A Quasi-Newton Primal-Dual Algorithm with Line Search*

Shida Wang[1][0000−0001−7521−8192], Jalal Fadili[2][0000−0002−8165−7578], and Peter Ochs[1][0000−0002−4880−7511]

[1] Department of Mathematics, University of Tübingen, Germany
[2] Normandie Univ, ENSICAEN, CNRS, GREYC, France

Abstract. Quasi-Newton methods refer to a class of algorithms at the interface between first and second order methods. They aim to progress as substantially as second order methods per iteration, while maintaining the computational complexity of first order methods. The approximation of second order information by first order derivatives can be expressed as adopting a variable metric, which for (limited memory) quasi-Newton methods is of type "identity ± low rank". This paper continues the effort to make these powerful methods available for non-smooth systems occurring, for example, in large scale Machine Learning applications by exploiting this special structure. We develop a line search variant of a recently introduced quasi-Newton primal-dual algorithm, which adds significant flexibility, admits larger steps per iteration, and circumvents the complicated precalculation of a certain operator norm. We prove convergence, including convergence rates, for our proposed method and outperform related algorithms in a large scale image deblurring application.

Keywords: quasi-Newton · primal-dual algorithm · line search · saddle-point problems · large scale optimization

1 Introduction

In modern optimization, the datasets and dimensionality of the problems and parameters is vastly increasing. In the early stages of large scale optimization, the shift from second order to first order optimization could cope with the increasing dimensionality of the problems. Second order methods achieve a significant progress per iteration at the cost of a high computational load, since the computation of the second derivative (Hessian) and oftentimes its inverse are required, which is intractable in the large scale regime. In contrast, first order methods have a low computational effort per iteration but also less information about the objective. The gradient cannot capture curvature information and hence may fail to provide directions that allow for large steps. Nevertheless, for the large scale regime this exchange pays off.

* We acknowledge funding by the ANR-DFG joint project TRINOM-DS under the number DFG OC150/5-1.

However, the ever increasing dimensionality of the considered problems and datasets asks for faster algorithms. Motivated by classical optimization and the discussion above, algorithms at the interface of first and second order methods are key to reach the next level. Tractability in the (nowadays extremely) large scale regime requires methods that are as cheap as first order methods, while progressing as substantially as second order algorithms: *Quasi-Newton methods*. While they are known for their outstanding performance in unconstrained smooth optimization, their development for non-smooth (or constrained smooth) problems is insufficiently understood. As we discuss in related work below, most algorithmic development is either too simplistic, in the sense that only a diagonal metric is admitted, which can hardly capture second order information of the objective, or too theoretical, in the sense that a good performance is proved in theory while the implementation cost is on a par with that of second order methods. Algorithmic subproblems (e.g. the evaluation of the proximal mapping) that are easy to solve with respect to the Euclidean metric may become intractable with respect to another metric.

We pursue the line of research initiated in [2,3] that considers both aspects as equally important. Key is the observation that quasi-Newton methods actually generate a metric of the specific type "identity $\pm$ low rank", which allows for an efficient proximal calculus (cf. Section 6) that unlocks the quasi-Newton power—well-known from classical optimization—in the area of optimization for Machine Learning. This idea was recently transferred to non-convex optimization in [15,16] and to monotone inclusion problems in [28]. A special case of the latter setting comprises the extremely broad class of convex–concave saddle point problems, which has numerous applications in Machine Learning, Computer Vision, Image Processing and Statistics [8,7,14,26,1].

In this paper, we restrict our interest to saddle-point problems only. This focus allows us to design a quasi-Newton primal–dual algorithm that is tailored to this setting and therefore highly efficient and adaptable thanks to an additional line search procedure. This has several advantages as compared to a fixed step size: (i) the oftentimes expensive computation of the operator norm can be avoided, (ii) the choice of metric need not obey any static spectral restrictions, and (iii) in many situations larger steps and thus a faster convergence is observed.

Our main contribution is reduction of the gap between the outstanding performance of quasi-Newton methods in classical optimization and quasi-Newton methods for (non-smooth) convex–concave saddle point problems for modern optimization in Machine Learning. In detail, our contribution is the following:

1. We extend the line-search based primal–dual algorithm in [20] (extension of [6] by line search) to incorporate a quasi-Newton metric with efficiently implementable proximal mapping; thereby aiming equally at theoretical convergence guarantees (including convergence rates) as well as highly efficient implementation.

2. We unlock the use of multi-memory quasi-Newton metrics (L-BFGS and SR1 method) via a compact representation for primal–dual algorithms, including their efficient implementation via a semi-smooth Newton solver.
3. The proposed algorithm outperforms the line-search based primal–dual algorithm (with identity metric) on a challenging image deblurring problem.

1.1 Related Work

Due to page limitations, for an extended discussion of quasi-Newton approaches in non-smooth optimization and the vast literature on primal-dual algorithms, we refer to [28].

Non-smooth quasi-Newton. For a class of non-smooth problems that are given as a composition of a smooth function h and a non-smooth function g, [21,25] combine quasi-Newton methods with forward–backward splitting via the forward–backward envelope. If g is an indicator function, [23,24] proposed a projected quasi-Newton method which requires either solving a complicated subproblem or is restricted to a diagonal metric. Later, their work was extended by [18] to a more general setting. [17,2,3] developed algorithms with efficient evaluation of the proximal operator with respect to a low-rank perturbed metric. It is worth to mention that in [2,3] the subproblem is a low dimensional root finding problem which can be solved efficiently. Inspired by [3], the work in [16] applied the limited-memory quasi-Newton method on non-convex problems.

PDHG. Primal-Dual Hybrid Gradient (PDHG) is widely used to solve saddle point problems [6,8]. However, in order to guarantee the convergence of PDHG, the computation of the norm of a operator K is required. To avoid this disadvantage, [20] combined line search with PDHG to get a new algorithm PDAL. For faster convergence, variable metric is being used [14]. It shows the potential of combining quasi-Newton methods and PDAL via a variable metric, however suffers again from the need to solve more complicated subproblems, which is remedied in [28] for the more general class of monotone inclusion problems and hence builds the grounding for our proposed line search variant.

2 Problem Setup: A class of saddle point problems

Let X and Y be finite dimensional real vector spaces with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$. We consider saddle point problems

$$\min_{x \in X} \max_{y \in Y} \langle Kx, y \rangle + g(x) + h(x) - f^*(y)\,, \tag{1}$$

where $g, h \colon X \to \overline{\mathbb{R}}$ are proper, lower semi-continuous (lsc), convex functions with h, additionally, having an L-Lipschitz continuous gradient, $f^* \colon Y \to \overline{\mathbb{R}}$ is a proper, lsc, convex function; the convex conjugate (Legendre–Fenchel conjugate) of a function f, and $K \colon X \to Y$ is a bounded linear operator with operator norm $L_K := \|K\|$ and adjoint K^*. Moreover, throughout this paper we assume that

(1) has a saddle point. We remark that (1) is equivalent to the *primal problem*

$$\min_{x\in X} f(Kx) + g(x) + h(x)\,, \tag{2}$$

and to the *dual problem*

$$\max_{y\in Y} -\Big(f^*(y) + (g^* \,\triangledown\, h^*)(-K^*y)\Big)\,, \tag{3}$$

where we use the fact that the conjugate of a sum of two function $(g+h)^*$ equals the infimal convolution of the conjugate functions $g^* \,\triangledown\, h^*$.

3 Our quasi-Newton Primal-Dual Algorithm with Line Search

The primal–dual algorithm that we develop in this paper is an extension of [20] to incorporate a variable metric of quasi-Newton type, which itself adds an efficient line search procedure to the primal–dual hybrid gradient (PDHG) algorithm [6] (aka Chambolle–Pock algorithm) and the extension in [19]. The handling of (a possibly) non-smooth functions essentially relies on evaluating the proximal mapping, which we define here for a function g and parameter τ with respect to an arbitrary metric $M \in \mathcal{S}_\alpha(X)$, $\alpha > 0$. Here, $\mathcal{S}_\alpha$ is the set of bounded self-adjoint linear operators from Hilbert space X to X such that $M - \alpha I$ is positive semi-definite for each $M \in \mathcal{S}_\alpha$. For simplicity, some notations are introduced:

$$\|x\|_M^2 = \langle Mx, x\rangle\,, \quad x \in X\,.$$

$$\operatorname{prox}_{\tau g}^M(\bar{x}) := \operatorname*{argmin}_{x\in X} g(x) + \frac{1}{2\tau}\|x-\bar{x}\|_M^2\,, \quad \text{and set } \operatorname{prox}_{\tau g} := \operatorname{prox}_{\tau g}^I\,,$$

where I is the identity (Euclidean) metric.

Algorithm 1 presents the proposed algorithm. The algorithm alternates between updates of the dual variable (4) and the primal variable (5), where the line search is only implemented in the primal update. Let us discuss the algorithm for a fixed iteration k, i.e., we are given x^k, y^{k-1}, σ_{k-1}, θ_{k-1} and a monotone decreasing sequence of β_k. The discussion of the quasi-Newton type variable metric in Step (i) is deferred to Section 5. Step (ii) is a standard dual update step. In Step (iii), we perform the line search. We select a basic step size $\bar{\sigma}_k$ and, in the ith loop of the line search, we perform a trial step (5) with the current $\sigma_k = \bar{\sigma}_k \cdot \mu^i$ and check if the breaking condition (6) is satisfied. If 'yes', the current iteration k is completed. If 'no', the new trial step size is reduced to $\sigma_k = \bar{\sigma}_k \cdot \mu^{i+1}$ (in the subsequent $(i+1)$th line search step). Here, $M_k \in \mathcal{S}_\alpha(X)$ is symmetric positive definite and $\alpha \in (0,1)$.If M_k is chosen as an identity, then we recover the breaking (stopping) criterion used in [20]. However, in this paper, we adopt a variable metric M_k which is generated by quasi-Newton methods to exploit the local geometry of the function h. As a result, it is more likely to obtain a larger step size σ_k and fewer inner loops for the line search procedure as compared to

Algorithm 1 Quasi-Newton PDHG with Line Search

Require: [initial data] $x^1 \in X$, [initial data] $y^0 \in Y$, [maximal iteration count] $N \geq 0$, [scaling of line search parameter] $\mu \in (0,1)$, [extrapolation parameter] θ_0, [initial dual step size] σ_0, [tolerance weight] $\delta \in (0,1)$, [primal-dual step ratio] $+\infty > \beta \geq \beta_{k+1} \geq \beta_k > 0$, $\forall k \in \mathbb{N}$.

Update for $k = 1, \ldots, N$:

(i) Compute M_k according to a quasi-Newton framework (cf. Section 5).

