectation.

1.4. **Nonlinear regression under ergodic sampling.** Let \( \mathcal{U} \) be a Polish space, and let \( \mathcal{F} = \{f_\theta : \theta \in \Theta\} \) be a family of functions
Table 1. Correspondence between objects in the general setting and objects in MLE under ergodic sampling.

| General setting | MLE under ergodic sampling |
|-----------------|-----------------------------|
| $\mathcal{X}$   | $\Theta$                   |
| $S: \mathcal{X} \to \mathcal{X}$ | $\text{Id}: \Theta \to \Theta$ |
| $\mathcal{Y}$   | $\mathcal{U}^\infty$      |
| $T: \mathcal{Y} \to \mathcal{Y}$ | Left shift $\tau$ on $\mathcal{U}^\infty$ |
| $\nu$           | Measure of process $\{U_i\}_{i \geq 0}$ |
| $c: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ | $(\theta, u) \mapsto -\log p_\theta(u_0)$ |
| $\varphi: \mathcal{X} \to \Theta$ | $\text{Id}: \Theta \to \Theta$ |

$f_\theta: \mathcal{U} \to \mathbb{R}$ indexed by a compact metric space $\Theta$ in such a way that $(\theta, u) \mapsto f_\theta(u)$ is a continuous map from $\Theta \times \mathcal{U}$ to $\mathbb{R}$. Suppose that we observe a stationary ergodic process $(U_0, V_0), (U_1, V_1), \ldots \in \mathcal{U} \times \mathbb{R}$ and wish to identify a function $f_\theta \in \mathcal{F}$ that best captures the marginal relation between $U$ and $V$ in the sense that

$$(1.7) \quad \mathbb{E} \ell_0(f_\theta(U), V) = \min_{\theta' \in \Theta} \mathbb{E} \ell_0(f_{\theta'}(U), V),$$

where $\ell_0: \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is a lower semicontinuous loss function. Let $\theta_n: (\mathcal{U} \times \mathbb{R})^n \to \Theta$, $n \geq 1$, be measurable estimators that minimize the average loss, in the sense that

$$(1.8) \quad \lim_{n} \frac{1}{n} \sum_{i=0}^{n-1} \ell_0(f_{\hat{\theta}_n}(U_i), V_i) = \lim_{n} \inf_{\theta \in \Theta} \frac{1}{n} \sum_{i=0}^{n-1} \ell_0(f_{\theta}(U_i), V_i) \quad \text{wp1}$$

where $\hat{\theta}_n = \hat{\theta}_n((U_0, V_0), \ldots (U_{n-1}, V_{n-1}))$. This problem can readily be expressed as a two-stage inference procedure, see Table 2 for the details. The following result is an easy consequence of Theorem 1.6; the proof is omitted. We note that for continuous losses the result can also be established by arguments based on uniform laws of large numbers.

**Theorem 1.8.** If $\mathbb{E} \sup_{\theta \in \Theta} \ell_0(f_{\theta}(U), V)$ is finite then $\hat{\theta}_n$ converges almost surely to the set $\Theta_0 = \arg\min_{\theta \in \Theta} \mathbb{E} \ell_0(f_{\theta}(U), V)$.

1.5. Estimating a transformation from quantized trajectories.

As a third application of variational analysis, we consider the problem of identifying a transformation from a known family based on quantized observations of its trajectories, possibly subject to noise. In this case, the reference system requires a non-trivial dynamical component.
Let $\mathcal{U}$ be a Polish space and let $\mathcal{R} = \{ R_\theta : \theta \in \Theta \}$ be a family of Borel measurable transformations $R_\theta : \mathcal{U} \to \mathcal{U}$ indexed by a compact metric space $\Theta$ in such a way that the map $(\theta, u) \mapsto R_\theta(u)$ is a Borel measurable function from $\Theta \times \mathcal{U}$ to $\mathcal{U}$. Let $\{A_0, A_1\}$ be a known, measurable partition of $\mathcal{U}$, and let $\pi : \mathcal{U} \to \{0, 1\}$ be the associated label function, i.e., $\pi(u) = j$ if $u \in A_j$. For each parameter $\theta \in \Theta$ and element $u \in \mathcal{U}$ there is an associated trajectory $u, R_{\theta_0} u, R_{\theta_0}^2 u, \ldots$ arising from repeated application of $R_{\theta}$. Application of the map $\pi$ gives rise to a corresponding binary label sequence

$$\text{lab}(\theta, u) = (\pi(u), \pi(R_{\theta_0} u), \pi(R_{\theta_0}^2 u), \ldots) \in \{0, 1\}^\mathbb{N}.$$  

Of interest here is whether, and in what sense, we can estimate the parameter $\theta$ from noisy observations of the label sequence $\text{lab}(\theta, u)$ when the state $u$ is drawn from an invariant ergodic measure for $R_{\theta}$.

In more detail, we assume that observations take the form of a binary stochastic process

\begin{equation}
Y_k = \pi(R_{\theta_0}^k U) \oplus \varepsilon_k, \quad k \geq 0.
\end{equation}

Here $\theta_0 \in \Theta$ is the parameter of the underlying transformation, $U$ is a $\mathcal{U}$-valued random variable whose distribution is invariant and ergodic under $R_{\theta_0}$, $\{\varepsilon_k\}$ is a sequence of independent Bernoulli($p$) random variables independent of $U$, and $\oplus$ denotes addition modulo 2. Thus $Y_k$ is equal to the label of $R_{\theta_0}^k U$ perturbed by noise; standard arguments ensure that $\{Y_k : k \geq 0\}$ is stationary and ergodic. Let $\theta_n : \{0, 1\}^n \to \Theta$, $n \geq 1$, be measurable estimators that minimize average Hamming (0-1).
risk in the sense that
\[
\lim_{n} \inf_{u \in \mathcal{U}} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{I}(\pi(R_{\hat{\theta}_n}^k u) \neq Y_k)
\]
\[
= \lim_{n} \inf_{\theta \in \Theta} \inf_{u \in \mathcal{U}} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{I}(\pi(R_{\theta}^k u) \neq Y_k) \text{ wp1,}
\]
where \(\hat{\theta}_n = \theta_n(Y_0, \ldots, Y_{n-1})\). (The existence of measurable estimators satisfying (1.10) is guaranteed by Lemma 4.1.) We are interested in the limiting behavior of \(\hat{\theta}_n\).

The estimation problem can be expressed as a two stage inference procedure. As in the previous examples, the observation process \(\{Y_k\}_{k \geq 0}\) can be represented as a measure preserving system \((\mathcal{Y}, T, \nu)\), where \(\mathcal{Y}\) is the sequence space \(\{0, 1\}^\mathbb{N}\), \(T\) is the left-shift on \(\{0, 1\}^\mathbb{N}\), and \(\nu\) is the process measure. Specification of the reference system and loss require more care. We let the state space of the reference system consist of parameter-sequence pairs,

\[
\mathcal{X} = \text{cl}\{(\theta, \text{lab}(\theta, u)) : \theta \in \Theta, u \in \mathcal{U}\} \subseteq \Theta \times \{0, 1\}^\mathbb{N},
\]
where \(\text{cl}A\) denotes the closure of \(A\) and we assume that \(\{0, 1\}^\mathbb{N}\) is equipped with the usual product topology. Thus \(\mathcal{X}\) is compact, and we let the reference transformation \(S\) be the restriction to \(\mathcal{X}\) of the product \(\text{id}_\Theta \times \tau\), where \(\text{id}_\Theta\) is the identity on \(\Theta\) and \(\tau\) is the left shift on \(\{0, 1\}^\mathbb{N}\). It is easy to see that \(\mathcal{X}\) is invariant under \(S\). We further define the loss \(\ell((\theta, a), b) = \mathbb{I}(a_0 \neq b_0)\), and let the parameter map \(\varphi\) be the projection onto \(\Theta\), namely \(\varphi(\theta, a) = \theta\). These correspondences are summarized in Table 3.

Theorem 1.6 characterizes the limiting behavior of the parameter estimates \(\hat{\theta}_n\). Let \(\nu_0\) be the distribution of the true label process \(\{\pi(R_{\theta_0}^k U) : k \geq 0\}\) on \(\{0, 1\}^\mathbb{N}\). Note that \(\nu_0\) is not equal to the process measure \(\nu\) of \(\{Y_k\}\) if the noise level \(p > 0\). For each \(\theta \in \Theta\) let \(\mathcal{X}_\theta = \{a : (\theta, a) \in \mathcal{X}\}\) be the \(\theta\)-section of \(\mathcal{X}\), and define

\[
\Theta_1 = \{\theta \in \Theta : \nu_0(\mathcal{X}_\theta) = 1\}
\]
to be the set of parameters for which \(\mathcal{X}_\theta\) supports the true label process. It is easy to see that \(\theta_0 \in \Theta_1\), so \(\Theta_1\) is non-empty.

It is clear that \(\Theta_1\) is a natural identifiability class for the estimates \(\hat{\theta}_n\) in the absence of noise \((p = 0)\). In fact, it continues to be so when noise is present \((p > 0)\), provided there are complexity constraints on the family of transformations \(\mathcal{R}\). To quantify these constraints, let \(\mathcal{L}\) be the closure (in \(\{0, 1\}^\mathbb{N}\)) of the set of all label sequences \(\{\text{lab}(\theta, u) :
\( \theta \in \Theta, u \in U \) of the transformations in \( \mathcal{R} \). Let

\[
(1.12) \quad h(\mathcal{R}) = \lim_{n \to \infty} \frac{1}{n} \log \# \{ a_0^{n-1} \in \{0, 1\}^n : a \in \mathcal{L} \},
\]

be the exponential growth rate of the number of distinct labeled trajectories of length-\( n \). A more detailed discussion of \( \Theta_1 \) and \( h(\mathcal{R}) \) can be found in Section 5, which also contains the proof of the following theorem.

| General setting | Quantized observations |
|----------------|-----------------------|
| \( \mathcal{X} \) | \( \text{cl}\{(\theta, \text{lab}(\theta, u)) : \theta \in \Theta, u \in U\} \) |
| \( S : \mathcal{X} \to \mathcal{X} \) | \( \text{id}_\Theta \times \tau \) restricted to \( \mathcal{X} \) |
| \( \mathcal{Y} \) | \( \{0, 1\}^\mathbb{N} \) |
| \( T : \mathcal{Y} \to \mathcal{Y} \) | \( \text{Left shift } \tau \) on \( \{0, 1\}^\mathbb{N} \) |
| \( \nu \) | Measure of process \( \{Y_k\}_{k \geq 0} \) |
| \( \ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \) | \( (\theta, a, b) \mapsto \mathbb{I}(a_0 \neq b_0) \) |
| \( \varphi : \mathcal{X} \to \Theta \) | \( (\theta, a) \mapsto \theta \) |

**Table 3.** Correspondence between objects in the general setting and objects in the estimation of a transformation with quantized observations.

