Submitted to the Annals of Statistics
arXiv: arXiv:1611:06173

EMPIRICAL RISK MINIMIZATION AND COMPLEXITY OF DYNAMICAL MODELS

By Kevin McGoff[*,†] and Andrew B. Nobel[‡,§]

University of North Carolina at Charlotte[†]
University of North Carolina at Chapel Hill[§]

A dynamical model consists of a continuous self-map $T : \mathcal{X} \to \mathcal{X}$ of a compact state space $\mathcal{X}$ and a continuous observation function $f : \mathcal{X} \to \mathbb{R}$. This paper considers the fitting of a parametrized family of dynamical models to an observed real-valued stochastic process using empirical risk minimization. The limiting behavior of the minimum risk parameters is studied in a general setting. We establish a general convergence theorem for minimum risk estimators and ergodic observations. We then study conditions under which empirical risk minimization can effectively separate signal from noise in an additive observational noise model. The key condition in the latter results is that the family of dynamical models has limited complexity, which is quantified through a notion of entropy for families of infinite sequences that connects covering number based entropies with topological entropy studied in dynamical systems. We establish close connections between entropy and limiting average mean widths for stationary processes, and discuss several examples of dynamical models.

1. Introduction. Empirical risk minimization is a common approach to model fitting and estimation in a variety of parametric and non-parametric problems. In this paper we investigate the use of empirical risk minimization to fit a family of dynamical models to an observed stochastic process. Formally, a dynamical model consists of a continuous transformation $T : \mathcal{X} \to \mathcal{X}$ on a compact metric space $\mathcal{X}$, and a continuous observation function $f : \mathcal{X} \to \mathbb{R}$. Let T^k denote the k-fold composition of T with itself, and let T^0 be the identity map on $\mathcal{X}$. From each initial state $x \in \mathcal{X}$ the dynamical model (T, f) yields a real-valued sequence $(f(T^k x))_{k \geq 0}$ obtained by applying the observation function f to the deterministic sequence of states (trajectory) generated by repeated application of the transformation T to

[*]KM acknowledges the support of NSF grant DMS-1613261.
[‡]ABN acknowledges the support of NSF grants DMS-1613072, DMS-1310002, and DMS-1613261.
MSC 2010 subject classifications: Primary 62M09
Keywords and phrases: emprirical risk minimization, dynamical models, joinings, topological entropy

the initial state x. In general, f need not be injective, so one cannot necessarily recover the underlying state sequence from the values of $(f(T^k x))_{k\geq 0}$.

In what follows we consider an indexed family $\mathcal{D} = \{(T_\theta, f_\theta) : \theta \in \Theta\}$ of dynamical models defined on a common compact metric space $\mathcal{X}$, and satisfying the following conditions:

(D1) the index set Θ is a compact metric space;

(D2) the map $(\theta, x) \mapsto T_\theta(x)$ from $\Theta \times \mathcal{X}$ to $\mathcal{X}$ is continuous;

(D3) the map $(\theta, x) \mapsto f_\theta(x)$ from $\Theta \times \mathcal{X}$ to $\mathbb{R}$ is continuous.

Condition (D2) ensures that each transformation T_θ is continuous and that the action of T_θ is continuous in θ. Condition (D3) ensures that each observation function f_θ is continuous and that observations vary continuously with θ. In particular, there exists a constant $K_\mathcal{D} > 0$ such that $|f_\theta(x)| \leq K_\mathcal{D}$ for every $x \in \mathcal{X}$ and $\theta \in \Theta$. Examples of families of systems satisfying these conditions are given in Section 3.

By definition, dynamical models are deterministic: the real-valued sequence $\big(f(T^k x)\big)_{k\geq 0}$ generated by a model (T, f) is fully determined once the initial condition $x \in \mathcal{X}$ is given. In this paper our primary interest is in dynamical models that represent low complexity regularities of potential interest, such as periodicity, multi-periodicity, constrained growth behavior, and hierarchical structure. Fitting a family of dynamical models to an observed stochastic process is a means of identifying and quantifying the corresponding low complexity regularities in the observed process. Examples of model families and references to existing applications are given in Section 3 below.

Due to the nature of the underlying dynamics and the possible presence of dynamic or observational noise, the observed process is likely to be complex, and one cannot expect a low-complexity model to capture all features of the observed process. Accordingly, our results do not assume that the observed process is generated from a model in the family $\mathcal{D}$ under study. The complexity of model families is quantified through a combinatorial notion of entropy with connections to empirical process theory and ergodic theory that is defined in Section 2 below.

1.1. *Minimum Risk Fitting of Dynamical Models.* Let $\mathcal{D}$ be a family of dynamical models that capture some behavior of interest, and let $\mathbf{Y} = Y_0, Y_1, \ldots \in \mathbb{R}$ be an observed stationary ergodic process. Suppose that we wish to identify regularities in $\mathbf{Y}$ by fitting the observed values of the process with models in $\mathcal{D}$. We do not assume that the observed process $\mathbf{Y}$ is generated by a process in $\mathcal{D}$. Let $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be a nonnegative loss

function that is jointly lower semicontinuous in its arguments. We require the following integrability condition:

$$\text{(C1)} \qquad \mathbb{E}\left[\sup_{|u|\le K_{\mathcal{D}}} \ell\big(u, Y_0\big)\right] \;<\; \infty,$$

where $K_{\mathcal{D}}$ is an upper bound on $\{|f_\theta(x)| : x \in \mathcal{X}, \theta \in \Theta\}$. If the supremum in (C1) is not measurable, then one may replace the expectation by an outer expectation. For each $n \ge 0$, $\theta \in \Theta$, and $x \in \mathcal{X}$ define

$$R_n(\theta : x) \;=\; \frac{1}{n}\sum_{k=0}^{n-1} \ell\Big(f_\theta \circ T_\theta^k(x), Y_k\Big),$$

which is the empirical risk of the model (T_θ, f_θ) with initial state x relative to the first n observations of $\mathbf{Y}$. We formalize empirical risk minimization as follows.

Definition 1.1. A sequence of measurable functions $\theta_n : \mathbb{R}^n \to \Theta$, $n \ge 1$, will be called (empirical) minimum risk estimates for $\mathcal{D}$ if there exists a corresponding sequence of measurable functions $x_n : \mathbb{R}^n \to \Theta$, $n \ge 1$, such that

$$\text{(1.1)} \qquad \lim_n R_n\big(\hat\theta_n, \hat x_n\big) = \lim_n \inf_{\theta,x} R_n\big(\theta, x\big) \quad \text{w.p.1},$$

where $\hat\theta_n = \theta_n(Y_0, \dots, Y_{n-1})$ and $\hat x_n = x_n(Y_0, \dots, Y_{n-1})$ depend only on the first n observations. We note that the existence of the limit on the right hand side of (1.1) follows from Kingman's subadditive ergodic theorem (under (C1)), whereas the existence of the limit on the left hand side of (1.1) is part of the definition.

Remark 1.2. The notion of minimum risk estimates formalizes empirical risk minimization when fitting dynamical models. The key difference between the definition above and minimum risk estimates in standard, non-dynamic settings, is the presence of the initial state $x \in \mathcal{X}$ in (1.1). Given observations $Y_0, \dots, Y_{n-1}$, one selects a parameter $\hat\theta_n \in \Theta$ *and* an initial state $\hat x_n \in \mathcal{X}$ so that $(\hat\theta_n, \hat x_n)$ is an approximate minimizer of $R_n(\theta : x)$. Our assumptions on ℓ, $\mathcal{X}$, and Θ ensure that an exact minimizer (θ_n^*, x_n^*) of $R_n(\theta : x)$ exists, and further that the pair (θ_n^*, x_n^*) may be chosen to depend measurably on the observations $Y_0, \dots, Y_{n-1}$ [7, Proposition 7.33, p.153].

The definition of minimum risk estimates does not require that $(\hat\theta_n, \hat x_n)$ be an exact minimizer of $R_n(\theta : x)$ for each n; it requires only that the

average loss is minimized asymptotically. Our results apply to any sequence $(\hat{\theta}_n, \hat{x}_n)$ of approximate minimizers. This generalization is important, as exact minimization of $R_n(\theta : x)$ will typically be difficult in practice. Similar computational difficulties arise even in simple non-dynamical settings, e.g., in the problem of classification where finding the best (minimum error) hyperplane separating a set of labeled vectors in Euclidean space is known to be NP hard.

In practice, it is natural to obtain minimum risk estimates $(\hat{\theta}_n, \hat{x}_n)$ by minimizing over finite ϵ_n-coverings of the joint parameter-state space $\Theta \times \mathcal{X}$, where the radius ϵ_n tends to zero with increasing n. The following proposition shows that this approach is possible (in principle) under mild continuity conditions on the loss ℓ. However, the size of the covering sets may grow rapidly (e.g., exponentially) as a function of n, and therefore this approach may still be limited by computational power.

In more detail, for each $n \geq 1$ define a pseudo-metric on the compact space $\Theta \times \mathcal{X}$ by

$$\rho_n\Big((\theta, x), (\theta', x')\Big) \;=\; \max_{0 \leq k < n} \Big| f_\theta \circ T_\theta^k(x) - f_{\theta'} \circ T_{\theta'}^k(x') \Big|.$$

Let $(\epsilon_n)_{n \geq 1}$ be a sequence of positive real numbers tending to zero, and for each n, let $\mathcal{C}_n \subset \Theta \times \mathcal{X}$ be a finite ϵ_n-cover of $\Theta \times \mathcal{X}$ with respect to ρ_n. Compactness of $\Theta \times \mathcal{X}$ ensures that such a finite cover exists. A proof of the following result appears in the Supplementary Material [47, Appendix B)].

PROPOSITION 1.3. *If the loss function ℓ is uniformly continuous and for each $n \geq 1$,*

$$(\hat{\theta}_n, \hat{x}_n) \in \operatorname*{argmin}_{(\theta, x) \in \mathcal{C}_n} R_n\big(\theta : x\big),$$

then $(\hat{\theta}_n)_{n \geq 1}$ is a sequence of minimum risk estimates.

REMARK 1.4. Minimum risk estimation involves the identification of an optimal, or near optimal, parameter-state pair $(\hat{\theta}_n, \hat{x}_n)$. In what follows we focus on the parameter estimates $\hat{\theta}_n$ rather than the initial states $\hat{x}_n$. As the results here show, under suitable conditions the parameter estimates exhibit regular limiting behavior. The same cannot be said about the initial states. For example, the negative results of [31, 32] give general conditions under which estimation of the initial state of a dynamical system is not possible.

The principal goal of this paper is to understand and characterize the limiting behavior of minimum risk estimates $\hat{\theta}_n$. Our analysis hinges on the observation that every dynamical model, and every family of such models, is

associated with a family of stationary processes. We focus on the misspecified case in which the observed process is not necessarily generated by a model in the family $\mathcal{D}$ under study. Two main results are presented. The first main result (Theorem 5.2) provides a variational characterization of the limiting behavior of minimum risk estimates. In particular, we establish that minimum risk estimates converge to a parameter set determined by the projection of the observed process onto the family of processes associated with the models in $\mathcal{D}$, where the projection is with respect to a divergence measure that depends on the loss. Identifiability of parameters is addressed automatically, through consideration of their associated processes.

The second main result (Theorem 5.3) provides conditions under which minimum risk estimation can effectively separate signal from noise in a simple signal plus noise setting when the model family $\mathcal{D}$ has limited complexity. Complementing the second, positive result, we establish a negative result (Proposition 5.9) showing that minimum risk estimation can be strongly inconsistent for complex families of dynamical models.

In establishing the results mentioned above, we study a notion of entropy that measures the complexity of a family of dynamical models. The entropy measure is based on the growth rate of ℓ_p-covering numbers of finite length sequences, and we show that it is closely related to the notion of topological entropy studied in dynamical systems. Furthermore, we show that the entropy measure is independent of the exponent p, and we establish a qualitative connection between entropy and stochastic mean widths studied in empirical process theory.

Both the statements and proofs of our results rely on the concept of joinings, which are stationary couplings of stochastic processes. Joinings, introduced by Furstenberg [17], have been well-studied in ergodic theory but have not been widely applied to problems of statistical inference. Our results show that joinings are intimately connected with minimum risk fitting of dynamical models. Several tools from the theory of joinings, including disjointness and relatively independent joinings, play an important role in our analysis of complexity and separation of signal and noise.

2. Complexity of Dynamical Models. Quantifying the complexity of a family of models is a key issue in nonparametric inference. Indeed, model complexity is a key factor in establishing consistency, convergence rates, and optimality conditions for a variety of common inference procedures. Although fitting nonlinear dynamical models differs from model fitting for classification or regression, complexity still plays a central role in the analysis of minimum risk estimation. In particular, as we demonstrate in Section 5,

model complexity has a close connection with the ability of minimum risk estimators to separate signal from noise.

