# Catalyst Meta Acceleration Framework: The history and the gist of it

Hongda Li

November 15, 2024

**Abstract**

Nesterov's accelerated gradient first appeared back in the 1983 has sparked numerous theoretical and practical advancements in Mathematics programming literatures. The idea behind Nesterov's acceleration is universal for convex objective, and it has concrete extension in the non-convex case. In this paper we survey specifically the Catalyst Acceleration that incorporated ideas from the Accelerated Proximal Point Method proposed by Guler back in 1993. The paper reviews Nesterov's classical analysis of accelerated gradient in the convex case. The paper will describe key aspects of the theoretical innovations involved to achieve the design of the algorithm in convex, and non-convex case.

## 1 Introduction

THIS REPORT IS CURRENTLY: UNFINISHED

The optimal algorithm named accelerated gradient descent method is proposed in Nesterov's seminal work back in 1983 [**?**]. The algorithm closed the upper bound and lower bound on the iteration complexity for all first order Lipschitz smooth convex function among all first order algorithms. For a specific definition of first order method, we refer reader to Chapter 2 of Nesterov's new book [5] for more information. Gradient descent has an upper bound of $\mathcal{O}(1/k)$ in iteration complexity that is slower than the lower iteration complexity $\mathcal{O}(1/k^2)$. Accelerated gradient descent has an upper bound of $\mathcal{O}(1/k^2)$, making it optimal.

It's tempting to believe that the existence of an optimal algorithm sealed the ceiling for the need of theories for convex first-order smooth optimization. It is correct but lacks the nuance

in understanding because Nesterov's accelerated gradient is a system of analysis techniques which is not a specific design paradigm for algorithms.

Guler's accelerated Proximal Point Method (PPM) [2] in 1993 used the technique of Nesterov's estimating sequence to accelerate PPM for convex objectives. Use $(\lambda_k)_{k\geq 0}$ to parameterize the proximal point evaluation to generate $(x_k)_{k\geq 0}$ given any initial guess $x_0$. Guler's prior work [1] showed the convergence of PPM method in the convex without acceleration is $\mathcal{O}\left(1/\sum_{i=1}^{n}\lambda_i\right)$. His new algorithm with acceleration has a rate of $\mathcal{O}\left(1/(\sum_{i=1}^{n}\sqrt{\lambda_i})^2\right)$. An inexact Accelerated PPM method using conditions described in Rockafellar's works in 1976 [7] is also discussed in the paper.

It's tempting to conclude that the results has reached the ceiling for extending Nesterov's acceleration. It is correct, but not from a practical point of view. Let $F:\mathbb{R}^n\to\mathbb{R}$ be our objective function, $\mathcal{J}_\lambda := (I+\lambda\partial F)^{-1}$ and $\mathcal{M}^\lambda(x;y) := F(x)+\frac{1}{2\lambda}\|x-y\|^2$ then the inexact proximal point considers with error $\epsilon_k$ has the following characterizations of inexactness as put forward by Guler [2]:

$$\begin{aligned}\tilde{x} &\approx \mathcal{J}_\lambda y\\ \operatorname{dist}(\mathbf{0},\partial\mathcal{M}^\lambda(\tilde{x};y)) &\leq \frac{\epsilon}{\lambda}.\end{aligned}$$

The difficulty comes from controlling the error $\epsilon$ at each iteration to ensure the overall convergence of accelerated PPM. In the paper $\epsilon_k\to 0$ at a specific rate. It requires knowledge about exact minimum of the Moreau envelope at each step and the optimal value of the Nesterov's estimating sequence. These quantities are intractable in practice making it impossible to formulate it to algorithms directly.

Introduced in Lin et al. [3, 4] is a series of papers on a concrete meta algorithm called Catalyst (It's called 4WD Catalyst for the non-convex extension in works by Paquette, Lin et al. [6]). The meta algorithm uses other first order algorithms to evaluate inexact proximal point method and then performs accelerated PPM, therefore the umbrella term: "meta". Major innovations include Tracking and controlling the errors made in the inexact PPM using Nesterov's estimating sequence throughout and an algorithm called accelerated MISO-Prox. Prior to Lin's paper, it was an open question on the conditions required to accelerate incremental method such as stochastic gradient descent can be accelerated.

