Noname manuscript No.
(will be inserted by the editor)

A relaxation of the augmented Lagrange method

Received: date / Accepted: date

Abstract We propose a splitting method for solving ....

Keywords Non-linear constraint · Non-convex · Smoothing · Primal-dual

Mathematics Subject Classification (2010) 47H05 · 49M29 · 49M27 · 90C25

1 Introduction

Various problems in engineering and computational sciences can be cast as non-linear optimization programs, and the design of efficient numerical algorithms to provably solve such problems is therefore of fundamental importance. cite? In this paper, we are particularly interested in solving the optimization program

$$\begin{cases} \min_u h(u), \\ A(u) = b, \\ u \in C, \end{cases} \tag{1}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ and $A : \mathbb{R}^d \to \mathbb{R}^m$ satisfy

$$\|\nabla h(u) - \nabla h(u')\| \leq \lambda_h \|u - u'\|, \qquad \|Dh(u) - Dh(u')\| \leq \lambda_A \|u - u'\|, \tag{2}$$

for every $u, u' \in \mathbb{R}^d$. Above, $\nabla h(u) \in \mathbb{R}^d$ is the gradient of h and $DA(u) \in \mathbb{R}^{m \times d}$ is the Jacobian of A. Moreover, $C \subset \mathbb{R}^d$ is non-empty, closed, and convex. Variants of Program (1) naturally arise in a broad range of applications in ?? **Note: Please add some representative applications above alongside some references.** For the sake of brevity, we showcase here one instance of Program (1).

Example 1 **(Burer-Monteiro factorization)** Let $\mathbb{S}^{d' \times d'}$ be the space of $d' \times d'$ symmetric matrices, equipped with the standard inner product $\langle x | y \rangle = \mathrm{tr}(x^* y)$. In particular, when $x \in \mathbb{S}^{d' \times d'}$ is positive semi-definite, we write that $x \succeq 0$. Consider the program

$$\begin{cases} \min_x h'(x) \\ A'(x) = b' \\ x \in C' \\ x \succeq 0, \end{cases} \tag{3}$$

Address(es) of author(s) should be given

where $h' : \mathbb{S}^{d' \times d'} \to \mathbb{R}$, $A' \colon \mathbb{S}^{d' \times d'} \to \mathbb{R}^m$, $b \in \mathbb{R}^m$, and $C' \subseteq \mathbb{R}^{d' \times d'}$.

Variants of Program (3) are popular in matrix completion and sensing [17], with a broad range of applications to problems in collaborative filtering, geophysics, and imaging, among others [8,7,21]. Two common choices for C' in Program (3) are $C' = \{x : x \geq 0\}$ and $C' = \{x : \mathrm{tr}(x) \leq 1\}$ [**?**].

Solving Program (3) with semi-definite programming is not scalable, becoming increasingly cumbersome as the dimension d' grows. To overcome this computational bottleneck, the factorized technique sets $x = uu^\top$ for $u \in \mathbb{R}^{d' \times r}$ and a sufficiently large r. The resulting non-convex program is then solved with respect to the much lower-dimensional variable u. If we also replace the constraint $uu^\top \in C'$ with $u \in C$ for a properly chosen convex set, the new problem in u matches Program (1) with $h(u) = h'(uu^\top)$ and $A(u) = A'(uu^\top)$. For our examples of C' above, we might choose $C = \{u : u \geq 0\}$ and $C = \{\|u\|_F^2 \leq 1\}$, respectively. Here, $\|\cdot\|_F$ stands for the Frobenius norm.

The *augmented Lagrangian method* [15] is a powerful approach to solve Program (1), see Section 4 for a review of the related literature as well as other approaches to solve Program (1). Indeed, for positive β, it is easy to verify that Program (1) is equivalent to

$$\min_{u \in C} \max_{y} \mathcal{L}_\beta(u, y), \tag{4}$$

where

$$\mathcal{L}_\beta(u, y) := h(u) + \langle A(u) - b, y \rangle + \frac{\|A(u) - b\|^2}{2\beta}, \tag{5}$$

is the augmented Lagrangian corresponding to Program (1). The equivalent formulation in Program (4) naturally suggests the following algorithm to solve Program (1):

$$u_{k+1} \in \operatorname*{argmin}_{u \in C} \mathcal{L}_\beta(u, y_k), \tag{6}$$

$$y_{k+1} = y_k + \frac{A(u_{k+1}) - b}{\beta}. \tag{7}$$

In fact, when the penalty parameter β is sufficiently small, the augmented Lagrangian has a local minimum point near the true optimal point. However, we do not know exactly how small β is. Hence, the choice of β plays a centreral role in practices. **Note: Is the last claim really true? Programs (1) and (4) seem to be equivalent.** In our nonlinear framework, updating u in the augmented Lagrangian method requires solving the non-convex Program (6) to global optimality, which is often intractable. **Note: We should discuss fixes to this issue, if any, and explain why they are not satisfactory.** The key contribution of this paper is to provably and efficiently address this challenge.

Contributions. In order to solve Program (1), this paper proposes to replace the (intractable) Program (6) with the update

$$u_{k+1} = P_C(u_k - \gamma_k \nabla \mathcal{L}_{\beta_k}(u_k, y_k)), \tag{8}$$

for carefully selected sequences $\{\beta_k, \gamma_k\}_k$. Here, P_C is the orthogonal projection onto the convex set C which is often easy to compute in various applications and consequently the update in (8) is inexpensive and fast.

Put differently, instead of fully solving Program (6), this paper proposes to apply one iteration of the projected gradient algorithm for every update. We provide the convergence guarantees for this fast and scalable new algorithm. **Note: We should summarize the guarantees.**

2 Preliminaries

Note: I think the whole of this section should move down. The actual results are hidden deep in the paper!

Notation. We use the notations $\langle \cdot \mid \cdot \rangle$ and $\|\cdot\|$ for the standard inner product and the associated norm on $\mathbb{R}^d$, respectively. The adjoint of a linear operator is denoted the superscript $\top$. Let $C \subset \mathbb{R}^d$ be nonempty, closed, and convex. The indicator function of C is denoted by $\iota_{\mathcal{C}}$, and the projection onto C is denoted by P_C. For $u \in C$, the tangent cone to C at u is

$$T_C(u) = \left\{ v \in \mathbb{R}^d : \exists t > 0 \text{ such that } u + tv \in C \right\}. \tag{9}$$

The corresponding normal cone $N_C(u)$ at u is the polar of the tangent cone, namely,

$$N_C(u) = \{ v' : \langle v, v' \rangle \leq 0, \, \forall v \in T_C(u) \}. \tag{10}$$

The sub-differential of a convex function f at u is defined as

$$\partial f(u) = \{ g : f(u') - f(u) \geq \langle g, u' - u \rangle, \ \forall u' \}. \tag{11}$$

In particular, if f is differentiable at u, $\partial f(u)$ is a singleton and denoted by $\nabla f(u)$.

Necessary Optimality Conditions. Necessary optimality conditions for Program (1) are well studied in the literature [18, Corollary 6.15]. Indeed, u is a (first-order) stationary point of Program (1) if there exists y for which

$$\begin{cases} -\nabla h(u) - DA(u)^\top y \in N_C(u) \\ A(u) = b. \end{cases} \tag{12}$$

Here, $DA(u)$ is the Jacobian of A at u. Recalling (5), we observe that (13) is equivalent to

$$\begin{cases} -\nabla_u \mathcal{L}_\beta(u, y) \in N_C(u) \\ A(u) = b, \end{cases} \tag{13}$$

which is in turn the necessary optimality condition for Program (4).

Gradient Mapping. In nonconvex optimization, the relation between the gradient mapping and stationarity is well-understood [19,13,3], which we review here for completeness.

Definition 1 Given u and $\gamma > 0$, define the gradient mapping

$$G_{\beta,\gamma}(\cdot; y) \colon u \to \frac{u - u^+}{\gamma}, \tag{14}$$

where $u^+ = P_C(u - \gamma \nabla \mathcal{L}_\beta(u, y))$.

In particular, if we remove the constraints of Program (1), the gradient mapping reduces to $G_{\beta,\gamma}(u, y) = \nabla h(u)$. The gradient mapping is closely related to $\mathcal{L}_\beta$. The following standard result is proved in Appendix A.

Lemma 1 *For fixed $y \in \mathbb{R}^m$, suppose that $\nabla_u \mathcal{L}_\beta(\cdot, y)$ is λ_β Lipschitz-continuous. For $u \in C$ and $\gamma \in (0, 1/\lambda_\beta)$, it holds that*

$$\|G_{\beta,\gamma}(u; y)\|^2 \leq \frac{2}{\gamma}(\mathcal{L}_\beta(u; y) - \mathcal{L}_\beta(u^+; y)), \tag{15}$$

where

$$\lambda_\beta \leq \lambda_h + \sqrt{m}\lambda_A \left(\|y\| + \frac{\|A(u)\|}{\beta} \right) + \frac{\|DA(u)\|_F^2}{\beta}, \tag{16}$$

where $DA(u)$ is the Jacobian of A at u.

