Gossip of Statistical Observations using Orthogonal Polynomials

Raphaël Berthier, Francis Bach, Pierre Gaillard

GOSSIP OF STATISTICAL OBSERVATIONS USING ORTHOGONAL POLYNOMIALS

RAPHAËL BERTHIER, FRANCIS BACH, PIERRE GAILLARD

INRIA - Département d'informatique de l'ENS
Ecole normale supérieure, CNRS, INRIA
PSL Research University, 75005 Paris, France

Abstract. Consider a network of agents connected by communication links, where each agent holds a real value. The gossip problem consists in estimating the average of the values diffused in the network in a distributed manner. Current techniques for gossiping are designed to deal with worst-case scenarios, which is irrelevant in applications to distributed statistical learning and denoising in sensor networks. We design second-order gossip methods tailor-made for the case where the real values are i.i.d. samples from the same distribution. In some regular network structures, we are able to prove optimality of our methods, and simulations suggest that they are efficient in a wide range of random networks. Our approach of gossip stems from a new acceleration framework using the family of orthogonal polynomials with respect to the spectral measure of the network graph.

1. Introduction

The averaging problem, or gossip problem, is a fundamental primitive of distributed algorithms. Given a network composed of agents and communication links between them, we assign to each agent a real value, called the observation. The goal is to design an iterative communication procedure allowing each agent to know the average of the initial observations in the network, as quickly as possible. This problem appears when distributing an optimization or learning task on several devices.

The landmark paper [BGPS06] suggests the natural following protocol to solve the averaging problem: at each iteration, each agent replaces his current observation by the average of the observations of its neighbors in the network. We will refer to this method in the following by the term *simple gossip*. The paper [BGPS06] proves the linear convergence of the observations to their average.

However, the rate of the linear convergence was shown to worsen significantly in many networks of interest as the size of the network increases. More precisely, define the diameter D of the network as the largest number of communication links needed to connect any two agents. While obviously, D steps of averaging are needed for any gossip method to spread information in the network, the simple gossip method may require up to $\Theta(D^2)$ communication steps to estimate the average, as for instance in the two-dimensional grid [RT17]. To reach the $O(D)$ bound, a diverse set of ideas were proposed, including second-order recursions [CSY06, RT17], message passing algorithms [MR05], lifted Markov chain techniques [Sha09] or inspiration arising from advection-diffusion processes [SGB10]. To our knowledge, all of these methods assume that the agents hold additional information about the graph, such as its spectral gap.

It is worth noting that all the aforementioned literature analyzes the gossip algorithms in a worst-case setting, meaning that no prior about the observations is assumed. However, in many applications, there is a statistical structure one should exploit. In sensor networks, observations are measurements of the environment corrupted by noise. The purpose of the gossip algorithm is to average observations to get a better estimate of the ground truth. Gossip algorithms are also used as building blocks in distributed statistical learning problems such as distributed optimization (see [NO09, SBB$^+$17, S$^+$14, RNV10, DAW12, CS12]) or distributed bandit algorithms (see [SBFH$^+$13,

E-mail address: raphael.berthier@inria.fr, francis.bach@inria.fr, pierre.gaillard@inria.fr.

LSL16, KSS16]). All of these problems have a statistical structure that simplifies the underlying gossip problem. For instance, in sensor networks, good estimates of the mean may not require using observations from nodes extremely far in the network.

This remark motivates that in this paper, we consider the special case where the observations are independent identically distributed (i.i.d.) random variables $\xi \sim \nu$, and the goal is to estimate the statistical mean $\mathbb{E}[\nu]$ of these random variables (and not the average). A more precise definition of this framework, called *statistical gossip*, is given in Section 2. We then make the following contributions to the statistical gossip problem.

Contributions. In Section 3, we use a parallel with random walks to show that in this framework too, simple gossip is suboptimal in networks of interest. The bulk of the paper focuses on accelerating the statistical gossip. In Section 4, we describe an acceleration framework that, given the spectral measure of the network graph, and the associated family of orthogonal polynomials, derives an accelerated algorithm suited for that network.

However, it is unrealistic in practice to assume that we know the whole spectrum of the graph, let alone that we are able to compute the corresponding orthogonal polynomials. We argue, with theoretical results and simulations, that one can wisely approximate the spectral measure by a simpler one with known orthogonal polynomials without any significant decrease in performance of the accelerated algorithm.

In d-regular graphs, and when the approximate spectral measure is the one of the infinite d-regular tree, we show in Section 5 that the resulting algorithm is equivalent to the message passing algorithm of [MR05], an optimal method on trees. In grids $\mathbb{Z}^d$, $d \geqslant 2$, we give an approximate spectral measure that gives asymptotically optimal results (see Section 6).

Section 7 provides simulations showing that the algorithm for grids is still efficient on random geometric graphs and the algorithm for trees is still efficient on random d-regular graphs. These results support that only little information about the spectral measure is needed to build efficient algorithms from our acceleration framework.

2. Problem setting

A network of agents is modeled by a undirected graph $G = (V, E)$, where V is the set of vertices of the graph, or agents, and E the set of edges, or communication links. Note that here, V can be either finite or infinite, but we will always assume G to be locally finite, meaning that for all $v \in V$, the degree $\deg v$ (i.e., the number of neighbors of v in the graph) is finite. The case where V is infinite must be thought as the large graph limit where we neglect border effects. Although the infinite network modelling might be surprising, in this paper it provides insights on the gossip problem, even in the finite case.

We consider a probability law ν on $\mathbb{R}$, and $\mu = \int_{\mathbb{R}} \xi \, d\nu(\xi)$ its statistical mean. Each agent $v \in V$ is given a sample from ν:

$$\xi_v, v \in V \underset{\text{i.i.d.}}{\sim} \nu \, .$$

A fundamental operation to estimate the mean μ consists in averaging the observations of neighbors in the network. We formalize this notion using a gossip matrix.

Definition 2.1. *A* gossip matrix $W = (W_{v,w})_{v,w \in V}$ *on the graph G is a matrix with entries indexed by the vertices of the graph satisfying the following properties:*

- *W is nonnegative: for all $v, w \in V$, $W_{v,w} \geqslant 0$.*
- *W is supported by the graph G: for all distinct vertices v, w such that $W_{v,w} > 0$, $\{v, w\}$ must be an edge of G.*
- *W is stochastic: for all $v \in V$, $\sum_{w \in V} W_{v,w} = 1$.*
- *W is symmetric: for all $v, w \in V$, $W_{v,w} = W_{w,v}$.*

If W is a gossip matrix and $x = (x_v)_{v \in V}$ is a set of values stored by the agents v, the product Wx is interpreted as the computation by each agent v of a weighted average of the values x_w of its neighbors w in the graph (and of its own value x_v). This average is computed simultaneously for all agents v; indeed in this paper we deal only with *synchronous* gossip. Note that we do not need the symmetry assumption on W to interpret W as an averaging operation. This assumption is usual in gossip frameworks as it allows one to use the spectral theory for W, on which our analysis relies heavily. It appears, for instance, in the works [BGPS06, CSY06, RT17].

In a d-regular graph G ($\forall v, \deg v = d$), a typical gossip matrix is $W = A(G)/d = (\mathbf{1}_{\{\{v,w\}\in E\}}/d)_{v,w\in V}$ where $A(G)$ is the adjacency matrix of the graph. We will call this matrix the simple gossip matrix as it is the transition matrix of the simple random walk on G.

Simple gossip. Simple gossip is a natural algorithm solving the gossip problem that consists in averaging repeatedly values in the graph. More precisely, we choose a gossip matrix W on the graph G, initialize $\hat{\mu}^0 = \xi = (\xi_v)_{v\in V}$, and at each communication round t, compute

$$\hat{\mu}^{t+1} = W\hat{\mu}^t\,. \tag{2.1}$$

We can rewrite this iteration as $\hat{\mu}^t = W^t\xi$. Simple gossip builds unbiased estimators $\hat{\mu}_v^t$ of the mean μ for each agent v. Indeed, W is stochastic, i.e., $W\mathbf{1} = \mathbf{1}$, thus

$$\mathbb{E}[\hat{\mu}^t] = W^t\mathbb{E}[\xi] = \mu W^t\mathbf{1} = \mu\mathbf{1}\,. \tag{2.2}$$

Note that in the last equation, we used the notation $.^t$ to denote both the index of $\hat{\mu}$ and the power of the square matrix W. We will frequently make use of the indexation $.^t$ when vectors indexed by the vertices (or the edges) also depend on time.

Polynomial gossip. Several acceleration schemes of gossip [CSY06, RT17] store some past iterates to compute higher-order recursions (that thus depend on powers of W). Without specifying the recursion for now, we define a polynomial gossip method as any method combining the past iterates of the simple gossip method:

$$\hat{\mu}^t = P_t(W)\xi\,, \tag{2.3}$$

where $P_t(W)$ is a polynomial of degree smaller or equal to t satisfying $P_t(1) = 1$. As in (2.2), the constraint $P_t(1) = 1$ ensures that the estimators $\hat{\mu}^t$ in (2.3) are unbiased. The constraint $\deg P_t \leqslant t$ ensures that the estimator $\hat{\mu}^t$ can be computed in t time steps. Simple gossip corresponds to the particular case of the polynomial $P_t(\lambda) = \lambda^t$.

The rest of this paper is devoted to the study of the variance of the unbiased estimators of simple gossip (2.1) and polynomial gossip (2.3), as well as the design of polynomials P_t that ensure small variances of the estimators. The next result is both a technical lemma linking the variance $\operatorname{var}\hat{\mu}^t$ to the polynomials P_t and a lower bound that we aim to match.

Proposition 2.2. *Let $\hat{\mu}^t$ be the unbiased estimator defined in* (2.3)*. Let $v \in V$. Denote $e_v = (\mathbf{1}_{\{w=v\}})_{w\in V} \in \mathbb{R}^V$. Then*

$$\frac{\operatorname{var}\hat{\mu}_v^t}{\operatorname{var}\nu} = \|P_t(W)e_v\|^2_{\ell^2(V)} = \left\langle e_v, P_t(W)^2 e_v\right\rangle_{\ell^2(V)}\,, \qquad \frac{\operatorname{var}\hat{\mu}_v^t}{\operatorname{var}\nu} \geqslant \frac{1}{|B_v(t)|}\,, \tag{2.4}$$

where $B_v(t) = \{w \in V \mid d(v,w) \leqslant t\}$ denotes the ball of radius t, centered in v, for the shortest path distance $d(.,.)$ in the graph G.