## 1.1 Contributions

The writing is expository and comprehensive, it will survey the history and major results, and innovations involved in conceiving and designing the Catalyst algorithm. We reviewed the literatures and faithfully reproduced important claims. In addition, we give insights and context to understand the claims in these papers and making connections to ideas in

optimization. Three papers by Guler [2] and Lin [3] and Paquette et al. [4] together with Nesterov's [5] method of estimating sequence are the targets of this report.

We will only cover innovations in the theoretical aspect of Catalyst Acceleration. Applications and specific example algorithms are out of the scope because there are too many papers on the separate topic of variance reduced stochastic algorithms.

# 2 Preliminaries

Throughout the writing, let the ambient space is $\mathbb{R}^n$. The optimization problem is

$$\min_{x\in\mathbb{R}^n} F(x).$$

This section introduces the Nesterov's estimating sequence technique. This technique is fundamental for in Guler's accelerated PPM method and Catalyst meta acceleration in the convex/strongly convex case.

## 2.1 Method of Nesterov's Estimating Sequence

**Definition 2.1 (Nesterov's estimating sequence)** *Let $(\phi_k : \mathbb{R}^n \to \mathbb{R})_{k\geq 0}$ be a sequence of functions. We call this sequence of function a Nesterov's estimating sequence when it satisfies the conditions:*

(i) *There exists another sequence $(x_k)_{k\geq 0}$ such that for all $k \geq 0$ it has $F(x_k) \leq \phi_k^* := \min_x \phi_k^*(x)$.*

(ii) *There exists a sequence of $(\alpha_k)_{k\geq 0}$ where $\alpha_k \in (0,1)$ $\forall k \geq 0$ such that for all $x \in \mathbb{R}^n$ it has $\phi_{k+1}(x) - \phi_k(x) \leq -\alpha_k(\phi_k(x) - F(x))$.*

**Observation 2.2** *If we define $\phi_k$, $\Delta_k(x) := \phi_k(x) - F(x)$ for all $x \in \mathbb{R}^n$ and assume that $F$ has minimizer $x^*$. Then observe that $\forall k \geq 0$:*

$$\begin{aligned} & \Delta_k(x) = \phi_k(x) - F(x) \geq \phi_k^* - F(x) \\ x = x_k \implies & \Delta_k(x_k) \geq \phi_k^* - F(x_k) \geq 0 \\ x = x_* \implies & \Delta_k(x_*) \geq \phi_k^* - F_* \geq F(x_k) - F_* \geq 0 \end{aligned}$$

*The function $\Delta_k(x)$ is non-negative at points: $x_*, x_k$. We can derive the convergence rate of $\Delta_k(x^*)$ because $\forall x \in \mathbb{R}^n$:*

$$\begin{aligned} & \phi_{k+1}(x) - \phi_k(x) \leq -\alpha_k(\phi_k(x) - F(x)) \\ \iff & \phi_{k+1}(x) - F(x) - (\phi_k(x) - F(x)) \leq -\alpha_k(\phi_k(x) - F(x)) \\ \iff & \Delta_{k+1}(x) - \Delta_k(x) \leq -\alpha_k \Delta_k(x) \\ \iff & \Delta_{k+1}(x) \leq (1 - \alpha_k)\Delta_k(x). \end{aligned}$$

*Unrolling the above recursion it yields:*

$$\Delta_{k+1}(x) \leq (1 - \alpha_k)\Delta_k(x) \leq \cdots \leq \left(\prod_{i=0}^{k}(1 - \alpha_i)\right)\Delta_0(x).$$

*Finally, by setting $x = x^*$, $\Delta_k(x^*)$ is non-negative and using the property of Nesterov's estimating sequence it gives:*

$$F(x_k) - F(x^*) \leq \phi_k^* - F(x^*) \leq \Delta_k(x^*) = \phi_k(x^*) - F(x^*) \leq \left(\prod_{i=0}^{k}(1 - \alpha_i)\right)\Delta_0(x^*).$$

Creativity is important in the construction of the estimating sequence $(\phi_k)_{k\geq 0}$.

# 3 Nesterov's accelerated proximal gradient

This section swiftly exposes the constructions of the Nesterov's estimating sequence for accelerated proximal gradient method. A similar accelerated projected gradient is Algorithm (2.2.63) in Nesterov's book [5]. We use accelerated proximal gradient algorithm as an example because its formulation is similar to the Catalyst Acceleration framework.

