Distributed averaging of observations in a graph: the gossip problem

Pierre Gaillard

May 31, 2018

Joint work with **Raphaël Berthier** and Francis Bach

Sierra project-team – INRIA Paris – Département d'informatique de l'ENS Paris

Motivation from optimization

Many machine learning problems can be cast as the minimization of an average:

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \,,$$

where $f_i(\theta)$ is the error of the parameter θ on a part of the data set indexed by i.

Typical machine learning problem

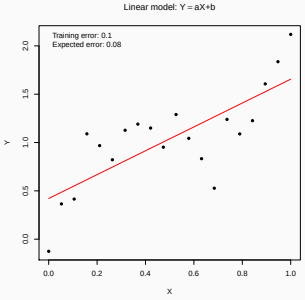
Optimize

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \, .$$

Example: Data set $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \dots, n$. One wants to find the best linear combination of the inputs $X_i \in \mathbb{R}^d$ to predict the outputs Y_i. The ordinary least square estimator is

$$\widehat{\theta}_n \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta^\top X_i)^2 \, .$$

In this case, $f_i(\theta) = (Y_i - \theta^\top X_i)^2$.



Typical machine learning problem

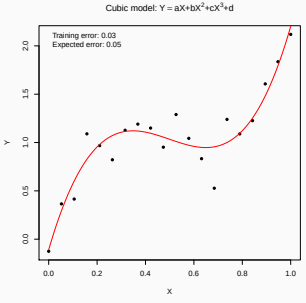
Optimize

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \, .$$

Example: Data set $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \ldots, n$. More complicated function spaces may be considered (polynomials, splines, kernels,...)

$$\widehat{\theta}_n \in \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - h_\theta(X_i)\big)^2 \, .$$

for some space of functions $\{h_\theta, \theta \in \Theta\}$.



Typical machine learning problem

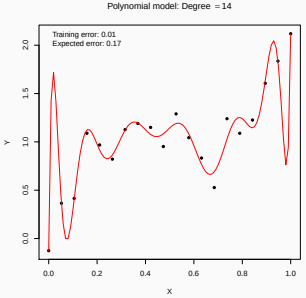
Optimize

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)\,.$$

Example: Data set $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \ldots, n$. More complicated function spaces may be considered (polynomials, splines, kernels,...)

$$\widehat{\theta}_n \in \operatorname*{arg\,min}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - h_\theta(X_i)\big)^2\,.$$

for some space of functions $\{h_\theta, \theta \in \Theta\}$.



Typical machine learning problem

Optimize

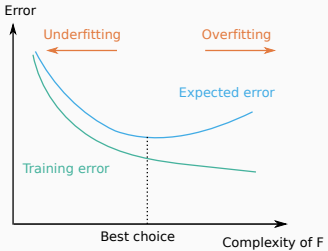
$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \, .$$

Example: Data set $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \dots, n$. More complicated function spaces may be considered (polynomials, splines, kernels,...)

$$\widehat{\theta}_n \in \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - h_\theta(X_i)\big)^2 + \mathrm{pen}(\theta) \, .$$

for some space of functions $\{h_\theta, \theta \in \Theta\}$.

To avoid overfitting, one needs to regularize.



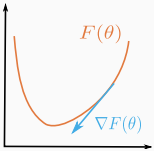
Classical optimization algorithm: gradient descent

Optimize

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \;=: F(\theta)\,.$$

Gradient descent: start from $\theta_0 \in \Theta$ and iterate

$$\theta_t = \theta_{t-1} - \gamma \nabla F(\theta_{t-1})\,.$$



For μ-strongly convex and L-smooth function F, it achieves ε-accuracy in $O(\kappa \log(1/\varepsilon))$ steps where $\kappa = L/\mu$ is the condition number.
Each step requires a full gradient computation.

Accelerated gradient descent (see Nesterov, 2004): improves it to $O(\sqrt{\kappa} \log(1/\varepsilon))$

$$\theta_t = \tilde{\theta}_{t-1} - \gamma \nabla F(\tilde{\theta}_{t-1}), \qquad \text{and} \qquad \tilde{\theta}_t = \theta_t + \gamma_t(\theta_t - \theta_{t-1})$$

Classical optimization algorithm: gradient descent

Optimize

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \; =: F(\theta) \, .$$

Classical algorithm: (Accelerated) Gradient descent $\theta_t = \theta_{t-1} - \gamma \nabla F(\theta_{t-1}) \, .$

Issue: it not distributed. Each update needs to compute the global gradient:

$$\nabla F(\theta_{t-1}) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\theta_{t-1}) \, .$$

Why would one want to use decentralized optimization? - speed up - privacy - robstness

Vast litterature on decentralized optimization.

The gossip problem: how to compute the mean of the values in a decentralized way?

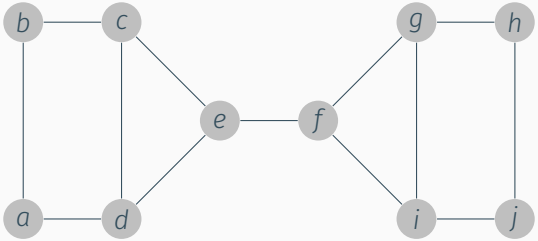
$$\frac{1}{n} \sum_{i=1}^{n} f_i(\theta) \qquad \text{or} \qquad \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\theta) \, .$$

Distributed learning in a graph – The Gossip problem

Some agent talk to another connected agents: information spreads.

Everyone has a piece of information

Can we find global information using this type of information propagation?

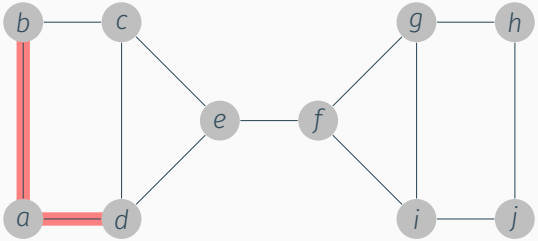


Distributed learning in a graph – The Gossip problem

Some agent talk to another connected agents: information spreads.