We begin this section by reviewing and discussing the notion of topological entropy for a topological dynamical system. Subsequently, we define and discuss two related notions of complexity for families of dynamical models that are used in our principal results. The first is a combinatorial entropy measure that captures the exponential growth rate of the real-valued sequences generated by the model family, and is closely related to topological entropy. The second is a limiting mean width, which arises when using the squared loss. In establishing consistency or rates of convergence for classification or regression methods, it is common, and typically necessary, to constrain the family of models under study by requiring the sub-exponential growth of its complexity, e.g., by assuming that the VC dimension of the model class is finite, or by imposing conditions on its covering/bracketing numbers [72]. The complexity conditions used in this paper serve as analogous constraints for families of dynamical models.

2.1. *Topological entropy of a topological dynamical system.* Before introducing our notion of entropy for a family of dynamical models, we discuss the notion of topological entropy for dynamical systems, originally introduced in [2]. Topological entropy has served as the central notion of complexity for dynamical systems for at least fifty years. The theory of topological entropy is developed in detail in many books, including [26] and [74]. For a thorough historical account, see [25].

DEFINITION 2.1. Let $(\mathcal{X}, d)$ be a compact metric space, and let $T : \mathcal{X} \to \mathcal{X}$ be continuous. For $n \geq 1$, let d_n be the metric on $\mathcal{X}$ given by

$$d_n(x, y) = \max\Big\{ d\Big(T^k x,\, T^k y\Big) : 0 \leq k \leq n-1 \Big\}.$$

For $\epsilon > 0$, let $B(x, n, \epsilon)$ denote the ball of radius ϵ around the point x with respect to the metric d_n. Then let $\mathrm{C}(n, \epsilon)$ denote the ϵ-covering number of $\mathcal{X}$ with respect to the metric d_n, which is the least natural number M such that there exist points $x_1, \ldots, x_M \in \mathcal{X}$ for which

$$\mathcal{X} \subset \bigcup_{i=1}^{M} B(x_i, n, \epsilon).$$

Finally, the topological entropy of the system $(\mathcal{X}, T)$ can be defined as

$$h_{top}(\mathcal{X}, T) = \lim_{\epsilon \searrow 0} \limsup_{n \to \infty} \frac{1}{n} \log \mathrm{C}(n, \epsilon).$$

The topological entropy serves as a quantitative measure of the exponential growth rate of the number of orbits within the system. A positive value of entropy is typically taken as an indicator of "chaos". In smooth systems, positive entropy is closely related to the existence of positive Lyapunov exponents (see [6]). Examples of systems with positive entropy include Axiom A attractors [9] and the classical Lorenz system [49]. There are also many interesting examples of systems with zero entropy, which have received substantial recent attention in the dynamics literature, including toral rotations (see [26]), interval exchange transformations (see [26]), and rational billiards (see [41]). In Section 3 we discuss some additional examples.

2.2. *Entropy of a family of dynamical models.* Let us now define the entropy of a family of dynamical models $\mathcal{D}$, which is assessed through the covering numbers of the real-valued sequences generated by its constituent models. Let $\mathbf{u} = (u_k)_{k\geq 0}$ and $\mathbf{v} = (v_k)_{k\geq 0}$ denote infinite sequences in $\mathbb{R}^{\mathbb{N}}$. For each $n \geq 1$ and $1 \leq p \leq \infty$ define pseudo-metrics $d_{n,p}(\cdot,\cdot)$ on $\mathbb{R}^{\mathbb{N}}$ as follows:

$$d_{n,p}(\mathbf{u},\mathbf{v}) = \begin{cases} \left(n^{-1}\sum_{k=0}^{n-1}|u_k - v_k|^p\right)^{1/p} & \text{if } 1 \leq p < \infty \\ \max_{0\leq k\leq n-1}|u_k - v_k| & \text{if } p = \infty. \end{cases}$$

Let $\mathcal{U} \subseteq \mathbb{R}^{\mathbb{N}}$ be a bounded family of infinite sequences, meaning that there is a $K > 0$ such that $\mathcal{U} \subset [-K, K]^{\mathbb{N}}$. For a fixed length n and radius $r > 0$ we can assess the effective number of initial n-sequences in $\mathcal{U}$ at radius r by the covering number $N(\mathcal{U}, r, d_{n,p})$, which is the minimal number of balls of radius r under the pseudo-metric $d_{n,p}(\cdot,\cdot)$ that are required to cover the set $\mathcal{U}$. In what follows we will be interested in the exponential growth rate of these covering numbers with increasing n, which is captured by the quantity

$$h_p(\mathcal{U}, r) \;=\; \limsup_{n\to\infty} \frac{1}{n} \log N(\mathcal{U}, r, d_{n,p}).$$

The ℓ_p entropy of the family $\mathcal{U}$ is defined to be the supremum of these growth rates over r, obtained by letting r tend to zero,

$$h_p(\mathcal{U}) \;=\; \lim_{r\searrow 0} h_p(\mathcal{U}, r).$$

The definition of $h_p(\mathcal{U})$ raises questions about an appropriate choice of p. From a statistical point of view, it is natural (and common) to consider empirical ℓ_2 covering numbers when assessing complexity. On the other hand, from a dynamical systems point of view, it is common to consider empirical

ℓ_∞ covering numbers, as is done with topological entropy. In fact, as the next proposition shows, all the ℓ_p entropies coincide. As we were unable to find this fact in the literature, a proof is given in the Supplementary Material [47, Appendix C].

PROPOSITION 2.2. *The ℓ_p entropies $h_p(\mathcal{U})$ for $1 \leq p \leq \infty$ are all equal.*

REMARK 2.3. Although it is not needed here, we note that Proposition 2.2 holds more generally for sets of sequences $\mathcal{U} \subseteq A^{\mathbb{N}}$, where (A, ρ) is any compact metric space such that

$$\lim_{r \searrow 0} r \log N(A, r, \rho) = 0,$$

and the pseudo metrics $d_{n,p}(\cdot, \cdot)$ are defined in terms of ρ.

DEFINITION 2.4 (Entropy of a model family). The *entropy* $h(\mathcal{D})$ of a family $\mathcal{D}$ of dynamical models is the common value of $h_p(\mathcal{U}_\mathcal{D})$, where

$$\mathcal{U}_\mathcal{D} = \left\{ \left(f_\theta \circ T_\theta^k(x) \right)_{k \geq 0} : x \in \mathcal{X},\, \theta \in \Theta \right\} \subseteq \mathbb{R}^{\mathbb{N}} \tag{2.1}$$

is the set of infinite sequences generated by models in $\mathcal{D}$.

In Lemma 4.2 we establish an equivalent expression for $h(\mathcal{D})$ in terms of the entropy rates of the processes associated with the model family $\mathcal{D}$.

REMARK 2.5. The entropy of a family of dynamical models has close connections with the notion of topological entropy. Indeed, under our hypotheses (conditions (D1)-(D3)), it is straightforward to show that the family of sequences $\mathcal{U}_\mathcal{D}$ is a compact subset of $\mathbb{R}^{\mathbb{N}}$ in the product topology. Moreover, if $\tau : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ is the left-shift map defined by $\tau(\mathbf{u})_k = u_{k+1}$ for $k \geq 0$, then it is easy to see that τ is continuous and that $\tau(\mathcal{U}_\mathcal{D}) \subseteq \mathcal{U}_\mathcal{D}$. Thus $(\mathcal{U}_\mathcal{D}, \tau)$ is a topological dynamical system that captures the dynamics of the *family* of dynamical models $\mathcal{D}$, and one may show that the entropy $h(\mathcal{D})$ defined above is equal to the topological entropy $h_{top}(\mathcal{U}_\mathcal{D}, \tau)$.

The following proposition provides a connection between the entropy of a family $\mathcal{D} = \{(T_\theta, f_\theta) : \theta \in \Theta\}$ and the topological entropies of the systems $(\mathcal{X}, T_\theta)$ in the family. A proof of this result appears in the Supplementary Material [47, Appendix E.3].

PROPOSITION 2.6. *Let $\mathcal{D}$ be a family of dynamical models satisfying (D1)-(D3), and suppose that for each $\theta \in \Theta$, the topological entropy of $(\mathcal{X}, T_\theta)$ is zero. Then $h(\mathcal{D}) = 0$.*

2.3. *Mean width.* Our analysis of minimum risk estimation under the squared loss $\ell(x,y) = (x-y)^2$ leads naturally to another measure of complexity for the family $\mathcal{D}$ that is based on the standard notion of mean width.

DEFINITION 2.7. Let $\mathcal{D}$ be a family of dynamical models, and let $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k\geq 0}$ be an i.i.d. process with mean zero and finite variance. The n-sample mean width of $\mathcal{D}$ relative to $\boldsymbol{\varepsilon}$ is

$$\kappa_n(\mathcal{D} : \boldsymbol{\varepsilon}) \;=\; \mathbb{E}\left[\sup_{x,\theta} \sum_{k=0}^{n-1} f_\theta \circ T_\theta^k(x) \cdot \varepsilon_k\right]. \tag{2.2}$$

Define the mean width of $\mathcal{D}$ relative to $\boldsymbol{\varepsilon}$ to be the limiting linear growth rate of the finite sample mean widths,

$$\kappa(\mathcal{D} : \boldsymbol{\varepsilon}) = \lim_n \frac{1}{n} \kappa_n(\mathcal{D} : \boldsymbol{\varepsilon}), \tag{2.3}$$

which exists by subadditivity (see Remark C.2). When $\varepsilon_i \sim N(0,1)$ we denote $\kappa(\mathcal{D} : \boldsymbol{\varepsilon})$ by $\kappa_G(\mathcal{D})$ and refer to this quantity as the Gaussian mean width of the family $\mathcal{D}$.

Finite sample mean widths have been widely studied in machine learning and empirical process theory, with an emphasis on Rademacher and Gaussian noise processes [8, 33]. As the next result shows, the mean width of $\mathcal{D}$ has connections with the entropy of $\mathcal{D}$. A proof of this result appears in the Supplementary Material [47, Appendix C].

THEOREM 2.8. *Let $\boldsymbol{\varepsilon} = (\epsilon_k)_{k\geq 0}$ be an i.i.d. sequence with mean zero and finite variance. If $h(\mathcal{D}) = 0$ then $\kappa(\mathcal{D} : \boldsymbol{\varepsilon}) = 0$. Moreover, the Gaussian mean width $\kappa_G(\mathcal{D}) = 0$ if and only if $h(\mathcal{D}) = 0$.*

REMARK 2.9. Theorem 2.8 establishes a qualitative relationship between asymptotic mean width and entropy: for a given family of dynamical models, they are either both zero or both positive. In general one cannot expect a more quantitative relationship between asymptotic mean width and entropy. While it is possible to provide upper and lower bounds on $\kappa_n(\mathcal{D} : \boldsymbol{\varepsilon})$ in terms of ℓ_2 covering numbers (as in the proof of Theorem 2.8), additional care must be taken when passing to the limits to obtain the mean width and the entropy. As it turns out, the presence of these limits in the definitions precludes any more quantitative dependence between these quantities.

3. Examples of Dynamical Models. In this section we discuss several families of dynamical models. These families capture different regularities that may be of interest when studying an observed dynamical system. These and related families have been fit to data by applied scientists (e.g. [10, 34, 43, 70]), without any theoretical guarantees of consistency. Each family described below satisfies assumptions (D1)-(D3) and, under suitable assumptions, has entropy zero. Thus the results of the next section apply to minimum risk estimates based on these families.

EXAMPLE 3.1. (Gene Regulatory Networks) Inference of gene regulatory networks from observed data is considered an important problem in systems biology [40]. In recent years, it has become increasingly feasible for experimentalists to assay the abundance of all the genes in a given system with regular frequency over time. In such cases, one would like to infer the structure of the underlying network from the observed gene expression dynamics. Here we present a family of dynamical models studied in [43].

Suppose one would like to investigate the gene regulatory network for genes $g_1, \ldots, g_N$. Let $x_i(t)$ denote the abundance of mRNA associated with gene g_i at time t. We are interested in modeling the feedback effects of genes on other genes. When the presence of proteins associated with gene g_i positively affects the rate of production of mRNA for g_j, we say that g_i *promotes* g_j. When the presence of proteins associated with g_i negatively affects the rate of production of gene g_j, we say that g_i *inhibits* g_j. If either of these relationships holds, then we say that g_i *controls* g_j. In order to constrain the complexity of the model class, we assume that for each gene g_j, there is at most one gene g_i that controls g_j.