Definition 2.3 (Optimality)**.** *In the case where $\operatorname{var}\hat{\mu}_v^t/\operatorname{var}\nu = \Theta(1/|B_v(t)|)$ as $t \to \infty$, we will say that the gossip method is asymptotically optimal (for v). In the stronger case where $\operatorname{var}\hat{\mu}_v^t/\operatorname{var}\nu = 1/|B_v(t)|$ for all t, we will say that the gossip method is exactly optimal.*

See Appendix A.1 for a proof of Proposition 2.2. Note that the above definition of asymptotic optimality is of interest only if the graph G is infinite. Indeed, if G is finite, $\operatorname{var}\hat{\mu}_v^t/\operatorname{var}\nu$ and $1/|B_v(t)|$ are both lower bounded by $1/|V|$ and upper bounded by 1, thus $\operatorname{var}\hat{\mu}_v^t/\operatorname{var}\nu = \Theta(1/|B_v(t)|)$ as $t \to \infty$.

3. Sub-optimality of simple gossip

For the reader's convenience, we recall that simple gossip is defined by the iteration

$$\hat{\mu}^0 = \xi = (\xi_v)_{v\in V}\,, \quad \hat{\mu}^{t+1} = W\hat{\mu}^t\,. \tag{3.1}$$

In this section, we derive a parallel between simple gossip and a random walk on V to show the sub-optimality of simple gossip. Indeed, as W is a stochastic matrix, it can be seen as a transition matrix of a random walk on V. Denote $V_t \in V$ the random walk with transition matrix W, and $\mathbb{E}_v$ (resp. $\mathbb{P}_v$) the corresponding expectation (resp. probability) when it is started from v.

Proposition 3.1. *Let $\hat{\mu}^t$ be the estimates of the simple gossip method* (3.1)*. Then*

$$\frac{\operatorname{var}\hat{\mu}_v^t}{\operatorname{var}\nu} = \left\langle e_v, W^{2t}e_v\right\rangle_{\ell^2(V)} = \mathbb{P}_v(V_{2t} = v)\,.$$

Using estimates of return probabilities of random walks provided by the probabilistic literature, we derive two applications of Proposition 3.1 showing that simple gossip is suboptimal.

Grids $\mathbb{Z}^d$. Let $\omega = (\omega_v)_{v\in\mathbb{Z}^d}$ be a centered probability distribution over $\mathbb{Z}^d$ with compact support such that the random walk on $\mathbb{Z}^d$ with law of increments ω is irreducible and aperiodic. Let $G(\omega) = (\mathbb{Z}^d, E)$ with $E = \{\{v, w\} \mid \omega_{w-v} \neq 0\}$ be the associated graph, and $W_{v,w} = \omega_{w-v}$ the associated gossip matrix. The local central limit theorem [LL10, chapter 2] gives the asymptotic equivalent of the return probability of the random walk on $\mathbb{Z}^d$ with increments of law ω. Combining with Proposition 3.1, we get for all $v \in \mathbb{Z}^d$,

$$\frac{\operatorname{var}\hat{\mu}_v^t}{\operatorname{var}\nu} \sim \frac{1}{2^d\pi^{d/2}\sqrt{\det Q}}\frac{1}{t^{d/2}} \qquad \text{as } t\to\infty\,, \tag{3.2}$$

where $Q \in \mathbb{R}^{d\times d}$ is the covariance matrix of ω. Note that in $G(\omega)$, we have the growth $|B_v(t)| = \Theta(t^d)$ as $t \to \infty$. Thus simple gossip is suboptimal on a wide range of "grid-like" graphs. For instance, if we choose $\omega_v = 1/(2d)$ for v being the $2d$ points in $\mathbb{Z}^d$ closest to 0, and $\omega_w = 0$ elsewhere, we get suboptimality of simple gossip on the d-dimensional grid, when $W = A(G(\omega))/(2d)$. This is the graph we will refer to as $G = \mathbb{Z}^d$ in rest of this paper.

Trees $\mathbb{T}_d$. ($d \geqslant 3$) Let $W = A(\mathbb{T}_d)/d$ be the gossip matrix on the infinite d-regular tree $\mathbb{T}_d$. The associated random walk on $\mathbb{T}_d$ is the simple random walk on $\mathbb{T}_d$, whose return probabilities are studied in [Saw78]. Combining their result with Proposition 3.1, we get

$$\frac{\operatorname{var}\hat{\mu}_v^t}{\operatorname{var}\nu} = \Theta\left(\frac{1}{|B_v(t)|}\left(2-\frac{2}{d}\right)^{2t}\frac{1}{t^{3/2}}\right) \qquad \text{as } t\to\infty\,.$$

Thus the simple gossip method is suboptimal on $\mathbb{T}_d$.

4. Design of best polynomial gossip algorithms

We now turn to the design of efficient estimators of the form $\hat{\mu}^t = P_t(W)\xi$. An important result of this section is that the best estimators of this form can be computed in an online fashion as they result from a second-order recurrence relation.

The analysis of gossip on finite graphs usually relies on the spectral theorem applied to an auto-adjoint finite matrix W. Here we will need the equivalent result on possibly infinite graphs. Fix $v \in V$. As W is an auto-adjoint operator, bounded by 1, acting on $\ell^2(V)$, there exists a unique positive measure $\sigma = \sigma(G, W, v)$ on $[-1, 1]$, called the *spectral measure*, such that for all polynomial P,

$$\langle e_v, P(W)e_v\rangle_{\ell^2(V)} = \int_{-1}^{1} P(\lambda)\mathrm{d}\sigma(\lambda)\,.$$

For a deeper presentation of spectral graph theory, see [MW89] and references therein. Note that when the graph G is finite, it is easy to check that the spectral measure is the discrete measure $\sigma(G, W, v) = \sum_{i=1}^{n}(u_v^i)^2\delta_{\lambda_i}$ where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of W and $u^1, \dots, u^n$ are the associated eigenvectors. When G is d-regular and we choose to take the simple gossip matrix $W = A(G)/d$, we will omit the dependence on W in $\sigma(G, v)$. Further, when the spectral measure does not depend on v, we will simply write $\sigma(G)$.

It follows from (2.4) that

$$\frac{\operatorname{var}\hat{\mu}_v^t}{\operatorname{var}\nu} = \left\langle e_v, P_t(W)^2 e_v\right\rangle_{\ell^2(V)} = \int_{-1}^{1} P_t(\lambda)^2\mathrm{d}\sigma(\lambda)\,.$$

The polynomial P_t^σ minimizing the variance $\operatorname{var}\hat{\mu}_v^t$ satisfies

$$P_t^\sigma \in \operatorname*{argmin}_{P(1)=1,\,\deg P\leqslant t}\int_{-1}^{1} P(\lambda)^2\mathrm{d}\sigma(\lambda)\,. \tag{4.1}$$

Two examples are given in Figure 1. Note how the polynomial P_t^σ adapts to be small where σ has mass while satisfying the constraint $P_t^\sigma(1) = 1$. The corresponding performance $\operatorname{var}(P_t^\sigma(W)\xi)/\operatorname{var}\nu$ is $\Lambda_t(\sigma, 1)$ where

$$\Lambda_t(\sigma, \lambda_0) = \min_{P(\lambda_0)=1,\,\deg P\leqslant t}\int_{-1}^{1} P(\lambda)^2\mathrm{d}\sigma(\lambda)$$

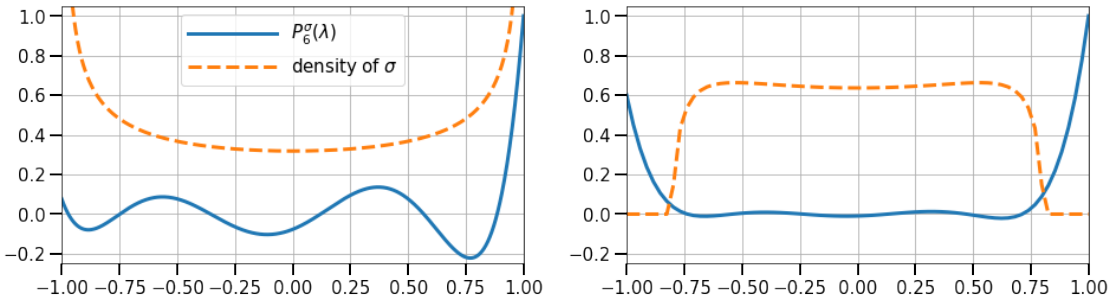


FIGURE 1. Two measures σ with densities and the corresponding optimal polynomial P_6^σ as defined in (4.1). Left: $\sigma(\mathrm{d}\lambda) = \sigma(\mathbb{Z})(\mathrm{d}\lambda) = \mathrm{d}\lambda/(\pi\sqrt{1-\lambda^2})$. Right: $\sigma(\mathrm{d}\lambda) = \sigma(\mathbb{T}_5)(\mathrm{d}\lambda)$ as explicited in (5.3).

is the Christoffel function associated with the measure σ, whose asymptotic as $t \to \infty$ is studied for instance in [Tot00, Nev79, Nev86, MNT91].

Remark 4.1. *One may not want to minimize the variance* $\operatorname{var}\hat{\mu}_v^t$ *at a single vertex* v *but a weighted average* $\sum_{w\in V}\gamma_w \operatorname{var}\hat{\mu}_w^t$. *This can easily be done by replacing* $\sigma = \sigma(G,W,v)$ *with* $\sigma = \sum_{w\in V}\gamma_w\sigma(G,W,w)$. *The discussions of this section and the following sections still apply in this case. However, to simplify the matter, we present the case* $\sigma = \sigma(G,W,v)$.

We will now show that the sequence of best polynomials $P_0^\sigma, P_1^\sigma, P_2^\sigma, \dots$ can be computed online as the result of a second order recursion, which leads to a second order gossip method, whose coefficients depend on σ. As noted in [CSY06], having estimators $\hat{\mu}^t$ that satisfy a low-order recurrence relation is valuable as it ensures that they can be computed online with limited memory cost. In order to prove this property for our estimators, we will need a sequence of orthogonal polynomials with respect to σ.

Definition 4.2 (Orthonormal polynomials w.r.t. σ)**.** *Let* σ *be a measure on* $\mathbb{R}$ *whose moments are all finite. Endow the set of polynomials* $\mathbb{R}[X]$ *with the scalar product*

$$\langle P, Q\rangle_\sigma = \int_{\mathbb{R}} P(\lambda)Q(\lambda)\mathrm{d}\sigma(\lambda)\,.$$

Denote $T \in \mathbb{N}\cup\{\infty\}$ *the cardinal of the support of* σ. *Then there exists a unique family* $\pi_0, \pi_1, ..., \pi_{T-1}$ *of polynomials with positive leading coefficient, such that for all* $t < T$, $\pi_0, \pi_1, ..., \pi_t$ *form an orthonormal basis of* $(\mathbb{R}_t[X], \langle ., .\rangle_\sigma)$, *where* $\mathbb{R}_t[X]$ *denotes the set of polynomials of degree smaller or equal to* t. *In other words, for all* $s, t < T$,

$$\deg \pi_t = t\,, \qquad\qquad \langle \pi_s, \pi_t\rangle_\sigma = \mathbf{1}_{\{s=t\}}\,.$$

$\pi_0, \pi_1, ..., \pi_{T-1}$ *are called the* orthonormal polynomials with respect to σ.