Everyone has a piece of information

Can we find global information using this type of information propagation?

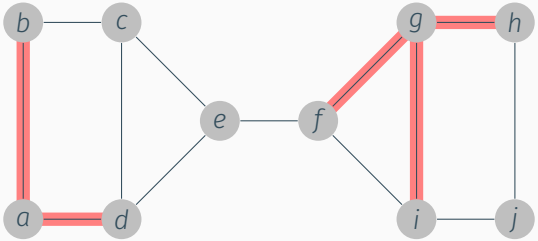


Distributed learning in a graph – The Gossip problem

Some agent talk to another connected agents: information spreads.

Everyone has a piece of information

Can we find global information using this type of information propagation?

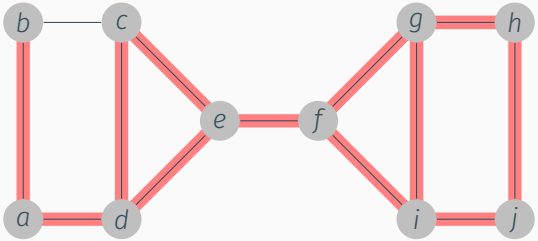


Distributed learning in a graph – The Gossip problem

Some agent talk to another connected agents: information spreads.

Everyone has a piece of information

Can we find global information using this type of information propagation?



Distributed learning in a graph – The Gossip problem

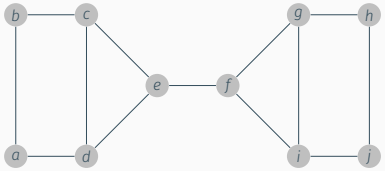
A network of agents is modeled by a undirected graph $G = (V, E)$

Each agent (node) $v \in V$ is given an observation $\xi(v) \in \mathbb{R}$ (typically $f_i(\theta)$ or $\nabla f_i(\theta)$).

Goal: distributive computation of the average

$$\mu := \frac{1}{|V|} \sum_{v \in V} \xi(v)$$

Applications: mixing gossip and optimization algorithms (Nedic and Ozdaglar (2009),...,Scaman et al. (2017))

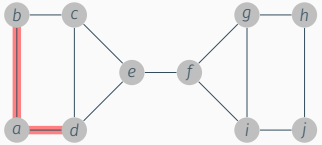


At each round, each agent $v \in V$ computes a weighted average of its neighbors:

$$\widehat{\mu}_0(v) = \xi(v), \qquad \widehat{\mu}_{t+1}(v) = \sum_{w \in V} W_{vw}\, \widehat{\mu}_t(w)\,.$$

where W is a Gossip matrix such that

- $W_{vw} \geqslant 0$ if $(v, w) \in E$ is an edge of the graph and $W_{vw} = 0$ otherwise
- W is symmetric and stochastic ($\sum_{w \in V} W_{vw} = 1$ for all v) $\quad \to$ Eigenvalues$(W) \subset [-1, 1]$



Pros:

- distributed (no sink)
- fault tolerant: failures of some node do not affect the protocol
- adapt to changing values: keep gossiping

Issues:

- converges to the true value but not exact
- convergence time: no guarantees before consensus is reached
- communication cost

At each round, each agent $v \in V$ computes a weighted average of its neighbors:

$$\widehat{\mu}_0(v) = \xi(v), \qquad \widehat{\mu}_{t+1}(v) = \sum_{w \in V} W_{vw}\, \widehat{\mu}_t(w)\,.$$

where W is a Gossip matrix such that

- $W_{vw} \geqslant 0$ if $(v, w) \in E$ is an edge of the graph and $W_{vw} = 0$ otherwise
- W is symmetric and stochastic ($\sum_{w \in V} W_{vw} = 1$ for all v) $\quad \to$ Eigenvalues$(W) \subset [-1, 1]$

Synchronous gossip (all nodes simultaneously) This can be re-written in the matrix form:

$$\widehat{\mu}_0 = \xi, \qquad \widehat{\mu}_{t+1} = W\, \widehat{\mu}_t = W^t \xi\,.$$

Convergence

Gossip: $\widehat{\mu}_t = W^t \xi$

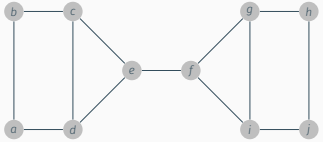
Convergence: The value of each node is diffusing in the network until uniform distribution

$$\widehat{\mu}_t(v) \underset{t \to +\infty}{\longrightarrow} \mu = \frac{1}{|V|} \sum_{v \in V} \xi(v)$$

How fast? The convergence rate depends on the eigengap of the Gossip matrix:

$$\gamma = \lambda_1(W) - \lambda_2(W) = 1 - \lambda_2(W)\,.$$

γ^{-1} is the mixing time associated with the Markov chain.



Classical gossip needs $\frac{1}{\gamma} \log \frac{1}{\varepsilon}$ iterations to reach precision ε.

Convergence

Gossip: $\widehat{\mu}_t = W^t \xi$

Convergence: Needs $\frac{1}{\gamma} \log \frac{1}{\varepsilon}$ iterations to reach precision ε.

Example: On the finite line of size n, we can take $W = I_n - \frac{1}{2}\Delta$ where $\Delta = D - A$ is the Laplacian of the graph.

In this case, we can show that the eigen-values are $\lambda_k(W) = \cos\left(2\pi \frac{k-1}{n}\right)$ and thus the eigengap

$$\gamma = \lambda_1(W) - \lambda_2(W) = 1 - \cos\left(\frac{2\pi}{n}\right) \approx \frac{1}{2}\left(\frac{2\pi}{n}\right)^2.$$

Convergence in $O(n^2 \log(1/\varepsilon))$.

What's wrong?

