2020 59th IEEE Conference on Decision and Control (CDC)
Jeju Island, Republic of Korea, December 14-18, 2020

Kullback-Leibler Average of von Mises Distributions in Multi-Agent Systems

Cody Fan, Tsang-Kai Chang, and Ankur Mehta

***Abstract*— We derive the Kullback-Leibler average of von Mises distributions to fuse circular informations in multi-agent systems. Similar to the covariance intersection for Gaussian distributions, the derived fusion protocol does not require the independence among incoming distributions but maintains the estimation consistency, if those von Mises distributions represent estimates. Therefore, this fusion protocol is especially useful in the distributed estimation problem to avoid the over-confidence problem. For example, together with von Mises filters, the derived fusion protocol can estimate the dynamic circular term in a distributed manner. In addition, we apply this fusion protocol to determine the consensus of von Mises distributions over a network. Since the fusion protocol can be easily achieved by calculating the weighted average of the associated complex numbers, the corresponding convergent conditions of the consensus algorithm are then elegantly determined.**

I. Introduction

A multi-agent system is comprised of locally interacting agents equipped with sensing, processing and communication capabilities. Without a centralized console to coordinate all agents, the distributed multi-agent system may perform suboptimally. However, owning to the same reason, such distributed systems are far more robust and efficient than the centralized ones, and are applied in various applications including sensor networks and multi-robot systems.

Even though the algorithms on multi-agent systems have been studied for a while, most of the works focus on sensing and processing data in Euclidean space. In realistic applications, instead, there are a lot of circular data that have drastically different properties than the Euclidean one. Directly applying the algorithms for Euclidean data on circular data leads to consequential inconsistency. Therefore, the algorithms for circular data need to be designed explicitly by acknowledging the difference [1], [2].

While the sensing of the circular data can be largely solved by the von Mises filter [3], the processing and the communication of circular data in a multi-agent system are relatively overlooked. In particular, the fusion algorithm for circular data is essential to combine local informations, which enables the multi-agent systems to estimate the circular quantity in a distributed manner. In [4], a fusion algorithm is proposed for the von Mises distributions of circular data, but this algorithm requires the independence of the incoming informations. In fact, tracking the data dependency in a distributed network is very costly, but ignoring the dependency in the estimation scenario inevitably leads to the over-confidence problem.

In this paper, we derive a fusion method for von Mises distributions by obtaining its Kullback-Leibler (KL) average [5]. The exact formulation is already applied for the Gaussian distribution in Euclidean space, and the result is well-known as the covariance intersection [6]. The KL average actually gives a conservative fusion result with respect to the incoming informations, which guarantees the estimation consistency and avoids the over-confidence problem [7]. As a result, the derived KL average of von Mises distributions can be applied in multi-agent systems where the dependency of those von Mises distributions are not known. For computational consideration, the KL average of von Mises distribution can be easily calculated as the weighted average of the associated complex numbers, which provides an efficient formula to obtain the fusion result.

The proposed KL average for von Mises distributions can not only be applied on a single agent to fuse the incoming informations, but can also be implemented in networked agents to reach a consensus. We furthermore investigate the conditions of network topology and of weight selection to reach a consensus. Those conditions are well-studied for the real-number cases, both in theory [8] and in applications [9], [10]. Since the derived KL average of von Mises distribution can be represented in the complex number calculation and admits a linear form, the previous results in real-number cases can be directly and elegantly applied in our scenarios.

The contributions of this paper include:

- the fusion protocol of the KL average for von Mises distributions,
- the conservative fusion of von Mises estimates without knowing the dependency, and
- the consensus algorithm for von Mises distributions over a network.

We organize this paper as follows: The KL average and the general barycenter interpretation will be reviewed in Sec. II. The main contribution, the KL average of von Mises distributions, will be derived in Sec. III. In the next section, the KL average will be applied in a network scenario to reach a consensus of von Mises distributions. In Sec V, two simulations are presented to demonstrate the conservative fusion and the network consensus, respectively. We conclude this paper is the last section.

II. The Kullback-Leibler Average

In this section, we present the background of the KL average, which belongs to a more general barycenter formulation.

The authors are with the Electrical and Computer Engineering Department, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: trifire@ucla.edu; tsangkaichang@ucla.edu; mehtank@ucla.edu).

The first two authors contribute equally to this work.

To motivate the general barycenter formulation, we begin with two simple weighted average problems of real vectors and of circular terms.