(ii) Compute dual update step:

$$y^k = \operatorname{prox}_{\sigma_{k-1} f^*}(y^{k-1} + \sigma_{k-1} K x^k)\,. \tag{4}$$

(iii) Select $\bar{\sigma}_k \in [\frac{\beta_{k-1}}{\beta_k}\sigma_{k-1}, \sqrt{(1+\theta_{k-1})}\frac{\beta_{k-1}}{\beta_k}\sigma_{k-1}]$ and compute the quasi-Newton primal update step by:

Line search: Find the smallest power $i = 0, 1, 2, \ldots$ such that

$$\begin{aligned} \bar{y}^k &= y^k + \theta_k (y^k - y^{k-1})\,, \\ x^{k+1} &= \operatorname{prox}^{M_k}_{\tau_k g}\big(x^k - \tau_k M_k^{-1} K^* \bar{y}^k - \tau_k M_k^{-1} \nabla h(x^k)\big) \end{aligned} \tag{5}$$

with

$$\sigma_k = \bar{\sigma}_k \cdot \mu^i\,, \quad \theta_k = \frac{\sigma_k}{\sigma_{k-1}}\,, \quad \text{and} \quad \tau_k = \beta_k \sigma_k$$

satisfy

$$\begin{aligned} \tau_k \sigma_k \|K x^{k+1} - K x^k\|^2 + 2\tau_k \Big(h(x^{k+1}) - h(x^k) - \Big\langle \nabla h(x^k), x^{k+1} - x^k \Big\rangle\Big) \\ \leq \delta \|x^{k+1} - x^k\|^2_{M_k}\,. \end{aligned} \tag{6}$$

End of for-loop

the Euclidean version ($M_k = I$). The employed metric is of type "identity $\pm$ low rank" for which the proximal mapping can be computed efficiently as shown in Section 6.

Remark 1. While the line search procedure is formulated for the primal problem, by duality, the primal problem can be interpreted as the dual of the dual problem and, thus, the dual problem as the primal problem. As a consequence, an equivalent algorithm with line search on the dual can be easily stated.

Discussion of computational cost for line search. In general, every loop of the line search procedure requires recomputing several quantities, including (5) and Kx^{k+1}, $h(x^{k+1})$ and $\|x^{k+1} - x^k\|^2_{M_k}$ in (6). While this seems to be expensive at first glance, often (6) is satisfied after 1–3 trial steps and hence large steps are taken with a low cost, as we underline in our experiments in Section 7. Nevertheless, the cost can be further reduced significantly in certain special cases, observed in [20] and generalized here to our setting, whenever $\operatorname{prox}^{M_k}_{\tau_k g}$ is a linear or affine operator.

1. If $g(x) = \langle c, x\rangle$, then $\operatorname{prox}^{M_k}_{\tau_k g}(u) = u - \tau_k M_k^{-1} c$ and therefore, we obtain

$$\begin{aligned} x^{k+1} &= \operatorname{prox}^{M_k}_{\tau_k g}(x^k - \tau_k M_k^{-1} K^* \bar{y}^k - \tau_k M_k^{-1} \nabla h(x^k)) \\ &= x^k - \tau_k [M_k^{-1} K^* \bar{y}^k + M_k^{-1} \nabla h(x^k) + M_k^{-1} c]\,. \end{aligned}$$

$$Kx^{k+1} = Kx^k - \tau_k [K M_k^{-1} K^* \bar{y}^k + \tau_k K M_k^{-1} \nabla h(x^k) + \tau_k K M_k^{-1} c]\,.$$

2. If $g(x) = \frac{1}{2}\|x - b\|^2$, then $\operatorname{prox}^{M_k}_{\tau_k g}(u) = (I + \tau_k M_k^{-1})^{-1}[u + \tau_k M_k^{-1} b]$ and therefore, we obtain

$$\begin{aligned} x^{k+1} &= \operatorname{prox}^{M_k}_{\tau_k g}(x^k - \tau_k M_k^{-1} K^* \bar{y}^k - \tau_k M_k^{-1} \nabla h(x^k)) \\ &= (I + \tau_k M_k^{-1})^{-1}[x^k - \tau_k M_k^{-1} K^* \bar{y}^k - \tau_k M_k^{-1} \nabla h(x^k) + \tau_k M_k^{-1} b]\,, \end{aligned}$$

$$Kx^{k+1} = K(I + \tau_k M_k^{-1})^{-1}[x^k - \tau_k (M_k^{-1} K^* \bar{y}^k + M_k^{-1} \nabla h(x^k) - M_k^{-1} b)]\,.$$

3. Let $g(x) = \delta_H(x)$, where H refers to the hyperplane $H := \{u : \langle u, a\rangle = b\}$. Then $\operatorname{prox}^{M_k}_{\tau_k g}(u) = u + \frac{b - \langle u, a\rangle}{\|a\|^2_{M_k^{-1}}} M_k^{-1} a$ and therefore, we obtain

$$\begin{aligned} x^{k+1} &= \operatorname{prox}^{M_k}_{\tau_k g}(x^k - \tau_k M_k^{-1} K^* \bar{y}^k - \tau_k M_k^{-1} \nabla h(x^k)) \\ &= x^k - \tau_k [M_k^{-1} K^* \bar{y}^k + M_k^{-1} \nabla h(x^k)] \\ &\quad + \frac{b - \left\langle x^k - \tau_k [M_k^{-1} K^* \bar{y}^k + M_k^{-1} \nabla h(x^k)], a\right\rangle}{\|a\|^2_{M_k^{-1}}} M_k^{-1} a\,, \end{aligned}$$

$$\begin{aligned} Kx^{k+1} &= Kx^k - \tau_k [K M_k^{-1} K^* \bar{y}^k + K M_k^{-1} \nabla h(x^k)] \\ &\quad + \frac{b - \left\langle x^k - \tau_k [M_k^{-1} K^* \bar{y}^k + \tau_k M_k^{-1} \nabla h(x^k)], a\right\rangle}{\|a\|^2_{M_k^{-1}}} K M_k^{-1} a\,. \end{aligned}$$

4 Convergence Analysis of Algorithm 1

Let us now analyze the convergence of Algorithm 1. As for most variable metric primal–dual algorithms (cf. [12,11,10,28]), we require the following restriction for the change of the metric from one iteration to the next. Under this condition, we can generalize all convergence results from [20] by adapting their proofs.

Assumption 1. *Let $\alpha \in (0, +\infty)$. $(M_k)_{k\in\mathbb{N}}$ is a sequence in $\mathcal{S}_\alpha(X)$ such that*

$$\begin{cases} \exists C_M \in \mathbb{R}, \text{s.t.} \sup_{k\in\mathbb{N}} ||M_k|| \le C_M < \infty\,, \\ (\exists (\eta_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N}))(\forall k \in \mathbb{N}): \quad (1+\eta_k) M_k \succeq M_{k+1}\,. \end{cases} \tag{7}$$

Lemma 1. *(i) There exists some $\sigma > 0$ such that $\sigma_k \ge \sigma$ for any $k \in \mathbb{N}$.*
(ii) The line search terminates.
(iii) If $\beta_k \equiv \beta$, θ_k is bounded from above by some θ for any $k \in \mathbb{N}$.

The proof is provided in Section B.1.

Theorem 1. *Consider Problem (1) and let the sequence $(x^k, y^k)_{k\in\mathbb{N}}$ be generated by Algorithm 1 with $\beta_k \equiv \beta$ where Assumption 1 holds. Then $(x^k, y^k)_{k\in\mathbb{N}}$ is a bounded sequence and its cluster points are solutions of (1). Furthermore, if $f^*|_{\mathrm{dom} f^*}$ is continuous and σ_k is bounded from above, then the whole sequence $(x^k, y^k)_{k\in\mathbb{N}}$ converges to a solution of (1).*

The proof is provided in Section B.2.

We obtain the same ergodic convergence rate as in [20], with respect to the primal–dual gap $\mathcal{G}_{\hat{x},\hat{y}}$ which is the difference (gap) between the optimal primal objective value in (2) and the optimal dual objective value (3).

Theorem 2. *Let the sequence $(x^k, y^k)_{k\in\mathbb{N}}$ be generated by Algorithm 1 with $\beta_k \equiv \beta$ where Assumption 1 holds and $(\hat{x}, \hat{y})$ be some saddle point of (1). Then it holds for a constant $D = \Pi_{k\in\mathbb{N}}(1+\eta_k) < +\infty$ that*

$$\mathcal{G}_{\hat{x},\hat{y}}(\bar{X}^N, \bar{Y}^N) \leq \frac{D}{s_N}\Big(\frac{1}{2\beta}\|x^1 - \hat{x}\|^2_{M_1} + \frac{1}{2}\|y^1 - \hat{y}\|^2 + \sigma_1\theta_1 D_{\hat{x},\hat{y}}(y^0)\Big) = O\Big(\frac{1}{N}\Big), \tag{8}$$

where $s_N := \sum_{k=1}^N \sigma_k$, $\bar{X}^N := \frac{\sum_{k=1}^N \sigma_k x^{k+1}}{s_N}$ and $\bar{Y}^N := \frac{\sigma_1\theta_1 y^0 + \sum_{k=1}^N \sigma_k \bar{y}^k}{\sigma_1\theta_1 + s_N}$. The last equality in (8) provides a simplified rate in terms of the big-O notation.

The proof is provided in Section B.3.

Under the additional assumption that g is strongly convex, improved convergence rates can be derived when $(\beta_k)_{k\in\mathbb{N}}$ is varied appropriately.

Theorem 3. *Assume g is γ-strongly convex and $(x^k, y^k)_{k\in\mathbb{N}}$ is generated by Algorithm 1 with*

$$\beta_k = \frac{\beta_{k-1}}{\min\{1 + \frac{\gamma}{C_M}\beta_{k-1}\sigma_{k-1}, C_\theta\}}, \ \forall k \in \mathbb{N}, \quad \text{and} \quad \beta_0 > 0, \tag{9}$$

where $C_\theta \in \mathbb{R}_+$ is a constant, Assumption 1 holds and $(\hat{x}, \hat{y})$ be some saddle point of (1). Then, we have $(\theta_k)_{k\in\mathbb{N}}$ is bounded from above. Furthermore, we obtain

$$\|x^N - \hat{x}\| = O(1/N) \quad \text{and} \quad \mathcal{G}_{\hat{x},\hat{y}}(\bar{X}^N, \bar{Y}^N) = O(1/N^2),$$

where $(\bar{X}^N, \bar{Y}^N)$ are the ergodic sequences defined in Theorem 2.

The proof is provided in Section B.4.

Remark 2. For the result in Theorem 3, $\delta = 1$ is also admitted.

5 Computing and Representing the quasi-Newton Metric

In this section, we abuse notation in order to follow the conventions of quasi-Newton methods[3]. The metric M_k is expected to be an approximation of the

[3] For example, the variable y^k defined in (12) is not the dual variable in Algorithm 1.

