A nonlinear ADMM for nonlinear composite problems

Bằng Công Vũ[1] and Dimitri Papadimitriou[1]

3NLab, Huawei Belgium Research Center (BeRC), Leuven, Belgium

vu.cong.bang@huawei.com, dpapadimitriou@3nlab.org

1 Introduction

This paper investigates a structured nonconvex optimization problem including a nonlinear composition. This general framework covers a wide class of problems in optimization such as semidefinite programming [6], distributed optimization over networks, robust principal component analysis [22], image processing [19], and machine learning [13].

Problem 1.1 Let $(\mathcal{X}, \langle\cdot \mid \cdot\rangle)$ and $(\mathcal{Y}, \langle\cdot \mid \cdot\rangle)$ be real Hilbert spaces, and $c\colon \mathcal{X} \to \mathcal{Y}$ be a **nonlinear**, differentiable mapping. Let $b \in \mathcal{Y}$. Let $h\colon \mathcal{X} \to \mathbb{R}$ be a differentiable nonconvex function, with L_h Lipschitz continuous gradient. Let $f\colon \mathcal{X} \to]-\infty, +\infty]$ and $g\colon \mathcal{Y} \to]-\infty, +\infty]$ be proper and lower semicontinuous convex functions not necessarily differentiable, and $g^\star$ denote its conjugate function. The problem is to

$$\min_{u\in\mathcal{X}} \varphi(u) = f(u) + h(u) + g(c(u) - b), \tag{1}$$

$$\text{or equivalently } \min_{u\in\mathcal{X}} \max_{v\in\mathcal{Y}} \psi(u, v) = f(u) + h(u) + \langle c(u) - b \mid v\rangle - g^\star(v). \tag{2}$$

The main challenges come from the nonlinearity of c and the nonconvexity of h. A typical approach to solve nonconvex as well as convex optimization problems is to investigate the first order optimal conditions. For Problem 1.1, the ϵ-first order optimality conditions for $(\overline{u}, \overline{v})$ solving the min-max problem is characterized by

$$c(\overline{u}) - b \in \partial(g^*)(\overline{v}) + B(0;\epsilon) \text{ and } -[\nabla c(\overline{u})]^*\overline{v} \in (\nabla h + \partial f)(\overline{u}) + B(0;\epsilon), \tag{3}$$

where $B(0;\epsilon)$ is the closed ball with radius ϵ, and $\partial f(w)$ denotes the sub-differential of f at w. When $\epsilon = 0$, (3) is known as the Karush-Kuhn-Tucker (KKT) conditions. Throughout this paper, we assume that such ϵ-saddle points exist.

Assumption 1.2 Let Y be subsets in $\mathcal{Y}$ and Z be a subset in $\mathcal{X}$. Assume that

1. Assume that $\mathrm{Im}(\partial g)$ and $(\mathrm{Im}(\mathrm{prox}_{\beta g^*}))_{\beta\in\mathbb{R}_+}$ are bounded by κ.
2. ∇c is μ_c-Lipschitz, and the uniform regularity of c on Z with constant $\zeta \in]0, +\infty[$ with respect to Y ; more precisely, $(\exists \zeta \in]0, +\infty[)(\forall (u, y) \in Z \times Y)\ \|\nabla c(u)^* y\| \geq \zeta \|y\|$.

2 Related Work and Contribution

When c is linear and h is convex, this primal-dual problem has been widely investigated in the literature ; see [7, 14] for detailed surveys on the subject. ADMM is the classic method proposed for solving Problem 1.1 for the case where c is a bounded linear operator and $h \equiv 0$ [9]. This method is an application of the Douglas-Rachford method to the dual problem [8]. The main drawback of standard ADMM is that it may still exhibit slow convergence since it requires exact solutions of the subproblems at each iteration. To overcome this issue, the first strategy, proposed in [18], refined in [2], is known as the alternating direction proximal method

of multipliers. The second strategy consists of using a linearization technique [15]. Furthermore, when $f \equiv 0$, some numerical methods have been recently proposed for solving Problem 1.1 even when g is nonconvex [16, 1]. When c is nonlinear and $h \equiv 0$, a modification of the well-known Chambolle-Pock's algorithm has been proposed in [19]. An alternative approach based on preconditioned ADMM can be found in [3], where the convergence of the iterates is proved under strong assumptions not fulfilled in our setting. However, it is unclear how these works can solve Problem 1.1 with $h \not\equiv 0$. Several algorithms based on the Augmented Lagrangian methods for constrained nonconvex problems have been recently investigated in [11, 20, 5, 17].
Contribution : We propose a primal-dual splitting method to solve Problem 1.1. We characterize the convergence guarantee in term of gradient mapping and feasibility. We then show that, under mild conditions, the gradient mapping and feasibility of the generated iteration converge to 0.

