Learning in Mean-Field Oscillator Games

Huibing Yin, Prashant G. Mehta, Sean P. Meyn and Uday V. Shanbhag

***Abstract*— This research concerns a noncooperative dynamic game with large number of oscillators. The states are interpreted as the phase angles for a collection of non-homogeneous oscillators, and in this way the model may be regarded as an extension of the classical coupled oscillator model of Kuramoto.**

We introduce approximate dynamic programming (ADP) techniques for learning approximating optimal control laws for this model. Two types of parameterizations are considered, each of which is based on analysis of the deterministic PDE model introduced in our prior research. In an offline setting, a Galerkin procedure is introduced to choose the optimal parameters. In an online setting, a steepest descent stochastic approximation algorithm is proposed. We provide detailed analysis of the optimal parameter values as well as the Bellman error with both the Galerkin approximation and the online algorithm.

Finally, a phase transition result is described for the large population limit when each oscillator uses the approximately optimal control law. A critical value of the control cost parameter is identified: Above this value, the oscillators are incoherent; and below this value (when control is sufficiently cheap) the oscillators synchronize. These conclusions are illustrated with results from numerical experiments.

I. Introduction

Computation of optimal or approximately optimal control laws in large population of coupled heterogeneous nonlinear systems is of interest in a number of applications, including neuroscience, networks, economics, and power markets. In this paper we introduce methods for approximating optimal control laws in a model of coupled oscillators. Our approach draws on the game-theoretic analysis in our recent paper [1].

As in [1], we consider a set of N oscillators, denoted by $\mathscr{N} := \{1,\dots,N\}$. The model for the i^{th} oscillator is:

$$\mathrm{d}\theta_i(t) = (\omega_i + u_i(t))\,\mathrm{d}t + \sigma\,\mathrm{d}\xi_i(t), \tag{1}$$

where $\theta_i(t)$ is the phase of the i^{th} oscillator at time t, $u_i(t)$ is the control input, and $\{\xi_i(t),\ i\in\mathscr{N}\}$ are mutually independent standard Wiener processes. The frequencies $\{\omega_i\}$ are chosen independently according to a fixed distribution with density g, which is supported on an interval of the form $\Omega = [1-\gamma, 1+\gamma]$ for some $\gamma < 1$. One important type of control is the *Kuramoto control* $u_i = u_i^{(\text{Kur})}(\theta_i,t) := -\kappa\frac{1}{N}\sum_{j\neq i}\sin(\theta_i-\theta_j(t))$ [2].

We consider a game-theoretic model. Specifically, we assume that the i^{th} oscillator minimizes its own performance objective, given the decisions of (competing) oscillators:

$$\eta_i^{(\text{POP})}(u_i;u_{-i}) = \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathsf{E}[c(\theta_i;\theta_{-i}) + \tfrac{1}{2}Ru_i^2]\,\mathrm{d}s, \tag{2}$$

where $\theta_{-i} = (\theta_j)_{j\neq i}$, $c(\cdot)$ is a cost function, $u_{-i} = (u_j)_{j\neq i}$ and R models the control penalty. The form of the function c and the value of R are assumed to be common to the entire population. A *Nash equilibrium* in control policies is given by $\{u_i^*\}$ such that u_i^* minimizes $\eta_i^{(\text{POP})}(u_i;u_{-i}^*)$ for $i = 1,\dots,N$.

The cost function c is separable, as shown below

$$c(\theta_i;\theta_{-i}) := \frac{1}{N}\sum_{j\neq i} c^\bullet(\theta_i,\theta_j(t)), \tag{3}$$

where $c^\bullet$ is a non-negative function on $\mathbb{R}^2$. In the remainder of this paper, we assume the following:

Assumption 1.1: The frequency distribution $g(\omega)$ is the uniform distribution on Ω and $c^\bullet(\theta,\vartheta) = \frac{1}{2}\sin^2\left(\frac{\theta-\vartheta}{2}\right)$.

In [1] we derived a deterministic PDE model in the large population ($N\to\infty$) limit. The modeling approach follows the seminal work of Huang et. al. [3] and Weintraub et. al. [4], and is based on a version of the mean-field approximation, central to the study of the interacting particle systems. The solutions of the PDE describe ε-Nash equilibria for the noncooperative game with a finite number of oscillators.

The bifurcation analysis of the PDE model reveals a phase transition depicted in Fig. 1: For $R > R_c$, the oscillators are incoherent, and for $R < R_c$ the oscillators synchronize. That is, the oscillators synchronize when the control is sufficiently cheap. Qualitatively, such a phase transition is believed to be important in a number of applications. For example, in thalamocortical circuits in the brain, transition to synchrony is associated with diseased brain states such as epilepsy [5].

The focus of this paper is to introduce approximate dynamic programming (ADP) based methods to synthesize approximately optimal feedback control laws for the noncooperative game. The motivation for this is three-fold: One, we are interested in a formulation that yields time-invariant *causal feedback* control laws for the game as opposed to the time-dependent distributed control laws that are obtained from solution of the PDEs. Two, the ADP formulation naturally suggests that each oscillator can *learn* an approximately optimal policy using simulation-based methods such as Q-learning. Learning schemes are important in applications of interest. Three, we would like to better understand the

H. Yin and P. G. Mehta are with the Coordinated Science Laboratory and the Department of Mechanical Science and Engineering at the University of Illinois at Urbana-Champaign (UIUC) yin3@illinois.edu; mehtapg@illinois.edu

S. P. Meyn is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory at UIUC meyn@illinois.edu

U. V. Shanbhag is with the Department of Industrial and Enterprise Systems Engineering at UIUC udaybag@illinois.edu

Financial support from the AFOSR grant FA9550-09-1-0190 and NSF grant CCF-0728863 are gratefully acknowledged. The authors thank Roland Malhame for many useful discussions.

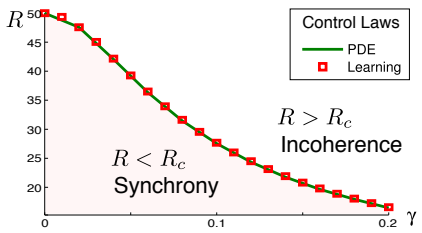

Fig. 1. Bifurcation diagram: Comparison of stability boundary obtained from the analysis of the PDE model introduced in [1] and from the analysis of the large population, where each oscillator applies the approximately optimal control law obtained using the learning algorithm.

relationship between the game theoretic solution described in [1] and the classical Kuramoto control [2].

Four types of analyses are presented in this paper:

1. Approximate dynamic programming. The key to the success of many of the learning algorithms is the specification of an appropriate parametrization for an approximation of the value function. The analysis of the PDE solution reveals that the game theoretic control law closely approximates the Kuramoto control law. This suggests two types of parametrizations that we introduce in this paper.