Hessian $\nabla^2 h(x^k)$ at x^k for the k-th iteration. The most popular quasi-Newton methods are the SR1 and BFGS methods (and their low-memory variants), which update M_k by adding a rank-one modification (SR1 method)

$$M_{k+1} := M_{k+1}^{SR1} = M_k + \frac{(y^k - M_k s^k)(y^k - M_k s^k)^\top}{(y^k - M_k s^k)^\top s^k}\,, \tag{10}$$

or a rank-two modification (BFGS method)

$$M_{k+1} := M_{k+1}^{BFGS} = M_k + \frac{y^k (y^k)^\top}{(s^k)^\top y^k} - \frac{M_k s^k (s^k)^\top M_k}{(s^k)^\top M_k s_k}\,, \tag{11}$$

respectively, where

$$s^k := x^{k+1} - x^k \quad \text{and} \quad y^k := \nabla h(x^{k+1}) - \nabla h(x^k)\,. \tag{12}$$

In order to apply quasi-Newton methods on large-scale problems, *m-limited memory quasi-Newton methods* are adopted [16], with the most popular version being L-BFGS [29]. It means that instead of generating M_k via all previous s^i and y^i for $i = 1, \ldots, k$ and M_0, for each k, the metric M_k is re-computed based on $M_{k,0}$ and the most recent m vectors s^i and y^i for $i = k-m+1, \ldots, k$, if $k \geq m$. Usually, m is very small, such that only a small storage will be required. As pointed out by [5], the matrices of the m-limited memory version of quasi-Newton methods have a compact representation of the form

$$M_k = M_{k,0} + A_k Q_k^{-1} A_k^\top\,, \tag{13}$$

where $M_{k,0} \in \mathbb{R}^{n\times n}$, $n = \dim(X)$, is a symmetric positive definite matrix, $A_k \in \mathbb{R}^{n\times m}$, and a symmetric and non-singular matrix $Q_k \in \mathbb{R}^{m\times m}$ ($m \ll n$). For limited memory BFGS (known as L-BFGS), we have the following block-matrix representation

$$A_k = \begin{bmatrix} M_{k,0} S_k & Y_k \end{bmatrix} \in \mathbb{R}^{n\times 2m} \quad \text{and} \quad Q_k = \begin{bmatrix} -S_k^* M_{k,0} S_k & -L_k \\ -L_k^* & D_k \end{bmatrix} \in \mathbb{R}^{2m\times 2m}\,, \tag{14}$$

where S_k and Y_k are matrices collecting the m most recent vectors in (12) as columns, $D_k := D(S_k^\top Y_k)$ and $L_k := L(S_k^\top Y_k)$ refer to the diagonal $D(\cdot)$ and the strict lower triangular $L(\cdot)$ part of the matrix $S_k^\top Y_k$, respectively. By using a spectral decomposition $Q^{-1} = V \Lambda V^\top$ with orthogonal $V \in \mathbb{R}^{s\times s}$ and diagonal $\Lambda \in \mathbb{R}^{s\times s}$, for some $s \in \mathbb{N}$, (13) is transformed into the compact representation

$$M_k = M_{k,0} + U_1 U_1^\top - U_2 U_2^\top\,, \tag{15}$$

for some $U_1 \in \mathbb{R}^{n\times m}$ and $U_2 \in \mathbb{R}^{n\times m}$. In detail, since Λ is a diagonal matrix with eigenvalues λ_i, $i = 1, 2, \ldots, s$ of Q_k^{-1} on the diagonal, we decompose $\Lambda = \Lambda_1 - \Lambda_2$ where Λ_1, given by $(\Lambda_1)_{i,i} = \max(\lambda_i, 0)$, $i = 1, 2, \ldots, s$, corresponds to the positive eigenvalues and Λ_2, given by $(\Lambda_2)_{i,i} = \max(-\lambda_i, 0)$, $i = 1, 2, \ldots, s$, corresponds to the negative eigenvalues. In this way, we obtain

$$U_1 := (A_k V)\Lambda_1^{1/2} \quad \text{and} \quad U_2 := (A_k V)\Lambda_2^{1/2}\,. \tag{16}$$

Theoretically, it is guaranteed that $M_k = M_{k,0} + U_1 U_1^\top - U_2 U_2^\top$ is positive definite [13] if $s^k y^k > 0$ for any $k \in \mathbb{N}$. However, in order to account for numerical rounding errors and the assumption that $M_k \in \mathcal{S}_\alpha$ is bounded from above by some C_M, we adopt a scaling version:

$$\begin{aligned} \tilde{M}_k &= M_{k,0} + \gamma_1 U_1 U_1^\top - \gamma_2 U_2 U_2^\top \,, \\ M_k &= \min\{\frac{C_M - \alpha}{\|\tilde{M}_k\|_2}, 1\}\tilde{M}_k + \alpha I \,, \end{aligned} \tag{17}$$

where $\|\tilde{M}_k\|_2$ denotes the l_2 norm of matrix $\tilde{M}_k$ and we set $\alpha = 0.01$,$\gamma_1 = 1$, $\gamma_2 = 1, C_M = 50$ in practice. There is an easy way to make sure that Assumption 1 is satisfied by setting $\gamma_1 = \frac{\eta_k}{\|U_1\|^2}$ and $\gamma_2 = \frac{\eta_k}{\|U_2\|^2}$ with arbitrary $\eta_k \in \ell_+^1$.

6 Proximal Calculus and Efficient Implementation

The transformation in Section 5 enables us to compute the proximal mapping with respect to the metric in the form of (15) via the proximal calculus developed in [3], which we state here for completeness.

Theorem 4. *Let $B = B_0 + U_1 U_1^\top - U_2 U_2^\top \in \mathcal{S}_\sigma(\mathbb{R}^n)$ with $\sigma > 0$, $B_0 \in \mathcal{S}_\sigma(\mathbb{R}^n)$ and $U_i \in \mathbb{R}^{n \times r_i}$ with rank r_i $(i = 1, 2)$. Set $B_1 = B_0 + U_1 U_1^\top$. Then, the following holds:*

$$\operatorname{prox}_g^B(x) = \operatorname{prox}_g^{B_0}(x + B_1^{-1} U_2 \alpha_2^* - B_0^{-1} U_1 \alpha_1^*) \,, \tag{18}$$

where $\alpha_i^*, i = 1, 2$, are the unique zeros of the coupled system $\mathcal{L}(\alpha) = \mathcal{L}(\alpha_1, \alpha_2) = 0$, where $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^{r_1 + r_2}$ and $\mathcal{L} = (\mathcal{L}_1, \mathcal{L}_2)$ is defined by

$$\begin{aligned} \mathcal{L}_1(\alpha_1, \alpha_2) &= U_1^\top (x + B_1^{-1} U_2 \alpha_2 - \operatorname{prox}_g^{B_0}(x + B_1^{-1} U_2 \alpha_2 - B_0^{-1} U_1 \alpha_1)) + \alpha_1 \,, \\ \mathcal{L}_2(\alpha_1, \alpha_2) &= U_2^\top (x - \operatorname{prox}_g^{B_0}(x + B_1^{-1} U_2 \alpha_2 - B_0^{-1} U_1 \alpha_1)) + \alpha_2 \,. \end{aligned} \tag{19}$$

Here, $\mathcal{L}\colon \mathbb{R}^{r_1 + r_2} \to \mathbb{R}^{r_1 + r_2}$ is Lipschitz continuous.

The computation of a possibly complicated proximal mapping $\operatorname{prox}_g^B(x)$ is reduced to a simple (by assumption) proximal mapping $\operatorname{prox}_g^{B_0}$ and a low dimensional root finding problem in (19), which we tackle by the semi-smooth Newton solver proposed in [3], formulated in Algorithm 2. It requires to solve Newton-like equations where the classic Jacobian at the current iterate α_k is replaced by the Clarke Jacobian $\partial^c \mathcal{L}(\alpha_k)$ (see [9]), where we account for inexact solutions of these subproblems in terms of an error e_k. For completeness, we also state the convergence result of [3] for Algorithm 2.

Theorem 5. *If g is in addition a tame function, then the Lipschitz continuous function $\mathcal{L}$ is semi-smooth [4] and all elements of $\partial^C \mathcal{L}(\alpha^*)$ are non-singular[3]. Therefore, if $\rho_k \leq \bar{\rho}$, $\forall k \in \mathbb{N}$, for some sufficiently small $\bar{\rho}$ and α_0 sufficiently close to α^*, then the sequence generated by the Algorithm 2 is well-defined and converges to α^* linearly. Additionally, if $\rho_k \to 0$, the convergence is superlinear.*

Algorithm 2 Semi-smooth Newton method to solve $\mathcal{L}(\alpha) = 0$ in (19)

Require: [initial data] $\alpha_0 \in \mathbb{R}^r$, [maximum iterations] N

Update for $k = 0, \cdots, N$:

(i) Select $G_k \in \partial^c \mathcal{L}(\alpha_k)$, compute α_{k+1} such that

$$\mathcal{L}(\alpha_k) + G_k(\alpha_{k+1} - \alpha_k) = e_k \,,$$

and $e_k \in \mathbb{R}^r$ is an error term satisfying $\|e_k\| \leq \rho_k \|G_k\|$ and $\rho_k \geq 0$.

(ii) **if** $\mathcal{L}(\alpha_k) = 0$ **then terminate.**

End of for-loop

The tameness assumption is extremely mild, as it includes basically any function that occurs in practical applications, by excluding pathological special cases. For example this class of functions comprises all semi-algebraic functions [4].