A much more comprehensive description of orthogonal polynomials from the point of view of applied mathematics can be found in [Gau04]. We now fix the spectral measure $\sigma = \sigma(G,W,v)$ and $\pi_0, \pi_1, \dots, \pi_{T-1}$ the corresponding orthogonal polynomials.

Proposition 4.3. *There exists a unique minimizing polynomial* P_t^σ *satisfying* (4.1) *and*

$$P_t^\sigma = \Big(\sum_{s=0}^t \pi_s(1)^2\Big)^{-1}\sum_{s=0}^t \pi_s(1)\pi_s\,, \tag{4.2}$$

$$\Lambda_t(\sigma, 1) = \frac{\operatorname{var}\left(P_t^\sigma(W)\xi\right)_v}{\operatorname{var}\nu} = \Big(\sum_{s=0}^t \pi_s(1)^2\Big)^{-1}. \tag{4.3}$$

This result is well-known and usually stated without proof [Nev86, Sections 3, 4.1], [Nev79, Section 2]; we give the short proof in Appendix A.2. A fundamental result on orthogonal polynomials states that they follow a second order recursion.

Proposition 4.4 ([Gau04, Section 1.5.1])**.** *There exist two sequences of coefficients* $(a_t)_{0\leqslant t\leqslant T-2}$ *and* $(b_t)_{0\leqslant t\leqslant T-2}$ *with* $a_t > 0$ *and* $b_t \in \mathbb{R}$ *such that (using the convention* $\pi_{-1} = 0$*)*

$$a_{t+1}\pi_{t+1}(\lambda) = (\lambda - b_t)\pi_t(\lambda) - a_t\pi_{t-1}(\lambda)\,.$$

As a consequence, the best polynomial gossip algorithm is a second order method whose coefficients are determined by the graph G, the gossip matrix W and the vertex v. Assuming the coefficients a_t, b_t, $t \geqslant 0$ are given, the computation of the best polynomial gossip $\hat{\mu}^t = P_t^\sigma(W)\xi$ goes:

Computation formula		**Result**	
$x^{-1} = 0\,,$	$x^0 = \xi\,,\quad x^{t+1} = \frac{1}{a_{t+1}}\left(Wx^t - b_t x^t - a_t x^{t-1}\right)$	$x^t = \pi_t(W)\xi$	(4.4)
$\rho_{-1} = 0\,,$	$\rho_0 = 1\,,\quad \rho_{t+1} = \frac{1}{a_{t+1}}\left((1-b_t)\rho_t - a_t\rho_{t-1}\right)$	$\rho_t = \pi_t(1)$	(4.5)
$u^0 = \xi\,,$	$u^{t+1} = u^t + \rho_{t+1}x^{t+1}$	$u^t = \sum_{s=0}^t \pi_s(1)\pi_s(W)\xi$	(4.6)
$v_0 = 1\,,$	$v_{t+1} = v_t + \rho_{t+1}^2$	$v_t = \sum_{s=0}^t \pi_s(1)^2$	(4.7)
$\hat{\mu}^t = u^t/v_t$		$\hat{\mu}^t = P_t^\sigma(W)\xi$	(4.8)

However, we warn the reader that this algorithm is unpractical for two reasons: it assumes the full knowledge of the measure σ, as well as our ability to compute the sequences a_t, b_t from the measure σ, which can be challenging (see [Gau04]). The rest of this paper will circumvent these difficulties by approximating the measure σ with a simpler measure $\tilde{\sigma}$, whose recursion coefficients are known. We will show that in some cases, substituting $P_t^{\tilde{\sigma}}(W)\xi$ to $P_t^\sigma(W)\xi$ worsens the variances negligibly.

Remark 4.5. *Our method is similar to the Chebychev acceleration scheme used in* [AS14, SBB$^+$17] *to accelerate gossip. The underlying idea is that the Chebychev polynomial of degree* t *can be properly rescaled to a polynomial* P_t *such that* $P_t(1) = 1$ *and* $\sup_{\lambda\in[-1,1-\gamma]} |P_t(\lambda)|$ *is small for some* $\gamma > 0$ *(typically the spectral gap of the matrix* W*), thus* P_t *is a good candidate for polynomial gossip. Our framework brings two benefits over this approach. First, it adapts the choice of the polynomial* P_t *to the type of network (characterized by the spectral measure). Second, the Chebychev acceleration method performs poorly when* γ *is small (i.e., the graph is large) and* t *is small (compared to* $1/\sqrt{\gamma}$*), a regime where our acceleration still gives good results (see the simulations of Section 7).*

5. Message passing seen as a polynomial gossip algorithm

The message passing algorithm of [MR05] (in its zero-temperature limit) defines quantities on the edges of the graph G with the following recursion: for $v, w \in V$ linked by an edge in the graph G, it defines $K_{vw}^0 = 0$, $M_{vw}^0 = 0$, and

$$K_{vw}^{t+1} = 1 + \sum_{u\in\mathcal{N}(v),\, u\neq w} K_{uv}^t\,, \qquad M_{vw}^{t+1} = \frac{1}{K_{vw}^{t+1}}\left(\xi_v + \sum_{u\in\mathcal{N}(v),\, u\neq w} K_{uv}^t M_{uv}^t\right), \tag{5.1}$$

where $\mathcal{N}(v)$ denotes the set of neighbors of v. K_{vw} and M_{vw} are interpreted as messages going from v to w in G: M_{vw}^t corresponds to an average of observations gathered by v and transmitted to w; K_{vw}^t is the corresponding number of observations. We recommend [MR05, Section II.A] and Lemma A.1 for a detailed description of this intuition. At each time step t, the output of the algorithm is

$$\hat{\mu}_v^t = \frac{\xi_v + \sum_{u\in\mathcal{N}(v)} K_{uv}^t M_{uv}^t}{1 + \sum_{u\in\mathcal{N}(v)} K_{uv}^t}\,. \tag{5.2}$$

It is an easy check that this gossip method is unbiased. Furthermore it is exactly optimal on trees, as shown by the following proposition.

Proposition 5.1. *Assume that* G *is a tree. Then for all* $t \geqslant 1$*,* $v \in V$*,*

$$\hat{\mu}_v^t = \frac{1}{|B_v(t)|}\sum_{w\in B_v(t)} \xi_w\,, \qquad \frac{\operatorname{var}\hat{\mu}_v^t}{\operatorname{var}\nu} = \frac{1}{|B_v(t)|}\,.$$

However, nothing prevents us from running the message passing recursion (5.1)-(5.2) in a graph G with loops. In the case of regular graphs, we are able to interpret the message passing algorithm as a polynomial algorithm using P_t^σ with a particular choice of σ.

Theorem 5.2. *Assume G is d-regular, meaning that each vertex has degree d, $d \geqslant 2$. Assume further that $W = A(G)/d$. Recall that $\sigma(\mathbb{T}_d)$ is the spectral measure of the infinite d-regular tree. Then the output $\hat{\mu}_v^t$ of the message passing algorithm (5.1)-(5.2) on G can also be obtained as $\hat{\mu}_v^t = P_t^{\sigma(\mathbb{T}_d)}(W)\xi$ where $P_t^{\sigma(\mathbb{T}_d)}$ is defined in (4.1).*

As message passing algorithms are often derived by neglecting loops in a graph, this theorem should not surprise the reader: message passing corresponds to the best polynomial gossip algorithm when one *believes* the graph is a tree.

An important feature of the spectral measure of $\mathbb{T}_d$ and its corresponding orthogonal polynomials is that they can be computed explicitly (combining [Sod07, Section 2.2] and Lemma B.1), thus making the corresponding polynomial algorithm practical:

$$\sigma(\mathbb{T}_d)(\mathrm{d}\lambda) = \frac{d}{2\pi(1-\lambda^2)}\left(\frac{4(d-1)}{d^2} - \lambda^2\right)^{1/2} \mathbf{1}_{[-2\sqrt{d-1}/d, 2\sqrt{d-1}/d]}(\lambda)\mathrm{d}\lambda\,, \tag{5.3}$$

$$a_1 = \frac{1}{\sqrt{d}}\,, \qquad a_t = \frac{\sqrt{d-1}}{d} \quad \text{for } t \geqslant 2, \qquad b_t = 0 \quad \text{for all } t\,. \tag{5.4}$$

Note that by merging Proposition 5.1 and Theorem 5.2, it follows that:

Corollary 5.3. *The best polynomial algorithm on $\mathbb{T}_d$ ($d \geqslant 2$) gives a practical and exactly optimal gossip algorithm on $\mathbb{T}_d$.*

Note that it is fairly easy to derive an elementary proof of Corollary 5.3 (without using the connection with message passing). In particular, when $d = 2$ ($\mathbb{T}_2 = \mathbb{Z}$), the spectral measure is the arcsine distribution $\sigma(\mathbb{Z})(\mathrm{d}\lambda) = \mathrm{d}\lambda/(\pi\sqrt{1-\lambda^2})\mathbf{1}_{(-1,1)}(\lambda)$. The corresponding orthonormal polynomials are, up to rescaling, the Chebychev polynomials. We believe that the interested reader could gain some insights about our framework by repeating explicitly the derivations of Sections 4 and 5 in this particular case.

6. Design of polynomial gossip algorithms for grids

We now focus on the case of the regular grid $G = (\mathbb{Z}^d, E)$, with $d \geqslant 2$ and $E = \{\{v, w\} \mid \|v - w\|_2 = 1\}$ that we equip with the simple gossip matrix $W = A(\mathbb{Z}^d)/(2d)$. We believe that it is difficult to compute the sequence of orthogonal polynomials with respect to the spectral measure $\sigma(\mathbb{Z}^d)$ in this case. However, we have the following information on the spectrum.

Proposition 6.1. *The spectral measure $\sigma(\mathbb{Z}^d)$ has a symmetric density $w_d : \mathbb{R} \mapsto \mathbb{R}_+$, with support in $[-1, 1]$, and $w_d(\lambda) \sim C_d(1-\lambda)^{d/2-1}$ as $\lambda \to 1$, for some constant C_d.*

As eigenvalues close to 1 of a graph Laplacian are known to correspond to the large-scale structure of a graph, a natural idea for our gossip algorithm is that we can approximate $\sigma(\mathbb{Z}^d)$ by another measure $\tilde{\sigma}_d$ with the same behavior in 1. We choose $\tilde{\sigma}_d(\mathrm{d}\lambda) = (1-\lambda)^{d/2-1}\mathrm{d}\lambda$ because the corresponding orthogonal polynomials are the well-known Gegenbauer polynomials. The recursion coefficients can be computed (see [Gau04, Section 1.5.1] for instance):

$$a_t = \left(\frac{t(t+d-2)}{(2t+d-1)(2t+d-3)}\right)^{1/2}, \qquad b_t = 0\,. \tag{6.1}$$

Building on work studying the decrease of the Christoffel function [Nev79, Section 6.2, Lemma 21], we prove that this approximation gives optimal variance.