2. Learning algorithm. We focus on a variant of a *Q*-learning algorithm [6], in which an approximation of a *Q-function* leads to an approximation of the optimal control law. In an offline setting, a Galerkin procedure is introduced to choose the optimal parameters. In an online setting, a steepest descent stochastic approximation algorithm is proposed.

3. Error analysis. A salient feature of this work is that we provide detailed analysis of both the Galerkin approximation and the learning algorithm: One, we characterize optimal parameter values and show them to be consistent. Two, an estimate of Bellman error is provided.

The final contribution of this paper is grounded in the large population (infinite-N continuum) limit:

4. Transition from incoherence to synchrony. The goal of the final analysis is to describe the population behavior if the local control laws obtained in step 2 are applied to each oscillator. The bifurcation diagram, depicted in Fig. 1, reveals a phase transition with two distinct types of population behavior:

Incoherence The control solution is $u_i^* \equiv 0$, which coincides with the solution of the game. The Bellman error in this case is zero.

Synchrony For $R < R_c$, the population synchronizes. Detailed comparison of the average cost with the game theoretic solution and the ADP solution are provided. The Bellman error is small if σ^2 is large.

The remainder of this paper is organized as follows. In Sec. II, the main results of our earlier paper [1] are briefly reviewed. The approximate dynamic programming framework is introduced in Sec. III, and the two parametrizations described in Sec. IV. The analysis with these two parametrizations is reported in Sec. V and Sec. VI, respectively. The main conclusions are illustrated with numerical experiments in Sec. VII.

II. Preliminaries

In this section we briefly summarize the main results of [1] for the noncooperative game (1) - (2).

A. Mean-field approximation

If N is large, the sum in (3) is expected to be nearly deterministic when the frequencies $\{\omega_i\}$ are independently sampled according to the density g. The law of large numbers suggests the approximation of $c(\vartheta;\theta_{-i}(t))$ by $\bar{c}(\vartheta,t)$:

$$\bar{c}(\vartheta,t) \approx \frac{1}{N}\sum_{j\neq i} c^\bullet(\vartheta,\theta_j(t)). \tag{4}$$

For the scalar model (1) with cost $\bar{c}(\vartheta,t)$ depending only on $\vartheta = \theta_i$, the game reduces to independent optimal control problems. The associated average-cost HJB equation is given by,

$$\min_{u_i}\{\bar{c}(\theta,t) + \tfrac{1}{2}Ru_i^2 + \mathscr{D}_{u_i}h_i(\theta,t)\} = \eta_i^*, \tag{5}$$

where $\mathscr{D}_u$ denotes the controlled generator, defined for C^2 functions f via,

$$\mathscr{D}_u f = \partial_t f + (\omega_i + u)\partial_\theta f + \frac{\sigma^2}{2}\partial_{\theta\theta}^2 f,$$

where ∂_t and ∂_θ denote the partial derivative with respect to t and θ, respectively, and $\partial_{\theta\theta}^2$ denotes the second derivative with respect to θ. Because the cost is quadratic in u_i and the dynamics are linear in u_i, this leads to the nonlinear HJB equation

$$\partial_t h_i + \omega\partial_\theta h_i = \frac{1}{2R}(\partial_\theta h_i)^2 - \bar{c}(\theta,t) + \eta_i^* - \frac{\sigma^2}{2}\partial_{\theta\theta}^2 h_i,$$

and the optimal control law

$$u_i^* = -\frac{1}{R}\partial_\theta h_i(\theta_i,t). \tag{6}$$

B. PDE model

The notation in the large population limit is a minor variant of the $N = 1$ solution: The relative value function is denoted by $h(\theta,t,\omega)$, which is a solution to the HJB equation,

$$\partial_t h + \omega\partial_\theta h = \frac{1}{2R}(\partial_\theta h)^2 - \bar{c}(\theta,t) + \eta^* - \frac{\sigma^2}{2}\partial_{\theta\theta}^2 h. \tag{7}$$

The evolution of the population is described through a Fokker-Planck-Kolmogorov (FPK) equation. For any i and any $t > 0$, the density $p(\cdot,t,\omega_i)$ is intended to approximate the probability density of the random variable $\theta_i(t)$, evolving according to the stochastic differential equation (1) with optimal control law (6). For the population, the density is denoted by $p(\theta,t,\omega)$ and the FPK equation is given by

$$\partial_t p + \partial_\theta\left[\left(\omega - \frac{1}{R}\partial_\theta h\right)p\right] = \frac{\sigma^2}{2}\partial_{\theta\theta}^2 p.$$

Finally, the two PDEs are coupled through an integral that defines the relationship between cost $\bar{c}$ and density p:

$$\bar{c}(\vartheta,t) = \int_\Omega\int_0^{2\pi} c^\bullet(\vartheta,\theta)p(\theta;t,\omega)g(\omega)\,\mathrm{d}\theta\,\mathrm{d}\omega. \tag{8}$$

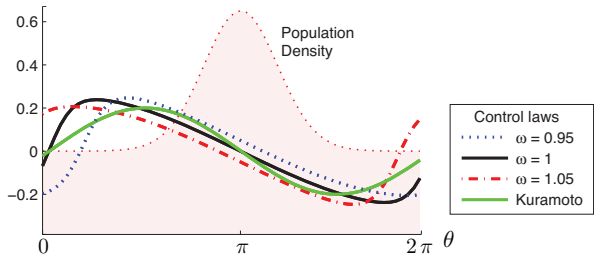

Fig. 2. Comparison of the control obtained from solving the PDE model (7) - (8) and from Kuramoto model.

C. ε-Nash equilibrium

For a finite population, each oscillator is controlled using the control solution in (6),

$$u_i^o = -\frac{1}{R}\partial_\theta h(\theta,t,\omega)\Big|_{\omega=\omega_i}.$$

Thm. 3.3 in [1], repeated below, shows that this control law is an ε-Nash equilibrium for (2).

Theorem 2.1: The oblivious control $\{u_i^o\}$ is an ε-Nash equilibrium for (2): For any adapted control u_i,

$$\eta_i^{(\mathrm{POP})}(u_i^o;u_{-i}^o) \le \eta_i^{(\mathrm{POP})}(u_i;u_{-i}^o) + O(\frac{1}{\sqrt{N}}). \qquad \blacksquare$$

D. Analysis of phase transition

means that we can obtain an ε-Nash equilibrium of the game (1) - (3) by considering the solution of the PDEs (7) - (8). Two types of solution are described in [1]:

(i) *Incoherence solution:*

$$p(\theta,t,\omega) = p_0(\theta) := \frac{1}{2\pi},\ h(\theta,t,\omega) = h_0(\theta) := 0,$$

with associated control law $u(t) \equiv 0$.