For an Euclidean space $\mathbb{R}^l$, the weighted average of points $x_1, \dots, x_N$ with weights $w_1, \dots, w_N$ is the solution of the minimization problem

$$\min_{x \in \mathbb{R}^l} \sum_{n=1}^{N} w_n (x - x_n)^2, \tag{1}$$

where $w_1, \dots, w_N$ are positive weights summing to 1. We can write $d(x, x') = (x - x')^2$ as the difference measure between two points. To extend to circular terms, we can formulate the similar minimization problem by using the circular distance

$$d(\theta, \theta') = 1 - \cos(\theta - \theta') \tag{2}$$

between two circular terms θ and θ' [11]. To be explicit, we consider N circular values $\theta_1, \dots, \theta_N$, which all lie in $[0, 2\pi)$. The weighted average θ_c is then the solution of the minimization problem, or

$$\theta_c = \arg\min_{\theta} \sum_{n=1}^{N} w_n \left(1 - \cos(\theta - \theta_n)\right).$$

It is not difficult to see that

$$\exp(i\theta_c) = \sum_{n=1}^{N} w_n \exp(i\theta_n). \tag{3}$$

A. KL Average

With a slight abuse of the metric notation, one can choose the Kullback-Leibler (KL) divergence to quantify the difference between two distributions $p(x)$ and $q(x)$, with definition

$$D_{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx. \tag{4}$$

The corresponding weighted average with KL divergence for probability density functions $p_1, \dots, p_N$ and weights $w_1, \dots, w_N$ can then be defined as

$$p_c = \arg\inf_{p \in \mathcal{P}} \sum_{n=1}^{N} w_n D_{KL}(p||p_n), \tag{5}$$

where $\mathcal{P}$ denotes the set of all probability density functions. The solution is also called the KL average in [5]. Furthermore, the authors of [5] show that p_c has an explicit form

$$p_c(x) = \frac{\prod_{n=1}^{N} [p_n(x)]^{w_n}}{\int \prod_{n=1}^{N} [p_n(x)]^{w_n} \, dx}, \tag{6}$$

which is exactly the normalized geometric mean of the density functions $p_1, \dots, p_N$. The properties of the geometric mean of density functions are investigated in [7]. Focusing on the fusion between two densities, the authors show that the solution (6) is both conservative and effective in combining information from dependent sources.

When only the Gaussian distributions are considered, the solution of the KL average turns out to be identical as the one of the covariance intersection. The covariance intersection is first proposed to fuse several Gaussian distributions and to guarantee the estimation consistency without knowing the correlations between them [6]. Due to its simplicity and its consistency, the covariance intersection is applied in multiagent systems [12] and in robotics [13].

B. Wasserstein Barycenter

The aforementioned minimization problems actually belong to a more general barycenter formulation. The concept of barycenter is introduced in astronomy to find the center of mass of several orbiting bodies. By generalizing the orbiting bodies to probability measures, we can define the barycenter of several probability measures with properly defined distance metric. One choice of the distance metric on probability distributions is the Wasserstein distance. The resulting Wasserstein barycenter is well studied [14], and is applied in computer graphics [15] and machine learning [16], [17]. The explicit solution of the Wasserstein barycenter for Gaussian distributions is given in [18], but the application is not fully understood yet.

Even though the KL divergence can be interpreted as the "distance" between two probability distributions, there are some fundamental differences between the KL divergence and the Wasserstein distance. For example, the two probability distributions of the KL divergence should be defined on the same probability space, while the Wasserstein distance can be defined on two different probability spaces.

III. The KL Average of von Mises Distributions

We now explicitly solve the KL average of von Mises distribution in this section. The von Mises distribution, denoted by $vM(\mu, \kappa), \kappa > 0$, has probability density function

$$g(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\theta - \mu)), \quad 0 \le \theta < 2\pi, \tag{7}$$

where I_n is the modified Bessel function of the first kind and order n, which can be defined by

$$I_n(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \cos(n\theta) \exp(\kappa \cos\theta) \, d\theta. \tag{8}$$

In (7), μ is the mean direction and κ is known as the concentration parameter.