**Theorem 1.9.** Let \( \{\hat{\theta}_n\} \) be a sequence of estimates satisfying (1.10). If either

1. \( p = 0 \) or
2. \( 0 < p < 1/2 \) and \( h(\mathcal{R}) = 0 \),

then \( \hat{\theta}_n \) converges almost surely to \( \Theta_1 \).

Note that the limit set \( \Theta_1 \) is the same for both the noisy and noise-free settings. Similar results hold when trajectories are quantized by arbitrary finite partitions and subject to more general noise, see [32]. Note also that Theorem 1.9 holds without continuity assumptions on the transformations \( R_\theta \) or their indexing by \( \theta \). However, the topology of \( \Theta \) does play an important role in this result, as it affects the closure operation that defines \( \mathcal{X} \), which in turn is used to define \( \Theta_1 \).

### 1.5.1. Circle Rotations

As a non-trivial application of the results in this section, consider the family \( \mathcal{R} \) of circle rotations \( R_\alpha : [0, 1) \to [0, 1) \) defined by \( R_\alpha(x) = x + \alpha \mod 1 \), with \( \alpha \in \Theta = [0, 1/2] \). If \( \alpha \) is rational
with reduced form \( m/n \), then \( R_\alpha^n = \text{Id} \), each orbit contains exactly \( n \) distinct points, and each ergodic measure is supported on a single orbit. On the other hand, if \( \alpha \) is irrational, then it is known that Lebesgue measure is the only ergodic Borel probability measure for \( R_\alpha \). Consider the partition of \([0,1)\) into sets \( A_0 = [0, \frac{1}{2}) \) and \( A_1 = [\frac{1}{2}, 1) \). The proof of the following result appears in Section 5.

**Proposition 1.10.** Let \( Y_i = \pi(R_{\alpha_0}^i U) \oplus \epsilon_i \), where \( \alpha_0 \in [0,1/2] \), the distribution of \( U \) is invariant and ergodic under \( R_{\alpha_0} \), and \( \{\epsilon_i\}_{i \geq 0} \) is an i.i.d. sequence of Bernoulli\((p)\) random variables that is independent of \( U \). If \( p < 1/2 \) and the estimates \( \hat{\theta}_n \) satisfy (1.10), then \( \hat{\theta}_n \) converges almost surely to \( \alpha_0 \).

1.6. **Related work.** The most closely related prior work concerns statistical inference in the context of ergodic observations and dynamical systems. In fact, several recent papers have considered this topic, including [14, 15, 25, 26, 29, 43]. While the variational framework considered here focuses on the problem of parameter estimation, the recent work cited above has focused on different aspects of statistical inference. For example, [14, 15, 43] give some results about forecasting dynamical systems with specified mixing rates, and [25, 26] contain both positive and negative results for filtering problems in the context of certain dynamical systems. For additional references and discussion, see the recent survey on statistical inference for dynamical systems [30].

In several applied fields, there is interest in fitting parametrized families of dynamical systems to observations. For some examples, see [3, 17, 27, 28, 33, 44] and references therein. As explained in greater detail in [31], the variational approach taken here may be useful in analyzing the fitting methods in settings such as these.

Some inferential questions have been considered in ergodic theory. Ornstein and Weiss [39] considered finitary estimation of a stationary ergodic process from samples of the process. They proposed a specific estimation scheme based on the empirical \( k \)-dimensional distributions of the process, with \( k \) growing as a function of \( n \). Then they showed that this scheme produces consistent estimates of the observed process if and only if the observed process is Bernoulli. Furthermore, they showed that there is no estimation scheme that produces universally consistent estimates. Note that consistent estimation may be possible for restricted classes of systems or processes, as we show in some of our results, despite that fact that consistency is impossible for larger classes of processes (as shown by Ornstein and Weiss). Other related work concerns finitary estimation of \( k \)-dimensional distributions for growing \( k \) [18] and finitary estimation of isomorphism invariants [13, 30].
The minimal expected loss \( L(S : \nu) \) and the set of optimal joinings \( J_{\min}(S : \nu) \) defined here have close analogies in the study of optimal transport; see the book by Villani [46]. The optimal transport cost associated with two probability measures \( \mu \) and \( \nu \) is the infimum of \( \mathbb{E} c(X, Y) \) over all couplings \( (X, Y) \) of \( \mu \) and \( \nu \). One of the main goals in optimal transport is to describe the properties of optimal couplings, that is, couplings that achieve the infimum of the expected cost. Such optimal couplings are somewhat analogous to the optimal joinings in \( J_{\min}(S : \nu) \). In some cases, notably in the case of Ornstein’s \( d \)-metric and its generalizations in ergodic theory and information theory (see the work of Gray, Neuhoff and Shields [12] and the book of Gray [11]), the measures \( \mu \) and \( \nu \) are taken to be process measures, and the couplings are required to be joinings. While similar in spirit, our results are distinct from this previous work, since we consider the family of joinings between a topological dynamical system and a measure-preserving system and we focus on applications to inference.

In the special case that the loss function does not depend on the observed trajectory, the tracking part of our two-stage procedure reduces to the problem of ergodic optimization, which has received considerable attention in the mathematical literature in recent years (see the survey of Jenkinson [16] for a thorough introduction to the topic). For some recent results, see the work of Quas and Seifken [41] and references therein.

1.7. **Organization of the paper.** The next section provides some background notation and preliminary lemmas needed for the proofs of the main results. Theorems 1.2 and 1.6 are established in Sections 3 and 4 respectively. Section 5 contains the proofs of consistency for the quantized observation problem presented above. Appendix A contains material on the set of optimal joinings, including the proof of Theorem 1.3.

2. **Definitions and background**

In this section we provide several preliminary definitions and facts required for the principal results of the paper.

2.1. **Dynamical systems and spaces of measures.** All topological spaces considered in this paper are Polish (separable and completely metrizable). We endow any such space with its Borel \( \sigma \)-algebra and suppress this choice in our notation. Let \( U \) be a Polish space. Following standard notation, \( \mathcal{M}(U) \) will denote the space of Borel probability measures on \( U \) endowed with the usual weak topology, under which
$M(\mathcal{U})$ is itself a Polish space. Recall that if $R : \mathcal{U} \to \mathcal{U}$ a measurable transformation, then $M(\mathcal{U}, R)$ denotes the set of measures $\mu \in M(\mathcal{U})$ that are invariant under $R$. If $h : \mathcal{U} \to \mathcal{V}$ is measurable, then we define the “push-forward” map $m_h(\cdot) : M(\mathcal{U}) \to M(\mathcal{V})$ by $m_h(\eta) = \eta \circ h^{-1}$. If $\mathcal{U}$ is a non-empty compact metric space and $R$ is continuous, then we refer to $(\mathcal{U}, R)$ as a topological dynamical system. It is well-known that in this case, $M(\mathcal{U}, R)$ is non-empty and compact in the weak topology (see [17]).

2.2. Product spaces and the shift map. The canonical projections of a product space $\mathcal{U} \times \mathcal{V}$ onto its constituent sets will be denoted by $\text{proj}_\mathcal{U}$ and $\text{proj}_\mathcal{V}$. If $\lambda$ is a measure on $\mathcal{U} \times \mathcal{V}$, then its marginal distributions on $\mathcal{U}$ and $\mathcal{V}$ will be denoted by $m_\mathcal{U}(\lambda)$ and $m_\mathcal{V}(\lambda)$, respectively. In several places throughout the paper we will consider infinite product spaces of the form $\mathcal{U}^N$, where $\mathcal{U}$ is a Polish space. In each case $\mathcal{U}^N$ is endowed with its product topology and associated Borel sigma-field; elements of $\mathcal{U}^N$ are denoted as sequences $u = (u_i)_{i \geq 0}$. For any product space, the left-shift map $\tau : \mathcal{U}^N \to \mathcal{U}^N$ is defined by by $\tau(u_0, u_1, \ldots) = (u_1, u_2, \ldots)$. Note that $\tau$ is continuous in the product topology. For any sequence $(u_i)_{i \geq 0}$ and $0 \leq i \leq j$, we define $u_i^j = (u_i, \ldots, u_j)$.

2.3. The process generated by a measure-preserving system.

Let the dynamical systems $(\mathcal{X}, S)$ and $(\mathcal{Y}, T, \nu)$ satisfy the standard assumptions of Section 1. By definition of the left shift $\tau$, any probability measure $\tilde{\nu} \in M(\mathcal{Y}^N, \tau)$ is the distribution of a one-sided stationary process with values in $\mathcal{Y}$. We will say that a measure $\tilde{\nu} \in M(\mathcal{Y}^N, \tau)$ is generated by the system $(\mathcal{Y}, T, \nu)$ if the one-dimensional marginal distribution of $\tilde{\nu}$ is $\nu$, and if $\tilde{\nu}$ is supported on trajectories of $T$ in the sense that

\begin{equation}
\tilde{\nu}\left(\{y \in \mathcal{Y}^N : y_{i+1} = Ty_i \text{ for } i \geq 0\}\right) = 1.
\end{equation}

The following technical lemma, which relates $J(S : \tilde{\nu})$ to $J(S : \nu)$, will be used in several proofs.

**Lemma 2.1.** If $\tilde{\nu} \in M(\mathcal{Y}^N, \tau)$ is generated by the system $(\mathcal{Y}, T, \nu)$, then $m_{\mathcal{X} \times \mathcal{Y}}(J(S : \tilde{\nu})) = J(S : \nu)$.