3 Algorithm

We use the smoothing technique of [21]. Let $\beta \in]0, +\infty[$ and $(u, \dot{y}) \in \mathcal{X} \times \mathcal{Y}$, let $F_\beta : u \mapsto h(u) + g_\beta(u, \dot{y})$ with $g_\beta(u, \dot{y}) = \sup_{y \in \mathcal{Y}} \Big(\langle c(u) - b \mid \dot{y} \rangle - g^*(y) - \frac{\beta}{2} \|y - \dot{y}\|^2 \Big)$. Further, we assume that $\nabla g_\beta(\cdot, \dot{y})$ is L_β-Lipschitz continuous. Then, we apply only one proximal gradient step on the resulting problem, reduce the parameter β and update the Lagrangian multiplier following the usual ADMM. The resulting iterative algorithm can be described as follows, where $\mathrm{prox}_{\beta g} \colon y \mapsto \mathrm{argmin}_{v \in \mathcal{Y}} (g(v) + \frac{1}{2\beta}\|v - y\|^2)$.

Let $\alpha \in]1/2, 1[$, $\beta_0 = 1$, $\sigma_0 > 0$, $\theta \gg 1$, $\chi > \sqrt{48}$, $\zeta \in]0, +\infty[$, $N \in \mathbb{N}_0$. Let $u_0 \in \mathrm{dom}(f)$, $\dot{y}_0 \in \mathcal{Y}$.

Step 1 : Primal step

$$\begin{cases} w_k = \mathrm{prox}_{\beta_k g}(c(u_k) - b + \beta_k \dot{y}_k) \\ y_k^\star = \dot{y}_k + 1/\beta_k (c(u_k) - b - w_k) \\ \gamma_k \in]0, 1/(L_{\beta_k} + L_h)[\\ u_{k+1} = \mathrm{prox}_{\gamma_k f}(u_k - \gamma_k (\nabla h(u_k) + \nabla c(u_k)^\star y_k^\star)) \end{cases}$$

Step 2 : Dual step

$$\begin{cases} a_{k+1} = c(u_{k+1}) - b \\ z_{k+1} = \mathrm{prox}_{\beta_k g}(a_{k+1} + \beta_k \dot{y}_k) \\ \dot{y}_{k+1} = \dot{y}_k + (1/\sigma_k)(a_{k+1} - z_{k+1}) \\ \beta_{k+1} = 1/(k+2)^\alpha \end{cases}$$

Step 3 : Find $\tau \leq \min\Big\{ \dfrac{\zeta \beta_{k+1}}{\chi \|\dot{y}_k\| \mu_c \sqrt{\gamma_k}}, \dfrac{\zeta \beta_{k+1}}{\chi \theta} \Big\}$ such that

$$\begin{cases} \|\nabla c(u_k)^\star y_k^\star - \nabla c(u_{k+1})^\star (\dot{y}_{k+1} + \dfrac{1}{\beta_{k+1}}(a_{k+1} - w_{k+1})\| \leq \dfrac{\theta}{\sqrt{\gamma_k}} \|u_{k+1} - u_k\| + \dfrac{\theta \beta_{k+1}}{(k+1)^\alpha \tau} \\ \dfrac{1}{\beta_{k+1}} \|\nabla c(u_{k+1})^\star\| \, \|a_{k+1} - w_{k+1}\| \leq \dfrac{\zeta \beta_{k+1}}{\tau \chi \sqrt{\gamma_k}} \|u_{k+1} - u_k\| \end{cases} \tag{4}$$

If $k > N$, Then STOP
Otherwise $\sigma_{k+1} = \tau$, $k \leftarrow k + 1$

4 Convergence Results

Given $\gamma > 0$, the gradient mapping is defined by $G_{\beta,\gamma}(\cdot, \dot{y}) \colon u \mapsto \gamma^{-1}(u - u^+)$ where $u^+ = \mathrm{prox}_{\gamma f}(u - \gamma \nabla F_\beta(u, \dot{y}))$. In nonconvex optimization, the relations between the gradient mapping and stationary points are well-understood [10, 12, 4]. The following result develops this idea to the context of saddle point problems and provides us a tool to find a saddle point.