(ii) *Synchrony solution:* The traveling wave equation,

$$p(\theta,t,\omega) = p(\theta-t,0,1),\ h(\theta,t,\omega) = h(\theta-t,0,\omega)$$

The two types of solution can be visualized using a bifurcation diagram in the (R,γ)-plane (see Fig. 1).

E. Comparison to Kuramoto model

The synchrony solution is obtained via numerical solution of the coupled PDEs (7) - (8). Fig. 2 depicts the optimal control laws $u^*(\theta,t) = -\frac{1}{R}\partial_\theta h(\theta,t,\omega)$ in relation to the population density. Also depicted is the Kuramoto control law $u_i^{(\mathrm{Kur})}(\theta,t) = -\frac{\kappa}{N}\sum_{j\neq i}\sin(\theta-\theta_j(t))$.

The comparison shows that the optimal control law is "close to" the Kuramoto control law. This provides motivation for the ADP architecture described in the next section.

III. Approximate Dynamic Programming

In this section we develop methods to construct approximate solutions to the Bellman equation (5). We assume a mean-field approximation (4), so $c(\theta;\theta_{-i}(t)) = \bar{c}(\theta,t)$, and denote

$$H_i(\theta,u_i,t) := c(\theta;\theta_{-i}(t)) + \tfrac{1}{2}Ru_i^2 + \mathscr{D}_{u_i}h_i(\theta,t). \qquad (9)$$

Assumption 3.1: The functions $h_i(\theta,t)$ and $\bar{c}(\theta,t)$ are periodic functions of time, with a common period.

Define

$$\underline{H}_i(\theta,t) := \min_{u_i}\{H_i(\theta,u_i,t)\}.$$

In this notation, the nonlinear HJB equation (5) is simply given by

$$\underline{H}_i(\theta,t) = \eta_i^* = \text{constant}.$$

As explained in [7], the function H_i defined in (9) is analogous to the Q-function that arises in the Q-learning algorithm.

A. Bellman error

The goal of Q-learning is to approximate the Q-function within a parameterized class:

$$H_i^\alpha(\theta,u,t) = c(\theta;\theta_{-i}(t)) + \tfrac{1}{2}Ru_i^2 + \mathscr{D}_{u_i}h_i^\alpha(\theta,t), \qquad \alpha\in\mathbb{R}^d, \qquad (10)$$

where $h_i^\alpha(\theta,t)$ will be constructed in a separable form, as shown below:

$$h_i^\alpha(\theta,t) = \frac{1}{N}\sum_{j\neq i}G^\alpha(\theta,\theta_j(t)).$$

Note that the term on the right hand side is a stochastic function, while $h_i^\alpha(\theta,t)$ is a deterministic function. This is justified for large N using a mean-field approximation.

Our goal is to choose the parameter α so that $H_i^\alpha \approx H_i$, where the approximation is with respect to a specific error criterion. The quality of the approximation crucially depends on the choice of the function $\{G^\alpha\}$, and N being large.

On denoting $\underline{H}_i^\alpha(\theta,t) := \min_{u_i} H_i^\alpha(\theta,u_i,t)$, the pointwise error in the DP equation is denoted by

$$\mathscr{L}^\alpha(\theta,t) := \underline{H}_i^\alpha(\theta,t) - \eta_i^\alpha,$$

where η_i^α is the mean value of the periodic function $\underline{H}_i^\alpha(\theta,t)$.

Throughout this paper we adopt a Hilbert space setting for approximation. On letting $\|\cdot\|_{\mathscr{H}}$ denote the associated norm on function space, we define the Bellman error

$$\varepsilon_{Bell}(\alpha) := \tfrac{1}{2}\|\mathscr{L}^\alpha\|_{\mathscr{H}}^2.$$

Computation is made possible by choosing the Hilbert space in terms of ergodic averages. For any real-valued function F on the product space $[0,2\pi]\times\mathbb{R}^+$ we denote,

$$\|F\|_{\mathscr{H}}^2 := \lim_{T\to\infty}\frac{1}{T}\int_0^T\int_0^{2\pi}\big(F(\theta,t)\big)^2\,\mathrm{d}\theta\,\mathrm{d}t$$

The Hilbert space $\mathscr{H}$ is defined to be the set of functions for which the limit exists and is finite. The associated inner product is given by

$$\langle F,G\rangle_{\mathscr{H}} = \lim_{T\to\infty}\frac{1}{T}\int_0^T\int_0^{2\pi}F(\theta,t)G(\theta,t)\,\mathrm{d}\theta\,\mathrm{d}t.$$

In the calculations that follow we will require the Fourier series of a function $F \in \mathscr{H}$:

$$[F(\cdot,t))]_0 = \frac{1}{2\pi}\int_{\theta=0}^{2\pi} F(\theta,t)\,d\theta,$$
$$[F(\cdot,t)]_c = \frac{1}{2\pi}\int_{\theta=0}^{2\pi} F(\theta,t)\cos(\theta)\,d\theta,$$
$$[F(\cdot,t)]_s = \frac{1}{2\pi}\int_{\theta=0}^{2\pi} F(\theta,t)\sin(\theta)\,d\theta\,.$$

B. Galerkin relaxation

Our goal is to choose $\alpha^* \in \mathbb{R}^d$ so that the Bellman error is zero, or nearly so. One possible approach to obtain a solution is to consider a Galerkin relaxation.

Let $\varphi(\theta,t) = (\varphi_i(\theta,t))^T$ denote a d-dimensional function on $[0,2\pi]\times\mathbb{R}^+$, with entries $\{\varphi_i : 1 \le i \le d\} \subset \mathscr{H}$. The Galerkin relaxation is obtained by setting the projection onto the associated d-dimensional subspace equal to zero:

$$\langle \mathscr{L}^\alpha, \varphi_i\rangle_{\mathscr{H}}\Big|_{\alpha=\alpha^*} = 0. \tag{11}$$

C. Gradient descent algorithm

One approach to compute α^* is to minimize the error,

$$e(\alpha) = \sum_{i=1}^{d} |\langle \mathscr{L}^\alpha, \varphi_i\rangle_{\mathscr{H}}|^2$$

Minimization can be accomplished using a gradient or Newton iteration, but this would require computation of inner-products at each stage of the algorithm.

The stochastic approximation algorithm is obtained on removing the integration: We define the point-wise error by

$$\tilde{e}(t;\alpha) = \sum_{i=1}^{d} \left([\mathscr{L}^\alpha(\,\cdot\,,t), \varphi_i]_0\right)^2.$$

The gradient descent algorithm is given by,

$$\frac{\mathrm{d}\alpha}{\mathrm{d}t} = -\varepsilon(t)\frac{\mathrm{d}\tilde{e}(t;\alpha)}{\mathrm{d}\alpha},$$

where $\varepsilon(\,\cdot\,)$ is a non-negative gain that satisfies the standard conditions,

$$\varepsilon(t) > 0, \quad \int_0^\infty \varepsilon^2(t)\,\mathrm{d}t < \infty, \quad \int_0^\infty \varepsilon(t)\,\mathrm{d}t = \infty. \tag{12}$$

Although formulated in a straight forward way, the steps are often challenging to analyze and apply in nonlinear non-convex settings. The objective of the remainder of the paper thus is to choose the basis functions gained via analysis of PDEs (7) - (8) to define parametrization that are easily used and theoretically justifiable.