Throughout this section we assume that: $F = f + g$ where $f$ is $L$-Lipschitz smooth and $\mu \geq 0$ strongly convex and $g$ is convex. Define

$$\begin{aligned} \mathcal{M}^{L^{-1}}(x; y) &:= g(x) + f(y) + \langle \nabla f(x), x - y\rangle + \frac{L}{2}\|x - y\|^2, \\ \widetilde{\mathcal{J}}_{L^{-1}} y &:= \operatorname*{argmin}_x \mathcal{M}^{L^{-1}}(x; y), \\ \mathcal{G}_{L^{-1}}(y) &:= L\left(I - \widetilde{\mathcal{J}}_{L^{-1}}\right) y. \end{aligned}$$

In the literature, $\mathcal{G}_{L^{-1}}$ is commonly known as the gradient mapping. The definition follows, we define the Nesterov's estimating sequence used to derive the accelerated proximal gradient method.

**Definition 3.1 (Accelerated proximal gradient estimating sequence)** *Define $(\phi_k)_{k\geq 0}$ be the Nesterov's estimating sequence recursively given by:*

$$l_F(x; y_k) := F\left(\widetilde{\mathcal{J}}_{L^{-1}} y_k\right) + \langle \mathcal{G}_{L^{-1}} y_k, x - y_k\rangle + \frac{1}{2L}\|\mathcal{G}_{L^{-1}} y_k\|^2,$$
$$\phi_{k+1}(x) := (1-\alpha_k)\phi_k(x) + \alpha_k\left(l_F(x; y_k) + \frac{\mu}{2}\|x - y_k\|^2\right).$$

*The Algorithm generates a sequence of vectors $y_k, x_k$, and scalars $\alpha_k$ satisfies the following:*

$$x_{k+1} = \widetilde{\mathcal{J}}_{L^{-1}} y_k,$$
$$\text{find } \alpha_{k+1} \in (0,1) : \alpha_{k+1} = (1-\alpha_{k+1})\alpha_k^2 + (\mu/L)\alpha_{k+1}$$
$$y_{k+1} = x_{k+1} + \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}(x_{k+1} - x_k).$$

*One of the possible base case can be $x_0 = y_0$ and any $\alpha_0 \in (0,1)$.*

**Observation 3.2** *Fix any $y$, for all $x \in \mathbb{R}^n$, $F(x) \geq l_F(x; y_k) + \mu/2\|x - y_k\|^2$ is the proximal gradient inequality. If $f \equiv 0$ then $\widetilde{\mathcal{J}}_{L^{-1}} y_k$ becomes resolvent $(I + L^{-1}\partial F)^{-1}$, which makes $x_k$ being an exact evaluation of PPM:*

$$\begin{aligned} l_F(x; y_k) &= F(\mathcal{J}_{L^{-1}} y_k) + \langle L(y - \mathcal{J}_{L^{-1}} y), x - y_k\rangle + \frac{L}{2}\|y_k - \mathcal{J}_{L^{-1}} y_k\|^2 \\ &= F(\mathcal{J}_{L^{-1}} y_k) + \langle L(y - \mathcal{J}_{L^{-1}} y), x - \mathcal{J}_{L^{-1}} y_k\rangle. \end{aligned}$$

*This is the proximal inequality with constant a step size: $L^{-1}$.*

To demonstrate the usage of Nesterov's estimating sequence here, consider sequence $(x_k)_{k\geq 0}$ such that $F(x_k) \leq \phi_k^*$. Assume the existence of minimizer $x^*$ for $F$, by definition of $\phi_k$ let $x = x^*$ then $\forall k \geq 0$:

$$\phi_{k+1}(x^*) = (1-\alpha_k)\phi_k(x^*) + \alpha_k\left(l_F(x^*; y_k) + \frac{\mu}{2}\|x^* - y_k\|^2\right)$$
$$\phi_{k+1}(x^*) - \phi_k(x^*) = -\alpha_k\phi_k(x^*) + \alpha_k\left(l_F(x^*; y_k) + \frac{\mu}{2}\|x^* - y_k\|^2\right)$$
$$\implies \phi_{k+1}(x^*) - F(x^*) + F(x^*) - \phi_k(x^*) \leq -\alpha_k(\phi_k(x^*) - F(x^*))$$
$$\implies F(x_{k+1}) - F(x^*) \leq \phi_{k+1}^* - F(x^*) \leq \phi_{k+1}(x^*) - F(x^*) \leq (1-\alpha_k)(\phi_k(x^*) - F(x^*)).$$

On the first inequality we used the fact that $l_F(x; y_k) + \mu/2\|x - y_k\|^2 \leq F(x)$. Unrolling the recurrence, we can get the convergence rate of $F(x_k) - F(x^*)$ to be on Big O of $\prod_{i=1}^k (1-\alpha_i)$.

