DECOMPOSITION IN OPTIMIZATION: ALGORITHMIC ADVANCES BEYOND ADMM

Terry Rockafellar
University of Washington, Seattle

26th MOM and Workshop on
Large-Scale Optimization and Applications
University of Waterloo, Ontario, Canada
November 8–9, 2024

Overview

The need for decomposition:

- Problems in imaging and machine learning can be very large, but structure suggests a breakdown into much smaller subproblems
- A solution might be found by repeatedly solving updated versions of the subproblems in parallel and coordinating the results

Status of current methodology for this:

- Algorithms utilizing augmented Lagrangians are popular
- They typically suffer from the fact that augmentation may interfere with the separability that supports decomposition

New outlook on improvements:

- The progressive decoupling algorithm avoids that trouble
- It exhibits linear convergence "generically" while also providing more flexibility in the articulation of proximal parameters

Structured Problems in Convex Optimization

Commonly chosen format: minimize $f(x) + g(Ax)$

for linear $A : \boldsymbol{R}^n \to \boldsymbol{R}^m$ and f, g, closed proper convex

feasible solutions: $x \in \operatorname{dom} f$ with $Ax \in \operatorname{dom} g$

Possibilities for finer structure:

$$f(x) = f_1(x_1) + \cdots + f_q(x_q), \quad x = (x_1, \ldots, x_q) \text{ with } x_j \in \boldsymbol{R}^{n_j}$$

$$g(z) = g_1(z_i) + \cdots + g_p(z_p), \quad z = (z_1, \ldots, z_p) \text{ with } z_i \in \boldsymbol{R}^{m_i}$$

Problem examples with such block separability:

minimize $\sum_{j=1}^{q} f_j(x_j) + g(\sum_{j=1}^{q} A_j x_j)$ maybe with $g = \delta_K$

minimize $f(x) + \sum_{i=1}^{p} g_i(A_i x)$ maybe with $f = \delta_X$, $A_i = I$,

or maybe $f = ||\cdot||_1$ and A_i selects several coordinates of x

$\longrightarrow$ **how to take advantage of this in computations?**

Augmented Lagrangian Framework

(P) primal problem: minimize $f(x) + g(Ax)$ for $x \in R^n$
(D) dual problem: maximize $-g^*(y) - f^*(-A^*y)$ for $y \in R^m$

Lagrangian: $L_0(x, y) = f(x) + y{\cdot}Ax - g^*(y)$
Augmented Lagrangian: with $r > 0$ and $g_r(z) = g(z) + \frac{r}{2}|z|^2$

$$L_r(x, y) = f(x) + y{\cdot}Ax + \frac{r}{2}|Ax|^2 - g_r^*(y + rAx)$$

Characterization of optimality for any $r \geq 0$ (under a mild c.q.)

$$\left.\begin{array}{l}\bar{x} \text{ solves (P)} \\ \bar{y} \text{ solves (D)}\end{array}\right\} \iff (\bar{x}, \bar{y}) \text{ is a saddle point of } L_r(x, y)$$

Common reformulation with simpler-looking Lagrangians:

$(\overline{P})$ minimize $f(x) + g(z)$ subject to $Ax - z = 0$

$$\overline{L}_r(x, z; y) = f(x) + g(z) + y{\cdot}[Ax - z] + \frac{r}{2}|Ax - z|^2$$

augmentation ruins block-separability in primal arguments

The Original Methods of Multipliers — ALM

well known procedures based on the proximal point algorithm, PPA

Basic ALM : apply the PPA to the dual problem

$$x^{k+1} \in \operatorname{argmin}_x L_r(x, y^k), \qquad y^{k+1} = y^k + r[z^{k+1} - Ax^{k+1}],$$
$$\text{where} \quad z^{k+1} = \operatorname{argmin}_z \left\{ g(z) - y^k \cdot z + \tfrac{r}{2}|z - Ax^{k+1}|^2 \right\}$$

Proximal ALM: apply the PPA to the saddle point problem

$$\text{same except} \quad x^{k+1} = \operatorname{argmin}_x \left\{ L_r(x, y^k) + \tfrac{1}{2r}|x - x^k|^2 \right\}$$

Characteristics: assuming $\exists$ optimal $\bar{x}$ and $\bar{y}$

- basic ALM gets convergence of $\{y^k\}$ to some $\bar{y}$, but might only get cluster points of $\{x^k\}$ as $\bar{x}$
- proximal ALM gets convergence of $\{(x^k, y^k)\}$ to some $(\bar{x}, \bar{y})$

Alternating Direction Method of Multipliers — ADMM

Basic ADMM: a more complicated PPA application

$$x^{k+1} \in \operatorname{argmin}_x \left\{ f(x) + y^k \cdot Ax + \frac{r}{2}|Ax - z^k|^2 \right\},$$

then **exactly as in ALM,** $y^{k+1} = y^k + r[z^{k+1} - Ax^{k+1}],$

where $z^{k+1} = \operatorname{argmin}_z \left\{ g(z) - y^k \cdot z + \frac{r}{2}|z - Ax^{k+1}|^2 \right\}$