**Proof.** Let $\tilde{\nu} \in M(\mathcal{Y}^N)$ be generated by $(T, \nu)$. Let $\tilde{\lambda} \in M(\mathcal{X} \times \mathcal{Y}^N)$ have marginal distribution $\tilde{\nu}$ on $M(\mathcal{Y}^N)$ and marginal distribution $\eta$ on $\mathcal{X} \times \mathcal{Y}$. Let $f : \mathcal{X} \times \mathcal{Y}^N \to \mathbb{R}$ be bounded and measurable, and define an associated bounded, measurable function $f_0 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ by
\[ f_0(x, y) = f(x, (y, Ty, \ldots)) \] Using (2.1), one may verify that
\[ \int f \, d\bar{\lambda} = \int f_0 \, d\eta, \]
and for \( \bar{\lambda} \)-almost every \((x, y)\),
\[ f \circ (S \times \tau)(x, y) = f_0(Sx, y_1) = f_0 \circ (S \times T)(x, y_0). \]

Now suppose that \( \bar{\lambda} \in \mathcal{J}(S : \nu) \) has marginal distribution \( \eta \) on \( X \times Y \). Let \( h : X \times Y \to \mathbb{R} \) be a bounded measurable function, and define \( f(x, y) = h(x, y_0) \). It then follows from (2.2), (2.3), and the invariance of \( \bar{\lambda} \) under \( S \times \tau \) that \( \int h \circ (S \times T) \, d\eta = \int h \, d\eta \). As \( h \) was arbitrary, \( \eta \) is \( S \times T \) invariant. Furthermore, it is easy to see that the \( Y \)-marginal of \( \eta \) is \( \nu \). Thus, \( \eta \in \mathcal{J}(S : \nu) \).

To establish the other direction, suppose that \( \eta \in \mathcal{J}(S : \nu) \), and define \( \bar{\lambda} \in \mathcal{M}(X \times Y^\infty) \) to be the distribution of \((X, Y_0, TY_0, \ldots)\), where \((X, Y_0) \sim \eta \). It is clear that \( m_{X \times Y}(\bar{\lambda}) = \eta \) and that the marginal distribution \( \bar{\nu} \) of \( \bar{\lambda} \) on \( Y^\infty \) is generated by \((T, \nu)\). Moreover, it follows from (2.2), (2.3), and the invariance of \( \eta \) under \( S \times T \) that \( \bar{\lambda} \) is invariant under \( S \times \tau \). Thus, \( \bar{\lambda} \in \mathcal{J}(S : \bar{\nu}) \), as desired. \( \square \)

2.4. A genericity lemma. The following lemma is standard when \( U \) is compact \( (e.g., \text{see } [4]) \). One may reduce the more general case of interest here to the compact case using the regularity of \( \mu \), the separability of \( C(K) \) for any compact \( K \), and the pointwise ergodic theorem. As the argument is straightforward, we omit the proof.

**Lemma 2.2.** Suppose \( U \) is a Polish space, equipped with the Borel \( \sigma \)-algebra, \( R : U \to U \) is measurable, and \( \eta \) is a Borel probability measure on \( U \) that is ergodic and invariant with respect to \( R \). Then there exists a measurable set \( E \subset U \) such that \( \eta(E) = 1 \) and if \( x \) is in \( E \), then for each bounded continuous function \( f : U \to \mathbb{R} \),
\[ \lim_{n} \frac{1}{n} \sum_{k=0}^{n-1} f \circ R^k(x) = \int f \, d\eta. \]

3. The Tracking Theorem

This section is devoted to the proof of Theorem 1.2. We first establish the finiteness of the optimal loss \( L(S : \nu) \). Recall that under the standard assumptions \( \ell^* \in L^1(\nu) \) is a measurable upper bound on \( \sup_x |\ell(x, y)| \).

**Lemma 3.1.** Under the standard assumptions, \( L(S : \nu) \in (-\infty, \infty) \).
Proof. As $X$ is non-empty and compact and $S$ is continuous, there exists at least one measure $\mu \in \mathcal{M}(X, S)$. Thus $\mu \otimes \nu$ is in $\mathcal{J}(S : \nu)$, and in particular, $\mathcal{J}(S : \nu)$ is non-empty. By assumption, $|\int \ell \, d\lambda| \leq \int \ell^* \, d\nu < \infty$, for each $\lambda \in \mathcal{J}(S : \nu)$, and as this bound is independent of $\lambda$, the lemma follows. □

The proof of Theorem 1.2 relies on Kingman’s subadditive ergodic theorem [19, 20, 21] and a weak compactness argument. We first establish the result when $T$ is continuous and then deduce the general result from this special case.

**Proof of Theorem 1.2.** We begin by establishing that, for $\nu$-almost every $y$,

$$ \lim \inf_{n} \frac{1}{n} \sum_{k=0}^{n-1} \ell(S^k x, T^k y) = \sup_{n} \frac{1}{n} \int \left( \inf_{x \in X} \sum_{k=0}^{n-1} \ell(S^k x, T^k y) \right) \, d\nu(y). $$

For each $n \in \mathbb{N}$ and $y \in \mathcal{Y}$, define

$$ G_n(y) = \inf_{x \in X} \sum_{i=0}^{n-1} \ell(S^i x, T^i y). $$

Note that the sequence $(G_n)_{n \geq 1}$ is super-additive in the sense that

$$ G_{m+n}(y) \geq \inf_{x \in X} \sum_{i=0}^{m-1} \ell(S^i x, T^i y) + \inf_{x \in X} \sum_{i=m}^{m+n-1} \ell(S^i x, T^i y) $$

$$ \geq G_m(y) + G_n(T^m y). $$

By Kingman’s subadditive ergodic theorem applied to $(-G_n)_{n \geq 1}$, there exists $\gamma \in (-\infty, \infty]$ such that for $\nu$-almost every $y$,

$$ \lim \frac{G_n(y)}{n} = \gamma = \sup_{n} \frac{1}{n} \int G_n \, d\nu. $$

This equation establishes the existence of the limit in (1.1) and the equality in (3.1).

We now establish that $L(S : \nu) = \gamma$. Let $\lambda$ be any element of $\mathcal{J}(S : \nu)$. As $\lambda$ is invariant under $S \times T$, for each $n \geq 1$,

$$ \int \ell \, d\lambda = \frac{1}{n} \int \sum_{i=0}^{n-1} \ell(S^i x, T^i y) \, d\lambda \geq \frac{1}{n} \int G_n \, d\nu. $$

It then follows from (3.2) that $\int \ell \, d\lambda \geq \gamma$. As $\lambda \in \mathcal{J}(S : \nu)$ was arbitrary, we conclude that $\gamma \leq \inf_\lambda \int \ell \, d\lambda = L(S : \nu)$, where the infimum is taken over $\lambda$ in $\mathcal{J}(S : \nu)$. 

To complete the proof, we establish the existence of a joining \( \lambda \in J(S: \nu) \) such that \( \int \ell \, d\lambda \leq \gamma \). To do this, we construct a suitable sequence of empirical measures on \( \mathcal{X} \times \mathcal{Y} \) and then use a weak compactness argument to identify a limit \( \lambda \) with the desired properties. Assume for the moment that \( T \) is continuous.

For each \( m \geq 1 \), let \( K_m \subset \mathcal{Y} \) be a compact set such that \( \nu(K_m) > 1 - \frac{1}{m} \). Using the arguments above, Lemma 2.2 and the ergodic theorem, one may identify a measurable set \( E \subseteq \mathcal{Y} \) such that \( \nu(E) = 1 \) and for every \( y \in E \), Equation (3.2) and each of the following relations holds as \( n \) tends to infinity:

\[
\nu_n := \frac{1}{n} \sum_{k=0}^{n-1} \delta_{T_k y} \text{ converges weakly to } \nu; \tag{3.4}
\]

\[
\nu_n(K_m) \to \nu(K_m) \text{ for each } m \geq 1; \tag{3.5}
\]

\[
\int_{\ell^* > m} \ell^* \, d\nu_n \to \int_{\ell^* > m} \ell^* \, d\nu \text{ for each } m \geq 1. \tag{3.6}
\]

Elements of the set \( E \) will be referred to as \( \nu \)-generic points.

Let \( y \) be a \( \nu \)-generic point in \( \mathcal{Y} \). By (3.2), there exists a sequence \( (x_n)_{n \geq 1} \) in \( \mathcal{X} \) such that \( n^{-1} \sum_{k=0}^{n-1} c(S^k x_n, T^k y) \to \gamma \). For each \( n \geq 1 \), define the discrete measure

\[
\lambda_n = \frac{1}{n} \sum_{k=0}^{n-1} \delta_{(S^k x_n, T^k y)}
\]
on \( \mathcal{X} \times \mathcal{Y} \). Note that \( \lim_n \int \ell \, d\lambda_n = \gamma \). We claim that the family \( \{ \lambda_n : n \in \mathbb{N} \} \) is tight. To this end, let \( \delta > 0 \) be given and choose \( N > 1/\delta \). By definition, \( K_N \) is compact and \( \nu(K_N) > 1 - \delta \). As \( y \in E \), for all \( n \) sufficiently large, \( \lambda_n(\mathcal{X} \times K_N) = \nu_n(K_N) > 1 - \delta \). As \( \delta > 0 \) was arbitrary and \( \mathcal{X} \) is compact, the claim follows.

Let \( \lambda \) be a weak limit of the family \( \{ \lambda_n : n \in \mathbb{N} \} \). We claim that \( \lambda \) is in \( J(S: \nu) \) and that \( \lambda \) achieves the infimum in the definition of \( L(S: \nu) \) (see 1.1). By passing to a subsequence if necessary, assume that \( \lambda_n \Rightarrow \lambda \). Let \( f \) be in \( C_b(\mathcal{X} \times \mathcal{Y}) \). Under the assumption that \( S \) and \( T \) are continuous, the composition \( f \circ (S \times T) \) is in \( C_b(\mathcal{X} \times \mathcal{Y}) \), and weak convergence then implies that

\[
\int f \circ (S \times T) \, d\lambda = \lim_n \int f \circ (S \times T) \, d\lambda_n = \lim_n \frac{1}{n} \sum_{i=1}^{n} f(S^i x_n, T^i y) \tag{3.12}
\]

\[
= \lim_n \frac{1}{n} \sum_{i=0}^{n-1} f(S^i x_n, T^i y) = \lim_n \int f \, d\lambda_n = \int f \, d\lambda.
\]
As \( f \in C_b(X \times Y) \) was arbitrary, it follows that \( \lambda \) is invariant under \( S \times T \). In particular, \( \lambda \in \mathcal{M}(X \times Y, S \times T) \) and therefore \( m_X(\lambda) \) is in \( \mathcal{M}(X, S) \). Furthermore, as \( y \in E, m_Y(\lambda_n) = \nu_n \) converges weakly to \( \nu \), and therefore \( m_Y(\lambda) = \nu \). Thus, \( \lambda \) is in \( \mathcal{J}(S : \nu) \).

