Asynchronism and convergence rates in distributed optimization

Hamid Reza Feyzmahdavian, Arda Aytekin and Mikael Johansson
KTH - Royal Institute of Technology

Optimization as iterative algorithms

Many optimization algorithms are iterations, *e.g.*

$$x(t+1) = x(t) - \gamma \nabla f(x(t)) := \mathcal{M}x(t)$$

Optimizer $x^\star$ is a fixed-point of $\mathcal{M}$.

Easy to analyze when $\mathcal{M}$ is a **contraction mapping**

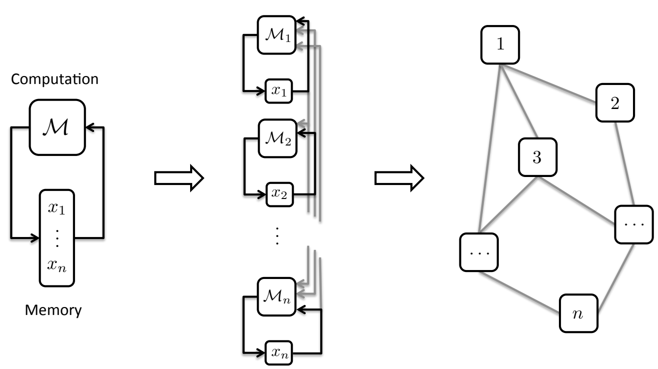
$$\|\mathcal{M}x - \mathcal{M}y\| \le c\|x-y\| \qquad \forall x, y \in \mathbb{R}^n$$

for some $c \in [0,1)$ and some norm $\|\cdot\|$. Then $\|x(t) - x^\star\| \le c^t \|x(0) - x^\star\|$

Ex. Gradient mapping when f is μ-strongly convex with L-Lipschitz gradient

Distributed implementations and asynchrony

Emerging applications require **distributed implementations**

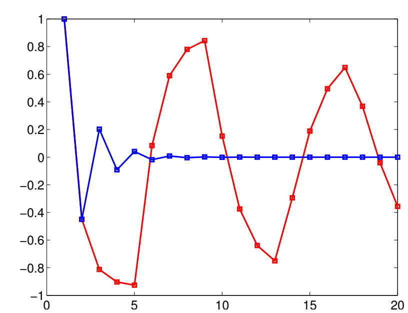


Communication delays, lack of synchronization ⇒ **asynchronous iterations**

The impact of asynchrony

Asynchrony can cause otherwise stable iterations to diverge, or slow down.

$$x_1(t+1) = x_1(t) - 0.75x_1(t) - 0.7x_2(t - \tau(t))$$
$$x_2(t+1) = x_2(t) - 0.75x_2(t) - 0.7x_1(t - \tau(t))$$



Need models and tools for asynchronous iterations!

A model for asynchronous iterations

A standard form for **asynchronous iterations**:

$$x_i(t+1) = \begin{cases} \mathcal{M}_i(x_1(\tau_1^i(t)), \ldots, x_n(\tau_n^i(t))) & \text{if } t \in \mathcal{T}^i \\ x_i(t) & \text{otherwise} \end{cases}$$

Here,

$\mathcal{T}^i$ is the set of times when node i executes an update, and

$\tau_j^i(t)$ is the time when the most recent version of x_j available to node i at time t was computed

Note: Can view $t - \tau_j^i(t)$ as information delay from node j to i at time t

Chazan and Miranker (1969), Baudet (1978), Bertsekas and Tsitsiklis (1989), ...

Partially asynchronous algorithms

The iteration

$$x_i(t+1) = \begin{cases} \mathcal{M}_i(x_1(\tau_1^i(t)), \ldots, x_n(\tau_n^i(t))) & \text{if } t \in \mathcal{T}^i \\ x_i(t) & \text{otherwise} \end{cases}$$

is called **partially asynchronous** if there exists $B > 0$ such that

a) For every i, t, at least one element of $\{t, t+1, \ldots, t+B-1\}$ is in $\mathcal{T}^i$.
b) For every i, j and all $t \in \mathcal{T}^i$, we have $0 \leq t - \tau_j^i(t) \leq B - 1$.
c) There holds $\tau_i^i(t) = t$ for all i and all $t \in \mathcal{T}^i$