IV. Parametrization

We begin by recalling the results of numerical experiments (see Fig. 2) where we showed the game-theoretic optimal control law (6),

$$u_i(\theta,t) = -\frac{1}{R}\partial_\theta h(\theta,t)$$

closely approximates the Kuramoto control law

$$u_i^{(\mathrm{Kur})}(\theta,t) = -\kappa\frac{1}{N}\sum_{j\neq i}\sin(\theta-\theta_j(t)).$$

We use this analogy to consider two basis functions

$$S^{(\phi)}(\theta,\theta_{-i}(t)) := \frac{1}{N}\sum_{j\neq i}\sin(\theta-\theta_j(t)-\phi),$$
$$C^{(\phi)}(\theta,\theta_{-i}(t)) := \frac{1}{N}\sum_{j\neq i}\cos(\theta-\theta_j(t)-\phi),$$

where ϕ is a phase variable.

Inspired by the Kuramoto law, we consider the following parameterized approximation of h_i:

$$h_i^{A,\phi}(\theta,t) = -AC^{(\phi)}(\theta,\theta_{-i}(t)) = -A\frac{1}{N}\sum_{j\neq i}\cos(\theta-\theta_j(t)-\phi). \tag{13}$$

Note the notation requires a mean-field approximation, which we assume. Before presenting the approximation of the Q-function, we make one additional assumption:

Assumption 4.1: The function $h_i^{A,\phi}(\theta,t) = \tilde{h}_i^{A,\phi}(\theta-t)$.

This assumption is motivated by the game theoretic solution (see Sec. II-D) where the synchrony solution is a traveling wave. The assumption is useful because it implies that

$$\partial_t h_i^{A,\phi} + \partial_\theta h_i^{A,\phi} = 0,$$

which serves to simplify the notation for the parameterized Q-function.

On substituting (13) in (10), we obtain a two dimensional parameterization with $\alpha^i = (A_i,\phi)$:

(P2)
$$\begin{aligned} &H_i^{\alpha^i}(\theta,u_i,t) \\ &\quad = c(\theta;\theta_{-i}(t)) + (\omega_i - 1 + u_i)A_iS^{(\phi)}(\theta,\theta_{-i}(t)) \\ &\qquad + \tfrac{1}{2}Ru_i^2 + \frac{\sigma^2}{2}A_iC^{(\phi)}(\theta,\theta_{-i}(t)). \end{aligned}$$

The phase ϕ for game-theoretic control law is seen to depend upon the frequency ω (See Fig. 2) with $\phi(1) = 0$, $\phi(\omega) > 0$ for $\omega < 1$ and $\phi(\omega) < 0$ for $\omega > 1$.

The Kuramoto law on the other hand has phase $\phi(\omega) \equiv 0$ and the control law is homogeneous for the population. This leads us to a simpler one-dimensional parametrization with $\alpha^i = A_i$, $h_i^A(\theta,t) = -AC^{(0)}(\theta,\theta_{-i}(t))$ and the Q-function:

(P1)
$$\begin{aligned} &H_i^{\alpha^i}(\theta,u_i,t) \\ &\quad = c(\theta;\theta_{-i}(t)) + \tfrac{1}{2}Ru_i^2 + u_iA_iS^{(0)}(\theta,\theta_{-i}(t)) \\ &\qquad + \frac{\sigma^2}{2}A_iC^{(0)}(\theta,\theta_{-i}(t)). \end{aligned}$$

One can interpret parametrization (P2) as the case where the oscillator knows its own frequency ω_i and (P1) as a simpler parametrization where frequency ω_i is not known.

In the following sections we present analysis and simulation results for these two parameterizations. We begin with the two-parameter case (P2).

V. Analysis of parametrization (P2)

On denoting $\underline{H}_i^{\alpha^i} = \min_{u_i}\{H_i^{\alpha^i}\}$, we have

$$\underline{H}_i^{\alpha^i}(\theta,t) = c(\theta;\theta_{-i}(t)) - \frac{1}{2R}\left(A_i S^{(\phi_i)}(\theta,\theta_{-i}(t))\right)^2 + (\omega_i - 1)A_i S^{(\phi_i)}(\theta,\theta_{-i}(t)) + \frac{\sigma^2}{2}A_i C^{(\phi_i)}(\theta,\theta_{-i}(t)).$$

The goal is to choose the parameters (A_i,ϕ_i) so that

$$\underline{H}_i^{\alpha^i}(\theta,t) = \text{constant}.$$

This may not be feasible, so instead we consider the Galerkin relaxation.

A. Galerkin relaxation

For the projection (11), we select the two functions as

$$\varphi_1(\theta,t) = \sin(\theta - t), \quad \varphi_2(\theta,t) = \cos(\theta - t). \tag{14}$$

The choice is motivated by the traveling wave solution observed in the synchrony. With this choice, simple trigonometric identities lead to the following representations for the projections:

$$\begin{aligned}\langle \mathscr{L}^{\alpha^i}, \varphi_1\rangle_{\mathscr{H}} &= \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathscr{L}_s^{\alpha^i}(t)\cos(t) - \mathscr{L}_c^{\alpha^i}(t)\sin(t)\,\mathrm{d}t,\\ \langle \mathscr{L}^{\alpha^i}, \varphi_2\rangle_{\mathscr{H}} &= \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathscr{L}_c^{\alpha^i}(t)\cos(t) + \mathscr{L}_s^{\alpha^i}(t)\sin(t)\,\mathrm{d}t,\end{aligned} \tag{15}$$

in which

$$\begin{aligned}\mathscr{L}_s^{\alpha^i} &:= [\mathscr{L}^{\alpha^i}(\cdot,t)]_s = (\omega_i - 1)A_i\bar{\psi}^c + \frac{\sigma^2}{2}A_i\bar{\psi}^s - \frac{1}{4}\psi^s,\\ \mathscr{L}_c^{\alpha^i} &:= [\mathscr{L}^{\alpha^i}(\cdot,t)]_c = -(\omega_i - 1)A_i\bar{\psi}^s + \frac{\sigma^2}{2}A_i\bar{\psi}^c - \frac{1}{4}\psi^c,\end{aligned} \tag{16}$$

where $\bar{\psi}^s$, $\bar{\psi}^c$, ψ^s, ψ^c are functions of the entire population:

$$\begin{aligned}\bar{\psi}^c &:= \frac{1}{N}\sum_{j\neq i}\cos(\theta_j(t)+\phi_i), \quad \bar{\psi}^s := \frac{1}{N}\sum_{j\neq i}\sin(\theta_j(t)+\phi_i),\\ \psi^c &:= \frac{1}{N}\sum_{j\neq i}\cos(\theta_j(t)), \quad \psi^s := \frac{1}{N}\sum_{j\neq i}\sin(\theta_j(t)).\end{aligned} \tag{17}$$