7 Numerical Experiment

We apply our proposed algorithm for solving a challenging non-smooth image deblurring problem under a Poisson noise assumption [27]. Given the observation $b \in \mathbb{R}^{n_x \times n_y}$ as $n_x \times n_y$-sized image, the task is the following popular problem:

$$\min_{x \in \mathbb{R}_+^{n_x \times n_y}} D_{KL}(b, Ax) + \gamma \|\mathcal{D}x\|_{2,1} \,, \tag{20}$$

which involves the Kullback–Leibler divergence as data fidelity measure $h(x) \coloneqq D_{\mathrm{KL}}(b, Ax) \coloneqq \sum_{i,j} (Ax)_{i,j} - b_{i,j} \log((Ax)_{i,j})$ with respect to the blurred reconstruction Ax with known blur operator and a discrete total variation regularization term $f(x) = \gamma \|\mathcal{D}x\|_{2,1}$ that is steered by a weight $\gamma > 0$ where $\mathcal{D}$ implements discrete spatial finite differences. We recast (20) into the saddle point problem:

$$\min_{x \in \mathbb{R}^{n_x \times n_y}} \max_{y \in \mathbb{R}^{2 \times n_x \times n_y}} \langle \mathcal{D}x, y \rangle + \delta_{\mathbb{R}_+^{n_x \times n_y}}(x) + D_{KL}(b, Ax) - \delta_{\|\cdot\|_{2,\infty} \leq \gamma}(y) \tag{21}$$

and set $g(x) \coloneqq \delta_{\mathbb{R}_+^{n_x \times n_y}}(x)$ and $K = \mathcal{D}$ in (1). Figure 1 compares several methods including PDHG with fixed stepsize (`PDHG`), PDHG with line search (`PDAL`), PDHG with fixed stepsize and variable metric (`VarPDHG`), PDHG with variable metric and line search (`VarPDAL`). The variable metric is generated by the limited memory BFGS method in Section 5. Figure 1 shows the primal gap where the optimal primal value was computed for 10000 iterations by running `PDHG`. For the update of the variable metric (17), we set $\gamma_1 = 1.0$ and $\gamma_2 = 0.99$. However, the Assumption 1 is not satisfied since it is not guaranteed by the construction of the metric that there is a sequence $(\eta_k)_k \in \ell_+^1(\mathbb{N})$ such that $M_{k+1} \preceq (1 + \eta_k) M_k$. Fortunately, we still observe convergence of PDHG with this variable metric. Figure 1 shows that a variable metric (`VarPDHG`) improves the convergence vs only using line search. However, our algorithm `VarPDAL` that combines both features is even faster, with the best performance when $m = 9$.

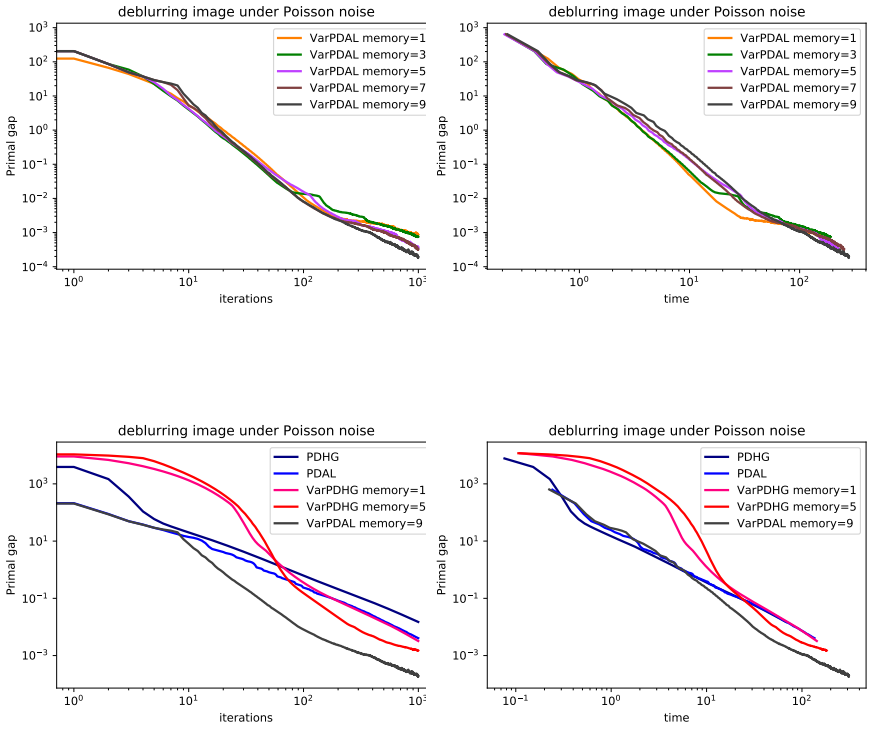

Fig. 1: Performance evaluation for the experiment in (21). Algorithms are described in the text. Our algorithms, which combine a quasi-Newton variable metric with line search, outperform all other algorithms (that either use a variable metric `VarPDHG` or line search `PDAL`; or none of the two `PDHG`).

As a second experiment to test out accelerated version, we consider $A = I$ and a constraint set $\mathcal{C} := \{x | x_{ij} \in [\epsilon, 255]\}$ since the grey value of each pixel should be less than 255 and be positive; We set $\epsilon = 0.1$. In this case $D_{KL}(b, x)$ restricted on $\mathcal{C}$ is a strongly convex function. We apply the accelerated version of Algorithm 1 in Theorem 3, and we use the notation `VarAPDAL`. Similary, `APDAL` denotes the accelerated version of `PDAL`. The dashed line in 2 corresponds to $O(1/N^2)$. We can observe that `VarPDAL`, `APDAL` can converge faster than $O(1/N^2)$ as predicted by the convergence theorem.

In order to record the exact algorithmic details for our experiments, all code for experiments from this paper is available at `https://github.com/wsdxiaohao/VarPDAL.git`.

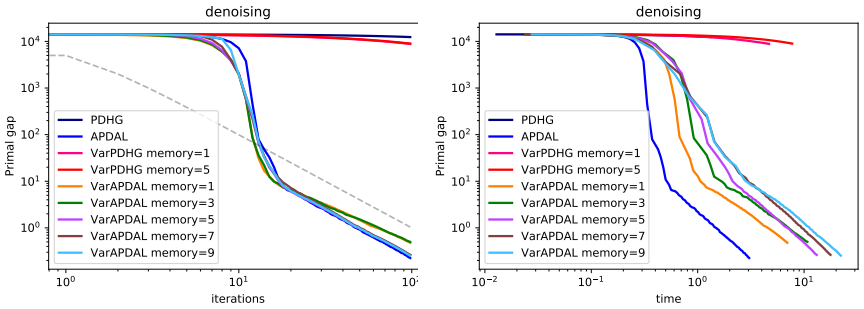

Fig. 2: Performance evaluation for the experiment for a strongly convex case. This figure reflects that our algorithms can retrieve $O(1/N^2)$ convergence rate as `PDAL`, which is theoretically guaranteed by Theorem 3.

8 Conclusion

In this paper, we introduced a line search variant of a recently introduced quasi-Newton primal-dual algorithm. In contrast to related work, the employed quasi-Newton metric is of type "identity $\pm$ low rank", which captures significantly more second order information than a commonly used diagonal metric. We equally care for both, theoretical convergence guarantees including convergence rates as well as efficient practical implementation. The additional line search procedure usually leads to larger steps at a computational cost that pays off, which is confirmed by our numerical experiments.

References

1. D. Applegate, M. Díaz, O. Hinder, H. Lu, M. Lubin, B. O'Donoghue, and W. Schudy. Practical large-scale linear programming using primal-dual hybrid gradient. *Advances in Neural Information Processing Systems*, 34, 2021.
2. S. Becker and J. Fadili. A quasi-Newton proximal splitting method. *Advances in Neural Information Processing Systems*, 25, 2012.
3. S. Becker, J. Fadili, and P. Ochs. On quasi-Newton forward-backward splitting: proximal calculus and convergence. *SIAM Journal on Optimization*, 29(4):2445–2481, 2019.
4. J. Bolte, A. Daniilidis, and A. Lewis. Tame functions are semismooth. *Mathematical Programming*, 117(1):5–19, 2009.
5. R.H. Byrd, J. Nocedal, and R.B. Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1):129–156, 1994.
6. A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.
7. A. Chambolle and T. Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

8. A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1):253–287, 2016.
9. F. H. Clarke. *Optimization and nonsmooth analysis*. Society for Industrial and Applied Mathematics, 1990.
10. P. Combettes, L. Condat, J.C. Pesquet, and B. Vu. A forward-backward view of some primal-dual optimization methods in image recovery. *IEEE International Conference on Image Processing*, 2014.
11. P. L. Combettes and B. C. Vũ. Variable metric forward–backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):1289–1318, 2014.
12. D. Davis. Convergence rate analysis of primal-dual splitting schemes. *SIAM Journal on Optimization*, 25(3):1912–1943, 2015.
13. R. Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.
14. T. Goldstein, M. Li, X. Yuan, E. Esser, and R. Baraniuk. Adaptive primal-dual hybrid gradient methods for saddle-point problems. *arXiv:1305.0546*, 2013.
15. C. Kanzow and T. Lechner. Globalized inexact proximal newton-type methods for nonconvex composite functions. *Computational Optimization and Applications*, 78:377–410, 2021.
16. C. Kanzow and T. Lechner. Efficient regularized proximal quasi-Newton methods for large-scale nonconvex composite optimization problems. Technical report, University of Würzburg, Institute of Mathematics, January 2022.
17. S. Karimi and S. Vavasis. IMRO: A proximal quasi-Newton method for solving l1-regularized least squares problems. *SIAM Journal on Optimization*, 27(2):583–615, 2017.
18. J.D. Lee, Y. Sun, and M.A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.
19. D. A. Lorenz and T. Pock. An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2):311–325, 2015.
20. Y. Malitsky and T. Pock. A first-order primal-dual algorithm with linesearch. *SIAM Journal on Optimization*, 28(1):411–432, 2018.
21. P. Patrinos, L. Stella, and A. Bemporad. Forward-backward truncated Newton methods for convex composite optimization. *arXiv:1402.6655*, 2014.
22. B. Polyak. *Introduction to optimization*. Optimization Software, 1987.
23. M. Schmidt, D. Kim, and S. Sra. Projected Newton-type methods in machine learning. *Optimization for Machine Learning*, (1), 2012.
24. M. Schmidt, D. Kim, and S. Sra. Projected newton-type methods in machine learning. *Optimization for Machine Learning*, 2012.
25. L. Stella, A. Themelis, and P. Patrinos. Forward–backward quasi-Newton methods for nonsmooth optimization problems. *Computational Optimization and Applications*, 67(3):443–487, 2017.
26. T. Valkonen. A primal–dual hybrid gradient method for nonlinear operators with applications to MRI. *Inverse Problems*, 30(5):055012, 2014.
27. Y. Vardi, L.A. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *Journal of the American statistical Association*, 80(389):8–20, 1985.
28. S. Wang, J. Fadili, and P. Ochs. Inertial quasi-newton methods for monotone inclusion: Efficient resolvent calculus and primal-dual methods. *arXiv:2209.14019*, 2022.
29. S. Wright and J. Nocedal. Numerical optimization. *Springer Science*, 1999.

A Preliminaries

There are several preliminaries we will use in the following section. The first one is a convergence result from [22, Lemma 2.2.2] of a special sequence which appears in B.2.

Lemma 2. *Let* $a_k \geq 0$ *and let*

$$\begin{aligned} &a_{k+1} \leq (1+\nu_k)a_k + \zeta_k, \quad \nu_k \geq 0,\ \zeta_k \geq 0, \\ &\sum_{k\in\mathbb{N}} \nu_k < \infty, \quad \sum_{k\in\mathbb{N}} \zeta_k < \infty. \end{aligned} \tag{22}$$

Then, $a_k \to A \geq 0$ *for some* $A < +\infty$.