- Convergence in $O(n^2)$, while we would like n (after n steps we have observed all the data points) $\rightarrow$ accelerated gossip
- No exact convergence (may be solved through message passing algorithms for trees)
- Transient behavior: no guarantee before all points are observed $\rightarrow$ statistical gossip

Convergence

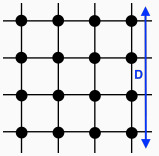
Gossip: $\widehat{\mu}_t = W^t \xi$

Convergence: Needs $\frac{1}{\gamma} \log \frac{1}{\varepsilon}$ iterations to reach precision ε.

How fast is the convergence? Hopefully:

convergence time $\approx$ diameter of the graph

so that each node get information from all other nodes.



This is not the case! Simple gossip is too slow.
In a d-dimensional grid: convergence is $\frac{1}{\gamma} \approx n^{2/d}$ while the diameter is $n^{1/d}$

Accelerated gossip

Gossip: $\widehat{\mu}_t = W^t \xi$

Convergence: Needs $\frac{1}{\gamma} \log \frac{1}{\varepsilon}$ iterations to reach precision ε.

Accelerared gossip:

- Chebytchev acceleration (Auzinger, 2011; Arioli and Scott, 2014)
- Shift-register gossip (Cao et al., 2006)
- Min-sum splitting (Rebeschini and Tatikonda, 2017)

The main idea of all these methods, similarly to Nesterov accelearation for gradient descent is to store past iterates

$$\tilde{\mu}_t = \sum_{s=1}^{t} \alpha_s \widehat{\mu}_s = \sum_{s=1}^{t} \alpha_s W^s \xi = P_t(W)\xi \,.$$

where P_t is a polynomial s.t. $\deg(P_t) \leqslant t$ and $P_t(1) = 1$.

$\rightarrow$ Replaces γ^{-1} with $\gamma^{-1/2}$ in the rates: convergence in $O\big(\frac{1}{\sqrt{\gamma}} \log \frac{1}{\varepsilon}\big)$

Back to our motivation from optimization

Goal of many optimization problems

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$$

Several works considered mixing gossip and gradient descent to perform decentralized optimization:

Nedic and Ozdaglar (2009); Duchi et al. (2012); Wei and Ozdaglar (2012); Iutzeler et al. (2013); Shi et al. (2015); Jakoveti´c et al. (2015); Nedich et al. (2016); Mokhtari et al. (2016); Colin et al. (2016); Scaman et al. (2017), etc.

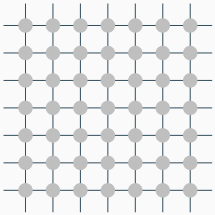
Using accelerated gossip Scaman et al. (2017) obtained optimal complexity:

Error ε achieved in $\sqrt{\kappa} \log \frac{1}{\varepsilon}$ gradient steps and $\sqrt{\kappa/\gamma} \log(1/\varepsilon)$ communication steps.

Problem for very large networks

Small eigengaps in many practical graphs:
for the d-dimensional grid $\gamma \approx n^{-\frac{2}{d}}$

If the network is very large $n \gg 1$, $O(\frac{1}{\sqrt{\gamma}})$ is still too long.



Problem of classical gossip analysis through eigengap: it does not take into account transient behaviors.

This is because classical gossip analysis is done for adversarial $\xi(v)$:

$$\widehat{\mu}_t(v) \underset{t\to+\infty}{\longrightarrow} \mu = \frac{1}{|V|}\sum_{v\in V}\xi(v)$$

$\rightarrow$ needs diffusion of the information through the entire graph before guarantees.

But in classical machine learning problems $\frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$ is not adversarial.

Statistical gossip (Berthier et al. 2018)

Random observations:

$$\xi(v) \quad \text{i.i.d.} \quad v \in V \qquad \text{with mean} \quad \mathbb{E}[\xi(v)] = \mu \quad \text{and variance} \quad \mathrm{Var}\big(\xi(v)\big) = \tau^2$$

Goal: estimate μ at every node as fast as possible

Criterion: minimize the expected squared error at every node

$$\mathbb{E}\big[(\widehat{\mu}_t(v) - \mu)^2\big] = \underbrace{\big(\mathbb{E}[\widehat{\mu}_t(v)] - \mu\big)^2}_{\text{bias squared}} + \underbrace{\mathbb{E}\Big[\big(\widehat{\mu}_t(v) - \mathbb{E}[\widehat{\mu}_t(v)]\big)^2\Big]}_{\text{variance: } \mathrm{Var}(\widehat{\mu}_t(v))}$$

If the estimator is unbiased $\mathbb{E}[\widehat{\mu}_t(v)] = \mu$, the goal is to minimize the variance $\mathrm{Var}\big(\widehat{\mu}_t(v)\big)$.

Statistical Gossip: algorithms

Simple Gossip:

$$\widehat{\mu}_0 = \xi, \quad \widehat{\mu}_t = W\widehat{\mu}_{t-1} \qquad \rightarrow \qquad \widehat{\mu}_t = W^t\xi$$

Polynomial Gossip:

$$\widehat{\mu}_t = P_t(W)\xi\,,$$

where P_t is a polynomial s.t. $\deg(P_t) \leqslant t$ and $P_t(1) = 1$.

Is simple gossip optimal?

How to choose the best polynomial?

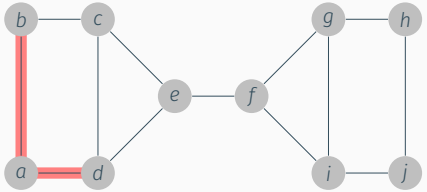
What performance can we expect? Lower-bound

Optimality

If $\widehat{\mu}_t = P_t(W)\xi$ is an unbiased estimator. For any $v \in V$ we have

$$\mathrm{Var}\big(\widehat{\mu}_t(v)\big) \geqslant \frac{\tau^2}{|B_t(v)|}$$

where $B_t(v) = \{w \in V : \ d(v,w) \leqslant t\}$ is the ball of radius t centered in v and $d(\cdot,\cdot)$ is the shortest path distance.