In practice, the Lipschitz constant λ_β is often hard to evaluate exactly and we might resort to the classic line search technique, reviewed below and proved in Appendix B for completeness.

Lemma 2 *Fix $\theta \in (0, 1)$ and γ_0. For $\gamma' > 0$, let $u^+_{\gamma'} = P_C(u - \gamma' \nabla \mathcal{L}_\beta(u, y))$ and define*

$$\gamma := \max \left\{ \gamma' = \gamma_0 \theta^i : \mathcal{L}_\beta(u^+_{\gamma'}, y) \leq \mathcal{L}_\beta(u, y) + \left\langle u^+_{\gamma'} - u, \nabla \mathcal{L}_\beta(u, y) \right\rangle + \frac{1}{2\gamma'} \|u^+_{\gamma'} - u\|^2 \right\}.$$

Then, (15) holds and moreover we have that

$$\gamma \geq \frac{\theta}{\lambda_\beta}. \tag{17}$$

Optimality conditions in Section 2 can also be expressed in terms of the gradient mapping. Indeed, it is straightforward to verify that u^+ is a first-order stationary point of Program (1) if

$$\begin{cases} G_{\beta,\gamma}(u, y) = 0 \\ A(u^+) = b. \end{cases} \tag{18}$$

Sufficient Optimality Conditions. Sufficient optimality conditions for Program (1) are also well understood in the litterature [15,18,16,12]. Indeed, u is a local minimizer of Program (1) if there exists y for which

$$\begin{cases} v^\top \left(\nabla_{uu} h(u) + \sum_{i=1}^m \nabla_{uu} A_i(u) \right) v \geq 0, & \forall v \in T_C(u), \\ A(u) = b. \end{cases} \tag{19}$$

Note: Why does above look different from sufficient cnds for Lagrangian? Suppose to be equivalent problems.

3 Algorithm & Convergence

3.1 Algorithm

We propose the following method for solving the problem (1) where, the main idea is that we do a projected gradient descent step on u to obtain u^+ and update the penalty parameter β^+ in such a way that the feasiblity $\frac{1}{2\beta^+\gamma}\|Lu^+ - b\|^2$ reduce faster than the gradient mapping up to some noise level ω:

$$\frac{1}{2\beta^+\gamma}\|Lu^+ - b\|^2 \leq \frac{1}{8}\|G_{\beta,\gamma}(u, y)\|^2 + \frac{\omega}{\gamma} \tag{20}$$

Then update the corresponding the multiplier y as in the classical ADMM:

$$y^+ = y + \frac{1}{\sigma}(LU^+ - b). \tag{21}$$

The formal algorithm is presented as follows.

Input: $\beta_0 > 0$, $c > 0$, $\alpha \in [0, 1[$, $u_0 \in \mathcal{C}$, $y_0 = 0$, $\epsilon_1 \in]0, 1[$. Given β_k, choose $\gamma_k \leq \frac{1-\epsilon_1}{\mathbb{L}_h + \mathbb{L}_{\beta_k}}$. Iterate
For k=0,1,...
1. Projected gradient: $u_{k+1} = P_{\mathcal{C}}(u_k - \gamma_k \nabla F_{\beta_k}(u_k, y_k))$.
2. Line search step
$s = 0, d_{k,0} = 2$, $\beta_{k+1,0} = \frac{1}{2}\|Lu_{k+1} - b\|^2 \left(\frac{\gamma_k}{8} \|G_{\beta_k,\gamma_k}(u_k, y_k)\|^2 + \frac{d_{k,s}}{(1+k)^{1+\epsilon_1}} \right)^{-1}$.
While $\beta_{k+1,s} \geq c/(k+1)^\alpha$ do

$$d_{k,s+1} = 2 * d_{k,s} \tag{22}$$

$$\beta_{k+1,s+1} = \frac{1}{2}\|Lu_{k+1} - b\|^2 \left(\frac{\gamma_k}{8} \|G_{\beta_k,\gamma_k}(u_k, y_k)\|^2 + \frac{d_{k,s+1}}{(1+k)^{1+\epsilon_1}} \right)^{-1} \tag{23}$$

$$s \leftarrow s + 1. \tag{24}$$

Endwhile
3. Update $\beta_{k+1} = \beta_{k+1,s}$.
4. Chose $\sigma_{k+1} \geq 2\beta_k$ and update $y_{k+1} = y_k + \frac{1}{\sigma_{k+1}}(Lu_{k+1} - b)$.

Remark 1 The updating rule of $(\beta_k)_{k\in\mathbb{N}}$ in (22) plays a role in our analysis. Intuitively, if u_{k+1} is solution then $Lu_{k+1} = b$ and (22) is trivially satisfied for any $\beta_{k+1} \geq 0$. Hence β_{k+1} enforces u_{k+1} close to $\{u \mid Lu = b\}$

Remark 2 When $\sigma_k \equiv \infty$, we get $y_k \equiv 0$ and hence the step 2 disappears. If we chose $\sigma_k = c(k+1)^{\alpha_1}\|Lu_k - b\|$ where c, α_1 is chosen such that $\sigma_k > 2\beta_{k-1}$, then

$$\|y_{k+1}\| \leq \|y_k\| + \frac{\|Lu_{k+1} - b\|}{\sigma_{k+1}} = \|y_k\| + \frac{1}{c(k+2)^\alpha}. \tag{25}$$

Since $\sum_{k\in\mathbb{N}} \frac{1}{c(k+2)^\alpha} < +\infty$, $(\|y_k\|)_{k\in\mathbb{N}}$ converges and hence bounded. Therefore,

$$b_0 = \inf_{k\in\mathbb{N}} \mathcal{L}_{\beta_k}(u_{k+1}, y_k) \geq \inf_k h(u_k), \tag{26}$$

which implies that $b_0 > -\infty$ whenever $(u_k)_{k\in\mathbb{N}}$ or $\mathrm{dom}(f)$ is bounded.

3.2 Convergence

In view of Lemma **??**, we need to estimate gradient mapping $\|G_{\beta_k,\gamma_k}(u_k, y_k)\|$ as well as feasibility $\|Lu_{n+1} - b\|^2$.

Theorem 1 *Suppose that* $b_0 = \inf_{k\in\mathbb{N}} \mathcal{L}_{\beta_k}(u_{k+1}, y_k) > -\infty$ *and that* $z_0 = \sum_{k=1}^{\infty} \frac{d_{k,s_k}}{(1+k)^{1+\epsilon_1}} < +\infty$, *where* s_k *be the smallest index such that* $\beta_{k,s_k} < c/(k+1)^\alpha$. *Then the following hold.*

$$\sum_{k=1}^{\infty} \gamma_k \|G_{\beta_k,\gamma_k}(u_k, y_k)\|^2 \leq 4(\mathcal{L}_{\beta_0}(u_1, y_0) + z_0 - b_0 + \frac{\gamma_0}{8}\|G_{\beta_0,\gamma_0}(u_0)\|^2), \tag{27}$$

and

$$\sum_{k=1}^{\infty} \frac{1}{\beta_{k+1}} \|Lu_{k+1} - b\|^2 \leq (\mathcal{L}_{\beta_0}(u_1, y_0) + 3z_0 - b_0 + \frac{\gamma_0}{8} \|G_{\beta_0,\gamma_0}(u_0)\|^2). \tag{28}$$

4 Related Work

To the best of our knowledge, the proposed method is new and different from existing methods in the literature.

As mentioned in Introduction, the connection to augmented Lagrange method is already mentioned. Our method is significantly different from the augmented Lagrange method, we perform only step of the projected gradient step on primal variable u instead of minimizing the augmented Lagrange fucntion. Furthermore, we update the penalty parameter β adaptively to make sure that the feasibility reduces faster than the gradient mapping.

In the case when $h = 0$, a modification of Chambolle-Pock's method is investigated in [22] and preconditioned ADMM [2] where the convergence of iterate is proved under strong assumptions not full-filling in our setting here.

ADMM is the classic method proposed for solving the problem 1 for the case where L is a linear operator and h is zero [10]. This method is an application of the Douglas-Rachford method to the dual problem [9]. One of the main drawback of the ADMM is the appearance of the term Lu in the update rule of u_{k+1}. To overcome this issue, some strategies were suggested. The first strategies is proposed in [20], refined in [1], known as alternating direction proximal method of multipliers. The second strategies is to use linearized technique [14]. We show here that our proposed method is closed related to updating rule as the linearized alternating direction method [14]. Assume that $h \equiv 0$ and L is a linear operator. Then the proposed method can be rewritten as

$$\begin{cases} u_{k+1} = \arg\min_{u \in C} \frac{1}{2\gamma_k} \|u - u_k + \gamma_k L^* \Big(\lambda_k + \frac{1}{\beta_k} \big(Lu_k - b \big) \Big) \|^2 \\ \beta_{k+1} = \frac{1}{2} \|Lu_{k+1} - b\|^2 \Big(\frac{\gamma_k}{8} \|G_{\beta_k,\gamma_k}(u_k, y_k)\|^2 + \frac{d_k}{(k+1)^\alpha} \Big)^{-1} \\ \text{Chose } \sigma_{k+1} \geq 2\beta_k \text{ and } \lambda_{k+1} = \lambda_k + \frac{1}{\sigma_{k+1}} (Lu_{k+1} - b), \end{cases}$$

which is a variant version of Linearized ADMM [14].

