Distributed Proximal Splitting Algorithms with Rates and Acceleration

Laurent Condat

King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Grigory Malinovsky

Peter Richtárik

arXiv:2010.00952

Distributed Optimization

$$\underset{x \in \mathcal{X}}{\text{minimize}} \left\{ \Psi(x) := \frac{1}{M} \sum_{m=1}^{M} \Big(F_m(x) + H_m(K_m x) \Big) + R(x) \right\}$$

Distributed Optimization

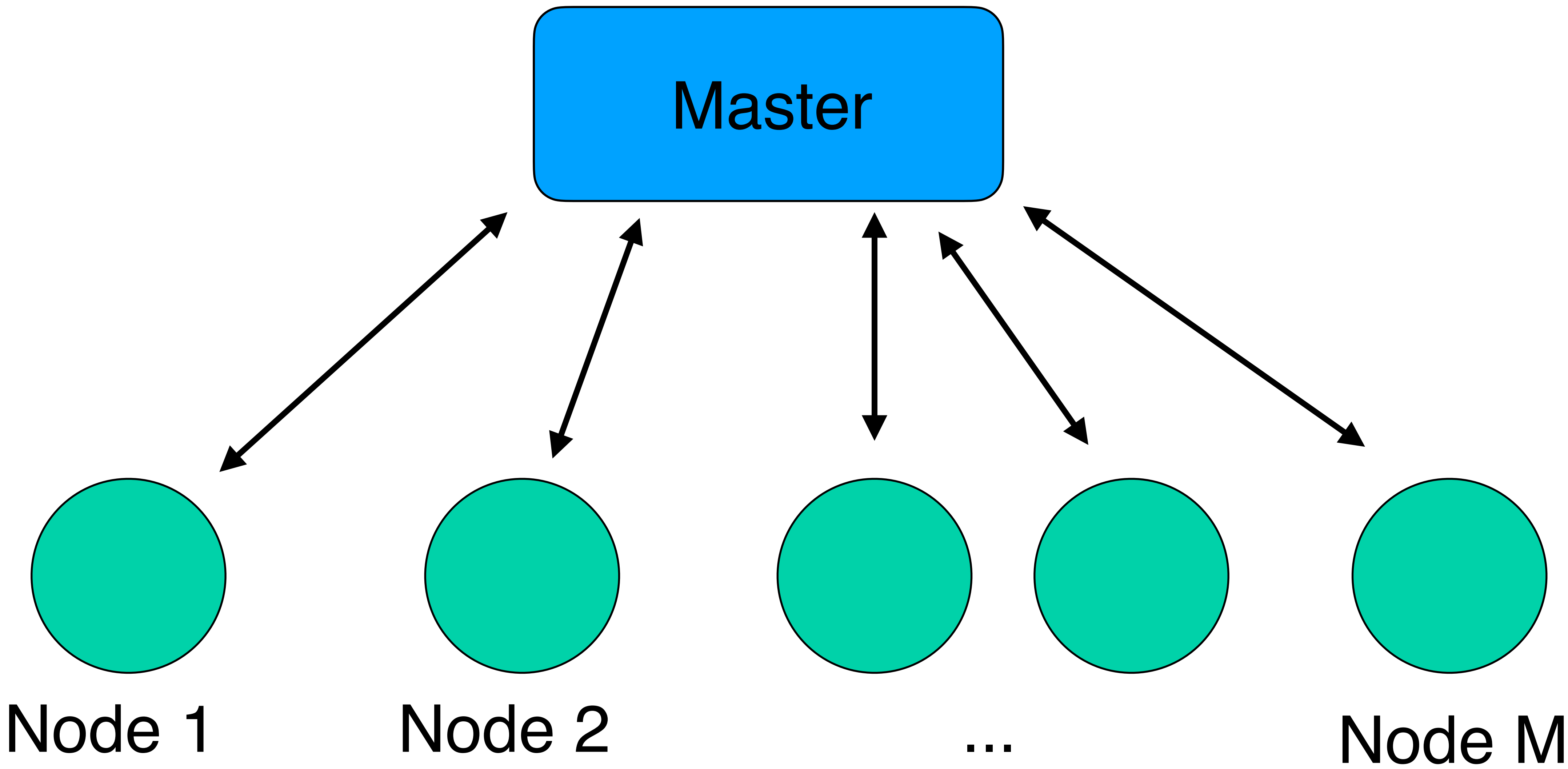
$$\underset{x \in \mathcal{X}}{\text{minimize}} \left\{ \Psi(x) := \frac{1}{M} \sum_{m=1}^{M} \Big(F_m(x) + H_m(K_m x) \Big) + R(x) \right\}$$