**Remark 3.3** The definition is a generalization of Nesterov's estimating sequence comes from (2.2.63) from Nesterov's book [5]. Compare to Nesterov's work, we used proximal gradient operator instead of projected gradient. The same inequality is called "Fundamental Proximal Gradient Inequality" in Amir Beck's book [**?**], Theorem 10.16.

For a proof for the Nesterov's estimating sequence $\phi_k$ and a derivation of the algorithm, see Appendix A.1. We warn the readers that the proof is long.

# 4 Guler 1993

This section introduces the setup of the Nesterov's estimating sequence used in Guler's accelerated Proximal Point method. Guler showed Nesterov's estimating sequence technique can accelerate proximal point method from Rockafellar in the convex settings.

Throughout this section, we assume that $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a convex function. We use the following list of notations:

$$\begin{aligned}\mathcal{M}^{\lambda}(x; y) &:= F(x) + \frac{1}{2\lambda}\|x - y\|^2 \\ \mathcal{J}_{\lambda} y &:= \operatorname*{argmin}_x \mathcal{M}^{\lambda}(x; y) \\ \mathcal{G}_{\lambda} &:= \lambda^{-1}(I - \mathcal{J}_{\lambda}).\end{aligned}$$

For notations simplicity, we use $\mathcal{G}_k, \mathcal{J}_k$ to denote the gradient mapping and the proximal point operator because under the context of the algorithm, the proximal point step is conductive iteratively with some arbitrary sequence that $(\lambda)_{k\geq 0}$ which we fixed at the start.

**Definition 4.1 (Accelerated PPM estimating sequence)** *The Nesterov's estimating sequence $(\phi_k)_{k\geq 0}$ for the accelerated proximal point method is defined by the following recurrence for all $k \geq 0$, any $A \geq 0$:*

$$\begin{aligned}\phi_0 &:= f(x_0) + \frac{A}{2}\|x - x_0\|^2, \\ \phi_{k+1}(x) &:= (1 - \alpha_k)\phi_k(x) + \alpha_k(F(\mathcal{J}_k y_k) + \langle \mathcal{G}_k y_k, x - \mathcal{J}_k y_k\rangle).\end{aligned}$$

*Let $(\lambda_k)_{k\geq 0}$ be the step size which defines the descent sequence $x_k = \mathcal{J}_{\lambda} y_k$. Then the descent sequence $x_k$, along with the auxiliary vector sequence $(y_k, v_k)$, scalar sequence $(\alpha_k, A_k)_{k\geq 0}$ will be made to satisfy for all $k \geq 0$, the conditions:*

$$\begin{aligned}\alpha_k &= \frac{1}{2}\left(\sqrt{(A_k\lambda_k)^2 + 4A_k\lambda_k} - A_k\lambda_k\right) \\ y_k &= (1 - \alpha_k)x_k + \alpha_k v_k \\ v_{k+1} &= v_k - \frac{\alpha_k}{A_{k+1}\lambda_k}(y_k - \mathcal{J}_k y_k) \\ A_{k+1} &= (1 - \alpha_k)A_k,\end{aligned}$$

**Remark 4.2** The auxiliary sequences $(A_k, v_k)$ parameterizes a canonical representation of the estimating sequence $(\phi_k)_{k\geq 0}$. Guler didn't simplify his results compare to what Nesterov did in his book.