To obtain an approximate analysis we identify two idealized models of behavior associated with the infinite population limit:

1. *Incoherence solution* where the population density is $p(\theta,t,\omega) = \frac{1}{2\pi}$. So $\bar{\psi}^c = \frac{1}{N}\sum_{j\neq i}\cos(\theta_j(t)+\phi_i) \approx \frac{1}{2\pi}\int_0^{2\pi}\cos(\theta+\phi_i)\,\mathrm{d}\theta = 0$, and similarly $\bar{\psi}^s \approx 0$, $\psi^c \approx 0$, $\psi^s \approx 0$.
2. *Synchrony solution* where the population density $p(\theta,t,\omega) \approx \delta(\theta - t)$. So $\bar{\psi}^c = \int_0^{2\pi}\cos(\theta+\phi_i)\delta(\theta - t)\,\mathrm{d}\theta \approx \cos(t+\phi_i)$. Similarly, expressions can be obtained for $\bar{\psi}^s$, ψ^c, and ψ^s. These are summarized in Tab. I.

Using these as approximations for the behavior with finite N we obtain approximations for the solution of the Galerkin relaxation. We find that in either case, the approximation

TABLE I

Galerkin projection results

	Incoherence	Synchrony
$\bar{\psi}^c$	0	$\cos(t+\phi_i)$
$\bar{\psi}^s$	0	$\sin(t+\phi_i)$
ψ^c	0	$\cos(t)$
ψ^s	0	$\sin(t)$
$\langle \mathscr{L}^{\alpha^i}, \varphi_1\rangle_{\mathscr{H}}$	0	$A_i\left[(\omega_i - 1)\cos\phi_i + \frac{1}{2}\sigma^2\sin\phi_i\right]$
$\langle \mathscr{L}^{\alpha^i}, \varphi_2\rangle_{\mathscr{H}}$	0	$A_i\left[-(\omega_i - 1)\sin\phi_i + \frac{1}{2}\sigma^2\cos\phi_i\right] - \frac{1}{4}$

yields a Galerkin parameter α^{i*} for which the corresponding control law is given by,

$$u_i^{\alpha^{i*}}(\theta_i,t) = -\frac{A_i^*}{R}\frac{1}{N}\sum_{j\neq i}\sin(\theta_i - \theta_j(t) - \phi_i^*). \tag{18}$$

Theorem 5.1: Consider the two-parameter parametrization (P2) combined with the Galerkin basis functions (14). We have the following conclusions for the infinite-population approximate model:

(1) *Incoherence:* If the population is in incoherence, then the optimal control law is identically zero ($u_i^{\alpha^i}(\cdot,t) = 0$, $\forall\, t$), which coincides with (18) in the infinite population limit under incoherence. The average cost is $\eta_i^{\alpha^i} = 1/4$, and the pointwise Bellman error is identically zero, $\mathscr{L}^{\alpha^i}(\theta,t) = 0$. This solution is obtained for an arbitrary value of $\alpha^i = (A_i,\phi_i)$.

(2) *Synchrony:* If the population is in synchrony, then the optimal value of the parameter $\alpha^{i*} = (A_i^*,\phi_i^*)$ is,

$$A_i^* = \frac{1}{4\sqrt{(\omega_i - 1)^2 + (\frac{\sigma^2}{2})^2}}, \quad \phi_i^* = -\tan^{-1}\left(\frac{2(\omega_i - 1)}{\sigma^2}\right). \tag{19}$$

The resulting control law (18) results in the average cost $\eta_i^{\alpha^{i*}} = \frac{1}{4} - \varepsilon(R,\omega_i)$, and the pointwise Bellman error

$$\mathscr{L}^{\alpha^i}(\theta,t) = \varepsilon(R,\omega_i)\cos 2(\theta - t - \phi_i^*),$$

where

$$\varepsilon(R,\omega_i) = \frac{1}{64R[(\omega_i - 1)^2 + \sigma^4/4]}.$$

Proof: The formulae are obtained by setting the projections in (15) equal to zero, where (16)-(17) are used together with expressions in Tab. I. ∎

We remark that the pointwise Bellman error is zero in the incoherence regime because the parametrization recovers the optimal control $u_i^* \equiv 0$, for any values of the parameters (A_i,ϕ_i).

In the following section we analyze the infinite-N limit population behavior with the Galerkin-based control (18).

B. Analysis of phase transition with Galerkin solution

Using the Galerkin-based control (18), the closed-loop system is given

$$\mathrm{d}\theta_i(t) = \left[\omega_i - \frac{A_i^*}{RN}\sum_{j\neq i}\sin(\theta_i - \theta_j(t) - \phi_i^*)\right]\mathrm{d}t + \sigma\,\mathrm{d}\xi_i(t). \tag{20}$$

Our interest is to characterize the behavior of the infinite-N limit (see e.g., [8]). The limiting FPK equation for the density is given by

$$\partial_t p + \partial_\theta [pv] = \frac{\sigma^2}{2} \partial^2_{\theta\theta} p, \tag{21}$$

where the velocity $v(\theta,t,\omega)$ is given by

$$v(\theta,t,\omega) = \omega - \frac{A^*(\omega)}{R} \int_\Omega \int_0^{2\pi} \sin(\theta - \vartheta - \phi^*(\upsilon)) \, p(\vartheta,t,\upsilon) g(\upsilon) \, \mathrm{d}\vartheta \, \mathrm{d}\upsilon,$$

where $A^*(\omega) = \frac{1}{4\sqrt{(\omega-1)^2 + (\frac{\sigma^2}{2})^2}}$, $\phi^*(\omega) = -\tan^{-1}(\frac{2(\omega-1)}{\sigma^2})$.