The following identity is called (cosine rule), which proves to be very useful.

$$2\langle a-b, c-a\rangle = \|b-c\|^2 - \|a-b\|^2 - \|a-c\|^2 \quad \forall a,b,c \in X\,. \tag{23}$$

Another inequality appears many times in B.2 is the characteristic property of the proximal operator with respect to a symmetric positive definite matrix M:

$$\hat{x} = \mathrm{prox}_g^M(\bar{x}) \iff \langle \hat{x}-\bar{x}, y-\hat{x}\rangle_M \geq g(\bar{x}) - g(y) \quad \forall y \in X\,. \tag{24}$$

If $M = I$ is an identity matrix, then (24) is the characteristic property of the standard proximal operator. Assume $(\hat{x},\hat{y})$ is a saddle point which solves (1). Then we obtain

$$\begin{aligned} &P_{\hat{x},\hat{y}}(x) = g(x) + h(x) - g(\hat{x}) - h(\hat{x}) + \langle K^*\hat{y}, x-\hat{x}\rangle \geq 0 \quad \forall x \in X\,, \\ &D_{\hat{x},\hat{y}}(y) = f^*(y) - f^*(\hat{y}) - \langle K\hat{x}, y-\hat{y}\rangle \geq 0 \quad \forall y \in Y\,, \end{aligned} \tag{25}$$

where $P_{\hat{x},\hat{y}}(x)$ and $D_{\hat{x},\hat{y}}(y)$ are convex. Then $\mathcal{G}_{\hat{x},\hat{y}}(x,y) := P_{\hat{x},\hat{y}}(x) + D_{\hat{x},\hat{y}}(y)$ is the primal-dual gap. Without ambiguity, in the proofs, we may omit the subscript in P and D.

B Collection of Proofs

B.1 Proof of Lemma 1

It is a similar argument with the one in [20].

(i)&(ii) σ_k is decreased by $\mu \in (0,1)$ and the inequality (6) is satisfied as long as $\sigma_k < \underline{\sigma}_k := \frac{-1+\sqrt{(4\delta\alpha)/\beta_k+1}}{2\hat{L}}$ where $\hat{L} = \max\{L, L_K\}$. We introduce a notation $\underline{\sigma} := \frac{-1+\sqrt{(4\delta\alpha)/\beta+1}}{2\hat{L}}$. Since $\beta_k < \beta$, we have $\underline{\sigma}_k \geq \underline{\sigma}$. We argument by induction. We assume $\sigma_0 > \mu\underline{\sigma}_0$ and $\sigma_{k-1} > \mu\underline{\sigma}_{k-1}$. For the case $\sigma_k = \bar{\sigma}_k$, then $\sigma_k \geq (\frac{\beta_{k-1}}{\beta_k})\sigma_{k-1} > \mu(\frac{\beta_{k-1}}{\beta_k})\underline{\sigma}_{k-1} > \mu\underline{\sigma}_k > \mu\underline{\sigma}$. For the case $\sigma_k = \mu^i\bar{\sigma}_k$, $\sigma'_k = \mu^{i-1}\bar{\sigma}_k$ does not satisfy (6). It follows $\sigma'_k > \underline{\sigma}_k$. Thus, $\sigma_k = \mu\sigma'_k > \mu\underline{\sigma}_k \geq \mu\underline{\sigma}$.

(iii) By $\sigma_k \leq \sigma_{k-1}\sqrt{1+\theta_{k-1}}$, we get $\theta_k \leq \sqrt{1+\theta_{k-1}}$. Thus, θ_k is bounded from above. □

B.2 Proof of Theorem 1

The following proof is adapted from [20]. Assume $(\hat{x}, \hat{y})$ is a saddle point of problem 1 and $\beta_k \equiv \beta$. By using (24), we obtain the following two inequalities:

$$\left\langle y^{k+1} - y^k - \sigma_k K x^{k+1}, \hat{y} - y^{k+1} \right\rangle \geq \sigma_k (f^*(y^{k+1}) - f^*(\hat{y})) \tag{26}$$

$$\left\langle x^{k+1} - x^k + \tau_k M_k^{-1} K^* \bar{y}^k + \tau_k M_k^{-1} \nabla h(x^k), \hat{x} - x^{k+1} \right\rangle_{M_k} \geq \tau_k (g(x^{k+1}) - g(\hat{x})) \tag{27}$$

By using $\tau_k = \beta \sigma_k$

$$\begin{aligned} &\left\langle \frac{1}{\beta}(x^{k+1} - x^k) + \sigma_k M_k^{-1} K^* \bar{y}^k + \sigma_k M_k^{-1} \nabla h(x^k), \hat{x} - x^{k+1} \right\rangle_{M_k} \\ &\geq \sigma_k (g(x^{k+1}) - g(\hat{x})) \end{aligned} \tag{28}$$

Similarly, we apply (24) on y^k and obtain

$$\left\langle y^k - y^{k-1} - \sigma_{k-1} K x^k, y - y^k \right\rangle \geq \sigma_{k-1} (f^*(y^k) - f^*(y)) \quad \forall y \in Y\,. \tag{29}$$

Setting $y = y^{k+1}$ and $y = y^{k-1}$ respectively, we obtain

$$\left\langle y^k - y^{k-1} - \sigma_{k-1} K x^k, y^{k+1} - y^k \right\rangle \geq \sigma_{k-1} (f^*(y^k) - f^*(y^{k+1})) \quad \forall y \in Y\,, \tag{30}$$

$$\left\langle y^k - y^{k-1} - \sigma_{k-1} K x^k, y^{k-1} - y^k \right\rangle \geq \sigma_{k-1} (f^*(y^k) - f^*(y^{k-1})) \quad \forall y \in Y\,. \tag{31}$$

We deduce from (30) $\times\, \theta_k$ and $\theta_k = \frac{\sigma_k}{\sigma_{k-1}}$ that:

$$\left\langle \theta_k (y^k - y^{k-1}) - \sigma_k K x^k, y^{k+1} - y^k \right\rangle \geq \sigma_k (f^*(y^k) - f^*(y^{k+1}))\,. \tag{32}$$

By (31) $\times\, \theta_k^2$, we also get:

$$\left\langle \theta_k (y^k - y^{k-1}) - \sigma_k K x^k, \theta_k (y^{k-1} - y^k) \right\rangle \geq \sigma_k (\theta_k f^*(y^k) - \theta_k f^*(y^{k-1}))\,. \tag{33}$$

Summing (32) and (33) together, by using $\bar{y}^k = y^k + \theta_k (y^k - y^{k-1})$, we obtain

$$\left\langle \bar{y}^k - y^k - \sigma_k K x^k, y^{k+1} - \bar{y}^k \right\rangle \geq \sigma_k ((1+\theta_k) f^*(y^k) - \theta_k f^*(y^{k-1}) - f^*(y^{k+1}))\,. \tag{34}$$

To sum up inequalties (26), (28) and (34) , we obtain

$$\begin{aligned} &\left\langle y^{k+1} - y^k - \sigma_k K x^{k+1}, \hat{y} - y^{k+1} \right\rangle \\ &+ \left\langle \frac{1}{\beta}(x^{k+1} - x^k) + \sigma_k M_k^{-1} K^* \bar{y}^k + \sigma_k M_k^{-1} \nabla h(x^k), \hat{x} - x^{k+1} \right\rangle_{M_k} \\ &+ \left\langle \bar{y}^k - y^k - \sigma_k K x^k, y^{k+1} - \bar{y}^k \right\rangle \\ &\geq \sigma_k (f^*(y^{k+1}) - f^*(\hat{y})) + \sigma_k (g(x^{k+1}) - g(\hat{x})) + \sigma_k ((1+\theta_k) f^*(y^k) - \theta_k f^*(y^{k-1}) \\ &- f^*(y^{k+1}))\,, \end{aligned} \tag{35}$$

Reorganizing the above inequality and using $\tau_k = \beta\sigma_k$, we have

$$\begin{aligned}
&\left\langle y^{k+1} - y^k, \hat{y} - y^{k+1} \right\rangle + \frac{1}{\beta}\left\langle x^{k+1} - x^k, \hat{x} - x^{k+1} \right\rangle_{M_k} + \left\langle \bar{y}^k - y^k, y^{k+1} - \bar{y}^k \right\rangle \\
&+ \left\langle -\sigma_k K x^k, y^{k+1} - \bar{y}^k \right\rangle + \left\langle -\sigma_k K x^{k+1}, \hat{y} - y^{k+1} \right\rangle \\
&+ \left\langle \sigma_k K^* \bar{y}^k + \sigma_k \nabla h(x^k), \hat{x} - x^{k+1} \right\rangle \\
&\geq \sigma_k (g(x^{k+1}) - g(\hat{x})) + \sigma_k ((1+\theta_k) f^*(y^k) - \theta_k f^*(y^{k-1}) - f^*(\hat{y}))\,,
\end{aligned} \tag{36}$$

As in [20], we still have:

$$\begin{aligned}
&\left\langle -\sigma_k K x^k, y^{k+1} - \bar{y}^k \right\rangle + \left\langle -\sigma_k K x^{k+1}, \hat{y} - y^{k+1} \right\rangle + \left\langle \sigma_k K^* \bar{y}^k, \hat{x} - x^{k+1} \right\rangle \\
&= \sigma_k \left\langle K x^k - K x^{k+1}, \bar{y}^k - y^{k+1} \right\rangle + \sigma_k \left\langle K\hat{x}, \bar{y}^k - \hat{y} \right\rangle - \sigma_k \left\langle K^* \hat{y}, x^{k+1} - \hat{x} \right\rangle
\end{aligned} \tag{37}$$

Adding $\sigma_k h(x^{k+1}) - \sigma_k h(\hat{x})$ on both sides of (36), we obtain:

$$\begin{aligned}
&\left\langle y^{k+1} - y^k, \hat{y} - y^{k+1} \right\rangle + \frac{1}{\beta}\left\langle x^{k+1} - x^k, \hat{x} - x^{k+1} \right\rangle_{M_k} + \left\langle \bar{y}^k - y^k, y^{k+1} - \bar{y}^k \right\rangle \\
&+ \left\langle -\sigma_k K x^k, y^{k+1} - \bar{y}^k \right\rangle + \left\langle -\sigma_k K x^{k+1}, \hat{y} - y^{k+1} \right\rangle \\
&+ \left\langle \sigma_k K^* \bar{y}^k + \sigma_k \nabla h(x^k), \hat{x} - x^{k+1} \right\rangle + \sigma_k h(x^{k+1}) - \sigma_k h(\hat{x}) \\
&\geq \sigma_k (g(x^{k+1}) - g(\hat{x}) + (1+\theta_k) f^*(y^k) - \theta_k f^*(y^{k-1}) - f^*(\hat{y}) + h(x^{k+1}) - h(\hat{x}))\,.
\end{aligned} \tag{38}$$