In light of the fact that \( \lim_n \int \ell d\lambda_n = \gamma \), it suffices to show that

\[
\text{(3.7)} \quad \lim_n \int \ell d\lambda_n \geq \int \ell d\lambda.
\]

If the loss \( \ell \) were bounded from below, this inequality would follow from the Portmanteau Theorem for weak convergence, since we have assumed that it is lower semicontinuous. For unbounded losses, we appeal to a truncation argument. Though the details are somewhat routine, we include them here for completeness. For \( m \in \mathbb{N} \), define the truncated loss

\[
\ell_m(x, y) = \begin{cases} 
\ell(x, y), & \text{if } |\ell(x, y)| \leq m \\
-m, & \text{if } \ell(x, y) \leq -m \\
-m, & \text{if } \ell(x, y) \geq m
\end{cases}
\]

Note that \( |\ell_m| \leq |\ell| \) and that \( \ell_m \to \ell \) as \( m \) tends to infinity. The integrability of \( \ell \) with respect to \( \lambda \) follows from that of \( \ell^* \) with respect to \( \nu \), and the dominated convergence theorem then ensures that \( \int \ell_m d\lambda \to \int \ell d\lambda \). Moreover, with \( \nu_n \) defined as in (3.4), it follows from the choice of \( y \) that

\[
\limsup_n \int |\ell - \ell_m| d\lambda_n \leq \limsup_n \int_{\ell^* > m} \ell^* d\nu_n = \int_{\ell^* > m} \ell^* d\nu.
\]

In order to establish (3.7), let \( \epsilon > 0 \) be fixed. By virtue of the results in the previous paragraph, there exist integers \( m \) and \( n_1 \) sufficiently large that for each \( n \geq n_1 \)

\[
\left| \int \ell_m d\lambda - \int \ell d\lambda \right| < \epsilon/3 \quad \text{and} \quad \int |\ell - \ell_m| d\lambda_n < \epsilon/3.
\]

Moreover, as \( \lambda_n \Rightarrow \lambda \) and \( \ell_m \) is lower semi-continuous and bounded, there exists \( n_2 \geq n_1 \) such that for each \( n \geq n_2 \),

\[
\int \ell_m d\lambda - \int \ell_m d\lambda_n < \epsilon/3.
\]

Combining the inequalities above, a straightforward bound shows that

\[
\int \ell d\lambda - \int \ell d\lambda_n < \epsilon
\]

for \( n > n_2 \). As \( \epsilon > 0 \) was arbitrary, the inequality (3.7) is established, and we conclude that \( \gamma \geq L(S : \nu) \).
Suppose now that the transformation $T$ is Borel measurable but not continuous. Let $\tilde{\nu}$ be the process measure on $\mathcal{Y}^\infty$ generated by $(\mathcal{Y}, T, \nu)$ (see Section 2.3), and let $\tilde{\ell} : \mathcal{X} \times \mathcal{Y}^\infty \to \mathbb{R}$ be defined by $\tilde{\ell}(x, y) = \ell(x, y_0)$. Note that $\tilde{\ell}$ is lower semicontinuous and that $\sup_y |\tilde{\ell}(x, y)|$ is bounded above by a $\tilde{\nu}$-integrable function. As the left-shift $\tau : \mathcal{Y}^\infty \to \mathcal{Y}^\infty$ is continuous, we may apply the arguments above to the systems $(\mathcal{X}, S)$ and $(\mathcal{Y}^\infty, \tau, \tilde{\nu})$ with loss $\tilde{\ell}$. Equation (3.2) and inequality (3.3) are the same for the original and shift systems. As for the inequality $\gamma \geq L(S : \nu)$, the arguments above show that there is a joining $\tilde{\lambda} \in J(S : \tilde{\nu})$ such that $\gamma = \int \tilde{\ell} d\tilde{\lambda} = \int \ell dm_{\mathcal{X} \times \mathcal{Y}}(\tilde{\lambda})$. By Lemma 2.1, $\lambda = m_{\mathcal{X} \times \mathcal{Y}}(\tilde{\lambda})$ is in $J(S : \nu)$. This establishes (1.1) and the existence of a joining $\lambda$ in $J(S : \nu)$ that achieves the infimum in the definition of $L(S : \nu)$.

4. General results for inference

The present section is devoted to the proof of Theorem 1.6. We begin with several preliminary lemmas, the first of which establishes the existence of optimal tracking schemes.

**Lemma 4.1.** If the systems $(\mathcal{X}, S)$ and $(\mathcal{Y}, T, \nu)$ and the loss $\ell$ satisfy the standard assumptions of Section 7, then there exists a measurable sequence of functions $f_n : \mathcal{Y}^n \to \mathcal{X}$ satisfying (1.2).

**Proof.** For each $n \geq 1$, define $\ell_n : \mathcal{X} \times \mathcal{Y}^n \to \mathbb{R}$ by $\ell_n(x, y_{0}^{n-1}) = \sum_{k=0}^{n-1} \ell(S^k x, y_k)$. Then it is easy to see that $\ell_n = s_n \circ \psi_n$, where

$$\psi_n(x, y_{0}^{n-1}) = ((x, \ldots, S^{n-1} x), y_{0}^{n-1})$$

and $s_n(x_{0}^{n-1}, y_{0}^{n-1}) = \sum_{k=0}^{n-1} \ell(x_k, y_k)$. Our assumptions on $S$ and $\ell$ ensure that $\psi_n$ is continuous and that $s_n$ is lower semicontinuous, and therefore $\ell_n$ is lower semicontinuous. It follows from 6 Proposition 7.33, p. 153] that there exists a Borel measurable function $f_n : \mathcal{Y}^n \to \mathcal{X}$ such that

$$\ell_n(f_n(y_{0}^{n-1}), y_{0}^{n-1}) = \inf_{x \in \mathcal{X}} \ell_n(x, y_{0}^{n-1}).$$

The definition of $\ell_n$ ensures that $(f_n)_{n \geq 1}$ satisfies the conclusions of the lemma. 

**Lemma 4.2.** Under the standard assumptions,

$$L(S : \nu) = \inf_{\theta \in \Theta} L(S_\theta : \nu)$$

where $S_\theta$ is the restriction of $S$ to $\mathcal{X}_\theta = \varphi^{-1}\{\theta\}$. Furthermore, the infimum is attained.
Proof. Since \((\mathcal{X}_0, S_0)\) is a subsystem of \((\mathcal{X}, S)\), it is immediate that \(L(S : \nu) \leq L(S_0 : \nu)\) for each \(\theta \in \Theta\). Thus, it suffices to show that \(L(S_0 : \nu) \leq L(S : \nu)\) for some \(\theta \in \Theta\). By Remark 1.4, there exists an ergodic joining \(\lambda\) in \(J_{\text{min}}(S : \nu)\). Define \(h : \mathcal{X} \times \mathcal{Y} \to \Theta\) by \(h = \varphi \circ \text{proj}_\mathcal{X}\) and let \(\eta = \lambda \circ h^{-1}\) be the push-forward measure of \(\lambda\) on \(\Theta\). The assumption that \(\varphi \circ S = \varphi\) ensures that \(\eta\) is ergodic with respect to the identity transformation on \(\Theta\), and therefore \(\eta\) is necessarily a point mass concentrated at some parameter \(\theta \in \Theta\). In particular, \(\lambda(\mathcal{X}_0 \times \mathcal{Y}) = 1\), so that \(\lambda \in J(S_0 : \nu)\). Thus \(L(S_0 : \nu) \leq \int \ell \, d\lambda = L(S : \nu)\), and the result follows.

Lemma 4.3. The set \(\Theta_{\text{min}} \subseteq \Theta\) is non-empty and compact.

Proof. As the infimum in (4.1) is achieved, \(\Theta_{\text{min}}\) is non-empty. Since \(\Theta\) is compact by assumption, it suffices to show that \(\Theta_{\text{min}}\) is closed. Let \((\theta_n)_{n \geq 1}\) be a sequence in \(\Theta_{\text{min}}\) that converges to a parameter \(\theta \in \Theta\). It follows from (4.2) that for each \(n\), there is a joining \(\lambda_n \in J_{\text{min}}(S : \nu)\) such that \(\lambda_n(\varphi^{-1}\{\theta_n\} \times \mathcal{Y}) = 1\). As \(J_{\text{min}}(S : \nu)\) is compact, the sequence \((\lambda_n)_{n \geq 1}\) has a convergent subsequence. Passing to a subsequence if necessary, suppose that \(\lambda_n\) converges to \(\lambda \in J_{\text{min}}(S : \nu)\).

Define \(h : \mathcal{X} \times \mathcal{Y} \to \Theta\) by \(h = \varphi \circ \text{proj}_\mathcal{X}\), and consider the associated push-forward measures \(\eta_n = \lambda_n \circ h^{-1}\) and \(\eta = \lambda \circ h^{-1}\) on \(\Theta\). Note that \(\eta_n \Rightarrow \eta\), as \(\lambda_n \Rightarrow \lambda\) and \(h\) is continuous. Our choice of \((\theta_n)\) ensures that \(\eta_n = \delta_{\theta_n} \Rightarrow \delta_\theta\), and as weak limits are unique, \(\eta = \delta_\theta\). Thus \(\lambda(\varphi^{-1}\{\theta\} \times \mathcal{Y}) = 1\), and as \(\lambda \in J_{\text{min}}(S : \nu)\) we conclude that \(\theta\) is an element of \(\Theta_{\text{min}}\).

In the following proof, we make use of the ergodic decomposition of an invariant measure and a related lemma, details of which may be found in Appendix A.

**Proof of Theorem 1.6.** Let \((\theta_n)_{n \geq 1}\) be a sequence of estimators of the form \(\theta_n = \varphi \circ f_n\), where \(\{f_n\}\) is a sequence of measurable functions satisfying (1.2), which exists by Lemma 4.1. As in the proof of Theorem 1.2, fix a set \(E \subset \mathcal{Y}\) of \(\nu\)-generic points having full measure. Let \(y \in E\), and then let \(\hat{\theta}_n = \theta_n(y, \ldots, T^{n-1}y)\). For \(n \geq 1\), define the state \(x_n = f_n(y, \ldots, T^{n-1}y)\) and the associated empirical measure \(\lambda_n = n^{-1} \sum_{k=0}^{n-1} \delta_{(S^k x_n, T^k y)}\) on \(\mathcal{X} \times \mathcal{Y}\). By arguments identical to those in the proof of Theorem 1.2, one may show that \((\lambda_n)\) is tight and that all of its weak limit points are in \(J_{\text{min}}(S : \nu)\).