Very recently, [4] proposes a framework with for solving the problem 1 with $C = \mathbb{R}^d$. In particular, a special case AlBUM3 (Proximal Linearized Alternating Minimization) in this work is closely related to us where their conditions are checkable only when L is linear. Moreover, our updating of β_k in [4] depending on the smallest eigenvalue L^*L. For nonlinear L, the application of their method remains a challenge.

The deflected subgradient method is investigated in [6] can be use to solve a special case of the Problem 1 for some a compact subset $\mathcal{C}$ in $\mathcal{X}$. The basis step of the deflected subgradient method to solve: given β, v,

$$u^* \in \arg\min_{u \in C} h(u) + \beta \boldsymbol{\sigma}(Lu - b) - \langle Kv \mid Lu - b \rangle \tag{29}$$

where $\boldsymbol{\sigma}$ is a continuous penalty function such as $\|\cdot\|$, and K is bounded linear operator. In general, there is no closed -form expression for u^* since it does not split f, h, L invididually. Hence, it is hard to implement deflected subgradient method. This is also a common drawback of the classic penalty method and its related works [11,5].

5 Numerical experiments

5.1 Hanging chain

References

1. S. Banert, R. I. Bot, and E. R. Csetnek. Fixing and extending some recent results on the admm algorithm. *arXiv preprint arXiv:1612.05057*, 2016.
2. M. Benning, F. Knoll, C.-B. Schönlieb, and T. Valkonen. Preconditioned admm with nonlinear operator constraint. In L. Bociu, J.-A. Désidéri, and A. Habbal, editors, *System Modeling and Optimization*, pages 117–126, Cham, 2016. Springer International Publishing.
3. J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization or nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.
4. J. Bolte, S. Sabach, and M. Teboulle. Nonconvex lagrangian-based optimization: monitoring schemes and global convergence. *Mathematics of Operations Research*, 2018.
5. R. S. Burachik, A. N. Iusem, and J. G. Melo. A primal dual modified subgradient algorithm with sharp lagrangian. *Journal of Global Optimization*, 46(3):347–361, 2010.
6. R. S. Burachik and C. Y. Kaya. A deflected subgradient method using a general augmented lagrangian duality with implications on penalty methods. In *Variational Analysis and Generalized Differentiation in Optimization and Control*, pages 109–132. Springer, 2010.
7. S. Burer and R. D. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, Feb 2003.
8. S. Burer and R. D. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, Jul 2005.
9. 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.
10. D. Gabay and B. 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.
11. R. N. Gasimov. Augmented lagrangian duality and nondifferentiable optimization methods in nonconvex programming. *Journal of Global Optimization*, 24(2):187–203, 2002.
12. H. Gfrerer and B. S. Mordukhovich. Complete characterizations of tilt stability in nonlinear programming under weakest qualification conditions. *SIAM Journal on Optimization*, 25(4):2081–2119, 2015.
13. W. Hare and C. Sagastizábal. Computing proximal points of nonconvex functions. *Mathematical Programming*, 116(1):221–258, 2009.
14. Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in neural information processing systems*, pages 612–620, 2011.
15. D. G. Luenberger, Y. Ye, et al. *Linear and nonlinear programming*, volume 2. 2007.
16. B. S. Mordukhovich. *Variational analysis and generalized differentiation I: Basic theory*, volume 330. Springer Science & Business Media, 2006.
17. D. Park, A. Kyrillidis, S. Bhojanapalli, C. Caramanis, and S. Sanghavi. Provable burer-monteiro factorization for a class of norm-constrained matrix problems. *arXiv preprint arXiv:1606.01316*, 2016.
18. R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.
19. H. X. S. Ghadimi, G. Lan. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program., Ser. A*, 155:267–305, 2016.
20. R. Shefi and M. 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.
21. S. Tu and J. Wang. Practical first order methods for large scale semidefinite programming. 2014.
22. T. Valkonen. A primal-dual hybrid gradient method for nonlinear operators with applications to mri. *Inverse Problems*, 30(5):055012, 2014.

A Proof of Lemma 1

Note that (15) follows immediately from an application of [3, Lemma 3.2, Remark 3.2(i)]. It only remains to compute the smoothness parameter of $\mathcal{L}_\beta(\cdot, y)$, namely, λ_β. To that end, note that

$$\mathcal{L}_\beta(u, y) = h(u) + \sum_{i=1}^{m} y_i(A_i(u) - b_i) + \frac{1}{2\beta}\sum_{i=1}^{m}(A_i(u) - b_i)^2, \tag{30}$$

which implies that

$$\begin{aligned}\nabla_u \mathcal{L}_\beta(u,y) &= \nabla h(u) + \sum_{i=1}^m y_i \nabla A_i(u) + \frac{1}{\beta}\sum_{i=1}^m (A_i(u)-b_i)\nabla A_i(u) \\ &= \nabla h(u) + DA(u)^\top y + \frac{DA(u)^\top (A(u)-b)}{\beta}, \end{aligned} \tag{31}$$

where $DA(u)$ is the Jacobian of A at u. Likewise,

$$\nabla_u^2 \mathcal{L}_\beta(u,y) = \nabla^2 h(u) + \sum_{i=1}^m \left(y_i + \frac{A_i(u)}{\beta}\right)\nabla^2 A_i(u) + \frac{1}{\beta}\sum_{i=1}^m \nabla A_i(u)\nabla A_i(u)^\top. \tag{32}$$

It follows that

$$\begin{aligned}\|\nabla_u^2 \mathcal{L}_\beta(u,y)\| &\le \|\nabla^2 h(u)\| + \max_i \|\nabla^2 A_i(u)\| \left(\|y\|_1 + \frac{\|A(u)-b\|_1}{\beta}\right) + \frac{1}{\beta}\sum_{i=1}^m \|\nabla A_i(u)\|^2 \\ &\le \lambda_h + \sqrt{m}\lambda_A \left(\|y\| + \frac{\|A(u)-b\|}{\beta}\right) + \frac{\|DA(u)\|_F^2}{\beta} \qquad (h \in \mathbb{L}(\lambda_h),\ A_i \in \mathbb{L}(\lambda_A)) \\ &= \lambda_h + \sqrt{m}\lambda_A \left(\|y\| + \frac{\|A(u)-b\|}{\beta}\right) + \frac{\|DA(u)\|_F^2}{\beta}, \end{aligned} \tag{33}$$

and, consequently,

$$\begin{aligned}\lambda_\beta &= \sup_u \|\nabla_u^2 \mathcal{L}_\beta(u,y)\| \\ &\le \lambda_h + \sqrt{m}\lambda_A \left(\|y\| + \frac{\|A(u)-b\|}{\beta}\right) + \frac{\|DA(u)\|_F^2}{\beta}, \end{aligned} \tag{34}$$

which completes the proof of Lemma 1.

B Proof of Lemma 2

Since $u, u_\gamma^+ \in C$, it holds that

$$u_\gamma^+ - u \in -T_C(u_\gamma^+). \tag{35}$$

Also, recalling u_γ^+ in Definition 1, we have that

$$u_\gamma^+ - u + \gamma \nabla \mathcal{L}_\beta(u,y) \in -N_C(u_\gamma^+). \tag{36}$$

Lastly, γ by definition in (2) satisfies

$$\begin{aligned}&\mathcal{L}_\beta(u_\gamma^+, y) \\ &\le \mathcal{L}_\beta(u,y) + \left\langle u_\gamma^+ - u, \nabla \mathcal{L}_\beta(u,y)\right\rangle + \frac{1}{2\gamma}\|u_\gamma^+ - u\|^2 \\ &= \mathcal{L}_\beta(u,y) + \frac{1}{\gamma}\left\langle u_\gamma^+ - u, u_\gamma^+ - u + \gamma\nabla \mathcal{L}_\beta(u,y)\right\rangle - \frac{1}{2\gamma}\|u_\gamma^+ - u\|^2 \\ &\le \mathcal{L}_\beta(u,y) - \frac{1}{2\gamma}\|u_\gamma^+ - u\|^2 \qquad \text{(see (35,36))} \\ &= \mathcal{L}_\beta(u,y) - \frac{\gamma}{2}\|G_{\beta,\gamma}(u,y)\|^2, \qquad \text{(see Definition 1)} \end{aligned} \tag{37}$$

which completes the proof of Lemma 2.

