On the mean square error of randomized averaging

Insights into the wisdom of crowds

Paolo Frasca

DISMA, Politecnico di Torino, Italy

based on joint work with J. M. Hendrickx (Uc Louvain)

Kolloquium Technische Kybernetik
Stuttgart, May 15, 2012

Outline

1. Averaging, consensus, and wisdom
2. Randomized averaging
3. Main result and applications
4. Remarks and conclusion

Populations and learning

- Unknown $\theta \in \mathbb{R}$ is the *state of the world*
- a population I of N agents takes noisy observations

$$y_i = \theta + n_i \qquad \text{for all } i \in I$$

- noises n_i are independent random variables:
  $\mathbb{E}[n_i] = 0$ and $\mathbb{E}[n_i^2] = \sigma^2$
- the population wishes to learn θ

  agents have beliefs $x_i(t)$, which are based on the observations and evolve in time through communication between agents, in order to (hopefully) approach θ

Collective estimation: an interdisciplinary issue

Social interpretation

People beliefs and evolving opinions about a topic

B. Golub and M. O. Jackson. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010

Technological interpretation

Sensor network: measurements and fusion/filtering

F. Garin and S. Zampieri. Mean square performance of consensus-based distributed estimation over regular geometric graphs. *SIAM Journal on Control and Optimization*, 50(1):306–333, 2012

In both cases, we ask the same **question**:

Does learning ensure that observation errors average away (when N is large)?

the answer depends on the belief dynamics!

Wisdom and averaging

A formal definition requires sequences of populations:

Definition (Wise population)

Take a sequence of populations $\{I_N\}_{N\in\mathbb{N}}$ of increasing size.
Assume $x_i(t) \to \alpha^{(N)}$ as $t \to \infty$ for all $i \in I$.
Then, I_N is said to be *wise* if

$$\lim_{N\to+\infty} \alpha^{(N)} = \theta \qquad \text{for all } i \in I$$

An ideal learning process would provide the population with the ML estimator of the state of the world: $\hat{\theta} = \frac{1}{N}\sum_i y_i$

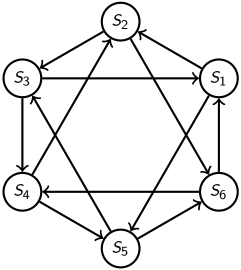
Note: $\mathbb{E}[(\theta - \hat{\theta})^2] = \frac{\sigma^2}{N} \quad \Rightarrow \quad$ if a population can compute $\hat{\theta}$, it is wise

In-network averaging: standard consensus algorithm

We represent the communication constraints among the agents by a network: communication is restricted to neighboring nodes.

An iterative averaging algorithm allows the population to compute $\hat{\theta}$:

$$\begin{cases} x_i(0) = y_i \\ x_i(t+1) = \sum_j a_{ij} x_j(t) \end{cases}$$



Note: synchronous communication,
$a_{ij} \geq 0$, $\sum_j a_{ij} = 1$
and $a_{ij} > 0$ according to the network

Proposition (Convergence and wisdom)

If the network is strongly connected and $\sum_i a_{ij} = 1$,

then $x_i(t) \to \alpha$ *as* $t \to \infty$ *for all* i, *and* $\alpha = \hat{\theta} = \frac{1}{N} \sum_i y_i$.

Obstacles to averaging: randomness

Issue: this simple algorithm may not always be used...

Examples:

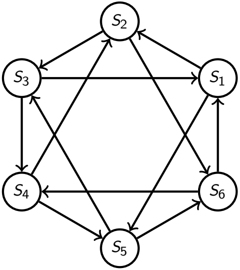
- packet losses (asymmetric link failures)
- asymmetric gossip approaches (by design, few links are active symultaneously)

In these cases, the population actually attempts to compute $\hat{\theta}$ by a rule which is properly described as stochastic:

- the a_{ij} are time-dependent random variables $a_{ij}(t)$,
- we can only know the statistics of $a_{ij}(t)$

then, $x_i(t+1) = \sum_j a_{ij}(t, \omega) x_j(t)$

Example: packet losses



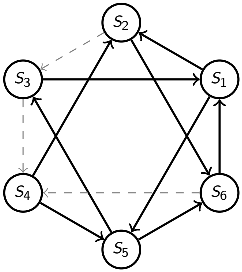
At each $t \in \mathbb{N}$:

- each message is lost with probability p;
- each node compensates missing information using her own state instead

F. Fagnani and S. Zampieri. Average consensus with packet drop communication. *SIAM Journal on Control and Optimization*, 48(1):102–133, 2009

P. Frasca and J. M. Hendrickx. Packet loss errors are negligible in large network averaging. In *IEEE Conf. on Decision and Control*, 2012. submitted