Let \(O \subseteq \Theta\) be an open neighborhood of \(\Theta_{\text{min}}\). Define the function \(\psi : \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to [0, 1]\) by \(\psi(\lambda) = \lambda(\varphi^{-1}(O) \times \mathcal{Y})\), and let \(V = \psi^{-1}\left(\frac{1}{2}, 1\right)\). As \(\varphi^{-1}(O) \times \mathcal{Y}\) is open, \(\psi\) is lower semi-continuous, and therefore \(V\) is open in \(\mathcal{M}(\mathcal{X} \times \mathcal{Y})\). We claim that \(J_{\text{min}}(S : \nu) \subset V\). To see this, let
\( \lambda \in \mathcal{J}_{\text{min}}(S : \nu) \) have ergodic decomposition \( \lambda = \int \eta \, d\xi \). By Lemma A.3, \( \xi \)-almost every measure \( \eta \) is an ergodic element of \( \mathcal{J}_{\text{min}}(S : \nu) \). Define \( h : \mathcal{X} \times \mathcal{Y} \to \Theta \) by \( h = \varphi \circ \text{proj}_x \). For every ergodic \( \eta \in \mathcal{J}_{\text{min}}(S : \nu) \), the push-forward measure \( \eta \circ h^{-1} \) must be ergodic with respect to the identity map on \( \Theta \), as in the proof of Lemma 4.2, and hence \( \eta \circ h^{-1} = \delta_\theta \) for some \( \theta \in \Theta \). Thus, for every such \( \eta \) there is a parameter \( \theta \in \Theta_{\text{min}} \) such that \( \eta(\varphi^{-1}\{\theta\} \times \mathcal{Y}) = 1 \), and therefore \( \eta(\varphi^{-1}(\Theta_{\text{min}}) \times \mathcal{Y}) = 1 \). It follows that \( \psi(\lambda) = 1 \), and we conclude that \( \lambda \) is in \( V \).

Since all the limit points of the family \( \{\lambda_n\}_{n \geq 1} \) are in the open set \( V \), there exists an integer \( n_1 \) such that \( \lambda_n \in V \) for all \( n \geq n_1 \). By construction, \( \lambda_n(\varphi^{-1}(\mathcal{O}) \times \mathcal{Y}) = \delta_{\theta_n}(\mathcal{O}) \). Thus for each \( n \geq n_1 \), \( \delta_{\theta_n}(\mathcal{O}) = \psi(\lambda_n) > 1/2 \), which implies that \( \theta_n \) is in \( \mathcal{O} \) as desired.

Let us now show that any parameter in \( \Theta_{\text{min}} \) is the limit of an optimal \( \varphi \)-estimation scheme. Let \( \theta_0 \) be any element of \( \Theta_{\text{min}} \). We will show that there exists a \( \varphi \)-optimal estimation scheme that converges almost surely to \( \theta_0 \). By Lemma 4.2, \( L(S : \nu) = L(S_{\theta_0} : \nu) \), and it follows from Remark 1.4 that there exists an ergodic joining \( \lambda \in \mathcal{J}_{\text{min}}(S_{\theta_0} : \nu) \). By Birkhoff’s ergodic theorem, there exists a set \( E \subset \mathcal{X}_0 \times \mathcal{Y} \) of \( \lambda \) measure one such that if \( (x, y) \in E \) then

\[
\lim_{n} \frac{1}{n} \sum_{k=0}^{n-1} \ell(S^kx, T^ky) = \int \ell \, d\lambda,
\]

which is equal to \( L(S : \nu) \) by construction. By the regularity of \( \lambda \), \( E \) contains a \( \sigma \)-compact set \( F \) of the same measure. The measurable selection theorem of Brown and Purves [5, Theorem 1] then implies that there is a measurable function \( f : \mathcal{Y} \to \mathcal{X}_{\theta_0} \) such that \( (f(y), y) \in F \) for \( \nu \)-almost every \( y \in \mathcal{Y} \). Let \( f_n : \mathcal{Y}^n \to \mathcal{X} \) be given by \( f_n(y_0^{n-1}) = f(y_0) \). By (4.2), the sequence \( \{f_n\}_{n \geq 1} \) satisfies (1.2), and therefore the constant estimator \( \theta_0 = \varphi \circ f_n \) is an optimal \( \varphi \)-estimation scheme. In particular, there exists an optimal \( \varphi \)-estimation scheme that converges to \( \theta_0 \), as desired.

5. Quantized Observations

This section is devoted to the proofs of our results concerning the estimation of transformations from quantized observations. We refer to the objects defined in Section 1.3 throughout this section.

Recall that the state space \( \mathcal{X} \) of the reference system is defined to be the closure of the set \( \{ (\theta, \text{lab}(\theta, u) : \theta \in \Theta, u \in \mathcal{U} \} \) inside of \( \Theta \times \{0, 1\}^\mathbb{N} \). Hence \( \mathcal{X} \) is compact and metrizable, and it is easy to see that \( \mathcal{X} \) is invariant under the map \( \text{Id} \times \tau \), where \( \tau \) is the left-shift map on \( \{0, 1\}^\mathbb{N} \).
Further recall that the family of transformations in \( \mathcal{R} \) and the topology on \( \Theta \) enter indirectly through the definition of \( \mathcal{X} \), which is used to define the target set \( \Theta_1 \).

First we establish some additional notation and another interpretation of \( h(\mathcal{R}) \), as the topological entropy of a dynamical system. Let \( \mathcal{L} \) be defined by

\[
\mathcal{L} = \text{closure of } \{(\pi(R^k_{\theta}u))_{k\geq 0} : \theta \in \Theta, u \in \mathcal{U}\} \text{ in } \{0,1\}^\mathbb{N}.
\]

Since \( \mathcal{L} \) is a closed subset of \( \{0,1\}^\mathbb{N} \), it is compact, and it is easy to see that \( \mathcal{L} \) is invariant under the left shift \( \tau \). Thus the pair \( (\mathcal{L},\tau) \) is a topological dynamical system. Now we may observe that the quantity \( h(\mathcal{R}) \) defined in Section 1.5 is actually the topological entropy of the system \( (\mathcal{L},\tau) \). See [47] for an introduction to topological entropy for dynamical systems.

In what follows, we find it useful to state our results in terms of the \( d \)-distance introduced by Ornstein [35, 37, 38] in the context of the isomorphism theory for Bernoulli processes. For a fixed shift-invariant measure \( \nu \) on \( \{0,1\}^\mathbb{N} \) and each \( \theta \in \Theta \), define

\[
d(\theta : \nu) = \inf_{\mu} d(\mu,\nu),
\]

where the infimum is taken over all \( \tau \)-invariant probability measures \( \mu \) such that \( \mu(\mathcal{X}_\theta) = 1 \). An application of Theorem 1.6 yields the following result.

**Theorem 5.1.** Let \( (B_i)_{i \geq 0} \) be a stationary ergodic \( \{0,1\} \)-valued process having distribution \( \nu \) on \( \{0,1\}^\mathbb{N} \). If the sequence of estimators \( (\hat{\theta}_n)_n \) satisfies (1.10), then \( \hat{\theta}_n(B_0,\ldots,B_{n−1}) \) converges almost surely to the set

\[
\Theta_0 = \arg\min_{\theta \in \Theta} d(\theta : \nu).
\]

**Proof.** By definition, \( \Theta_{\min} = \arg\min_{\theta} \mathcal{L}(S_{\theta} : \nu) \). Using the correspondences in Table 3 it is easy to see that \( S_{\theta} \) is the restriction of \( \text{Id} \times \tau \) to the set \( \{\theta\} \times \mathcal{X}_\theta \), and therefore

\[
\mathcal{L}(S_{\theta} : \nu) = \inf_{\lambda \in \mathcal{L}(S_{\theta} : \nu)} \ell d\lambda = \inf_{\mu \in \mathcal{M}(\mathcal{X}_\theta,\tau)} d(\mu,\nu) = d(\theta : \nu).
\]

Thus \( \Theta_{\min} = \arg\min_{\theta} d(\theta : \nu) = \Theta_0 \), as desired. \( \square \)

By associating parameters \( \theta \) with the set of \( \tau \)-invariant measures on \( \mathcal{X}_\theta \), we may view the set \( \Theta_0 \) as the projection of the observation process onto the parameter set \( \Theta \) under the \( d \)-metric. The following lemma, which characterizes when \( d(\theta : \eta) \) equals zero, will be used to prove our consistency results.
Lemma 5.2. If \( \nu \) is a shift-invariant Borel probability measure on \( \{0, 1\}^\mathbb{N} \), then \( \overline{d}(\theta : \nu) = 0 \) if and only if \( \nu(X_\theta) = 1 \).

Proof. If \( \nu(X_\theta) = 1 \), then clearly \( 0 \leq \overline{d}(\theta : \nu) \leq \overline{d}(\nu, \nu) = 0 \). Now suppose that \( \overline{d}(\theta : \eta) = 0 \). By Theorem 1.2, the infimum defining \( \overline{d}(\theta : \eta) = L(S_\theta : \eta) \) is achieved, and it follows that there is a measure \( \mu \in \mathcal{M}(X_\theta, \tau) \) such that \( \overline{d}(\mu, \nu) = 0 \). As \( \overline{d} \) is a metric, \( \mu = \nu \) and therefore \( \nu(X_\theta) = 1 \). \( \square \)

Proof of Theorem 1.9, Part (1). By Theorem 5.1, the estimates \( \hat{\theta}_n(Y_0, \ldots, Y_{n-1}) \) converge almost surely to the set \( \Theta_0 = \arg\min_{\theta \in \Theta} \overline{d}(\theta : \nu_0) \). Thus it suffices to show that \( \Theta_0 = \Theta_1 \). As the measure \( \nu_0 \) of the observation process is supported on \( X_{\theta_0} \), we have \( \overline{d}(\theta_0 : \nu_0) = 0 \). Then by Lemma 5.2 we see that \( \Theta_0 = \{ \theta : \overline{d}(\theta : \nu_0) = 0 \} \) is identical to \( \Theta_1 = \{ \theta : \nu_0(X_\theta) = 1 \} \). \( \square \)

Before turning to the proof of Part (2) of Theorem 1.9 we require some additional definitions.