Bounded update intervals/information delays, direct access to "own" state

Totally asynchronous algorithms

The iteration

$$x_i(t+1) = \begin{cases} \mathcal{M}_i(x_1(\tau_1^i(t)), \ldots, x_n(\tau_n^i(t))) & \text{if } t \in \mathcal{T}^i \\ x_i(t) & \text{otherwise} \end{cases}$$

is called **totally asynchronous** if

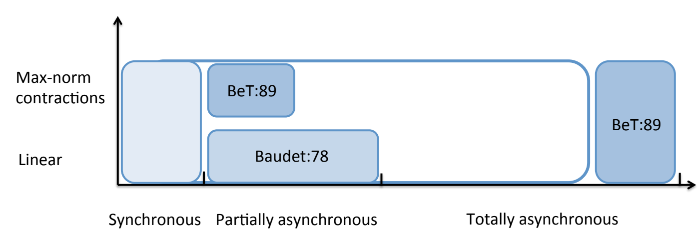
a) every set $\mathcal{T}^i$ is an infinite subset of $\mathbb{N}_0$
b) for every sequence $\{t_k\}$ of elelements of $\mathcal{T}^i$ that tends to infinity, it holds that $\lim_{k\to\infty} \tau_j^i(t_k) = \infty$ for all i, j.

No node ceases to update, old information eventually purged out of system.

Challenge: quantify the impact of asynchronism

We address two key questions:

1. quantify how B impacts convergence of partially asynchonus iterations
2. establish convergence rates for classes of totally asynchonous iterations



We then use this insight to design delay-insensitive optimization algorithms

Outline

Problem formulation

Consider iterations

$$x(t+1) = \mathcal{M}x(t)$$

where $\mathcal{M}$ is a **pseudo-contraction**

$$\|\mathcal{M}x - x^\star\| \leq c\|x - x^\star\| \qquad \forall x \in \mathbb{R}^n$$

with respect to a **block-maximum norm**

$$\|x\|_b^w = \max_{1 \leq i \leq m} \frac{\|x_i\|_i}{w_i}$$

(here $x = (x_1, \dots, x_m) \in \mathbb{R}^n$, $x_i \in \mathbb{R}^{n_i}$ and $\|\cdot\|_i$ is any norm)

Challenge: Quantify the impact of asynchrony on the iterates.

Our approach

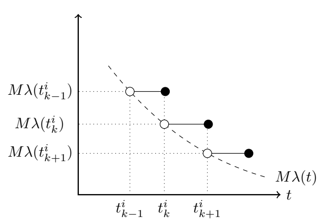
Use a continuous decreasing function $\lambda : \mathbb{R}_+ \mapsto \mathbb{R}_+$ satisfying

$$\lim_{t \to \infty} \lambda(t) = 0$$

and show that there is $M > 0$ such that

$$\frac{1}{w_i}\|x_i(t) - x_i^\star\|_i \leq M\lambda(t_k^i), \qquad \forall t \in (t_k^i, t_{k+1}^i]$$

for all i, all t and every pair of consecutive elements t_k^i and t_{k+1}^i in $\mathcal{T}^i$.



Main result

Theorem 1. If

a) $\mathcal{M}$ is pseudo-contraction with modulus c w.r.t. block-maximum norm

b) There exist functions $\beta^i : \mathbb{R}_+ \mapsto \mathbb{R}_+$ and $\Delta \in \mathbb{N}_0$ such that, $\forall t \geq \Delta$

$$t - t_k^i \leq \beta^i(t) \leq t \qquad t \in (t_k^i, t_{k+1}^i]$$

for every two consecutive elements t_k^i and t_{k+1}^i in $\mathcal{T}^i$.

c) There is a decreasing function $\lambda : \mathbb{R}_+ \mapsto \mathbb{R}_+$ with $\lim_{t \to \infty} \lambda(t) = 0$ and

$$c \lim_{t \to \infty} \frac{\lambda(\tau_j^i(t) - \beta^j(\tau_j^i(t)))}{\lambda(t)} < 1 \qquad \forall i, j$$

Then, the sequence generated by (2) under total asynchronism satisfies

$$\frac{1}{w_i}\|x_i(t) - x_i^\star\|_i \leq M\lambda(t_k^i), \qquad t \in (t_k^i, t_{k+1}^i]$$

for all i and all t, where M is a positive constant.