with:

- convex functions F_m, H_m, R
- F_m is L_{F_m}-smooth
- linear operators $K_m : \mathcal{X} \to \mathcal{U}_m$

Distributed Optimization

$$\underset{x \in \mathcal{X}}{\text{minimize}} \left\{ \Psi(x) := \frac{1}{M} \sum_{m=1}^{M} \Big(F_m(x) + H_m(K_m x) \Big) + R(x) \right\}$$



Distributed Optimization

$$\underset{x \in \mathcal{X}}{\text{minimize}} \left\{ \Psi(x) := \frac{1}{M} \sum_{m=1}^{M} \Big(F_m(x) + H_m(K_m x) \Big) + R(x) \right\}$$

Full splitting proximal algorithms:

- Iterative fixed-point algorithms with calls to ∇F_m, prox_{H_m}, prox_R, K_m, K_m^*
- No other operation

Distributed Optimization

$$\underset{x \in \mathcal{X}}{\text{minimize}} \left\{ \Psi(x) := \frac{1}{M} \sum_{m=1}^{M} \Big(F_m(x) + H_m(K_m x) \Big) + R(x) \right\}$$

Full splitting proximal algorithms:

- Iterative fixed-point algorithms with calls to ∇F_m, prox_{H_m}, prox_R, K_m, K_m^*
- No other operation

Condat et al., "Proximal Splitting Algorithms: A Tour of Recent Advances, with New Twists," arXiv:1912.00137

Proximal Splitting Algorithms

$$\text{Find } x^\star \in \arg\min_{x \in \mathcal{X}} \Big\{ F(x) + R(x) + H(Kx) \Big\}$$

with $K : \mathcal{X} \to \mathcal{U}$

Proximal Splitting Algorithms

$$\text{Find } x^\star \in \arg\min_{x \in \mathcal{X}} \left\{ F(x) + R(x) + H(Kx) \right\}$$

Fermat's rule ☞

$$0 \in \nabla F(x^\star) + \partial R(x^\star) + K^* \partial H(Kx^\star)$$

Proximal Splitting Algorithms

Find $x^\star \in \arg\min_{x \in \mathcal{X}} \left\{ F(x) + R(x) + H(Kx) \right\}$

Fermat's rule ☞

$$0 \in \nabla F(x^\star) + \partial R(x^\star) + K^* \partial H(Kx^\star)$$

$$\equiv$$

$$\begin{cases} 0 \in \nabla F(x^\star) + \partial R(x^\star) + K^* u^\star \\ 0 \in -Kx^\star + \partial H^*(u^\star) \end{cases}$$

Proximal Splitting Algorithms

$$\text{Find } x^\star \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \left\{ F(x) + R(x) + H(Kx) \right\}$$

Fermat's rule

$$0 \in \nabla F(x^\star) + \partial R(x^\star) + K^* \partial H(Kx^\star)$$

$$\equiv$$

$$\begin{cases} x^\star = \operatorname{prox}_{\gamma R}\left(x^\star - \gamma \nabla F(x^\star) - \gamma K^* u^\star\right) \\ u^\star = \operatorname{prox}_{H^*/(\gamma\eta)}\left(u^\star + \frac{1}{\eta\gamma} K x^\star\right) \end{cases}$$