To make the model precise, we parametrize it as follows. If g_i activates g_j, we assume that the functions x_i, x_j satisfy a differential equation of the following form:

$$\frac{dx_j}{dt} = A_\alpha(x_i(t)) - \gamma x_j,$$

where $\gamma > 0$ is a degradation rate and $A_\alpha(x)$ is a parametrized nonlinear "activation" function with parameter α. Similarly, if g_i inhibits g_j, then we assume that the functions x_i, x_j satisfy a differential equation of the following form:

$$\frac{dx_j}{dt} = I_\alpha(x_i(t)) - \gamma x_j,$$

where $\gamma > 0$ is a degradation rate and $I_\alpha(x)$ is a parameterized nonlinear "inhibition" function with parameter α. We also assume that all parameters (α, γ) are constrained to lie in a compact set $\mathcal{K} \subset \mathbb{R}^p$. Note that the entire system of differential equations governing $(x_i(t))_{i=1}^N$ can be specified by a

compact parameter space Θ, consisting of the discrete variables indicating the type of control (activation or inhibition) of each gene, along with all the associated continuous parameters (α, γ). Let $(x_i^\theta(t))_{i=1}^N$ denote the solution of the system of equations with parameter θ at time t.

Observations of the system are assumed to have the following structure: there is a time step $\Delta > 0$ such that at times $t_k = k\Delta$, for $k = 0, \ldots, T$, the abundance of mRNA associated with each gene g_i is observed to be $y_{i,k}$. The parameters in the system of differential equations are then fit to such observations by attempting to minimize the sum of squared differences between the observations and the integrated solutions of the equations at the associated time points.

We can place this model class in the general framework of dynamical models as follows. Under appropriate conditions on the activation and inhibition functions (boundedness, smoothness, and monotonicity in x), there will exist a compact set $\mathcal{X} \subset \mathbb{R}^N$ such that for all $\theta \in \Theta$, if $(x_1^\theta(0), \ldots, x_N^\theta(0)) \in \mathcal{X}$, then $(x_1^\theta(t), \ldots, x_N^\theta(t)) \in \mathcal{X}$ for all time. Let $T_\theta : \mathcal{X} \to \mathcal{X}$ be the time-Δ map, defined as follows: given a point $\mathbf{x} \in \mathcal{X}$, let $T_\theta(\mathbf{x})$ be the solution $(x_1^\theta(\Delta), \ldots, x_N^\theta(\Delta))$ at time Δ of the system of equations with parameter θ and initial condition $(x_1^\theta(0), \ldots, x_N^\theta(0)) = \mathbf{x}$.

We remark that the sparsity constraint on the model class (that each gene be controlled by at most one other gene) ensures that any corresponding family of dynamical models will have zero entropy [13]. A full proof of this fact is beyond the scope of the present paper, but an outline can be stated as follows. First, any directed network with maximal in-degree equal to one can be decomposed into a collection of disjoint directed cycles, along with disjoint paths out from the cycles. Then the Poincaré-Bendixon theorem for monotone cyclic feedback systems [39] implies that the dynamics of any such system must be severely constrained, and in particular, it must have zero topological entropy. Finally, by Proposition 2.6, the entire family of dynamical models must have zero entropy.

EXAMPLE 3.2. (Subcritical logistic family and ecology) Since at least the early work of May [42], simple parametric families of dynamical systems have been used by ecologists as models of the population dynamics of many species [36]. In many instances, various types of deterministic models have been fit to ecological data (e.g., [70]).

The prototypical family in this context is the logistic family, which may be parametrized as follows. Consider the state space $\mathcal{X} = [0, 1]$ and the family of maps $T_a : [0, 1] \to [0, 1]$, where $T_a(x) = ax(1 - x)$ for $a \in [0, 4]$. If we restrict a to the region $[0, a_{max}]$, where $a_{max} = 1 + \sqrt{5}$, then each system in the

family has zero topological entropy [24], and therefore any associated family of dynamical models will have zero entropy by Proposition 2.6. This situation is thought to occur in many naturally occurring populations (see results and discussion from [23]). In examples such as these, the state variable x typically represents the (rescaled) population size. The overall structure of the logistic family captures the idea that the reproductive rate depends on the density of the population, taking into account effects such as competition for limited resources. Given observations of population size over time, researchers are interested in fitting logistic dynamical models to the observations in order to identify the parameter a. For examples that involve fitting this family or similar families, see [23, 70].

EXAMPLE 3.3. (Symbolic dynamics and quasicrystals) Symbolic dynamical systems, also known as subshifts, are a useful family of models that arise in the study of dynamical systems through discretization of the state space. Informally, if $T : X \to X$ is a dynamical system and $\{A_1, \ldots, A_N\}$ is a finite partition of X, then the associated symbolic system consists of the label sequences $\{(\pi(T^k x))_{k\geq 0} : x \in X\}$ under the left shift map, where $\pi : X \to \{1, \ldots, N\}$ is defined by the relation $x \in A_{\pi(x)}$. Symbolic systems have been widely studied for their own sake [37], for the purpose of understanding other dynamical systems [9], and for their connections to other disciplines, e.g., physics [64]. Due to their combinatorial nature, they can be used to model a variety of regularities in physical systems. For example, they have been used in communications, coding and information theory to capture the rules by which binary strings should be encoded on magnetic tapes and compact discs in order to minimize errors [37].

As another example, symbolic dynamical systems have recently been used by several researchers [12, 63, 67] as a mathematical model of crystallographic structures known as quasicrystals, which were discovered by Shechtman [66]. Quasicrystals are characterized by the presence of long-range aperiodic order, in contrast to crystals, which are characterized by long-range periodic order. Substitution systems [60] are special cases of symbolic dynamical systems that are constructed by enforcing a rigid hierarchical structure at all scales, and they have been shown to possess long-range aperiodic order similar to that observed in quasicrystals. As such, substitution systems have been studied as theoretical models for quasicrystals.

Let us now present detailed definitions of symbolic dynamical systems and substitution systems. Let $\mathcal{A}$ be a finite set, known as the alphabet. The set $\mathcal{A}^{\mathbb{Z}}$ is known as the full-shift on $\mathcal{A}$. We endow $\mathcal{A}$ with the discrete topology and $\mathcal{A}^{\mathbb{Z}}$ with the product topology, making it a compact completely

metrizable space. We let $\tau : \mathcal{A}^{\mathbb{Z}} \to \mathcal{A}^{\mathbb{Z}}$ be the left-shift: if $\mathbf{a} = (a_k)_{k\in\mathbb{Z}}$, then $\tau(\mathbf{a})_k = a_{k+1}$. Note that τ is continuous. A subshift on alphabet $\mathcal{A}$ is a subset $\mathcal{X} \subset \mathcal{A}^{\mathbb{Z}}$ such that $\mathcal{X}$ is closed and invariant, i.e., $\tau(\mathcal{X}) = \mathcal{X}$. Note that if $\mathcal{X}$ is a subshift and $\mathcal{F} \subset C(\mathcal{X})$ is any compact set of continuous functions from $\mathcal{X}$ to $\mathbb{R}$ with respect to the uniform metric on $C(\mathcal{X})$, then $\mathcal{D} = \{(\tau, f) : f \in \mathcal{F}\}$ is a continuous family of dynamical models satisfying (D1)-(D3).

Substitution systems provide interesting examples of subshifts with entropy zero. Let $m \geq 1$, and let $s : \mathcal{A} \to \mathcal{A}^m$. The map s is called the substitution map. We extend s to words of any length ℓ by concatenation, $s(a_1 \dots a_n) = s(a_\ell) \dots s(a_\ell)$. In this way, we may refer to iterates $s^k : \mathcal{A} \to \mathcal{A}^{km}$, defined by induction $s^{k+1}(a) = s(s^k(a))$. To a substitution map s, one may associate a subshift $\mathcal{X}$ as follows: a sequence $\mathbf{a} = (a_k)_{k\in\mathbb{Z}} \in \mathcal{A}^{\mathbb{Z}}$ is in $\mathcal{X}$ if for each $i < j$ in $\mathbb{Z}$, there exists a symbol $b \in \mathcal{A}$ and a power $k \geq 1$ such that $a_i \dots a_j$ appears as a subword of $s^k(b)$. Under some mild conditions on the substitution map s (see [60]), any sequence $\mathbf{a} = (a_k)_{k\in\mathbb{Z}}$ in $\mathcal{X}$ can be uniquely decomposed as $\dots s(b_{-2})s(b_{-1})s(b_0)s(b_1)s(b_2)\dots$ for some sequence $\mathbf{b} = (b_k)_{k\in\mathbb{Z}}$ in $\mathcal{X}$. Here the sequence $\mathbf{b}$ is interpreted as giving the structure of $\mathbf{a}$ at a larger scale (blocks of length m). As this decomposition may be repeated with $\mathbf{b}$ in the role of $\mathbf{a}$ and continued in this way, one may interpret substitution systems as having a rigid hierarchical structure. This rigid hierarchical structure leads to low complexity: if $\mathcal{X}$ is a substitution system, then it can be shown that $\mathcal{X}$ has zero topological entropy (see [11]). Consequently, any continuous family $\mathcal{D}$ of dynamical models on $\mathcal{X}$ will have $h(\mathcal{D}) = 0$ by Proposition 2.6. Such models could then be fit to a observations of quasicrystals in an effort to identify particular hierarchical structure. Although this type of fitting has not yet been used in statistical studies of quasicrystals, our results provide some theoretical grounding for potential work in that direction.

Example 3.4. (Toral rotations and almost periodicity) Let the state space $\mathcal{X}$ be the d-dimensional torus $\mathbb{T}^d$, which is the direct product of d circles, $\mathbb{T}^d = S^1 \times \dots \times S^1$. For a vector $\alpha \in \mathbb{T}^d$, define the transformation $R_\alpha : \mathbb{T}^d \to \mathbb{T}^d$ to be the rotation of $\mathbb{T}^d$ by the angle vector α, i.e. $R_\alpha(x) = x + \alpha$ (addition in $\mathbb{T}^d$). Then let $\mathcal{F} \subset C(\mathbb{T}^d)$ be a compact set of continuous functions from $\mathbb{T}^d$ to $\mathbb{R}$ (with respect to the topology induced by the supremum norm). Let $\Theta = \mathbb{T}^d \times \mathcal{F}$, and define the family of dynamical models $\mathcal{D} = \{(R_\alpha, f) : (\alpha, f) \in \Theta\}$.

With these definitions, $\mathcal{D}$ is a continuous family of dynamical models satisfying (D1)-(D3). Furthermore, it is well-known that any toral rotation

has zero topological entropy (see, e.g., [26]), and therefore $h(\mathcal{D}) = 0$ by Proposition 2.6.

Fitting this family to an observed process amounts to looking for periodic or "almost periodic" (also known as "quasi-periodic") structure in the observations. Intuitively, one is looking for up to d independent "periods" in a process. A process would have d independent "periods" if there were d periodic processes with incommensurate periods and the observed process is a function of all d of these periodic processes. For example, consider the classical situation in celestial mechanics in which two planets orbit a star and do not interact with each other. As each planet's trajectory will form an ellipse, the natural state space for the combined system is a two-dimensional torus, and the dynamics may be naturally desribed as a rotation of the torus, with the vector α being related to the periods of the two planets.

4. Background concepts and notation. This section introduces several key concepts, along with associated notation, that will be used in what follows. We begin by detailing the important connection between dynamical models and stationary processes, and then we establish a relationship between the entropy of a family and the entropy rates of its associated processes. We conclude by defining the joining of two stationary processes and a related, loss-based measure of divergence that will play a key role in characterizing the limiting behavior of minimum risk estimates.

4.1. *Processes associated with dynamical models.* Let (T, f) be a dynamical model on a compact metrizable state space $\mathcal{X}$. Recall that a Borel probability measure μ on $\mathcal{X}$ is said to be *invariant* under T if $\mu(T^{-1}A) = \mu(A)$ for all Borel sets $A \subseteq \mathcal{X}$. Let $\mathcal{M}(\mathcal{X}, T)$ be the set of Borel measures on $\mathcal{X}$ that are invariant under T, which is nonempty (see [74, p.152]). To each measure $\mu \in \mathcal{M}(\mathcal{X}, T)$ there is an associated real-valued process

$$\mathbf{U} = f(X), f(TX), f(T^2X), \ldots$$

where $X \in \mathcal{X}$ has distribution μ. The invariance of μ under T ensures that $\mathbf{U}$ is stationary. Here and in what follows we will regard real-valued processes as measures on the infinite product space $\mathbb{R}^{\mathbb{N}}$ equipped with its Borel sigma-field in the standard product topology.

Definition 4.1. Let $\mathcal{D} = \{(T_\theta, f_\theta) : \theta \in \Theta\}$ be a family of dynamical models. For each $\theta \in \Theta$ let

$$\mathcal{Q}_\theta \;=\; \Big\{\mathbf{U} = \big(f_\theta \circ T_\theta^k(X)\big)_{k\geq 0} : X \sim \mu \text{ for some } \mu \in \mathcal{M}(\mathcal{X}, T_\theta)\Big\},$$

the set of stationary processes associated with (T_θ, f_θ). Further, let $\mathcal{Q}_\mathcal{D} = \bigcup_{\theta\in\Theta} \mathcal{Q}_\theta$, the set of processes associated with the entire family of models $\mathcal{D}$.