Combining (37) and (38), we have

$$\begin{aligned}
&\left\langle y^{k+1} - y^k, \hat{y} - y^{k+1} \right\rangle + \frac{1}{\beta}\left\langle x^{k+1} - x^k, \hat{x} - x^{k+1} \right\rangle_{M_k} + \left\langle \bar{y}^k - y^k, y^{k+1} - \bar{y}^k \right\rangle \\
&\sigma_k \left\langle K x^k - K x^{k+1}, \bar{y}^k - y^{k+1} \right\rangle + \sigma_k \left\langle K\hat{x}, \bar{y}^k - \hat{y} \right\rangle - \sigma_k \left\langle K^* \hat{y}, x^{k+1} - \hat{x} \right\rangle \\
&+ \left\langle \sigma_k \nabla h(x^k), \hat{x} - x^{k+1} \right\rangle + \sigma_k h(x^{k+1}) - \sigma_k h(\hat{x}) \\
&\geq \sigma_k (g(x^{k+1}) - g(\hat{x}) + (1+\theta_k) f^*(y^k) - \theta_k f^*(y^{k-1}) - f^*(\hat{y}) + h(x^{k+1}) - h(\hat{x}))\,.
\end{aligned} \tag{39}$$

By the definition of $D(y)$ (25) and $\bar{y}^k = y^k + \theta_k(y^k - y^{k-1})$, we have

$$\begin{aligned}
&(1+\theta_k) f^*(y^k) - \theta_k f^*(y^{k-1}) - f^*(\hat{y}) - \left\langle K\hat{x}, \bar{y}^k - \hat{y} \right\rangle \\
&= (1+\theta_k)(f^*(y^k) - f^*(\hat{y}) - \left\langle K\hat{x}, y^k - \hat{y} \right\rangle) - \theta_k (f^*(y^{k-1}) - f^*(\hat{y}) \\
&- \left\langle K\hat{x}, y^{k-1} - \hat{y} \right\rangle) \\
&= (1+\theta_k) D(y^k) - \theta_k D(y^{k-1})\,.
\end{aligned} \tag{40}$$

Using (40) and the definition of $P(x)$, we deduce from (39) that

$$\begin{aligned}
&\left\langle y^{k+1}-y^k, \hat{y}-y^{k+1}\right\rangle + \frac{1}{\beta}\left\langle x^{k+1}-x^k, \hat{x}-x^{k+1}\right\rangle_{M_k} + \left\langle \bar{y}^k-y^k, y^{k+1}-\bar{y}^k\right\rangle \\
&+ \sigma_k\left\langle Kx^k-Kx^{k+1}, \bar{y}^k-y^{k+1}\right\rangle + \left\langle \sigma_k\nabla h(x^k), \hat{x}-x^{k+1}\right\rangle \\
&+ \sigma_k h(x^{k+1}) - \sigma_k h(\hat{x}) \\
&\geq \sigma_k(P(x^{k+1}) + (1+\theta_k)D(y^k) - \theta_k D(y^{k-1}))\,.
\end{aligned} \tag{41}$$

From the line search condition (6), we have

$$\begin{aligned}
&\sigma_k(h(x^{k+1}) - h(x^k) - \left\langle \nabla h(x^k), x^{k+1}-x^k\right\rangle) \\
&\leq \frac{\delta}{2\beta}\|x^{k+1}-x^k\|^2_{M_k} - \frac{1}{2}\sigma_k^2\|Kx^{k+1}-Kx^k\|^2\,.
\end{aligned} \tag{42}$$

Additionally, by the convexity of $h(x)$, we also have

$$h(x^k) - h(\hat{x}) + \left\langle \nabla h(x^k), \hat{x}-x^k\right\rangle \leq 0\,. \tag{43}$$

Combining (42) and $\sigma_k \times$ (43), we get

$$\begin{aligned}
&\sigma_k(h(x^{k+1}) - h(\hat{x}) - \left\langle \nabla h(x^k), x^{k+1}-\hat{x}\right\rangle) \\
&\leq \frac{\delta}{2\beta}\|x^{k+1}-x^k\|^2_{M_k} - \frac{1}{2}\sigma_k^2\|Kx^{k+1}-Kx^k\|^2\,.
\end{aligned} \tag{44}$$

Thus, it follows from (41) and (44) that

$$\begin{aligned}
&\left\langle y^{k+1}-y^k, \hat{y}-y^{k+1}\right\rangle + \frac{1}{\beta}\left\langle x^{k+1}-x^k, \hat{x}-x^{k+1}\right\rangle_{M_k} + \left\langle \bar{y}^k-y^k, y^{k+1}-\bar{y}^k\right\rangle \\
&+ \sigma_k\left\langle Kx^k-Kx^{k+1}, \bar{y}^k-y^{k+1}\right\rangle + \frac{\delta}{2\beta}\|x^{k+1}-x^k\|^2_{M_k} - \frac{1}{2}\sigma_k^2\|Kx^{k+1}-Kx^k\|^2 \\
&\geq \sigma_k(P(x^{k+1}) + (1+\theta_k)D(y^k) - \theta_k D(y^{k-1}))\,.
\end{aligned} \tag{45}$$

Using Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
&\left\langle y^{k+1}-y^k, \hat{y}-y^{k+1}\right\rangle + \frac{1}{\beta}\left\langle x^{k+1}-x^k, \hat{x}-x^{k+1}\right\rangle_{M_k} + \left\langle \bar{y}^k-y^k, y^{k+1}-\bar{y}^k\right\rangle \\
&+ \frac{1}{2}\sigma_k^2\|Kx^k-Kx^{k+1}\|^2 + \frac{1}{2}\|\bar{y}^k-y^{k+1}\|^2 + \frac{\delta}{2\beta}\|x^{k+1}-x^k\|^2_{M_k} \\
&- \frac{1}{2}\sigma_k^2\|Kx^{k+1}-Kx^k\|^2 \\
&\geq \sigma_k(P(x^{k+1}) + (1+\theta_k)D(y^k) - \theta_k D(y^{k-1}))\,.
\end{aligned} \tag{46}$$

Applying (23), we deduce from (46) that

$$
\begin{aligned}
&(\frac{1}{2}\|y^k-\hat{y}\|^2-\frac{1}{2}\|y^{k+1}-y^k\|^2-\frac{1}{2}\|\hat{y}-y^{k+1}\|^2)\\
&+(\frac{1}{2\beta}\|x^k-\hat{x}\|^2_{M_k}-\frac{1}{2\beta}\|x^{k+1}-x^k\|^2_{M_k}-\frac{1}{2\beta}\|\hat{x}-x^{k+1}\|^2_{M_k})\\
&+(\frac{1}{2}\|y^k-y^{k+1}\|^2-\frac{1}{2}\|\bar{y}^k-y^k\|^2-\frac{1}{2}\|y^{k+1}-\bar{y}^k\|^2)\\
&+\frac{1}{2}\|\bar{y}^k-y^{k+1}\|^2+\frac{\delta}{2\beta}\|x^{k+1}-x^k\|^2_{M_k}\\
&\geq \sigma_k(P(x^{k+1})+(1+\theta_k)D(y^k)-\theta_k D(y^{k-1}))\,.
\end{aligned}
\tag{47}
$$

Reorganizing the above inequalities, we obtain

$$
\begin{aligned}
&\frac{1}{2}\|y^k-\hat{y}\|^2+\frac{1}{2\beta}\|x^k-\hat{x}\|^2_{M_k}-\frac{1-\delta}{2\beta}\|x^{k+1}-x^k\|^2_{M_k}\\
&+\sigma_k\theta_k D(y^{k-1})-\frac{1}{2}\|\bar{y}^k-y^k\|^2\\
&\geq \sigma_k(P(x^{k+1})+(1+\theta_k)D(y^k))+\frac{1}{2}\|\hat{y}-y^{k+1}\|^2+\frac{1}{2\beta}\|\hat{x}-x^{k+1}\|^2_{M_k}\,.
\end{aligned}
\tag{48}
$$

It follows from $\bar{\sigma}_k \leq \sqrt{1+\theta_{k-1}}\sigma_{k-1}$ that $\sigma_k\theta_k \leq \frac{\sigma_k^2}{\sigma_{k-1}} \leq \frac{\bar{\sigma}_k^2}{\sigma_{k-1}} \leq (1+\theta_{k-1})\sigma_{k-1}$. Thus,

$$
\begin{aligned}
&\frac{1}{2}\|y^k-\hat{y}\|^2+\frac{1}{2\beta}\|x^k-\hat{x}\|^2_{M_k}-\frac{1-\delta}{2\beta}\|x^{k+1}-x^k\|^2_{M_k}\\
&+\sigma_{k-1}(1+\theta_{k-1})D(y^{k-1})-\frac{1}{2}\|\bar{y}^k-y^k\|^2\\
&\geq \sigma_k(1+\theta_k)D(y^k)+\frac{1}{2}\|\hat{y}-y^{k+1}\|^2+\frac{1}{2\beta}\|\hat{x}-x^{k+1}\|^2_{M_k}\,.
\end{aligned}
\tag{49}
$$

Since $(1+\eta_k)M_k \succeq M_{k+1}$, we can obtain the following key inequality:

$$
\begin{aligned}
&\frac{1}{2}\|y^k-\hat{y}\|^2+\frac{1}{2\beta}\|x^k-\hat{x}\|^2_{M_k}-\frac{1-\delta}{2\beta}\|x^{k+1}-x^k\|^2_{M_k}\\
&+\sigma_{k-1}(1+\theta_{k-1})D(y^{k-1})-\frac{1}{2}\|\bar{y}^k-y^k\|^2\\
&\geq \sigma_k(1+\theta_k)D(y^k)+\frac{1}{2}\|\hat{y}-y^{k+1}\|^2+\frac{1}{2\beta(1+\eta_k)}\|\hat{x}-x^{k+1}\|^2_{M_{k+1}}\,.
\end{aligned}
\tag{50}
$$

Set $A_k \coloneqq \frac{1}{2}\|y^k-\hat{y}\|^2+\sigma_{k-1}(1+\theta_{k-1})D(y^{k-1})+\frac{1}{2\beta}\|x^k-\hat{x}\|^2_{M_k}$. Then, we deduce from (50) that

$$
A_{k+1} \leq (1+\eta_k)A_k\,. \tag{51}
$$

By Lemma 2, A_k is bounded from above by some constant C. Thus, $\|y^k-\hat{y}\|$ and $\|x^k-\hat{x}\|_{M_k}$ are both bounded. By the assumption that M_k is uniformly

bounded, $\|x^k - \hat{x}\|$ is also bounded. As a result, we deduce from (50) that

$$\begin{aligned}\sum_k \big(\frac{1-\delta}{2\beta}\|x^{k+1}-x^k\|^2_{M_k} + \frac{1}{2}\|\bar{y}^k - y^k\|^2\big) &\le \sum_k \big((1+\eta_k)A_k - A_{k+1}\big) \\ &\le C\sum_k \eta_k + A_0 < +\infty\,.\end{aligned} \tag{52}$$