Theorem 1. *Given N von Mises distributions $p_1, \dots, p_N$ with the means $\mu_1, \dots, \mu_N$ and the concentration parameters $\kappa_1, \dots, \kappa_N$, together with N positive weights $w_1, \dots, w_N$ summing to 1, the barycentric solution of the KL divergence*

$$g_c = \arg\inf_{g \in \mathcal{P}_{vM}} \sum_{n=1}^{N} w_n D_{KL}(g||g_n), \tag{9}$$

with the mean μ_c and concentration parameter κ_c satisfying

$$\kappa_c \exp(i\mu_c) = \sum_{n=1}^{N} w_n \kappa_n \exp(i\mu_n). \tag{10}$$

When all incoming concentration parameters $\kappa_1, \dots, \kappa_N$ are equal, the solution of the von Mises consensus (10) degenerates to that of the simple weighted averaged of the circular terms (3). Even though the KL average is calculated over several distributions, Theorem 1 shows that the result is only the average over the associated complex numbers, which greatly simplifies the calculation. Moreover, as a KL average, the von Mises distribution from the solution (10) is a conservative fusion of the incoming distributions. If the incoming von Mises distributions are estimations of a circular value, the proposed fusion scheme guarantees the estimation consistency, since no information is doubly-counted during the fusion [7].

Before proving Theorem 1, we need the function $A(\kappa)$ defined as

$$A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}. \tag{11}$$

We also need the following property of $A(\kappa)$ to complete the proof of the main theorem.

Lemma 1. *For $\kappa > 0$, $A(\kappa) > 0$ and $A'(\kappa) > 0$.*

We provide the proof of this lemma in the appendix.

Proof of Theorem 1: Since the solution of (9) is also a von Mises distribution, which is characterized by two parameters, we then define an objective function as

$$\begin{aligned} J(\mu, \kappa) &= \sum_{n=1}^{N} w_n D_{KL}\left(g(\mu, \kappa) || g_n\right) \\ &= \sum_{n=1}^{N} w_n \left(\log \frac{I_0(\kappa_n)}{I_0(\kappa)} + A(\kappa)\left[\kappa - \kappa_n \cos(\mu - \mu_n)\right] \right), \end{aligned}$$

where the KL divergence between two von Mises distributions is given in [19]. The parameters μ_c and κ_c that minimize $J(\mu, \kappa)$ are then the mean and the concentration parameter of the distribution g_c, respectively.

Mean: We take the derivative with respect to μ and set to zero. Interchanging differentiation and integration is allowed according to the dominated convergence theorem.

$$\frac{\partial}{\partial \mu} J(\mu, \kappa) = \sum_{n=1}^{N} w_n A(\kappa)\left[\kappa_n \sin(\mu - \mu_n)\right] = 0,$$

where the solution is μ_c. By sorting out the equation,

$$\sin \mu_c \left(\sum_{n=1}^{N} w_n \kappa_n \cos \mu_n \right) = \cos \mu_c \left(\sum_{n=1}^{N} w_n \kappa_n \sin \mu_n \right).$$

This yields the result of the mean

$$\mu_c = \arg\left(\sum_{n=1}^{N} w_n \kappa_n \exp(i\mu_n) \right). \tag{12}$$

We now check the second derivative

$$\begin{aligned} \frac{\partial^2}{(\partial \mu)^2} J(\mu, \kappa)\bigg|_{\mu=\mu_c} &= A(\kappa) \sum_{n=1}^{N} w_n \left[\kappa_n \cos(\mu - \mu_n)\right] \bigg|_{\mu=\mu_c} \\ &= A(\kappa)\kappa_c \cos(\mu - \mu_c)\bigg|_{\mu=\mu_c} > 0, \end{aligned}$$

where

$$\kappa_c = \sqrt{\left(\sum_{n=1}^{N} w_n \kappa_n \cos \mu_n \right)^2 + \left(\sum_{n=1}^{N} w_n \kappa_n \sin \mu_n \right)^2}.$$

The κ_c here is only a constant, but we will prove that the exact κ_c is the minimizer of (9) in the following.