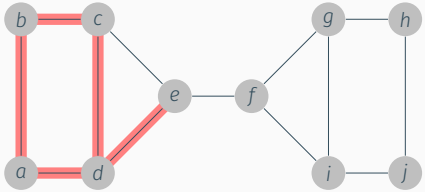
What performance can we expect? Lower-bound

Optimality

If $\widehat{\mu}_t = P_t(W)\xi$ is an unbiased estimator. For any $v \in V$ we have

$$\mathrm{Var}\big(\widehat{\mu}_t(v)\big) \geqslant \frac{\tau^2}{|B_t(v)|}$$

where $B_t(v) = \{w \in V : \ d(v,w) \leqslant t\}$ is the ball of radius t centered in v and $d(\cdot,\cdot)$ is the shortest path distance.



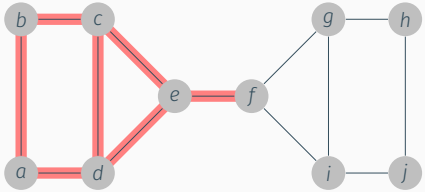
What performance can we expect? Lower-bound

Optimality

If $\widehat{\mu}_t = P_t(W)\xi$ is an unbiased estimator. For any $v \in V$ we have

$$\mathrm{Var}\big(\widehat{\mu}_t(v)\big) \geqslant \frac{\tau^2}{|B_t(v)|}$$

where $B_t(v) = \{w \in V : \ d(v,w) \leqslant t\}$ is the ball of radius t centered in v and $d(\cdot,\cdot)$ is the shortest path distance.



Polynomial Gossip: performance analysis

Polynomial Gossip: $\widehat{\mu}_t = P_t(W)\xi$, where P_t is a polynomial s.t. $\deg(P) \leqslant t$ and $P_t(1) = 1$.

Unbiased estimator: If $P_t(1) = 1$ **Performance**: minimize $\mathrm{Var}(\widehat{\mu}_t(v))$

Proposition

$$\mathrm{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} P_t(\lambda)^2 d\sigma_v(\lambda)$$

where $d\sigma_v(\lambda)$ is the spectral measure of W at v.

Finite graph of size n:

- W can be decomposed $W = \sum_{i=1}^{n} \lambda_i u_i u_i^\top$ and the spectral measure is discrete

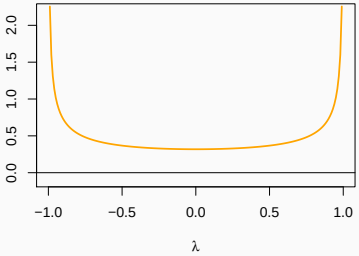
$$d\sigma_v = \frac{1}{n} \sum_{i=1}^{n} (u_i(v))^2 \delta_{\lambda_i} .$$

- The average variance is $\frac{1}{n} \sum_{v \in V} \mathrm{Var}(\widehat{\mu}_t(v)) = \frac{1}{n} \sum_{i=1}^{n} P_t(\lambda_i)^2$ where λ_i are the eigenvalues of W.
- Classical gossip analysis with eigengap $\gamma = 1 - \lambda_2$ $\quad \rightarrow$ here: all the measure

Example: infinite line (or one-dimensional grid)

Spectral measure: $d\sigma(\lambda) = \frac{d\lambda}{\pi\sqrt{1-\lambda^2}}$ $\rightarrow$ no eigengap

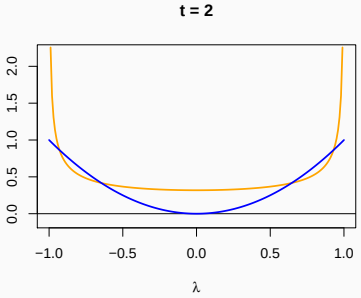
Simple gossip: $\mathrm{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx \frac{1}{\sqrt{t}}$ $\rightarrow$ suboptimal



Example: infinite line (or one-dimensional grid)

Spectral measure: $d\sigma(\lambda) = \frac{d\lambda}{\pi\sqrt{1-\lambda^2}}$ $\rightarrow$ no eigengap

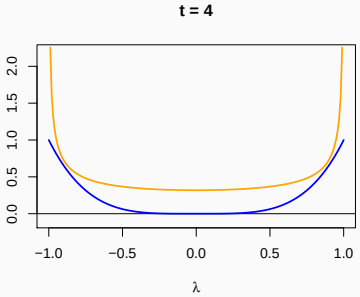
Simple gossip: $\mathrm{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx \frac{1}{\sqrt{t}}$ $\rightarrow$ suboptimal



Example: infinite line (or one-dimensional grid)

Spectral measure: $d\sigma(\lambda) = \frac{d\lambda}{\pi\sqrt{1-\lambda^2}}$ $\rightarrow$ no eigengap

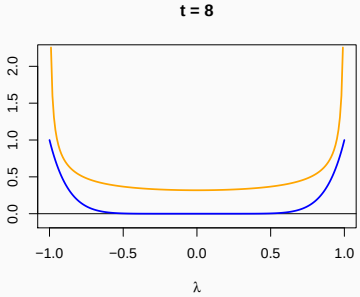
Simple gossip: $\mathrm{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx \frac{1}{\sqrt{t}}$ $\rightarrow$ suboptimal



Example: infinite line (or one-dimensional grid)

Spectral measure: $d\sigma(\lambda) = \frac{d\lambda}{\pi\sqrt{1-\lambda^2}}$ $\rightarrow$ no eigengap

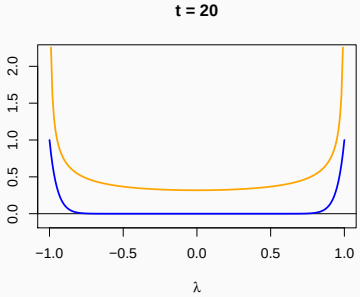
Simple gossip: $\mathrm{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx \frac{1}{\sqrt{t}}$ $\rightarrow$ suboptimal



