Modelling Language Evolution

Owen Biesel
Department of Mathematics
University of Washington
Seattle, WA 98195-4350
email: obiesel@u.washington.edu

August 14, 2007

Contents

1 Introduction

A child grows up to speak the language of its neighbors. Members of a society learn to agree on how to associate monetary values to goods. People learn to recognize arbitrary handwritten letters, and write their own so that others understand them. In each of these cases, a group of agents must reach a common way of associating objects (thoughts, goods, letters) to signals (sentences, values, symbols) when there is no clear leader whom everyone follows. In "Modeling Language Evolution" [2], Steve Smale attempts to model this convergence behavior.

We will begin by describing the space of objects X and the space of signals Y; languages will be continuous functions from X to Y. We will also assume that for our society of k members, or "agents," there exists a $k \times k$ matrix Γ representing the extent to which members of the society communicate with each other. The "state" $\mathbf{f}^{(t)}$ of the society at each time $t \in \mathbb{N}$ will be a list of the languages of each member. During each stage of the evolution, each of the agents receives a sample of "object-signal pairs" from the other agents. The agents then modify their languages so that they match all the signals received as closely as possible; in this way, each member of the society learns from every other member at each stage, and we call this process a learning dynamic. At first, we will consider the "ideal dynamic," in which the agents adjust their languages to the weighted average of all the languages used at time t. Then, we will move to a more sophisticated scenario in which the samples received are distributed in a way that only approximates this average. Denoting by Δ the set of states in which every agent has the same language, and using the L^2 distance function to compare two languages, we will culminate in a proof of a quantitative version of the following theorem:

Main Theorem. *Suppose that the communication matrix Γ satisfies a condition called weak irreducibility. Then for any $\varepsilon > 0$, $\delta > 0$, and stage t, there is a sample size m so that every agent receives at least m object-signal pairs at every stage, then the distance from $\mathbf{f}^{(t)}$ to Δ will be less than ε with confidence $1 - \delta$.*

In other words, by exchanging enough samples, we can be arbitrarily sure that the society will be arbitrarily close to sharing a common language.

2 Definitions Toward a Language-Learning Dynamic

2.1 Languages and Societies

Although the following is a far cry from a linguist's definition of a language, it will suffice to serve our purposes, which are in some senses both more specific and more general than the linguist's. Certain qualities of language, however, will be clearly recognizable.

2.1.1 Linguistic Settings

We begin with two Borel spaces: X a compact subset of $\mathbb{R}^n$ and Y a subset of $\mathbb{R}^l$. The set X denotes the space of *objects* or *meanings,* which are to be represented by elements of Y, the space of *signals.* (It is important to keep in mind which space is which, since the applications are rather widespread: In modeling human language, it may be appropriate to let Y be the set of utterable words. On the other hand, in a handwriting recognition program, a script converts words to written signals, and so in that case X would be the space of words.) A particular language, then, would be a function from X to Y, which associates to each object a point in the space of signals. We now have enough to define a *Linguistic Setting*:

Definition 2.1. A linguistic setting $\mathcal{L}$ is a triple $\langle (X, \pi), Y, \mathcal{F} \rangle$ where X is a compact subset of $\mathbb{R}^n$ and Y is a subset of $\mathbb{R}^l$ (where $\mathbb{R}^l$ has the full inner product structure). The function π is a Borel probability measure on X, and is interpreted as describing the relative frequency of objects occurring in *object-signal pairs* (elements of $Z = X \times Y$). Continuous functions from X to Y are called *language-like functions,* and $\mathcal{F}$ is a compact set of language-like functions (with respect to the uniform norm) which we call *languages.*

It is important to note at this time that this treatment of languages is robust enough to handle languages in which an object is not consistently represented with the same signal. Specifically, if the signal used to represent x is randomly chosen from a finite set $\{y_1, y_2, \ldots y_m\} \subset Y$ with probabilities $P_1, P_2, \ldots P_m$ respectively, then it may be appropriate to associate that language to a language-like function f in which $f(x) = \sum_1^m P_j y_j$. For this reason, the space of signals Y is often taken to be convex. From this point forward, we will assume the convexity of both Y and $\mathcal{F}$.

2.1.2 Linguistic Societies

With a set of k *agents,* we may form a linguistic society through their interaction:

Definition 2.2. A linguistic society $\mathcal{P}$ is a triple $\langle \{1, \ldots, k\}, \mathcal{L}, \Gamma \rangle$ where $\mathcal{L}$ is a linguistic setting and Γ is a $k \times k$ matrix with real, nonnegative entries γ_{ij}, measuring the linguistic influence of agent j on agent i. Γ is called the communication matrix of the society.

We will also assume that for $i = 1, \ldots, k$, $\sum_{j=1}^{k} \gamma_{ij} > 0$. Then we can always produce from Γ a *normalized communication matrix* Λ with entries

$$\lambda_{ij} = \frac{\gamma_{ij}}{\sum_{l=1}^{k} \gamma_{il}}, \tag{1}$$

so that $\sum_{j=1}^{k} \lambda_{ij} = 1$ for each i.

A *state* of a linguistic society is a characterization of each agent's language, i.e. a k-tuple of elements of $\mathcal{F}$. Those states in which every member of the society has the same language belong to the *diagonal*, the subset of $\mathcal{F}^k$ in which every agent's language is identical: $\Delta_{\mathcal{F}} = \{(f, \ldots, f) \in \mathcal{F}^k\}$.

2.2 Learning Dynamics, Part I

With a linguistic society $\mathcal{P}$ in hand, we can now define the evolution of that system by using a learning dynamic. Consider time indexed in discrete stages by the natural numbers, so that the state of the society at time t is $\mathbf{f}^{(t)} = (f_1^{(t)}, \ldots, f_k^{(t)})$. Then the dynamic of the system is a method for producing $\mathbf{f}^{(t+1)}$ from $\mathbf{f}^{(t)}$. In a *learning dynamic*, this process is executed in several stages.

2.2.1 Exchange Functions and Linguistic Fitness

The first stage of the dynamic is to define at each time t and for each agent i a language-like function $F_i^{(t)}$ that describes the overall direction in which the society is pulling the agent's language. These language-like functions, which I call the *exchange functions*, are defined as follows:

$$F_i^{(t)} = \frac{\sum_{j=1}^{k} \gamma_{ij} f_j^{(t)}}{\sum_{j=1}^{k} \gamma_{ij}} = \sum_{j=1}^{k} \lambda_{ij} f_j^{(t)}, \tag{2}$$

i.e. $\mathbf{F}^{(t)} = \Lambda \mathbf{f}^{(t)}$. These exchange functions will also belong to $\mathcal{F}$, since they are convex combinations of languages in $\mathcal{F}$. If a linguistic society is in the state $\mathbf{f} = (f_1, f_2, \ldots, f_k)$, we can also define the *linguistic fitness* Φ_i of a language f for agent i by

$$\Phi_i(f) = -\int_X \left(\sum_{j=1}^{k} \gamma_{ij} \|f(x) - f_j(x)\|^2 \right) d\pi(x), \tag{3}$$

where we are using the norm on Y induced by the Euclidean inner product on $\mathbb{R}^l$. We can think of linguistic fitness as the extent to which agent i would be able to communicate with all the members of the society (including i itself) if i were to use the language f instead of f_i. Then we have the following result:

Proposition 2.3. *For all $f \in \mathcal{F}$, $\Phi_i(f) \leq \Phi_i(F_i)$, where F_i is the exchange function defined by (2). Equality is true only when $f = F_i$ π-a.e.*

Proof. Let $f \in \mathcal{F} \subset C(X, \mathbb{R}^l)$. Then for each $x \in X$, we have

$$\begin{aligned}
\sum_{j=1}^{k} \gamma_{ij} \|f(x) - f_j(x)\|^2 &= \sum_j \gamma_{ij} \|(f_j(x) - F_i(x)) + (F_i(x) - f(x))\|^2 \\
&= \begin{array}{l} \sum_j \gamma_{ij} \left[\|f_j(x) - F_i(x)\|^2 + \|F_i(x) - f(x)\|^2 \right. \\ \left. +2\langle f_j(x) - F_i(x), F_i(x) - f(x)\rangle \right] \end{array} \\
&= \begin{array}{l} \sum_j \gamma_{ij} \|f_j(x) - F_i(x)\|^2 + \left(\sum_j \gamma_{ij}\right) \|F_i(x) - f(x)\|^2 \\ +2\left\langle \sum_j \gamma_{ij}(f_j(x) - F_i(x)), F_i(x) - f(x) \right\rangle \end{array}
\end{aligned}$$

But we have defined F_i so that

$$\sum_j \gamma_{ij}(f_j - F_i) = \sum_j \gamma_{ij} f_j - \left(\sum_j \gamma_{ij}\right) F_i = 0$$

Hence we have

$$\begin{aligned}
\sum_j \gamma_{ij} \|f_j(x) - f(x)\|^2 &= \sum_j \gamma_{ij} \|f_j(x) - F_i(x)\|^2 + \left(\sum_j \gamma_{ij}\right) \|F_i(x) - f(x)\|^2 \\
&\geq \sum_j \gamma_{ij} \|f_j(x) - F_i(x)\|^2,
\end{aligned}$$

where equality holds iff $f(x) = F_i(x)$ since $\sum_j \gamma_{ij} > 0$. Hence

$$\int_X \sum_{j=1}^{k} \gamma_{ij} \|f_j(x) - f(x)\|^2 d\pi(x) \geq \int_X \sum_{j=1}^{k} \gamma_{ij} \|f_j(x) - F_i(x)\|^2 d\pi(x),$$

i.e. $\Phi_i(f) \leq \Phi_i(F_i)$ with equality iff $f = F_i$ π-a.e. This justifies the previous statement that F_i represents the overall direction in which the society is "pulling" agent i's language. □

We will also find that in the "ideal" dynamic, given by $\mathbf{f}^{(t+1)} = \mathbf{F}^{(t)}$ for each stage t, $\mathbf{f}^{(t)}$ will always converge in L^2 to a point in the diagonal with a predictable rate, but we defer the proof of this statement until §3.3.

The next stage of the learning dynamic involves producing sampling measures from the exchange functions, but before we can explore that step we need to develop some machinery to handle these new measures.

2.3 A Digression: Measures on Product Spaces

At each stage in the dynamic, some probability measure ρ on the product of the measurable spaces $(X, \mathcal{M}) \times (Y, \mathcal{N})$ produces the sample object-signal pairs used to define the languages in the next stage. It is important to see what structure this measure induces on the spaces X and Y.