Remark 5.3. In the proof of Theorem 1.9 below, we make use of a standard construction, called the relatively independent joining, to “glue together” two joinings along a common factor (see [6, Section 3.1]). In more detail, suppose we have two measure-preserving systems \((\mathcal{U}_i, R_i, \eta_i)\) for \( i = 1, 2 \). Further suppose that these systems have a common factor \((\mathcal{U}, R, \eta)\), meaning that there exist measurable maps \( \psi_i : \mathcal{U}_i \to \mathcal{U} \) such that \( \eta = \eta_i \circ \psi_i^{-1} \) and \( \psi_i \circ R_i = R \circ \psi_i \) for \( i = 1, 2 \). Now let \( \eta_i = \int \eta_{i,u} d\eta(u) \) be the disintegration of \( \eta_i \) over \( \eta \), and define the measure \( \lambda \) on \( \mathcal{U}_1 \times \mathcal{U}_2 \) by

\[
\lambda = \int \eta_{1,u} \otimes \eta_{2,u} d\eta(u).
\]

Then it is not difficult to check that \( \lambda \) is a joining of \((\mathcal{U}_1, R_1, \eta_1)\) and \((\mathcal{U}_2, R_2, \eta_2)\) such that \( \lambda\{(u_1, u_2) \in \mathcal{U}_1 \times \mathcal{U}_2 : \psi_1(u_1) = \psi_2(u_2) \} = 1 \).

Remark 5.4. The proof of Theorem 1.9 also requires some elementary facts concerning the entropy of dynamical systems (see [47] for a thorough treatment of the subject). Let \( \Sigma \subset \{0, 1\}^\mathbb{N} \) be a closed, shift-invariant set of label sequences. The topological entropy \( h_{\text{top}}(\Sigma, \tau) \) of
the system \((\Sigma, \tau)\) is given in \([1.12]\). If \(\eta\) is a shift-invariant Borel probability measure on such a set \(\Sigma\), then the measure theoretic (Kolmogorov-Sinai) entropy of the system \((\Sigma, \tau, \eta)\) is defined by

\[
h(\eta) = \lim_{n \to \infty} \frac{1}{n+1} \sum_{a_0^n \in \Sigma_n} \eta([a_0^n]) \log \eta([a_0^n]),
\]

where \(\Sigma_n = \{a_0^n \in \{0,1\}^{n+1} : a \in \Sigma\}\) and \([a_0^n]\) denotes the cylinder set of sequences \(a \in \{0,1\}^N\) whose first \(n+1\) coordinates are \(a_0^n\). The well-known Variational Principle (see \([17]\)) states that

\[
h_{\text{top}}(\Sigma, \tau) = \sup_{\eta \in \mathcal{M}(\Sigma, \tau)} h(\eta).
\]

Thus if \(h_{\text{top}}(\Sigma, \tau) = 0\), then \(h(\eta) = 0\) for any measure \(\eta\) in \(\mathcal{M}(\Sigma, \tau)\).

**Proof of Theorem 1.9, Part (2).** By Theorem 5.1 any sequence of estimates satisfying (1.10) converges almost surely to the parameter set \(\Theta_0 = \text{argmin}_{\theta \in \Theta} \bar{d}(\theta : \nu)\), where \(\nu\) is the measure of the observation process \((Y_i)_{i \geq 0}\), which involves errors. By contrast, we have \(\Theta_1 = \{\theta \in \Theta : \nu_0(L_\theta) = 1\}\), where \(\nu_0\) is the measure for the error-free process \((\pi(R_0 U))_{i \geq 0}\). It therefore suffices to show that \(\theta\) minimizes \(\bar{d}(\theta : \nu)\) if and only if \(\nu_0(\mathcal{X}_\theta') = 1\).

Let \(p = \mathbb{P}(\epsilon_0 = 1)\), which is less than 1/2 by hypothesis. We claim that for each \(\theta \in \Theta\),

\[
\bar{d}(\theta : \nu) \geq p + \bar{d}(\theta : \nu_0)(1 - 2p).
\]

To see this, fix \(\theta \in \Theta\). Then Theorem 1.2 with \((\mathcal{X}, S) = (\mathcal{X}_\theta, \tau)\) ensures that there is an optimal joining \(\lambda_1\) of \(\theta\) and \(\nu\). In detail, there exists \(\lambda_1 \in \mathcal{M}(\mathcal{X}_\theta \times \{0,1\}^N, \tau \times \tau)\) such that its first marginal, \(\mu\) say, is supported on \(\mathcal{X}_\theta\), its second marginal is equal to \(\nu\), and \(\bar{d}(\theta : \nu) = \lambda_1\{(a,b) : a_0 \neq b_0\}\).

Let \(\eta\) be the process measure for the noise process \((\epsilon_i)_{i \geq 0}\), and let \(\lambda_2\) denote the product measure \(\nu_0 \otimes \eta\). Define the maps \(\psi_1 : \{0,1\}^N \times \{0,1\}^N \to \{0,1\}^N\) and \(\psi_2 : \{0,1\}^N \times \{0,1\}^N \to \{0,1\}^N\) by \(\psi_1(a,b) = b\) and \(\psi_2(v,\epsilon) = (v \oplus \epsilon)_{i \geq 0}\). Note that \(\nu\) is a common factor of \(\lambda_1\) and \(\lambda_2\) under the maps \(\psi_1\) and \(\psi_2\), respectively. Using the construction discussed in Remark 5.3 one may construct a joining \(\tilde{\lambda}\) of \(\lambda_1\) and \(\lambda_2\) such that

\[
\tilde{\lambda}\{(a,b,v,\epsilon) : b_i = v_i \oplus \epsilon_i \text{ for all } i \geq 0\} = 1.
\]

Let \((A_i, B_i, V_i, \epsilon_i)_{i \geq 0}\) denote the multi-label process having distribution \(\tilde{\lambda}\). By construction, the following hold:

1. \((A_i)_{i \geq 0} \sim \mu\) and is supported on \(\mathcal{X}_\theta\);
(2) \((B_i)_{i \geq 0} \sim \nu;\)
(3) \((V_i)_{i \geq 0} \sim \nu_0;\)
(4) \((\epsilon_i)_{i \geq 0} \sim \eta\) is a copy of the i.i.d. noise process;
(5) \((A_i, B_i)_{i \geq 0} \sim \lambda_1\), the joining of \(\mu\) and \(\nu;\)
(6) \(B_i = V_i \oplus \epsilon_i\) almost surely.

From these properties and elementary arguments we see that
\[
\overline{d}(\theta : \nu) = \lambda_1(A_0 \neq B_0) = \tilde{\lambda}(A_0 \neq (V_0 \oplus \epsilon_0))
\geq \tilde{\lambda}\left(\{\epsilon_0 = 1, A_0 = V_0\} \cup \{\epsilon_0 = 0, A_0 \neq V_0\}\right)
= \tilde{\lambda}(\epsilon_0 = 1, A_0 = V_0) + \tilde{\lambda}(\epsilon_0 = 0, A_0 \neq V_0).
\]

As the measures \(\mu\) and \(\nu_0\) are supported on \(\mathcal{L}\), the assumption that \(h(\mathcal{R}) = 0\) implies that \(h(\mu) = 0\) and \(h(\nu_0) = 0\) (see Remark 5.4). Let \(\lambda_3\) be the joining of \(\mu\) and \(\nu_0\) given by the marginal distribution of \(\tilde{\lambda}\) on \((a, v)\). By a standard bound on entropy, \(h(\lambda_3) \leq h(\mu) + h(\nu_0) = 0\), and therefore \(h(\lambda_3) = 0\). It follows from a classical result of Furstenberg [9, Theorem I.2] that the only joining between the zero-entropy measure \(\lambda_3\) and the i.i.d. measure \(\eta\) is the product (independent) joining \(\lambda_3 \otimes \eta\). Consequently, \((A_0, V_0)\) and \(\epsilon_0\) are independent under \(\lambda\). Therefore
\[
\tilde{\lambda}(\epsilon_0 = 1, A_0 = V_0) = \tilde{\lambda}(\epsilon_0 = 1) \tilde{\lambda}(A_0 = V_0) = p (1 - \tilde{\lambda}(A_0 \neq V_0)),
\]
and
\[
\tilde{\lambda}(\epsilon_0 = 0, A_0 \neq V_0) = \tilde{\lambda}(\epsilon_0 = 0) \tilde{\lambda}(A_0 \neq V_0) = (1 - p) \tilde{\lambda}(A_0 \neq V_0).
\]
Combining the previous three displays gives
\[
\overline{d}(\theta : \nu) = \lambda_1(A_0 \neq B_0)
\geq p (1 - \tilde{\lambda}(A_0 \neq V_0)) + (1 - p)\tilde{\lambda}(A_0 \neq V_0)
= p + \tilde{\lambda}(A_0 \neq V_0)(1 - 2p).
\]
Under the joining \(\tilde{\lambda}\), \((A_i)_{i \geq 0}\) is distributed according to \(\mu\), which is supported on \(\mathcal{X}_\theta\), and \((V_i)_{i \geq 0}\) is distributed according to \(\nu_0\). Thus \(\tilde{\lambda}(A_0 \neq V_0) \geq \overline{d}(\theta : \nu_0)\), and the inequality (5.1) follows from the previous display as \(p < 1/2\).

It follows from (5.1) that \(\overline{d}(\theta : \nu) \geq p\) for all \(\theta\). We now show that \(\overline{d}(\theta_0 : \nu) = p\), from which it follows that \(\Theta_0 = \{\theta : \overline{d}(\theta : \nu) = p\}\). Let \(\lambda = \nu_0 \otimes \eta \in \mathcal{M}(\{0, 1\}^\mathbb{N} \times \{0, 1\}^\mathbb{N})\) and let \(\psi : \{0, 1\}^\mathbb{N} \times \{0, 1\}^\mathbb{N} \rightarrow \{0, 1\}^\mathbb{N} \times \{0, 1\}^\mathbb{N}\) be defined by \(\psi(a, \epsilon) = (a, (a_i \oplus \epsilon_i)_{i \geq 0})\). It is straightforward to show that \(\lambda_4 = \lambda \circ \psi^{-1}\) is a joining of \(\nu_0\) with \(\nu\).
such that $\lambda_4(\{(a, b) : a_0 \neq b_0\}) = p$, and therefore $\overline{d}(\theta_0 : \nu) = p$ as desired.

