Dualize, Split, Randomize: Fast Nonsmooth Optimization Algorithms

Adil Salim Laurent Condat Konstantin Mishchenko Peter Richtárik

KAUST, Thuwal, Saudi Arabia.

Goal

$$\min_{x\in\mathbb{R}^d} F(x) + R(x) + H(Lx), \qquad (1)$$

where F, R, H convex functions, F smooth and R, H nonsmooth proximable.

We propose a new algorithm, the **Primal–Dual Davis–Yin (PDDY)**, to solve (1). PDDY is obtained as a carefully designed instance of the Davis–Yin Splitting between monotone operators.

We establish **convergence rates** for PDDY, when the algorithm is implemented with a **Variance Reduced (VR) stochastic gradient** of F.

In particular: **Linear rate for strongly convex minimization under linear constraints** (without projecting on the constraints space).

Primal–Dual optimality

Let $x^\star$ be a solution to Problem (1). Under a standard qualification condition,

$$0 \in \nabla F(x^\star) + \partial R(x^\star) + L^*\partial H(Lx^\star),$$

i.e., there exists $y^\star \in \partial H(Lx^\star)$ such that

$$0 = \nabla F(x^\star) + \partial R(x^\star) + L^* y^\star.$$

Since $Lx^\star \in \partial H^*(y^\star)$,

$$\begin{bmatrix}0\\0\end{bmatrix} \in \begin{bmatrix}\nabla F(x^\star) + \partial R(x^\star) + L^* y^\star\\ -Lx^\star \qquad\qquad + \partial H^*(y^\star)\end{bmatrix}.$$

Monotone operator

$$M(x,y) := \begin{bmatrix}\nabla F(x) + \partial R(x) + Ly\\ -Lx \qquad\qquad + \partial H^*(y)\end{bmatrix}.$$

Then, $0 \in M(x^\star, y^\star)$. Moreover, M is a monotone operator: $\langle M(x,y) - M(x',y'), (x,y)-(x',y')\rangle \geq 0$. Indeed, M is the sum of a skew symmetric operator and the subdifferential of $F(x) + R(x) + H^*(y)$.

Davis–Yin Splitting

Solving Problem (1) is equivalent to solving the inclusion $0 \in M(x^\star, y^\star)$. One idea could be to decompose

$$M(x,y) = \underbrace{\begin{bmatrix}\partial R(x)\\ 0\end{bmatrix}}_{:=A(x,y)} + \underbrace{\begin{bmatrix}L^* y\\ -Lx + \partial H^*(y)\end{bmatrix}}_{:=B(x,y)} + \underbrace{\begin{bmatrix}\nabla F(x)\\ 0\end{bmatrix}}_{:=C(x,y)}, \qquad (2)$$

and apply the **Davis–Yin Splitting (DYS)** algorithm [2] which can solve monotone inclusions of the form $0 \in (A+B+C)(x^\star, y^\star)$, see below. DYS generalizes the standard proximal gradient algorithm and relies on the computation of the resolvent of B, denoted $J_B(x,y) = (I+B)^{-1}(x,y)$.

In other words, $(x', y') = J_B(x,y)$ is equivalent to $(x',y') \in (x,y) - B(x',y')$, which is intractable in general. Hence one cannot apply DYS directly.

Primal–Dual Davis–Yin

The idea is preconditioning: let P a positive definite symmetric matrix. Then $0 \in M(x^\star, y^\star)$ is equivalent to $0 \in P^{-1}M(x^\star, y^\star)$. Besides, $P^{-1}M = P^{-1}A + P^{-1}B + P^{-1}C$. Finally $P^{-1}A, P^{-1}B, P^{-1}C$ are monotone operators under the inner product induced by P. DYS applied to the inclusion $0 \in (P^{-1}A + P^{-1}B + P^{-1}C)(x^\star, y^\star)$ relies on the computation of the resolvent of $P^{-1}B$.