Concentration parameter: We now take the derivative of J with respect to κ.

$$\frac{\partial}{\partial \kappa} J(\mu, \kappa)\bigg|_{\mu=\mu_c} = A'(\kappa) \sum_{n=1}^{N} w_n \left[\kappa - \kappa_n \cos(\mu_c - \mu_n)\right]$$

with the identity

$$\frac{\partial}{\partial \kappa} I_0(\kappa) = I_1(\kappa).$$

We set the partial derivative to zero. Since $A'(\kappa) > 0$, the solution κ_c of the equation follows

$$\begin{aligned} \kappa_c &= \sum_{n=1}^{N} w_n \kappa_n \cos(\mu_c - \mu_n) \\ &= \sqrt{\left(\sum_{n=1}^{N} w_n \kappa_n \cos \mu_n \right)^2 + \left(\sum_{n=1}^{N} w_n \kappa_n \sin \mu_n \right)^2}. \end{aligned} \tag{13}$$

The second derivative gives

$$\begin{aligned} &\frac{\partial^2}{(\partial \kappa)^2} J(\mu, \kappa)\bigg|_{\mu=\mu_c, \kappa=\kappa_c} \\ &= A'(\kappa_c) + A''(\kappa_c) \sum_{n=1}^{N} w_n \left[\kappa_c - \kappa_n \cos(\mu_c - \mu_n)\right] \\ &= A'(\kappa_c) > 0. \end{aligned}$$

where the inequality comes from from Lemma 1. By expressing μ_c and κ_c in the complex form, we arrive at (10).

IV. The von Mises Consensus over a Network

In this section, we apply the KL average algorithm for von Mises distributions in (10) on a network of agents. We then show the condition on network topology as well as on weight selection required for those networked agents to reach a consensus. In order to characterize the network topology, we first introduce some graph notations.

A. Network Topology

A directed graph $\mathcal{G} = (V, E_{\mathcal{G}})$ is applied to characterize the information flow over a network. In the graph $\mathcal{G}$, the vertex set V contains all the nodes, and an edge $(k, j) \in E_{\mathcal{G}}$, $k \neq j$, represents that information can be sent from node k to node j. The *neighborhood* of node j is defined as $N_{\mathcal{G}}(j) = \{k | (k, j) \in E_{\mathcal{G}}\}$. In other words, the neighborhood of node j contains all nodes that send its information to node j. The *inclusive neighborhood* is defined by $N^*_{\mathcal{G}}(j) = N_{\mathcal{G}}(j) \cup \{j\}$.

B. Consensus over a Network

We can now consider M nodes that form a network, and use a graph $\mathcal{G}$ to describe the information flow among those nodes. Each node holds a von Mises distribution, and each node combine the incoming informations from other nodes by (10). To be specific, we use $w_{jk} > 0$ to stand for the weight that node j use for the distribution from node k. By definition, $\sum_{k \in N_{\mathcal{G}}^*(j)} w_{jk} = 1$.

Since Theorem 1 states that the KL average of von Mises distributions can be calculated by the weighted average of the corresponding complex numbers, each consensus iteration over the network can now be represented by the matrix multiplication. To be explicit, the complex vector

$$v_t = \begin{bmatrix} v_{1,t} \\ \vdots \\ v_{M,t} \end{bmatrix} = \begin{bmatrix} \kappa_{1,t} \exp(i\mu_{i,t}) \\ \vdots \\ \kappa_{M,t} \exp(i\mu_{M,t}) \end{bmatrix} \tag{14}$$

can represent the von Mises distributions over the network at time t, where j-th element of v_t represents the von Mises distribution at node j at time t. We can also put the weights into a nonnegative matrix W where

$$[W]_{jk} = \begin{cases} w_{jk}, & \text{if } k \in N_{\mathcal{G}}^*(j). \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Consequently, each consensus iteration can be expressed as

$$v_{t+1} = W v_t. \tag{16}$$

To discuss the asymptotic behavior of v_t, we need to first characterize that of W^t, which is well-studied by the Perron-Frobenius theorem [8].

Lemma 2 (Perron-Frobenius theorem for stochastic matrices)**.** *For an irreducible primitive stochastic matrix P,*

$$\lim_{t \to \infty} P^t = 1_M u_P^\mathsf{T}, \tag{17}$$

where u_P is the Perron vector of P^T.

By definition (15), W is not only a nonnegative matrix but it is also stochastic, or

$$1_M = W 1_M,$$

where 1_M is an all-1 vector of size M. We denote the Perron vector of W^T as u_W, and therefore $1_N^\mathsf{T} u_W = 1$. Based on the Perron-Frobenius theorem, we can directly have the following two propositions.

