Distributed Optimization with Flexible Communications

Franck Iutzeler LJK, Univ. Grenoble Alpes

PGMO Days 2020

Composite minimization

$$\text{Find} \quad x^\star \in \arg\min_{x\in\mathbb{R}^n} \quad \mathcal{R}\left(x; \{a_i, b_i\}_{i=1}^m\right) \quad + \quad \lambda\, r(x)$$

$$\text{Find} \quad x^\star \in \arg\min_{x\in\mathbb{R}^n} \quad f(x) \quad + \quad g(x)$$

$$\text{smooth} \qquad\qquad \text{non-smooth}$$

- f: differentiable surrogate of the empirical risk $\Rightarrow$ **Gradient**
  non-linear smooth function that depends on all the data
- g: non-smooth but chosen regularization $\Rightarrow$ **Proximity operator**
  non-differentiability on some manifolds implies structure on the solutions

$$\mathbf{prox}_{\gamma g}(u) = \arg\min_{y\in\mathbb{R}^n} \left\{ g(y) + \tfrac{1}{2\gamma}\|y-u\|_2^2 \right\}$$

closed form/easy for many regularizations:

- $g(x) = \|x\|_1$
- $g(x) = TV(x)$
- $g(x) = indicator_C(x)$

Natural optimization method: **proximal gradient**

$$\begin{cases} u_{k+1} = x_k - \gamma \nabla f(x_k) \\ x_{k+1} = \mathbf{prox}_{\gamma g}(u_{k+1}) \end{cases}$$

and its stochastic variants: proximal sgd, etc.

Example: LASSO

$$\begin{array}{rrcccc} \text{Find} & x^\star \in \arg\min\limits_{x\in\mathbb{R}^n} & \mathcal{R}\left(x; \{a_i, b_i\}_{i=1}^m\right) & + & \lambda\ r(x) \\ \text{Find} & x^\star \in \arg\min\limits_{x\in\mathbb{R}^n} & \frac{1}{2}\left\|Ax-b\right\|_2^2 & + & \lambda\|x\|_1 \\ & & \text{smooth} & & \text{non-smooth} \end{array}$$

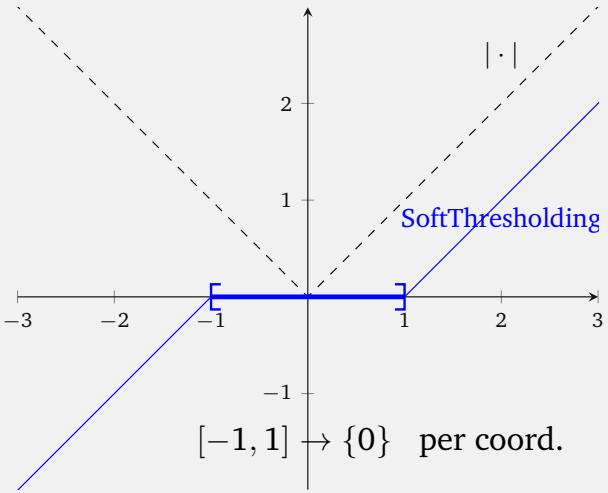
Coordinates	**Structure**	$\leftrightarrow$	**Optimality conditions**
$\forall i$	$x_i^\star = 0$	$\Leftrightarrow$	$A_i^\top (Ax^\star - b) \in [-\lambda, \lambda]$

Proximity Operator: per coordinate

$$\left[\mathbf{prox}_{\gamma\lambda\|\cdot\|_1}(u)\right]_i = \begin{cases} u_i - \lambda\gamma & \text{if } u_i > \lambda\gamma \\ 0 & \text{if } u_i \in [-\lambda\gamma; \lambda\gamma] \\ u_i + \lambda\gamma & \text{if } u_i < -\lambda\gamma \end{cases}$$

Proximal Gradient (aka ISTA):

$$\begin{cases} u_{k+1} = x_k - \gamma A^\top (Ax_k - b) \\ x_{k+1} = \mathbf{prox}_{\gamma\lambda\|\cdot\|_1}(u_{k+1}) \end{cases}$$



Example: LASSO

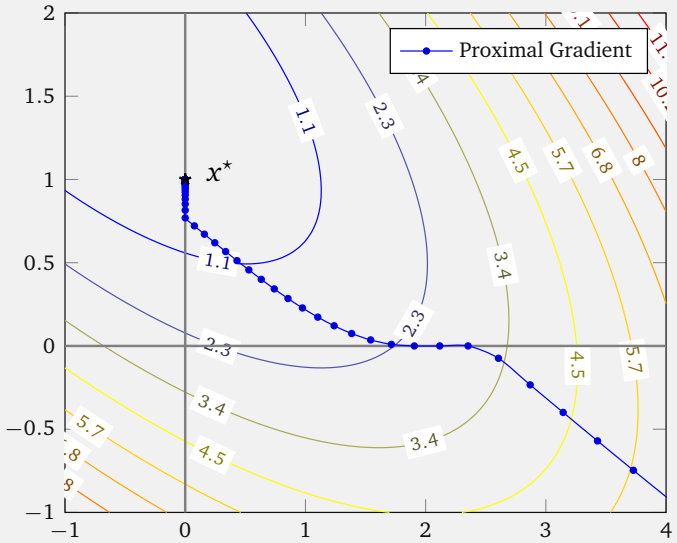
$$\begin{array}{rrcll} \text{Find} & x^\star \in \arg\min\limits_{x\in\mathbb{R}^n} & \mathcal{R}\left(x; \{a_i, b_i\}_{i=1}^m\right) & + & \lambda\ r(x) \\ \text{Find} & x^\star \in \arg\min\limits_{x\in\mathbb{R}^n} & \frac{1}{2}\left\|Ax-b\right\|_2^2 & + & \lambda\|x\|_1 \\ & & \text{smooth} & & \text{non-smooth} \end{array}$$