2.3.1 Induced Measures on the Factor Spaces

Naturally, ρ induces a probability measure ρ_X on $(X, \mathcal{M})$ defined by:

$$\text{For any } A \in \mathcal{M},\ \rho_X(A) = \rho(A \times Y). \tag{4}$$

For each fixed $x \in X$, ρ also produces a probability measure on $\{x\} \times Y$, but more subtly. We would like to interpret this measure as the conditional probability measure that the Y-component is in $B \in \mathcal{N}$ given that the X-component is x, i.e. as

$$\rho(B|x) = \frac{\rho(\{x\} \times B)}{\rho(\{x\} \times Y)}. \qquad \text{(False)}$$

However, this last expression is undefined for nearly all values of x, since $\rho_X(\{x\}) \neq 0$ for at most countably many values of x. We can, however define the measure as follows:

Definition 2.4. Let $x \in X$ and $B \in \mathcal{N}$. Then the function $\rho_B : \mathcal{M} \to \mathbb{R} : A \mapsto \rho(A \times B)$ is a measure that is absolutely continuous with respect to ρ_X, and we may define $\rho(B|x)$ as the Radon-Nikodym derivative of ρ_B with respect to ρ_X, evaluated at x.[1] Then $\rho(\cdot|x)$ is a probability measure on Y.

The assertion that ρ_B is absolutely continuous with respect to ρ_X follows immediately from the fact that these are both positive measures and $\rho_B(A) = \rho(A \times B) \leq \rho(A \times Y) = \rho_X(A)$. We also have, since $B = Y \Rightarrow \rho_B = \rho_X$,

$$\rho(Y|x) = \left.\frac{d\rho_B}{d\rho_X}\right|_x = 1.$$

Lastly, suppose that B is the disjoint union of a sequence $\langle B_n \rangle \subset \mathcal{N}$. Then letting $f_n = d\rho_{B_n}/d\rho_X$, $f = \sum_1^\infty f_n$, we have for any $A \in \mathcal{M}$:

$$\begin{aligned}
\rho_B(A) &= \rho(A \times B) \\
&= \sum_1^\infty \rho(A \times B_n) \\
&= \sum_1^\infty \rho_{B_n}(A) \\
&= \sum_1^\infty \int_A f_n \; d\rho_X \\
&= \int_A f \; d\rho_X,
\end{aligned}$$

where the last equality follows from the Dominated Convergence Theorem since $f_n \leq 1 \in L^1(X)$. This implies that

$$\begin{aligned}
\rho(B|x) &= \left.\frac{d\rho_B}{d\rho_X}\right|_x \\
&= f(x) = \sum_1^\infty f_n(x) \\
&= \sum_1^\infty \left.\frac{d\rho_{B_n}}{d\rho_X}\right|_x \\
&= \sum_1^\infty \rho(B_n|x)
\end{aligned}$$

[1] Unfortunately, the Radon-Nikodym derivative is unique only up to modification on any ρ_X-null set. For this definition, then, we must choose some representative function and consistently use it, which fortunately does not matter in any of the uses that follow.

demonstrating countable additivity of $\rho(B|x)$. Hence $\rho(\cdot|x)$ is a probability measure on Y for each x. Furthermore, in the case in which X and Y are Borel subspaces of $\mathbb{R}^n$, we find from [3] that for ρ_X-a.e. x:

$$\rho(B|x) = \lim_{r\to 0} \frac{\rho(A_r \times B)}{\rho(A_r \times Y)} \qquad \text{(True)}$$

for every family $\{A_r\}$ that shrinks nicely to x. This makes it clear why $\rho(B|x)$ can be interpreted as a conditional probability of B given x.

2.3.2 Fubini's Theorem for Measures on Product Spaces

We also have the following analogue to Fubini's Theorem:

Theorem 2.5. *Let ρ be a probability measure on $(X, \mathcal{M}) \times (Y, \mathcal{N})$. If $f \in L^1(\rho)$, then the function $x \mapsto \int_Y f(x,y)\ d\rho(y|x)$ is well-defined a.e. and is in $L^1(\rho_X)$, and*

$$\int_{X\times Y} f\ d\rho = \int_X \left[\int_Y f(x,y)\ d\rho(y|x) \right]\ d\rho_X(x) \tag{5}$$

Proof. First, suppose that f is the characteristic function of a measurable set E. Then let $\mathcal{C}$ be the collection of all E for which the theorem is true with $f = 1_E$, that is, for which $\rho(E) = \int_X \rho(E_x|x)\ d\rho_X(x)$. First suppose $E = A \times B$. Then we must show that $\rho(A\times B) = \int_A \rho(B|x)\ d\rho_X(x)$, but this follows obviously from the definition $\rho(B|\cdot) = d\rho_B/d\rho_X$. Hence products of measurable sets are in $\mathcal{C}$, and so are finite unions of them by additivity, so if we can show $\mathcal{C}$ is a monotone class then the theorem will be true for all characteristic functions by the Monotone Class Lemma. If E is the union of an increasing sequence $\langle E_n \rangle$ in $\mathcal{C}$, then continuity from below of $\rho(\cdot, x)$ shows that the sequence of functions $f_n(x) = \rho((E_n)_x|x)$ converges pointwise to $f(x) = \rho(E_x|x)$, so by the Monotone Convergence Theorem:

$$\begin{aligned}
\int_X \rho(E_x|x)\ d\rho_X(x) &= \lim_{n\to\infty} \int_X \rho((E_n)_x|x)\ d\rho_X(x) \\
&= \lim_{n\to\infty} \rho(E_n) \\
&= \rho(E)
\end{aligned}$$

So then $E \in \mathcal{C}$ as well. Similarly, if E is the intersection of a decreasing sequence $\langle E_n \rangle$ in $\mathcal{C}$, then the function $x \mapsto \rho((E_n)_x|x)$ is measurable and in L^1 since its integral is at most 1, so this time the Dominated Convergence

Theorem demonstrates the result. Hence $\mathcal{C}$ is a monotone class, so the theorem holds for all measurable $E \subset X \times Y$.

Now, all that remains is to show that the theorem holds for arbitrary functions. We know now that it holds for characteristic functions, so it holds for all simple functions by linearity. Then if f is a nonnegative measurable function on $X \times Y$, then there is an increasing sequence of simple functions converging pointwise to f, so the theorem follows by the Monotone Convergence Theorem. Then for an arbitrary function f in $L^1(\rho)$, we may simply apply the previous to the positive and negative parts of f. This proof closely mimics that of Fubini's Theorem in [3], and can easily be extended to apply to σ-finite measures on $X \times Y$. □

Although it is easiest to imagine cases where $\rho(B|x)$ is independent of x, i.e. ρ decomposes as a product measure, this is the least interesting possibility. Another extreme is when each $\rho(\cdot|x)$ is a Dirac measure for some point $y = f(x)$; in general the function $u_\rho(x) = \int_Y y \; d\rho(y|x)$ is called the *regression function* of ρ, and indicates the expectation value of y given a supplied value of x.

2.4 Learning Dynamics, Part II

Armed with the machinery of measures on $Z = X \times Y$, we can now constrain next stage of the learning dynamic: the sampling measures $\rho_i^{(t)}$.

2.4.1 Sampling Measures

The sampling measures are the first objects not to be necessarily defined, or even stochastically produced, by the linguistic setting and initial state of the society. We will assume that at each step t and for each agent i there is a probability measure $\rho_i^{(t)}$ on $(X, \mathcal{M}) \times (Y, \mathcal{N})$ that satisfies the following three conditions:

1. The measure $\rho_{iX}^{(t)}$, induced on X by $\rho_i^{(t)}$, is equal to π, the object probability measure, for every agent i and time t.

2. The regression function $x \mapsto \int_Y y \; d\rho_i^{(t)}(y|x)$ is equal to $F_i^{(t)}$.

3. There exists a constant M, independent of i and t, so that for every $f \in \mathcal{F}$,
$$\|f(x) - y\| \leq M \text{ for } \rho_i^{(t)}\text{-a.e. } (x, y) \in Z. \tag{6}$$

These sampling measures are used to sample packets of communication in the next stage.

2.4.2 Producing Samples

Now that we have measures on $Z = X \times Y$, we can use them to produce m samples at each time t (here m is allowed implicitly to depend on t). For each agent i, let $S_i^{(t)}$ be an m-tuple of independent Z-valued random variables:

$$S_i^{(t)} = \left\{ \left(x_{i1}^{(t)}, y_{i1}^{(t)} \right), \ldots, \left(x_{im}^{(t)}, y_{im}^{(t)} \right) \right\}$$

where every $z_{ij}^{(t)}$ is distributed according to $\rho_i^{(t)}$, i.e.

$$P\left(z_{ij}^{(t)} \in E \right) = \rho_i^{(t)}(E)$$

One consequence of this is that the X-coordinates of the z_i's are distributed according to π, that is

$$P\left(z_{ij}^{(t)} \in A \times Y \right) = \rho_i^{(t)}(A \times Y) = \pi(A)$$

by Condition 1. We also find, given that the X-coordinate of $z_{ij}^{(t)}$ is x, that the expected value of the Y-coordinate is:

$$E\left(y_{ij}^{(t)} | x_{ij}^{(t)} = x \right) = \int_Y y \; d\rho_i^{(t)}(y|x) = F_i^{(t)}(x)$$

Returning to our interpretation of Z as the space of object-signal pairs, we see that $S_i^{(t)}$ first samples m objects from X according to π, and then to each object x pairs a signal so that the expected value of the signal is $F_i^{(t)}(x)$.

One possible way to define $\rho_i^{(t)}$ is to write that $\rho_i^{(t)}(\cdot|x)$ is a sum of Dirac measures at $f_j^{(t)}(x)$ with weights λ_{ij}, since then

$$\int_Y y \; d\rho_i^{(t)}(y|x) = \sum_{j=1}^{k} \lambda_{ij} f_j^{(t)}(x) = F_i^{(t)}(x).$$

Then $\left(x_{i1}^{(t)}, y_{i1}^{(t)} \right)$ can be interpreted as first choosing an object randomly from (X, μ) (as always), and then choosing a member of the society at random (weighted probabilistically by λ_{ij}) to pair it with a signal. This can be likened to agent i receiving noiseless signals from each of the members of the society. Of course, in general $\rho_i^{(t)}$ will not be so simple, to allow for various types of noise (random Y-valued variables with zero expectation) to be added.

2.4.3 Minimizing the Empirical Error

The final step in the learning dynamic is to produce $f_i^{(t+1)}$ from the sample $S_i^{(t)}$. This last step is the easiest to state: simply choose $f_i^{(t+1)}$ from $\mathcal{F}$ so that the *empirical error*

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{j=1}^{m} \left\| f(x_{ij}^{(t)}) - y_{ij}^{(t)} \right\|^2 \text{ is minimized.} \tag{7}$$

(This minimum is guaranteed to be attained in $\mathcal{F}$ so long as we assume its compactness. On the other hand, minimizing functions are not necessarily unique, in which case we can choose any of them without affecting the results.) In this way, the map from $\mathcal{F}^k$ to $\mathcal{F}^k$ is completed, and so the state of the linguistic society progresses through an arbitrary state t. We are now farther along in stating the main theorem:

Main Theorem. *If Λ satisfies a condition called "weak irreducibility," then for every $\delta > 0$ there exist constants $\alpha_* < 1$ and $\mathbf{C} > 0$, as well as a sample size $m(t)$, such that with any initial state $\mathbf{f}^{(0)}$, the following holds with probability at least $1-\delta$:*

$$d(\mathbf{f}^{(t)}, \Delta_{\mathcal{F}}) \leq \mathbf{C}\alpha_*^t d(\mathbf{f}^{(0)}, \Delta_{\mathcal{F}}),$$

where d is the L^2 metric on $\mathcal{F}$.