KAUST

Proximal Splitting Algorithms

Find $x^\star \in \underset{x \in \mathcal{X}}{\arg\min} \left\{ F(x) + R(x) + H(Kx) \right\}$

Fermat's rule ☞

$$0 \in \nabla F(x^\star) + \partial R(x^\star) + K^* \partial H(Kx^\star)$$

$$\equiv$$

$$\begin{cases} x^\star = \mathrm{prox}_{\gamma R}\left(x^\star - \gamma \nabla F(x^\star) - \gamma K^* u^\star\right) \\ u^\star = \mathrm{prox}_{H^*/(\gamma\eta)}\left(u^\star + \frac{1}{\eta\gamma} K x^\star\right) \end{cases}$$

algorithm: iterate $(x^k, u^k) \mapsto (x^{k+1}, u^{k+1})$

Proximal Splitting Algorithms

Condat–Vu algorithm form I

$$\left\lfloor \begin{array}{l} x^{k+1} = \mathrm{prox}_{\gamma R}\big(x^k - \gamma \nabla F(x^k) - \gamma K^* u^k\big) \\ u^{k+1} = \mathrm{prox}_{H^*/(\gamma\eta)}\big(u^k + \frac{1}{\gamma\eta} Kx(2x^{k+1} - x^k)\big) \end{array} \right.$$

Condat–Vu algorithm form II

$$\left\lfloor \begin{array}{l} u^{k+1} = \mathrm{prox}_{H^*/(\gamma\eta)}\big(u^k + \frac{1}{\gamma\eta} Kx^k\big) \\ x^{k+1} = \mathrm{prox}_{\gamma R}\big(x^k - \gamma \nabla F(x^k) - \gamma K^* (2u^{k+1} - u^k)\big) \end{array} \right.$$

Condat, “A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms”, 2013

Vu, “A splitting algorithm for dual monotone inclusions involving cocoercive operators”, 2013

Proximal Splitting Algorithms

PD3O algorithm

$$\left\lfloor \begin{array}{l} x^{k+1} = \mathrm{prox}_{\gamma R}\big(x^k - \gamma \nabla F(x^k) - \gamma K^* u^k\big) \\ u^{k+1} = \mathrm{prox}_{H^*/(\gamma\eta)}\big(u^k + \frac{1}{\gamma\eta} Kx(2x^{k+1} - x^k - \gamma \nabla F(x^{k+1}) + \gamma \nabla F(x^k))\big) \end{array} \right.$$

PDDY algorithm

$$\left\lfloor \begin{array}{l} u^{k+1} = \mathrm{prox}_{H^*/(\gamma\eta)}\big(u^k + \frac{1}{\gamma\eta} Kx^k\big) \\ x^{k+1} = \mathrm{prox}_{\gamma R}\big(x^k - \gamma \nabla F(x^k - \gamma K^*(u^{k+1} - u^k)) - \gamma K^*(2u^{k+1} - u^k)\big) \end{array} \right.$$

Yan, “A new primal-dual algorithm for minimizing the sum of three functions with a linear operator”, 2018

Salim, Condat, Mishchenko, Richtárik, “Dualize, split, randomize: Fast nonsmooth optimization algorithms”, arXiv:2004.02635

Proximal Splitting Algorithms

PD3O algorithm

$$\left\lfloor \begin{array}{l} x^{k+1} = \mathrm{prox}_{\gamma R}\big(x^k - \gamma \nabla F(x^k) - \gamma K^* u^k\big) \\ u^{k+1} = \mathrm{prox}_{H^*/(\gamma\eta)}\big(u^k + \frac{1}{\gamma\eta} Kx(2x^{k+1} - x^k - \gamma\nabla F(x^{k+1}) + \gamma\nabla F(x^k))\big) \end{array} \right.$$