It implies that $\|x^{k+1}-x^k\|_{M_k} \to 0$ and $\|\bar{y}^k - y^k\| \to 0$. So does $\|x^{k+1}-x^k\| \to 0$, since $(M_k)_{k\in\mathbb{N}} \subset \mathcal{S}_\alpha(X)$. Since $\sigma_k > \sigma$ for some σ which is shown in Lemma 1 and $\beta > 0$ is fixed,

$$\begin{aligned}&\frac{y^{k+1}-y^k}{\sigma_k} = \frac{\bar{y}^{k+1}-y^{k+1}}{\sigma_{k+1}} \to 0 \quad \text{as } k \to +\infty\,,\\ &\frac{\|x^{k+1}-x^k\|^2_{M_k}}{\tau_k} \to 0 \quad \text{as } k \to +\infty\,.\end{aligned} \tag{53}$$

Since $(x^k, y^k)_{k\in\mathbb{N}}$ is bounded, we can extract a subsequence $(x^{k_i}, y^{k_i})_{i\in\mathbb{N}}$ converging to some cluster point (x^*, y^*). As in [20], similarly, by using the lower semi-continuity of functions g and f^* and the continuity of function h, we can pass the following two inequalities to the limit:

$$\begin{aligned}&\Big\langle \frac{y^{k_i+1}-y^{k_i}}{\sigma_{k_i}} - Kx^{k_i+1}, y - y^{k_i+1}\Big\rangle \ge (f^*(y^{k_i+1}) - f^*(y)) \quad \forall y \in Y\,,\\ &\Big\langle \frac{x^{k_i+1}-x^{k_i}}{\tau_{k_i}} + M_{k_i}^{-1}K^*\bar{y}^{k_i} + M_{k_i}^{-1}\nabla h(x^{k_i}), x - x^{k_i+1}\Big\rangle_{M_{k_i}}\\ &= \Big\langle \frac{M_{k_i}(x^{k_i+1}-x^{k_i})}{\tau_{k_i}}, x - x^{k_i+1}\Big\rangle + \big\langle K^*\bar{y}^{k_i} + \nabla h(x^{k_i}), x - x^{k_i+1}\big\rangle\\ &\ge (g(x^{k_i+1}) - g(x)) \quad \forall x \in X\,.\end{aligned} \tag{54}$$

Thus, (x^*, y^*) is the saddle point of (1). If, additionally, $f^*(y)|_{dom_{f^*}}$ is continuous, then $f^*(y^{k_i}) \to f^*(y^*)$ and $D(y^{k_i}) \to 0$ as $i \to +\infty$. From (50), we have $\frac{1}{\Pi_{j=1}^k(1+\eta_j)}A_k$ is monotone. Setting $\hat{x} = x^*$ and $\hat{y} = y^*$ in (50), by the boundedness of σ_k and θ_k, it follows that

$$\lim_{k\to\infty}\frac{A_k}{\Pi_{i=1}^\infty(1+\eta_i)} \le \lim_{k\to\infty}\frac{A_k}{\Pi_{i=1}^k(1+\eta_i)} = \lim_{i\to\infty}\frac{A_{k_i}}{\Pi_{j=1}^{k_i}(1+\eta_j)} \le \lim_{i\to\infty} A_{k_i} = 0 \tag{55}$$

Since $\Pi_{i=1}^\infty(1+\eta_i) < +\infty$, we have $\lim_{k\to+\infty} A_k \to 0$ which means $x^k \to x^*$ and $y^k \to y^*$ as $k \to +\infty$. □

B.3 Proof of Theorem 2

We adapt the corresponding proof in [20]. Let $\epsilon_k := \sigma_k\big(P(x^{k+1}) + (1+\theta_k)D(y^k) - \theta_k D(y^{k-1})\big)$. Then we obtain the following inequality from (47),

$$\frac{1}{2}\|y^k - \hat{y}\|^2 - \frac{1}{2}\|y^{k+1} - \hat{y}\|^2 + \frac{1}{2\beta}\|x^k - \hat{x}\|^2_{M_k} - \frac{1}{2\beta}\|x^{k+1} - \hat{x}\|^2_{M_k} - \frac{1}{2}\|\bar{y}^k - y^k\|^2 \ge \epsilon_k\,. \tag{56}$$

By the assumption 1, we get

$$\frac{1}{2}\|y^k-\hat{y}\|^2-\frac{1}{2}\|y^{k+1}-\hat{y}\|^2+\frac{1}{2\beta}\|x^k-\hat{x}\|^2_{M_k}-\frac{1}{2\beta}\frac{\|x^{k+1}-\hat{x}\|^2_{M_{k+1}}}{(1+\eta_k)}-\frac{1}{2}\|\bar{y}^k-y^k\|^2 \geq \epsilon_k\,. \tag{57}$$

Since $(1+\eta_k)\geq 1$, it follows

$$\frac{1}{2}\|y^k-\hat{y}\|^2-\frac{1}{2}\frac{\|y^{k+1}-\hat{y}\|^2}{(1+\eta_k)}+\frac{1}{2\beta}\|x^k-\hat{x}\|^2_{M_k}-\frac{1}{2\beta}\frac{\|x^{k+1}-\hat{x}\|^2_{M_{k+1}}}{(1+\eta_k)}\geq \epsilon_k\,. \tag{58}$$

Let both sides of the above inequality be divided by $\Pi_{i=1}^{k-1}(1+\eta_i)$ and it is common to assume that an empty product yields identity i.e. $\Pi_{i=1}^{0}(1+\eta_i)=1$. Thus,

$$\begin{aligned}&\frac{1}{2}\frac{\|y^k-\hat{y}\|^2}{\Pi_{i=1}^{k-1}(1+\eta_i)}-\frac{1}{2}\frac{\|y^{k+1}-\hat{y}\|^2}{\Pi_{i=1}^{k}(1+\eta_i)}+\frac{1}{2\beta}\frac{\|x^k-\hat{x}\|^2_{M_k}}{\Pi_{i=1}^{k-1}(1+\eta_i)}-\frac{1}{2\beta}\frac{\|x^{k+1}-\hat{x}\|^2_{M_{k+1}}}{\Pi_{i=1}^{k}(1+\eta_i)}\\ &\geq\frac{\epsilon_k}{\Pi_{i=1}^{k}(1+\eta_i)}\,.\end{aligned} \tag{59}$$

Summing up (59) for $k=1,\cdots,N$, we obtain

$$\frac{1}{2}\|y^1-\hat{y}\|^2+\frac{1}{2\beta}\|x^1-\hat{x}\|^2_{M_1}\geq\sum_{k=1}^{N}\frac{\epsilon_k}{\Pi_{i=1}^{k}(1+\eta_i)}\geq\sum_{k=1}^{N}\frac{\epsilon_k}{C}\,. \tag{60}$$

Here, we used the $C=\sum_{k\in\mathbb{N}}(1+\eta_k)<+\infty$.

The following steps are similar with the ones in [20].

$$\begin{aligned}\sum_{k=1}^{N}\epsilon_k=&\sigma_N(1+\theta_N)D(y^k)+\sum_{k=2}^{N}[(1+\theta_{k-1})\sigma_{k-1}-\theta_k\sigma_k]D(y^{k-1})\\ &-\theta_1\sigma_1 D(y^0)+\sum_{k=1}^{N}\sigma_k P(x^{k+1})\,.\end{aligned} \tag{61}$$

Since D is convex,

$$\begin{aligned}&\sigma_N(1+\theta_N)D(y^N)+\sum_{k=2}^{N}[(1+\theta_{k-1})\sigma_{k-1}-\theta_k\sigma_k]D(y^{k-1})\\ &\geq(\sigma_1\theta_1+s_N)D(\frac{\sigma_1(1+\theta_1)y^1+\sum_{k=2}^{N}\sigma_k\bar{y}^k}{\sigma_1\theta_1+s_N})\\ &=(\sigma_k\theta_1+s_N)D(\frac{\sigma_1\theta_1 y^0+\sum_{k=1}^{N}\sigma_k\bar{y}^k}{\sigma_1\theta_1+s_N})\\ &\geq s_N D(\bar{Y}^N)\,,\end{aligned} \tag{62}$$

where $s_N = \sum_{k=1}^{N} \sigma_k$. Similarly,

$$\sum_{k=1}^{N} \sigma_k P(x^{k+1}) \geq s_N P(\frac{\sum_{k=1}^{N} \sigma_k x^{k+1}}{s_N}) = s_N P(\bar{X}^N)\,. \tag{63}$$

As a result,

$$\mathcal{G}(\bar{X}^N, \bar{Y}^N) = P(\bar{X}^N) + D(\bar{Y}^N) \leq \frac{C}{s_N}\big(\frac{1}{2\beta}\|x^1 - \hat{x}\|^2_{M_1} + \frac{1}{2}\|y^1 - \hat{y}\|^2 + \sigma_1\theta_1 D(y^0)\big)\,. \tag{64}$$

□

B.4 Proof of Theorem 3

The proof is also adapted from [20]. From the update formula of β_k, it follows that β_k is decreasing. First, we are going to prove that θ_k is bounded from above. It is not difficult but tedious. We know that if there exists a $C \in \mathbb{R}_+$ s.t $\theta_k \leq C\sqrt{1+\theta_{k-1}}$ then θ_k is bounded. From this, it is sufficient to prove that $\frac{\beta_{k-1}}{\beta_k}$ is uniformly bounded from above by some C_θ. According to

$$\beta_k = \frac{\beta_{k-1}}{\min\{1 + \frac{\gamma}{C_M}\beta_{k-1}\sigma_{k-1}, C_\theta\}}\,,\ \forall k \in \mathbb{N}\,, \quad \text{and} \quad \beta_0 > 0\,, \tag{65}$$

we have that $\frac{\beta_{k-1}}{\beta_k} = \min\{1 + \frac{\gamma}{C_M}\beta_{k-1}\sigma_{k-1}, C_\theta\} \leq C_\theta$.