Main result (partially asynchronous iterations)

Theorem 2. Let $\mathcal{M}$ be a pseudo-contraction in the block-maximum norm. Then, the iterates generated by (2) under partial asynchronism satisfy

$$\frac{1}{w_i}\|x_i(t) - x_i^\star\| \leq M\rho^{t_k^i} \qquad t \in (t_k^i, t_{k+1}^i]$$

for every pair of consecutive elements t_k^i and t_{k+1}^i in $\mathcal{T}^i$. Moreover,

$$\rho = c^{\frac{1}{2B-1}}$$

Note. Convergence rate still linear. Slows down with increasing B.

Proof uses Theorem 1 with $\beta^i(t) = B$ and $\lambda(t) = \rho^t$.

Example ("retarding divider")

Consider the iteration

$$x(t+1) = \begin{cases} \frac{1}{2}x(t), & t \in \mathcal{T} \\ x(t), & t \notin \mathcal{T} \end{cases}$$

where $x(t) \in \mathbb{R}$ and $\mathcal{T} = \{2^k \mid k \in \mathbb{N}_0\}$.

Since $t_{k+1} - t_k = 2^k$, there is no uniform upper bound on inter-update times.

However, since

$$t - t_k \leq \frac{t}{2} \leq t \qquad \forall t \in (t_k, t_{k+1}]$$

$\beta(t) = t/2$ and $\lambda(t) = 1/t$ satisfy conditions of Theorem 1. It follows that

$$|x(t)| \leq \frac{M}{t_k}, \qquad t \in (t_k, t_{k+1}]$$

Main result (linearly bounded delays)

Theorem 3. If

a) $\mathcal{M}$ is a pseudo-contraction with modulus c w.r.t. a block-maximum norm

b) For each $t \in \mathcal{T}^i$, there exists $t' \in \mathcal{T}^i$ such that $1 \leq t' - t \leq B$.

c) It holds that $0 \leq t - \tau_j^i(t) \leq \alpha t$ for all i, j and all $t \geq t_\alpha$.

Then, the sequence generated by (2) under total asynchronism satisfies

$$\frac{1}{w_i}\|x_i(t) - x_i^\star\|_i \leq M\left(\frac{t_k^i}{B} + 1\right)^{-\zeta} \qquad t \in (t_k^i, t_{k+1}^i]$$

where $\zeta = \ln c / \ln(1-\alpha)$.

Note. Bounded by polynomial function of time. Slower as delays increase.

Discussion: iterate time vs. physical time

Upper bound decreases only at iteration times, stays constant in between.

In physical time, convergence rate depends on how update times grow large.

For partially asynchronous iterations $t - B \leq t_k^i$ for $t \in (t_k^i, t_{k+1}^i]$, so

$$M\rho^{t_k^i} \leq M\rho^{t-B} := M'\rho^t, \quad t \in (t_k^i, t_{k+1}^i]$$

Thus,

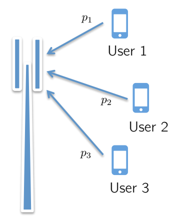
$$\frac{1}{w_i}\|x_i(t) - x_i^\star\|_i \leq M'\rho^t,$$

so error decays as $O(\rho^t)$

Application: wireless power control

User i transmits at power p_i, tries to maintain SINR target γ_i

$$\text{SINR}_i = \frac{g_{ii}p_i}{\sum_{j\neq i} g_{ij}p_j + \nu_i} \geq \gamma_i$$



Transmit powers that minimize total energy satisfy

$$\frac{g_{ii}p_i}{\sum_{j\neq i} g_{ij}p_j + \nu_i} = \gamma_i$$

or, equivalently

$$p_i = I_i(p)$$

where $I_i : \mathbb{R}^n_+ \to \mathbb{R}_+$ is the **interference function**.