4.2. *Connection between the family of processes $\mathcal{Q}_\mathcal{D}$ and the entropy $h(\mathcal{D})$.* The definition of the entropy $h(\mathcal{D})$ of a family of dynamical models is combinatorial in nature, and it does not involve measures. Nevertheless, $h(\mathcal{D})$ may also be characterized in a measure-theoretic way, using the entropy rates of the stationary processes $\mathbf{U} \in \mathcal{Q}_\mathcal{D}$. Let $\pi = \{A_1, \ldots, A_k\}$ be a finite Borel partition of $\mathbb{R}$ with k cells, and define $\pi(x)$ to be the index j of the cell A_j that contains x. Let $\mathbf{U}$ be any stationary process taking values in $\mathbb{R}$. For notation, let $[k] = \{1, \ldots, k\}$. Then for $n \geq 1$ and $b_1, \ldots, b_n \in [k]$ define

$$p(b_1, \ldots, b_n) = \mathbb{P}\big(\pi(U_1) = b_1, \ldots, \pi(U_n) \in b_n\big).$$

Further define the associated Shannon entropy of the sequence $\pi(U_1), \ldots, \pi(U_n)$,

$$H_n(\mathbf{U} : \pi) = -\sum_{b_1,\ldots,b_n} p(b_1, \ldots, b_n) \log p(b_1, \ldots, b_n).$$

The entropy rate of the process $(\pi(U_i))_{i\geq 0}$ is $H(\mathbf{U} : \pi) = \lim_n n^{-1} H_n(\mathbf{U}, \pi)$, where the existence of the limit holds as a result of subadditivity. The entropy rate of the process $\mathbf{U}$ is then given by

$$H(\mathbf{U}) = \sup_\pi H(\mathbf{U}, \pi).$$

where the supremum is over all finite Borel partitions of $\mathbb{R}$. The following lemma, which mirrors and makes use of the standard variational formula relating topological and measure theoretic entropy for dynamical systems (see [74, p. 190]), is established in the Supplementary Material [47, Appendix E.1].

LEMMA 4.2. *If $\mathcal{D}$ is any family of dynamical models satisfying (D1)-(D3), then*

$$h(\mathcal{D}) = \sup_{\mathbf{U}\in\mathcal{Q}_\mathcal{D}} H(\mathbf{U}).$$

4.3. *Joinings and divergence for stationary processes.* The statements and proofs of our principal results rely critically on stationary couplings of stationary processes, which are known as joinings.

DEFINITION 4.3. A *joining* of two stationary processes $\mathbf{U} = (U_k)_{k\geq 0}$ and $\mathbf{V} = (V_k)_{k\geq 0}$ is a stationary process $\mathbf{W} = \big((\tilde{U}_k, \tilde{V}_k)\big)_{k\geq 0}$ such that

$\tilde{\mathbf{U}} = (\tilde{U}_k)_{k\geq 0}$ has the same distribution as $\mathbf{U}$, and $\tilde{\mathbf{V}} = (\tilde{V}_k)_{k\geq 0}$ has the same distribution as $\mathbf{V}$. Let $\mathcal{J}(\mathbf{U}, \mathbf{V})$ denote the family of all joinings of $\mathbf{U}$ and $\mathbf{V}$.

By definition, a joining of two stationary processes is a coupling of the processes that is itself stationary. Note that the family $\mathcal{J}(\mathbf{U}, \mathbf{V})$ always contains the the so-called independent joining under which $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ are independent copies of $\mathbf{U}$ and $\mathbf{V}$, respectively, defined on the same probability space. Joinings were introduced by Furstenberg [17] in a general measure theoretic setting, and they have been widely studied in ergodic theory [14, 18]. For notational convenience, we use $[\mathbf{U}, \mathbf{V}]$ to denote a joining of $\mathbf{U}$ with $\mathbf{V}$. The joining of three or more stationary processes may be defined analogously. Several simple, but non-trivial, examples of joinings are given in the Supplementary Material [47, Appendix A].

DEFINITION 4.4. Let $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be a nonnegative loss function. The ℓ-divergence between two stationary processes $\mathbf{U}$ and $\mathbf{V}$ is the minimum expected loss of $(\tilde{U}_0, \tilde{V}_0)$ over all joinings of $\mathbf{U}$ and $\mathbf{V}$,

$$\gamma_\ell(\mathbf{U}, \mathbf{V}) \;=\; \inf_{\mathcal{J}(\mathbf{U},\mathbf{V})} \mathbb{E}\Big[\ell(\tilde{U}_0, \tilde{V}_0)\Big].$$

REMARK 4.5. The ℓ-divergence $\gamma_\ell(\mathbf{U}, \mathbf{V})$ is nonnegative, and it is symmetric whenever ℓ is symmetric. Furthermore, it also satisfies the triangle inequality whenever ℓ does. In the special case that $\ell(u, v) = (u - v)^2$ is the standard squared loss, $\gamma_\ell(\mathbf{U}, \mathbf{V})^{1/2}$ is a metric on the space of $\mathbb{R}$-valued stationary stochastic processes.

REMARK 4.6. Joinings were used in an analogous manner by Ornstein [55, 56, 57] to define the $\overline{d}$-distance between finite alphabet stationary processes based on the Hamming metric $\mathbb{I}(u_0 \neq v_0)$. The $\overline{d}$-distance was then extended by Gray et al. [19] to stationary processes with general alphabets and to arbitrary metrics $\rho(U_0, V_0)$. The divergence $\gamma_\ell(\cdot, \cdot)$ is simply the generalization of these distances to nonnegative loss functions $\ell(\cdot, \cdot)$ that need not be metrics.

REMARK 4.7. The fact that the infimum defining $\gamma_\ell(\cdot, \cdot)$ runs over the set of joinings, rather than the set of couplings, is critical. A minimizing joining makes the average loss between elements of the process as small as possible over the entire future of the process. By contrast, a minimizing coupling would make the processes as close as possible at time zero, without regard to their behavior in the future. The stationarity assumption constrains the

set of possible joinings: in some cases (arising in the arguments below) the only joining of two processes $\mathbf{U}$ and $\mathbf{V}$ is the independent joining, and the processes $\mathbf{U}$ and $\mathbf{V}$ are then said to be disjoint. This phenomenon appears even in the simple example presented in the Supplementary Material [47, Appendix A].

5. Convergence of Minimum Risk Estimates. This section is devoted to the asymptotic behavior of minimum risk estimates. The analysis relies on the joining based divergence defined in the previous section. We begin with a general theorem concerning the convergence of minimum risk estimators and then specialize, first to the case where the observed process has a signal plus noise structure, and then further to the case of squared loss. A key requirement in the signal plus noise setting is that the family $\mathcal{D}$ have entropy zero. In the final subsection we give a counterexample showing that the zero entropy conditions cannot be dropped.

5.1. *General convergence result.* As detailed in the previous section, a family $\mathcal{D}$ of dynamical models corresponds to a family $\mathcal{Q}_{\mathcal{D}} = \bigcup_{\theta\in\Theta} \mathcal{Q}_\theta$ of stationary processes. With this correspondence in mind, the problem of fitting models in $\mathcal{D}$ to an observed ergodic sequence $\mathbf{Y} = Y_0, Y_1, \ldots$ using the empirical risk $R_n(\theta : x)$ has a population analog in which one seeks processes $\mathbf{U} \in \mathcal{Q}_{\mathcal{D}}$ that minimize the divergence $\gamma_\ell(\mathbf{U}, \mathbf{Y})$ with the observed process $\mathbf{Y}$. The solution set of the population problem is the γ_ℓ-projection of $\mathbf{Y}$ onto $\mathcal{Q}_{\mathcal{D}}$, and the corresponding set of parameters is a natural limit set for empirical risk estimators. This leads to the following definition.

DEFINITION 5.1. Let $\mathcal{D}$ be a family of dynamical models parametrized by $\theta \in \Theta$, and let $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be a loss function. For any stationary ergodic process $\mathbf{Y}$, define

$$\Theta_\ell(\mathbf{Y}) \;=\; \operatorname*{argmin}_{\theta\in\Theta} \min_{\mathbf{U}\in\mathcal{Q}_\theta} \gamma_\ell(\mathbf{U}, \mathbf{Y}),$$

the set of parameters θ such that some process in $\mathcal{Q}_\theta$ minimizes the divergence with $\mathbf{Y}$.

The following theorem shows that the limiting behavior of minimum risk estimators is *fully characterized* by the set $\Theta_\ell(\mathbf{Y})$. The proof, which relies on results of McGoff and Nobel [46], is presented in Section 7.

THEOREM 5.2. *Let $\mathcal{D}$ be a family of dynamical models satisfying (D1)-(D3), and let ℓ be a lower semicontinuous loss function. If $\mathbf{Y}$ is a stationary ergodic process satisfying (C1), then $\Theta_\ell(\mathbf{Y})$ is non-empty and compact. Moreover,*

(a) Any sequence $\{\hat{\theta}_n = \theta_n(Y_0, \ldots, Y_{n-1})\}$ *of minimum risk estimates converges almost surely to* $\Theta_\ell(\mathbf{Y})$*;*

(b) For each $\theta \in \Theta_\ell(\mathbf{Y})$*, there exists a sequence of minimum risk estimates that converges almost surely to* θ*.*

Theorem 5.2 reduces the asymptotic analysis of empirical risk minimization to the analysis of the parameter set $\Theta_\ell(\mathbf{Y})$, which may not be a singleton. The conditions of the theorem place no restrictions on the relationship between the observation process $\mathbf{Y}$ and the family $\mathcal{D}$. The identifiability of optimal parameters is determined by the divergence γ_ℓ and the process families $\mathcal{Q}_\theta$. We show below how analysis of the limit set $\Theta_\ell(\mathbf{Y})$ yields both positive results (e.g. consistency) and negative results (inconsistency) in a signal plus noise setting.

5.2. *Signal Plus Noise.* In many situations it is natural to assume that the observed process $\mathbf{Y}$ is the componentwise sum of an underlying signal process $\mathbf{V}$ and an i.i.d. noise process $\boldsymbol{\varepsilon}$, neither of which is known. In this setting, one would like to know that the limiting behavior of minimum risk estimation is determined solely by the structure of the signal process $\mathbf{V}$ and is unaffected by the presence of the noise process $\boldsymbol{\varepsilon}$. In this subsection and the next we establish sufficient conditions for decoupling of signal and noise.

Assume that for each $k \geq 0$ the observation Y_k takes the form $Y_k = V_k + \varepsilon_k$ where $\mathbf{V} = \{V_k : k \geq 0\}$ is a stationary ergodic process and $\boldsymbol{\varepsilon} = \{\varepsilon_k : k \geq 0\}$ is an i.i.d. zero-mean noise process that is independent of $\mathbf{V}$. As a shorthand, we will write $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$. It is *not* assumed that $\mathbf{V}$ belongs to the family $\mathcal{Q}_\mathcal{D}$ of processes generated by the model family $\mathcal{D}$. We require the following integrability conditions:

$$\text{(C1)} \qquad \mathbb{E}\left[\sup_{|x| \leq K_\mathcal{D}} \ell(x, Y_0)\right] < \infty;$$

$$\text{(C2)} \qquad \mathbb{E}\left[\sup_{|x| \leq K_\mathcal{D}} \ell(x, V_0)\right] < \infty;$$

$$\text{(C3)} \qquad \mathbb{E}\,\ell(u, v + \varepsilon_0) < \infty \quad \text{for all } u, v \in \mathbb{R}.$$

Condition (C1) is the same condition that we require in the general setting, whereas (C2) and (C3) involve only the processes $\mathbf{V}$ and $\boldsymbol{\varepsilon}$, respectively. These conditions ensure integrability of the loss with respect to the three

processes $\mathbf{Y}$, $\mathbf{V}$, and $\boldsymbol{\varepsilon}$. If ℓ is the squared loss and $\mathbf{V}$ and $\boldsymbol{\varepsilon}$ have finite second moments, then conditions (C1)-(C3) are satisfied.

Let $\mathcal{D} = \{(T_\theta, f_\theta) : \theta \in \Theta\}$ be a family of dynamical models. Theorem 5.2 ensures that any sequence of minimum risk estimators for $\mathcal{D}$ based on observations of $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$ will converge to the set $\Theta_\ell(\mathbf{Y})$ of optimal parameters for $\mathbf{Y}$. In this setting, it is reasonable to ask if minimum risk estimation can recover optimal parameters for the underlying *signal process* $\mathbf{V}$ rather than the observed process $\mathbf{Y}$. The following result answers this question in the affirmative in two different cases. In the first case the signal process $\mathbf{V}$ is general, but we assume that the loss $\ell(u, v) = D_F(v, u)$ is the Bregman divergence of a continuously differentiable convex function $F : \mathbb{R} \to \mathbb{R}$, that is, $\ell(u, v) = F(v) - F(u) - (v - u)F'(u)$. In this case, we establish that minimum risk estimators converge to the optimal parameter set $\Theta_\ell(\mathbf{V})$ for the process $\mathbf{V}$. In the second case, we assume that the signal process $\mathbf{V}$ is generated by a dynamical model in the family $\mathcal{D}$ and impose a condition on the joint behavior of the loss function and the noise. In this case, we establish that minimum risk estimators converge to the set of parameters θ such that the set of processes $\mathcal{Q}_\theta$ contains $\mathbf{V}$.