To prove this, we need to define weak irreducibility and explore its implications, as well as consider a few results from learning theory.

3 The Normalized Communication Matrix Λ and its Action...

Recall that any $k \times k$ normalized communication matrix Λ satisfies the following two properties:

1. $\lambda_{ij} \geq 0$ for each $1 \leq i, j \leq k$.

2. $\sum_{j=1}^{k} \lambda_{ij} = 1$ for each $1 \leq i \leq k$.

In general, a matrix satisfying those two properties is called a *stochastic matrix.* Note that the second condition implies that the vector $\mathbf{e} = (1, 1, \ldots, 1)$ is an eigenvector of Λ with eigenvalue 1. Along similar lines, we have the following theorem about stochastic matrices, due to Perron and Frobenius[5]:

Lemma 3.1. *Let Λ be a stochastic matrix. Then the eigenvalues of Λ are all no more than 1 in modulus.*

Proof. Let α be an arbitrary eigenvalue of Λ, and $\mathbf{x}$ be a corresponding eigenvector of Λ, and without loss of generality normalize $\mathbf{x}$ so that for some index i_0, $|x_j| \leq |x_{i_0}| = 1$ for all j. Then we have the following:

$$\begin{aligned} |\alpha| &= |\alpha x_{i_0}| \\ &= \left| \sum_{j=1}^{k} \lambda_{i_0 j} x_j \right| \\ &\leq \sum_{j=1}^{k} \lambda_{i_0 j} |x_j| \\ &\leq \sum_{j=1}^{k} \lambda_{i_0 j} = 1. \end{aligned}$$

□

If 1 is a simple eigenvalue, and every other eigenvalue is strictly less than 1 in modulus, then Λ is called *weakly irreducible*, which amounts to a kind of connectedness property for the linguistic society. For example, if Λ decomposes as

$$\Lambda = \begin{bmatrix} \Lambda_I & 0 \\ 0 & \Lambda_J \end{bmatrix}$$

with Λ_I an $n \times n$ square matrix and Λ_J a $(k-n) \times (k-n)$ matrix, then the two vectors $\mathbf{e_1} = (1, \ldots, 1, 0, \ldots, 0)$ (with n ones and $k - n$ zeroes) and $\mathbf{e_2} = \mathbf{e} - \mathbf{e_1}$ will be two linearly independent eigenvectors of Λ, both with eigenvalue 1.

3.1 ...On $\mathbb{R}^k$

Since we can describe the exchange functions at a stage t by $\mathbf{F}^{(t)} = \Lambda \mathbf{f}^{(t)}$, it is useful to explore the properties of Λ as a linear map from $\mathcal{F} \to \mathcal{F}$.

However, it is easiest to begin with the map $\Lambda : \mathbb{R}^k \to \mathbb{R}^k$. First, let us consider the following fact:

Lemma 3.2. *For any real square matrix A and $\varepsilon > 0$, there exists a real nonsingular matrix C such that $D = C^{-1}AC$ is of the block form*

$$\begin{bmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_n \end{bmatrix}$$

where each D_i is an $k_i \times k_i$ square matrix of either the form

$$\begin{bmatrix} \lambda & \varepsilon & 0 & 0 & \cdots & 0 \\ 0 & \lambda & \varepsilon & 0 & \cdots & 0 \\ 0 & 0 & \lambda & \varepsilon & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \lambda \end{bmatrix} \quad or \quad \begin{bmatrix} \begin{bmatrix} \mu & -\nu \\ \nu & \mu \end{bmatrix} & \begin{bmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix} & \cdots & 0 \\ 0 & \begin{bmatrix} \mu & -\nu \\ \nu & \mu \end{bmatrix} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \begin{bmatrix} \mu & -\nu \\ \nu & \mu \end{bmatrix} \end{bmatrix}$$

where either λ is a real eigenvalue of A or $\mu \pm i\nu$ are conjugate complex eigenvalues of A.

Proof. With $\varepsilon = 1$ this is just the statement of the fact that every real square matrix has a real Jordan Canonical Form. The general version with arbitrary nonzero ε follows by conjugating with respect to

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \varepsilon & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varepsilon^{k_i} \end{bmatrix}$$

in the first case, or in the second:

$$\begin{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & 0 & \cdots & 0 \\ 0 & \begin{bmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \begin{bmatrix} \varepsilon^{k_i/2} & 0 \\ 0 & \varepsilon^{k_i/2} \end{bmatrix} \end{bmatrix}$$

□

Let Δ_k be the subspace of $\mathbb{R}^k$ consisting of multiples of $\mathbf{e}$, that is, with all components equal. Then we can now state a few properties Λ has:

Proposition 3.3. *For any $k \times k$ weakly irreducible stochastic matrix Λ, there exists a subspace $W \subset \mathbb{R}^k$, a norm $\|\cdot\|_\Lambda$ (induced by an inner product $\langle\cdot,\cdot\rangle_\Lambda$), and a constant $\alpha_* < 1$ such that:*

1. *$\mathbb{R}^k$ decomposes into the direct sum $\Delta_k \oplus W$.*
2. *If $\mathbf{v} \in \Delta_k$ then $\Lambda\mathbf{v} = \mathbf{v}$.*
3. *If $\mathbf{w} \in W$ then $\Lambda\mathbf{w} \in W$.*
4. *If $\mathbf{v} \in \Delta_k$ and $\mathbf{w} \in W$, then $\langle\mathbf{v},\mathbf{w}\rangle_\Lambda = 0$.*
5. *If $\mathbf{w} \in W$ then $\|\Lambda\mathbf{w}\|_\Lambda \leq \alpha_*\|\mathbf{w}\|_\Lambda$.*

Proof.

1. Let $\varepsilon > 0$ be small enough that $|\lambda| + \varepsilon < 1$ for every eigenvalue $\lambda \neq 1$. (This ε exists by our assumption of Λ's weak irreducibility.) Then let C be a nonsingular matrix (with column vectors $\mathbf{C}_1$ through $\mathbf{C}_k$), conjugation by which puts Λ into the form shown in Lemma 3.2, and without loss of generality let $\mathbf{C}_1$ be the eigenvector $\mathbf{e}$. (Hence the first block in $D = C^{-1}\Lambda C$ is a 1×1 square with the single entry 1.) Then $\mathbb{R}^k$ decomposes naturally as the direct sum of Δ_k, which is spanned by $\mathbf{C}_1 = \mathbf{e}$, and W, which we define as the span of $\mathbf{C}_2$ through $\mathbf{C}_k$. For any $\mathbf{x} \in \mathbb{R}^k$, define $\widetilde{\mathbf{x}} = C^{-1}\mathbf{x}$ so that $\mathbf{x} = \sum_{i=1}^k \widetilde{x}_i\mathbf{C}_i$.

2. If $\mathbf{v} \in \Delta_k$, then $\Lambda\mathbf{v} = \mathbf{v}$ simply because $\mathbf{e}$ is an eigenvector of Λ with eigenvalue 1, and Δ_k is the corresponding eigenspace.

3. Since W is the space spanned by $\mathbf{C}_2$ through $\mathbf{C}_k$, we can rewrite it as
$$W = \{\mathbf{x} \in \mathbb{R}^k : \widetilde{x}_1 = 0\}. \tag{8}$$
Then we find that:
$$\begin{aligned}
\mathbf{w} \in W \quad &\Rightarrow \quad \widetilde{w}_1 = 0 \\
&\Rightarrow \quad (D\widetilde{\mathbf{w}})_1 = 0 \text{ due to our choice of the block-diagonal structure of } D \\
&\Rightarrow \quad (C^{-1}\Lambda C C^{-1}\mathbf{w})_1 = 0 \\
&\Rightarrow \quad (C^{-1}\Lambda\mathbf{w})_1 = 0 \\
&\Rightarrow \quad \widetilde{(\Lambda\mathbf{w})}_1 = 0 \\
&\Rightarrow \quad \Lambda\mathbf{w} \in W
\end{aligned}$$
Hence Λ maps W into itself.

4. Now define an inner product on $\mathbb{R}^k$ by

$$\langle \mathbf{x}, \mathbf{y} \rangle_\Lambda = \mathbf{x}^T (C^{-1})^T (C^{-1}) \mathbf{y}. \tag{9}$$

   Then in the basis consisting of column vectors of C, we find that the vector $\mathbf{x}$ has components $\widetilde{x}_i = (C^{-1}\mathbf{x})_i$, so that we may write

$$\langle \mathbf{x}, \mathbf{y} \rangle_\Lambda = \widetilde{\mathbf{x}}^T \widetilde{\mathbf{y}} = \langle \widetilde{\mathbf{x}}, \widetilde{\mathbf{y}} \rangle, \tag{10}$$

   where the inner product on the right is the Euclidean inner product. Define the natural norm induced by this inner product to be $\|\mathbf{x}\|_\Lambda = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_\Lambda} = \|\widetilde{\mathbf{x}}\|$. Note that we can express Δ_k the following way: $\mathbf{x} \in \Delta_k$ iff $\widetilde{x}_j = 0$ for $j \neq 1$. Together with the alternate expression W from Equation 8, this makes it apparent that if $\mathbf{v} \in \Delta_k$ and $\mathbf{w} \in W$, then $\langle \mathbf{v}, \mathbf{w} \rangle_\Lambda = \langle \widetilde{\mathbf{v}}, \widetilde{\mathbf{w}} \rangle = 0$.