Theorem 6.2. *Equip the grid $\mathbb{Z}^d$, $d \geqslant 2$ with the simple gossip matrix $W = A(\mathbb{Z}^d)/(2d)$. Apply the gossip method corresponding to the Gegenbauer polynomials, that is $\hat{\mu}^t = P_t^{\tilde{\sigma}_d}(W)\xi$. Then the gossip method is asymptotically optimal, that is,*

$$\frac{\operatorname{var}\hat{\mu}_v^t}{\operatorname{var}\nu} = \Theta\left(\frac{1}{t^d}\right).$$

7. Simulations

In this section, we run our methods on random regular graphs, grids and random geometric graphs; the latter being a widely used model for real-world networks [Pen03, Section 1.1]. We compare our polynomial gossip algorithm with the simple gossip method (2.1), the shift-register algorithm of [CSY06] and the splitting algorithm of [RT17]. The last two algorithms are state-of-the-art accelerated methods for the averaging problem. Some theoretical results (see [LACM13, Theorem 2] for the former, [RT17, Theorem 4] for the latter) determine the best tuning of their parameters as a function of the spectral gap of W; this is the tuning we use in our simulations.

In all our simulations, we only specify the graph G that we run our algorithms on. It will be implicit that we built a gossip matrix on G with the formulas: $W_{vw} = \max(\deg v, \deg w)^{-1}$ if $v \in \mathcal{N}(w)$ and $W_{vv} = 1 - \sum_{w \in \mathcal{N}(v)} \max(\deg v, \deg w)^{-1}$. We measure the performance of all estimators $\hat{\mu}^t$ using the empirical variance averaged over the graph, $\|\hat{\mu}^t - \mu\|_2^2/n$. We compare with the best possible estimator $\hat{\mu}_v^t = (1/|B_v(t)|) \sum_{w \in B_v(t)} \xi_w$, which gives us a lower bound on the the performance achievable. All simulations are run with $\xi_w \sim \mathcal{N}(0,1)$.

Two-dimensional grid. We run simulations on a 40×40 square lattice ($n = 1600$ vertices). Approximating the finite grid with its infinite counterpart $\mathbb{Z}^2$, and following Section 6, we choose to run the algorithm using the Gegenbauer orthogonal polynomials for $\tilde{\sigma}_2$. The results are plotted in Figure 2A.

Two-dimensional random geometric graph. We build a random geometric graph G by sampling $n = 1600$ points uniformly in the unit square $[0,1]^2$ and linking pairs closer than $3/\sqrt{n} \approx 0.0548$. This tuning was chosen so that the resulting graph is connected with high probability. This ensures that the spectral gap is positive and the accelerated methods for averaging apply. As this random geometric graph has a structure "close to" $\mathbb{Z}^2$, we take again the algorithm using the Gegenbauer orthogonal polynomials w.r.t. $\tilde{\sigma}_2$. The results, averaged over 10 realizations of the graph, are shown in Figure 2B.

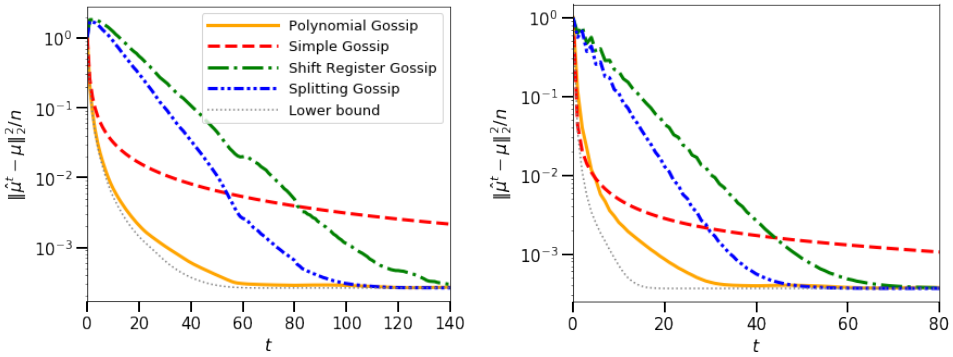
Three-dimensional grid. We run simulations on a $12 \times 12 \times 12$ cubic lattice ($n = 1728$ vertices). Approximating the finite lattice with its infinite counterpart $\mathbb{Z}^3$, and following Section 6, we choose to run the algorithm using the Gegenbauer orthogonal polynomials for $\tilde{\sigma}_3$. The results are plotted in Figure 2C.

Three-dimensional random geometric graph. We build a three-dimensional random geometric graph G by sampling $n = 1728$ points in the unit cube $[0,1]^3$ and linking pairs closer than $1.8/n^{1/3} = 0.15$. Again, this tuning was chosen so that the resulting graph is connected with high probability. As this random geometric graph has a structure "close to" $\mathbb{Z}^3$, we take the algorithm using the Gegenbauer orthogonal polynomials w.r.t. $\tilde{\sigma}_3$. The results, averaged over 10 realizations of the graph, are shown in Figure 2D.

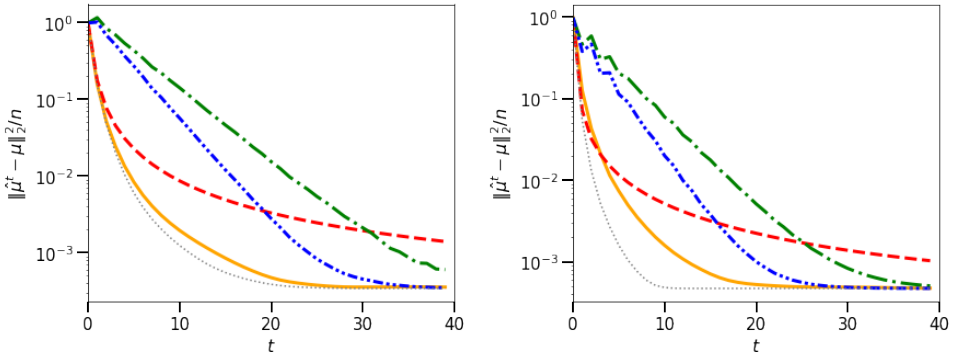
Note that in the four simulations of Figure 2, all curves converge to the same consensus limit $\hat{\mu}^t = (1/|V|) \sum_{v \in V} \xi_v$. As suggested by Section 3, simple gossip is slow in reaching that limit. In comparison, the shift-register and splitting methods reach consensus faster, but decrease slower in the first phase. Our polynomial gossip algorithm enjoys the fast decrease of simple gossip at the beginning, reaches consensus faster than the accelerated methods and is more efficient than existing algorithms in the intermediate regime. Our polynomials estimator $\hat{\mu}^t = P_t^{\tilde{\sigma}_2}\xi$ matches closely the lower bound of the best possible estimator in the case of grids, and more loosely in the case of random geometric graphs. This could mean that our approximation of the spectral measure of the random geometric graph by $\tilde{\sigma}_2$ or $\tilde{\sigma}_3$ is too crude.

Random regular graph. We also run our algorithmic comparisons on a random 3-regular graph. To be precise, we fix $n = 2000$ and pick a graph uniformly from the collection of all 3-regular graphs with n vertices. McKay's Theorem [Sod07, Theorem 1.1] shows that as $n \to \infty$, the spectrum of such a random graph converges to the spectral measure of the infinite 3-regular tree $\mathbb{T}_3$ (in distribution, for the weak convergence topology). According to Section 5, this is a strong incentive to use the message passing algorithm (5.1)-(5.2) of [MR05], that is equivalent to a polynomial algorithm (see Theorem 5.2). The results are plotted in Figure 3.

The message passing algorithm outperforms all other algorithms and matches closely the lower bound. This is due to the fact that the spectrum of the random regular graph is well-understood, and the corresponding orthogonal polynomials are known. This simulation encourages a deeper



(A) 2D grid (B) 2D random geometric graph



(C) 3D grid (D) 3D random geometric graph

FIGURE 2. Performance of different gossip algorithms running on graphs with an underlying low-dimensional geometry, as measured by $\|\hat{\mu}^t - \mu\|_2^2/n$.

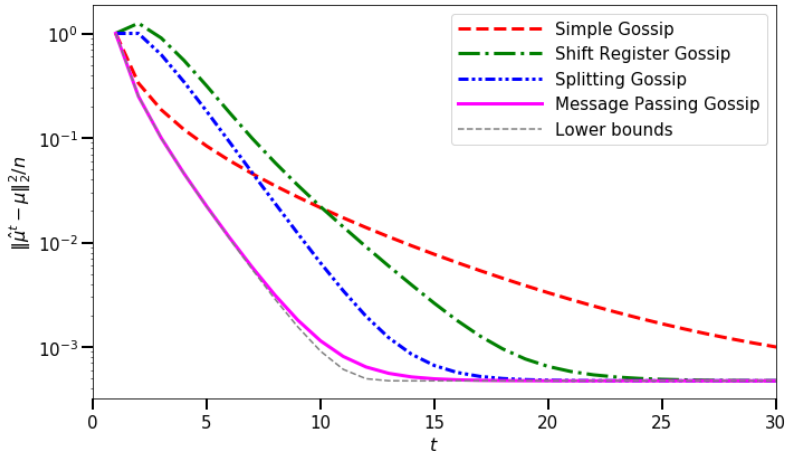


FIGURE 3. Performance of different gossip algorithms running a random regular graph, as measured by $\|\hat{\mu}^t - \mu\|_2^2/n$.

study of the orthogonal polynomials w.r.t the spectrum of some classes of random graphs, that we hope to conduct in future work.

8. Conclusion

In the averaging problem, the goal is to reach consensus (full averaging of the values in the network) as quickly as possible. However, the methods designed for this objective are paradoxically

bad at averaging locally in the intermediate regime before consensus in reached (see Figure 2). In statistical applications, local averaging may be sufficient and the network may be too large to hope for full averaging in the network. We proposed a new framework, better suited for these applications and new methods that are efficient *at all times*.

Our acceleration framework using orthogonal polynomials provides new tools for gossip and a new point of view on message passing. We hope to bring more theoretical support and understanding of the framework in future work. To what extent can one approximate the spectral measure σ of the graph when deriving the algorithm? Can we prove some results on large random graphs? Can we give interpretations of the recursion coefficients a_t and b_t in terms of the graph?

Acknowledgements

We acknowledge support from the European Research Council (grant SEQUOIA 724063). We also thank Loucas Pillaud-Vivien for his enlightening advice on this project.