Example: infinite line (or one-dimensional grid)

Spectral measure: $d\sigma(\lambda) = \frac{d\lambda}{\pi\sqrt{1-\lambda^2}}$ $\rightarrow$ no eigengap

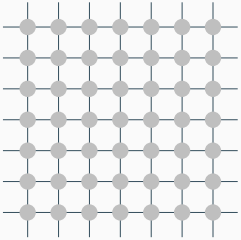
Simple gossip: $\mathrm{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx \frac{1}{\sqrt{t}}$ $\rightarrow$ suboptimal

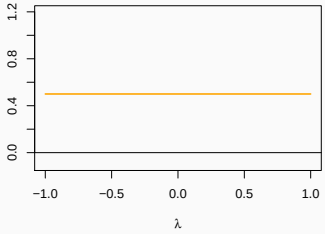


Example: two-dimensional grid

Spectral measure: $d\sigma(\lambda) = \text{Uniform}([-1, 1])$

Simple gossip: $\text{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx \frac{1}{t}$ $\rightarrow$ suboptimal

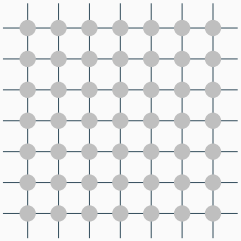


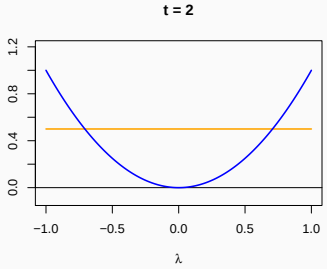


Example: two-dimensional grid

Spectral measure: $d\sigma(\lambda) = \text{Uniform}([-1, 1])$

Simple gossip: $\text{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx \frac{1}{t}$ $\rightarrow$ suboptimal

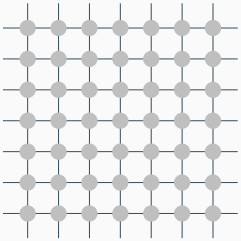


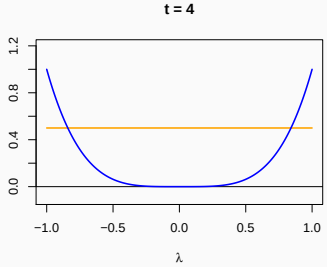


Example: two-dimensional grid

Spectral measure: $d\sigma(\lambda) = \text{Uniform}([-1, 1])$

Simple gossip: $\text{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx \frac{1}{t}$ $\rightarrow$ suboptimal

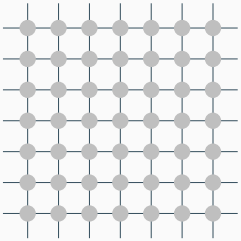


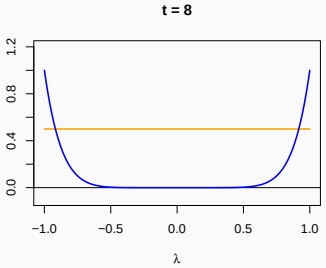


Example: two-dimensional grid

Spectral measure: $d\sigma(\lambda) = \text{Uniform}([-1, 1])$

Simple gossip: $\text{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx \frac{1}{t}$ $\rightarrow$ suboptimal

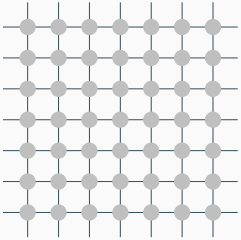


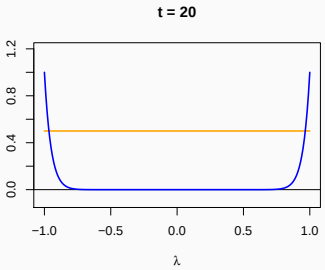


Example: two-dimensional grid

Spectral measure: $d\sigma(\lambda) = \text{Uniform}([-1, 1])$

Simple gossip: $\text{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx \frac{1}{t}$ $\rightarrow$ suboptimal

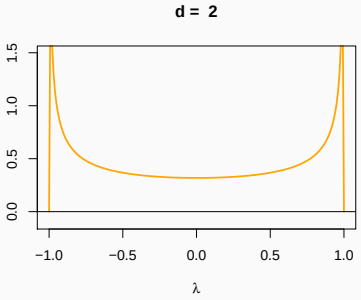




Example: d-regular tree

Spectral measure: support included in $\left[-2\frac{\sqrt{d-1}}{d}, 2\frac{\sqrt{d-1}}{d}\right]$ $\rightarrow$ eigengap for $d \geqslant 3$

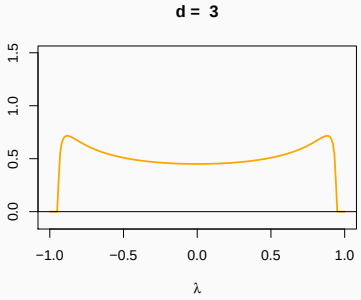
Simple gossip: $\mathrm{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx C^t$ for $d \geqslant 3$ $\rightarrow$ suboptimal



Example: d-regular tree

Spectral measure: support included in $\left[-2\frac{\sqrt{d-1}}{d}, 2\frac{\sqrt{d-1}}{d}\right]$ $\rightarrow$ eigengap for $d \geqslant 3$

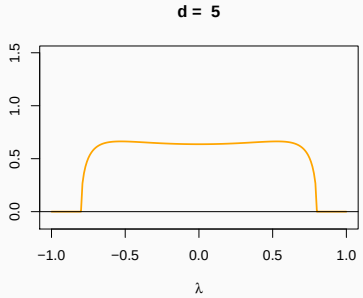
Simple gossip: $\mathrm{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx C^t$ for $d \geqslant 3$ $\rightarrow$ suboptimal