The FPK equation (21) has an incoherence solution $p_0(\theta,t,\omega) = \frac{1}{2\pi}$ for all θ, t and ω. We investigate stability and possible bifurcation by taking a linearization of (21) about the incoherence solution p_0. A perturbation of the solution is denoted $p = p_0 + \tilde{p}$. Since $p = p_0 + \tilde{p}$ is a probability density, the perturbation satisfies the normalization condition $\int_0^{2\pi} \tilde{p}(\theta,t,\omega) \, \mathrm{d}\theta = 0$ for any t, ω. When $\tilde{p}$ is small, its evolution is approximated by the linear equation,

$$\partial_t \tilde{p} = -\omega \partial_\theta \tilde{p} + \frac{1}{2\pi} \partial_\theta \tilde{v} + \frac{\sigma^2}{2} \partial^2_{\theta\theta} \tilde{p} =: \mathscr{L}_R \tilde{p} \tag{22}$$

where

$$\tilde{v} = \frac{A^*(\omega)}{R} \int_\Omega \int_0^{2\pi} \sin(\theta - \vartheta - \phi^*(\upsilon)) \tilde{p}(\vartheta,t,\upsilon) g(\upsilon) \, \mathrm{d}\vartheta \, \mathrm{d}\upsilon.$$

The following theorem describes the spectrum of the linear operator $\mathscr{L}_R$:

Theorem 5.2: The discrete spectrum of operator $\mathscr{L}_R : L^2([0,2\pi]) \to L^2([0,2\pi])$ is the set

$$\mathbb{S}_d := \left\{ \lambda \in \mathbb{C} \;\middle|\; 1 = \frac{1}{2R} \int_\Omega \frac{e^{\mp i\phi^*(\omega)}}{\lambda + \frac{\sigma^2}{2} \pm i\omega} A^*(\omega) g(\omega) \, \mathrm{d}\omega \right\},$$

where $A^*(\omega) = \frac{1}{4\sqrt{(\omega-1)^2 + (\frac{\sigma^2}{2})^2}}$, $\phi^*(\omega) = -\tan^{-1}(\frac{2(\omega-1)}{\sigma^2})$. Furthermore, the continuous spectrum contains the set

$$\mathbb{S}_c := \left\{ \lambda \in \mathbb{C} \middle| \lambda = -\frac{\sigma^2}{2} k^2 - k\omega i, \, \omega \in \Omega, \, k = \pm 1, \pm 2, \ldots \right\}.$$

Proof: The eigenvalue calculation is straightforward. The proof of the continuous spectrum is omitted. It is similar to the proof of a corresponding result (Theorem 4.1) in [1]. ∎

If the noise is not zero, i.e., $\sigma \neq 0$, then the continuous spectrum is always in the strict left half plane. Hence stability of (22) is solely determined by the discrete spectrum $\mathbb{S}_d$. Analysis of discrete spectrum as a function of R and γ allows us to obtain the phase transition boundary for the closed-loop system (20) in the infinite-N limit.

Theorem 5.3: Consider the closed-loop system (20) where A_i^*, ϕ_i^* are defined in (19). Suppose ω_i is sampled from uniform distribution on $\Omega := [1-\gamma, 1+\gamma]$. Define

$$R_c(\gamma) = \begin{cases} \frac{1}{2\sigma^4} & \text{if } \gamma = 0, \\ \frac{1}{4\sigma^2 \gamma} \tan^{-1}\left(\frac{2\gamma}{\sigma^2}\right) & \text{if } \gamma > 0. \end{cases} \tag{23}$$

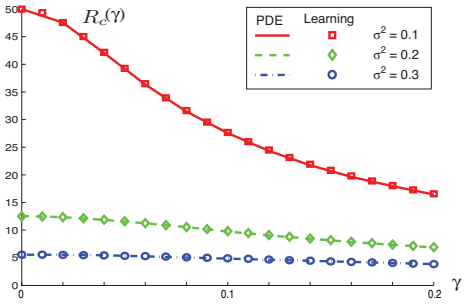

Fig. 3. Comparison of $R_c(\gamma)$ defined in (23), and the stability boundary obtained using the PDE model in [1].

Then in the infinite-N limit: For $R > R_c(\gamma)$, the system is in incoherence, and for $R < R_c(\gamma)$, the incoherence solution loses stability, via a Hopf bifurcation, to synchrony.

Proof: The stability boundary is obtained by setting the real part of the eigenvalue (in set $\mathbb{S}_d$) equal to zero. The Hopf bifurcation result is verified using certain transversality conditions [9]. These calculations are omitted here on account of space. We provide a numerical verification instead in Sec. VII. ∎

We plot the critical value R_c from (23) in the $R-\gamma$ plane with 3 different noise levels, i.e., $\sigma^2 = 0.1, 0.2, 0.3$. The results are labeled as "Learning" in Fig. 3. We also depict the critical value R_c obtained from game-theoretic model (labeled as "PDE") in the figure. The plots show that the two match extremely well.

C. Stochastic approximation algorithm

We now introduce a stochastic approximation procedure to compute $\alpha^{i*} = (A_i^*, \phi_i^*)$. For a given approximation $\alpha = (A_i, \phi_i)$ we define the error via

$$\tilde{e}(t;\alpha) = ([\mathscr{L}^\alpha(\cdot,t), \varphi_1]_0)^2 + ([\mathscr{L}^\alpha(\cdot,t), \varphi_2]_0)^2. \tag{24}$$

Estimates $\{A_i(t), \phi_i(t)\}$ are specified according to the gradient descent algorithm,

$$\frac{\mathrm{d}A_i(t)}{\mathrm{d}t} = -\varepsilon(t) \frac{\mathrm{d}\tilde{e}}{\mathrm{d}A_i}(t;\alpha(t)), \quad \frac{\mathrm{d}\phi_i(t)}{\mathrm{d}t} = -\varepsilon(t) \frac{\mathrm{d}\tilde{e}}{\mathrm{d}\phi_i}(t;\alpha(t)), \tag{25}$$

in which $\varepsilon(t) > 0$ satisfy (12), and the derivatives are obtained using (24):

$$\begin{aligned} \frac{\mathrm{d}\tilde{e}}{\mathrm{d}A_i} &= 2A_i \left((\omega_i - 1)^2 + (\frac{\sigma^2}{2})^2 \right) \left((\bar{\psi}^s)^2 + (\bar{\psi}^c)^2 \right) \\ &\quad + \frac{1}{2} \left[(\omega_i - 1)(\bar{\psi}^s \psi^c - \bar{\psi}^c \psi^s) - \frac{\sigma^2}{2} (\bar{\psi}^s \psi^s + \bar{\psi}^c \psi^c) \right], \\ \frac{\mathrm{d}\tilde{e}}{\mathrm{d}\phi_i} &= \frac{A_i}{2} \left[(\omega_i - 1)(\bar{\psi}^s \psi^s + \bar{\psi}^c \psi^c) + \frac{\sigma^2}{2} (\bar{\psi}^s \psi^c - \bar{\psi}^c \psi^s) \right], \end{aligned} \tag{26}$$

where $\bar{\psi}^s$, $\bar{\psi}^c$, ψ^s, ψ^c are defined as before in (16).