References

[AGZ10] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni, *An introduction to random matrices, volume 118 of cambridge studies in advanced mathematics*, 2010. 15

[AS14] M. Arioli and J. Scott, *Chebyshev acceleration of iterative refinement*, Numerical Algorithms **66** (2014), no. 3, 591–608. 6

[BGPS06] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah, *Randomized gossip algorithms*, IEEE transactions on information theory **52** (2006), no. 6, 2508–2530. 1, 2

[Car88] Donald I Cartwright, *Some examples of random walks on free products of discrete groups*, Annali di matematica pura ed applicata **151** (1988), no. 1, 1–15. 15

[CS12] Jianshu Chen and Ali H Sayed, *Diffusion adaptation strategies for distributed optimization and learning over networks*, IEEE Transactions on Signal Processing **60** (2012), no. 8, 4289–4305. 1

[CSY06] Ming Cao, Daniel A. Spielman, and Edmund M. Yeh, *Accelerated gossip algorithms for distributed computation*, 44th Annual Allerton Conference on Communication, Control, and Computation, 2006, pp. 952–959. 1, 2, 3, 5, 8

[DAW12] John C Duchi, Alekh Agarwal, and Martin J Wainwright, *Dual averaging for distributed optimization: Convergence analysis and network scaling*, IEEE Transactions on Automatic control **57** (2012), no. 3, 592–606. 1

[Gau04] Walter Gautschi, *Orthogonal polynomials: computation and approximation*, Oxford University Press on Demand, 2004. 5, 6, 7, 16

[KSS16] Nathan Korda, Balázs Szörényi, and Li Shuai, *Distributed clustering of linear bandits in peer to peer networks*, International Conference on Machine Learning, vol. 48, 2016, pp. 1301–1309. 2

[LACM13] Ji Liu, Brian D. O. Anderson, Ming Cao, and A. Stephen Morse, *Analysis of accelerated gossip algorithms*, Automatica **49** (2013), no. 4, 873–883. 8

[LL10] Gregory F Lawler and Vlada Limic, *Random walk: a modern introduction*, vol. 123, Cambridge University Press, 2010. 4

[LSL16] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard, *On distributed cooperative decision-making in multiarmed bandits*, Control Conference (ECC), 2016 European, IEEE, 2016, pp. 243–248. 2

[MNT91] Attila Máté, Paul Nevai, and Vilmos Totik, *Szego's extremum problem on the unit circle*, Annals of Mathematics (1991), 433–453. 5

[MR05] Ciamac C Moallemi and Benjamin Van Roy, *Consensus propagation*, Advances on Neural Information Processing Systems, MIT Press, 2005, pp. 899–906. 1, 2, 6, 8

[MW89] Bojan Mohar and Wolfgang Woess, *A survey on spectra of infinite graphs*, Bulletin of the London Mathematical Society **21** (1989), no. 3, 209–234. 4, 14

[Nev79] Paul G Nevai, *Orthogonal polynomials*, vol. 23, American Mathematical Soc., 1979. 5, 7, 15

[Nev86] Paul Nevai, *Géza Freud, orthogonal polynomials and Christoffel functions. a case study*, Journal of Approximation Theory **48** (1986), no. 1, 3–167. 5

[NO09] Angelia Nedic and Asuman Ozdaglar, *Distributed subgradient methods for multi-agent optimization*, IEEE Transactions on Automatic Control **54** (2009), no. 1, 48–61. 1

[Pen03] Mathew Penrose, *Random geometric graphs*, no. 5, Oxford university press, 2003. 8

[RNV10] S Sundhar Ram, Angelia Nedić, and Venugopal V Veeravalli, *Distributed stochastic subgradient projection algorithms for convex optimization*, Journal of optimization theory and applications **147** (2010), no. 3, 516–545. 1

[RT17] Patrick Rebeschini and Sekhar Tatikonda, *Accelerated consensus via min-sum splitting*, Advances on Neural Information Processing Systems, 2017. 1, 2, 3, 8, 13

[S+14] Ali H Sayed et al., *Adaptation, learning, and optimization over networks*, Foundations and Trends® in Machine Learning **7** (2014), no. 4-5, 311–801. 1

[Saw78] Stanley Sawyer, *Isotropic random walks in a tree*, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **42** (1978), no. 4, 279–292. 4

[SBB+17] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié, *Optimal algorithms for smooth and strongly convex distributed optimization in networks*, International Conference on Machine Learning, 2017. 1, 6

[SBFH+13] Balazs Szorenyi, Róbert Busa-Fekete, István Hegedus, Róbert Ormándi, Márk Jelasity, and Balázs Kégl, *Gossip-based distributed stochastic bandit algorithms*, International Conference on Machine Learning, 2013, pp. 19–27. 2

[SGB10] Stefania Sardellitti, Massimiliano Giona, and Sergio Barbarossa, *Fast distributed average consensus algorithms based on advection-diffusion processes*, IEEE Transactions on Signal Processing **58** (2010), no. 2, 826–842. 1

[Sha09] Devavrat Shah, *Gossip algorithms*, Foundations and Trends® in Networking **3** (2009), no. 1, 1–125. 1

[Sod07] Sasha Sodin, *Random matrices, nonbacktracking walks, and orthogonal polynomials*, Journal of Mathematical Physics **48** (2007), no. 12, 123503. 7, 8

[Sze39] Gabor Szegö, *Orthogonal polynomials*, vol. 23, American Mathematical Soc., 1939. 16

[Tot00] Vilmos Totik, *Asymptotics for christoffel functions for general measures on the real line*, Journal d'Analyse Mathématique **81** (2000), no. 1, 283–303. 5

Appendix A. Proofs of the results

A.1. Proof of Proposition 2.2. Using that W is symmetric and that the ξ_w, $w \in V$ are i.i.d. random variables, we get.

$$\operatorname{var} \hat{\mu}_v^t = \operatorname{var} \langle P_t(W)\xi, e_v\rangle_{\ell^2(V)} = \operatorname{var} \langle \xi, P_t(W)e_v\rangle_{\ell^2(V)} = \|P_t(W)e_v\|^2_{\ell^2(V)} \left(\operatorname{var} \nu\right),$$

and using again that W is symmetric,

$$\|P_t(W)e_v\|^2_{\ell^2(V)} = \langle P_t(W)e_v, P_t(W)e_v\rangle_{\ell^2(V)} = \left\langle e_v, P_t(W)^2 e_v\right\rangle_{\ell^2(V)} .$$

This proves the left part of Eq. (2.4).

Note that the intuition lying behind the right part of Eq. (2.4) is very simple: the unbiased estimator $\hat{\mu}_v^t$ is a linear combination of observations corresponding to vertices in the ball $B_v(t)$, thus it must have variance greater than $\operatorname{var}\nu/|B_v(t)|$.

A more rigorous argument goes as follows: using that W is a gossip matrix, it is easy to show by induction that for all $s \geqslant 0$ and $v, w \in V$, if $(W^s)_{vw} > 0$, then there exists a path of length s linking v to w in G. As $\deg P_t \leqslant t$, this implies that $P_t(W)e_v$ has at most $|B_v(t)|$ non-zero entries. Furthermore, the entries of $P_t(W)e_v$ sum to 1 because $W\mathbf{1} = \mathbf{1}$ and $P_t(1) = 1$. Thus, using the Cauchy-Schwarz inequality,

$$\begin{aligned} 1 = \left(\sum_{w\in V} (P_t(W)e_v)_w\right)^2 &= \left(\sum_{w\in V} (P_t(W)e_v)_w \mathbf{1}_{\{(P_t(W)e_v)_w > 0\}}\right)^2 \\ &\leqslant \|P_t(W)e_v\|^2_{\ell^2(V)} \sum_{w\in V} \mathbf{1}_{\{(P_t(W)e_v)_w>0\}} \leqslant \|P_t(W)e_v\|^2_{\ell^2(V)} \, |B_v(t)| . \end{aligned}$$

Combining with the previous expression, this gives the right part of (2.4).

A.2. Proof of Proposition 4.3. Eq. (4.1) can be rewritten as

$$P_t^\sigma \in \operatorname*{argmin}_{P(1)=1,\, \deg P\leqslant t} \langle P, P\rangle_\sigma . \tag{A.1}$$

We parametrize the minimization problem (A.1) using the orthogonal polynomials $\pi_0, \dots, \pi_t$ with respect to σ: as $\pi_0, \dots, \pi_t$ form a basis of $\mathbb{R}_t[X]$, one can uniquely decompose a polynomial P such that $\deg P \leqslant t$ as $P = \alpha_0\pi_0 + \dots + \alpha_t\pi_t$. Then $P(1) = \alpha_0\pi_0(1) + \dots + \alpha_t\pi_t(1)$, and by orthogonality of the orthogonal polynomials with respect to σ, $\langle P, P\rangle_\sigma = \alpha_0^2 + \dots + \alpha_t^2$. Denoting $\beta = (\pi_0(1), \dots, \pi_t(1))$, our minimization problem becomes

$$\min_{\langle\alpha,\beta\rangle=1} \|\alpha\|_2^2 .$$

This is the minimization of a strictly convex function over a linear subspace, thus the minimizer is unique and given by $\alpha^* = \beta/\|\beta\|_2^2$. One can then derive equations (4.2) and (4.3) from $P_t^\sigma = \alpha_0^*\pi_0 + \dots + \alpha_t^*\pi_t$.

A.3. Proof of Proposition 4.4. The polynomial $\lambda\pi_t(\lambda)$ of the variable λ is of degree $t+1$, thus it can be decomposed over the orthonormal basis $\pi_0(\lambda), \pi_1(\lambda), \dots, \pi_{t+1}(\lambda)$:

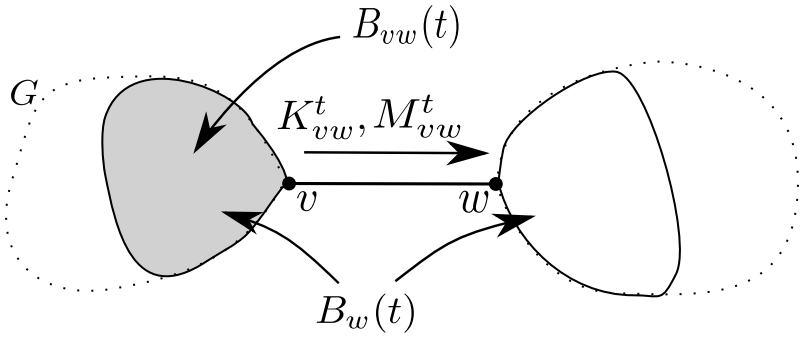
$$\lambda\pi_t(\lambda) = \sum_{s=0}^{t+1} \langle \lambda\pi_t, \pi_s\rangle_\sigma \pi_s(\lambda)\,.$$

Note that $\langle \lambda\pi_t, \pi_s\rangle_\sigma = \int \lambda\pi_t(\lambda)\pi_s(\lambda)\mathrm{d}\sigma(\lambda) = \langle \pi_t, \lambda\pi_s\rangle_\sigma = 0$ when $s \leqslant t-2$ because in this case $\lambda\pi_s(\lambda) \in \mathbb{R}_{t-1}[X]$ and π_t is orthogonal to $\mathbb{R}_{t-1}[X]$. Thus

$$\lambda\pi_t(\lambda) = \langle \lambda\pi_t, \pi_{t+1}\rangle_\sigma \pi_{t+1}(\lambda) + \langle \lambda\pi_t, \pi_t\rangle_\sigma \pi_t(\lambda) + \langle \pi_t, \lambda\pi_{t-1}\rangle_\sigma \pi_{t-1}(\lambda)\,,$$

where we keep the convention $\pi_{-1} = 0$. Denoting $a_t = \langle \pi_t, \lambda\pi_{t-1}\rangle_\sigma$ and $b_t = \langle \lambda\pi_t, \pi_t\rangle_\sigma$, we get the recursion formula. Note that a_t must also be the ratio of the leading coefficients of $\pi_{t-1}(\lambda)$ and $\pi_t(\lambda)$, which are both positive. Hence it must be positive.