Coordinates	**Structure**	$\leftrightarrow$	**Optimality conditions**	$\leftrightarrow$	**Proximity operation**
$\forall i$	$x_i^\star = 0$	$\Leftrightarrow$	$A_i^\top (Ax^\star - b) \in [-\lambda, \lambda]$	$\Leftrightarrow$	$\left[\mathbf{prox}_{\gamma\lambda\|\cdot\|_1}(u^\star)\right]_i = 0$
					$u^\star = x^\star - \gamma A^\top (Ax^\star - b)$

$$\left[\mathbf{prox}_{\gamma\lambda\|\cdot\|_1}(u)\right]_i = \begin{cases} u_i - \lambda\gamma & \text{if } u_i > \lambda\gamma \\ 0 & \text{if } u_i \in [-\lambda\gamma; \lambda\gamma] \\ u_i + \lambda\gamma & \text{if } u_i < -\lambda\gamma \end{cases}$$

Proximal Gradient (aka ISTA):

$$\begin{cases} u_{k+1} = x_k - \gamma A^\top (Ax_k - b) \\ x_{k+1} = \mathbf{prox}_{\gamma\lambda\|\cdot\|_1}(u_{k+1}) \end{cases}$$



Iterates (x_k) reach the **same structure as $x^\star$** in **finite** time!
In practice: After that **identification**: fixed structure and better rate

Algorithm

Worker i updates w/ local data (f_i)

$$x_i^{k+1/2} = x^k - \gamma \nabla f_i(x^k)$$

for all $i = 1, .., M$

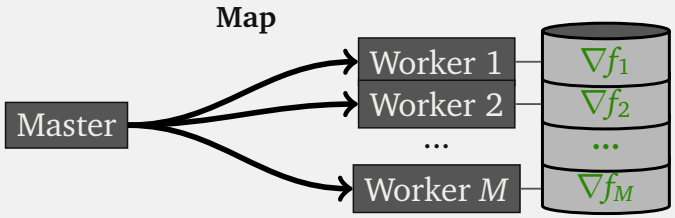
Master gathers the local variables

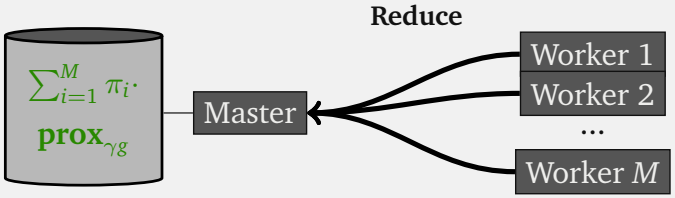
$$\bar{x}^{k+1} = \sum_{i=1}^M \pi_i x_i^{k+1/2}$$

Master performs a proximity operation

$$x_1^{k+1} = .. = x_M^{k+1} = \mathbf{prox}_{\gamma g}\left(\bar{x}^{k+1}\right)$$

Implementation





Distributed Proximal Gradient

Master:

Initialize $\bar{x} = \bar{x}^0$,
while *not converged* **do**
  when *all* workers have finished:
  Receive (x_i) from each of them
  $\bar{x} \leftarrow \sum_{i=1}^M \pi_i x_i$
  $x \leftarrow \mathbf{prox}_{\gamma g}(\bar{x})$
  Broadcast x to all agents
  $k \leftarrow k + 1$
Interrupt all slaves
Output x

Worker i:

Initialize $x = x_i = \bar{x}$,
while *not interrupted by master* **do**
  Receive the most recent x
  $x_i \leftarrow x - \gamma \nabla f_i(x)$
  Send x_i to the master

$$f_i(x) = \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \ell_j(x)$$

Local risk at worker i

Communications may soon become the bottleneck in distributed learning, hence the rise of asynchronous methods.

DAve-PG

Master:

Initialize $\bar{x}$
while *not converged* **do**
  when a worker finishes:
  Receive adjustment Δ from it
  $\bar{x} \leftarrow \bar{x} + \Delta$
  $x \leftarrow \mathbf{prox}_{\gamma g}(\bar{x})$
  Send x to the agent in return
  $k \leftarrow k + 1$
Interrupt all slaves
Output x

Worker i:

Initialize $x = x_i = \bar{x}$,
while *not interrupted by master* **do**
  Receive the most recent x
  $x_i \leftarrow x - \gamma \nabla f_i(x)$
  $\Delta \leftarrow \pi_i \left(x_i - x_i^{prev} \right) \quad x_i^{prev} \leftarrow x_i$
  Send adjustment Δ to master

$$f_i(x) = \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \ell_j(x)$$

Local risk at worker i

- With *sparsity inducing* regularizers (eg. ℓ_1 norm), *master-to-worker* communications will eventually become sparse $\Rightarrow$ **Use it!**
  identification of proximal methods
- Unfortunately, *worker-to-master* communications stay dense...
  Idea: sparsify adaptively!

- Machine Learning problems often have a *noticeable structure*;
  sparsity, low rank
- This structure is identified progressively by *proximal methods*;
  + CD, Var. Red., Distributed methods, etc.
- For most problem, we *do not know* if the identified structure is optimal;
  adaptivity is key
- Whenever structure appears, it can often be used numerically
  storage, communications
- Importance of conditionning and adaptation frequency
  to achieve best theoretical performance

▷ Grishchenko, I., Malick, Amini: *Distributed Learning with Sparse Communications by Identification*, 2020 `https://arxiv.org/abs/1812.03871`

Thanks to IDEX UGA IRS *DOLL* & PGMO

Thank you! – Franck IUTZELER `http://www.iutzeler.org`