Application: wireless power control

Transmit power control implements fixed-point iteration

$$p_i(t+1) = I_i(p(t))$$

Definition 1. $I : \mathbf{R}^n_+ \mapsto \mathbf{R}^n_+$ is a c-**contractive interference function** if

a) $I_i(p) \geq 0$

b) If $p \geq p'$ then $I_i(p) \geq I_i(p')$

c) There exists $c \in [0, 1)$ and a vector $v > 0$ such that for all $\epsilon > 0$

$$I_i(p + \epsilon v) \leq I_i(p) + c\epsilon v_i$$

Proposition. If $I : \mathbb{R}^n_+ \mapsto \mathbb{R}^n_+$ is a c-contractive interference function, then it has a unique fixed-point $p^\star \in \mathbb{R}^n_+$ and

$$\|I(p) - I(p')\|^v_\infty \leq c\|p - p'\|^v_\infty$$

Application: wireless power control

Corollary. Consider the asynchronous power control iteration, and assume

a) every mobile updates its power at least once every B time units, and

b) no information is more than $D_{\max}$ time units old.

If $I(p)$ is a c-contractive interference function, then

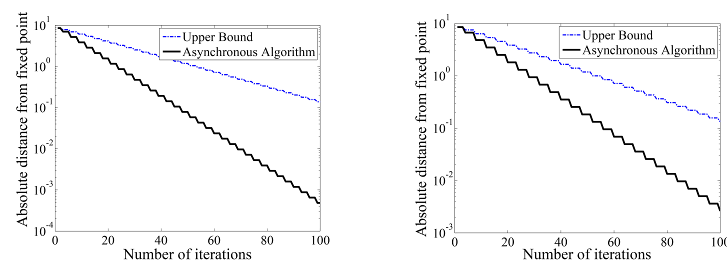
$$\frac{1}{v_i}|p_i(t) - p_i^\star| \leq M\rho^{t^i_k}, \qquad t \in (t^i_k, t^i_{k+1}]$$

where $M > 0$ and t^i_k and t^i_{k+1} are consecutive elements of $\mathcal{T}^i$. Moreover,

$$\rho = c^{\frac{1}{B+D_{\max}}}$$

Application: wireless power control

Simulations and bounds for two users in a four-user scenario



Linear interference functions, $B = D_{\max} = 4$.

Bounds valid, but not tight (for these users)

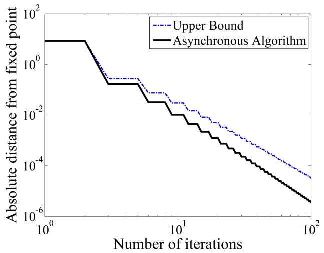
Application: wireless power control

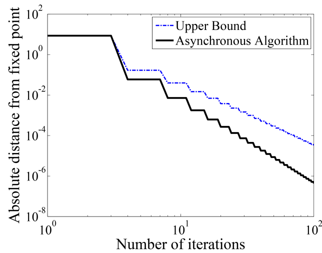
Assume that information delay for user 1 grows increasingly large

$$t - \tau_j^1(t) = t - \tau_1^j(t) = \lfloor 0.1t \rfloor$$

while other delays, execution times remain unchanged.

Simulations and bounds from Theorem 3.





Proof sketch

Theorem 1 (recollection and interpretation) If

a) $\mathcal{M}$ is pseudo-contraction with modulus c w.r.t. block-maximum norm

b) There exist functions $\beta^i : \mathbb{R}_+ \mapsto \mathbb{R}_+$ and $\Delta \in \mathbb{N}_0$ such that, $\forall t \geq \Delta$

$$t - t_k^i \leq \beta^i(t) \leq t \qquad t \in (t_k^i, t_{k+1}^i]$$

for every two consecutive elements t_k^i and t_{k+1}^i in $\mathcal{T}^i$.

c) There is a decreasing function $\lambda : \mathbb{R}_+ \mapsto \mathbb{R}_+$ with $\lim_{t\to\infty} \lambda(t) = 0$ and

$$c \lim_{t\to\infty} \frac{\lambda(\tau_j^i(t) - \beta^j(\tau_j^i(t)))}{\lambda(t)} < 1 \qquad \forall i, j$$

Then, the sequence generated by (2) under total asynchronism satisfies

$$\frac{1}{w_i}\|x_i(t) - x_i^\star\|_i \leq M\lambda(t_k^i), \qquad t \in (t_k^i, t_{k+1}^i]$$

for all i and all t, where M is a positive constant.