A.4. Proof of Proposition 5.1. Let $t \geqslant 0$ and $v, w \in V$ be two vertices linked by an edge in G. Define $B_{vw}(t)$ as the set of vertices u in $B_w(t)$ such that all paths in the tree G going from u to w pass though v.



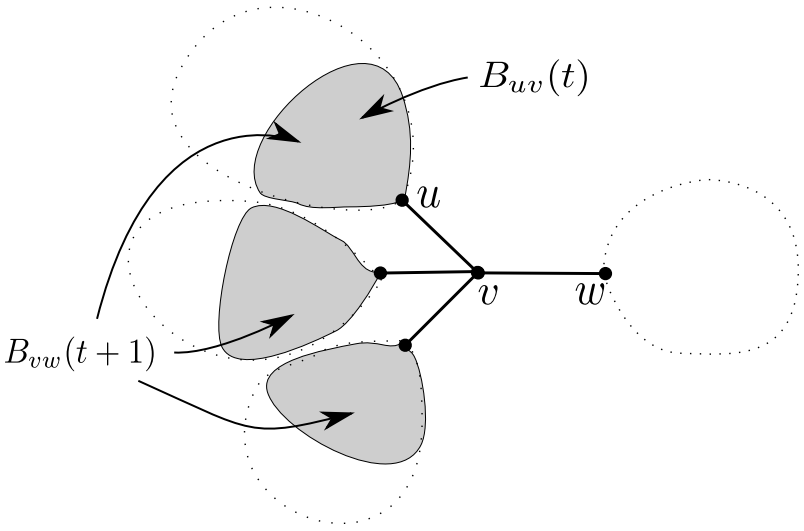
Lemma A.1. *For all $t \geqslant 0$, for all $v, w \in V$ linked by an edge in G,*

$$K^t_{vw} = |B_{vw}(t)|\,, \qquad \text{and} \quad \text{if } t \geqslant 1, \quad M^t_{vw} = \frac{1}{|B_{vw}(t)|} \sum_{u \in B_{vw}(t)} \xi_u\,.$$

Proof. The proof goes by induction. The statement is trivial for $t = 0, 1$. For the induction, assume the result at time t and note that

$$B_{vw}(t+1) = \{v\} \cup \left(\bigcup_{u \in \mathcal{N}(v),\, u \neq w} B_{uv}(t) \right)\,, \tag{A.2}$$

where all unions are disjoint. This essentially comes from the fact that G has no loops.



Taking cardinal, we get that

$$|B_{vw}(t+1)| \overset{(A.2)}{=} 1 + \sum_{u\in\mathcal{N}(v),\, u\neq w} |B_{uv}(t)| \overset{\text{(induction)}}{=} 1 + \sum_{u\in\mathcal{N}(v),\, u\neq w} K^t_{uv} \overset{(5.1)}{=} K^{t+1}_{vw}\,.$$

This proves the induction for the first equality. The proof for the second equality is similar:

$$\begin{aligned}\frac{1}{|B_{vw}(t+1)|}\sum_{u\in B_{vw}(t+1)} \xi_u &\overset{(A.2)}{=} \frac{1}{K^{t+1}_{vw}}\left(\xi_v + \sum_{u\in\mathcal{N}(v),\, u\neq w}\ \sum_{x\in B_{uv}(t)} \xi_x\right)\\ &\overset{\text{(induction)}}{=} \frac{\xi_v + \sum_{u\in\mathcal{N}(v),\, u\neq w} |B_{uv}(t)| M^t_{uv}}{K^{t+1}_{vw}}\\ &\overset{\text{(induction)}}{=} \frac{\xi_v + \sum_{u\in\mathcal{N}(v),\, u\neq w} K^t_{uv}M^t_{uv}}{K^{t+1}_{vw}} \overset{(5.1)}{=} M^{t+1}_{vw}\,.\end{aligned}$$

□

We now end the proof of Proposition 5.1. As $B_v(t) = \{v\} \cup \left(\bigcup_{u\in\mathcal{N}(v)} B_{uv}(t)\right)$ with disjoint unions, using Lemma A.1, we get

$$\frac{1}{|B_v(t)|}\sum_{w\in B_v(t)} \xi_w = \frac{\xi_v + \sum_{u\in\mathcal{N}(v)}\sum_{w\in B_{uv}(t)} \xi_w}{1 + \sum_{u\in\mathcal{N}(v)} |B_{uv}(t)|} = \frac{\xi_v + \sum_{u\in\mathcal{N}(v)} K^t_{uv}M^t_{uv}}{1 + \sum_{u\in\mathcal{N}(v)} K^t_{uv}} \overset{(5.2)}{=} \hat{\mu}^t_v\,.$$

A.5. Proof of Theorem 5.2. As noted by [RT17], the message passing iteration (5.1)-(5.2) indexed by the edges of the graph can be written as an iteration indexed by the vertices of the graph. We repeat here the elementary derivation of this statement in our particular case of d-regular graphs.

First, because G is d-regular, it is an easy check from (5.1) that K^t_{vw} does not depend on the edge (v,w) (thus we denote it K^t) and it satisfies the recursion $K^0 = 0$, $K^{t+1} = 1 + (d-1)K^t$.

Let us now denote $S^t_v = \xi_v + \sum_{u\in\mathcal{N}(v)} K^t_{uv}M^t_{uv}$ and $L_t = 1 + dK^t$ so that $\hat{\mu}^t_v = S^t_v/L_t$. We will now find recursions for L_t and S^t:

$$L_{t+1} = 1 + dK^{t+1} \overset{(5.1)}{=} 1 + d(1 + (d-1)K^t) = 2 + (d-1)(1 + dK^t) = 2 + (d-1)L_t\,,$$

and

$$\begin{aligned}S^{t+1}_v &= \xi_v + \sum_{u\in\mathcal{N}(v)} K^{t+1}M^{t+1}_{uv} \overset{(5.1)}{=} \xi_v + \sum_{u\in\mathcal{N}(v)}\left(\xi_u + \sum_{w\in\mathcal{N}(u), w\neq v} K^t M^t_{wu}\right)\\ &= \xi_v + \sum_{u\in\mathcal{N}(v)} \left(S^t_u - K^t M^t_{vu}\right)\,.\end{aligned}$$

As

$$\begin{aligned}\sum_{u\in\mathcal{N}(v)} K^t M^t_{vu} &\overset{(5.1)}{=} d\xi_v + \sum_{u\in\mathcal{N}(v)}\ \sum_{w\in\mathcal{N}(v), w\neq u} K^{t-1}M^{t-1}_{wv}\\ &= d\xi_v + (d-1)\sum_{w\in\mathcal{N}(v)} K^{t-1}M^{t-1}_{wv} = \xi_v + (d-1)S^{t-1}_v\,,\end{aligned}$$

we finally get

$$S^{t+1} = A(G)S^t - (d-1)S^{t-1}\,.$$

To sum up, we now have the simpler formulas for the message passing algorithm:

$$L_{t+1} = 2 + (d-1)L_t \qquad L_0 = 1 \tag{A.3}$$

$$S^{t+1} = A(G)S^t - (d-1)S^{t-1} \qquad S^0 = \xi \qquad S^1 = \xi + A(G)\xi \tag{A.4}$$

$$\hat{\mu}^t = S^t/L_t \tag{A.5}$$

We now want to compare this message passing algorithm to the polynomial algorithm (4.4)-(4.8) for d-regular trees (see the coefficients a_t, b_t in (5.4)), whose recursion is the following:

$$x^{-1} = 0 \qquad x^0 = \xi \qquad x^1 = \sqrt{d} W\xi \qquad x^2 = \frac{d}{\sqrt{d-1}} W x^1 - \sqrt{\frac{d}{d-1}}\xi \tag{A.6}$$

$$x^{t+1} = \frac{d}{\sqrt{d-1}} W x^t - x^{t-1} \quad (t \geqslant 2) \tag{A.7}$$

$$\rho_{-1} = 0 \qquad \rho_0 = 1 \qquad \rho_1 = \sqrt{d} \qquad \rho_2 = \frac{d}{\sqrt{d-1}}\rho_1 - \sqrt{\frac{d}{d-1}} \tag{A.8}$$

$$\rho_{t+1} = \frac{d}{\sqrt{d-1}}\rho_t - \rho_{t-1} \quad (t \geqslant 2) \tag{A.9}$$

$$u^0 = \xi \qquad u^{t+1} = u^t + \rho_{t+1} x^{t+1} \tag{A.10}$$

$$v^0 = 1 \qquad v^{t+1} = v^t + \rho_{t+1}^2 \tag{A.11}$$

$$m^t = u^t / v^t \tag{A.12}$$

Here, we denote m^t the output of the algorithm (as opposed to (4.8), where it is denoted $\hat{\mu}^t$) as we want to keep the notation $\hat{\mu}^t$ for the output of the message passing algorithm. The theorem states that $\hat{\mu}^t = m^t$. We will prove the stronger fact that $L_t = v_t$ and $S^t = u^t$.