In application of this algorithm we simulate a finite population model. A value of i between 1 and N is selected, and the i^{th} oscillator uses the control $u_i^{\alpha^i}$ given in (18). The

remaining oscillators apply the Kuramoto control law. The resulting dynamical equations are given by,

$$\mathrm{d}\theta_j = \left(\omega_j - \frac{\kappa}{N}\sum_{\ell\neq j}\sin(\theta_j-\theta_\ell)\right)\mathrm{d}t + \sigma\,\mathrm{d}\xi_j(t), \quad j\neq i \tag{27}$$

$$\mathrm{d}\theta_i = \left(\omega_i - \frac{A_i(t)}{R}S^{(\phi_i(t))}(\theta,\theta_{-i}(t))\right)\mathrm{d}t + \sigma\,\mathrm{d}\xi_i(t), \tag{28}$$

in which $\{\xi_k(t), k\in\mathcal{N}\}$ denote independent Wiener process with instantaneous variance 1.

Theorem 5.4 describes possible equilibria of the algorithm in the two idealized solution regimes for the population (see Sec. V-A):

Theorem 5.4: Consider the system (27)-(28), where the i^{th}-oscillator updates its parameters $A_i(t),\phi(t)$ according to the algorithm (25)-(26). In the infinite-N limit, we have the following conclusions:

(1) If the population $\{\theta_j(t)\}_{j\neq i}$ is in incoherence, $\mathrm{d}\tilde{e}/\mathrm{d}A_i = \mathrm{d}\tilde{e}/\mathrm{d}\phi_i = 0$. So any $\alpha^i = (A_i,\phi_i)$ is an equilibrium solution.
(2) If the population $\{\theta_j(t)\}_{j\neq i}$ is in synchrony, the equilibrium is given by $A_i = A_i^*$, $\phi_i = \phi_i^*$, where A_i^*, ϕ_i^* are defined in (19).

Proof: The formulae are obtained by setting the right hand side of (26) equal to zero, where expressions in Tab. I are used in the two idealized population regimes. ∎

For the infinite-N idealized models of population behavior, the right hand side of (26) do not explicitly depend upon time. Analysis of equilibria thus is straightforward. For the finite-N case, this is not true. Here, convergence of the stochastic algorithm will require analysis of the averaged ODEs [10]. This is a subject of future work.

We now describe some numerical results for the stochastic approximation algorithm (25). In each simulation the population consisted of $N = 200$ oscillators. For the distinguished value i, the frequency of the i^{th} oscillator was taken to be $\omega_i = 1.1$. The remaining $N-1$ frequencies were sampled independently from the uniform distribution on $\Omega = [0.9, 1.1]$. The parameters A_i and ϕ_i are updated according to (25) in the following two cases:

(1) $\kappa = 0.01$: Population is in incoherence.
(2) $\kappa = 1$: Population is in synchrony.

Fig. 4 depicts $A_i(t)$ and $\phi_i(t)$ in the two cases. Given the conclusions of Thm. 5.4 it may be surprising to see that the algorithm is consistent in the incoherence regime. The explanation is that $\bar{\psi}^s = \bar{\psi}^c = \psi^s = \psi^c = 0$ in the incoherence regime only in the limiting case as $N\to\infty$. For finite N, the terms are only approximately zero, so there is sufficient information for the i^{th} oscillator to learn the optimal values. In the synchronous regime the parameters converge *quickly* to the optimal values predicted by Theorem 5.4. In the incoherence regime, the convergence is expected to get progressively slower as N increases.

The form of Kuramoto control law for the population is not particularly important, except that it allows us to make the point about the influence of the population on learning. The

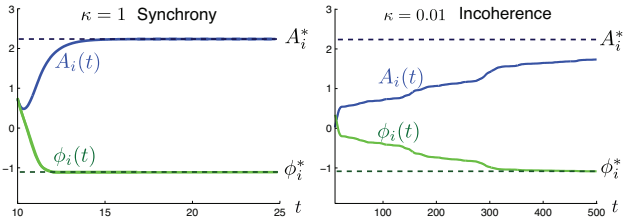

Fig. 4. Learning simulation: Parametrization (P2).

results of this section are *robust*: For instance, if we changed the control law of the j^{th} oscillator in the population to

$$u_j = -\frac{A_j^*}{R}S^{(\phi_j^*)}(\theta_j;\theta_{-j}).$$

Such a choice will be consistent with the Galerkin-based control law with parametrization (P2).

VI. Analysis of Parametrization (P1)

The analysis of one-dimensional parameterization is conceptually no different. We summarize the main conclusions:

1. The Galerkin-based control, counterpart of (18), is:

$$u_i^{\alpha^{i*}}(\theta_i,t) = -\frac{A_i^*}{R}\frac{1}{N}\sum_{j\neq i}\sin(\theta_i-\theta_j(t)), \tag{29}$$

where $A_i^* = \frac{1}{2\sigma^2}$. Note (29) is the Kuramoto control law.
2. The parameter value A_i^* can be computed in an online fashion by using a gradient descent algorithm

$$\frac{\mathrm{d}A_i}{\mathrm{d}t} = -\varepsilon\frac{\mathrm{d}\tilde{e}}{\mathrm{d}A_i},$$
$$\frac{\mathrm{d}\tilde{e}}{\mathrm{d}A_i} = \left(A_i\frac{\sigma^4}{2} - \frac{\sigma^2}{4}\right)\left((\psi^s)^2 + (\psi^c)^2\right),$$

where ψ^s, ψ^c are defined in (16).

VII. Comparison of Galerkin and PDE Results

Recall the optimal control (6) is $u^*(\theta,t;\omega) = -\frac{1}{R}\partial_\theta h^*(\theta,t,\omega)$. It is obtained as a numerical solution of the coupled PDEs (7)-(8) (see [1] for details on the numerical algorithm). The Galerkin control (18) is $u_i^{\alpha^{i*}}(\theta,t;\omega_i) = -\frac{A_i^*}{R}\frac{1}{N}\sum_{j\neq i}\sin(\theta-\theta_j(t)-\phi_i^*)$, where A_i^*, ϕ_i^* are defined in (19). In the experiments described here we have taken $N = 200$ oscillators, whose frequencies are sampled from a uniform distribution on $\Omega = [0.9, 1.1]$.