PDDY algorithm

$$\left\lfloor \begin{array}{l} u^{k+1} = \mathrm{prox}_{H^*/(\gamma\eta)}\big(u^k + \frac{1}{\gamma\eta} Kx^k\big) \\ x^{k+1} = \mathrm{prox}_{\gamma R}\big(x^k - \gamma \nabla F(x^k - \gamma K^*(u^{k+1} - u^k)) - \gamma K^*(2u^{k+1} - u^k)\big) \end{array} \right.$$

Convergence if $\gamma \in (0, 2/L_F)$ and $\eta \geq \|K\|^2$

Proximal Splitting Algorithms

PD3O algorithm

$$\left\lfloor \begin{array}{l} x^{k+1} = \mathrm{prox}_{\gamma R}\big(x^k - \gamma \nabla F(x^k) - \gamma K^* u^k\big) \\ u^{k+1} = \mathrm{prox}_{H^*/(\gamma\eta)}\big(u^k + \frac{1}{\gamma\eta} Kx(2x^{k+1} - x^k - \gamma \nabla F(x^{k+1}) + \gamma \nabla F(x^k))\big) \end{array} \right.$$

PDDY algorithm

$$\left\lfloor \begin{array}{l} u^{k+1} = \mathrm{prox}_{H^*/(\gamma\eta)}\big(u^k + \frac{1}{\gamma\eta} Kx^k\big) \\ x^{k+1} = \mathrm{prox}_{\gamma R}\big(x^k - \gamma \nabla F(x^k - \gamma K^*(u^{k+1} - u^k)) - \gamma K^*(2u^{k+1} - u^k)\big) \end{array} \right.$$

Condat et al., “Proximal Splitting Algorithms: A Tour of Recent Advances, with New Twists,” arXiv:1912.00137

Goal

$$\underset{x \in \mathcal{X}}{\text{minimize}} \left\{ \Psi(x) := \frac{1}{M} \sum_{m=1}^{M} \Big(F_m(x) + H_m(K_m x) \Big) + R(x) \right\}$$

☞

- Distributed versions of these algorithms?
- Convergence rates?

Distributed PD3O and PDDY algs.

Distributed PD3O Algorithm

input: $(\gamma_k)_{k\in\mathbb{N}}$, $\eta \geq \|\widehat{K}\|^2$, $(\omega_m)_{m=1}^M$, $(q_m^0)_{m=1}^M \in \mathcal{X}^M$, $(u_m^0)_{m=1}^M \in \mathcal{U}^M$
initialize: $a_m^0 := q_m^0 - K_m^* u_m^0$, $m = 1 \dots M$
for $k = 0, 1, \dots$ **do**
  at master, **do**
    $x^{k+1} := \mathrm{prox}_{\gamma_k R}\big(\frac{\gamma_k}{M}\sum_{m=1}^M a_m^k\big)$
    broadcast x^{k+1} to all nodes
  at all nodes, for $m = 1, \dots, M$, **do**
    $q_m^{k+1} := \frac{M\omega_m}{\gamma_{k+1}} x^{k+1} - \nabla F_m(x^{k+1})$
    $u_m^{k+1} := \mathrm{prox}_{M\omega_m H_m^*/(\gamma_{k+1}\eta)}\Big(u_m^k + \frac{1}{\eta} K_m\big(\frac{M\omega_m}{\gamma_k} x^{k+1} + q_m^{k+1} - q_m^k\big)\Big)$
    $a_m^{k+1} := q_m^{k+1} - K_m^* u_m^{k+1}$
    transmit a_m^{k+1} to master
end for