Proof of $L_t = v_t$. From (A.8)-(A.9), it is an easy check by induction that for $t \geqslant 1$, $\rho_t^2 = d(d-1)^t$. Using this, let us now show by induction that $L_t = v_t$. It is true that $L_0 = v_0$ and $L_1 = v_1$, so now we take $t \geqslant 1$, assume that $L_{t-1} = v_{t-1}$ and $L_t = v_t$, and show that $L_{t+1} = v_{t+1}$. Indeed,

$$\begin{aligned} L_{t+1} &\overset{\text{(A.3)}}{=} 2 + (d-1)L_t + L_t - (2 + (d-1)L_{t-1}) = L_t + (d-1)(L_t - L_{t-1}) \\ &\overset{\text{(induction)}}{=} v_t + (d-1)(v_t - v_{t-1}) \overset{\text{(A.11)}}{=} v_t + (d-1)\rho_t^2 = v_t + \rho_{t+1}^2 \overset{\text{(A.11)}}{=} v_{t+1}\,. \end{aligned}$$

Proof of $S^t = u^t$. Again, we proceed by induction. It is true that $S^0 = u^0$, $S^1 = u^1$, and $S^2 = u^2$, so now we take $t \geqslant 2$, assume that $S^{t-2} = u^{t-2}$, $S^{t-1} = u^{t-1}$ and $S^t = u^t$, and show that $S^{t+1} = u^{t+1}$. Indeed,

$$\begin{aligned} u^{t+1} &\overset{\text{(A.10)}}{=} u^t + \rho_{t+1} x^{t+1} \overset{\text{(A.7)}}{=} u^t + \rho_{t+1}\frac{d}{\sqrt{d-1}} W x^t - \rho_{t+1} x^{t-1} \\ &= u^t + d\rho_t W x^t - (d-1)\rho_{t-1} x^{t-1} \overset{\text{(A.10)}}{=} u^t + dW(u^t - u^{t-1}) - (d-1)(u^{t-1} - u^{t-2})\,. \end{aligned}$$

Then using the induction hypothesis and the fact that $dW = A(G)$, we get:

$$u^{t+1} = S^t + A(G)S^t - A(G)S^{t-1} - (d-1)S^{t-1} + (d-1)S^{t-2} \overset{\text{(A.4)}}{=} S^t + S^{t+1} - S^t = S^{t+1}\,.$$

A.6. Proof of Proposition 6.1. The spectrum of $\mathbb{Z}$ is well known and can be expressed in closed-form (see for instance [MW89, Section 7.A]):

$$\sigma(\mathbb{Z})(\mathrm{d}\lambda) = w_1(\lambda)\mathrm{d}\lambda\,, \qquad w_1(\lambda) = \frac{1}{\pi}\frac{1}{\sqrt{1-\lambda^2}}\mathbf{1}_{(-1,1)}(\lambda)\,.$$

The graph $\mathbb{Z}^d$ is the Cartesian product $\mathbb{Z} \times \cdots \times \mathbb{Z}$ (d times). According to [MW89, Theorem 4.10], the spectral measure of the adjacency matrix of the Cartesian product of two graphs is the convolution of the spectral measures of the adjacency matrices of the two graphs. This implies that $\sigma(\mathbb{Z}^d, A(\mathbb{Z}^d)) = \sigma(\mathbb{Z}, A(\mathbb{Z}))^{*d}$. (The notation $.^{*d}$ denotes the convolution power d of an object.) Here a rescaling is needed as the spectral measure we consider is $\sigma(\mathbb{Z}^d) = \sigma(\mathbb{Z}^d, A(\mathbb{Z}^d)/(2d))$, the image measure of $\sigma(\mathbb{Z}^d, A(\mathbb{Z}^d))$ by the map $\lambda \mapsto \lambda/(2d)$. Overall, it is easy to check that $\sigma(\mathbb{Z}^d)$ has a density w_d, which satisfies

$$w_d(\lambda) = d w_1^{*d}(d\lambda)\,,$$

(the expression $w_1^{*d}(d\lambda)$ is well-defined Lebesgue almost-everywhere). The symmetry of w_d follows from the symmetry of w_1.

To estimate $w_d(\lambda)$ as $\lambda \to 1$, we use the estimates on the resolvent given by [Car88, Lemma 4]: for $z \in \mathbb{C}\backslash[-1,1]$,

$$S_d(z) := \int \frac{w_d(\lambda)\mathrm{d}\lambda}{z-\lambda} = \begin{cases} (z-1)^{d/2-1} g_d(z) + h_d(z) & \text{if } d \text{ is odd,} \\ (z-1)^{d/2-1}\mathrm{Log}(z-1) g_d(z) + h_d(z) & \text{if } d \text{ is even,} \end{cases}$$

where g_d and h_d are two analytic functions on a neighborhood of 1, and $g_d(1) \neq 0$. We then use the Stieltjes-Perron inversion formula (see for instance [AGZ10, Theorem 2.4.3]):

$$\begin{aligned} w_d(\lambda) &= \lim_{\varepsilon \to 0+} \frac{1}{2i\pi} \left[S_d(\lambda - i\varepsilon) - S_d(\lambda + i\varepsilon) \right] \\ &= \begin{cases} (-1)^{(d-1)/2} \frac{1}{\pi} (1-\lambda)^{d/2-1} g_d(\lambda) & \text{if } d \text{ is odd,} \\ (-1)^{d/2} (1-\lambda)^{d/2-1} g_d(\lambda) & \text{if } d \text{ is even.} \end{cases} \end{aligned}$$

As g_d is continuous in 1 and $g_d(1) \neq 0$, this concludes the proof.

A.7. Proof of Theorem 6.2. In section A.6, we have seen that the density of $\sigma(\mathbb{Z}^d)$ is defined Lebesgue almost-everywhere by

$$w_d(\lambda) = d w_1^{*d}(d \cdot \lambda)\,, \qquad\qquad w_1(\lambda) = \frac{1}{\pi\sqrt{1-\lambda^2}} \mathbf{1}_{[-1,1]}(\lambda)\,.$$

The following technical lemma gives upper bounds for w_d.

Lemma A.2. *There exists constants $D_1, D_2, D_3, \dots$ such that the following inequalities hold almost everywhere:*

(i) $w_2(\lambda) \leqslant D_1 \log \frac{1}{|\lambda|} + D_2$,

(ii) $w_d(\lambda) \leqslant D_d$ *for* $d \geqslant 3$.

The proof of this lemma is given at the end of this section. We now finish the proof of Theorem 6.2.

If $d \geqslant 3$, we know from Lemma A.2.(ii) that $w_d(\lambda) \leqslant D_d$ for all $\lambda \in [-1,1]$, and from Proposition 6.1 that $w_d(\lambda) \sim C_d(1-\lambda)^{d/2-1}$ as $\lambda \to 1$. Moreover w_d is symmetric, so there exists a constant C'_d such that for all $\lambda \in [-1,1]$,

$$w_d(\lambda) \leqslant C'_d (1-\lambda)^{d/2-1}\,.$$

This leads to a bound on the variance of the estimator $\hat{\mu}_v^t = P_t^{\tilde{\sigma}_d}(W)\xi$:

$$\frac{\operatorname{var} \hat{\mu}_v^t}{\operatorname{var} \nu} = \int_{-1}^1 P_t^{\tilde{\sigma}_d}(\lambda)^2 w_d(\lambda)\mathrm{d}\lambda \leqslant C'_d \int_{-1}^1 P_t^{\tilde{\sigma}_d}(\lambda)^2 \tilde{\sigma}_d(\mathrm{d}\lambda) = C'_d \Lambda_t(\tilde{\sigma}_d, 1)\,.$$

The decrease of the Christoffel function for the Jacobi weights is given by [Nev79, Section 6.2, Lemma 21]: it gives $\Lambda_t(\tilde{\sigma}_d, 1) = O(1/t^d)$, which concludes the proof in this case.

If $d = 2$, note that $\tilde{\sigma}_2 = \mathbf{1}_{[-1,1]}(\lambda)\mathrm{d}\lambda$ is simply the Lebesgue measure restricted to $[-1,1]$. This case is harder because w_2 is not uniformly bounded on $[-1,1]$, so we cannot use only the decrease of the Christoffel function $\Lambda_t(\tilde{\sigma}_2, 1)$. We only have Lemma A.2.(i), which gives

$$\begin{aligned} \frac{\operatorname{var} \hat{\mu}_v^t}{\operatorname{var} \nu} &= \int_{-1}^1 P_t^{\tilde{\sigma}_2}(\lambda)^2 w_2(\lambda)\mathrm{d}\lambda = \int_{-1}^1 P_t^{\tilde{\sigma}_2}(\lambda)^2 \left(D_1 \log \frac{1}{|\lambda|} + D_2 \right) \mathrm{d}\lambda \\ &= D_1 \int_{-1}^{1/2} P_t^{\tilde{\sigma}_2}(\lambda)^2 \log \frac{1}{|\lambda|} \mathrm{d}\lambda + D_1 \int_{1/2}^1 P_t^{\tilde{\sigma}_2}(\lambda)^2 \log \frac{1}{|\lambda|} \mathrm{d}\lambda + D_2 \int_{-1}^1 P_t^{\tilde{\sigma}_2}(\lambda)^2 \mathrm{d}\lambda\,. \end{aligned}$$

Again, [Nev79, Section 6.2,,Lemma 21] proves that

$$\int_{-1}^1 P_t^{\tilde{\sigma}_2}(\lambda)^2 \mathrm{d}\lambda = \Lambda_t(\tilde{\sigma}_2, 1) = O\left(\frac{1}{t^2}\right),$$

and

$$\int_{1/2}^1 P_t^{\tilde{\sigma}_2}(\lambda)^2 \log \frac{1}{|\lambda|} \mathrm{d}\lambda \leqslant (\log 2) \int_{1/2}^1 P_t^{\tilde{\sigma}_2}(\lambda)^2 \mathrm{d}\lambda = O\left(\frac{1}{t^2}\right).$$

To conclude, we only need to show that $I_t := \int_{-1}^{1/2} P_t^{\tilde{\sigma}_2}(\lambda)^2 \log \frac{1}{|\lambda|} \mathrm{d}\lambda = O(1/t^2)$. In the case $d = 2$, the orthonormal polynomials $\pi_0, \pi_1, \dots$ w.r.t. the uniform measure $\tilde{\sigma}_2$ are closely related to the classical family of Legendre polynomials $P_0, P_1, \dots$ by the relation

$$\pi_s = \sqrt{\frac{2s+1}{2}} P_s \,, \tag{A.13}$$

(see [Gau04, Section 1.5.1] for instance). This will allow us to show $I_t = O(1/t^2)$, using that

$$P_s(1) = 1 \,, \qquad \max_{\lambda \in [-1,1]} |P_s(\lambda)| \leqslant 1 \,. \tag{A.14}$$

(see [Sze39, Section 7.21] for a proof). According the the Christoffel-Darboux formula [Gau04, Section 1.3.3],

$$\begin{aligned}
\sum_{s=0}^{t} \pi_s(1)\pi_s(\lambda) &= a_{t+1} \frac{\pi_{t+1}(1)\pi_t(\lambda) - \pi_t(1)\pi_{t+1}(\lambda)}{1-\lambda} \\
&\overset{(6.1)}{=} \left(\frac{(t+1)^2}{(2t+3)(2t+1)} \right)^{1/2} \frac{\sqrt{\frac{2t+3}{2}}\pi_t(\lambda) - \sqrt{\frac{2t+1}{2}}\pi_{t+1}(\lambda)}{1-\lambda} \\
&= \frac{t+1}{2} \frac{\sqrt{\frac{2}{2t+1}}\pi_t(\lambda) - \sqrt{\frac{2}{2t+3}}\pi_{t+1}(\lambda)}{1-\lambda} \\
&= \frac{t+1}{2} \frac{P_t(\lambda) - P_{t+1}(\lambda)}{1-\lambda} \,.
\end{aligned}$$