Example: d-regular tree

Spectral measure: support included in $\left[-2\frac{\sqrt{d-1}}{d}, 2\frac{\sqrt{d-1}}{d}\right]$ $\rightarrow$ eigengap for $d \geqslant 3$

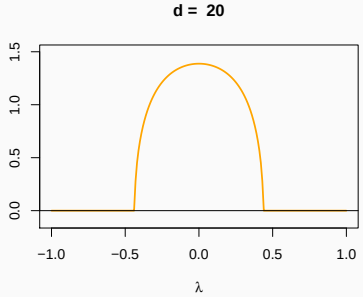
Simple gossip: $\mathrm{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx C^t$ for $d \geqslant 3$ $\rightarrow$ suboptimal



Example: d-regular tree

Spectral measure: support included in $\left[-2\frac{\sqrt{d-1}}{d}, 2\frac{\sqrt{d-1}}{d}\right]$ $\rightarrow$ eigengap for $d \geqslant 3$

Simple gossip: $\mathrm{Var}(\widehat{\mu}_t(v)) = \tau^2 \int_{-1}^{1} \lambda^t d\sigma(\lambda) \approx C^t$ for $d \geqslant 3$ $\rightarrow$ suboptimal

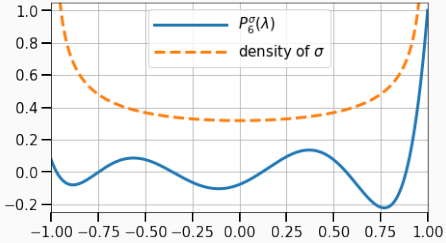


Polynomial Gossip: how to find the best polynomial?

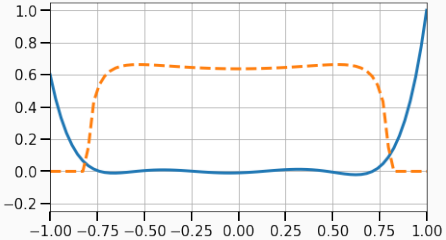
If we know σ, we want to find the polynomial minimizing:

$$P_t^{(\sigma)} \in \underset{P:\deg(P)\leqslant t, P(1)=1}{\arg\min} \int_{-1}^{1} P_t(\lambda)^2 d\sigma(\lambda)$$

The line $\mathbb{Z}$



The 5-regular tree: $\mathbb{T}_5$



How to find the best polynomial

If we know σ, we want to find the polynomial minimizing:

$$P_t^{(\sigma)} \in \underset{P:\deg(P)\leqslant t, P(1)=1}{\arg\min} \int_{-1}^{1} P_t(\lambda)^2 d\sigma(\lambda)$$

Classical problem in numerical linear algebra (Fischer (1996); Diekmann et al. (1999))

This functional is the square norm with respect to a scalar product:

$$\langle P, Q \rangle_\sigma = \int_{-1}^{1} P(\lambda)Q(\lambda)d\sigma(\lambda)\,.$$

Computation of $P_t^{(\sigma)}$ from orthogonal basis The minimization can then be done thanks to the orthogonal basis $\pi_0, \pi_1, \ldots, \pi_t$ of the set of polynomial with respect to $\langle \cdot\,, \cdot \rangle_\sigma$:

$$P_t^{(\sigma)} = \frac{1}{\sum_{s=0}^{t} \pi_s(1)} \sum_{s=0}^{t} \pi_s(1)\pi_s \qquad \leftarrow \quad \text{closed form solution}$$

Recursive computation of the orthogonal basis $\pi_0, \pi_1, \ldots, \pi_t$

Distributive computation of the best polynomial

Polynomial with the best polynomial can be computed via a second-order recursion (see Berthier et al. 2018 for details)

Computation formula			Result	
$x^{-1} = 0\,,$	$x^0 = \xi\,,$	$x^{t+1} = \frac{1}{a_{t+1}}\left(Wx^t - b_t x^t - a_t x^{t-1}\right)$	$x^t = \pi_t(W)\xi$	(1)
$\rho_{-1} = 0\,,$	$\rho_0 = 1\,,$	$\rho_{t+1} = \frac{1}{a_{t+1}}\left((1-b_t)\rho_t - a_t\rho_{t-1}\right)$	$\rho_t = \pi_t(1)$	(2)
$u^0 = \xi\,,$	$u^{t+1} = u^t$	$+\,\rho_{t+1}x^{t+1}$	$u^t = \sum_{s=0}^t \pi_s(1)\pi_s(W)\xi$	(3)
$v_0 = 1\,,$	$v_{t+1} = v_t$	$+\,\rho_{t+1}^2$	$v_t = \sum_{s=0}^t \pi_s(1)^2$	(4)
$\widehat{\mu}^t = u^t/v_t$			$\widehat{\mu}^t = P_t^\sigma(W)\xi$	(5)

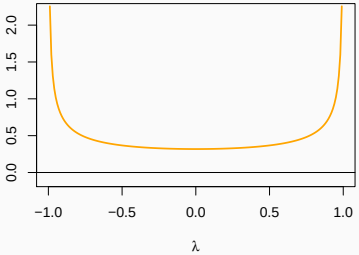
Issue: computation can be hard for some measures σ

Example: infinite line (or one-dimensional grid)

Spectral measure: $d\sigma(\lambda) = \frac{d\lambda}{\pi\sqrt{1-\lambda^2}}$

Simple gossip: $\frac{1}{\sqrt{t}}$ $\rightarrow$ suboptimal

Polynomial gossip with P_t^σ: $\frac{1}{t}$ $\rightarrow$ optimal

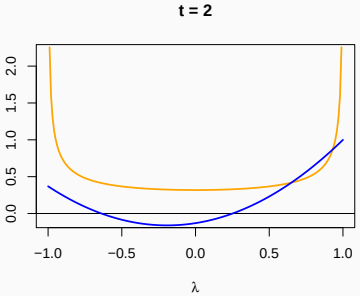


Example: infinite line (or one-dimensional grid)