Figure 5 provides a comparison of the two control laws, plotted as a function of ω, using $R = 9$, and two values of the variance $\sigma^2 = 0.1$ and 0.2. The Galerkin-based control law (18) qualitatively captures the main features of the optimal control (6):

1. The Galerkin-based control leads to synchronization of the population for values of R smaller than the critical value R_c. In synchrony, the population density is a traveling wave with wave speed 1. This justifies, a posteriori, the Assumption 4.1.
2. The control is zero when $\omega = 1$, and θ lies at its mean value (equal to π in this figure, for the particular value

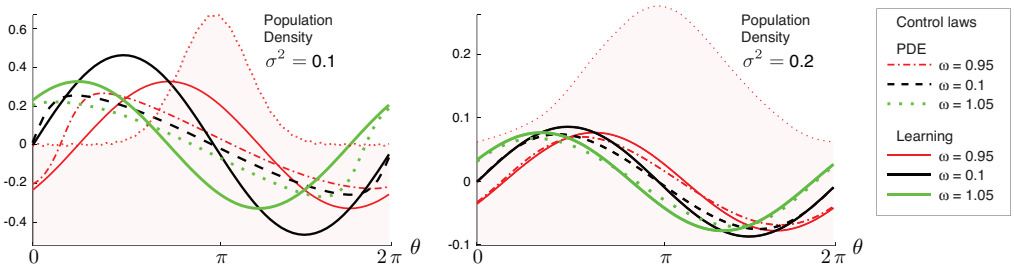

Fig. 5. Comparison of the optimal control $(u^*(\theta,t))$ and the Galerkin-based control $(u_i^{\alpha^{i*}}(\theta_i,t))$ for (a) $\sigma^2 = 0.1$ and (b) $\sigma^2 = 0.2$.

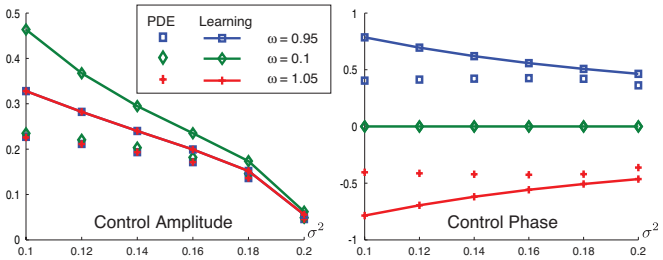

Fig. 6. Comparison of the magnitude and the phase of the first harmonic of the optimal control $(u^*(\theta,t))$ and the Galerkin-based control $(u_i^{\alpha^{i*}}(\theta_i,t))$.

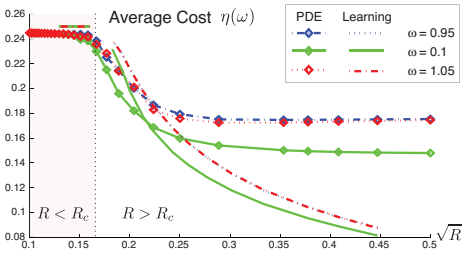

Fig. 7. Bifurcation diagram: the average cost as a function of $\frac{1}{\sqrt{R}}$.

of t chosen). At this mean value, the control is positive (negative) if $\omega < 1$ ($\omega > 1$).

The approximation improves as the variance of the noise σ^2 increases. This is qualitatively consistent with the estimate of the Bellman error given in Theorem 5.1.

For any control $u(\theta,t)$, the magnitude and phase of its first harmonic are computed as $\sqrt{([u]_c)^2+([u]_s)^2}$ and $\tan^{-1}([u]_c/[u]_s)\big|_{\omega=1} - \tan^{-1}([u]_c/[u]_s)\big|_{\omega}$, respectively. Fig. 6 provides a comparison of the optimal and Galerkin control laws in terms of these features. Qualitatively, the main difference is that the Galerkin-based control is more aggressive (larger magnitude and phase) than the optimal control.

Figure 7 compares the average cost $\eta(\omega)$ as a function of $1/\sqrt{R}$, with $\sigma^2 = 0.1$. When using optimal control, the data was obtained from a numerical solution of the coupled PDEs (7)-(8). For $R > R_c = 39.1$, the average cost is independent of frequency, $\eta(\omega) = \eta_0 = \frac{1}{4}$, which is consistent with the incoherence solution. For $R < R_c$ the average cost is reduced, and for such R the value of $\eta(\omega) < \eta_0$ depends upon the frequency ω. Its minimal value is attained uniquely when $\omega = 1$, which is the mean frequency under g.

With Galerkin-based control (18), the data was obtained as before, using a numerical simulation with $N = 200$ oscillators. The figure shows that the approximately optimal Galerkin-based control law (18) qualitatively captures the main features of the phase transition:

1. For $R > R_c$, $\eta(\omega) \approx \frac{1}{4}$ and the population is in incoherence. The slightly lower value of the average cost is due to finite number of oscillators in the simulation.
2. At $R = R_c$, there is a phase transition as predicted by Theorem 5.3. For $R < R_c$ the value of $\eta(\omega) < \eta_0$ depends upon the frequency ω. Its minimal value is attained uniquely when $\omega = 1$.

Consistent with the approximate formulation, the average cost is expected to be larger with the Galerkin-based control. This is indeed the case when R is much larger than R_c. For values of $R \approx R_c$, there are slight discrepancies. These are numerical artifacts because of the sensitive nature of the PDE solution in the vicinity of the bifurcation point, and because of the finite number of oscillators used in simulating the Galerkin-based control law.

REFERENCES

[1] H. Yin, P. G. Mehta, S. P. Meyn, and U. V. Shanbhag, "Synchronization of coupled oscillators is a game," To appear in Proc. of American Control Conference, 2010. [Online]. Available: http://mechse.illinois.edu/media/uploads/web_sites/67/files/sync_game_ymms.20100124.4b5d050d77dc00.29251135.pdf

[2] Y. Kuramoto, "Self-entrainment of a population of coupled non-linear oscillators," *In H. Araki, editor, International Symposium on Mathematical Problems in Theoretical Physics*, vol. 39, p. 420, 1975, ser. Lecture Notes in Physics, Springer.

[3] M. Huang, P. E. Caines, and R. P. Malhame, "Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ε-Nash equilibria," *IEEE Trans. Automat. Control*, vol. 52, no. 9, pp. 1560–1571, 2007.

[4] G. Y. Weintraub, L. Benkard, and B. V. Roy, "Oblivious equilibrium: A mean field approximation for large-scale dynamic games," in *Advances in Neural Information Processing Systems*, vol. 18. MIT Press, 2006.

[5] M. Steriade, D. A. McCormick, and T. J. Sejnowski, "Thalamocortical Oscillations in the Sleeping and Aroused Brain," *Science*, vol. 262, pp. 679–685, Oct. 1993.

[6] D. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Cambridge, Mass: Atena Scientific, 1996.

[7] P. G. Mehta and S. P. Meyn, "Q-learning and Pontryagin's Minimum Principle," in *Proc. of 48th IEEE Conference on Decision and Control*, December 2009, pp. 3598–3603.

[8] S. H. Strogatz and R. E. Mirollo, "Stability of incoherence in a population of coupled oscillators," *Journal of Statistical Physics*, vol. 63, pp. 613–635, May 1991.

[9] H. Kielhöfer, *Bifurcation Theory. An Introduction with Applications to PDEs*, ser. Applied Mathematical Sciences. Springer, 2003, no. 156.

[10] V. S. Borkar and S. P. Meyn, "The ODE method for convergence of stochastic approximation and reinforcement learning," *SIAM J. Control Optim.*, vol. 38, no. 2, pp. 447–469, 2000.