Example: packet losses



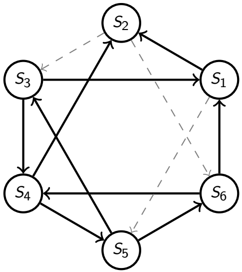
At each $t \in \mathbb{N}$:

- each message is lost with probability p;
- each node compensates missing information using her own state instead

F. Fagnani and S. Zampieri. Average consensus with packet drop communication. *SIAM Journal on Control and Optimization*, 48(1):102–133, 2009

P. Frasca and J. M. Hendrickx. Packet loss errors are negligible in large network averaging. In *IEEE Conf. on Decision and Control*, 2012. submitted

Example: packet losses



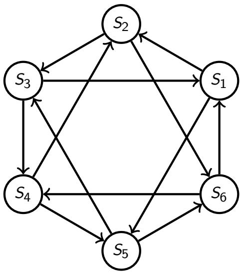
At each $t \in \mathbb{N}$:

- each message is lost with probability p;
- each node compensates missing information using her own state instead

F. Fagnani and S. Zampieri. Average consensus with packet drop communication. *SIAM Journal on Control and Optimization*, 48(1):102–133, 2009

P. Frasca and J. M. Hendrickx. Packet loss errors are negligible in large network averaging. In *IEEE Conf. on Decision and Control*, 2012. submitted

Example: Broadcast gossip



At each $t \in \mathbb{N}$:

- a node i is randomly chosen;
- her neighbors update as

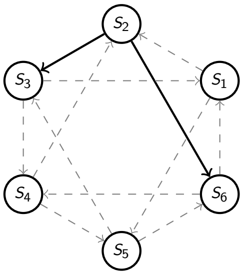
$$x_j(t+1) = (1-q)x_j(t) + q\, x_i(t)$$

for some $q \in (0,1)$

A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010

F. Fagnani and P. Frasca. The asymptotical error of broadcast gossip averaging algorithms. In *IFAC World Congress*, pages 10027–10031, Milan, Italy, August 2011

Example: Broadcast gossip



At each $t \in \mathbb{N}$:

- a node i is randomly chosen;
- her neighbors update as

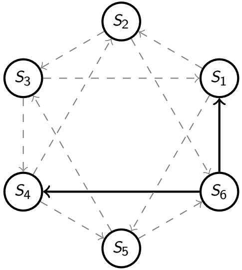
$$x_j(t+1) = (1-q)x_j(t) + q\,x_i(t)$$

for some $q \in (0,1)$

A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010

F. Fagnani and P. Frasca. The asymptotical error of broadcast gossip averaging algorithms. In *IFAC World Congress*, pages 10027–10031, Milan, Italy, August 2011

Example: Broadcast gossip



At each $t \in \mathbb{N}$:

- a node i is randomly chosen;
- her neighbors update as

$$x_j(t+1) = (1-q)x_j(t) + q\, x_i(t)$$

for some $q \in (0,1)$

A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010

F. Fagnani and P. Frasca. The asymptotical error of broadcast gossip averaging algorithms. In *IFAC World Congress*, pages 10027–10031, Milan, Italy, August 2011

Drawbacks of randomization

The population aims to compute $\hat{\theta}$ by a randomized rule:

- the a_{ij} are time-dependent random variables $a_{ij}(t)$,
- we can only know the statistics of $a_{ij}(t)$

Then, $x_i(t+1) = \sum_j a_{ij}(t,\omega)x_j(t)$

Effect: with probability 1, each $x_i(t)$ converges to α, but $\alpha \neq \hat{\theta}$

Question: How large is the induced error?

Mean square error estimate

Theorem (Probabilistic wisdom condition)

Let $A(t)$ *be the update matrix such that* $[A(t)]_{ij} = a_{ij}(t)$, I *the identity matrix, and* $\mathbf{1}$ *a vector of 1s of length* N. *If*

- $\sum_i \mathbb{E}[a_{ij}(t)] = 1$ *(1st order condition)*
- *it exists* $\gamma > 0$ *such that*
  $$\mathbb{E}[A(s)^* \mathbf{1}\mathbf{1}^* A(s)] \leq \gamma \left(I - \mathbb{E}[A(s)^* A(s)]\right) \qquad \text{(2st order condition)}$$

then
$$\mathbb{E}[(\frac{1}{N}\sum_i x_i(t) - \frac{1}{N}\sum_i x_i(0))^2] \leq \frac{\gamma}{N}(\frac{1}{N}\sum_i x_i^2(0)) \qquad \forall\, t \geq 0$$

Mean square error estimate

Theorem (Probabilistic wisdom condition)