Proposition 1. *If W is irreducible and primitive,*

$$\lim_{t \to \infty} v_{j,t} = u_W^\mathsf{T} v_0. \tag{18}$$

Since we require the weights in Theorem 1 to be positive, the irreducibility and the primitivity in Proposition 1 can be fully determined by the underlying graph $\mathcal{G}$ [8]. In other words, as long as the network topology is given, the condition of Proposition 1 is determined, regardless of the weight selection. Proposition 1 states that, even though the proposed fusion scheme is conservative, with the irreducibility and the primitivity of the matrix W, the concentration parameters of all von Mises distributions will not keep decreasing but converge to the same value. Since the elements of the Perron vector u are positive and summed to 1, the convergent value $u_W^\mathsf{T} v_0$ can be considered as the weighted averaged of the initial vector v_0.

As the convergence of the consensus is fully determined by the network topology, the converged consensus can be determined by the weight selection.

Definition 1. *A $M \times M$ matrix P is doubly-stochastic if $1_M = P 1_M$ and $1_M = P^\mathsf{T} 1_M$.*

Proposition 2. *If W is irreducible, primitive, and doubly-stochastic,*

$$\lim_{t \to \infty} v_{j,t} = \frac{1}{M} \sum_{k=1}^{M} v_{k,0}. \tag{19}$$

Proposition 2 states that under those conditions, the distributions of all nodes converge to the same distribution, which is the average of the initial distributions. In addition to the conditions from the network topology, Proposition 2 requires the weight selection to ensure that W is doubly-stochastic. The condition of doubly-stochastic matrices over a graph is well-investigated [20]. One of the weight choices to ensure that W is doubly-stochastic is given by the Metropolis weights:

$$w_{jk} = \begin{cases} \frac{1}{\max\left(|N_{\mathcal{G}}^*(j)|, |N_{\mathcal{G}}^*(k)|\right)}, & \text{if } k \in N_{\mathcal{G}}(j). \\ 1 - \sum_{k \neq j} w_{jk}, & \text{if } k = j. \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

The consensus weight selection is another weight choice that ensures that W is doubly-stochastic, which is given by

$$w_{jk} = \begin{cases} \epsilon, & \text{if } k \in N_{\mathcal{G}}(j). \\ 1 - \epsilon |N_{\mathcal{G}}(j)|, & \text{if } k = j. \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

For the consensus weights, ϵ should small enough to ensure that $w_{jk} > 0$ for all j, k.

The convergence rate of both propositions is determined by the second largest eigenvalue of W [8]. The authors in [9] even optimize the convergence rate to reach the distributed average. However, such optimization remains difficult to be achieved in a distributed network. Finally, even though the theory of both the propositions is well-known in the linear algebra [8] and in the literature of consensus problems [9], [10], we are not fusing real numbers but actually von Mises distributions in this paper.

V. Simulation

In this section, we show that the KL average for von Mises distributions is able to fuse dependent distributions in the first simulation. We consider two estimators that use the overlapping observation data, and the direct fusion of the estimates leads to the over-confidence problem. On the contrary, the proposed fusion algorithm can avoid the over-confidence problem, and maintain the estimation consistency.

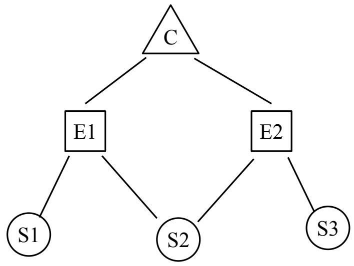

Fig. 1. The topology of the fusion between 2 dependent von Mises filters. S1, S2, and S3 are the sensors that observe θ_t by (23). E1 and E2 are the von Mises filters. The von Mises filters E1 and E2 operate the time and the observation updates for 20 times, then send its own estimate to the fusion center C for the final fused result. For the KL average fusion, the weights are fixed as $w_1 = 0.6$ and $w_2 = 0.4$.

In the second simulation, we show that the proposed fusion protocol can reach the consensus over a distributed network, and we demonstrate the effect of the weight selection on the resulting consensus.

A. von Mises Filters with Overlapping Sensors

To demonstrate that the proposed KL average can fuse dependent distributions, we consider a scenario with 2 estimators. These two estimators try to estimate a dynamic circular value θ_t that evolves according to

$$\theta_t = \theta_{t-1} + u_{t-1} + w_{t-1}, \tag{22}$$

where u_{t-1} is the input and w_{t-1} is the process noise modeled by $vM(0, \kappa_w)$. These two estimators do not directly observe θ_t, but receive the observation data from distributed sensors, which observe θ_t by

$$o_{k,t} = \theta_t + \nu_{k,t}, \tag{23}$$

where $o_{k,t}$ is the observation from sensor k. In (23), $\nu_{k,t}$ is the observation noise of sensor k, and is modeled by $vM(0, \kappa_{\nu,k})$. All process noises and observation noises are independent of the rest of the system.