Spectral measure: $d\sigma(\lambda) = \frac{d\lambda}{\pi\sqrt{1-\lambda^2}}$

Simple gossip: $\frac{1}{\sqrt{t}}$ $\rightarrow$ suboptimal

Polynomial gossip with P_t^σ: $\frac{1}{t}$ $\rightarrow$ optimal

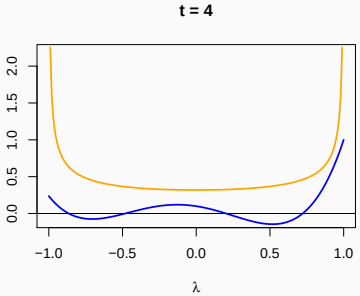


Example: infinite line (or one-dimensional grid)

Spectral measure: $d\sigma(\lambda) = \frac{d\lambda}{\pi\sqrt{1-\lambda^2}}$

Simple gossip: $\frac{1}{\sqrt{t}}$ $\rightarrow$ suboptimal

Polynomial gossip with P_t^σ: $\frac{1}{t}$ $\rightarrow$ optimal

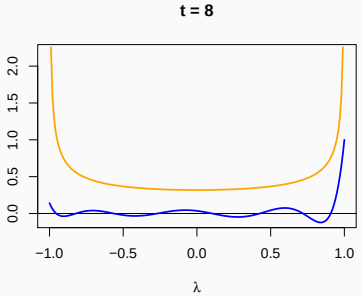


Example: infinite line (or one-dimensional grid)

Spectral measure: $d\sigma(\lambda) = \frac{d\lambda}{\pi\sqrt{1-\lambda^2}}$

Simple gossip: $\frac{1}{\sqrt{t}}$ $\rightarrow$ suboptimal

Polynomial gossip with P_t^σ: $\frac{1}{t}$ $\rightarrow$ optimal

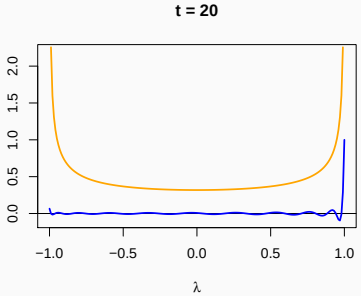


Example: infinite line (or one-dimensional grid)

Spectral measure: $d\sigma(\lambda) = \frac{d\lambda}{\pi\sqrt{1-\lambda^2}}$

Simple gossip: $\frac{1}{\sqrt{t}}$ $\rightarrow$ suboptimal

Polynomial gossip with P_t^σ: $\frac{1}{t}$ $\rightarrow$ optimal

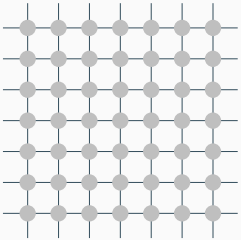


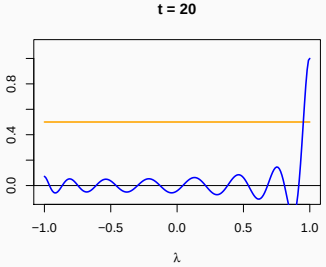
Example: two-dimensional grid

Spectral measure: $d\sigma(\lambda) = \text{Uniform}([-1, 1])$

Simple gossip: $\frac{1}{t}$ $\rightarrow$ suboptimal

Polynomial gossip with P_t^σ: $\frac{1}{t^2}$ $\rightarrow$ optimal

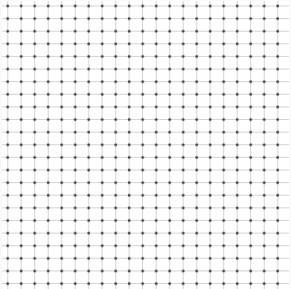


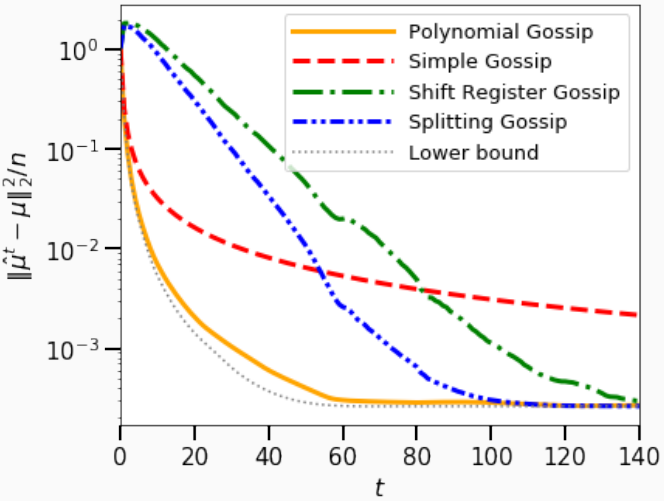


Simulations $2D$-Grid

2D-grid: 40×40

$\xi_v \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$



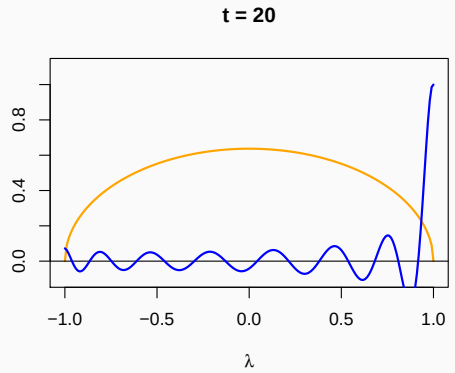


Example: d-dimensional grid

Spectral measure: $d\sigma(\lambda) \propto (1-\lambda)^{d/2-1} d\lambda$

Simple gossip: $t^{-d/2}$ $\rightarrow$ suboptimal

Polynomial gossip with P_t^σ: t^{-d} $\rightarrow$ optimal

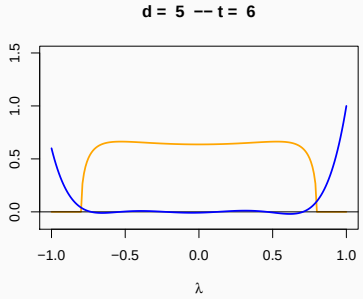


Example: d-regular tree ($d \geqslant 3$)