Let $A(t)$ be the update matrix such that $[A(t)]_{ij} = a_{ij}(t)$, I the identity matrix, and $\mathbf{1}$ a vector of 1s of length N. If

- $\sum_i \mathbb{E}[a_{ij}(t)] = 1$ *(1st order condition)*
- *it exists $\gamma > 0$ such that*
  $$\mathbb{E}[A(s)^* \mathbf{1}\mathbf{1}^* A(s)] \leq \gamma \left(I - \mathbb{E}[A(s)^* A(s)]\right) \qquad \text{(2st order condition)}$$

then
$$\mathbb{E}[(\frac{1}{N}\sum_i x_i(t) - \frac{1}{N}\sum_i x_i(0))^2] \leq \frac{\gamma}{N}(\frac{1}{N}\sum_i x_i^2(0)) \qquad \forall\, t \geq 0$$

Actually, at convergence:

$$\mathbb{E}[(\alpha - \hat{\theta})^2] \leq \frac{\gamma}{N}\sigma^2$$

Corollary & examples

A population satisfying the above condition is wise:

$$\mathbb{E}[(\alpha-\theta)^2] \le \mathbb{E}[(\underbrace{\alpha-\hat{\theta}}_{\text{randomization error}} + \underbrace{\hat{\theta}-\theta}_{\text{estimator error}})^2] \le \frac{1+\gamma}{N}\sigma^2$$

(cf: increasing the number of samples improves the estimate)

Corollary & examples

A population satisfying the above condition is wise:

$$\mathbb{E}[(\alpha-\theta)^2] \leq \mathbb{E}[(\underbrace{\alpha-\hat\theta}_{\text{randomization error}} + \underbrace{\hat\theta-\theta}_{\text{estimator error}})^2] \leq \frac{1+\gamma}{N}\sigma^2$$

(cf: increasing the number of samples improves the estimate)

Examples (assuming balanced networks):

Packet loss: let packet loss probability p, $\bar a = \max_i \sum_{j\neq i} a_{ij}$

$$\gamma = \frac{\bar a}{1-\bar a}(1-p).$$

Broadcast: let q update gain, $d_{\max}$ largest degree

$$\gamma = \frac{q}{1-q}\frac{d_{\max}}{N}$$

Remarks

- the proof uses a probabilistic method, based on a key remark: the current average is a martingale, *i.e.*

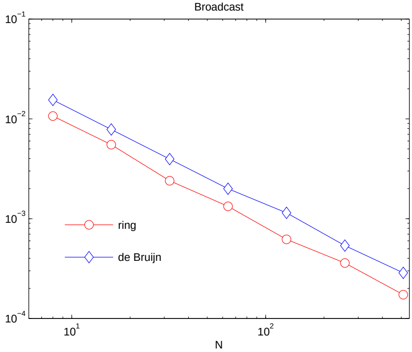
$$\mathbb{E}[\frac{1}{N}\sum_i x_i(t+1)|x(t)] = \frac{1}{N}\sum_i x_i(t)$$

Remarks

- the proof uses a probabilistic method, based on a key remark: the current average is a martingale, *i.e.*

$$\mathbb{E}[\frac{1}{N}\sum_i x_i(t+1)|x(t)] = \frac{1}{N}\sum_i x_i(t)$$

- the result is independent of convergence properties



Remarks II

- "robustness" result:
  under mild assumptions, *asymmetric asynchronous* averaging is effective!

- tight bounds (compared with simulations)

- wide application:
  available results cover algorithms featuring
  - small number of concurrent updates
  - little correlation between updates

  (over balanced graphs)

P. Frasca and J. M. Hendrickx. On the mean square error of randomized averaging algorithms. *Automatica*, November 2011. submitted

Current and future work

1. General result on the role of correlation between entries of $P(t)$ (cf. law of large numbers)

2. Extension to non-doubly-stochastic $\mathbb{E}[P(t)]$ (implies $\mathbb{E}[\alpha] \neq \theta$, useful for unbalanced graphs)

3. Social science applications
   a. Naïve learning and random interactions
   b. Resilience to fluctuations in economic networks

      D. Acemoglu, V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi. The network origins of aggregate fluctuations. *Econometrica*, 2012. to appear

Current and future work

1. General result on the role of correlation between entries of $P(t)$ (cf. law of large numbers)

2. Extension to non-doubly-stochastic $\mathbb{E}[P(t)]$ (implies $\mathbb{E}[\alpha] \neq \theta$, useful for unbalanced graphs)

3. Social science applications
   a. Naïve learning and random interactions
   b. Resilience to fluctuations in economic networks

   D. Acemoglu, V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi. The network origins of aggregate fluctuations. *Econometrica*, 2012. to appear

Thank you for your attention