We consider a system with 2 estimators and 3 sensors, with the system topology in Fig. 1. We choose a constant input $u_t = 0.7$ with $\kappa_w = 7$. For the sensor parameters, $\kappa_{\nu,1} = 3.3$, $\kappa_{\nu,2} = 4.4$, and $\kappa_{\nu,3} = 2.2$. The two estimators E1 and E2 use the von Mises filters to dynamically estimate θ_t, but the estimates are dependent since both estimators use the observation from S2. These two estimators run 20 iterations of both the time update and the observation update, and then send their own estimate to the fusion center C.

We compare various fusion protocols at the fusion center C, including our KL average fusion protocol and the fusion equation assuming independence from [3], in Fig. 2. As a benchmark, we also plot the optimal fusion where the estimator can directly obtain the raw observation data, and fuse them. However, the optimal fusion is not practical in distributed networks.

In Fig. 2, the concentration parameter of the independence fusion is larger than that of the optimal fusion. Therefore, the independence fusion is definitely over-confident, which

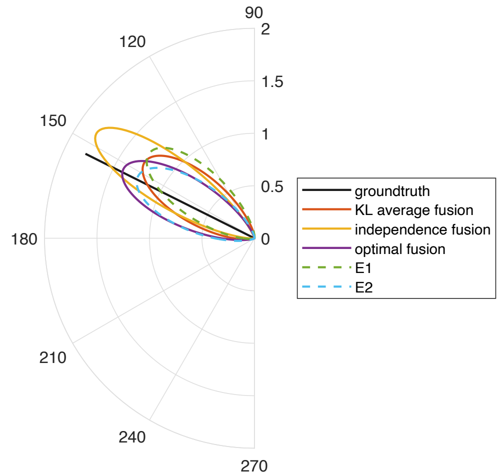

Fig. 2. The density function of von Mises distributions under different fusion methods. The independence fusion produces over-confident estimates, since its concentration parameter is larger than that of the optimal fusion. On the contrary, the KL average fusion gives reasonable estimation without even knowing the exact independent observation data, which shows its efficiency in distributed networks.

TABLE I
THE STATISTICS OF THE CONCENTRATION PARAMETERS κ UNDER DIFFERENT FUSION METHODS OVER 200 TRIALS

Fusion method	Mean	Standard deviation
Optimal fusion	13.467	1.210
KL average fusion	11.042	0.771
Independence fusion [3]	21.825	1.491

effects the estimation reliability. We further execute the identical simulation for 200 trials, and summarize the statistics of the concentration parameters of these three fusion methods in Table I. The concentration parameters from the independence fusion are significantly higher that those of the optimal fusion. On the contrary, the result from the derived KL average gives reasonable concentration parameter. In Table I, the concentration parameters of the KL average are smaller than those of the optimal fusion, since the proposed KL average fusion is a conservative fusion scheme. Moreover, no independence is required in the derived KL average fusion, which shows its efficiency in distributed networks.

B. The von Mises Consensus over a Network

We simulate the consensus problem discussed in Sec. IV. The graph $\mathcal{G}$ is depicted in Fig. 3, and we consider the edges are bi-directional. The initial distribution of each node is summarized in Table II.

When the graph is given, the irreducibility and primitivity of the weight matrix W are also determined, since we require that the weights on all incoming informations are positive. We first pick the weights equally for all incoming informations, or

$$w_{jk} = \begin{cases} \frac{1}{|N^*_{\mathcal{G}}(j)|}, & \text{if } k \in N^*_{\mathcal{G}}(j). \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

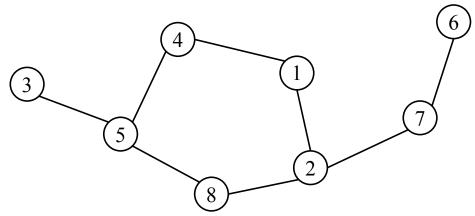

Fig. 3. The network topology of nodes to reach a consensus. The graph associated with this topology guarantees the irreducibility and primitivity of the weight matrix W.