Proximal ADMM: same except

$$x^{k+1} = \operatorname{argmin}_x \left\{ f(x) + y^k \cdot Ax + \frac{r}{2}|Ax - z^k|^2 + \frac{1}{2r}|x - x^k|^2 \right\}$$

Relationships with ALM: the augmented Lagrangian expression

$$L_r(x, y^k) = f(x) + y^k \cdot Ax + \frac{r}{2}|Ax|^2 - g_r^*(y^k + rAx)$$

is simplified in ADMM by an affine substitute for the final term

convergence characterics are similar

$\{x^k\}$, $\{z^k\}$, $\{y^k\}$ converge to some optimal $\bar{x}$, $\bar{z} = A\bar{x}$, and $\bar{y}$

Shortcomings of ALM and ADMM for Decomposition

Case of primal block-separability: $x = (x_1, \ldots, x_q)$

$$f(x) = f_1(x_1) + \cdots + f_q(x_q), \qquad Ax = A_1x_1 + \cdots + A_qx_q$$

Expressions to be minimized iteratively: **no separability!**

ALM $\sum_{j=1}^q [f_j(x_j) + y^k \cdot A_jx_j] + \frac{r}{2}|\sum_{j=1}^q A_jx_j|^2 - g_r\left(y^k + r\sum_{j=1}^q A_jx_j\right)$

ADMM $\sum_{j=1}^q [f_j(x_j) + y^k \cdot A_jx_j] + \frac{r}{2}|\sum_{j=1}^q A_jx_j - z^k|^2$

Adopted remedy: *only rough minimization*, Gauss-Seidel mode

$$\forall j: \quad x_j^{k+1} \in \operatorname{argmin}_{x_j} \left\{ f_j(x_j) + y^k \cdot A_jx_j + \frac{r}{2}|A_jx_j + a_j^k - z^k|^2 \right\}$$

where a_j^k = the sum of the terms $A_{j'}x_{j'}^k$ for $j' \neq j$

$\Longrightarrow$ no longer a PPA application, problematical convergence

Decomposition Achieved Via "Progressive Decoupling"

actually another indirect application of the PPA

New algorithm: generating $\{x_j\}^k$, $\{z^k\}$, $\{y^k\}$, and now $\{w_j^k\}$

$$\forall j: \; x_j^{k+1} = \operatorname*{argmin}_{x_j} \left\{ f_j(x_j) + y^k \cdot A_j x_j + \frac{r_k}{2} |A_j x_j - w_j^k|^2 + \frac{s_k}{2} |x_j - x_j^k|^2 \right\}$$

$$z^{k+1} = \operatorname*{argmin}_{z} \left\{ g(z) - y^k \cdot z + \frac{r_k}{2} |z - \textstyle\sum_{j=1}^q w_j^k|^2 \right\}$$

then update

$$\left.\begin{aligned} y^{k+1} &= y^k + \frac{r_k}{q+1} \Delta^{k+1} \\ w_j^{k+1} &= A_j x_j^{k+1} - \frac{1}{q+1} \Delta^{k+1} \end{aligned}\right\} \text{ for } \Delta^{k+1} = \textstyle\sum_{j=1}^q A_j x_j^{k+1} - z^{k+1}$$

"residual" (0 in optimality)

Features to notice: in comparison to proximal ADMM

- w_j^k substitutes for z^k in the decomposed x-minimizations
- $\sum_{j=1}^q w_j^k$ substitutes for z^k in z-minimization
- r is now r_k, and the proximal term has parameter s_k

Improvements in the New Decomposition Approach

"rough" minimization using Gauss-Seidel is avoided
without asking for calculations any harder than ADMM

Global convergence characteristics:

- $x_j^k \to \bar{x}_j, \quad y^k \to \bar{y}, \quad z^k \to \bar{z}, \quad w_j^k \to A_j \bar{x}_j$
  assuming that $r_k \to r_\infty \in (0, \infty)$, $s_k \to s_\infty \in (0, \infty)$
- moreover **a linear rate is "generic"** in a certain sense
  (in contrast, linear convergence for ADMM is "very special")
- separate proximal parameters r_k and s_k allow more influence
  but there are trade-offs, superlinear convergence is out of reach
- approximate minimization allowed, with stopping criteria

Decoupling of Linkages in Optimization More Generally

convex case for now, but nonconvex case later

Linkage problem (L): minimize $\varphi(u)$ over $u \in S \subset \mathbf{R}^N$

φ is closed proper convex, S is a subspace giving linkages

Optimality condition: $\bar{u} \in S, \quad \bar{v} \in S^\perp, \quad \bar{v} \in \partial\varphi(\bar{u})$

$$\longrightarrow \quad \bar{u} \in \operatorname{argmin}_u \{\varphi(u) - \bar{v}\cdot u\} \quad \text{(under convexity)}$$