5. Lastly, we will show that Λ is a contraction on W. Let α_* be ε plus the magnitude of the largest eigenvalue of Λ not equal to 1; by assumption, $\alpha_* < 1$. Then we must show that $\|\Lambda \mathbf{w}\|_\Lambda \leq \alpha_* \|\mathbf{w}\|_\Lambda$. The norm on the right is equal to $\sqrt{\widetilde{\mathbf{w}}^T \widetilde{\mathbf{w}}} = \|\widetilde{\mathbf{w}}\|$ with the Euclidean norm; the norm on the left is the square root of:

$$\begin{aligned}
\|\Lambda \mathbf{w}\|_\Lambda^2 &= \mathbf{w}^T \Lambda^T (C^{-1})^T C^{-1} \Lambda \mathbf{w} \text{ by 9} \\
&= \mathbf{w}^T \left[(C^T)^{-1} C^T \right] \Lambda^T (C^{-1})^T C^{-1} \Lambda \left[C C^{-1} \right] \mathbf{w} \\
&= \left[\mathbf{w}^T (C^{-1})^T \right] \left[C^T \Lambda^T (C^{-1})^T \right] \left[C^{-1} \Lambda C \right] \left[C^{-1} \mathbf{w} \right] \\
&= \widetilde{\mathbf{w}}^T D^T D \widetilde{\mathbf{w}} \text{ by definition of } \widetilde{\mathbf{w}} \text{ and } D \\
&= \|D \widetilde{\mathbf{w}}\|^2 \text{ with the Euclidean norm.}
\end{aligned}$$

   Hence $\mathbf{w} \in W \Rightarrow \|\Lambda \mathbf{w}\|_\Lambda \leq \alpha_* \|\mathbf{w}\|_\Lambda$ if and only if $x_1 = 0 \Rightarrow \|D\mathbf{x}\| \leq \alpha_* \|\mathbf{x}\|$. We will now prove this second statement. Recall from Lemma 3.2 that D has the following block structure:

$$D = \begin{bmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_n \end{bmatrix}$$

   where D_i, $i = 1, \ldots, n$ is a square $k_i \times k_i$ of either of the two forms listed above. (Remember that we chose $D_1 = [1]$.) Now given a k-vector $\mathbf{x}$, decompose it into n separate k_i-vectors $\mathbf{x}_i$, formed so that the concatenation of all these vectors from $i = 1$ to n is just $\mathbf{x}$. Then for any $\mathbf{x}$ with $\mathbf{x}_1 = x_1 = 0$, we have (in block format):

$$\begin{aligned}
\|D\mathbf{x}\|^2 &= \begin{bmatrix} 0 & \mathbf{x}_2^T & \cdots & \mathbf{x}_n^T \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & D_2^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_n^T \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_n \end{bmatrix} \begin{bmatrix} 0 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} \\
&= \sum_{i=2}^{n} \mathbf{x}_i^T D_i^T D_i \mathbf{x}_i \\
&= \sum_{i=2}^{n} \|D_i \mathbf{x}_i\|^2.
\end{aligned}$$

Hence, if we can show that for each i from 2 to n, $\|D_i\mathbf{x}_i\| \leq \alpha_* \|\mathbf{x}_i\|$, then we will have $\|D\mathbf{x}\|^2 \leq \alpha_*^2 \sum_{i=2}^n \|\mathbf{x}_i\|^2 = \alpha_*^2 \|\mathbf{x}\|^2$ as desired. Consider the first case, in which

$$D_i = \begin{bmatrix} \lambda & \varepsilon & 0 & \cdots & 0 \\ 0 & \lambda & \varepsilon & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{bmatrix}$$

Then letting $\widehat{\mathbf{x}}_i$ be the n_i-vector with $(\widehat{\mathbf{x}}_i)_j = (\mathbf{x}_i)_{j+1}$ for $1 \leq j < n_i$ and $(\widehat{\mathbf{x}}_i)_{n_i} = 0$, we have

$$\begin{aligned}
\|D_i \mathbf{x}_i\| &= \|\lambda \mathbf{x}_i + \varepsilon \widehat{\mathbf{x}}_i\| \\
&\leq |\lambda| \|\mathbf{x}_i\| + \varepsilon \|\widehat{\mathbf{x}}_i\| \\
&\leq |\lambda| \|\mathbf{x}_i\| + \varepsilon \|\mathbf{x}_i\| \leq \alpha_* \|\mathbf{x}_i\|.
\end{aligned}$$

Similarly, suppose

$$D_i = \begin{bmatrix} \begin{bmatrix} \mu & -\nu \\ \nu & \mu \end{bmatrix} & \begin{bmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix} & \cdots & 0 \\ 0 & \begin{bmatrix} \mu & -\nu \\ \nu & \mu \end{bmatrix} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \begin{bmatrix} \mu & -\nu \\ \nu & \mu \end{bmatrix} \end{bmatrix},$$

with $\lambda = \mu + i\nu$ a complex eigenvalue of Λ. Then noticing that if we identify real column 2-vectors $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ with complex numbers $x_1 + ix_2$

(denoting this correspondence by $\simeq$), we have

$$\begin{aligned}\begin{bmatrix} \mu & -\nu \\ \nu & \mu \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} x_1\mu - x_2\nu \\ x_1\nu + x_2\mu \end{bmatrix} \\ &\simeq (\mu + i\nu)(x_1 + ix_2)\end{aligned}$$

Now if we convert any k_i-vector $\mathbf{x}$ into a complex $\frac{k_i}{2}$-vector $\mathbf{z}$ by taking pairs of adjacent elements, we find that

$$\begin{aligned} D_i\mathbf{x} &= \begin{bmatrix} \begin{bmatrix} \mu & -\nu \\ \nu & \mu \end{bmatrix} & \begin{bmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix} & \cdots & 0 \\ 0 & \begin{bmatrix} \mu & -\nu \\ \nu & \mu \end{bmatrix} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \begin{bmatrix} \mu & -\nu \\ \nu & \mu \end{bmatrix} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ \begin{bmatrix} x_3 \\ x_4 \end{bmatrix} \\ \vdots \\ \begin{bmatrix} x_{k_i-1} \\ x_{k_i} \end{bmatrix} \end{bmatrix} \\ &\simeq \begin{bmatrix} \mu + i\nu & \varepsilon & \cdots & 0 \\ 0 & \mu + i\nu & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu + i\nu \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{k_i/2} \end{bmatrix} \end{aligned}$$

Since $\simeq$ preserves the norms of vectors, the result follows in the same way as the previous case, but with $\lambda = \mu + i\nu$. Hence $\|D_i\mathbf{x}_i\| \leq \alpha_*\|\mathbf{x}_i\|$ for all $i = 2, \ldots, n$, and so $\|D\mathbf{x}\| \leq \alpha_*\|\mathbf{x}\|$ for any $\mathbf{x}$ with $x_1 = 0$, and so $\|\Lambda\mathbf{w}\|_\Lambda \leq \alpha_*\|\mathbf{w}\|_\Lambda$ for any $w \in W$.

□

It is worth noting that since all norms on $\mathbb{R}^k$ are equivalent, there exist positive constants C_Λ and C'_Λ such that for every $\mathbf{v} \in \mathbb{R}^k$,

$$C'_\Lambda\|\mathbf{v}\| \leq \|\mathbf{v}\|_\Lambda \leq C_\Lambda\|\mathbf{v}\|.$$

3.2 ...On $(\mathbb{R}^l)^k$

Now that we have considered the action of Λ on $\mathbb{R}^k$, we can extend this to an action on $(\mathbb{R}^l)^k$ for an arbitrary $l \in \mathbb{N}$. Elements of $(\mathbb{R}^l)^k$ are $k \times l$ matrices:

$$\vec{\mathbf{v}} = \begin{bmatrix} \vec{v}_1 \\ \vec{v}_2 \\ \vdots \\ \vec{v}_k \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_l \end{bmatrix}$$

where each $\vec{v}_i$, $i = 1, \ldots, k$ is a row l-vector, and each $\mathbf{v}_j$ is a column k-vector for $j = 1, \ldots, l$. The individual components of $\vec{\mathbf{v}}$ are v_{ij} for $1 \leq i \leq k$ and $1 \leq j \leq l$. It is worthwhile to note that if A is any $k \times k$ matrix, it acts on $k \times l$ matrices $\mathbf{v} \in (\mathbb{R}^l)^k$ by ordinary matrix multiplication:

$$\begin{aligned} A\vec{\mathbf{v}} &= \begin{bmatrix} \sum_{i=1}^k A_{1i}\vec{v}_i \\ \sum_{i=1}^k A_{2i}\vec{v}_i \\ \vdots \\ \sum_{i=1}^k A_{ki}\vec{v}_i \end{bmatrix} = \sum_{i=1}^{k} \begin{bmatrix} A_{1i}\vec{v}_i \\ A_{2i}\vec{v}_i \\ \vdots \\ A_{ki}\vec{v}_i \end{bmatrix} = \sum_{i=1}^{k} \mathbf{A}_i \otimes \vec{v}_i & (11) \\ &= \begin{bmatrix} A\mathbf{v}_1 & A\mathbf{v}_2 & \cdots & A\mathbf{v}_l \end{bmatrix} & (12) \end{aligned}$$

where $\mathbf{A}_i$ is the ith column k-vector of A. Also, define

$$\begin{aligned} \Delta_{kl} &= \{\vec{\mathbf{v}} : \vec{v}_i = \vec{v}_j \text{ for all } 1 \leq i, j \leq k\} & (13) \\ &= \{\mathbf{e} \otimes \vec{v} : \vec{v} \in \mathbb{R}^l\} & (14) \\ &= \{\vec{\mathbf{v}} : \mathbf{v}_j \in \Delta_k \text{ for all } 1 \leq j \leq l\} & (15) \end{aligned}$$

Proposition 3.4. *For any $k \times k$ weakly irreducible stochastic matrix Λ acting on $(\mathbb{R}^l)^k$, there exists a subspace $W \subset (\mathbb{R}^l)^k$, a norm $\|\cdot\|_\Lambda$ (induced by an inner product $\langle \cdot, \cdot \rangle_\Lambda$), and a constant $\alpha_* < 1$ such that:*

1. *$(\mathbb{R}^l)^k$ decomposes into the direct sum $\Delta_{kl} \oplus W$.*

2. *If $\vec{\mathbf{v}} \in \Delta_{kl}$ then $\Lambda\vec{\mathbf{v}} = \vec{\mathbf{v}}$.*

3. *If $\vec{\mathbf{w}} \in W$ then $\Lambda\vec{\mathbf{w}} \in W$.*

4. *If $\vec{\mathbf{v}} \in \Delta_{kl}$ and $\vec{\mathbf{w}} \in W$, then $\langle \vec{\mathbf{v}}, \vec{\mathbf{w}} \rangle_\Lambda = 0$.*

5. *If $\vec{\mathbf{w}} \in W$ then $\|\Lambda\vec{\mathbf{w}}\|_\Lambda \leq \alpha_* \|\vec{\mathbf{w}}\|_\Lambda$.*

Proof. 1. Let α_* and C be defined as in the proof of Proposition 3.3. Then because C is nonsingular, we may define $\widetilde{\vec{\mathbf{x}}}$ for any $\vec{\mathbf{x}} \in (\mathbb{R}^l)^k$ by $\widetilde{\vec{\mathbf{x}}} = (C^{-1})\vec{\mathbf{x}}$. This relationship between $\vec{\mathbf{x}}$ and $\widetilde{\vec{\mathbf{x}}}$ can also be expressed as

$$\begin{aligned} \vec{\mathbf{x}} &= C\widetilde{\vec{\mathbf{x}}} \\ &= \sum_{i=1}^{k} \mathbf{C}_i \otimes \widetilde{\vec{x}}_i \text{ by (12),} \end{aligned}$$

and the invertibility of C guarantees that this expansion is unique. (These statements are in analogy with those in the proof of Proposition 3.3, where $\mathbf{x} = \sum_{i=1}^{k} \widetilde{x}_i \mathbf{C}_i$.) Then

$$\begin{aligned} \mathbf{x} \in \Delta_{kl} \quad &\iff \quad \mathbf{x} = \mathbf{e} \otimes \vec{v} \text{ for some } \vec{v} \in \mathbb{R}^l \text{ by (14)} \\ &\iff \quad \mathbf{x} = \mathbf{C}_1 \otimes \widetilde{\vec{x}}_1 \text{ (since } \mathbf{C}_1 = \mathbf{e}) \\ &\iff \quad \widetilde{\vec{x}}_i = 0 \text{ for all } 2 \leq i \leq k. \end{aligned}$$

Then if we define W as $\{\vec{\mathbf{x}} : \widetilde{\vec{x}}_1 = 0\}$, we find that $(\mathbb{R}^l)^k$ decomposes naturally as $\Delta_{kl} \oplus W$.