TABLE II

INITIAL CONFIGURATION OF EACH NODE

Node	Mean μ	Concentration parameter κ
1	2	10
2	1	5
3	0	7
4	-2	2
5	1.5	4
6	3	15
7	-3	6
8	2.2	9

In Fig. 4, we can see that the distribution of each node converges to the same distribution with only local information exchange, since the condition of Proposition 1 is satisfied. The converged distribution can be explicitly determined by (18), which is the weighted average of all the initial distributions.

To further ensure that the convergent distribution is the exact average of the initial distributions, we have to ensure that W is doubly-stochastic. With the same network topology and the initial distributions, we simulate the consensus with the Metropolis weights (20), and plot the result in Fig. 5. As Proposition 2 suggests, all the distributions converge to the same distribution, and the corresponding complex number of this distribution is exactly the average of all the associated complex numbers of the initial distributions.

In addition to the weight selection, Fig. 4 and Fig. 5 also show that even though the derived KL average is conservative, as long as the graph-induced weight matrix is irreducible and primitive, the distribution of each node can reach an equilibrium with convergent concentration parameter. Therefore, with properly designed network topology, we can apply the derive KL average to fuse von Mises distributions, and the resulting distributions will converge.

VI. CONCLUSION

We study the KL average of von Mises distributions in this paper. The derived fusion formula can combine several von Mises distributions in simple matrix operation. Furthermore, if those von Mises distributions represent estimation distributions, the derived fusion protocol ensures the estimation consistency, without knowing the exact dependency among those distributions. Therefore, the derived fusion formula is especially useful in multi-agent systems and distributed networks.

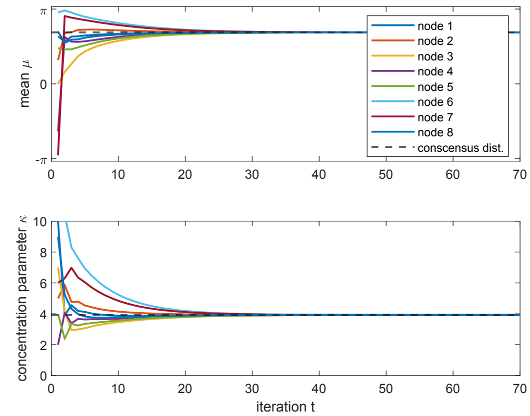

Fig. 4. The dynamics of von Mises distributions in a network with irreducible and primitive W. The weights are chosen equally for all incoming information. According to Proposition 1, the distribution in each node will converges to the same distribution, which is determined by the Perron vector u and the initial distributions. The mean and the concentration parameter of the consensus distribution are 2.162 and 3.921, respectively.

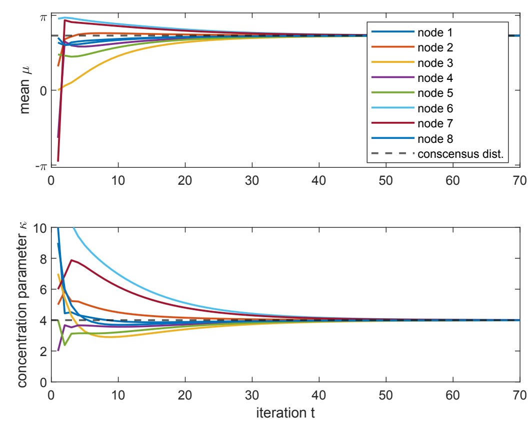

Fig. 5. The dynamics of von Mises distributions in a network with irreducible, primitive and doubly-stochastic W. The weights are the Metropolis weights to ensure the doubly-stochasticity of W. All the individual distributions converge to the distribution given by (19). The mean and the concentration parameter of the consensus distribution are 2.292 and 3.996, respectively.

Theoretically, this work provides a geometric interpretation and manipulation towards the consensus problem. Moving towards more general distributions, including higher dimensional directional distributions as well as the exponential family, we believe that information geometry would be an useful perspective in the ongoing investigation [21]. In the future, we will also investigate the joint estimation scheme with the covariance intersection and the proposed fusion protocol.

APPENDIX: PROOF OF LEMMA 1

Lemma 3 (Modified Bessel functions [22])**.** *For real* $p \geq 0$,

$$\frac{-p+\sqrt{p^2+\kappa^2}}{\kappa} < \frac{I_p(\kappa)}{I_{p-1}(\kappa)}. \tag{25}$$

Also, for $p \geq 1/2$, the inequality $I_p(\kappa)/I_{p-1}(\kappa) < 1$ holds.