We substitute this inequality in (4.2):

$$P_t^{\tilde{\sigma}_2}(\lambda) = \left(\sum_{s=0}^{t} \pi_s(1)^2 \right)^{-1} \frac{t+1}{2} \frac{P_t(\lambda) - P_{t+1}(\lambda)}{1-\lambda} = \frac{1}{t+1} \frac{P_t(\lambda) - P_{t+1}(\lambda)}{1-\lambda} \,.$$

Thus from (A.14),

$$\left| P_t^{\tilde{\sigma}_2}(\lambda) \right| \leqslant \frac{1}{t+1} \frac{2}{1-\lambda}$$

and finally

$$\begin{aligned}
I_t &= \int_{-1}^{1/2} P_t^{\tilde{\sigma}_2}(\lambda)^2 \log \frac{1}{|\lambda|} \mathrm{d}\lambda \leqslant \int_{-1}^{1/2} \frac{4}{(t+1)^2} \frac{1}{(1-\lambda)^2} \log \frac{1}{|\lambda|} \mathrm{d}\lambda \\
&\leqslant \frac{16}{(t+1)^2} \int_{-1}^{1/2} \log \frac{1}{|\lambda|} \mathrm{d}\lambda = O\left(\frac{1}{t^2} \right) .
\end{aligned}$$

This concludes the proof of Theorem 6.2.

Proof of Lemma A.2. **(i).** As w_2 is symmetric, we only need to prove (i) for $\lambda > 0$. For all $\lambda > 0$,

$$\begin{aligned}
(w_1 * w_1)(\lambda) &= \int_{-(1-\lambda)}^{1} w_1(\rho) w_1(\lambda - \rho) \mathrm{d}\rho \\
&= 2 \int_{\lambda/2}^{1} w_1(\rho) w_1(\lambda - \rho) \mathrm{d}\rho \qquad \text{using the symmetry of } w_1.
\end{aligned}$$

Note now that

$$\frac{1}{\sqrt{1-\lambda^2}} \leqslant \begin{cases} \frac{1}{\sqrt{1-\lambda}} & \text{if } \lambda \in [0,1), \\ \frac{1}{\sqrt{1+\lambda}} & \text{if } \lambda \in (-1,0], \end{cases}$$

thus for $\rho > \lambda/2 > 0$ we have

$$\begin{aligned}
w_1(\rho) &\leqslant \frac{1}{\pi} \frac{1}{\sqrt{1-\rho}} \,, \\
w_1(\lambda - \rho) &\leqslant \frac{1}{\pi} \left(\frac{1}{\sqrt{1-(\lambda-\rho)}} + \frac{1}{\sqrt{1+(\lambda-\rho)}} \right) .
\end{aligned}$$

We get that

$$(w_1 * w_1)(\lambda) \leqslant \frac{2}{\pi^2}\left(\int_{\lambda/2}^1 \frac{1}{\sqrt{1-\rho}}\frac{1}{\sqrt{1-(\lambda-\rho)}}\mathrm{d}\rho + \int_{\lambda/2}^1 \frac{1}{\sqrt{1-\rho}}\frac{1}{\sqrt{1+(\lambda-\rho)}}\mathrm{d}\rho\right). \tag{A.15}$$

These integrals can be computed explicitly using primitives:

$$\begin{aligned}
\int_{\lambda/2}^1 \frac{1}{\sqrt{1-\rho}}\frac{1}{\sqrt{1-(\lambda-\rho)}}\mathrm{d}\rho &= \left[-2\arctan\left(\frac{\sqrt{1-\rho}}{\sqrt{\rho-\lambda+1}}\right)\right]_{\lambda/2}^1 = \pi/2\,,\\
\int_{\lambda/2}^1 \frac{1}{\sqrt{1-\rho}}\frac{1}{\sqrt{1+(\lambda-\rho)}}\mathrm{d}\rho &= \left[-2\log\left(\sqrt{\lambda-\rho+1}+\sqrt{1-\rho}\right)\right]_{\lambda/2}^1\\
&= \log\frac{1}{\lambda} + 2\log\left(\sqrt{1+\frac{\lambda}{2}}+\sqrt{1-\frac{\lambda}{2}}\right).
\end{aligned}$$

Substituting in (A.15), we get

$$(w_1 * w_1)(\lambda) \leqslant D_1' \log\frac{1}{\lambda} + D_2'$$

with $D_1' = 2/\pi^2$ and $D_2' = 1/\pi + 4\log(2\sqrt{2})/\pi^2$. Finally,

$$w_2(\lambda) = 2(w_1 * w_1)(2\lambda) \leqslant 2D_1'\log\frac{1}{2\lambda} + 2D_2' = D_1\log\frac{1}{\lambda} + D_2$$

with $D_1 = 2D_1'$ and $D_2 = 2D_2' - 2D_1'\log 2$, which proves (i).
(ii). We start by proving by induction over $d \geqslant 3$ that there exists a constant D_3' such that for all $d \geqslant 3$, for all $\lambda \in \mathbb{R}$, $w_1^{*d}(\lambda) \leqslant D_3'$.
Initialization. $d = 3$. Then for $\lambda \in \mathbb{R}$,

$$\begin{aligned}
w_1^{*3}(\lambda) &= \int w_1(\rho) w_1^{*2}(\lambda-\rho)\mathrm{d}\rho\\
&\leqslant \int \frac{1}{\pi}\frac{1}{\sqrt{1-\rho^2}}\mathbf{1}_{[-1,1]}(\rho)\left(D_1'\log\frac{1}{|\lambda-\rho|} + D_2'\right)\mathbf{1}_{[-1,1]}(\lambda-\rho)\mathrm{d}\rho\\
&\leqslant \frac{D_1'}{\pi}\int \frac{1}{\sqrt{1-\rho^2}}\mathbf{1}_{[-1,1]}(\rho)\log\frac{1}{|\lambda-\rho|}\mathbf{1}_{[-1,1]}(\lambda-\rho)\mathrm{d}\rho + D_2'\,.
\end{aligned}$$

To bound the last integral independently of λ, we use Hölder's inequality:

$$\begin{aligned}
&\int \frac{1}{\sqrt{1-\rho^2}}\mathbf{1}_{[-1,1]}(\rho)\log\frac{1}{|\lambda-\rho|}\mathbf{1}_{[-1,1]}(\lambda-\rho)\mathrm{d}\rho\\
&\leqslant \left(\int \frac{1}{(1-\rho^2)^{3/4}}\mathbf{1}_{[-1,1]}(\rho)\mathrm{d}\rho\right)^{2/3}\left(\int \log^3\frac{1}{|\lambda-\rho|}\mathbf{1}_{[-1,1]}(\lambda-\rho)\mathrm{d}\rho\right)^{1/3}\\
&= \left(\int_{-1}^1 \frac{1}{(1-\rho^2)^{3/4}}\mathrm{d}\rho\right)^{2/3}\left(\int_{-1}^1 \log^3\frac{1}{|\rho|}\mathrm{d}\rho\right)^{1/3} =: D_1''\,,
\end{aligned}$$

where the last two integrals are finite and independent of λ. Thus denoting $D_3' = D_1'D_1''/\pi + D_2'$, we get the initialization.
Iteration. Let $d \geqslant 3$. We assume that $w_1^{*d}(\lambda) \leqslant D_3'$ for all $\lambda \in \mathbb{R}$. Then for $\lambda \in \mathbb{R}$,

$$w_1^{*(d+1)}(\lambda) = \int w_1^{*d}(\lambda-\rho)w_1(\rho)\mathrm{d}\rho \leqslant D_3'\int w_1(\rho)\mathrm{d}\rho = D_3'\,.$$

Thus we get the result for $w^{*(d+1)}$.

Finally, we have $w_d(\lambda) = dw_1^{*d}(d\cdot\lambda) \leqslant dD_3'$, thus we get (ii) with constant $D_d = dD_3'$. □

Appendix B. A rescaling lemma for orthogonal polynomials

Lemma B.1. *Let σ be a measure on $\mathbb{R}$, $\pi_0, \ldots, \pi_{T-1}$ the corresponding orthonormal polynomials and*

$$a_{t+1}\pi_{t+1}(\lambda) = (\lambda - b_t)\pi_t(\lambda) - a_t\pi_{t-1}(\lambda) \tag{B.1}$$

their recurrence formula (see Definition 4.2 and Theorem 4.4).

Let α be a positive real, $\varphi : \lambda \mapsto \alpha\lambda$ a linear function and $\tilde{\sigma}$ be the image measure of σ by φ (which means that for all measurable set A, $\tilde{\sigma}(A) = \sigma(\varphi^{-1}(A))$). Again, denote $\tilde{\pi}_0, \dots, \tilde{\pi}_{T-1}$ the orthonormal polynomials corresponding to $\tilde{\sigma}$ and

$$\tilde{a}_{t+1}\tilde{\pi}_{t+1}(\tilde{\lambda}) = (\tilde{\lambda} - \tilde{b}_t)\tilde{\pi}_t(\tilde{\lambda}) - \tilde{a}_t\tilde{\pi}_{t-1}(\tilde{\lambda})$$

the recursion formula. Then:

$$\tilde{\pi}_t(\tilde{\lambda}) = \pi_t(\tilde{\lambda}/\alpha)\,, \qquad \tilde{a}_t = \alpha a_t\,, \qquad \tilde{b}_t = \alpha b_t\,.$$

Proof. By change of variable,

$$\int \pi_t\left(\frac{\tilde{\lambda}}{\alpha}\right)\pi_s\left(\frac{\tilde{\lambda}}{\alpha}\right)d\tilde{\sigma}(\tilde{\lambda}) = \int \pi_t\left(\frac{\varphi(\lambda)}{\alpha}\right)\pi_s\left(\frac{\varphi(\lambda)}{\alpha}\right)\mathrm{d}\sigma(\lambda) = \int \pi_t(\lambda)\pi_s(\lambda)\mathrm{d}\sigma(\lambda) = \mathbf{1}_{\{s=t\}}\,,$$

and $\pi_t(\tilde{\lambda}/\alpha)$ has positive leading coefficient thus by uniqueness of the orthonormal polynomials, $\tilde{\pi}_t(\tilde{\lambda}) = \pi_t(\tilde{\lambda}/\alpha)$. The recurrence relation for $\tilde{\pi}_t$ follows by evaluating the recurrence relation (B.1) for π_t in $\tilde{\lambda}/\alpha$. □