C Draft of convergence proof

For convenience, let us recall the updates of the algorithm in iteration k, namely,

$$\begin{aligned} u_{k+1} &= P_C(u_k - \gamma_k \nabla \mathcal{L}_{\beta_k}(u_k, y_k)) \\ &= P_C\left(u_k - \gamma_k \nabla h(u_k) - \gamma_k DA(u_k)^\top \left(y_k + \frac{A(u_k) - b}{\beta_k}\right)\right), \qquad \text{(see (5))} \end{aligned} \tag{38}$$

$$y_{k+1} = y_k + \frac{A(u_{k+1}) - b}{\sigma_{k+1}}, \tag{39}$$

$$G_k = G_{\beta_k, \gamma_k}(u_k, y_k) = \frac{u_k - u_{k+1}}{\gamma_k}. \qquad \text{(see (14))} \tag{40}$$

For integers $k_0 \leq k_1$, consider the interval

$$K = [k_0 : k_1] = \{k_0, \cdots, k_1\}. \tag{41}$$

Since γ_k is determined by the line search subroutine in Lemma 2, we may now apply Lemma 1 for every iteration in this interval to find that

$$\begin{aligned} \frac{\gamma_k \|G_k\|^2}{2} &\leq \mathcal{L}_{\beta_k}(u_k, y_k) - \mathcal{L}_{\beta_k}(u_{k+1}, y_k) \qquad \text{(see Lemma 1)} \\ &= h(u_k) - h(u_{k+1}) + \langle A(u_k) - A(u_{k+1}), y_k \rangle + \frac{\|A(u_k) - b\|^2 - \|A(u_{k+1}) - b\|^2}{2\beta_k}, \qquad \text{(see (5))} \end{aligned} \tag{42}$$

for every $k \in K$. On the other hand,

$$y_k = y_{k_0} + \sum_{i=k_0+1}^{k} \frac{A(u_i) - b}{\sigma_i}, \qquad \text{(see (39))} \tag{43}$$

which, after substituting in (42), yields that

$$\frac{\gamma_k \|G_k\|^2}{2} \leq h(u_k) - h(u_{k+1}) + \left\langle A(u_k) - A(u_{k+1}), y_{k_0} + \sum_{i=k_0+1}^{k} \frac{A(u_i) - b}{\sigma_i} \right\rangle + \frac{\|A(u_k) - b\|^2 - \|A(u_{k+1}) - b\|^2}{2\beta_k}. \tag{44}$$

Additionally, let us assume that

$$\beta_k = \frac{\beta}{\sqrt{k}}, \qquad \sigma_k = \beta k, \qquad \forall k \in K, \tag{45}$$

with $\beta > 0$. For the record, the above assumptions imply that

$$\frac{1}{\beta_k} - \frac{1}{\beta_{k-1}} \leq \frac{1}{\beta\sqrt{k}}, \qquad \frac{1}{\sigma_k} \leq \frac{1}{\beta\sqrt{k}}, \qquad \forall k \in K, \tag{46}$$

for sufficiently large k_0. By summing up the key inequality in (44) over k from k_0 to k_1 and using (45), we find that

$$
\begin{aligned}
&\sum_{k=k_0}^{k_1} \frac{\gamma_k \|G_k\|^2}{2} \\
&\le h(u_{k_0}) - h(u_{k_1+1}) + \langle A(u_{k_0}) - A(u_{k_1+1}), y_{k_0}\rangle + \sum_{k=k_0}^{k_1} \sum_{i=k_0+1}^{k} \left\langle A(u_k) - A(u_{k+1}), \frac{A(u_i)-b}{\sigma_i} \right\rangle \\
&\quad + \sum_{k=k_0}^{k_1} \frac{\|A(u_k)-b\|^2}{2\beta_k} - \sum_{k=k_0}^{k_1} \frac{\|A(u_{k+1})-b\|^2}{2\beta_k} \qquad \text{(see (44))} \\
&= h(u_{k_0}) - h(u_{k_1+1}) + \langle A(u_{k_0}) - A(u_{k_1+1}), y_{k_0}\rangle + \sum_{k=k_0}^{k_1} \sum_{i=k_0+1}^{k} \left\langle A(u_k) - A(u_{k+1}), \frac{A(u_i)-b}{\sigma_i} \right\rangle \\
&\quad + \sum_{k=k_0}^{k_1} \frac{\|A(u_k)-b\|^2}{2\beta_k} - \sum_{k=k_0+1}^{k_1+1} \frac{\|A(u_k)-b\|^2}{2\beta_{k-1}} \\
&\le h(u_{k_0}) - h(u_{k_1+1}) + \langle A(u_{k_0}) - A(u_{k_1+1}), y_{k_0}\rangle + \frac{\|A(u_{k_0})-b\|^2}{2\beta_{k_0}} + \sum_{i=k_0+1}^{k_1} \sum_{k=i}^{k_1} \left\langle A(u_k) - A(u_{k+1}), \frac{A(u_i)-b}{\sigma_i} \right\rangle \\
&\quad + \sum_{k=k_0+1}^{k_1} \left(\frac{1}{2\beta_k} - \frac{1}{2\beta_{k-1}} \right) \|A(u_k)-b\|^2 \\
&\le \mu + \sum_{i=k_0+1}^{k_1} \left\langle A(u_i) - A(u_{k_1+1}), \frac{A(u_i)-b}{\sigma_i} \right\rangle + \sum_{k=k_0+1}^{k_1} \left(\frac{1}{2\beta_k} - \frac{1}{2\beta_{k-1}} \right) \|A(u_k)-b\|^2 \qquad \text{(see (48))} \\
&= \mu + \sum_{k=k_0+1}^{k_1} \left(\frac{1}{\sigma_k} + \frac{1}{2\beta_k} - \frac{1}{2\beta_{k-1}} \right) \|A(u_k)-b\|^2 - \sum_{k=k_0+1}^{k_1} \left\langle A(u_{k_1+1}) - b, \frac{A(u_k)-b}{\sigma_k} \right\rangle \\
&\le \mu + \sum_{k=k_0+1}^{k_1} \left(\frac{1}{\sigma_k} + \frac{1}{2\beta_k} - \frac{1}{2\beta_{k-1}} \right) \|A(u_k)-b\|^2 + \sum_{k=k_0+1}^{k_1} \frac{1}{\sigma_k} \|A(u_{k_1+1})-b\| \|A(u_k)-b\| \\
&\le \mu + \sum_{k=k_0+1}^{k_1} \left(\frac{1}{\beta\sqrt{k}} + \frac{1}{2\beta_k} - \frac{1}{2\beta_{k-1}} \right) \|A(u_k)-b\|^2 + \sum_{k=k_0+1}^{k_1} \frac{1}{\beta\sqrt{k}} \|A(u_{k_1+1})-b\| \|A(u_k)-b\| \qquad \text{(see (46))} \\
&\le \mu + \sum_{k=k_0+1}^{k_1} \frac{3}{2\beta\sqrt{k}} \|A(u_k)-b\|^2 + \sum_{k=k_0+1}^{k_1} \frac{1}{\beta\sqrt{k}} \|A(u_{k_1+1})-b\| \|A(u_k)-b\| \qquad \text{(see (46))} \\
&\le \mu + \sum_{k=k_0+1}^{k_1} \left(\frac{3}{2\beta\sqrt{k}} + \frac{1}{2\beta} \right) \|A(u_k)-b\|^2 + \sum_{k=k_0+1}^{k_1} \frac{1}{2\beta k} \|A(u_{k_1+1})-b\|^2 \qquad \left(2ab \le ca^2 + c^{-1}b^2\right) \\
&\le \mu + \sum_{k=k_0+1}^{k_1} \frac{2}{\beta} \|A(u_k)-b\|^2 + \sum_{k=k_0+1}^{k_1} \frac{1}{2\beta k} \|A(u_{k_1+1})-b\|^2 \\
&\le \mu + \sum_{k=k_0+1}^{k_1} \frac{2}{\beta} \|A(u_k)-b\|^2 + \frac{\log k_1}{\beta} \|A(u_{k_1+1})-b\|^2 \qquad \left(\sum_{k=1}^{k_1} \frac{1}{k} \le 2 \int_1^{k_1} \frac{d\kappa}{\kappa} = 2 \log k_1 \right) \\
&\le \mu + \frac{2}{\beta} \sum_{k=k_0+1}^{k_1+1} \log k \|A(u_k)-b\|^2, \qquad (47)
\end{aligned}
$$

where we assumed that

$$
\mu := \sup_k h(u_{k_0}) - h(u_k) + \langle A(u_{k_0}) - A(u_k), y_{k_0}\rangle + \frac{\|A(u_{k_0})-b\|^2}{2\beta_{k_0}} < \infty. \qquad (48)
$$