Theorem 4.1 *Suppose that $\|a_{k+1} - z_{k+1}\| \leq \epsilon$ and $\gamma_k(L_{\beta_k} + L_h) \leq 1$. Then $(u_{k+1}, \dot{y}_{k+1})$ is a ϵ-saddle point.*

Theorem 4.2 [Convergence results of the proposed algorithm] *Suppose that φ is bounded below and Assumption 1.2 is satisfied with $(u_k)_{k \in \mathbb{N}} \subset Z$, $(\dot{y}_{k+1} - \dot{y}_k)_{k \in \mathbb{N}} \subset Y$. Suppose that $\sigma_k \leq (2\chi/3)\sigma_{k+1}$. Then $(\varphi(x_k))_{k \in \mathbb{N}}$ is a convergent sequence, and every strong cluster point $(u_k)_{k \in \mathbb{N}}$ is the stationary point. Moreover, $\sum_{k \in \mathbb{N}} (\frac{1}{2\beta_k}\|a_k - w_k\|^2 + \frac{1}{\gamma_{k1}}\|u_{k+1} - u_k\|^2) < +\infty$. Therefore, under the condition, $(\gamma_k)_{k \in \mathbb{N}}$ is monotone decreasing, for every $N \in \mathbb{N}$,*

$$\min_{1 \leq k \leq N} \|(u_{k+1} - u_k)/\gamma_k\|^2 = \mathcal{O}(1/(N\gamma_N)) \text{ and } \min_{1 \leq k \leq N} \|(a_k - w_k)/\sqrt{\beta_k}\|^2 = \mathcal{O}(1/N). \tag{5}$$

Références

[1] Marco Artina, Massimo Fornasier, and Francesco Solombrino. Linearly constrained nonsmooth and nonconvex minimization. *SIAM Journal on Optimization*, 23(3) :1904–1937, 2013.

[2] Sebastian Banert, Radu Ioan Boţ, and Ernö Robert Csetnek. Fixing and extending some recent results on the admm algorithm. *Numerical Algorithms*, pages 1–23, 2016.

[3] Martin Benning, Florian Knoll, Carola-Bibiane Schönlieb, and Tuomo Valkonen. Preconditioned admm with nonlinear operator constraint. In Lorena Bociu, Jean-Antoine Désidéri, and Abderrahmane Habbal, editors, *System Modeling and Optimization*, pages 117–126, Cham, 2016. Springer International Publishing.

[4] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization of nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2) :459–494, 2014.

[5] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Nonconvex lagrangian-based optimization : monitoring schemes and global convergence. *Mathematics of Operations Research*, 2018.

[6] Samuel Burer and Renato D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2) :329–357, Feb 2003.

[7] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25 :161–319, 2016.

[8] D. Gabay. Applications of the method of multipliers to variational inequalities. *in : M. Fortin and R. Glowinski (eds.), Augmented Lagrangian Methods : Applications to the Solution of Boundary-Value Problems, North-Holland, Amsterdam*, 1983.

[9] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Applications*, 2(1) :17–40, 1976.

[10] G. Ghadimi, S. Lan and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program., Ser. A*, 155 :267–305, 2016.

[11] Geovani Nunes Grapiglia and Ya-xiang Yuan. On the complexity of an augmented lagrangian method for nonconvex optimization. *IMA Journal of Numerical Analysis*, 2019.

[12] Warren Hare and Claudia Sagastizábal. Computing proximal points of nonconvex functions. *Mathematical Programming*, 116(1) :221–258, 2009.

[13] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4) :142–336, 2017.

[14] Nikos Komodakis and Jean-Christophe Pesquet. Playing with duality : An overview of recent primal ? dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine*, 32(6) :31–54, 2015.

[15] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in neural information processing systems*, pages 612–620, 2011.

[16] Thomas Mollenhoff, Evgeny Strekalovskiy, Michael Moeller, and Daniel Cremers. The primal-dual hybrid gradient method for semiconvex splittings. *SIAM Journal on Imaging Sciences*, 8(2) :827–857, 2015.

[17] Mehmet Fatih Sahin, Armin Eftekhari, Ahmet Alacaoglu, Fabian Latorre, and Volkan Cevher. An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints. In *Proc. of 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, pages 13943–13955, 2019.

[18] Ron Shefi and Marc Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM Journal on Optimization*, 24(1) :269–297, 2014.

[19] Tuomo Valkonen. A primal-dual hybrid gradient method for nonlinear operators with applications to mri. *Inverse Problems*, 30(5) :055012, 2014.

[20] Yue Xie and Stephen J Wright. Complexity of proximal augmented lagrangian for non-convex optimization with nonlinear equality constraints. *Journal of Scientific Computing*, 86(3) :1–30, 2021.

[21] Alp Yurtsever, Quoc Tran Dinh, and Volkan Cevher. A universal primal-dual convex optimization framework. In *Advances in Neural Information Processing Systems*, pages 3150–3158, 2015.

[22] Zihan Zhou, Xiaodong Li, John Wright, Emmanuel Candes, and Yi Ma. Stable principal component pursuit. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1518–1522. IEEE, 2010.