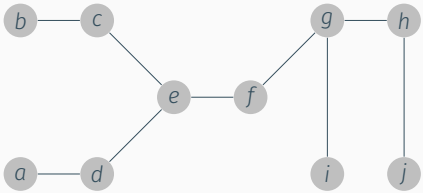
Spectral measure: support included in $\left[-2\frac{\sqrt{d-1}}{d}, 2\frac{\sqrt{d-1}}{d}\right]$ $\rightarrow$ eigengap for $d \geqslant 3$

Simple gossip: C^t $\rightarrow$ suboptimal

Polynomial gossip with P_t^σ: $(d-1)^t$ $\rightarrow$ optimal



Optimal algorithm for trees



For trees, the message passing algorithm (Moallemi and Roy, 2005) is exactly optimal.
→ No backtracking information.

Proposition

The polynomial gossip algorithm obtained from the optimal polynomial for the d-regular tree is exactly the message passing algorithm on any d-regular graph.

But we do not know σ for general graphs...

What if we do not know σ?

- Current solution: use an approximation (works well in practice)
- Future work: estimate as the process goes on in a distributive manner

In some graphs we can use approximations

Random regular graph: if G is a random d-regular graph on n vertices, its spectral measure converges to

$$d\sigma(\lambda) = \frac{d}{2\pi} \frac{\sqrt{\frac{4(d-1)}{d^2} - \lambda^2}}{1 - \lambda^2}$$

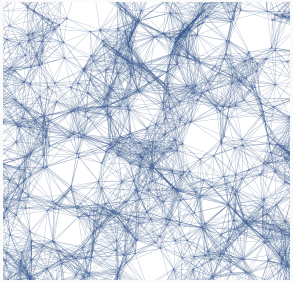
and we know the orthogonal polynomials with respect to σ!

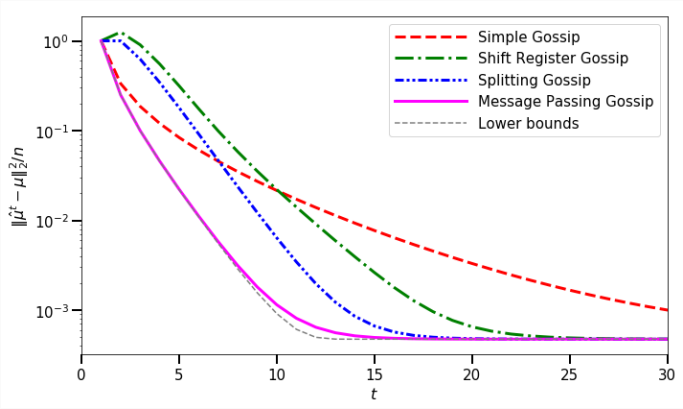
Simulations on random d-regular graphs

Random 3-regular graph:

$n = 2000$

$\xi_v \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$





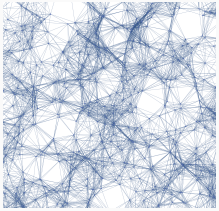
Simulations two-dimensional random geometric graphs

Random geometric graph: $n = 1600$

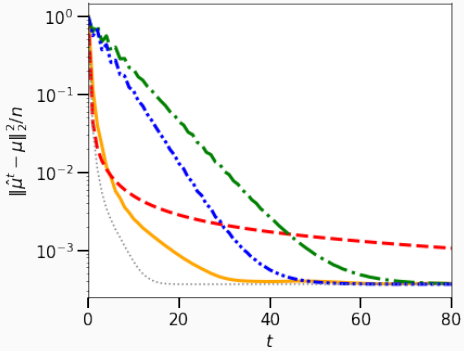
Vertices are sampled uniformly in $[0, 1]^2$

Edges between points closer than $3/\sqrt{n}$

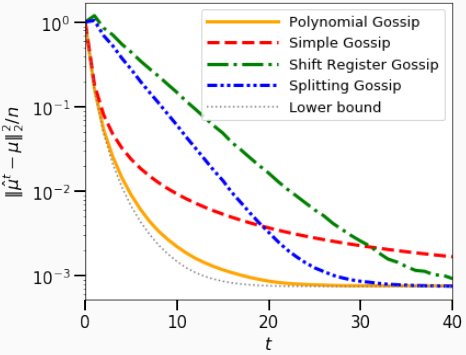
$\xi_v \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$

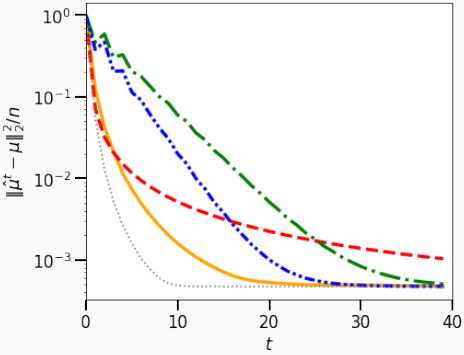


Using polynomials obtained from two-dimensional grid



Simulations 3D grid – geometric graphs





Conclusion

In averaging problem in a graph, the goal is to **reach consensus** (full averaging of the values in the graph) as quickly as possible.
→ Here new frame work for i.i.d. observation to provide better local guarantees before consensus is reached.

Future work:

- theoretical results on other large random graphs
- adaptive approximation of σ by the algorithm on arbitrary graph
- non i.i.d. observations ξ_v but smooth distribution with respect to the graph metric
- asynchronous algorithm

Thank you!


Berthier, R., F. Bach, and P. Gaillard. “Gossip of Statistical Observations using Orthogonal Polynomials”. In: *arXiv preprint arXiv:1805.08531* (2018).