Note that (47) bounds the gradient mapping with the feasibility gap. We next find a converse, thus bounding the feasibility gap with the gradient mapping. To that end, let $T_C(u)$ and $P_{T_C(u)}$ be the tangent cone of C at $u \in C$ and orthogonal projection onto this subspace, respectively. Likewise, let $N_C(u)$ and $P_{N_C(u)}$ be the normal cone of C at u and the corresponding orthogonal projection. The update rule for u_k in (38) immediately implies that

$$G_k - \nabla h(u_k) - DA(u_k)^\top y_k - \frac{1}{\beta_k} DA(u_k)^\top (A(u_k) - b) \in N_C(u_{k+1}). \tag{49}$$

By definition in (40), we have that $G_k \in T_C(u_{k+1})$ which, together with (49), imply that

$$\begin{aligned} G_k &= P_{T_C(u_{k+1})} \left(-\nabla h(u_k) - DA(u_k)^\top y_k - \frac{1}{\beta_k} DA(u_k)^\top (A(u_k) - b) \right) \\ &= P_{T_C(u_{k+1})}(-\nabla h(u_k)) + P_{T_C(u_{k+1})}(-DA(u_k)^\top y_k) + \frac{1}{\beta_k} P_{T_C(u_{k+1})}(-DA(u_k)^\top (A(u_k) - b)) \\ &= P_{T_C(u_{k+1})}(-\nabla h(u_k)) + P_{T_C(u_{k+1})}(-DA(u_k)^\top y_{k-1}) + \left(\frac{1}{\beta_k} + \frac{1}{\sigma_k} \right) P_{T_C(u_{k+1})}(-DA(u_k)^\top (A(u_k) - b)), \end{aligned} \tag{50}$$

where the last line above uses (39). After rearranging and applying the triangle inequality above, we reach

$$\begin{aligned} \frac{1}{\beta_k} \| P_{T_C(u_{k+1})}(DA(u_k)^\top (A(u_k) - b)) \| &\leq \left(\frac{1}{\sigma_k} + \frac{1}{\beta_k} \right) \| P_{T_C(u_{k+1})}(DA(u_k)^\top (A(u_k) - b)) \| \\ &\leq \| \nabla h(u_k) \| + \| DA(u_k) \| \cdot \| y_{k-1} \| + \| G_k \| \\ &\leq \lambda_h' + \eta_{\max} \| y_{k-1} \| + \| G_k \|, \end{aligned} \tag{51}$$

where we set

$$\lambda_h' := \max_{u \in C} \| \nabla h(u) \|, \qquad \eta_{\max} = \max_{u \in C} \| DA(u) \|. \tag{52}$$

We next translate (51) into an upper bound on $\| A(u_k) - b \|$.

Lemma 3 *For an integer k_0, let*

$$S_K \supseteq \bigcup_{k \in K} T_C(u_k), \tag{53}$$

and, with some abuse of notation, let S_K also denote an orthonormal basis for this subspace. For $\rho > 0$, suppose that there exists $\eta_{\min}$ such that

$$0 < \eta_{\min} := \begin{cases} \min_u \left\| S_K^\top P_{T_C(u)}(DA(u)^\top v) \right\| \\ \| v \| = 1 \\ \| A(u) - b \| \leq \rho \\ u \in C. \end{cases} \tag{54}$$

Suppose also that

$$\sup_{k \in K} \| A(u_k) - b \| \leq \rho, \tag{55}$$

$$\operatorname{diam}(C) \leq \frac{\eta_{\min}}{2 \lambda_A}. \tag{56}$$

Then it holds that

$$\| A(u_k) - b \| \leq \frac{2 \beta_k}{\eta_{\min}} \left(\lambda_h' + \eta_{\max} \| y_{k-1} \| + \| G_k \| \right), \qquad \forall k \in K. \tag{57}$$

Roughly speaking, (57) states that the feasibility gap is itself bounded by the gradient map. In order to apply Lemma 3, let us assume that (55) holds. Lemma 3 is then in force and we may now substitute (57) back into (47) to find that

$$\begin{aligned}
\sum_{k=k_0}^{k_1} \gamma_k \|G_k\|^2 &\leq \frac{4}{\beta} \sum_{k=k_0+1}^{k_1+1} \log k \|A(u_k) - b\|^2 + 2\mu \qquad \text{(see (47))} \\
&\leq \frac{16}{\beta \eta_{\min}^2} \sum_{k=k_0+1}^{k_1+1} \beta_k^2 \log k \left(\lambda_h' + \eta_{\max}\|y_{k-1}\| + \|G_k\|\right)^2 + 2\mu \qquad \text{(see (57))} \\
&\leq \frac{16\beta}{\eta_{\min}^2} \sum_{k=k_0+1}^{k_1+1} \frac{\log k}{k} \left(\lambda_h' + \eta_{\max}\|y_{k-1}\| + \|G_k\|\right)^2 + 2\mu \qquad \text{(see (45))} \\
&\leq \frac{48\beta}{\eta_{\min}^2} \sum_{k=k_0+1}^{k_1+1} \frac{\log k}{k} \left(\lambda_h'^2 + \eta_{\max}^2\|y_{k-1}\|^2 + \|G_k\|^2\right) + 2\mu. \qquad \left((a+b+c)^2 \leq 3(a^2+b^2+c^2)\right) \qquad (58)
\end{aligned}$$

To simplify the above expression, let us assume that

$$\frac{48\beta \log k}{\eta_{\min}^2 k} \leq \frac{\gamma_k}{2}, \qquad \forall k \in K. \tag{59}$$

Let $|K| = k_1 - k_0 + 1$ be the size of the interval K. After rearranging (58) and applying (59), we arrive at

$$\begin{aligned}
&\frac{|K|}{2} \cdot \min_{k \in K} \gamma_k \|G_k\|^2 \\
&\leq \sum_{k=k_0}^{k_1} \frac{\gamma_k}{2} \|G_k\|^2 \\
&\leq \sum_{k=k_0}^{k_1} \left(\gamma_k - \frac{48\beta \log k}{\eta_{\min}^2 k}\right) \|G_k\|^2 \qquad \text{(see (59))} \\
&\leq \frac{48\beta\lambda_h'^2}{\eta_{\min}^2} \sum_{k=k_0+1}^{k_1+1} \frac{\log k}{k} + \frac{48\beta\eta_{\max}^2}{\eta_{\min}^2} \sum_{k=k_0+1}^{k_1+1} \frac{\|y_{k-1}\|^2 \log k}{k} + \frac{48\beta \log(k_1+1)\|G_{k_1+1}\|^2}{\eta_{\min}^2 (k_1+1)} + 2\mu \qquad \text{(see (58))} \\
&=: \frac{48\beta\lambda_h'^2}{\eta_{\min}^2} \sum_{k=k_0+1}^{k_1+1} \frac{\log k}{k} + \frac{48\beta\eta_{\max}^2}{\eta_{\min}^2} \sum_{k=k_0+1}^{k_1+1} \frac{\|y_{k-1}\|^2 \log k}{k} + 2\mu' \\
&\leq \frac{48\beta\lambda_h'^2 \log(k_1+1)}{\eta_{\min}^2} \sum_{k=k_0+1}^{k_1+1} \frac{1}{k} + \frac{48\beta\eta_{\max}^2 \log(k_1+1)}{\eta_{\min}^2} \sum_{k=k_0+1}^{k_1+1} \frac{\|y_{k-1}\|^2}{k} + 2\mu' \\
&\leq \frac{96\beta\lambda_h'^2 \log^2(k_1+1)}{\eta_{\min}^2} + \frac{48\beta\eta_{\max}^2 \log(k_1+1)}{\eta_{\min}^2} \sum_{k=k_0}^{k_1} \frac{\|y_k\|^2}{k+1} + 2\mu' \\
&\leq \frac{96\beta \log^2(k_1+1)}{\eta_{\min}^2} \left(\lambda_h'^2 + \eta_{\max}^2 \sum_{k=k_0}^{k_1} \frac{\|y_k\|^2}{k+1}\right) + 2\mu', \qquad (60)
\end{aligned}$$

or, equivalently,

$$\min_{k \in K} \gamma_k \|G_k\|^2 \leq \frac{192\beta \log^2(k_1+1)}{\eta_{\min}^2 |K|} \left(\lambda_h'^2 + \eta_{\max}^2 |K| B_K\right) + \frac{4\mu'}{|K|}, \tag{61}$$

where

$$\mu' := \frac{24\beta \log(k_1+1)\|G_{k_1+1}\|^2}{\eta_{\min}^2 (k_1+1)} + \mu, \tag{62}$$

$$B_K := \frac{1}{|K|}\sum_{k=k_0}^{k_1}\frac{\|y_k\|^2}{k+1}, \tag{63}$$

and we will later estimate both μ', B_K. In turn, the bound above on the gradient mapping controls the feasibility gap, namely,