2. If $\vec{\mathbf{v}} \in \Delta_{kl}$, then the jth column $(1 \leq j \leq l)$ of $\Lambda\vec{\mathbf{v}}$ is $\Lambda\mathbf{v}_j$ by (12). But $\mathbf{v}_j \in \Delta_k$ for each j by (15), so $(\Lambda\vec{\mathbf{v}})_j = \Lambda\mathbf{v}_j = \mathbf{v}_j$, and hence $\Lambda\vec{\mathbf{v}} = \vec{\mathbf{v}}$.

3. Suppose $\vec{\mathbf{w}} \in W$. Then by definition, $\widetilde{\vec{w}}_1 = 0$. Again, because the first block in D was chosen to be the 1×1 square [1], this implies that the first row of $D\widetilde{\vec{\mathbf{w}}}$ is zero as well. Since $D = C^{-1}\Lambda C$, this is the same as saying the first row of $C^{-1}\Lambda C\widetilde{\vec{\mathbf{w}}} = C^{-1}\Lambda\vec{\mathbf{w}}$ is zero, i.e. $\widetilde{(\Lambda\vec{\mathbf{w}})}_1 = 0$. Hence $\Lambda\vec{\mathbf{w}} \in W$.

4. Define an inner product on $(\mathbb{R}^l)^k$ by

$$\langle \vec{\mathbf{x}}, \vec{\mathbf{y}} \rangle_\Lambda = \sum_{j=1}^{l} \langle \mathbf{x}_j, \mathbf{y}_j \rangle_\Lambda$$

where the inner product on the right-hand side is that from Proposition 3.3. This inner product has the alternate expression

$$\begin{aligned} \langle \vec{\mathbf{x}}, \vec{\mathbf{y}} \rangle_\Lambda &= \sum_{j=1}^{l} \langle \mathbf{x}_j, \mathbf{y}_j \rangle_\Lambda \\ &= \sum_{j=1}^{l} \widetilde{\mathbf{x}}_j^T \widetilde{\mathbf{y}}_j \qquad \text{(N.B.}^2) \\ &= \sum_{j=1}^{l} \sum_{i=1}^{k} (\widetilde{\vec{\mathbf{x}}})_{ij} (\widetilde{\vec{\mathbf{y}}})_{ij} \\ &= \sum_{i=1}^{k} (\widetilde{\vec{x}}_i)(\widetilde{\vec{y}}_i)^T \end{aligned}$$

Now, if $\vec{\mathbf{v}} \in \Delta_{kl}$ and $\vec{\mathbf{w}} \in W$, this last sum vanishes since for any value of i, either $\widetilde{\vec{v}_i}$ or $\widetilde{\vec{w}_i}$ is zero, so $\langle \vec{\mathbf{v}}, \vec{\mathbf{w}} \rangle_\Lambda = 0$.

5. Suppose $\vec{\mathbf{w}} \in W$. This means that $\widetilde{\vec{w}_1} = 0$, i.e. that the first entry in $\widetilde{\mathbf{w}}_j$ is zero for any j. This is the same as saying that $\mathbf{w}_j$ is in the subspace $W \subset \mathbb{R}^k$ from Proposition 3.3 for every j. Then we have:

$$\begin{aligned} \|\Lambda \vec{\mathbf{w}}\|_\Lambda^2 &= \sum_{j=1}^{l} \|\Lambda \mathbf{w}_j\|_\Lambda^2 \\ &\leq \alpha_*^2 \sum_{j=1}^{l} \|\mathbf{w}_j\|_\Lambda^2 \\ &= \alpha_*^2 \|\vec{\mathbf{w}}\|_\Lambda^2 \end{aligned}$$

as desired, where the norm inside the sums is the one defined in Proposition 3.3.

□

3.3 ...On $C(X, (\mathbb{R}^l)^k)$

We have one more extension of our proposition, to Λ acting on k-tuples in $C(X, (\mathbb{R}^l)^k)$, where X has a Borel probability measure π. We will now use the following notation: elements $\vec{\mathbf{f}}$ of $C(X, (\mathbb{R}^l)^k)$ are continuous functions from X to $(\mathbb{R}^l)^k$, that is,

$$\vec{\mathbf{f}} = \begin{bmatrix} \vec{f}_1 \\ \vec{f}_2 \\ \vdots \\ \vec{f}_k \end{bmatrix}$$

with each $\vec{f}_i \in C(X, \mathbb{R}^l)$. (It is obvious in this notation that for any $k \times k$ matrix A, $A(\vec{\mathbf{f}}(x)) = (A\vec{\mathbf{f}})(x)$. The parentheses in these situations will now no longer be observed.) We also define

$$\begin{aligned} \Delta_{C(X,\mathbb{R}^l)} &= \{\vec{\mathbf{f}} \in C(X, \mathbb{R}^l) : \vec{f}_i = \vec{f}_j \text{ for all } 1 \leq i, j \leq k\} \\ &= \{\vec{\mathbf{f}} \in C(X, \mathbb{R}^l) : \vec{\mathbf{f}}(x) \in \Delta_{kl} \text{ for all } x \in X\}. \end{aligned}$$

[2] The astute reader will notice that mixing the notations in Propositions 3.3 and 3.4 presents an ambiguity: the equality to the preceding line suggests that $\widetilde{\mathbf{x}}_j$ is formed from $\mathbf{x}_j$ as in Proposition 3.3, while the equality following this line interprets $\widetilde{\mathbf{x}}_j$ as the jth column of the $k \times l$ matrix $\widetilde{\vec{\mathbf{x}}}$ formed directly from $\vec{\mathbf{x}}$. However, since in both interpretations $\widetilde{\mathbf{x}}_j = C^{-1}\mathbf{x}_j$, this is a legal ambiguity, and prevents our already cluttered notation from becoming more so.

Proposition 3.5. *For any $k \times k$ weakly irreducible stochastic matrix Λ, there exists a subspace $W \subset C(X, (\mathbb{R}^l)^k)$, a norm $\|\cdot\|_\Lambda$ (induced by an inner product $\langle \cdot, \cdot \rangle_\Lambda$), and a constant $\alpha_* < 1$ such that:*

1. *$C(X, (\mathbb{R}^l)^k)$ decomposes into the direct sum $\Delta_{C(X,\mathbb{R}^l)} \oplus W$.*
2. *If $\vec{\mathbf{f}} \in \Delta_{C(X,\mathbb{R}^l)}$ then $\Lambda\vec{\mathbf{f}} = \vec{\mathbf{f}}$.*
3. *If $\vec{\mathbf{g}} \in W$ then $\Lambda\vec{\mathbf{g}} \in W$.*
4. *If $\vec{\mathbf{f}} \in \Delta_{C(X,\mathbb{R}^l)}$ and $\vec{\mathbf{g}} \in W$, then $\langle \vec{\mathbf{f}}, \vec{\mathbf{g}} \rangle_\Lambda = 0$.*
5. *If $\vec{\mathbf{g}} \in W$ then $\|\Lambda\vec{\mathbf{g}}\|_\Lambda \leq \alpha_* \|\vec{\mathbf{g}}\|_\Lambda$.*

Proof.

1. Let C and α_* be as before, but this time, let W be the space of all $\vec{\mathbf{f}}$ such that $\vec{\mathbf{f}}(x)$ belongs to the W from Proposition 3.4 for every $x \in X$. Then it is once again apparent that $\vec{\mathbf{f}} \in W$ iff the first entry in $\widetilde{\vec{\mathbf{f}}} = C^{-1}\vec{\mathbf{f}}$ is identically zero. On the other hand, $\vec{\mathbf{f}} \in \Delta_{C(X,\mathbb{R}^l)}$ exactly when every entry, except possibly the first, in $\widetilde{\vec{\mathbf{f}}}$ is zero, demonstrating the desired decomposition.
2. Let $\vec{\mathbf{f}} \in \Delta_{C(X,\mathbb{R}^l)}$. Then $\vec{\mathbf{f}}(x) \in \Delta_{kl}$ for every $x \in X$, which implies that $\Lambda\vec{\mathbf{f}}(x) = \vec{\mathbf{f}}(x)$, and hence $\Lambda\vec{\mathbf{f}} = \vec{\mathbf{f}}$.
3. Let $\vec{\mathbf{g}} \in W$. Then for every $x \in X$, $\vec{\mathbf{g}}(x)$ is in the W from Proposition 3.4, which implies $\Lambda\vec{\mathbf{g}}(x)$ is in that W, and hence $\Lambda\vec{\mathbf{g}} \in W$.
4. Define an inner product $\langle \vec{\mathbf{f}}, \vec{\mathbf{g}} \rangle_\Lambda$ by $\int_X \langle \vec{\mathbf{f}}(x), \vec{\mathbf{g}}(x) \rangle_\Lambda d\pi(x)$, where the inner product in the integrand is that from Proposition 3.4. Then if $\vec{\mathbf{f}} \in \Delta_{C(X,\mathbb{R}^l)}$, $\vec{\mathbf{g}} \in W$, we have
$$\begin{aligned} \langle \vec{\mathbf{f}}, \vec{\mathbf{g}} \rangle_\Lambda &= \int_X \langle \vec{\mathbf{f}}(x), \vec{\mathbf{g}}(x) \rangle_\Lambda d\pi(x) \\ &= \int_X 0 \, d\pi(x) = 0. \end{aligned}$$
5. Suppose $\vec{\mathbf{g}} \in W$. Then we have
$$\begin{aligned} \|\Lambda\vec{\mathbf{g}}\|_\Lambda^2 &= \int_X \|\Lambda\vec{\mathbf{g}}(x)\|_\Lambda^2 d\pi(x) \\ &\leq \int_X \alpha_*^2 \|\vec{\mathbf{g}}(x)\|_\Lambda^2 d\pi(x) \\ &= \alpha_*^2 \|\vec{\mathbf{g}}\|_\Lambda^2. \end{aligned}$$

□

3.4 ...On $\mathcal{F}^k$

Now we can finally restrict our attention to Λ acting on $\mathcal{F}^k$, where $\mathcal{F}$ is a convex subset of $C(X, \mathbb{R}^l)$. Since $\lambda_{ij} \geq 0$ and $\sum_j \lambda_{ij} = 1$ for each i, the rows of $\Lambda\vec{\mathbf{f}}$ are convex combinations of the rows $\vec{f_i}$, so if $\vec{f_i} \in \mathcal{F}$ for each i, so is $(\Lambda\vec{\mathbf{f}})_i$, and hence Λ can be considered a linear map $\Lambda : \mathcal{F}^k \to \mathcal{F}^k$. Henceforth, we will no longer use arrows above function names, and denote elements of $\mathcal{F}$ by f or g, and elements of $\mathcal{F}^k$ by $\mathbf{f}$ or $\mathbf{g}$.