THEOREM 5.3. *Let $\mathcal{D}$ be a family of dynamical models satisfying (D1)-(D3) with entropy $h(\mathcal{D}) = 0$. Let $\{\hat{\theta}_n : n \geq 1\}$ be any sequence of ℓ-minimum risk estimates for $\mathcal{D}$ based on an observed ergodic process $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$ satisfying (C1)-(C3).*

(a) If $\ell(u, v) = D_F(v, u)$ is the Bregman divergence of a continuously differentiable convex function then $\hat{\theta}_n$ converges almost surely to $\Theta_\ell(\mathbf{V})$.

(b) Suppose that $\mathbf{V}$ is an ergodic process in $\mathcal{Q}_\mathcal{D}$ and $\mathbb{E}\,\ell(u, v + \varepsilon_0) \geq \mathbb{E}\,\ell(0, \varepsilon_0)$ for all u, v, with equality if and only if $u = v$. Then $\hat{\theta}_n$ converges almost surely to $\{\theta \in \Theta : \mathbf{V} \in \mathcal{Q}_\theta\}$.

REMARK 5.4. The loss condition in part (b) of the theorem holds, e.g., if $\ell(u, v) = |u - v|$ is the absolute loss and the distribution of the noise has a unique median at zero.

REMARK 5.5. Recall that dynamical models include arbitrary (e.g., nonlinear) continuous observation functions $f_\theta : \mathcal{X} \to \mathbb{R}$. In the signal plus noise setting of the present section, these functions allow one to consider observation models of the form $Y = f(X) + \varepsilon$, where X is the underlying state and ε is independent noise. In short, the setting includes general observation functions, but the function must be applied *before* noise is

added. Nonetheless, we note that Theorem 5.3 remains true if the observation model has the form $\mathbf{Y} = F(\mathbf{V} + \boldsymbol{\varepsilon})$ where F is *linear*. Indeed, in this case $F(\mathbf{V} + \boldsymbol{\varepsilon}) = F(\mathbf{V}) + F(\boldsymbol{\varepsilon}) = \mathbf{V}' + \boldsymbol{\varepsilon}'$, and $\mathbf{V}'$ and $\boldsymbol{\varepsilon}'$ can be seen to satisfy the hypotheses of the theorem. Investigation of more general noise models is interesting but beyond the scope of the present paper.

EXAMPLE 5.6. Consider the estimation of a rotational period, as in Example 3.4. For concreteness, consider the case $d = 1$, and consider a set of angles $\Theta = [0, \pi]$. Suppose that there is a single vector $h \in \mathbb{R}^2$ such that the observation function f_θ is given by $f(x) = \langle x, h \rangle$ for all θ. Now suppose that ε is a random variable in $\mathbb{R}^2$ with zero mean and finite variance, and suppose that $\boldsymbol{\varepsilon}$ is an i.i.d. process whose marginals have the same distribution as ε. Finally, suppose that the observation process is given by $\mathbf{Y} = (Y_k)_{k \geq 0}$, where

$$Y_k = f\Big(R^k_{\theta_0}(X) + \varepsilon_k\Big),$$

$\theta_0 \in [0, \pi]$ is irrational, and X is uniformly distributed on the circle $S^1 \subset \mathbb{R}^2$. Since f is linear, the previous remark applies, and we conclude that the hypotheses of Theorem 5.3 (both parts) are satisfied with the squared loss $\ell(u, v) = (u - v)^2$. Additionally, one may check that θ_0 is identifiable within Θ (i.e., it is the only parameter θ such that the signal process is contained in $\mathcal{Q}_\theta$). Then we conclude that any sequence $(\theta_n)_{n \geq 1}$ of minimum risk estimates converges almost surely to θ_0.

As the squared loss $\ell_2(u, v) = (u - v)^2$ is a Bregman divergence, minimum risk fitting of a zero entropy family will converge to the optimal parameter set for the signal by Theorem 5.3. The next result extends this result to the more general case in which the mean width of the family is zero.

THEOREM 5.7. *Let $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$, where $\mathbf{V}$ is ergodic and $\boldsymbol{\varepsilon}$ is an i.i.d. process with mean zero and finite variance. If the mean width $\kappa(\mathcal{D} : \boldsymbol{\varepsilon}) = 0$, then any sequence of least squares estimates converges almost surely to $\Theta_{\ell_2}(\mathbf{V})$.*

In our final result of this section, we establish the consistency of least squares estimation for a family of transformations on a compact state space in $\mathbb{R}^d$ where each observation function is the identity. Suppose $\mathcal{X} \subset \mathbb{R}^d$ is compact and $\{T_\theta : \theta \in \Theta\}$ is a family of transformations on $\mathcal{X}$ such that Θ is a compact metric space and $(\theta, x) \mapsto T_\theta(x)$ is continuous. Further, suppose that $Y_k = T^k_{\theta^*}(X) + \varepsilon_k$ where X is distributed according to an ergodic measure $\mu \in \mathcal{M}(\mathcal{X}, T_{\theta^*})$, and $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k \geq 0}$ is i.i.d. with mean zero and finite variance and is independent of X.

COROLLARY 5.8. *If the topological entropy of T_θ is zero for all $\theta \in \Theta$, then any sequence of least squares estimates converges almost surely to the set $\{\theta \in \Theta : \mu(T_\theta = T_{\theta^*}) = 1\}$.*

The limit set in Corollary 5.8 contains the true parameter θ^*, and serves as the natural identifiability class of θ^* in this context.

5.3. *A negative result.* It is assumed above that the model family $\mathcal{D}$ has zero entropy. The next proposition shows that the zero entropy assumption is, in general, necessary for consistent estimation. In particular, if $h(\mathcal{D})$ is positive then least squares estimation can fail to identify the optimal parameters of the signal process $\mathbf{V}$, even if the signal process is generated by a dynamical model in the family. The underlying idea is that a family $\mathcal{D}$ with positive entropy is capable of tracking the noise, and consequently least squares estimates will overfit the observed sequence.

Let us say that a family $\mathcal{D}$ is inseparable from Gaussian noise if there exists an ergodic process $\mathbf{V}$ in $\mathcal{Q}_\mathcal{D}$ and $\sigma_0 > 0$ such that for every $\sigma > \sigma_0$ the limiting parameter set $\Theta_{\ell_2}(\mathbf{Y})$ of least squares estimates derived from $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$ with $\varepsilon_i \sim N(0, \sigma^2)$ does not capture $\mathbf{V}$ in the sense that $\mathbf{V} \notin \mathcal{Q}_\theta$ for each $\theta \in \Theta_{\ell_2}(\mathbf{Y})$.

PROPOSITION 5.9. *Let $\mathcal{D}$ be a family of dynamical models with entropy $h(\mathcal{D}) > 0$. If there exists $\theta_0 \in \Theta$ such that $h(\{(T_{\theta_0}, f_{\theta_0})\}) = 0$ and $\mathcal{Q}_{\theta_0} \setminus \cup_{\theta \neq \theta_0} \mathcal{Q}_\theta$ contains an ergodic process, then $\mathcal{D}$ is inseparable from Gaussian noise.*

REMARK 5.10. It is relatively easy to construct families $\mathcal{D}$ satisfying the conditions of Proposition 5.9. See the Supplementary Material [47, Example D.1]. Informally, the inconsistency phenomenon illustrated by this example can be shown to occur for any family in which there is a parameter θ_0 with zero entropy, there is a parameter with positive entropy, and the processes associated to θ_0 are distinct from the processes associated to all parameter values with positive entropy.

6. Discussion of related work. The results of this paper have points of overlap with recent work in the statistics and machine learning literature concerning estimation, forecasting, and prediction from dependent observations. While some of this work, for example Morvai and Weiss [53, 52], Nobel [54], and Adams and Nobel [1], is focused on asymptotics for general ergodic observations, a number of papers provide rates of convergence or finite sample bounds under more stringent assumptions.

A complete survey of all recent work on learning from dependent data is beyond the scope of the present paper, but we nonetheless mention several recent directions of research in this area. In representative early work, Arcones and Yu [5] prove central limit theorems for empirical and U-processes of stationary mixing processes. More recently, Modha and Masry [50], Meir [48], and Alquier and Wintenberger [4] establish oracle inequalities and finite sample bounds for predicting the next value of a stationary process. Agarwal and Duchi [3], Kuznetsov and Mohri [28, 30, 29], and Zimin and Lampert [78] establish finite sample performance bounds on the conditional risk of online learning algorithms for predicting dependent time series. Each of the papers cited above imposes mixing conditions (as in [77]) on the observations, as well as regularity conditions on the loss function and model family of interest. Shalizi and Kontorovich [65] consider learning mixtures of stationary processes, while Kontorovich [27] studies statistical estimation using finite automata with bounded memory. Mohri and Rostamizadeh [51] provide stability-based generalization bounds from ϕ-mixing and β-mixing processes. Hang and Steinwart [20] and Steinwart and Christmann [68] obtain rates of convergence for empirical risk minimization from α-mixing observations, while Wong *et al.* [75] establish finite sample bounds for Lasso-based inference under β-mixing conditions. In another direction, Rakhlin *et al.* [62] and Rakhlin and Sridharan [61] have established exponential inequalities for suprema of martingale difference sequences by using and extending ideas from machine learning, including Rademacher complexity and deterministic regret inequalities. Finally, let us mention that both Dean *et al.* [15] and Tu *et al.* [69] provide finite sample bounds for system estimation in the context of certain control problems.

As noted in the introduction, the problem of fitting dynamical models differs from the inference problems above as both the observations *and* the models under study can exhibit dynamical behavior and long-range dependence. Moreover, our principal results make no assumptions concerning mixing properties of the observed process $\mathbf{Y}$, mixing properties or stationary distributions of the dynamical models $\mathcal{D}$, smoothness of the loss $\ell(\cdot, \cdot)$ (beyond lower semicontinuity), or the relationship between the observed process and the family of models being fit. This general setting enables us to study the asymptotic behavior of minimum risk estimation for dynamical models in a variational framework where the roles of the observed process, the loss, and (most critically) the model family are clearly delineated.

The results here provide a framework for, and initial progress towards, the detailed analyses of specific problems and model families that might lead to rates of convergence, or finite sample performance bounds. It is evident from

the papers above that stronger results, e.g., rates of convergence, will require substantially stronger assumptions, including mixing conditions (with geometric or polynomial rates) on the observed process, smoothness (possibly with convexity) of the loss function, and stronger, covering-based complexity constraints on the family of dynamical models. If mixing type conditions are required for the dynamical models themselves, then these conditions would require even stronger assumptions, as mixing conditions typically hold only for distinguished invariant measures and observation functions.

A number of the papers cited above make use of exponential probability bounds, typically Azuma-Hoeffding type inequalities, to control error terms that are sums of martingale differences. Martingale differences do not arise in the theoretical analysis of this paper, but we note that there are some uses of reverse martingale methods in the dynamics literature [38]. Investigating martingale approaches to the problems considered here represents an interesting direction for future research.

Our work is also related to a line of research concerning least squares estimation of individual sequences from noisy observations, see for example [59, 71, 76]. Pollard and Radchenko [59] use empirical process theory to establish consistency and asymptotic normality of least squares estimation for individual sequences from signal plus noise. In the present work, we consider sets of individual sequences that arise from a continuous family of dynamical models, as in (2.1), and we are interested in estimation of a dynamical invariant parameter (i.e. θ), rather than the signal sequence itself.

Ornstein and Weiss [58] studied the estimation of a stochastic process from its samples. They proposed an inference procedure, based on matching k-block frequencies, and characterize when it produces consistent estimates of the observed stochastic process in the d-bar metric.

Furstenberg's original work on joinings [17] includes an application of joinings to a nonlinear filtering problem. Beyond this application, we are not aware of other uses of joinings in the literature on statistical inference.

Some of Furstenberg's original results are extended in recent work of Lev, Peled, and Peres [35]. Given an infinite sequence equal to a target signal plus noise, they consider the problem of detecting whether the signal is non-zero, and the problem of recovering the signal from the given sequence. Target sequences are assumed to belong to a known family (as in [59]), and their analysis places no restrictions (beyond measurability) on the detection and filtering procedures, which can be functions of the entire sequence of observations.

Finally, we mention that statistical inference in the context of dynamical systems has been considered in a variety of subject areas; see the survey [45]

for a broad overview and references. Dynamical systems in the observational noise setting have been studied in [31, 32, 44], and statistical prediction in the context of dynamical systems has been considered in [21, 22, 68, 73].