Let $\theta \in \Theta_0$. The arguments above show that $\overline{d}(\theta : \nu) = p$, and then it follows from (5.1) that $\overline{d}(\theta : \nu_0) = 0$. Hence $\nu_0(X_{\theta}) = 1$ by Lemma 5.2 and we conclude that $\theta \in \Theta_1$. This shows that $\Theta_0 \subseteq \Theta_1$. For the reverse inclusion, we note that if $\nu_0(X_{\theta}) = 1$, then (5.1) and the joining $\lambda_4$ can be used to show that $\overline{d}(\theta : \nu) = p$ (as we did for $\theta_0$), which implies that $\theta$ is in $\Theta_0$.

\[ \square \]

If the partition $\pi$ does not not resolve differences between the generative transformation $R_{\theta_0}$ and other transformations on the support of $\nu_0$, then $\Theta_1$ may not be equal to $\{\theta_0\}$. The following result provides conditions under which $\Theta_1$ is a singleton.

**Proposition 5.5.** Suppose that for all $\theta' \neq \theta_0$ there exists a neighborhood $\mathcal{O}$ of $\theta'$ and an integer $N$ depending on $\theta_0$ and $\mathcal{O}$ such that for all $u, v \in \mathcal{U}$ and all $\theta \in \mathcal{O}$, $\pi(R^k_{\theta_0}u) \neq \pi(R^k_{\theta}v)$ for some $k \in [0, N]$. Then $\Theta_1 = \{\theta_0\}$.

**Proof.** Suppose the hypotheses of the proposition hold, and let $\theta' \neq \theta_0$. We will show that $\nu_0(X_{\theta'}) = 0$, and therefore $\theta' \notin \Theta_1$. Let $\mathcal{O}$ and $N$ be as in the statement of the proposition. For $\theta \in \Theta$ and $u \in \mathcal{U}$ define

\[ C^N(\theta, u) = \{a \in \{0, 1\}^\mathbb{N} : a_k = \pi(R^k_{\theta}u) \text{ for } 0 \leq k \leq N\}, \]

and for $\Theta' \subseteq \Theta$ let $C^N(\Theta')$ be the union of $C^N(\theta, u)$ over $\theta \in \Theta'$ and $u \in \mathcal{U}$. The cylinder sets $C^N(\theta, u)$ are closed and open, and since $N$ is fixed, there are finitely many of them. As $\mathcal{O}$ is an open neighborhood of $\theta'$ it is clear that $X_{\theta'} \subset C^N(\Theta')$. Moreover, as $\nu_0$ is supported on the set of label sequences generated by $\theta_0$ and $C^N(\theta_0)$ is closed, $\text{supp}(\nu_0) \subset C^N(\Theta_0)$. The hypotheses of the proposition imply that that $C^N(\mathcal{O})$ and $C^N(\theta_0)$ are disjoint, and therefore $X_{\theta'}$ is disjoint from $\text{supp}(\nu_0)$.

\[ \square \]

**Proposition 5.6.** Under the hypotheses in Section 1.5.1, we have $\Theta_1 = \{\alpha_0\}$.

**Proof.** Recall that $\Theta = [0, \frac{1}{2}]$. We wish to apply Proposition 5.5. Let $\alpha_1 \neq \alpha_2$ and assume without loss of generality that $\alpha_1 < \alpha_2$. Fix $0 < \epsilon < (\alpha_2 - \alpha_1)/2$ and let $\mathcal{O} = [\alpha_2 - \epsilon, \alpha_2 + \epsilon] \cap \Theta$ and $N \geq 3/2(\alpha_2 - \alpha_1 - \epsilon)$. Let $u, v \in [0, 1)$ and $\alpha \in \mathcal{O}$. Define

\[ k = \inf \left\{ j \geq 0 : |(u + j \alpha) - (v + j \alpha_1)| \geq \frac{1}{2} \right\}. \]
Our choice of $N$ ensures that $N(\alpha - \alpha_1) \geq 3/2 \geq 1/2 + v - u$. Therefore

$$u + N\alpha - (v + N\alpha_1) \geq \frac{1}{2},$$

so that $k \leq N$. We claim that $\pi(R^k_\alpha u) \neq \pi(R^k_\alpha v)$. If $k = 0$ then $\frac{1}{2} \leq |u - v| < 1$, which implies that $\pi(u) \neq \pi(v)$. Suppose that $k \geq 1$. Using the definition of $k$ (twice) and the triangle inequality, we see that

$$\frac{1}{2} \leq |(u + k\alpha) - (v + k\alpha_1)|$$

$$\leq |(u + (k - 1)\alpha) - (v + (k - 1)\alpha_1)| + |\alpha - \alpha_1|$$

$$\leq \frac{1}{2} + \frac{1}{2} = 1,$$

and therefore $\pi(R^k_\alpha u) \neq \pi(R^k_\alpha v)$. Now Proposition 5.6 yields the result.

\[\Box\]

Proof of Proposition 1.10. We first show that $h(\mathcal{R}) = 0$. For $u$ in $[0, 1)$ and $\alpha$ in $\Theta$, let $\pi_n(u, \alpha)$ denote the element $w = w_0 \ldots w_{n-1}$ of $\{0, 1\}^n$ such that $R^k_\alpha u \in A_{w_k}$ for $k = 0, \ldots, n - 1$. Define

$$C(n) = \left| \left\{ w \in \{0, 1\}^n : \exists u \in [0, 1), \exists \alpha \in \Theta, w = \pi_n(u, \alpha) \right\} \right|,$$

and note that

$$h(\mathcal{R}) \leq \limsup_n \frac{1}{n} \log C(n).$$

It is known (see [8]) that $C(n) \leq Kn^4$ for some constant $K$, and it then follows from the previous display that $h(\mathcal{R}) = 0$. By Theorem 1.9, any estimates satisfying (1.10) converge almost surely to $\Theta_1$, and $\Theta_1 = \{\alpha_0\}$ by Proposition 5.6.

\[\Box\]

Acknowledgements

The authors thank Sayan Mukherjee and Karl Petersen for productive conversations. Both authors acknowledge the support of the National Science Foundation under grant DMS 1613261, and AN also acknowledges support under grant DMS 1310002.

References

[1] Terrence M Adams and Andrew B Nobel. Uniform convergence of Vapnik–Chervonenkis classes under ergodic sampling. The Annals of Probability, 38(4):1345–1367, 2010.

[2] Pierre Alquier and Olivier Wintenberger. Model selection for weakly dependent time series forecasting. Bernoulli, 18(3):883–913, 2012.
[3] Andrew F Bennett. Inverse modeling of the ocean and atmosphere. Cambridge University Press, 2005.
[4] Dimitri P Bertsekas and Steven E Shreve. Stochastic Optimal Control: The Discrete-Time Case. Athena Scientific, 1996.
[5] L. D. Brown and R. Purves. Measurable selections of extrema. Ann. Statist., 1(5):902–912, 09 1973.
[6] Thierry de la Rue. An introduction to joinings in ergodic theory. Discrete and Continuous Dynamical Systems, 15(1):121–142, 2006.
[7] Manfred Denker, Christian Grillenberger, and Karl Sigmund. Ergodic theory on compact spaces. Springer-Verlag, 1976.
[8] Frid, A. and Jamet, D. The number of binary rotation words. RAIRO-Theor. Inf. Appl., 48(4):453–465, 2014.
[9] Harry Furstenberg. Disjointness in ergodic theory, minimal sets, and a problem in diophantine approximation. Theory of Computing Systems, 1(1):1–49, 1967.
[10] Eli Glasner. Ergodic theory via joinings. American Mathematical Soc., 2003.
[11] Robert M Gray. Entropy and information theory. Springer Science & Business Media, 2011.
[12] Robert M Gray, David L Neuhoff, and Paul C Shields. A generalization of Ornstein’s d-bar distance with applications to information theory. The Annals of Probability, pages 315–328, 1975.
[13] Yonatan Gutman and Michael Hochman. On processes which cannot be distinguished by finite observation. Israel Journal of Mathematics, 164(1):265–284, 2008.
[14] H Hang and I Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. Ann. Statist., 2016.
[15] H Hang, I Steinwart, Y Feng, and J Suykens. Kernel density estimation for dynamical systems. arXiv:1607.03792, 2016.
[16] Oliver Jenkinson. Ergodic optimization. Discrete and Continuous Dynamical Systems, 15(1):197, 2006.
[17] Eugenia Kalnay. Atmospheric modeling, data assimilation and predictability. Cambridge university press, 2003.
[18] Paul C. Shields Katalin Marton. Entropy and the consistent estimation of joint distributions. The Annals of Probability, 22(2):960–977, 1994.
[19] JFC Kingman. The ergodic theory of subadditive stochastic processes. Journal of the Royal Statistical Society. Series B (Methodological), pages 499–510, 1968.
[20] JFC Kingman. Subadditive ergodic theory. The Annals of Probability, pages 883–899, 1973.
[21] JFC Kingman. Subadditive processes. In Ecole d’Eté de Probabilités de Saint-Flour V-1975, pages 167–223. Springer, 1976.
[22] Nicolas Kryloff and Nicolas Bogoliouboff. La théorie générale de la mesure dans son application à l’étude des systèmes dynamiques de la mécanique non linéaire. Annals of Mathematics, pages 65–113, 1937.
[23] Yury A Kutoyants. Statistical inference for ergodic diffusion processes. Springer Science & Business Media, 2013.
[24] Vitaly Kuznetsov and Mehryar Mohri. Time series prediction and online learning. In 29th Annual Conference on Learning Theory, pages 1190–1213, 2016.
[25] Steven P. Lalley. Beneath the noise, chaos. The Annals of Statistics, 27(2):461–479, 1999.
[26] Steven P. Lalley and A. B. Nobel. Denoising deterministic time series. Dyn. Partial Differ. Equ., 3(4):259–279, 2006.
[27] Kody Law, Andrew Stuart, and Kostas Zygalakis. Data Assimilation. Springer, 2015.
[28] Simon A Levin, Stephen R Carpenter, H Charles J Godfray, Ann P Kinzig, Michel Loreau, Jonathan B Losos, Brian Walker, and David S Wilcove. The Princeton guide to ecology. Princeton University Press, 2009.
[29] Kevin McGoff, Sayan Mukherjee, Andrew Nobel, and Natesh Pillai. Consistency of maximum likelihood estimation for some dynamical systems. The Annals of Statistics, 43(1):1–29, 2015.
[30] Kevin McGoff, Sayan Mukherjee, and Natesh Pillai. Statistical inference for dynamical systems: A review. Statist. Surv., 9:209–252, 2015.
[31] Kevin McGoff and Andrew Nobel. Empirical risk minimization and dynamical models, 2016. arXiv:1611.06173.
[32] Kevin McGoff and Andrew Nobel. Optimal tracking for dynamical systems, 2016. arXiv:1601.05033v1.
[33] Kevin A McGoff, Xin Guo, Anastasia Deckard, Christina M Kelliher, Adam R Leman, Lauren J Francey, John B Hogenesch, Steven B Haase, and John L Harer. The local edge machine: inference of dynamic models of gene regulation. Genome biology, 17(1):214, 2016.
[34] Gusztáv Morvai and Benjamin Weiss. Prediction for discrete time series. Probability theory and related fields, 132(1):1–12, 2005.
[35] Donald Ornstein. Bernoulli shifts with the same entropy are isomorphic. Advances in Mathematics, 4(3):337–352, 1970.
[36] Donald Ornstein and Benjamin Weiss. Entropy is the only finitely observable invariant. Journal of Modern Dynamics, 1(1):93–105, 2007.
[37] Donald S Ornstein. An application of ergodic theory to probability theory. The Annals of Probability, 1(1):43–58, 1973.
[38] Donald S Ornstein. Ergodic theory, randomness, and dynamical systems. Yale University Press, 1974.
[39] Donald S Ornstein and Benjamin Weiss. How sampling reveals a process. The Annals of Probability, pages 905–930, 1990.
[40] Robert R Phelps. Lectures on Choquet’s theorem, volume 1757. Springer Science & Business Media, 2001.
[41] Anthony Quas and Jason Siefken. Ergodic optimization of super-continuous functions on shift spaces. Ergodic Theory and Dynamical Systems, 32(06):2071–2082, 2012.
[42] Omri Sarig. Lecture Notes on Ergodic Theory. 2008.
[43] Ingo Steinwart and Marian Anghel. Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise. The Annals of Statistics, pages 841–875, 2009.
[44] Peter Turchin. Complex population dynamics: a theoretical/empirical synthesis, volume 35. Princeton University Press, 2013.
[45] Aad W Van der Vaart. Asymptotic statistics, volume 3. Cambridge University Press, 2000.
[46] Cédric Villani. Optimal transport: old and new, volume 338. Springer Science & Business Media, 2008.
Appendix A. Structure of the set of optimal joinings