Using the notation that any $\mathbf{f} \in \mathcal{F}^k$ can be written uniquely as

$$\mathbf{f} = \mathbf{f}_\Delta + \mathbf{f}_W,$$

where $\mathbf{f}_\Delta \in \Delta_{C(X,\mathbb{R}^l)}$ and $\mathbf{f}_W \in W$, then $\Lambda\mathbf{f} = \Lambda(\mathbf{f}_\Delta + \mathbf{f}_W) = \mathbf{f}_\Delta + \Lambda\mathbf{f}_W$, and $\Lambda\mathbf{f}_W \in W$, so $(\Lambda\mathbf{f})_\Delta = \mathbf{f}_\Delta$. Then denoting by $d_\Lambda(\mathbf{f}, \mathbf{g})$ the Λ-norm of the difference between $\mathbf{f}$ and $\mathbf{g}$, and $d_\Lambda(\mathbf{f}, S) = \inf_{\mathbf{g}\in S}\{d_\Lambda(\mathbf{f}, \mathbf{g})\}$, we have the following:

Lemma 3.6. $d_\Lambda(\mathbf{f}, \Delta_{C(X,\mathbb{R}^l)}) = d_\Lambda(\mathbf{f}, \mathbf{f}_\Delta) = \|\mathbf{f}_W\|_\Lambda$.

Proof. This follows easily by writing, for any $\mathbf{g} \in \Delta_{C(X,\mathbb{R}^l)}$:

$$\begin{aligned} d_\Lambda(\mathbf{f}, \mathbf{g})^2 &= \|\mathbf{f} - \mathbf{g}\|_\Lambda^2 \\ &= \|\mathbf{f}_W + \mathbf{f}_\Delta - \mathbf{g}\|_\Lambda^2 \\ &= \|\mathbf{f}_W\|_\Lambda^2 + \|\mathbf{f}_\Delta - \mathbf{g}\|_\Lambda^2 \end{aligned}$$

(by the orthogonality of $\Delta_{C(X,\mathbb{R}^l)}$ and W with respect to $\langle\cdot,\cdot\rangle_\Lambda$)

$$\geq \|\mathbf{f}_W\|_\Lambda^2,$$

with equality in the last statement exactly when $\mathbf{g} = \mathbf{f}_\Delta$. □

Corollary 3.7. *The following hold for any* $\mathbf{f} \in \mathcal{F}$*:*

1. $d_\Lambda(\Lambda\mathbf{f}, \mathbf{f}_\Delta) \leq \alpha_* d_\Lambda(\mathbf{f}, \mathbf{f}_\Delta)$
2. $d_\Lambda(\Lambda\mathbf{f}, \Delta_{C(X,\mathbb{R}^l)}) \leq \alpha_* d_\Lambda(\mathbf{f}, \Delta_{C(X,\mathbb{R}^l)})$
3. $d_\Lambda(\Lambda^n\mathbf{f}, \mathbf{f}_\Delta) \to 0$ *as* $n \to \infty$
4. $\Lambda^n\mathbf{f} \to \mathbf{f}_\Delta \in \Delta_{C(X,\mathbb{R}^l)}$ *as* $n \to \infty$ *with respect to the* L^2 *norm*

Hence in the ideal dynamic $\mathbf{f}^{(t+1)} = \Lambda\mathbf{f}^{(t)}$, we find that $\mathbf{f}^{(t)} \to \mathbf{f}_\Delta^{(0)}$ as $t \to \infty$, where the distance (in the Λ-norm) to the diagonal shrinks by at least a factor of α_* at each stage.

4 Learning Theory

4.1 Some Notions from Learning Theory

We need to develop a few concepts from learning theory before we can finish the proof of the main theorem. One such idea is that of the *error* of a function $f \in \mathcal{F}$ with respect to a probability measure ρ on $Z = X \times Y$ ($Y \subset \mathbb{R}^l$), defined as

$$\mathcal{E}(f) = \int_Z \|f(x) - y\|^2 \; d\rho(x,y).$$

This quantity can be thought of as the deviation of f from the pair (x,y), where we have sampled (x,y) over all of Z, with respect to the measure ρ. If we define

$$u_\rho(x) = \mathbf{E}(y|x) = \int_Y y \; d\rho(y|x),$$

then we have the following alternate characterization of the error:

Lemma 4.1. $\mathcal{E}(f) = \mathcal{E}(u_\rho) + \int_X \|f(x) - u_\rho(x)\|^2 \; d\rho_X(x).$

Proof. We have

$$\begin{aligned}
\mathcal{E}(f) &= \int_Z \|f(x) - y\|^2 \; d\rho(x,y) \\
&= \int_Z \|f(x) - u_\rho(x) + u_\rho(x) - y\|^2 \; d\rho(x,y) \\
&= \int_X \int_Y \|f(x) - u_\rho(x)\|^2 \; d\rho(y|x) \; d\rho_X(x) \\
&+ \int_X \int_Y \|u_\rho(x) - y\|^2 \; d\rho(y|x) \; d\rho_X(x) \\
&+ 2\int_X \int_Y \langle f(x) - u_\rho(x), u_\rho(x) - y\rangle_Y \; d\rho(y|x) \; d\rho_X(x).
\end{aligned}$$

Consider these three integrals one at a time. The first integrand does not actually depend on Y, so its integral is equal to $\int_X \|f(x) - u_\rho(x)\|^2 \; d\rho_X(x)$. The second integral is $\mathcal{E}(u_\rho)$ by definition. The integral over Y in the third term can be shifted into the second slot of the inner product in the integrand, but since $\int_Y (u_\rho(x) - y) \; d\rho(y|x)$ is defined to vanish, the entire third term does as well. □

In the event that we have a sample $\mathbf{z} = ((x_1, y_1), \ldots, (x_m, y_m))$ in Z^m, we can also define the *empirical error* of a function f by

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^{m} \|f(x_i) - y_i\|^2,$$

which measures how well the values assumed by f match the sample. Since $\mathcal{F}$ is compact, there exists a (not necessarily unique) minimizing function $f_{\mathbf{z}}$ such that $\mathcal{E}_{\mathbf{z}}(f) \geq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})$ for all $f \in \mathcal{F}$. Sometimes we will want to refer to only the error due to f being different from u_ρ, and so we will define

$$\begin{aligned} \mathcal{E}_\rho(f) &= \mathcal{E}(f) - \mathcal{E}(u_\rho) \\ \mathcal{E}_{\mathbf{z},\rho}(f) &= \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(u_\rho) \end{aligned}$$

Note that Lemma 4.1 guarantees that $\mathcal{E}_\rho(f) = \int_X \|f(x) - u_\rho(x)\|^2 \, d\rho_X(x) \geq 0$ for all $f \in \mathcal{F}$, but $\mathcal{E}_{\mathbf{z},\rho}(f)$ can in general assume negative values. In this section, we will also use the probabilist's notation $\text{Prob}\{\mathbf{z} : \mathbf{z} \in A\}$ to indicate the probability of choosing a sample from $A \subset Z^m$, and the expectation value $\mathbf{E}(f)$ to indicate the integral of f over the appropriate probability space. Then in this notation, for example, $\mathbf{E}(\|f(x) - y\|^2) = \mathcal{E}(f)$, $\mathbf{E}(\|f(x) - u_\rho(x)\|^2) = \mathcal{E}_\rho(f)$, and $\mathbf{E}(\mathcal{E}_{\mathbf{z}}(f)) = \mathcal{E}(f)$.

4.2 Pointwise Probability Estimates

For the remainder of this section, let ρ be one of the sampling measures on Z produced in the learning dynamic defined in §2.4.1, so that $\rho_X = \pi$, $u_\rho \in \mathcal{F}$, and $\|f(x) - y\| \leq M$ for every $f \in \mathcal{F}$ and ρ-a.e. $(x, y) \in Z$. We can also use the following result, due to Bernstein:

Theorem 4.2 (Bernstein's Inequality). *Let ξ be a random variable on a probability space Z with mean $\mathbf{E}(\xi) = \mu$ and variance σ^2. If $|\xi(z) - \mu| \leq M$ for a.e $z \in Z$, then for all $\varepsilon > 0$,*

$$\text{Prob}\left\{\mathbf{z} \in Z^m : \mu - \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) \geq \lambda\right\} \leq \exp\left(\frac{-m\lambda^2}{2\left(\sigma^2 + \frac{1}{3}M\varepsilon\right)}\right)$$

Proof. This inequality, like many similar inequalities, follows by assuming $\mu = 0$, writing the probability as $\text{Prob}\left\{\exp[\frac{1}{m}\sum t\xi(z_i)] \geq \exp(t\varepsilon)\right\}$ for arbitrary $t > 0$, applying Chebyshev's Inequality, and then carefully estimating the bound produced while choosing a clever value for t. For a complete proof, see [4]. □

We can apply this result to get the following proposition:

Proposition 4.3. *Let $f \in \mathcal{F}$, and let M satisfy* (6)*. Then for all $\varepsilon > 0$, $0 < \alpha \leq 1$,*

$$\operatorname{Prob}\left\{\mathbf{z} \in Z^m : \frac{\mathcal{E}_\rho(f) - \mathcal{E}_{\mathbf{z},\rho}(f)}{\mathcal{E}_\rho(f) + \varepsilon} \geq \alpha\right\} \leq \exp\left(\frac{-\alpha^2 m\varepsilon}{8M^2}\right)$$

Proof. For any $f \in \mathcal{F}$ let $\ell(f)$ be a random variable on Z defined by

$$\ell(f)(x,y) = \|f(x) - y\|^2 - \|u_\rho(x) - y\|^2.$$

Then we have the following identities:

$$\begin{aligned}
\mathbf{E}(\ell(f)) &= \mathbf{E}(\|f(x) - y\|^2) - \mathbf{E}(\|u_\rho(x) - y\|^2) = \mathcal{E}(f) - \mathcal{E}(u_\rho) = \mathcal{E}_\rho(f) \\
\frac{1}{m}\sum_{i=1}^{m} \ell(f)(z_i) &= \mathbf{E}_{\mathbf{z}}(\|f(x) - y\|^2) - \mathbf{E}_{\mathbf{z}}(\|u_\rho(x) - y\|^2) = \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(u_\rho) = \mathcal{E}_{\mathbf{z},\rho}(f)
\end{aligned}$$

Letting $\mu = \mathcal{E}_\rho(f)$, we also find that $|\ell(f)(z)| \leq M^2$ for a.e. z, and hence $|\ell(f)(z) - \mu| \leq M^2 + \mu$.[3] Also, denote by σ^2 the variance of $\ell(f)$. Now we can apply Bernstein's Inequality to $\ell(f)$ and we obtain:

$$\begin{aligned}
\operatorname{Prob}\left\{\frac{\mathcal{E}_\rho(f) - \mathcal{E}_{\mathbf{z},\rho}(f)}{\mu + \varepsilon} \geq \alpha\right\} &= \operatorname{Prob}\left\{\mathbf{E}[\ell(f)] - \frac{1}{m}\sum_{j=1}^{m} \ell(f)(z_i) \geq \alpha(\mu + \varepsilon)\right\} \\
&\leq \exp\left[\frac{-m\alpha^2(\mu + \varepsilon)^2}{2\left(\sigma^2 + \frac{1}{3}(M^2 + \mu)\alpha(\mu + \varepsilon)\right)}\right]
\end{aligned}$$