6.1. *Generalizations and future work.* Generalization of all of the definitions and results of the paper to $\mathbb{R}^d$-valued models and processes is straightforward, requiring only minor changes of notation. We omit the details. In a different direction, one could analyze families of dynamical models defined on a non-compact state space $\mathcal{X}$ with uniformly bounded observation functions, requiring only measurability of the maps $(\theta, x) \mapsto T_\theta(x)$ and $(\theta, x) \mapsto f_\theta(x)$. For families $\mathcal{D}$ of this more general type, the set $\mathcal{U}_\mathcal{D}$ of associated sequences would not necessarily be a closed (hence compact) subset of $\mathbb{R}^\mathbb{N}$, and in this case one needs to consider the closure of $\mathcal{U}_\mathcal{D}$ in $\mathbb{R}^\mathbb{N}$, along with all the stationary processes supported on this set. The analysis here can be carried out in this more general setting, but the corresponding results are difficult to interpret in the context of the original problem.

7. Optimal tracking and proof of Theorem 5.2. In this section we detail connections between fitting dynamical models and the optimal tracking problem studied in [46]. In particular, we construct a single dynamical system that captures the family $\mathcal{D}$ of dynamical models, and we analyze this system using optimal tracking.

7.1. *Optimal tracking.* The tracking problem for dynamical systems concerns two systems: a model system consisting of a compact metric space $\mathcal{Z}$ and a continuous map $T : \mathcal{Z} \to \mathcal{Z}$, and an observed system consisting of a complete separable metric space $\mathcal{Y}$ and a Borel measurable map $S : \mathcal{Y} \to \mathcal{Y}$. Given the initial segment of a trajectory $y, S(y), \ldots, S^{n-1}(y)$ of the observed system, one seeks a corresponding initial condition $z_n \in \mathcal{Z}$ such that the trajectory $z_n, T(z_n), \ldots, T^{n-1}(z_n)$ from the model system "tracks" the given trajectory from the observed system. An optimal tracking trajectory is chosen by minimizing an additive cost functional

$$\sum_{k=0}^{n-1} c\big(S^k y, T^k z\big),$$

where $c : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ is a fixed lower semicontinuous cost function.

We will appeal to results from [46] for optimal tracking, which we summarize here for completeness. Suppose that the initial condition y of the observed trajectory is drawn from an ergodic measure $\nu \in \mathcal{M}(\mathcal{Y}, S)$, and that $\sup_z |c(y, z)|$ is bounded above by a function in $L^1(\nu)$. Let Θ be a compact metrizable space, and let $\varphi : \mathcal{Z} \to \Theta$ be a continuous map that is

invariant with respect under the transformation T, i.e. $\varphi \circ T = \varphi$, so that φ is invariant on trajectories of T. Let $\mathbf{V} = (S^k(V_0))_{k\geq 0}$, where $V_0 \sim \nu$. Also, for $\theta \in \Theta$, let $\mathcal{M}_\theta(\mathcal{Z}, T)$ denote the set of measures $\mu \in \mathcal{M}(\mathcal{Z}, T)$ such that $\mu(\varphi^{-1}\{\theta\}) = 1$, i.e., μ is invariant under T and supported on the (closed) set of z in $\mathcal{Z}$ for which $\varphi(z) = \theta$. Finally, let $\mathcal{J}(\mathbf{V} : \theta)$ denote the set of all joinings of the process $\mathbf{V}$ with a process $\mathbf{U} = (T^k(U_0))_{k\geq 0}$, where $U_0 \sim \mu$ for some $\mu \in \mathcal{M}_\theta(\mathcal{Z}, T)$.

THEOREM A ([46]). *Let $z_n : \mathcal{Y}^n \to \mathcal{Z}$, $n \geq 1$, be Borel measurable functions. If for ν almost every $y \in \mathcal{Y}$ the sequence $\hat{z}_n = z_n(y, \dots, S^{n-1}y)$ optimally tracks $y, Sy, \dots$ in the sense that*

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} c\Big(S^k y, T^k \hat{z}_n\Big) = \lim_n \inf_{z\in\mathcal{Z}} \frac{1}{n} \sum_{k=0}^{n-1} c\Big(S^k y, T^k z\Big), \tag{7.1}$$

then $\hat{\theta}_n = \varphi(\hat{z}_n)$ converges ν almost surely to the non-empty, compact set

$$\Theta_{min} = \operatorname*{argmin}_{\theta\in\Theta} \min_{\mathcal{J}(\mathbf{V}:\theta)} \mathbb{E}\big[c(\tilde{V}_0, \tilde{U}_0)\big]. \tag{7.2}$$

Furthermore, for any $\theta \in \Theta_{min}$, there exists $\hat{z}_n$ such that (7.1) holds and $\hat{\theta}_n = \varphi(\hat{z}_n)$ converges to θ.

7.2. *Proof of Theorem 5.2.* To begin, we describe how fitting a family of dynamical models to an observed stochastic process can be cast as a tracking problem. As a first step, we define a single dynamical system that encapsulates the family of dynamical models. Consider the state space

$$\mathcal{Z} = \left\{ \Big(\theta,\ \big(f_\theta \circ T_\theta^k(x)\big)_{k\geq 0}\Big) : \theta \in \Theta,\ x \in \mathcal{X} \right\} \subseteq \Theta \times \mathbb{R}^{\mathbb{N}},$$

and define the transformation $T : \mathcal{Z} \to \mathcal{Z}$ by $T(\theta, (u_k)_{k\geq 0}) = (\theta, (u_{k+1})_{k\geq 0})$. The next lemma establishes some basic properties of the dynamical system $(\mathcal{Z}, T)$. Here and in what follows $\mathbb{R}^{\mathbb{N}}$ is equipped with the usual product topology.

LEMMA 7.1. *The set $\mathcal{Z}$ is a compact subset of $\Theta \times \mathbb{R}^{\mathbb{N}}$, and the map T is continuous. If μ is an ergodic element of $\mathcal{M}(\mathcal{Z}, T)$, then there exists $\theta \in \Theta$ and an ergodic process $\mathbf{U} \in \mathcal{Q}_\theta$ with distribution ξ such that $\mu = \delta_\theta \otimes \xi$.*

PROOF. Continuity of the map T follows from continuity of the left shift $\tau : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$. As both the parameter space Θ and the state space $\mathcal{X}$ are

compact by assumption, the product $\Theta \times \mathcal{X}$ is compact. Define a map $\pi : \Theta \times \mathcal{X} \to \Theta \times \mathbb{R}^{\mathbb{N}}$ by

$$\pi(\theta, x) = \Big(\theta, \big(f_\theta \circ T_\theta^k(x)\big)_{k \geq 0}\Big).$$

It is easy to see that $\mathcal{Z}$ is the image of $\Theta \times \mathcal{X}$ under π. To establish that $\mathcal{Z}$ is compact, it therefore suffices to show that π is continuous.

Let $\{(\theta_n, x_n)\}_{n \geq 1}$ be a sequence converging to (θ, x) in $\Theta \times \mathcal{X}$, and let $K \geq 1$. The continuity conditions (D2) and (D3) imply that for $0 \leq k \leq K$,

$$\lim_n f_{\theta_n} \circ T_{\theta_n}^k(x_n) = f_\theta \circ T_\theta^k(x).$$

As K was arbitrary, it follows that $\{\pi(\theta_n, x_n)\}_{n \geq 1}$ converges to $\pi(\theta, x)$ in $\Theta \times \mathbb{R}^{\mathbb{N}}$ with the product topology, and therefore π is continuous.

For the last statement of the lemma, define the map $R : \Theta \times \mathcal{X} \to \Theta \times \mathcal{X}$ by the rule $R(\theta, x) = (\theta, T_\theta(x))$, which is known as the skew-product over the identity in the dynamics literature. By construction, we have $\pi \circ R = T \circ \pi$. In the dynamics literature, π is called a factor map from $(\Theta \times \mathcal{X}, R)$ onto $(\mathcal{Z}, T)$. It is a standard fact [16, p. 19] that the associated map from $\mathcal{M}(\Theta \times \mathcal{X}, R)$ to $\mathcal{M}(\mathcal{Z}, T)$ defined by $\eta \mapsto \eta \circ \pi^{-1}$ is a surjection.

Now let $\mu \in \mathcal{M}(\mathcal{Z}, T)$ be ergodic. Since the map from $\mathcal{M}(\Theta \times \mathcal{X}, R)$ to $\mathcal{M}(\mathcal{Z}, T)$ given by $\eta \mapsto \eta \circ \pi^{-1}$ is a surjection, there exists $\eta \in \mathcal{M}(\Theta \times \mathcal{X}, R)$ such that $\eta \circ \pi^{-1} = \mu$. As $\mathrm{proj}_\Theta \circ T = \mathrm{proj}_\Theta$, the induced measure $\eta \circ (\mathrm{proj}_\Theta \circ \pi)^{-1} = \mu \circ \mathrm{proj}_\Theta^{-1}$ on Θ must be invariant under the identity map. Also, it must be ergodic, since μ is ergodic. As the only ergodic measures for the identity map are the point masses, we see that there exists $\theta \in \Theta$ such that $\eta \circ (\mathrm{proj}_\Theta \circ \pi)^{-1} = \delta_\theta$. Then $\eta = \delta_\theta \otimes \xi'$ for some invariant measure $\xi' \in \mathcal{M}(\mathcal{X}, T_\theta)$, and therefore $\mu = \eta \circ \pi^{-1} = \delta_\theta \otimes \xi$, where ξ is the distribution of a stationary process in $\mathcal{Q}_\theta$. To see that ξ is ergodic, note that $\xi = \mu \circ (\mathrm{proj}_{\mathbb{R}^{\mathbb{N}}})^{-1}$ and μ is ergodic. □

We now proceed with the proof of Theorem 5.2. The observed process $\mathbf{Y}$ gives rise to an observed dynamical system in the tracking problem, where $\mathcal{Y} = \mathbb{R}^{\mathbb{N}}$, $S : \mathcal{Y} \to \mathcal{Y}$ is the left shift $S((u_k)_{k \geq 0}) = (u_{k+1})_{k \geq 0}$, and ν is the distribution of $\mathbf{Y}$ on $\mathbb{R}^{\mathbb{N}}$. Define the cost function $c : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ by

$$c\Big(\mathbf{v}, \big(\theta, \mathbf{u}\big)\Big) = \ell(u_0, v_0).$$

By Lemma 7.1, the hypotheses of Theorem A are satisfied. Then an application of Theorem A shows that any sequence of minimum ℓ-risk parameters $(\hat{\theta}_n)_{n \geq 1}$ converges almost surely to the set Θ_{min}. Additionally, using the

second sentence of Lemma 7.1 and the variational characterizations of Θ_{min} and $\Theta_\ell(\mathbf{Y})$, we see that $\Theta_{min} = \Theta_\ell(\mathbf{Y})$. Furthermore, the conclusions of Theorem A give that the projection $\Theta_\ell(\mathbf{Y})$ is nonempty and compact, and the ("converse") statement in Theorem 5.2 (b) holds. We have thus proved Theorem 5.2.

8. Proofs for Signal Plus Noise. This section contains the proofs of Theorems 5.3 and 5.7, and Proposition 5.9 concerning the behavior of minimum risk estimation in the signal plus noise setting. In addition, we state and establish several auxiliary results that may be of independent interest. We begin with a straightforward extension of Theorem A. A proof is included in the Supplementary Material [47, Appendix E.2] for completeness.

THEOREM 8.1. *Let* $\mathbf{U}$, $\mathbf{V}$, *and* $\mathbf{W}$ *be real valued stationary processes such that* $H(\mathbf{U}) = 0$, $\mathbf{V}$ *is ergodic, and* $\mathbf{W}$ *is i.i.d. If* $[\mathbf{U}, \mathbf{V}, \mathbf{W}]$ *is any joining of these three processes such that* $\mathbf{V}$ *and* $\mathbf{W}$ *are independent, then the joint process* $[\mathbf{U}, \mathbf{V}]$ *is independent of* $\mathbf{W}$.

The proofs below require the concept of a relatively independent joining, which is a standard construction in ergodic theory (see [18, p. 126] or [14]). Let $\mathbf{U}$ and $\mathbf{W}$ be stationary processes taking values in complete separable metric spaces $\mathcal{U}$ and $\mathcal{W}$, respectively. A measurable map $f : \mathcal{U} \to \mathcal{W}$ is said to map $\mathbf{U}$ onto $\mathbf{W}$ if the process $f(\mathbf{U}) := (f(U_k))_{k\geq 0}$ has the same distribution as $\mathbf{W}$.