In this section, we investigate the structure of the set of optimal joinings in Theorem 1.2 and provide the proof of Theorem 1.3. Our results rely on a general version of the ergodic decomposition for invariant probability measures. The following version, a restatement of [42, Theorem 2.5], is sufficient for our purposes.

**Theorem (The Ergodic Decomposition).** Suppose that $R : \mathcal{U} \to \mathcal{U}$ is a Borel measurable map of a Polish space $\mathcal{U}$ and that $\lambda \in \mathcal{M}(\mathcal{U}, R)$. Then there exists a Borel probability measure $\xi$ on $\mathcal{M}(\mathcal{U})$ such that

1. $\xi(\{\eta \text{ is invariant and ergodic for } R\}) = 1$
2. If $f \in L^1(\lambda)$, then $f \in L^1(\eta)$ for $\xi$-almost every $\eta$, and

\[ \int f \, d\lambda = \int \left( \int f \, d\eta \right) \, d\xi(\eta). \]

Whenever (2) holds, we write $\lambda = \int \eta \, d\xi$.

**Remark A.1.** Suppose $f : \mathcal{U} \to \mathcal{V}$ is a Borel measurable map between Polish spaces $\mathcal{U}$ and $\mathcal{V}$ and that $\lambda \in \mathcal{M}(\mathcal{U})$ satisfies $\lambda = \int \eta \, d\xi$. Then one may readily check that $m_f(\lambda) = \int m_f(\eta) \, d\xi$.

In the remainder of this section we assume that the systems $(\mathcal{X}, S)$ and $(\mathcal{Y}, T, \nu)$ and the loss $\ell$ satisfy the standard assumptions in Section 1. The following three lemmas will be used to prove Theorem 1.3.

**Lemma A.2.** If $\lambda \in \mathcal{J}(S : \nu)$ has ergodic decomposition $\lambda = \int \eta \, d\xi$, then $\xi$-almost every $\eta$ is in $\mathcal{J}(S : \nu)$.

**Proof.** Let $\lambda \in \mathcal{J}(S : \nu)$ have ergodic decomposition $\lambda = \int \eta \, d\xi$. Then $\xi$-almost every $\eta$ is in $\mathcal{M}(\mathcal{X} \times \mathcal{Y}, S \times T)$, and for these measures $m_X(\eta)$ is necessarily in $\mathcal{M}(\mathcal{X}, S)$. It follows from Remark A.1 that $\nu = m_Y(\lambda) = \int m_Y(\eta) \, d\xi$. Since $\nu$ is ergodic, it is an extreme point of the convex set $\mathcal{M}(\mathcal{Y}, T)$ (see [10, Proposition 12.4]), and therefore $m_Y(\eta) = \nu$ for $\xi$-almost every $\eta$. \qed

**Lemma A.3.** If $\lambda \in \mathcal{J}_{\text{min}}(S : \nu)$ has ergodic decomposition $\lambda = \int \eta \, d\xi$, then $\xi$-almost every $\eta$ is ergodic and contained in $\mathcal{J}_{\text{min}}(S : \nu)$.
Proof. Ergodicity of $\xi$-almost every $\eta$ follows from the definition of the ergodic decomposition. By assumption, $\ell \in L^1(\lambda)$, and the ergodic decomposition yields
\[ \int \ell \, d\lambda = \int \left( \int \ell \, d\eta \right) \, d\xi(\eta). \]

By Lemma A.2 $\xi$-almost every $\eta \in J(S : \nu)$ and in this case $\int \ell \, d\lambda \leq \int \ell \, d\eta$, as $\lambda \in J_{\min}(S : \nu)$. It follows that $\int \ell \, d\lambda = \int \ell \, d\eta$ for $\xi$-almost every $\eta$. □

Lemma A.4. The functional $\phi_\ell : J(S : \nu) \to \mathbb{R}$ defined by
\[ \phi_\ell(\lambda) = \int \ell \, d\lambda \]
is lower semi-continuous. If the loss $\ell$ is bounded from below, then the conclusion of this lemma follows immediately from the Portmanteau Theorem for weak convergence. In the general case, the result may be established by a truncation argument; as the argument is very similar to that in the proof of Theorem 1.2, we omit the details.

Proposition A.5. The set of measures $J(S : \nu)$ is compact in the weak topology.

Proof. We address the general case of measurable $T$. If $T$ is continuous, then the proof may be simplified. Let $\tilde{\nu} \in \mathcal{M}(Y^N)$ be the process measure generated by $(T, \nu)$ (see Section 2.3), and let $J(S : \tilde{\nu})$ be the associated set of joinings.

We claim that $J(S : \tilde{\nu})$ is compact in $\mathcal{M}(X \times Y^N)$. To see this, note that the direct product of $S$ with the left-shift $\tau$ on $Y^N$ is continuous, since each of these maps is continuous. It then follows that the induced push-forward map $m_{S \times \tau}$ from $\mathcal{M}(X \times Y^N)$ to itself is continuous in the weak topology. The set of fixed points of any continuous map is closed. Hence the set $C_1 = \mathcal{M}(X \times Y^N, S \times \tau)$ of fixed points of $m_{S \times \tau}$ is closed. Additionally, the continuity of the projection $m_{Y^N} : \mathcal{M}(X \times Y^N) \to \mathcal{M}(Y^N)$ ensures that $C_2 = m_{Y^N}^{-1}\{\tilde{\nu}\}$ is also closed, and therefore our set of interest $J(S : \tilde{\nu}) = C_1 \cap C_2$ is closed. Now let $\epsilon > 0$. Since $\{\tilde{\nu}\}$ is tight, there exists a compact set $K \subset Y^N$ such that $\tilde{\nu}(K) > 1 - \epsilon$. By assumption $X$ is compact, and therefore $X \times K$ is compact. Also, for any $\lambda \in J(S : \tilde{\nu})$, we have $\lambda(X \times K) = \tilde{\nu}(K) > 1 - \epsilon$. Since $\epsilon > 0$ was arbitrary, we see that $J(S : \tilde{\nu})$ is tight. Then, since it is both tight and closed, we may conclude that it is compact.

As the projection $(x, y) \mapsto (x, y_0)$ is continuous, the induced push-forward map $m_{X \times Y} : \mathcal{M}(X \times Y^N) \to \mathcal{M}(X \times Y)$ is continuous. By Lemma 2.1 $m_{X \times Y}(J(S : \tilde{\nu})) = J(S : \nu)$, and therefore $J(S : \nu)$
is compact, as it is the image of a compact set under a continuous map.

**Proof of Theorem 1.3.** Let $\phi_\ell: \mathcal{J}(S: \nu) \to \mathbb{R}$ be the functional defined by $\phi_\ell(\lambda) = \int t \, d\lambda$. The convexity of $\mathcal{J}_{\text{min}}(S: \nu)$ follows from the fact that $\phi_\ell(t\lambda_1 + (1-t)\lambda_2) = t\phi_\ell(\lambda_1) + (1-t)\phi_\ell(\lambda_2)$. As $\mathcal{J}(S: \nu)$ is compact in the weak topology (Proposition A.5) and $\phi_\ell$ is lower semicontinuous (Lemma A.4), the set $\mathcal{J}_{\text{min}}(S: \nu)$ where $\phi_\ell$ attains its minimum is compact.

It remains to identify the extreme points of $\mathcal{J}_{\text{min}}(S: \nu)$. Any ergodic measure in the (convex) set $\mathcal{M}(\mathcal{X} \times \mathcal{Y}, S \times T)$ is an extreme point of this set (see [40, Proposition 12.4]), and it follows that any ergodic measure $\lambda \in \mathcal{J}_{\text{min}}(S: \nu)$ is an extreme point of $\mathcal{J}_{\text{min}}(S: \nu)$. Suppose now that $\lambda$ is an extreme point of $\mathcal{J}_{\text{min}}(S: \nu)$, and let $\lambda = \int \eta \, d\xi$ be its ergodic decomposition. By Lemma A.3, $\xi$-almost every $\eta$ is in $\mathcal{J}_{\text{min}}(S: \nu)$, and as $\lambda$ is an extreme point of $\mathcal{J}_{\text{min}}(S: \nu)$, it follows that $\xi$-almost every $\eta$ equals $\lambda$. Then it follows from the ergodic decomposition that $\lambda$ is ergodic. \hfill \Box