We need to find an estimate for σ^2. To do this, we first expand note that for any two functions f and g in $\mathcal{F}$,

$$\begin{aligned}
\left(\|f(x) - y\|^2 - \|g(x) - y\|^2\right)^2 &= \left(\|f(x)\|^2 - 2\langle f(x), y\rangle + \|y\|^2 - \|g(x)\|^2 + 2\langle g(x), y\rangle - \|y\|^2\right)^2 \\
&= \left(\|f(x)\|^2 - \|g(x)\|^2 - 2\langle f(x), y\rangle + 2\langle g(x), y\rangle\right)^2 \\
&= (\langle f(x) - g(x), f(x) + g(x)\rangle - \langle f(x) - g(x), 2y\rangle)^2 \\
&= \langle f(x) - g(x), f(x) + g(x) - 2y\rangle^2 \\
&\leq \|f(x) - g(x)\|^2 \|(f(x) - y) + (g(x) - y)\|^2,
\end{aligned}$$

[3] In [2], the latter estimate is not made and Bernstein's Inequality is incorrectly applied. Fortunately, the proposition is still true, and the argument has been adjusted to account for the extra term μ.

where the last inequality comes from the Cauchy-Schwarz inequality for the Euclidean inner product on Y. Then

$$\begin{aligned}\left(\|f(x)-y\|^2-\|g(x)-y\|^2\right)^2 &\leq \|f(x)-g(x)\|^2\left(\|f(x)-y\|+\|g(x)-y\|\right)^2\\ &\leq \|f(x)-g(x)\|^2(M+M)^2\\ &= 4M^2\|f(x)-g(x)\|^2 \qquad (16)\end{aligned}$$

So by writing $\sigma^2=\mathbf{E}[\ell(f)^2]-\mathbf{E}[\ell(f)]^2$ we have

$$\begin{aligned}\sigma^2 &= \mathbf{E}[\ell(f)^2]-\mu^2\\ &= \mathbf{E}[(\|f(x)-y\|^2-\|u_\rho(x)-y\|^2)^2]-\mu^2\\ &\leq 4M^2\mathbf{E}[\|f(x)-u_\rho(x)\|^2]-\mu^2 \text{ by (16)}\\ &= 4M^2\mu-\mu \qquad (17)\end{aligned}$$

So $\sigma^2\leq 4M^2\mu-\mu^2$, and in particular $\mu\leq 4M^2$ in order that σ^2 be nonnegative. To finish the proof, we need to show that

$$\frac{\varepsilon}{8M^2}\leq\frac{(\mu+\varepsilon)^2}{2\left(\sigma^2+\frac{1}{3}(M^2+\mu)\alpha(\mu+\varepsilon)\right)},$$

which is true iff

$$\frac{\varepsilon}{4M^2}\left(\sigma^2+\frac{1}{3}(M^2+\mu)\alpha(\mu+\varepsilon)\right)\leq(\mu+\varepsilon)^2,$$

i.e.

$$\frac{\varepsilon\sigma^2}{4M^2}+\left(\varepsilon\mu+\varepsilon^2\right)\left(\frac{\alpha}{12}\left(1+\frac{\mu}{M^2}\right)\right)\leq(\mu+\varepsilon)^2$$

The first term on the left is at most $\frac{\varepsilon}{4M^2}(4M^2\mu-\mu^2)\leq\varepsilon\mu$ by (17), and the quantity $\frac{\alpha}{12}\left(1+\frac{\mu}{M^2}\right)$ is at most $\frac{5}{12}<1$ since $\mu\leq 4M^2$ and $\alpha\leq 1$. Then we must only show that $2\varepsilon\mu+\varepsilon^2\leq(\mu+\varepsilon)^2$, but this is obvious since the difference is $\mu^2\geq 0$. Hence

$$\text{Prob}\left\{\mathbf{z}\in Z^m:\frac{\mathcal{E}_\rho(f)-\mathcal{E}_{\mathbf{z},\rho}(f)}{\mathcal{E}_\rho(f)+\varepsilon}\geq\alpha\right\}\leq\exp\left(\frac{-m\alpha^2\varepsilon}{8M^2}\right)\text{ as desired. } \square$$

4.3 Uniform Probability Estimates

In Proposition 4.3, we find that for each $f\in\mathcal{F}$, a condition on $\mathbf{z}$ holds with a particular confidence. We would like a similar result that handles all $f\in\mathcal{F}$ simultaneously. To do this, we first require a lemma:

Lemma 4.4. *Let* $0 < \alpha < 1$, $\varepsilon > 0$, $f \in \mathcal{F}$ *and* $\mathbf{z} \in Z^m$ *so that* $\frac{\mathcal{E}_\rho(f)-\mathcal{E}_{\mathbf{z},\rho}(f)}{\mathcal{E}_\rho(f)+\varepsilon} < \alpha$. *Then for any* $g \in \mathcal{F}$ *with* $\|f(x)-g(x)\| \leq \frac{\alpha\varepsilon}{4M}$, *we have the similar estimate* $\frac{\mathcal{E}_\rho(g)-\mathcal{E}_{\mathbf{z},\rho}(g)}{\mathcal{E}_\rho(g)+\varepsilon} < 3\alpha$.

Proof. Rewrite the expression involving g as follows:

$$\begin{aligned}
\frac{\mathcal{E}_\rho(g)-\mathcal{E}_{\mathbf{z},\rho}(g)}{\mathcal{E}_\rho(g)+\varepsilon} &= \frac{\mathcal{E}(g)-\mathcal{E}(u_\rho)-\mathcal{E}_{\mathbf{z}}(g)+\mathcal{E}_{\mathbf{z}}(u_\rho)}{\mathcal{E}_\rho(g)+\varepsilon} \\
&= \frac{\mathcal{E}(f)-\mathcal{E}(u_\rho)-\mathcal{E}_{\mathbf{z}}(f)+\mathcal{E}_{\mathbf{z}}(u_\rho)}{\mathcal{E}_\rho(g)+\varepsilon} \\
&\quad + \frac{\mathcal{E}(g)-\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(g)+\mathcal{E}_{\mathbf{z}}(f)}{\mathcal{E}_\rho(g)+\varepsilon} \\
&= \frac{\mathcal{E}_\rho(f)-\mathcal{E}_{\mathbf{z},\rho}(f)}{\mathcal{E}_\rho(g)+\varepsilon} + \frac{\mathcal{E}(g)-\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(g)+\mathcal{E}_{\mathbf{z}}(f)}{\mathcal{E}_\rho(g)+\varepsilon}
\end{aligned}$$

Consider the numerator of the second term. First, we know that:

$$\begin{aligned}
\left| \|f(x)-y\|^2 - \|g(x)-y\|^2 \right| &\leq 2M\|f(x)-g(x)\| \text{ by (16)} \\
&\leq 2M\frac{\alpha\varepsilon}{4M} = \frac{\alpha\varepsilon}{2}
\end{aligned}$$

Hence $|\mathcal{E}(f)-\mathcal{E}(g)| \leq \frac{\alpha\varepsilon}{2}$ and $|\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(g)| \leq \frac{\alpha\varepsilon}{2}$, so magnitude of the numerator of the second term is

$$|\mathcal{E}(f)-\mathcal{E}(g)-\mathcal{E}_{\mathbf{z}}(f)+\mathcal{E}_{\mathbf{z}}(g)| \leq \alpha\varepsilon.$$

Since the denominator of the second term is $\mathcal{E}_\rho(g)+\varepsilon \geq \varepsilon$, the whole second term is at most α.

Now consider the first term. We just showed that

$$\mathcal{E}_\rho(f)-\mathcal{E}_\rho(g) = \mathcal{E}(f)-\mathcal{E}(g) \leq \frac{\alpha\varepsilon}{2} < \varepsilon \leq \varepsilon + \mathcal{E}_\rho(g),$$

i.e. $\mathcal{E}_\rho(f)+\varepsilon < 2\mathcal{E}_\rho(g)+2\varepsilon$, so $\frac{\mathcal{E}_\rho(f)+\varepsilon}{\mathcal{E}_\rho(g)+\varepsilon} < 2$. On the other hand, by assumption $\frac{\mathcal{E}_\rho(f)-\mathcal{E}_{\mathbf{z},\rho}(f)}{\mathcal{E}_\rho(f)+\varepsilon} < \alpha$, so

$$\frac{\mathcal{E}_\rho(f)-\mathcal{E}_{\mathbf{z},\rho}(f)}{\mathcal{E}_\rho(g)+\varepsilon} = \left(\frac{\mathcal{E}_\rho(f)-\mathcal{E}_{\mathbf{z},\rho}(f)}{\mathcal{E}_\rho(f)+\varepsilon}\right)\left(\frac{\mathcal{E}_\rho(f)+\varepsilon}{\mathcal{E}_\rho(g)+\varepsilon}\right) < 2\alpha.$$

Hence the sum of the two terms is less than 3α. □

We can now extend the result of Proposition 4.3:

Proposition 4.5. *Let $\mathcal{F} \subset C(X,Y)$ be compact in the uniform topology, so that $\mathcal{N}(\mathcal{F},\delta)$ (the minimum number of discs (in the uniform norm) of radius δ needed to cover $\mathcal{F}$) is finite for each $\delta > 0$, and let M satisfy* (6). *Then for all $\varepsilon > 0$, $0 < \alpha < 1$,*

$$\operatorname{Prob}\left\{\mathbf{z} \in Z^m : \sup_{f\in\mathcal{F}} \frac{\mathcal{E}_\rho(f) - \mathcal{E}_{\mathbf{z},\rho}(f)}{\mathcal{E}_\rho(f) + \varepsilon} \geq 3\alpha\right\} \leq \mathcal{N}\left(\mathcal{F}, \frac{\alpha\varepsilon}{4M}\right) \exp\left(\frac{-m\alpha^2\varepsilon}{8M^2}\right).$$

Proof. Let $\ell = \mathcal{N}\left(\mathcal{F}, \frac{\alpha\varepsilon}{4M}\right)$, so that there exist $D_1, D_2, \ldots, D_\ell \subset \mathcal{F}$ discs of radius $\frac{\alpha\varepsilon}{4M}$ centered at $f_1, \ldots, f_\ell$ that together cover $\mathcal{F}$. Then we have:

$$\begin{aligned}
\operatorname{Prob}\left\{\sup_{f\in\mathcal{F}} \frac{\mathcal{E}_\rho(f) - \mathcal{E}_{\mathbf{z},\rho}(f)}{\mathcal{E}_\rho(f) + \varepsilon} \geq 3\alpha\right\} &\leq \sum_{j=1}^{\ell} \operatorname{Prob}\left\{\sup_{f\in D_j} \frac{\mathcal{E}_\rho(f) - \mathcal{E}_{\mathbf{z},\rho}(f)}{\mathcal{E}_\rho(f) + \varepsilon} \geq 3\alpha\right\} \\
&\leq \sum_{j=1}^{\ell} \operatorname{Prob}\left\{\frac{\mathcal{E}_\rho(f_j) - \mathcal{E}_{\mathbf{z},\rho}(f_j)}{\mathcal{E}_\rho(f_j) + \varepsilon} \geq \alpha\right\} \\
&\leq \sum_{j=1}^{\ell} \exp\left(\frac{-m\alpha^2\varepsilon}{8M^2}\right) \\
&= \mathcal{N}\left(\mathcal{F}, \frac{\alpha\varepsilon}{4M}\right) \exp\left(\frac{-m\alpha^2\varepsilon}{8M^2}\right)
\end{aligned}$$