THEOREM B (Relatively Independent Joining). *Suppose* $\mathbf{U}$, $\mathbf{V}$, *and* $\mathbf{W}$ *are stationary processes taking values in (possibly distinct) complete separable metric spaces. If there are Borel measurable maps* f *and* g *such that* $f(\mathbf{U})$ *and* $g(\mathbf{V})$ *each have the same distribution as* $\mathbf{W}$, *then there is a joining* $[\mathbf{U}, \mathbf{V}]$ *of* $\mathbf{U}$ *and* $\mathbf{V}$ *such that* $f(\mathbf{U}) = g(\mathbf{V})$ *almost surely.*

The joining $[\mathbf{U}, \mathbf{V}]$ in Theorem B is called the relatively independent joining of $\mathbf{U}$ and $\mathbf{V}$ (relative to $\mathbf{W}$).

The following general result establishes the ability of minimum risk estimation to separate signal from noise for zero entropy model families. Separation is relative to an auxiliary loss function L that depends on both the given loss function ℓ and the noise process.

PROPOSITION 8.2. *Let* $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$ *satisfy (C1)-(C3), and let* $\mathcal{D}$ *satisfy (D1)-(D3). If* $h(\mathcal{D}) = 0$, *then any sequence of minimum* ℓ*-risk estimates converges almost surely to* $\Theta_L(\mathbf{V})$, *where* $L(u, v) := \mathbb{E}\,\ell(u, v + \varepsilon_0)$.

REMARK 8.3. Since $\ell(\cdot,\cdot)$ is nonnegative and lower semicontinuous, the auxiliary loss function $L(\cdot,\cdot)$ has the same properties (using Fatou's Lemma for the lower semicontinuity). If a given process $\mathbf{Y}$ can be expressed in two different ways as $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$ and $\mathbf{Y} = \mathbf{V}' + \boldsymbol{\varepsilon}'$, then the proof of Proposition 8.2 shows that $\Theta_L(\mathbf{V}) = \Theta_{L'}(\mathbf{V}')$, where L' is defined using $\boldsymbol{\varepsilon}'$ in place of $\boldsymbol{\varepsilon}$.

PROOF. By Theorem 5.2 any sequence of minimal ℓ-risk estimates converges almost surely to $\Theta_\ell(\mathbf{Y})$. It therefore suffices to show that $\Theta_\ell(\mathbf{Y}) = \Theta_L(\mathbf{V})$. To this end, let $\mathbf{U}$ be any process in $\mathcal{Q}_\mathcal{D}$, and let $[\mathbf{U}, \mathbf{Y}]$ be a joining of $\mathbf{U}$ and $\mathbf{Y}$ that is optimal in the sense that $\mathbb{E}\big[\ell(U_0, Y_0)\big] = \gamma_\ell(\mathbf{U}, \mathbf{Y})$. It follows from Lemma 4.2 and the assumption that $h(\mathcal{D}) = 0$ that the entropy rate $H(\mathbf{U}) = 0$. Let $[\mathbf{V}, \boldsymbol{\varepsilon}]$ be the independent joining of $\mathbf{V}$ and $\boldsymbol{\varepsilon}$. As $\mathbf{Y}$ and $\mathbf{V} + \epsilon$ have the same distribution, Theorem B ensures the existence of a joining $[\mathbf{U}, \mathbf{Y}, \mathbf{V}, \boldsymbol{\varepsilon}]$ such that $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$ almost surely. Projecting this joining onto its first, third, and fourth coordinates, we obtain a joining $[\mathbf{U}, \mathbf{V}, \boldsymbol{\varepsilon}]$ satisfying the conditions of Theorem 8.1. In particular, $[\mathbf{U}, \mathbf{V}]$ is independent of $\boldsymbol{\varepsilon}$. By conditioning on $[\mathbf{U}, \mathbf{V}]$ we find that

$$\begin{aligned}\mathbb{E}\big[\ell(U_0, Y_0)\big] = \mathbb{E}\big[\ell(U_0, V_0 + \varepsilon_0)\big] &= \mathbb{E}\Big[\mathbb{E}\big[\ell(u, v + \varepsilon_0) \mid U_0 = u, V_0 = v\big]\Big] \\ &= \mathbb{E}\Big[L(U_0, V_0)\Big] \geq \gamma_L(\mathbf{U}, \mathbf{V}),\end{aligned}$$

from which it follows that $\gamma_\ell(\mathbf{U}, \mathbf{Y}) \geq \gamma_L(\mathbf{U}, \mathbf{V})$.

Now let $[\mathbf{U}, \mathbf{V}]$ be a joining such that $\mathbb{E}\big[L(U_0, V_0)\big] = \gamma_L(\mathbf{U}, \mathbf{V})$. Let $[\mathbf{U}, \mathbf{V}, \boldsymbol{\varepsilon}]$ be the independent joining of $[\mathbf{U}, \mathbf{V}]$ with $\boldsymbol{\varepsilon}$, and let $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$. Then

$$\begin{aligned}\mathbb{E}\big[L(U_0, V_0)\big] &= \mathbb{E}\Big[\mathbb{E}\big[\ell(u, v + \varepsilon_0) \mid U_0 = u, V_0 = v\big]\Big] \\ &= \mathbb{E}\big[\ell(U_0, V_0 + \varepsilon_0)\big] = \mathbb{E}\big[\ell(U_0, Y_0)\big] \geq \gamma_\ell(\mathbf{U}, \mathbf{Y}).\end{aligned}$$

Thus $\gamma_L(\mathbf{U}, \mathbf{V}) \geq \gamma_\ell(\mathbf{U}, \mathbf{Y})$, and we conclude that $\gamma_L(\mathbf{U}, \mathbf{V}) = \gamma_\ell(\mathbf{U}, \mathbf{Y})$. As $\mathbf{U} \in \mathcal{Q}_\mathcal{D}$ was arbitrary, it follows that $\Theta_\ell(\mathbf{Y}) = \Theta_L(\mathbf{V})$, and the proof is complete. □

PROOF OF THEOREM 5.3. By Proposition 8.2 any sequence of minimum risk parameters converges almost surely to $\Theta_L(\mathbf{V})$, so it suffices to examine this set under the conditions of parts (a) and (b) the theorem.

Part (a): Based on the assumption that $\mathbb{E}(\varepsilon_0) = 0$ and $\ell(u, v) = D_F(v, u)$,

we have

$$\begin{aligned} L(u,v) \;=\; \mathbb{E}\ell(u, v+\varepsilon_0) \;&=\; \mathbb{E}F(v+\varepsilon_0) - F(u) - (v-u)F'(u) \\ &=\; \ell(u,v) + G(v), \end{aligned}$$

where $G(v) = \mathbb{E}F(v+\varepsilon_0) - F(v)$ depends only on v and the distribution of ε_0, and is non-negative since F is convex. Thus, for any $\theta \in \Theta$ and any $\mathbf{U} \in \mathcal{Q}_\theta$,

$$\begin{aligned} \gamma_L(\mathbf{U},\mathbf{V}) \;&=\; \inf_{\mathcal{J}(\mathbf{U},\mathbf{V})} \mathbb{E}L(U_0,V_0) \;=\; \inf_{\mathcal{J}(\mathbf{U},\mathbf{V})} \Big\{ \mathbb{E}\ell(U_0,V_0) + \mathbb{E}G(V_0) \Big\} \\ &=\; \gamma_\ell(\mathbf{U},\mathbf{V}) + \mathbb{E}G(V_0). \end{aligned}$$

It follows that $\Theta_L(\mathbf{V}) = \Theta_\ell(\mathbf{V})$, which establishes Part (a).

Part (b): Suppose that $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$ where $\mathbf{V} \in \mathcal{Q}_{\theta_0}$ is ergodic. Let $\mathbf{U} \in \mathcal{Q}_{\mathcal{D}}$ and let $[\mathbf{U},\mathbf{V}]$ be a joining of $\mathbf{U}$ and $\mathbf{V}$. The assumption that $\mathbb{E}\,\ell(x, y+\varepsilon_0) \geq \mathbb{E}\,\ell(0,\varepsilon_0)$ with equality if and only if $x = y$ ensures that

$$\begin{aligned} \mathbb{E}\big[L(U_0,V_0)\big] \;&=\; \mathbb{E}\Big[\mathbb{E}\big[\ell(u, v+\epsilon_0) \mid U_0 = u, V_0 = v\big]\Big] \\ &\geq\; \mathbb{E}\big[\ell(0,\epsilon_0)\big], \end{aligned}$$

with equality if and only if $U_0 = V_0$ almost surely. As $[\mathbf{U},\mathbf{V}]$ is a joining, and is therefore stationary, $U_0 = V_0$ almost surely if and only if $\mathbf{U} = \mathbf{V}$ almost surely. Thus, we have shown that $\gamma_L(\mathbf{U},\mathbf{V}) \geq \mathbb{E}\big[\ell(0,\epsilon_0)\big]$, with equality if and only if $\mathbf{U} = \mathbf{V}$. Therefore the set of θ minimizing the quantity $\min_{\mathbf{U}\in\mathcal{Q}_\theta} \gamma_L(\mathbf{U},\mathbf{V})$ is exactly the set of θ such that $\mathbf{V} \in \mathcal{Q}_\theta$. Hence $\Theta_L(\mathbf{V}) = \{\theta \in \Theta : \mathbf{V} \in \mathcal{Q}_\theta\}$, which finishes the proof of Part (b).

□

8.1. *Least squares estimation.* Here we provide proofs of Theorem 5.7 and Corollary 5.8 concerning least squares estimation. It is possible to give more direct proofs of these results that avoid a direct appeal to Theorem 5.2, at the expense of greater length, but we make use of this general result in the arguments below.

Proof of Theorem 5.7. By Theorem 5.2, any sequence of least squares parameters converges almost surely to $\Theta_{\ell_2}(\mathbf{Y})$, so it suffices to show that $\Theta_{\ell_2}(\mathbf{Y}) = \Theta_{\ell_2}(\mathbf{V})$.

Fix a parameter $\theta \in \Theta$ and a process $\mathbf{U} \in \mathcal{Q}_\theta$. Let $[\mathbf{U},\mathbf{Y}]$ be any joining of $\mathbf{U}$ with the observed process $\mathbf{Y}$, and let $[\mathbf{V},\boldsymbol{\varepsilon}]$ be the independent joining

of the signal and noise processes. Using Theorem B, let $[\mathbf{U}, \mathbf{Y}, \mathbf{V}, \boldsymbol{\varepsilon}]$ be the relatively independent joining of $[\mathbf{U}, \mathbf{Y}]$ with $[\mathbf{V}, \boldsymbol{\varepsilon}]$ such that $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$ almost surely. Then under this joining

$$\begin{aligned}\mathbb{E}\Big[\big|U_0 - Y_0\big|^2\Big] &= \mathbb{E}\Big[\big|U_0 - (V_0 + \varepsilon_0)\big|^2\Big] \\ &= \mathbb{E}\Big[\big|U_0 - V_0\big|^2\Big] - 2\mathbb{E}\big[(U_0 - V_0) \cdot \varepsilon_0\big] + \mathbb{E}\Big[\varepsilon_0^2\Big].\end{aligned}$$

As the mean width $\kappa(\mathcal{D} : \boldsymbol{\varepsilon}) = 0$ by assumption, Proposition C.3 ensures that $\mathbb{E}[U_0 \cdot \varepsilon_0] = 0$. Moreover, since V_0 is independent of ε_0 and ε_0 has zero mean, $\mathbb{E}[V_0 \cdot \varepsilon_0] = 0$. It follows from the previous display that

$$\mathbb{E}\Big[\big|U_0 - Y_0\big|^2\Big] = \mathbb{E}\Big[\big|U_0 - V_0\big|^2\Big] + \mathbb{E}\Big[\varepsilon_0^2\Big]. \tag{8.1}$$

Since $\mathbb{E}[\epsilon_0^2]$ is a constant that depends only on ϵ_0 (and not on the joining), and since $\mathbf{U} \in \mathcal{Q}_\theta$ was arbitrary, we conclude that

$$\min_{\mathbf{U} \in \mathcal{Q}_\theta} \gamma_{\ell_2}(\mathbf{U}, \mathbf{Y}) \geq \min_{\mathbf{U} \in \mathcal{Q}_\theta} \gamma_{\ell_2}(\mathbf{U}, \mathbf{V}) + \mathbb{E}[\epsilon_0^2]. \tag{8.2}$$

Now let $\mathbf{U}'$ be any process in $\mathcal{Q}_\theta$ that minimizes the divergence $\gamma_{\ell_2}(\mathbf{U}, \mathbf{V})$, and let $[\mathbf{U}', \mathbf{V}]$ be a joining that achieves the divergence. Let $\boldsymbol{\varepsilon}$ be a copy of the noise process that is independent of $[\mathbf{U}', \mathbf{V}]$, and define $\mathbf{Y} = \mathbf{V} + \boldsymbol{\varepsilon}$. One may readily show that the paired process $(\mathbf{U}', \mathbf{Y})$ is in fact a joining. Moreover, the choices above ensure that

$$\mathbb{E}\Big[\big|U_0' - Y_0\big|^2\Big] = \mathbb{E}\Big[\big|U_0' - V_0\big|^2\Big] + \mathbb{E}\Big[\varepsilon_0^2\Big] = \min_{\mathbf{U} \in \mathcal{Q}_\theta} \gamma_{\ell_2}(\mathbf{U}, \mathbf{V}) + \mathbb{E}\Big[\varepsilon_0^2\Big]. \tag{8.3}$$