Proof sketch

Step 1. Find initial time $\bar{t}$ such that hypotheses satisfied for $t = 0, \ldots, \bar{t}$:

Let t_0^i be smallest element of $\mathcal{T}^i$. By total asynchronism, there is $\hat{t}$ such that

$$\tau_j^i(t) \geq \max\{\Delta, \max_{1\leq i\leq m} t_0^i + 1\} \qquad \forall t \geq \hat{t}$$

By condition c), we can find $\tilde{t}$ such that

$$c\lambda\left(\tau_j^i(t) - \beta^j(\tau_j^i(t))\right) \leq \lambda(t) \qquad \forall t \geq \tilde{t}$$

Let $\bar{t} = \max\{\hat{t}, \tilde{t}\}$ and define $M = \|x(0) - x^\star\|_b^w / \lambda(\bar{t})$.

Since $\{x \mid \|x(t) - x^\star\|_b^w \leq \|x(0) - x^\star\|_b^w\}$ is invariant and $\lambda(t)$ decreasing

$$\frac{1}{w_i}\|x_i(t) - x_i^\star\|_i \leq M\lambda(t_k^i), \qquad t \in (t_k^i, t_{k+1}^i].$$

for all $t = 0, \ldots, \bar{t}$.

Proof sketch

Step 2. Induction: assume true until t', show that it holds for $t' + 1$.

First consider $t' \in \mathcal{T}^i$, and define $k' : t' \in (t'_k, t'_k + 1]$. Then, by a)

$$\frac{1}{w_i}\|x_i(t'+1) - x_i^\star\|_i \leq c \max_{1\leq j\leq m}\left\{\frac{1}{w_j}\|x_j(\tau_j^i(t')) - x_j^\star\|_j\right\}$$

Noting that $\tau_j^i(t') \leq t'$, we apply the induction hypothesis and find

$$\frac{1}{w_j}\|x_j(\tau_j^i(t')) - x_j^\star\|_j \leq M\lambda(t_{k_\tau}^j) \leq M\lambda(t_j^i(t') - \beta^j(\tau_j^i(t'))) \leq \frac{M}{c}\lambda(t')$$

It thus holds

$$\frac{1}{w_i}\|x_i(t'+1) - x_i^\star\|_i \leq M\lambda(t') = M\lambda(t_{k'+1}^i)$$

Since $t' + 1 \in (t_{k+1}^i, t_{k+2}^i]$, the assertion holds for $t' + 1$. ($t' \notin \mathcal{T}^i$ trivial)

Outline

1. Motivation
2. Problem formulation
3. Convergence rates of asynchronous iterations
4. Example: power control in wireless systems
5. A delayed incremental gradient method with linear convergence rate
6. Conclusions

So far. . .

Established rather general convergence estimates for asynchronous iterations.

Psuedo-contraction in block-maximum norm essential to analysis.

When the gradient iteration

$$x(t+1) = x(t) - \gamma \nabla f(x(t))$$

is a contraction mapping, this is typically w.r.t. the Euclidean norm.

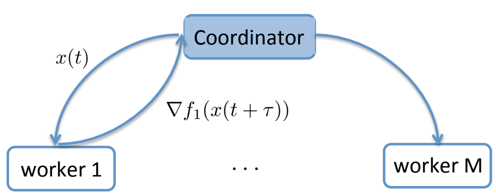
Can we use our insight to **design** delay-insensitive optimization algorithms?

Delayed incremental gradient methods

Common set-up in machine-learning applications:

$$\text{minimize} \quad \frac{1}{M}\sum_{m=1}^{M} f_m(x)$$

Centralized coordinator, workers that compute delayed (partial) gradients



Computational delay time-varying, update order sometimes stochastic

Agarwal and Duchi (2011), Niu, Recht et al (2011),

State-of-the art

The Hogwild! algorithm by Niu, Recht, et al (2011)

$$i(t) = \mathcal{U}[1, M]$$
$$x(t+1) = x(t) - \gamma \nabla f_{i(t)}(x(t-\tau(t)))$$

Converges linearly to ball around origin.

Limitations:

- Analysis asumes strong convexity **and** bounded gradients (!)
- Convergence proof valid for **one** particular value of γ.
- Step-size depends on M, max-delay and gradient norms at optimum

Note. Iterations mixing delayed and current states often hard to analyze.