To handle the inexact evaluation of the PPM, Guler cited Rockafellar [7] for condition (A') in his text which is the following:

$$x_{k+1} \approx \mathcal{J}_k y_k \text{ be such that: } \operatorname{dist}\left(\mathbf{0}, \partial \mathcal{M}^k(x_{k+1}; y_k)\right) \leq \frac{\epsilon_k}{k}$$
$$\implies \|x_{k+1} - \mathcal{J}_k y_k\| \leq \epsilon_k.$$

Condition A' also characterizes the property of $(\epsilon_k)_{k\geq 0}$ so inexact PPM converges. Guler strengthens it in his context and proved the following:

**Theorem 4.3 (Guler's inexact proximal point error bound)**
*Define Moreau Envelope at $y_k$ as $\mathcal{M}_k^* := \min_z \mathcal{M}^{\lambda_k}(z; y_k)$. If $x_{k+1}$ is an inexact evaluation under condition (A'), then the estimating sequence admits the conditions:*

$$\frac{1}{2\lambda_k}\|x_{k+1} - \mathcal{J}_k y_k\|^2 \leq \mathcal{M}_k(x_{k+1}, y_k) - \mathcal{M}_k^* \leq \frac{\epsilon_k^2}{2\lambda_k}.$$

The next theorem is Theorem 3.3 of Guler's 1993 papers which is a major result for inexact accelerated PPM method.

**Theorem 4.4 (Guler's accelerated inexact PPM convergence results)** *If the error sequence $(\epsilon_k)_{k\geq 0}$ for condition A' is bounded by $\mathcal{O}(1/k^\sigma)$ for some $\sigma > 1/2$, then the accelerated proximal point method has for any feasible $x \in \mathbb{R}^n$:*

$$f(x_k) - f(x) \leq \mathcal{O}(1/k^2) + \mathcal{O}(1/k^{2\sigma-1}) \to 0.$$

*When $\sigma \geq 3/2$ then the method converges at a rate of $\mathcal{O}(1/k^2)$.*

The theorem looks exciting, but Lin 2015 [3] page 11 pointed out that $\mathcal{G}_k^*, \mathcal{J}_{\lambda_k} y_k$ are both intractable quantities. In Guler's work, these intractable quantities were built into the Nesterov's estimating sequence making it unclear how to control $\epsilon_k \to 0$. If we use the inexact formulation from Guler and his estimating sequence, it will result in algorithm that contains intractable quantities $\mathcal{J}_{\lambda_k} y_k$.

# 5 Lin 2015

The section introduces the Nesterov's estimating sequence in Lin 2015 [3]. We warn the readers about the followings:

(i) The proofs in HongZhou Lin's original paper of Universal Catalyst is depressingly long and complicated. It is a result of using the constructive approach of Nesterov's estimating sequence.

(ii) Controlling the errors of inexact proximal point evaluations is context specific. Lin hinted at ways to track the errors such as using duality and non-negativity assumption of the objective. He illustrated the use of the meta acceleration on their own method called: "Accelerated MISO-Prox", in general there is not a universal solution.

(iii) We will provide proofs to clarify some of their proofs and compare with existing proofs and drawing references in the literatures in the appendix.

Let's assume $F$ is a $\mu \geq 0$ strongly convex function. Throughout this section we make the following notations

$$\mathcal{M}^{\kappa^{-1}}(x; y) := F(x) + \frac{\kappa}{2}\|x - y\|^2,$$
$$\mathcal{J}_{\kappa^{-1}} y := \operatorname*{argmin}_x \mathcal{M}^{\kappa^{-1}}(x, y).$$

Their algorithm is almost exactly the same as Nesterov's 2.2.20 [5] which we stated in the definition below:

**Definition 5.1 (Lin's accelerated proximal point method)** *Let the initial estimate be $x_0 \in \mathbb{R}^n$, fix parameters $\kappa$ and $\alpha_0$. Let $(\epsilon_k)_{k\geq 0}$ be an error sequence chosen for the evaluation for inexact proximal point method. Initialize $x_0 = y_0$, then the algorithm generates $(x_k, y_k)$ satisfies for all $k \geq 1$*

$$\begin{aligned}
&\text{find } x_k \approx \mathcal{J}_{\kappa^{-1}} y_{k-1} \text{ such that } \mathcal{M}^{\kappa^{-1}}(x_k, y_{k-1}) - \mathcal{M}^{\kappa^{-1}}(\mathcal{J}_{\kappa^{-1}} y_{k-1}, y_{k-1}) \leq \epsilon_k \\
&\text{find } \alpha_k \in (0, 1) \text{ such that } \alpha_k^2 = (1 - \alpha_k)\alpha_{k-1}^2 + (\mu/(\mu + \kappa)) \\
&y_k = x_k + \frac{\alpha_{k-1}(1 - \alpha_{k-1})}{\alpha_{k-1}^2 + \alpha_k}(x_k - x_{k-1}).
\end{aligned}$$

**Remark 5.2** The algorithm is similarity to Definition 3.1. In contrast, it has an inexact proximal point step controlled by $\epsilon_k$, and the Lipschitz constant $L$ is absent instead it has $\kappa + \mu$. Evaluating $x_k \approx \mathcal{J}_{\kappa^{-1}} y_{k-1}$ is also possible because the function $\mathcal{M}^{\kappa^{-1}}(\cdot, y_{k-1})$ is strongly convex, hence its optimality gap can be bounded via trackable quantity $\partial\mathcal{M}^{\kappa^{-1}}(x_k, y_{k-1})$.