$$\begin{aligned}
&|K|\min_{k-1\in K}\log k\|A(u_k)-b\|^2\\
&\le \sum_{k=k_0+1}^{k_1+1}\log k\|A(u_k)-b\|^2\\
&\le \frac{12\beta^2\lambda_h'^2}{\eta_{\min}^2}\sum_{k=k_0+1}^{k_1+1}\frac{\log k}{k}+\frac{12\beta^2\eta_{\max}^2}{\eta_{\min}^2}\sum_{k=k_0+1}^{k_1+1}\frac{\|y_{k-1}\|^2\log k}{k}+\sum_{k=k_0+1}^{k_1+1}\frac{12\beta^2\log k}{\eta_{\min}^2 k}\|G_k\|^2 \qquad \text{(see (58))}\\
&\le \frac{24\beta^2\lambda_h'^2\log^2(k_1+1)}{\eta_{\min}^2}+\frac{12\beta^2\eta_{\max}^2\log(k_1+1)}{\eta_{\min}^2}\cdot|K|B_K+\frac{\beta}{8}\sum_{k=k_0}^{k_1}\gamma_k\|G_k\|^2+\frac{12\beta^2\log(k_1+1)\|G_{k_1+1}\|^2}{\eta_{\min}^2(k_1+1)} \qquad \text{(see (59,63))}\\
&\le \frac{24\beta^2\log^2(k_1+1)}{\eta_{\min}^2}\left(\lambda_h'^2+\eta_{\max}^2|K|B_K\right)+\frac{24\beta^2\log^2(k_1+1)}{\eta_{\min}^2}\left(\lambda_h'^2+\eta_{\max}^2|K|B_K\right)+\frac{\beta\mu'}{2}+\frac{12\beta^2\log(k_1+1)\|G_{k_1+1}\|^2}{\eta_{\min}^2(k_1+1)} \qquad \text{(see (60))}\\
&\le \frac{24\beta^2\log^2(k_1+1)}{\eta_{\min}^2}\left(\lambda_h'^2+\eta_{\max}^2|K|B_K\right)+\frac{24\beta^2\log^2(k_1+1)}{\eta_{\min}^2}\left(\lambda_h'^2+\eta_{\max}^2|K|B_K\right)+\beta\mu' \qquad \text{(see (62))}\\
&= \frac{48\beta^2\log^2(k_1+1)}{\eta_{\min}^2}\left(\lambda_h'^2+\eta_{\max}^2|K|B_K\right)+\beta\mu',
\end{aligned} \tag{64}$$

which in turn implies that

$$\min_{k-1\in K}\|A(u_k)-b\|^2 \le \frac{48\beta^2\log^2(k_1+1)}{\eta_{\min}^2|K|}\left(\lambda_h'^2+\eta_{\max}^2|K|B_K\right)+\frac{\beta\mu'}{|K|}, \tag{65}$$

if $k_0\ge 2$. Let us now revisit and simplify the condition imposed in (55). To that end, we first derive a weaker but uniform bound on the feasibility gap. For every $k-1\in K$, it holds that

$$\begin{aligned}
\|A(u_k)-b\|^2 &\le \sum_{i=k_0+1}^{k_1+1}\log i\|A(u_i)-b\|^2 \qquad (k_0\ge 2)\\
&\le \frac{48\beta^2\log^2(k_1+1)}{\eta_{\min}^2}\left(\lambda_h'^2+\eta_{\max}^2|K|B_K\right)+\beta\mu'. \qquad \text{(see (65))}
\end{aligned} \tag{66}$$

Therefore, we may replace (55) with the assumption that

$$\|A(u_{k_0})-b\|\le\rho, \qquad \frac{48\beta^2\log^2(k_1+1)}{\eta_{\min}^2}\left(\lambda_h'^2+\eta_{\max}^2|K|B_K\right)+\beta\mu'\le\rho^2, \tag{67}$$

which ensures that

$$\|A(u_k)-b\|\le\rho, \qquad \forall k\in[k_0:k_1+1]. \tag{68}$$

In order to interpret (60,65,67), we next estimate B_K in (63). To that end, let us first control the growth of the dual sequence $\{y_k\}_k$. Recalling (39) and for every $k\in[k_0:k_1+1]$, we write that

$$\begin{aligned}
\|y_k\| &\le \|y_{k_0}\|+\sum_{i=k_0+1}^{k}\frac{\|A(u_i)-b\|}{\sigma_k} \qquad \text{(see (39))}\\
&\le \|y_{k_0}\|+\sum_{i=k_0+1}^{k}\frac{\rho}{\beta k} \qquad \text{(see (45,68))}\\
&\le \|y_{k_0}\|+\frac{2\rho\log k}{\beta}.
\end{aligned} \tag{69}$$

With the growth of the dual sequence uncovered, we evaluate B_K as

$$\begin{aligned}
B_K &= \frac{1}{|K|}\sum_{k=k_0}^{k_1}\frac{\|y_k\|^2}{k+1} && \text{(see (63))}\\
&\le \frac{1}{|K|}\sum_{k=k_0}^{k_1}\frac{1}{k+1}\left(\|y_{k_0}\|+\frac{2\rho\log k}{\beta}\right)^2 && \text{(see (69))}\\
&\le \frac{1}{|K|}\sum_{k=k_0}^{k_1}\frac{2\|y_{k_0}\|^2}{k+1}+\frac{8\rho^2\log^2 k}{\beta^2(k+1)} && ((a+b)^2\le 2a^2+2b^2)\\
&\le \frac{4\|y_{k_0}\|^2\log(k_1+1)}{|K|}+\frac{16\rho^2\log^3(k_1+1)}{|K|}\\
&\le \frac{16\log^3(k_1+1)}{|K|}\left(\|y_{k_0}\|^2+\rho^2\right). && (70)
\end{aligned}$$

In order to interpret (60,65,67), it still remains to estimate μ' in (62). To that end, we first derive a lower bound on the step sizes $\{\gamma_k\}_k$. To invoke (17), we in turn need to gauge how smooth the augmented Lagrangian $\mathcal{L}_{\beta_k}(\cdot,y_k)$ is. For every $k\in[k_0:k_1+1]$, note that

$$\begin{aligned}
\lambda_{\beta_k} &\le \lambda_h+\sqrt{m}\lambda_A\left(\|y_k\|+\frac{\|A(u_k)-b\|}{\beta_k}\right)+\frac{\|DA(u_k)\|_F^2}{\beta_k} && \text{(see (16))}\\
&\le \lambda_h+\sqrt{m}\lambda_A\left(\|y_{k_0}\|+\frac{2\rho\log(k_1+1)}{\beta_k}+\frac{\rho}{\beta_k}\right)+\frac{m\eta_{\max}^2}{\beta_k} && \text{(see (52,68,69))}\\
&= \lambda_h+\sqrt{m}\lambda_A\|y_{k_0}\|+\frac{1}{\beta_k}(2\sqrt{m}\lambda_A\rho\log(k_1+1)+\sqrt{m}\lambda_A\rho+m\eta_{\max}^2)\\
&\le \frac{2}{\beta_k}(2\sqrt{m}\lambda_A\rho\log(k_1+1)+\sqrt{m}\lambda_A\rho+m\eta_{\max}^2), && (71)
\end{aligned}$$

where the last line holds if k_0 is sufficiently large. We are now in position to invoke (17) by writing that

$$\begin{aligned}
\gamma_k &\ge \frac{\theta}{\lambda_{\beta_k}} && \text{(see (17))}\\
&\ge \frac{\theta\beta_k}{4\sqrt{m}\lambda_A\rho\log(k_1+1)+2\sqrt{m}\lambda_A\rho+2m\eta_{\max}^2} && \text{(see (71))}\\
&= \frac{\theta\beta}{(4\sqrt{m}\lambda_A\rho\log(k_1+1)+2\sqrt{m}\lambda_A\rho+2m\eta_{\max}^2)\sqrt{k}}, && \text{(see (45))} && (72)
\end{aligned}$$

for every $k\in[k_0:k_1+1]$. In particular, the lower bound on γ_{k_1+1} above allows us to estimate μ' by writing that

$$\begin{aligned}
\mu' &= \frac{24\beta\log(k_1+1)\|G_{k_1+1}\|^2}{\eta_{\min}^2(k_1+1)}+\mu && \text{(see (62))}\\
&= \frac{24\beta\log(k_1+1)\|u_{k_1+2}-u_{k_1+1}\|^2}{\eta_{\min}^2(k_1+1)\gamma_{k_1+1}^2}+\mu && \text{(see (40))}\\
&\le \frac{24\beta\log(k_1+1)\mathrm{diam}(C)^2}{\eta_{\min}^2(k_1+1)\gamma_{k_1+1}^2}+\mu && (\{u_k\}\subset C)\\
&\le \frac{24\beta\log(k_1+1)\mathrm{diam}(C)^2(4\sqrt{m}\lambda_A\rho\log(k_1+1)+2\sqrt{m}\lambda_A\rho+2m\eta_{\max}^2)^2}{\eta_{\min}^2\theta^2\beta^2}+2\mu && \text{(see (72))}\\
&=: \mu''. && (73)
\end{aligned}$$

