Dualize, Split, Randomize: Fast Nonsmooth Optimization Algorithms

Adil Salim  Laurent Condat  Konstantin Mishchenko  Peter Richtárik

KAUST, Thuwal, Saudi Arabia.

### Goal

\[
\min_{x \in \mathbb{R}^k} F(x) + R(x) + H(Lx),
\]

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

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

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

\[
0 = \nabla F(x^*) + R(x^*) + L^* y^*.
\]

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

\[
\begin{bmatrix}
\nabla F(x^*) + R(x^*) + L^* y^* \\
-Lx^* \\
+ \partial H(y^*)
\end{bmatrix} = 0.
\]

### Monotone operator

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

Then, \( 0 \in M(x^*, y^*) \). 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 \( 0 \in M(x^*, y^*) \). One idea could be to decompose

\[
M(x, y) = \begin{bmatrix}
\partial R(x) \\
0 \\
-\xi(x, y)
\end{bmatrix} + \begin{bmatrix}
L y \\
-Lx \\
-\partial H'(y)
\end{bmatrix} + \begin{bmatrix}
\nabla F(x) \\
0 \\
0
\end{bmatrix},
\]

and apply the Davis–Yin Splitting (DYS) algorithm (2) which can solve monotone inclusions of the form \( 0 \in (A + B + C)(x^*, y^*) \), 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 \( 0 \in M(x^*, y^*) \) is equivalent to \( 0 \in P^{-1}M(x^*, y^*) \). Besides, \( P^{-1}A = P^{-1}A + P^{-1}B + P^{-1}C \) are monotone operators under the inner product induced by \( P \). DYS applied to the inclusion \( 0 \in (P^{-1}A + P^{-1}B + P^{-1}C)(x', y') \) relies on the resolvent of \( P^{-1}B \).

In other words, \( (x^*, y^*) = J_B(x, y) \) is equivalent to \( P(x, y) \in P(x, y) - B(x', y') \), which only relies on the proximity of \( H \) denoted

\[
\text{prox}_H(x) = \text{arg min}_{y \in \mathbb{R}^d} H(y) + \frac{1}{2} \|x - y\|^2.
\]

if \[1\] \( P := \begin{bmatrix} I & 0 \\ 0 & \frac{1}{2}I - \gamma^2 L^* L \end{bmatrix}. \]

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

### Davis–Yin Algorithm

\( \text{DYS}(A, B, C) \) \[2\]

**Stochastic PDDY algorithm (proposed)**

\( \text{stochastic version: } \phi^{k+1} = \nabla F(x^k) \)

1. **Input:** \( y^m \in \mathbb{Y}, \gamma > 0 \)
2. for \( k = 0, 1, 2, \ldots \) do
3. \( x^k = J_{\gamma y^m}(x^k) \)
4. \( y^{k+1} = \text{prox}_{\gamma L^2}(y^k + \tau L(p^k - \gamma L^2 y^k)) \)
5. \( x^{k+1} = x^k + y^{k+1} - \gamma^2 L^* L x^k \)
6. end for

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

### VR stochastic gradient

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

There exist \( a, b, \delta \geq 0 \), \( p \in [0, 1] \) and a stochastic process denoted by \( (\xi_k) \), s.t.,

\[
\mathbb{E}_{\xi}(\phi^{k+1}) = \nabla F(x^k),
\]

\[
\mathbb{E}_{\xi}(\|y^{k+1} - \nabla F(x^k)\|^2) \leq 2aD_F(x^k, x^*) + \beta\sigma^2_k,
\]

\[
\mathbb{E}_{\xi}(\|\phi^{k+1}\|^2) \leq (1 - \rho)\sigma^2_k + 2\Delta D_F(x^k, x^*),
\]

where \( D_F \) Bergman divergence of \( F \).

### Convergence rates

Assume \( \gamma \) small enough and \( \gamma \tau \|L\|^2 < 1 \). Then,

\[
\mathbb{E}_{\xi}(\phi^{k+1}) + \mathbb{E}_{\xi}(\phi^{k+1}, y^k) + \mathbb{E}_{\xi}(s^{k+1}, s^*) = O(1/k).
\]

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

### References


### Contact

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