Controlling the error sequence $\epsilon_k$ however is a whole new business. Lin 2015 [3] commented on the second last paragraph on page 4, and here we quote:

"The choice of the sequence $(\epsilon_k)_{k\geq 0}$ is also subjected to discussion since the quantity $F(x_0) - F^*$ is unknown beforehand. Nevertheless, an upper bound may be used instead, which will only affects the corresponding constant in (7). Such an upper bounds can typically be obtained by computing a duality gap at $x_0$, or by using additional knowledge about the objective. For instance, when $F$ is non-negative, we may simply choose $\epsilon_k = (2/9)F(x_0)(1 - \rho)^k$".

This comment has upmost practical importance because it tells us how to bound the error $\epsilon_k$ to achieve accelerated convergence rate. In theory, $\epsilon_k$ decreases at a rate related to $F(x_0) - F^*$. It requires some knowledge about $F^*$ in prior. Therefore, controlling $\epsilon_k$ is still elusive in general in a practical context. To see how the error is controlled for the inexact proximal point evaluation, we refer the readers to Lemma B.1 in Lin's 2015 paper [3].

For theoretical interests, there is a major difference between Lin's approach and Guler's approach. Lin didn't formulate any of the intractable quantities in the definitions for his Nesterov's estimating sequence $\phi_k$. One major innovation is Lemma A.7 in Lin's 2015 paper [3]. The lemma allows the analysis Nesterov's estimating sequence to be carried through without using intractable quantities: $\mathcal{M}^{\kappa^{-1}}(\mathcal{J}_{\kappa^{-1}}y_{k-1}, y_{k-1}), \mathcal{J}_{\kappa^{-1}}y_{k-1}$.

**Lemma 5.3 (Lin's inexact proximal inequality)** *Let $F$ be a $\mu \geq 0$ strongly convex. Suppose $x_k$ is an inexact proximal point evaluation of $x_k \approx \mathcal{J}_{\kappa^{-1}}y_{k-1}$ with $\kappa$ fixed. Assume the approximation error is characterized by $\mathcal{M}^{\kappa^{-1}}(x_k; y_{k-1}) - \mathcal{M}^{\kappa^{-1}}(\mathcal{J}_{\kappa^{-1}}y_{k-1}, y_{k-1}) \leq \epsilon_k$. Denote $x_k^* = \mathcal{J}_{\kappa^{-1}}y_{k-1}$ to be the exact evaluation of the proximal point then for all $x$:*

$$F(x) \geq F(x_k) + \kappa\langle y_{k-1} - x_k, x - x_k\rangle + \frac{\mu}{2}\|x - x_k\|^2 + (\kappa + \mu)\langle x_k - x_k^*, x - x_k\rangle - \epsilon_k.$$

**Remark 5.4** The lemma plays a key role because it allows Lin to denote his Nesterov's estimating sequence to be for all $k \geq 0$:

$$\phi_k(x) := (1 - \alpha_{k-1})\phi_{k-1}(x) + \alpha_{k-1}\left(F(x_k) + \kappa\langle y_{k-1} - x_k, x - x_k\rangle + \frac{\mu}{2}\|x - x_k\|^2\right).$$

It is void of intractable quantities.

# 6 Non-convex extension of Catalyst acceleration

The non-convex extension of Catalyst acceleration by Lin 2018 [4] is similar to the convex case in his 2015 paper [3]. The new algorithm handles function with unknown weak convexity constant $\rho$ using a process called Auto Adapt subroutine. They only claimed convergence to stationary point is claimed for the weakly convex objective.

Fix $\kappa$ we use the following notations:

$$\mathcal{M}(x; y) := F(x) + \frac{\kappa}{2}\|x - y\|^2$$
$$\mathcal{J}y := \operatorname*{argmin}_x \mathcal{M}(x; y).$$

We define the algorithm and then its convergence claim below.