The second inequality follows from Lemma 4.4, since for any $\eta > 0$ there exists $f \in D_j$ with $\frac{\mathcal{E}_\rho(f)-\mathcal{E}_{\mathbf{z},\rho}(f)}{\mathcal{E}_\rho(f)+\varepsilon} \geq 3\alpha - 3\eta$, and so we must have $\frac{\mathcal{E}_\rho(f_j)-\mathcal{E}_{\mathbf{z},\rho}(f_j)}{\mathcal{E}_\rho(f_j)+\varepsilon} \geq \alpha - \eta$, and we can let $\eta \to 0$. The third inequality is just Proposition 4.3. □

We can now take a special case of this proposition at a single point:

Theorem 4.6. *Suppose that $\mathcal{F} \subset C(X, \mathbb{R}^l)$ is compact and convex, and that the probability measure ρ on $X \times Y \subset X \times \mathbb{R}^l$ (with regression function u_ρ satisfies: $\forall f \in \mathcal{F}$, $\|f(x) - y\| \leq M$ ρ_X-a.e. If for each $\mathbf{z}$, $f_\mathbf{z}$ is any minimizer of $\mathcal{E}_\mathbf{z}$, then for all $\varepsilon > 0$,*

$$\operatorname{Prob}\left\{\mathbf{z} \in Z^m : \int_X \|f_\mathbf{z}(x) - u_\rho(x)\|^2 \, d\rho_X(x) \geq \varepsilon\right\} \leq \mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{24M}\right) \exp\left(\frac{-m\varepsilon}{288M^2}\right).$$

Proof. Let $\alpha = \frac{1}{6}$ and $f = f_\mathbf{z}$ in Proposition 4.5 to obtain

$$\operatorname{Prob}\left\{\mathbf{z} \in Z^m : \frac{\mathcal{E}_\rho(f_\mathbf{z}) - \mathcal{E}_{\mathbf{z},\rho}(f_\mathbf{z})}{\mathcal{E}_\rho(f_\mathbf{z}) + \varepsilon} \geq \frac{1}{2}\right\} \leq \mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{24M}\right) \exp\left(\frac{-m\varepsilon}{288M^2}\right).$$

(Note that since $f_{\mathbf{z}}$ depends on $\mathbf{z}$, we could not use Proposition 4.3.) Now if $\frac{\mathcal{E}_\rho(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z},\rho}(f_{\mathbf{z}})}{\mathcal{E}_\rho(f_{\mathbf{z}})+\varepsilon} < \frac{1}{2}$, we can rewrite this as $\mathcal{E}_\rho(f_{\mathbf{z}}) < 2\mathcal{E}_{\mathbf{z},\rho}(f_{\mathbf{z}}) + \varepsilon$. However, $\mathcal{E}_{\mathbf{z},\rho}(f_{\mathbf{z}}) = \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(u_\rho) \leq 0$ since $f_{\mathbf{z}}$ minimizes $\mathcal{E}_{\mathbf{z}}$. Hence in this case $\mathcal{E}_\rho(f_{\mathbf{z}}) < \varepsilon$. Therefore, if $\mathcal{E}_\rho(f_{\mathbf{z}}) \geq \varepsilon$, we must have $\frac{\mathcal{E}_\rho(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z},\rho}(f_{\mathbf{z}})}{\mathcal{E}_\rho(f_{\mathbf{z}})+\varepsilon} \geq \frac{1}{2}$, and so

$$\text{Prob}\left\{\mathbf{z} \in Z^m : \mathcal{E}_\rho(f_{\mathbf{z}}) \geq \varepsilon\right\} \leq \mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{24M}\right) \exp\left(\frac{-m\varepsilon}{288M^2}\right).$$

□

5 Proof of the Main Theorem

We are now ready to prove the main theorem:

Main Theorem. *Suppose Λ is weakly irreducible. Then there exist constants α_* and $\mathbf{C}$ such that whenever $m \geq \frac{288M^2}{\varepsilon}\left[\ln\left(\delta^{-1}tk\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{24M}\right)\right)\right]$, where $\varepsilon = k^{-1}\alpha_*^{2t} d(\mathbf{f}^{(0)}, \Delta_{\mathcal{F}})^2(1-\alpha_*)^2$, we have*

$$d(\mathbf{f}^{(t)}, \Delta_{\mathcal{F}}) \leq \mathbf{C}\alpha_*^t d(\mathbf{f}^{(0)}, \Delta_{\mathcal{F}})$$

with probability at least $1-\delta$.

Proof. Let us consider stage t of a learning dynamic discussed in §2. Then since to produce $f_i^{(t)}$ from samples distributed by $\rho_i^{(t-1)}$, the regression function for which is $F_i^{(t-1)}$, we can rephrase Theorem 4.6 as

$$\text{Prob}\left\{\mathbf{z}_i^{(t)} \in Z^m : \int_X \|f_i^{(t)}(x) - F_i^{(t-1)}(x)\|^2 \, d\pi(x) \geq \varepsilon\right\} \leq \mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{24M}\right) \exp\left(\frac{-m\varepsilon}{288M^2}\right).$$

Applying this to each $i = 1, \ldots, k$, we find that

$$\text{Prob}\left\{\|\mathbf{f}^{(t)} - \mathbf{F}^{(t-1)}\|^2 \geq k\varepsilon\right\} \leq k\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{24M}\right) \exp\left(\frac{-m\varepsilon}{288M^2}\right),$$

where the norm comparing $\mathbf{f}^{(t)}$ to $\mathbf{F}^{(t-1)}$ is the sum of the L^2 norms comparing $f_i^{(t)}$ to $F_i^{(t-1)}$. Suppose now that we have chosen samples so that $d(\mathbf{f}^{(t)}, \mathbf{F}^{(t-1)})^2 = \|\mathbf{f}^{(t)} - \mathbf{F}^{(t-1)}\|^2 < k\varepsilon$. Recalling that we chose constants C_Λ and C'_Λ so that $C'_\Lambda\|\mathbf{v}\| \leq \|\mathbf{v}\|_\Lambda \leq C_\Lambda\|\mathbf{v}\|$ for every $\mathbf{v} \in \mathbb{R}^k$, it follows that for every $\mathbf{f}, \mathbf{g} \in \mathcal{F}$, $C'_\Lambda d(\mathbf{f}, \mathbf{g}) \leq d_\Lambda(\mathbf{f}, \mathbf{g}) \leq C_\Lambda d(\mathbf{f}, \mathbf{g})$ where d_Λ is the metric defined in §3.4. This means that

$$d_\Lambda(\mathbf{f}^{(t)}, \mathbf{F}^{(t-1)}) \leq C_\Lambda\|\mathbf{f}^{(t)} - \mathbf{F}^{(t-1)}\| < C_\Lambda\sqrt{k\varepsilon}.$$

This allows us to estimate:

$$\begin{aligned} d_\Lambda(\mathbf{f}^{(t)}, \Delta_{\mathcal{F}}) &\leq d_\Lambda(\mathbf{f}^{(t)}, \mathbf{F}^{(t-1)}) + d_\Lambda(\mathbf{F}^{(t-1)}, \Delta_{\mathcal{F}}) \\ &= d_\Lambda(\mathbf{f}^{(t)}, \mathbf{F}^{(t-1)}) + d_\Lambda(\Lambda\mathbf{f}^{(t-1)}, \Delta_{\mathcal{F}}) \\ &\leq C_\Lambda\sqrt{k\varepsilon} + \alpha_* d_\Lambda(\mathbf{f}^{(t-1)}, \Delta_{\mathcal{F}}) \end{aligned}$$

Then a simple induction on t shows that

$$\begin{aligned} d_\Lambda(\mathbf{f}^{(t)}, \Delta_{\mathcal{F}}) &\leq C_\Lambda\sqrt{k\varepsilon}\sum_{n=0}^{t-1}\alpha_*^n + \alpha_*^t d_\Lambda(\mathbf{f}^{(0)}, \Delta_{\mathcal{F}}) \\ &\leq \frac{C_\Lambda\sqrt{k\varepsilon}}{1-\alpha_*} + \alpha_*^t d_\Lambda(\mathbf{f}^{(0)}, \Delta_{\mathcal{F}}), \text{ i.e.} \end{aligned}$$

$$d(\mathbf{f}^{(t)}, \Delta_{\mathcal{F}}) \leq \frac{C_\Lambda}{C'_\Lambda}\left[\frac{\sqrt{k\varepsilon}}{1-\alpha_*} + \alpha_*^t d(\mathbf{f}^{(0)}, \Delta_{\mathcal{F}})\right]$$

What is the probability of this failing to hold? We know that each of the t induction steps will fail with probability at most $k\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{24M}\right)\exp\left(\frac{-m\varepsilon}{288M^2}\right)$, so our confidence in the above statement is at least

$$1 - tk\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{24M}\right)\exp\left(\frac{-m\varepsilon}{288M^2}\right).$$

We can make this confidence $1-\delta$ by letting

$$m \geq \frac{288M^2}{\varepsilon}\left[\ln\left(\delta^{-1}tk\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{24M}\right)\right)\right]$$

Then letting $\varepsilon = k^{-1}\alpha_*^{2t}d(\mathbf{f}^{(0)}, \Delta_{\mathcal{F}})^2(1-\alpha_*)^2$ and $\mathbf{C} = 2\frac{C_\Lambda}{C'_\Lambda}$, we have

$$d(\mathbf{f}^{(t)}, \Delta_{\mathcal{F}}) \leq \mathbf{C}\alpha_*^t d(\mathbf{f}^{(0)}, \Delta_{\mathcal{F}}) \text{ with confidence } 1-\delta$$

as advertised. □

References

[1] Cucker, F. and S. Smale (2002). On the Mathematical Foundations of Learning. *Bulletin of the American Mathematical Society, Volume 39*, 1-49.

[2] Cucker, F., S. Smale, and D.-X. Zhou (2004). Modelling Language Evolution. *Found. Comput. Math.*, 4:315-343.

[3] Folland, Gerald B. (1999). *Real Analysis: Modern Techniques and Applications.* John Wiley and Sons.

[4] D. Pollard (1984). *Convergence of stochastic processes.* Springer-Verlag. MR 86i:60074

[5] Seneta, E. (1973). *Non-Negative Matrices.* John Wiley and Sons.