Having estimated B_K and μ', we can rewrite the bound on the feasibility gap as

$$\begin{aligned}
\min_{k-1\in[K]}\|A(u_k)-b\|^2 &\le \frac{48\beta^2\log^2(k_1+1)}{\eta_{\min}^2|K|}\left(\lambda_h'^2+\eta_{\max}^2|K|B_K\right)+\frac{\beta\mu'}{|K|} && \text{(see (65))}\\
&\le \frac{48\beta^2\log^2(k_1+1)}{\eta_{\min}^2|K|}\left(\lambda_h'^2+16\eta_{\max}^2(\|y_{k_0}\|^2+\rho^2)\log^3(k_1+1)\right)+\frac{\beta\mu''}{|K|} && \text{(see (70,73))} && (74)
\end{aligned}$$

Moreover, we can simplify the assumption in (67). To be specific, thanks to (70,73), we can replace (67) with the assumption

$$\|A(u_{k_0})-b\|\le\rho,\qquad \frac{48\beta^2\log^2(k_1+1)}{\eta_{\min}^2}\left(\lambda_h'^2+16\eta_{\max}^2(\|y_{k_0}\|^2+\rho^2)\log^3(k_1+1)\right)+\beta\mu''\le\rho^2. \tag{75}$$

The lower bound on the step sizes in (72) has two other consequences. First, we find that (59) automatically holds if k_0 is sufficiently large. Second, it allows us to improve (61) by writing that

$$\begin{aligned}
&\min_{k\in K}\|G_k\|^2\\
&\le\frac{\min_{k\in K}\gamma_k\|G_k\|^2}{\min_{k\in K}\gamma_k}\\
&\le\frac{1}{\min_{k\in K}\gamma_k}\left(\frac{192\beta\log^2(k_1+1)}{\eta_{\min}^2|K|}\left(\lambda_h'^2+\eta_{\max}^2|K|B_K\right)+\frac{4\mu'}{|K|}\right)\qquad\text{(see (61))}\\
&\le\frac{1}{\min_{k\in K}\gamma_k}\left(\frac{192\beta\log^2(k_1+1)}{\eta_{\min}^2|K|}\left(\lambda_h'^2+16\eta_{\max}^2(\|y_{k_0}\|^2+\rho^2)\log^3(k_1+1)\right)+\frac{4\mu''}{|K|}\right)\qquad\text{(see (70,73))}\\
&\le\frac{(4\sqrt{m}\lambda_A\rho\log(k_1+1)+2\sqrt{m}\lambda_A\rho+2m\eta_{\max}^2)\sqrt{k}}{\theta\beta}\left(\frac{192\beta\log^2(k_1+1)}{\eta_{\min}^2|K|}\left(\lambda_h'^2+16\eta_{\max}^2(\|y_{k_0}\|^2+\rho^2)\log^3(k_1+1)\right)+\frac{4\mu''}{|K|}\right)\quad\text{(see (72))}\\
&\le\frac{4\sqrt{m}\lambda_A\rho\log(k_1+1)+2\sqrt{m}\lambda_A\rho+2m\eta_{\max}^2}{c\theta\beta\sqrt{k_1}}\left(\frac{192\beta\log^2(k_1+1)}{\eta_{\min}^2}\left(\lambda_h'^2+16\eta_{\max}^2(\|y_{k_0}\|^2+\rho^2)\log^3(k_1+1)\right)+4\mu''\right),
\end{aligned}\tag{76}$$

where the last line holds if there exists $c>0$ for which $k_1-k_0+1\ge ck_1$.

Theorem 2 *Let γ_k be the output of the line search subroutine in our algorithm in iteration k. For integers $k_0\le k_1$, consider the interval $K=[k_0:k_1]$ and suppose that*

$$\beta_k=\frac{\beta}{\sqrt{k}},\qquad\sigma_k=\beta k,\qquad\forall k\in K. \tag{77}$$

We impose the following geometric requirements on the constraints. Let $P_{T_C(u)}$ and $P_{N_C(u)}$ denote the orthogonal projection onto the tangent and normal cones at $u\in C$, respectively. Consider a subspace $S_K\subseteq\mathbb{R}^d$ such that

$$S_K\supseteq\bigcup_{k\in K}T_C(u_k), \tag{78}$$

and, with some abuse of notation, let S_K also denote an orthonormal basis for this subspace. For $\eta_{\min}$, we assume that the nonconvex Slater's condition holds, namely, that there exists $\rho>0$ such that

$$0<\eta_{\min}:=\begin{cases}\min_u\left\|S_{k_0}^\top P_{T_C(u)}(DA(u)^\top v)\right\|\\ \|v\|=1\\ \|A(u)-b\|\le\rho\\ u\in C.\end{cases} \tag{79}$$

Suppose that

$$\|A(u_{k_0})-b\|\le\rho,\qquad\operatorname{diag}(C)\le\frac{2\eta_{\min}}{\lambda_A},\qquad\rho\ge\rho_{\text{low}}(C,A,\beta)\log^{\frac{5}{2}}k_1, \tag{80}$$

where $\rho_{\text{low}}(C,A,\beta)$ depends only on C,A,β and is specified in the proof, see (75). Then it holds that

$$\min_{k\in K}\|G_{\beta_i,\gamma_i}(u_i,y_i)\|^2=\frac{O(\log^5 k_1)}{\sqrt{k_1}},\qquad\forall k\in K, \tag{81}$$

$$\min_{k\in K}\|A(u_k)-b\|=\frac{O(\log^5 k_1)}{\sqrt{k_1}},\qquad\forall k\in K, \tag{82}$$

provided that $k_0=\Omega(k_1)$ is sufficiently large and

$$\inf_k h(u_k)+\langle A(u_k)-b,y_{k_0}\rangle>-\infty. \tag{83}$$

Example: Max-cut Consider the factorized max-cut program, namely,

$$\begin{cases} \min \langle UU^\top, H \rangle \\ \operatorname{diag}(UU^\top) = 1, \end{cases} \tag{84}$$

where $U \in \mathbb{R}^{d' \times r}$. For every i, let $u_i \in \mathbb{R}^r$ denote the ith row of U. Let us form $u \in \mathbb{R}^d$ with $d = d'r$ by vectorizing U, namely,

$$u = [u_1^\top \cdots u_{d'}^\top]^\top. \tag{85}$$

We can therefore cast the above program as Program (1) with

$$h(u) = \sum_{i,j} H_{i,j} \langle u_i, u_j \rangle, \tag{86}$$

$$A : u \to [\|u_1\|^2 \cdots \|u_{d'}\|^2]^\top. \tag{87}$$

It is easy to verify that

$$DA(u) = \begin{bmatrix} u_1^\top & \cdots & 0 \\ \vdots & & \\ 0 & \cdots & u_{d'}^\top \end{bmatrix} \in \mathbb{R}^{d' \times d}. \tag{88}$$

In particular, if we take $S_{k_0} = \mathbb{R}^d$ and $\rho < 1$, we have $P_{T_C(u)} = I_d$ and thus

$$\begin{aligned} \eta_{\min} &= \begin{cases} \min_u \eta_{\min}\left(DA(u)\right) \\ \|A(u) - 1\| \le \rho \end{cases} \\ &= \begin{cases} \min_u \min_i \|u_i\|^2 \\ |\|u_i\|^2 - 1| \le \rho \qquad \forall i \end{cases} \\ &\ge 1 - \rho > 0. \end{aligned} \tag{89}$$

Above, $\eta_{\min}(DA(u))$ returns the smallest singular value of $DA(u)$. Consequently, the nonconvex Slater's condition holds for the max-cut problem.