In other words, $(x', y') = J_{P^{-1}B}(x,y)$ is equivalent to $P(x',y') \in P(x,y) - B(x',y')$, which only relies on the proximity operator of H denoted

$$\mathrm{prox}_H(x) = \arg\min_{y\in\mathbb{R}^d} H(y) + \frac{1}{2}\|x-y\|^2,$$

if [1]

$$P := \begin{bmatrix} I & 0\\ 0 & \frac{\gamma}{\tau}I - \gamma^2 LL^*\end{bmatrix}.$$

The resulting algorithm is the **PDDY algorithm**. It inherits *the convergence properties of DYS.*

Davis–Yin Algorithm $\mathrm{DYS}(A,B,C)$ [2]

Input: $v^0 \in \mathcal{Z}$, $\gamma > 0$
for $k = 0, 1, 2, \ldots$ **do**
  $z^k = J_{\gamma B}(v^k)$
  $u^{k+1} = J_{\gamma A}(2z^k - v^k - \gamma C(z^k))$
  $v^{k+1} = v^k + u^{k+1} - z^k$
end for

Stochastic PDDY algorithm (proposed)
(deterministic version: $g^{k+1} = \nabla F(x^k)$)

Input: $p^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$, $\gamma > 0$, $\tau > 0$
for $k = 0, 1, 2, \ldots$ **do**
  $y^{k+1} = \mathrm{prox}_{\tau H^*}\left(y^k + \tau L(p^k - \gamma L^* y^k)\right)$
  $x^k = p^k - \gamma L^* y^{k+1}$
  $s^{k+1} = \mathrm{prox}_{\gamma R}\left(2x^k - p^k - \gamma g^{k+1}\right)$
  $p^{k+1} = p^k + s^{k+1} - x^k$
end for

Other primal–dual algorithms like Condat-Vũ I, Condat-Vũ II and PD3O can be derived from DYS as well.

Contact

adil.salim,laurent.condat,konstantin.mishchenko,peter.richtarik@kaust.edu.sa

VR stochastic gradient

Several VR stochastic gradient estimators used in the literature satisfy the following [3].

There exist $\alpha, \beta, \delta \geq 0$, $\rho \in (0,1]$ and a stochastic process denoted by $(\sigma_k)_k$, s.t.,

$$\mathbb{E}_k(g^{k+1}) = \nabla F(x^k)$$
$$\mathbb{E}_k(\|g^{k+1} - \nabla F(x^\star)\|^2) \leq 2\alpha D_F(x^k, x^\star) + \beta\sigma_k^2$$
$$\mathbb{E}_k(\sigma_{k+1}^2) \leq (1-\rho)\sigma_k^2 + 2\delta D_F(x^k, x^\star),$$

where D_F Bergman divergence of F.

Convergence rates

Assume γ small enough and $\gamma\tau\|L\|^2 < 1$.

Then, $\mathbb{E}D_F(\bar{x}^k, x^\star) + \mathbb{E}D_{H^*}(\bar{y}^{k+1}, y^\star) + \mathbb{E}D_R(\bar{s}^{k+1}, s^\star) = \mathcal{O}(1/k)$.

If R strongly convex and H smooth, then $\mathbb{E}\|x^k - x^\star\|^2 + \mathbb{E}\|y^k - y^\star\|^2$ converges linearly.

If F strongly convex, $R \equiv 0$ and $H(x) = \infty$ except at $H(b) = 0$, $\mathbb{E}\|x^k - x^\star\|^2 + \mathbb{E}\|y^k - y^\star\|^2$ **converges linearly to zero** ($x^\star$ is the solution to $\min F$ s.t. $Lx = b$). Complexity: $\mathcal{O}(\kappa + \chi\log(1/\varepsilon))$, where κ (resp. χ) condition number of F (resp. L^*L).

References

[1] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi. Proximal splitting algorithms: A tour of recent advances, with new twists. preprint arXiv:1912.00137, 2019.

[2] D. Davis and W. Yin. A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25(4):829–858, 2017.

[3] E. Gorbunov, F. Hanzely, and P. Richtárik. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pages 680–690, 2020.