Distributed PDDY Algorithm

input: $(\gamma_k)_{k\in\mathbb{N}}$, $\eta \geq \|\widehat{K}\|^2$, $(\omega_m)_{m=1}^M$, $x_R^0 \in \mathcal{X}$, $(u_m^0) \in \mathcal{U}^M$
initialize: $p_m^0 := K_m^* u_m^0$, $m = 1, \dots, M$
for $k = 0, 1, \dots$ **do**
  at all nodes, for $m = 1, \dots, M$, **do**
    $u_m^{k+1} := \mathrm{prox}_{M\omega_m H_m^*/(\gamma_k\eta)}\Big(u_m^k + \frac{M\omega_m}{\gamma_k\eta} K_m x_R^k\Big)$
    $p_m^{k+1} := K_m^* u_m^{k+1}$
    $x_m^{k+1} := x_R^k - \frac{\gamma_k}{M\omega_m}(p_m^{k+1} - p_m^k)$
    $a_m^k := M\omega_m x_m^{k+1} - \gamma_{k+1}\nabla F_m(x_m^{k+1}) - \gamma_{k+1} p_m^{k+1}$
    transmit a_m^k to master
  at master, **do**
    $x_R^{k+1} := \mathrm{prox}_{\gamma_{k+1} R}\big(\frac{1}{M}\sum_{m=1}^M a_m^k\big)$
    broadcast x_R^{k+1} to all nodes
end for

Convergence Rates

Theorem 1 – convergence rate of the Distributed PD3O Algorithm. Suppose that $\gamma_k \equiv \gamma \in (0, 2/L_{\widehat{F}})$ and $\eta \geq \|\widehat{K}\|^2$. Suppose that every H_m is continuous on a ball around $K_m x^\star$. Then the following hold:

$$\text{(i)} \quad \Psi(x^k) - \Psi(x^\star) = o(1/\sqrt{k}).$$

Define the weighted ergodic iterate $\bar{x}^k = \frac{2}{k(k+1)} \sum_{i=1}^{k} i x^i$, for every $k \geq 1$. Then

$$\text{(ii)} \quad \Psi(\bar{x}^k) - \Psi(x^\star) = O(1/k).$$

Furthermore, if every H_m is L_m-smooth for some $L_m > 0$,

$$\text{(iii)} \quad \min_{i=1,\ldots,k} \Psi(x^i) - \Psi(x^\star) = o(1/k).$$

Convergence Rates

Theorem 2 – accelerated Distributed PD3O Algorithm. Suppose that $\mu_{\widehat{F}} + \mu_R > 0$. Let $x^\star$ be the unique solution. Let $\kappa \in (0, 1)$ and $\gamma_0 \in (0, 2(1-\kappa)/L_{\widehat{F}})$. Set $\gamma_1 = \gamma_0$ and $\gamma_{k+1} = \big(-\gamma_k^2 \mu_{\widehat{F}}\kappa + \gamma_k \sqrt{(\gamma_k \mu_{\widehat{F}}\kappa)^2 + 1 + 2\gamma_k \mu_R}\big)/(1 + 2\gamma_k \mu_R)$, for every $k \geq 1$. Then there exists $\hat{c}_0 > 0$ such that, for every $k \geq 2$,

$$\|x^k - x^\star\|^2 \leq \frac{\gamma_k^2}{1 - \gamma_k \mu_{\widehat{F}}\kappa} \hat{c}_0 = O\big(1/k^2\big).$$

Theorem 3 – similar result for the accelerated Distributed PDDY Algorithm.

Convergence Rates

Theorem 4 – linear convergence of the Distributed PD3O Algorithm. Suppose that $\mu_{\widehat{F}} + \mu_R > 0$, that every H_m is L_m-smooth, for some $L_m > 0$, that $\gamma \in (0, 2/L_{\widehat{F}})$. Then there exists $\rho \in (0, 1]$ and $\hat{c}_0 > 0$ such that, for every $k \in \mathbb{N}$,

$$\|x^{k+1} - x^\star\|^2 \leq (1 - \rho)^k \hat{c}_0.$$

Theorem 5 – similar result for the Distributed PDDY Algorithm.