Delayed gradient iterations

Instead of updating based on delayed gradient

$$x(t+1) = x(t) - \gamma \nabla f(x(t-\tau(t)))$$

we consider updating based on delayed gradient mapping,

$$x(t+1) = x(t-\tau(t)) - \gamma \nabla f(x(t-\tau(t))) \quad (1)$$

Proposition 1. Let f be μ-strongly convex and have L-Lipschitz continuous gradient. If $0 \le \tau(t) \le \tau_{\max}$ for all t, then $\{x(t)\}$ generated by (1) satisfies

$$\|x(t) - x^\star\| \le \left(\frac{\kappa - 1}{\kappa + 1}\right)^{\frac{t}{\tau_{\max}+1}}$$

where $\kappa = L/\mu$.

Delayed gradient iterations: quadratic objective functions

Consider minimization of the quadratic function

$$f(x) = \frac{1}{2}(Lx_1^2 + \mu x_2^2)$$

with $\tau(t) = 1$ for all t.

Then, delayed gradient iteration has convergence factor

$$c_g = \frac{\kappa}{\kappa + 1}$$

while the delayed prox iteration has convergence factor

$$c_p = \frac{\sqrt{\kappa^2 - 1}}{\kappa + 1} < c_g$$

Potentially faster **and** easier to analyze

Our algorithm

To minimize

$$f(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x)$$

we propose the following algorithm

$$\begin{aligned} i(t) &= \mathcal{U}[1, M] \\ s(t) &= x(t-\tau(t)) - \gamma \nabla f_{i(t)}(x(t-\tau(t))) \\ x(t+1) &= (1-\theta)x(t) + \theta s(t) \end{aligned}$$

Main result

Theorem 4. Assume that

a) each f_m is convex and has L_m-Lipschitz gradient on $\mathbb{R}^n$

b) the overall objective f is μ-strongly convex

Then, if $\gamma \in (0, \mu / \max_m L_m^2)$ the iterates generated by our method satisfy

$$\mathbf{E}_{t-1}[f(x(t))] - f^\star \le c^t(f(x(0)) - f^\star) + e$$

with

$$c = \left(1 - 2\gamma\mu\theta\left(1 - \gamma\frac{\max_m L_m^2}{\mu}\right)\right)^{1/(\tau_{\max}+1)}$$

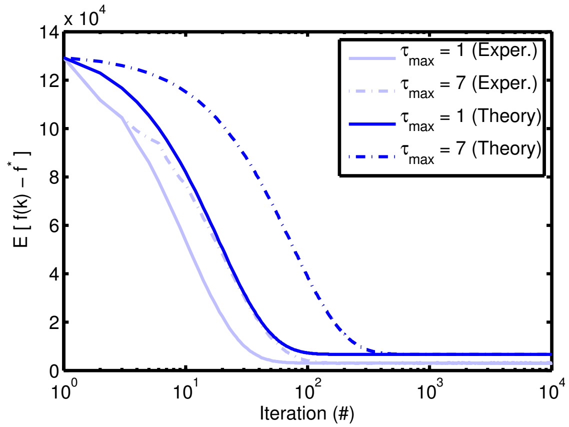
and

$$e = \frac{\gamma \max_m L_m}{2M(\mu - \gamma \max_m L_m^2)} \sum_{m=1}^{M} \|\nabla f_m(x^\star)\|$$

Note. Linear convergence to ball around optimum. Error/speed trade-off.

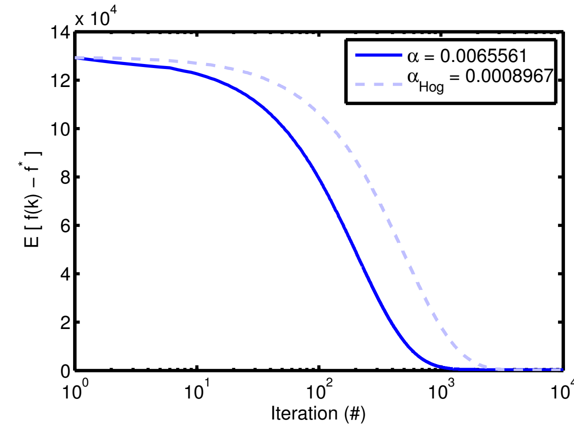
Numerical results

Representative convergence behaviour



Numerical results

Comparison with Hogwild!



Our algorithm converges faster with theoretically justified stepsizes.