Second part, we are going to show the convergence rate. Since g is strongly convex, we obtain:

$$\begin{aligned}&\left\langle \frac{x^{k+1} - x^k}{\tau_k} + M_k^{-1}K^*\bar{y}^k + M_k^{-1}\nabla h(x^k), \hat{x} - x^{k+1} \right\rangle_{M_k} \\ &\geq (g(x^{k+1}) - g(\hat{x})) + \frac{\gamma}{2}\|x^{k+1} - \hat{x}\|^2\,.\end{aligned} \tag{66}$$

From Assumption 1, it follows that for any $k \in \mathbb{N}$,

$$\frac{\gamma}{2}\|x^{k+1} - \hat{x}\|^2 \geq \frac{\gamma}{2C_M}\|x^{k+1} - \hat{x}\|^2_{M_{k+1}}\,. \tag{67}$$

Following the same way in which we got equation (48), by equation (66) and the assumption that $(1+\eta_k)M_k \succeq M_{k+1}$, we obtain

$$\begin{aligned}&\frac{1}{2}\|y^k - \hat{y}\|^2 - \frac{1}{2}\|y^{k+1} - \hat{y}\|^2 + \frac{1}{2\beta_k}\|x^k - \hat{x}\|^2_{M_k} - \frac{1-\delta}{2\beta_k}\|x^{k+1} - x^k\|^2_{M_k} \\ &\quad - \frac{1}{2\beta_k}\frac{\|x^{k+1} - \hat{x}\|^2_{M_{k+1}}}{(1+\eta_k)} - \frac{1}{2}\|\bar{y}^k - y^k\|^2 \geq \epsilon_k + \frac{\gamma\sigma_k}{2}\|x^{k+1} - \hat{x}\|^2\,.\end{aligned} \tag{68}$$

In order to obtain the following inequality, it is sufficient to assume $\delta \leq 1$. Thus,

$$\frac{1}{2}\|y^k - \hat{y}\|^2 - \frac{1}{2}\|y^{k+1} - \hat{y}\|^2 + \frac{1}{2\beta_k}\|x^k - \hat{x}\|^2_{M_k} - \frac{1}{2\beta_k}\frac{\|x^{k+1} - \hat{x}\|^2_{M_{k+1}}}{(1+\eta_k)} - \frac{1}{2}\|\bar{y}^k - y^k\|^2 \geq \epsilon_k + \frac{\gamma\sigma_k}{2}\|x^{k+1} - \hat{x}\|^2 . \tag{69}$$

Since $\delta \leq 1$, by dividing the above inequality with σ_k, we have

$$\frac{1}{2\sigma_k}\|y^k - \hat{y}\|^2 - \frac{1}{2\sigma_k}\|y^{k+1} - \hat{y}\|^2 + \frac{1}{2\tau_k}\|x^k - \hat{x}\|^2_{M_k} - \frac{1}{2\tau_k}\frac{\|x^{k+1} - \hat{x}\|^2_{M_{k+1}}}{(1+\eta_k)} - \frac{1}{2\sigma_k}\|\bar{y}^k - y^k\|^2 \geq \frac{\epsilon_k}{\sigma_k} + \frac{\gamma}{2}\|x^{k+1} - \hat{x}\|^2 , \tag{70}$$

where, we used $\tau_k = \beta_k\sigma_k$. By using (67), from the above inequality, we obtain that

$$\frac{1}{2\sigma_k}\|y^k - \hat{y}\|^2 - \frac{1}{2\sigma_k}\|y^{k+1} - \hat{y}\|^2 + \frac{1}{2\tau_k}\|x^k - \hat{x}\|^2_{M_k} - \frac{1}{2\tau_k}\frac{\|x^{k+1} - \hat{x}\|^2_{M_{k+1}}}{(1+\eta_k)} - \frac{1}{2\sigma_k}\|\bar{y}^k - y^k\|^2 \geq \frac{\epsilon_k}{\sigma_k} + \frac{\gamma}{2C_M}\|x^{k+1} - \hat{x}\|^2_{M_{k+1}} . \tag{71}$$

It follows from the above inequality that

$$\begin{aligned}
&\frac{1}{2\sigma_k}\|y^k - \hat{y}\|^2 - \frac{1}{2\sigma_k}\|y^{k+1} - \hat{y}\|^2 + \frac{1}{2\tau_k}\|x^k - \hat{x}\|^2_{M_k} - \frac{1}{2\sigma_k}\|\bar{y}^k - y^k\|^2 \\
&\qquad \geq \frac{\epsilon_k}{\sigma_k} + \frac{1 + (1+\eta_k)\tau_k\gamma/C_M}{2\tau_k(1+\eta_k)}\|x^{k+1} - \hat{x}\|^2_{M_{k+1}} , \\
&\frac{1}{2\sigma_k}\|y^k - \hat{y}\|^2 - \frac{1}{2\sigma_k}\|y^{k+1} - \hat{y}\|^2 + \frac{1}{2\tau_k}\|x^k - \hat{x}\|^2_{M_k} - \frac{1}{2\sigma_k}\|\bar{y}^k - y^k\|^2 \\
&\qquad \geq \frac{\epsilon_k}{\sigma_k} + \frac{\tau_{k+1}(1 + \tau_k\gamma/C_M)}{\tau_k}\frac{\|x^{k+1} - \hat{x}\|^2_{M_{k+1}}}{2\tau_{k+1}(1+\eta_k)} ,
\end{aligned} \tag{72}$$

For convenience, we set $\tilde{\gamma} = \gamma/C_M$. From the update step of β_k, it follows that

$$\frac{\tau_{k+1}(1 + \tilde{\gamma}\tau_k)}{\tau_k} \geq \frac{\tau_{k+1}\min\{C_\theta, (1 + \tilde{\gamma}\tau_k)\}}{\tau_k} = \frac{\sigma_{k+1}}{\sigma_k} \tag{73}$$

Set $B_k := \frac{1}{2\tau_k}\|x^k - \hat{x}\|^2_{M_k} + \frac{1}{2\sigma_k}\|y^k - \hat{y}\|^2$ and $\tilde{B}_k := \frac{B_k}{\Pi_{i=1}^{k-1}(1+\eta_i)}$. From (72), we have:

$$\frac{\sigma_{k+1}}{\sigma_k(1+\eta_k)}B_{k+1} + \frac{\epsilon_k}{\sigma_k} \leq B_k - \frac{1}{2\sigma_k}\|\bar{y}^k - y^k\|^2 \tag{74}$$

By dividing the above inequality by $\Pi_{i=1}^{k-1}(1+\eta_i) \geq 1$, we obtain

$$\frac{\sigma_{k+1}}{\sigma_k}\tilde{B}_{k+1} + \frac{\epsilon_k}{\sigma_k \Pi_{i=1}^{k-1}(1+\eta_i)} \leq \tilde{B}_k - \frac{1}{2\sigma_k \Pi_{i=1}^{k-1}(1+\eta_i)}\|\bar{y}^k - y^k\|^2 \tag{75}$$

By multiplying σ_k on both sides, we have

$$\sigma_{k+1}\tilde{B}_{k+1} + \frac{\epsilon_k}{\Pi_{i=1}^{k-1}(1+\eta_i)} \leq \sigma_k \tilde{B}_k - \frac{1}{2\Pi_{i=1}^{k-1}(1+\eta_i)}\|\bar{y}^k - y^k\|^2 . \tag{76}$$

By Assumption 1, $C = \Pi_{i\in\mathbb{N}}(1+\eta_i) < +\infty$, we have

$$\sigma_{k+1}\tilde{B}_{k+1} + \frac{\epsilon_k}{C} \leq \sigma_k \tilde{B}_k - \frac{1}{2C}\|\bar{y}^k - y^k\|^2 . \tag{77}$$

Summing up (77) from $k = 1, \cdots, N$, we obtain

$$\sigma_{N+1}\tilde{B}_{N+1} + \sum_{k=1}^{N}\frac{\epsilon_k}{C} \leq \sigma_1 \tilde{B}_1 - \frac{1}{2C}\sum_{k=1}^{N}\|\bar{y}^k - y^k\|^2 . \tag{78}$$

Since σ_k is bounded by some σ for any $k \in \mathbb{N}$, $\tilde{B}_k$ is bounded from above. Since $C = \Pi_{i\in\mathbb{N}}(1+\eta_i) < +\infty$, B_k is also bounded from above. So, y^k is also bounded with $\lim_{k\to\infty}\|\bar{y}^k - y^k\|^2 = 0$. Thus, using the similar argument and notations in the proof B.2, we retrieve the same key inequality as the one in [20]:

$$\begin{aligned} \mathcal{G}(\bar{X}^N, \bar{Y}^N) &\leq \frac{C}{s_N}(\sigma_1 B_1 + \theta_1\sigma_1 P(x^0)), \\ \|x^{N+1} - \hat{x}\|^2_{M_{N+1}} &\leq \frac{C\tau_{N+1}}{\sigma_{N+1}}(\sigma_1 A_1 + \theta_1\tau_1 P(x^0)) = C\beta_{N+1} , \end{aligned} \tag{79}$$

Using the same argument from [20], we know from B.1 that σ_k is bounded by $\mu\underline{\sigma}_k = \mu(\frac{-1+\sqrt{(4\delta\alpha)/\beta_k+1}}{2\hat{L}})$ where $\hat{L} = \max\{L, L_K\}$. We claim that there exists a constant C_β such that, $\beta_k = C_\beta(1/k^2)$.

i If $\alpha\delta/(\beta_k) \leq 1$, by $\sigma_k \geq \mu\underline{\sigma}_k \geq \mu\underline{\sigma}$, we have

$$\beta_{k+1} = \frac{\beta_k}{\min\{C_\theta, 1+\tilde{\gamma}\beta_k\sigma_k\}} \leq \frac{\beta_k}{\min\{C_\theta, 1+\mu\underline{\sigma}\delta\alpha\tilde{\gamma}\}} . \tag{80}$$

In this case, β_k decreases linearly. Thus, $\beta_{k+1} \leq C_\beta/(k+1)^2$ for k sufficiently large.

ii If $\alpha\delta/(\beta_k) \geq 1$, then $\sigma_k > \mu\underline{\sigma}_k > \frac{\mu}{2\hat{L}}\sqrt{\frac{\delta\alpha}{\beta_k}}$. Therefore, for k large enough, we have

$$\beta_{k+1} = \frac{\beta_k}{\min\{C_\theta, 1+\tilde{\gamma}\beta_k\sigma_k\}} \leq \frac{\beta_k}{\min\{C_\theta, 1+\frac{\mu\sqrt{\delta\alpha}\tilde{\gamma}}{2\hat{L}}\sqrt{\beta_k}\}} = \frac{\beta_k}{1+\frac{\mu\sqrt{\delta\alpha}\tilde{\gamma}}{2\hat{L}}\sqrt{\beta_k}} . \tag{81}$$

In this case, by induction $\beta_k \leq \frac{C_\beta}{k^2}$ for some constant $C_\beta > 0$.

From $\sigma_k > \mu\underline{\sigma}_k > \mu\underline{\sigma}$, we have $s_N = \sum_{k=1}^{N}\sigma_k > \sum_{k=1}^{N}\underline{\sigma}_k > \sum_{k=1}^{N}O(k) \sim N^2$ since $\beta_k \leq C_\beta/k^2$ for k sufficiently large. Then, we conclude the results. □