Combining displays (8.2) and (8.3), we find that

$$\min_{\mathbf{U} \in \mathcal{Q}_\theta} \gamma_{\ell_2}(\mathbf{U}, \mathbf{Y}) \;=\; \min_{\mathbf{U} \in \mathcal{Q}_\theta} \gamma_{\ell_2}(\mathbf{U}, \mathbf{V}) + \mathbb{E}\big[\varepsilon_0^2\big].$$

Minimizing over θ, we find that $\Theta_{\ell_2}(\mathbf{Y}) = \operatorname{argmin}_\theta \min_{\mathbf{U} \in \mathcal{Q}_\theta} \gamma_{\ell_2}(\mathbf{U}, \mathbf{V})$, as was to be shown. □

Proof of Corollary 5.8. By Proposition 2.6, the hypothesis that the topological entropy $h_{top}(\mathcal{X}, T_\theta) = 0$, for each $\theta \in \Theta$, gives that $h(\mathcal{D}) = 0$. Then by Theorem 5.3 any sequence of least squares parameters converges almost surely to the set $\{\theta \in \Theta : \mathbf{V} \in \mathcal{Q}_\theta\}$. Now suppose $\mathbf{V} \in \mathcal{Q}_\theta$. Then there exists a measure $\mu_0 \in \mathcal{M}(\mathcal{X}, T_\theta)$ such that the process $\mathbf{U} = (T_\theta^k(X))_{k \geq 0}$ with $X \sim \mu_0$ has the same distribution as $\mathbf{V}$. Hence X has the same distribution as V_0, which is given by μ, and therefore $\mu = \mu_0$. Furthermore, $(X, T_\theta(X))$ must have the same distribution as (V_0, V_1), which implies that $T_\theta(x) = T_{\theta^*}(x)$ for μ almost every x. We have thus shown that $\{\theta \in \Theta : \mathbf{V} \in \mathcal{Q}_\theta\} \subset \{\theta : \mu(T_\theta = T_{\theta^*}) = 0\}$. The reverse inclusion is obvious. □

Acknowledgments. We wish to thank the anonymous reviewers for comments leading to improvements in the exposition of the paper, and one of the reviewers for suggesting Example 5.6. We would also like to thank Sayan Mukherjee for many helpful discussions.

SUPPLEMENTARY MATERIAL

Supplementary Material
(doi: COMPLETED BY THE TYPESETTER; .pdf). The supplementary document contains Appendices A - E, which consist of proofs of statements made in this work.

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] Roy L Adler, Alan G Konheim, and M Harry McAndrew. Topological entropy. *Transactions of the American Mathematical Society*, 114(2):309–319, 1965.

[3] Alekh Agarwal and John C Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013.

[4] Pierre Alquier and Olivier Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913, 2012.

[5] Miguel Angel Arcones and Bin Yu. Central limit theorems for empirical and U-processes of stationary mixing sequences. *Journal of Theoretical Probability*, 7(1):47–71, 1994.

[6] Luis Barreira and Yakov Pesin. *Nonuniform hyperbolicity: Dynamics of systems with nonzero Lyapunov exponents*, volume 115. Cambridge University Press, 2007.

[7] Dimitri P Bertsekas and Steven E Shreve. *Stochastic Optimal Control: The Discrete–Time Case*. Athena Scientific, 1996.

[8] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[9] Rufus Bowen and Jean-René Chazottes. *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, volume 470. Springer, 1975.

[10] Chris A Brackley, Oliver Ebenhöh, Celso Grebogi, Jürgen Kurths, Alessandro de Moura, M Carmen Romano, and Marco Thiel. Introduction to focus issue: dynamics in systems biology, 2010.

[11] Michael Brin and Garrett Stuck. *Introduction to dynamical systems*. Cambridge university press, 2002.

[12] David Damanik, Mark Embree, and Anton Gorodetski. Spectral properties of Schrödinger operators arising in the study of quasicrystals. In *Mathematics of Aperiodic Order*, pages 307–370. Springer, 2015.

[13] Christopher Daube and Kevin McGoff. On the dynamics of gene regulatory networks with sparsity constraints. Unpublished manuscript, 2018.

[14] Thierry de la Rue. An introduction to joinings in ergodic theory. *Discrete and Continuous Dynamical Systems*, 15(1):121–142, 2006.

[15] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*, 2017.

[16] Manfred Denker, Christian Grillenberger, and Karl Sigmund. *Ergodic theory on compact spaces*. Springer-Verlag, 1976.

[17] Harry Furstenberg. Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. *Theory of Computing Systems*, 1(1):1–49, 1967.

[18] Eli Glasner. *Ergodic theory via joinings*. American Mathematical Soc., 2003.

[19] 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.

[20] Hanyuan Hang and Ingo Steinwart. Fast learning from α-mixing observations. *Journal of Multivariate Analysis*, 127:184–199, 2014.

[21] Hanyuan Hang and Ingo Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *Ann. Statist.*, 2016.

[22] Hanyuan Hang, Ingo Steinwart, Yunlong Feng, and Johan AK Suykens. Kernel density estimation for dynamical systems. *arXiv:1607.03792*, 2016.

[23] Michael P Hassell, John H Lawton, and Robert M May. Patterns of dynamical behaviour in single-species populations. *The Journal of Animal Ecology*, pages 471–486, 1976.

[24] Franz Hofbauer. The topological entropy of the transformationx? ax (1- x). *Monatshefte für Mathematik*, 90(2):117–141, 1980.

[25] Anatole Katok. Fifty years of entropy in dynamics: 1958–2007. *J. Mod. Dyn*, 1(4):545–596, 2007.

[26] Anatole Katok and Boris Hasselblatt. *Introduction to the modern theory of dynamical systems*, volume 54. Cambridge university press, 1997.

[27] Leonid Aryeh Kontorovich. Statistical estimation with bounded memory. *Statistics and Computing*, 22(5):1155–1164, 2012.

[28] Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. In *Advances in neural information processing systems*, pages 541–549, 2015.

[29] Vitaly Kuznetsov and Mehryar Mohri. Time series prediction and online learning. In *29th Annual Conference on Learning Theory*, pages 1190–1213, 2016.

[30] Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

[31] Steven P. Lalley. Beneath the noise, chaos. *The Annals of Statistics*, 27(2):461–479, 1999.

[32] Steven P. Lalley and Andrew B. Nobel. Denoising deterministic time series. *Dyn. Partial Differ. Equ.*, 3(4):259–279, 2006.

[33] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[34] Benjamin Letham, Portia A Letham, Cynthia Rudin, and Edward P Browne. Prediction uncertainty and optimal experimental design for learning dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(6):063110, 2016.

[35] Nir Lev, Ron Peled, and Yuval Peres. Separating signal from noise. *Proceedings of the London Mathematical Society*, 110(4):883–931, 2015.

[36] 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.

[37] Douglas Lind and Brian Marcus. *An introduction to symbolic dynamics and coding*. Cambridge university press, 1995.

[38] Carlangelo Liverani. Central limit theorem for deterministic systems. In *International Conference on Dynamical Systems (Montevideo, 1995)*, volume 362, pages 56–75, 1996.

[39] John Mallet-Paret and Hal L Smith. The Poincaré-Bendixson theorem for monotone cyclic feedback systems. *Journal of Dynamics and Differential Equations*, 2(4):367–421, 1990.

[40] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

[41] Howard Masur and Serge Tabachnikov. Rational billiards and flat structures. In *Handbook of dynamical systems*, volume 1, pages 1015–1089. Elsevier, 2002.

[42] Robert M May. Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. *Science*, 186(4164):645–647, 1974.

[43] Kevin McGoff, Xin Guo, Anastasia Deckard, Christina Kelliher, Adam Leman, Lauren Francey, John Hogenesch, Steven Haase, and John Harer. Local Edge Machine: inference of dynamic models of gene regulation. *Genome Biology*, 2016.

[44] 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.

[45] Kevin McGoff, Sayan Mukherjee, and Natesh Pillai. Statistical inference for dynamical systems: A review. *Statist. Surv.*, 9:209–252, 2015.

[46] Kevin McGoff and Andrew Nobel. Optimal tracking for dynamical systems. arXiv:1601.05033, 2016.

[47] Kevin McGoff and Andrew Nobel. Supplement to "Empirical risk minimization and complexity of dynamical models". 2019.

[48] Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine learning*, 39(1):5–34, 2000.

[49] Konstantin Mischaikow, Marian Mrozek, and Andrzej Szymczak. Chaos in the Lorenz equations: A computer assisted proof part iii: Classical parameter values. *Journal of Differential Equations*, 169(1):17–56, 2001.

[50] Dharmendra S Modha and Elias Masry. Memory-universal prediction of stationary random processes. *IEEE transactions on information theory*, 44(1):117–133, 1998.

[51] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(Feb):789–814, 2010.

[52] Gusztáv Morvai and Benjamin Weiss. On classifying processes. *Bernoulli*, pages 523–532, 2005.

[53] Gusztáv Morvai and Benjamin Weiss. Prediction for discrete time series. *Probability theory and related fields*, 132(1):1–12, 2005.

[54] Andrew B Nobel. Hypothesis testing for families of ergodic processes. *Bernoulli*, pages 251–269, 2006.

[55] Donald Ornstein. Bernoulli shifts with the same entropy are isomorphic. *Advances in Mathematics*, 4(3):337–352, 1970.

[56] Donald S Ornstein. An application of ergodic theory to probability theory. *The Annals of Probability*, 1(1):43–58, 1973.

[57] Donald S Ornstein. *Ergodic theory, randomness, and dynamical systems*. Yale University Press, 1974.

[58] Donald S Ornstein and Benjamin Weiss. How sampling reveals a process. *The Annals of Probability*, pages 905–930, 1990.

[59] David Pollard and Peter Radchenko. Nonlinear least-squares estimation. *Journal of Multivariate Analysis*, 97(2):548–562, 2006.

[60] Martine Queffélec. *Substitution dynamical systems-spectral analysis*, volume 1294. Springer, 2010.

[61] Alexander Rakhlin and Karthik Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1704–1722, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

[62] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015.

[63] E Arthur Robinson. The dynamical theory of tilings and quasicrystallography. *London Mathematical Society Lecture Note Series*, pages 451–474, 1996.

[64] David Ruelle. *Thermodynamic formalism: the mathematical structure of equilibrium statistical mechanics*. Cambridge University Press, 2004.

[65] Cosma Shalizi and Aryeh Kontorovich. Predictive PAC learning and process decompositions. In *Advances in neural information processing systems*, pages 1619–1627, 2013.

[66] Dan Shechtman, Ilan Blech, Denis Gratias, and John W Cahn. Metallic phase with long-range orientational order and no translational symmetry. *Physical Review Letters*, 53(20):1951, 1984.

[67] Boris Solomyak. Spectrum of dynamical systems arising from Delone sets. *Quasicrystals and Discrete Geometry, ed. J. Patera, Fields Institute Monographs*, 10:265–275, 1998.

[68] 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.

[69] Stephen Tu, Ross Boczar, Andrew Packard, and Benjamin Recht. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

[70] Peter Turchin. *Complex population dynamics: a theoretical/empirical synthesis*, volume 35. Princeton University Press, 2013.

[71] Sara Van de Geer. Estimating a regression function. *The Annals of Statistics*, pages 907–924, 1990.

[72] A.W. Van der Vaart and Jon Wellner. *Weak convergence and empirical processes*. Springer, 2000.

[73] Divakar Viswanath, Xuan Liang, and Kirill Serkh. Metric entropy and the optimal prediction of chaotic signals. *SIAM Journal on Applied Dynamical Systems*, 12(2):1085–1113, 2013.

[74] Peter Walters. *An introduction to ergodic theory*, volume 79. Springer-Verlag, 1982.

[75] Kam Chung Wong, Zifan Li, and Ambuj Tewari. Regularized estimation in high dimensional time series under mixing conditions. *arXiv preprint arXiv:1602.04265*, 2016.

[76] Chien-Fu Wu. Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics*, pages 501–513, 1981.

[77] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

[78] Alexander Zimin and Christoph Lampert. Learning theory for conditional risk minimization. In *Artificial Intelligence and Statistics*, pages 213–222, 2017.

Kevin McGoff
UNC Charlotte
9201 University City Blvd.
Charlotte, NC 28223
E-mail: kmcgoff1@uncc.edu

Andrew B. Nobel
UNC Chapel Hill
308 Hanes Hall
Chapel Hill, NC 27599
E-mail: nobel@email.unc.edu