Proof sketch

Lemma 5. Let $\{V(t)\}$ be a sequence of real numbers satsfying

$$V(t+1) \leq pV(t) + q \max_{t-\tau(t)\leq s\leq t} V(s) + r$$

for some non-negative numbers p, q and r. If $p+q<1$, and

$$0 \leq \tau(t) \leq \tau_{\max}$$

Then,

$$V(t) \leq c^t V(0) + e$$

where $c = (p+q)^{1/(1+\tau_{\max})}$ and $e = r/(1-p-q)$.

Proof sketch

Proof of Lemma 5. First note that since $p+q<1$,

$$1 \leq (p+q)^{-\tau_{\max}/(1+\tau_{\max})}$$

so, since $c=(p+q)^{1/(1+\tau_{\max})}$,

$$p + qc^{-\tau_{\max}} = p + q(p+q)^{-\frac{\tau_{\max}}{1+\tau_{\max}}} \leq (p+q)(p+q)^{-\frac{\tau_{\max}}{1+\tau_{\max}}} = c$$

Assertion holds for $t=0$. Assume that it holds for $t=0,\ldots\bar{t}$. Then

$$V(\bar{t}) \leq c^{\bar{t}}V(0)+e, \qquad V(s) \leq c^s V(0) + e \quad s = \bar{t}-\tau_{\max},\ldots,\bar{t}$$

We then have

$$V(\bar{t}+1) \leq pc^{\bar{t}}V(0) + pe + q(\max_{\bar{t}-\tau(\bar{t})\leq s\leq \bar{t}} c^s)V(0) + qe + r$$
$$\leq pc^{\bar{t}}V(0) + pe + qc^{\bar{t}-\tau_{\max}}V(0) + qe + r = c^{\bar{t}+1}V(0) + e.$$

Proof sketch

Proof of Theorem 4. Consider

$$V(t+1) = \mathbf{E}_t f(x(t+1)) - f^\star = \mathbf{E}_{t-1}\left[\mathbf{E}_{t|t-1}[f(x(t+1))]\right] - f^\star$$

Since f is convex and $\theta \in [0,1]$,

$$\begin{aligned} f(x(t+1)) - f^\star &= f((1-\theta)x(t) + \theta s(t)) - f^\star \\ &\leq (1-\theta)(f(x(t)) - f^\star) + \theta(f(s(t)) - f^\star) \end{aligned}$$

We establish the following bound on $f(s(t)) - f^\star$:

$$\begin{aligned} \mathbf{E}_{t|t-1}[f(s(t))] - f^\star &\leq \left(1 - 2\mu\gamma\left(1 - \frac{\alpha \max_m L_m^2}{\mu}\right)\right)(f(x(t-\tau(t))) - f^\star) \\ &\quad + \frac{\gamma^2 \max_m L_m}{M}\sum_{m=1}^{M}\|\nabla f_m(x^\star)\|^2 \end{aligned}$$

Allows to express $V(t+1)$ in terms of $V(t), \dots V(t-\tau_{\max})$ plus error term.

Proof sketch

Specifically,

$$\begin{aligned} V(t+1) &\leq (1-\theta)V(t) \\ &\quad + \theta\left(1 - 2\mu\gamma\left(1 - \frac{\gamma \max_m L_m^2}{\mu}\right)V(t-\tau(t))\right) \\ &\quad + \frac{\theta\gamma^2 \max_m L_m}{M}\sum_{m=1}^{M}\|\nabla f_m(x^\star)\|^2 \end{aligned}$$

So Lemma 5 now yields

$$V(t) \leq c^t V(0) + e \qquad \forall t \in \mathbb{N}_0$$

with the desired convergence factors and error terms.

Conclusions

- Convergence analysis of asynchronous iterations
- A general theorem covering both totally and partially asynchronism
- Asynchronism affects rates, not only factors

- A delayed incremental gradient method
- Running averages of delayed incremental gradient mappings
- Converges faster, and under less restrictive assumptions, than alternatives

- Not everything is in "the book" - many open problems!

References

Complete statements and proofs can be found in

H. R. Feyzmahdavian and M. Johansson,
"On the convergence rate of asynchronous iterations",
In IEEE CDC 2014, Los Angeles, CA, December 2014.

H. R. Feyzmahdavian, A. Aytekin and M. Johansson,
"A delayed proximal gradient method with linear convergence rate",
In IEEE MLSP 2014, Reims, France, September 2014.