By Lemma 3, we have $A(\kappa) < 1$ and for $\kappa > 0$,

$$A(\kappa) > \frac{\sqrt{\kappa^2+1}-1}{\kappa} > 0, \tag{26}$$

which gives the lower bound of $A(\kappa)$. In [11], we have the recurrent equation

$$A'(\kappa) = 1 - A(\kappa)\left(A(\kappa) + \frac{1}{\kappa}\right). \tag{27}$$

By combining (26) and (27), $A'(\kappa) > 0$ is obtained.

REFERENCES

[1] Y. Ban, X. Alameda-Pineda, C. Evers, and R. Horaud, "Tracking multiple audio sources with the von Mises distribution and variational EM," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 798–802, Jun. 2019.

[2] I. Marković and I. Petrović, "Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering," *Robotics and Autonomous Systems*, vol. 58, no. 11, pp. 1185–1196, Nov. 2010.

[3] M. Azmani, S. Reboul, J.-B. Choquel, and M. Benjelloun, "A recursive fusion filter for angular data," in *Proceeding of the IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec. 2009, pp. 882–887.

[4] G. Stienne, S. Reboul, M. Azmani, J. B. Choquel, and M. Benjelloun, "A multi-temporal multi-sensor circular fusion filter," *Information Fusion*, vol. 18, pp. 86–100, Jul. 2014.

[5] G. Battistelli and L. Chisci, "Kullback-Leibler average, consensus on probability densities, and distributed state estimation with guaranteed stability," *Automatica*, vol. 50, no. 3, pp. 707–718, Mar. 2014.

[6] S. Julier and J. Uhlmann, "A non-divergent estimation algorithm in the presence of unknown correlations," in *Proceedings of the American Control Conference*, vol. 4, Jun. 1997, pp. 2369–2373.

[7] T. Bailey, S. Julier, and G. Agamennoni, "On conservative fusion of information with unknown non-Gaussian dependence," in *Proceeding of the International Conference on Information Fusion (FUSION)*, Jul. 2012, pp. 1876–1883.

[8] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. USA: Society for Industrial and Applied Mathematics, 2000.

[9] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, Sep. 2004.

[10] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, Jan. 2007.

[11] K. V. Mardia and P. E. Jupp, *Directional Statistics*. New York: Wiley, 1999.

[12] T.-K. Chang and A. Mehta, "Control-theoretical and topological analysis of covariance intersection based distributed Kalman filter," *IEEE Control Systems Letters*, vol. 2, no. 4, pp. 665–670, Oct. 2018.

[13] ——, "Optimal scheduling for resource-constrained multirobot cooperative localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1552–1559, Jul. 2018.

[14] M. Agueh and G. Carlier, "Barycenters in the Wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, Jan. 2011.

[15] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas, "Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains," *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 66:1–66:11, Jul. 2015.

[16] P. Dvurechenskii, D. Dvinskikh, A. Gasnikov, C. Uribe, and A. Nedich, "Decentralize and randomize: Faster algorithm for Wasserstein barycenters," in *Advances in Neural Information Processing Systems (NeurIPS)*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 10 760–10 770.

[17] C. A. Uribe, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and A. Nedi, "Distributed computation of Wasserstein barycenters over networks," in *Proceeding of the IEEE Conference on Decision and Control (CDC)*, Dec. 2018, pp. 6544–6549.

[18] A. N. Bishop, "Information fusion via the Wasserstein barycenter in the space of probability measures: Direct fusion of empirical measures and Gaussian fusion with unknown correlation," in *Proceedings of the International Conference on Information Fusion (FUSION)*, Jul. 2014, pp. 1–7.

[19] G. Kurz and U. D. Hanebeck, "Trigonometric moment matching and minimization of the Kullback-Leibler divergence," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 4, pp. 3480–3484, Oct. 2015.

[20] B. Gharesifard and J. Cortés, "When does a digraph admit a doubly stochastic adjacency matrix?" in *Proceedings of the 2010 American Control Conference*, Jun. 2010, pp. 2440–2445, iSSN: 2378-5861.

[21] S.-i. Amari, *Information Geometry and Its Applications*. Springer Japan, 2016.

[22] A. Laforgia and P. Natalini, "Some inequalities for modified Bessel functions," *Journal of Inequalities and Applications*, vol. 2010, no. 1, p. 253035, Dec. 2010.