Example: Clustering Consider the factorized clustering problem, namely,

$$\begin{cases} \min \langle UU^\top, H \rangle \\ UU^\top 1 = 1 \\ \|U\|_F \le \sqrt{k} \\ U \ge 0, \end{cases} \tag{90}$$

where $U \in \mathbb{R}^{d' \times r}$ and k is the number of clusters. We form $u \in \mathbb{R}^d$ as before. Note that the above program can be cast as Program (1) with the same h as before and

$$A : u \to \left[\, u_1^\top \textstyle\sum_j u_j \, \cdots u_{d'}^\top \sum_j u_j \,\right]^\top \in \mathbb{R}^{d'}, \tag{91}$$

and also $C = \sqrt{k} B_{2+}$, where $B_{2+} \subset \mathbb{R}^{d'}$ is the intersection of the unit ℓ_2-ball with the positive orthant. Note that

$$DA(u) = \begin{bmatrix} w_{1,1} u_1^\top & \cdots & w_{1,d'} u_1^\top \\ \vdots & & \\ w_{d',1} u_{d'}^\top & \cdots & w_{d',d'} u_{d'}^\top \end{bmatrix}, \tag{92}$$

where $w_{i.i} = 2$ and $w_{i,j} = 1$ for $i \neq j$. Let us start with the case where $u \in \partial C$ belongs to the boundary of C, namely, $\|u\| = \sqrt{k}$. We also assume that $u > 0$. Under these assumptions, note that

$$T_C(u) = \{z \in \mathbb{R}^{d'} : \langle u, z \rangle = 0\}, \tag{93}$$

and, consequently,

$$P_{T_C(u)} = I_d - \frac{uu^\top}{\|u\|^2} = I_d - \frac{uu^\top}{k}. \tag{94}$$

For simplicity, let us assume that $\rho = 0$, namely, u is a feasible point of Program (1). Then we find that

$$\begin{aligned}
\eta_{\min}(P_{T_C(u)} DA(u)^\top) &= \eta_{\min}((I - \frac{uu^\top}{k})DA(u)^\top) \\
&\geq \eta_{\min}(DA(u)) - \frac{1}{k}\|uu^\top DA(u)^\top\| \qquad \text{(Weyl's inequality)} \\
&= \eta_{\min}(DA(u)) - \frac{1}{\sqrt{k}}\|DA(u)u\| \qquad \left(\|u\| = \frac{1}{\sqrt{k}}\right).
\end{aligned} \tag{95}$$

We evaluate each term in the last line above separately. By its definition in (92), first note that

$$\begin{aligned}
\eta_{\min}(DA(u)) &\geq \eta_{\min}\left(\begin{bmatrix} u_1 & \cdots & u_{d'} \\ \vdots & & \\ u_1 & \cdots & u_{d'} \end{bmatrix}\right) - \max_i \|u_i\| \qquad \text{(Weyl's inequality)} \\
&= \sqrt{d'}\,\eta_{\min}(U) - \max_i \|u_i\| \\
&\geq \sqrt{d'}\eta_{\min}(U) - 1,
\end{aligned} \tag{96}$$

where in the last line follows from the assumptions that $u_i^\top \sum_j u_j = 1$ for all i and that $u > 0$ to see that $\max_i \|u_i\| \leq 1$. Note also that

$$\begin{aligned}
\|DA(u)u\| &\leq \left\| \begin{bmatrix} u_1^\top & \cdots & u_1^\top \\ \vdots & & \\ u_{d'}^\top & \cdots & u_{d'}^\top \end{bmatrix} u \right\| + \sqrt{\sum_{i=1}^{d'} \|u_i\|^4} \\
&= \|1_{d'}\| + \max_i \|u_i\| \cdot \|u\| \\
&\leq \sqrt{d'} + \sqrt{k},
\end{aligned} \tag{97}$$

where the last line follows because $\max_i \|u_i\| \leq 1$ and $\|u\| = \sqrt{k}$ by assumption. Consequently, we reach

$$\eta_{\min}(P_{T_C(u)} DA(u)^\top) \geq \sqrt{d'}\left(\eta_{\min}(U) - \frac{1}{\sqrt{k}}\right) - 2. \qquad \text{(see (95))} \tag{98}$$

By continuity, we extend the results to the case where u might have zero entries. Lastly, in the case where $u \in \mathrm{int}(C)$ (namely, $\|u\| < \sqrt{k}$ and $u > 0$), we have that $T_C(u) = \mathbb{R}^d$ and it directly follows from (96) that

$$\begin{aligned}
\eta_{\min}(P_{T_C(u)} DA(u)) &= \eta_{\min}(DA(u)) \\
&\geq \sqrt{d'}\eta_{\min}(U) - 1. \qquad \text{(see (96))}
\end{aligned} \tag{99}$$

Having studied all cases for u for $\rho = 0$, we conclude that

$$\eta_{\min} \geq \sqrt{d'}\left(\eta_{\min}(U) - \frac{1}{\sqrt{k}}\right) - 2. \qquad \text{(see (79))} \tag{100}$$

Roughly speaking, as long as $\eta_{\min}(U) \gtrsim 1/\sqrt{k}$, the right-hand side is greater than zero and the nonconvex Slater's condition holds. We assumed for simplicity that $\rho = 0$ but this is expected to hold for ρ sufficiently small as well by continuity.

New Slater's condition Here we describe a variant of the Slater's condition for Program (1).

Definition 2 (Nonconvex Slater's condition) Let $\theta_{\min}$ be the smallest angle between to subspace and define ψ to be

$$\begin{aligned}
\psi_{A,C} &:= \begin{cases} \inf_u \sin\left(\theta_{\min}(\mathrm{null}(A), T_C(u))\right) \\ Au \neq 0 \\ u \in \partial C \end{cases} \\
&= \begin{cases} \inf_u \eta_{\min}\left(P_{T_C(u)} A^\top\right) \\ Au \neq 0 \\ u \in \partial C \end{cases}
\end{aligned} \tag{101}$$

where ∂C is the boundary of C, and $\eta_{\min}$ returns the smallest singular value. We say that Program (1) satisfies the Slater's condition if $\psi_{A,C} > 0$.

As a sanity check, we have the following result.

Proposition 1 *The nonconvex Slater's condition for Program (1) implies the standard Slater's condition when A is a linear operator and Program (1) is feasible.*

Proof Suppose that the standard Slater's condition does not hold, namely, that

$$\operatorname{relint}(\operatorname{null}(A) \cap C) = \operatorname{null}(A) \cap \operatorname{relint}(C) = \emptyset. \tag{102}$$

Since Program (1) is feasible, there exists a feasible u, namely, $Au = 0$ and $u \in C$. By (102), it must be that $u \in \partial C$ and that $\operatorname{null}(A)$ supports C at u **Why?**. In particular, it follows that $\operatorname{null}(A) \cap T_C(u) \neq \{0\}$ or, equivalently, $\operatorname{row}(A) \cap N_C(u) \neq \{0\}$. That is, there exists a unit-norm vector v such that

$$P_{T_C(u)} A^\top v = 0, \tag{103}$$

and consequently

$$\eta_{\min}(P_{T_C(u)} A^\top) = 0. \tag{104}$$

Because $\eta_{\min}(P_{T_C(u)} A^\top)$ is a continuous function of u **(Why?)**, we conclude that $\psi_{A,C} = 0$, namely, the nonconvex Slater's condition also does not hold, thereby completing the proof of Proposition 1.

D Proof of Lemma 3

If $A(u_k) = b$, then (57) holds trivially. Otherwise, for an integer k_0, consider a subspace

$$S_K \supseteq \bigcup_{k \in K} T_C(u_k), \tag{105}$$

and let S_K with orthonormal columns denote a basis for this subspace, with some abuse of notation. We then assume that

$$0 < \eta_{\min} := \begin{cases} \min_u \ \left\| S_K^\top P_{T_C(u)} (DA(u)^\top v) \right\| \\ \|v\| = 1 \\ \|A(u) - b\| \leq \rho \\ u \in C. \end{cases} \tag{106}$$

If $\max_{k \in K} \|A(u_k) - b\| \leq \rho$, then (106) is in force and, for every $k \geq k_0$, we may write that

$$\begin{aligned}
&\left\| P_{T_C(u_{k+1})} (DA(u_k)^\top (A(u_k) - b)) \right\| \\
&\geq \left\| P_{T_C(u_{k+1})} (DA(u_{k+1})^\top (A(u_k) - b)) \right\| - \left\| (DA(u_{k+1}) - DA(u_k))^\top (A(u_k) - b) \right\| \qquad \text{(non-expansiveness of projection)} \\
&\geq \eta_{\min} \|A(u_k) - b\| - \|DA(u_{k+1}) - DA(u_k)\| \, \|A(u_k) - b\| \qquad \text{(see (106))} \\
&\geq \eta_{\min} \|A(u_k) - b\| - \lambda_A \|u_{k+1} - u_k\| \cdot \|A(u_k) - b\| \qquad \text{(see (2))} \\
&= (\eta_{\min} - \lambda_A \gamma_k \|G_k\|) \, \|A(u_k) - b\| \qquad \text{(see (40))} \\
&\geq \frac{\eta_{\min}}{2} \|A(u_k) - b\|,
\end{aligned} \tag{107}$$

where the last line above uses the observation that

$$\begin{aligned}
\lambda_A \gamma_k \|G_k\| &= \lambda_A \|u_{k+1} - u_k\| \\
&\leq \lambda_A \operatorname{diam}(C) \\
&\leq \frac{\eta_{\min}}{2}. \qquad \text{(see (56))}
\end{aligned} \tag{108}$$

We can now lower bound (51) by using (107), namely,

$$\begin{aligned}
\frac{\eta_{\min}}{2\beta_k} \|A(u_k) - b\| &\leq \frac{1}{\beta_k} \|P_{T_C(u_{k+1})} (DA(u_k)^\top (A(u_k) - b))\| \qquad \text{(see (107))} \\
&\leq \lambda_h' + \eta_{\max} \|y_{k-1}\| + \|G_k\|. \qquad \text{(see (51))}
\end{aligned} \tag{109}$$

which completes the proof of Lemma 3.