$\bar{v}$ thus **decouples** by neutralizing the constraint $u \in S$

Example: the instance behind the proposed alternative to ADMM

target: minimizing $\sum_{j=1}^q f_j(x_j) + g(\sum_{j=1}^q A_j x_j)$

$$\varphi(x_1,\ldots,x_q,z,w_1,\ldots,w_q) = \sum_{j=1}^q [f_j(x_j) + \delta_0(A_j x_j - w_j)] + g(z)$$

$$S = \{(x_1,\ldots,x_q,z,w_1,\ldots,w_q) \mid \sum_{j=1}^q w_j = z\}$$

$$S^\perp = \{(0,\ldots,0,y,-y,\ldots,-y)\} \qquad \delta_0 = \text{indicator of } \{0\}$$

Solution Methodology with Executable Projections P_S, $P_{S^\perp}$

$\longrightarrow$ aiming for $\bar{u} \in S$, $\bar{v} \in S^\perp$, with $\bar{u} \in \operatorname{argmin}\{\varphi(u) - \bar{v}{\cdot}u\}$

Linkage-compatible norms: $||u||_M = \sqrt{u{\cdot}Mu}$

M symmetric, pos.-definite, with $v{\cdot}Mu = 0$ when $u \in S$, $v \in S^\perp$

Progressive decoupling algorithm, PDA (with recent improvements)

Generate $\{u^k\} \subset S$, $\{v^k\} \subset S^\perp$, from chosen norms $||\cdot||_{M_k}$ by

(1) $\hat{u}^{k+1} = \operatorname{argmin}_u \{ \varphi(u) - v^k{\cdot}u + ||u - u^k||^2_{M_k} \}$,

(2) $u^{k+1} = P_S(\hat{u}^{k+1})$, $\quad v^{k+1} = v^k - M_k P_{S^\perp}(\hat{u}^{k+1})$

this applies PPA to a partial inverse of the $\partial\varphi$, like Spingarn

Convergence: under a mild assumption about choice of $\{M_k\}$

- $(u^k, v^k) \to (\bar{u}, \bar{v})$, and a **linear** rate is **generic** relative to the (a, b)-parametric embedding with $\varphi_{a,b}(u) = \varphi(u + a) - b{\cdot}u$
- inexact minimization allowed in (1) under a stopping criterion

Progressive Decoupling in Nonconvex Optimization

Example: for minimizing $\sum_{j=1}^{q} f_j(x_j) + g\Big(\sum_{j=1}^{q} F_j(x_j)\Big)$

Linkage problem (L): minimize $\varphi(u)$ over $u \in S \subset R^N$

φ closed proper, maybe nonconvex, S still a subspace

First-order optimality: $\bar{u} \in S,\ \bar{v} \in S^{\perp},\ \bar{v} \in \partial\varphi(\bar{u})$

Second-order: **strong variational sufficiency**

$\exists$ elicitation level $e \geq 0$ such that the function $\varphi_e = \varphi + e\,\mathrm{dist}_S^2$
is strongly convex **variationally** at $\bar{u}$ for the subgradient $\bar{v}$

this corresponds in NLP to classical strong sufficiency

Progressive decoupling algorithm, localized

- same, but initiated close enough to locally optimal $\bar{u}$, $\bar{v}$
- then all the same convergence properties will be obtained

Ongoing research challenge: how first to get close enough?

Background to Progressive Decoupling

- 1970s: convex ALM emerged, but without decomposition
- early 1980s: Spingarn (my PhD from 1976) got a "splitting method" from taking partial inverses of monotone mappings
- late 1980s: Wets and I built around this, for stochastic programming, the Progressive Hedging Algorithm, PHA (it caught on numerically, but other Spingarn splitting didn't)
- early 1990s: Eckstein and Bertsekas proposed ADMM for decomposition inspired by Gabay, Mercier, Lions, Glowinski
- starting 2016: I realized Spingarn's scheme can lead to lots more, and in localization even to nonconvex decomposition
- early 2020s: my efforts to understand "variational sufficiency" and to refine the Proximal Point Algorithm for local usage
- recently: my work on stopping criteria and prox-term flexibility through an extension of the proximal method of multipliers

Some References

[1] (2019) "Progressive decoupling of linkages in optimization and variational inqualities with elicitable convexity or monotonicity," *Set-Valued and Variational Analysis* 27 (2019), 863–893

[2] (2022) "Augmented Lagrangians and hidden convexity in sufficient conditions for local optimality," *Mathematical Programming* 198 (2023), 159–194 (published online June 2022)

[3] (2022) "Generic linear convergence of the proximal point algorithm even in variable metric implementation," *Computational Optimization and Applications* (accepted 2023)

[4] (2023) "Generalizations of the proximal method of multipliers in convex optimization," *Computational Optimization and Applications* (accepted 2023)

downloads: sites.washington.edu/∼rtr/mypage.html