**Definition 6.1 (Basic 4WD Catalyst Algorithm)** *Find any $x_0 \in dom(F)$. Initialize the algorithm with $\alpha_1 = 1, v_0 = x_0$. For $k \geq 1$, the iterates $(x_k, y_k, v_k)$ are generated by the procedures:*

$$\textit{find } \bar{x}_k \approx \operatorname*{argmin}_x \{\mathcal{M}(x; x_{k-1})\}$$

$$\textit{such that: } dist(\mathbf{0}, \partial\mathcal{M}(\bar{x}_k; x_{k-1})) \leq \kappa\|\bar{x}_k - x_{k-1}\|, \mathcal{M}(\bar{x}_k; x_{k-1}) \leq F(x_{k-1});$$

$$y_k = \alpha_k v_{k-1} + (1 - \alpha_k)x_{k-1};$$

$$\textit{find } \tilde{x}_k \approx \operatorname*{argmin}_x \{\mathcal{M}(x; y_k)\} \textit{ such that: } \operatorname{dist}(\mathbf{0}, \partial\mathcal{M}(\tilde{x}_k; y_k)) \leq \frac{\kappa}{k+1}\|\tilde{x}_k - y_k\|;$$

$$v_k = x_{k-1} + \frac{1}{\alpha_k}(\tilde{x}_k - x_{k-1});$$

$$\textit{find } \alpha_{k+1} \in (0, 1) : \frac{1 - \alpha_{k+1}}{\alpha_{k+1}^2} = \frac{1}{\alpha_k^2};$$

$$\textit{choose } x_k \textit{ such that: } f(x_k) = \min(f(\bar{x}_k), f(\tilde{x}_k)).$$

**Theorem 6.2 (Basic 4WD Catalyst Convergence)** *Let $(x_k, v_k, y_k)$ be generated by the basic Catalyst algorithm. If $F$ is $\kappa$ weakly convex and bounded below, then $x_k$ converges to a stationary point where*

$$\min_{j=1,\cdots,N} \operatorname{dist}^2(\mathbf{0}, \partial F(\bar{x}_j)) \leq \frac{8\kappa}{N}(F(x_0) - F^*).$$

*And when $F$ is convex, $F(x_k) - F^*$ converges at a rate of $\mathcal{O}(k^{-2})$.*

**Remark 6.3**

# References


[1] O. Guler, *On the convergence of the proximal point algorithm for convex minimization*, 29, p. 17. Num Pages: 17 Place: Philadelphia, United States Publisher: Society for Industrial and Applied Mathematics.

[2] ——, *New proximal point algorithms for convex minimization*, SIAM Journal on Optimization, 2 (1992), pp. 649–664.

[3] H. Lin, J. Mairal, and Z. Harchaoui, *A universal catalyst for first-order optimization*, MIT Press, Dec. 2015, pp. 33–84.

[4] ——, *Catalyst acceleration for first-order convex optimization: from theory to practice*, in Journal of Machine Learning Research, vol. 18, 2018, pp. 1–54.

[5] Y. Nesterov, *Lectures on Convex Optimization*, vol. 137 of Springer Optimization and Its Applications, Springer International Publishing, Cham, 2018.

[6] C. Paquette, H. Lin, D. Drusvyatskiy, J. Mairal, and Z. Harchaoui, *Catalyst for gradient-based nonconvex optimization*, in Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR, Mar. 2018, pp. 613–622.

[7] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.

# A Appendix

## A.1 Theorems and claims for accelerated proximal gradient

Throughout this section, $F = g + f$ is an additive composite objective function with $g$ convex, $f$ $L$-lipschitz smooth and $\mu \geq 0$ strongly convex. The notations here are

$$\begin{aligned}
\mathcal{M}^{L^{-1}}(x;y) &:= F(x) + \frac{L}{2}\|x-y\|^2 \\
\widetilde{\mathcal{M}}^{L^{-1}}(x;y) &:= g(x) + f(y) + \langle \nabla f(x), x-y\rangle + \frac{L}{2}\|x-y\|^2 \\
\widetilde{\mathcal{J}}_{L^{-1}} y &:= \operatorname*{argmin}_x \widetilde{\mathcal{M}}^{L^{-1}}(x;y) \\
\widetilde{\mathcal{G}}_{L^{-1}}(y) &:= L\left(I - \widetilde{\mathcal{J}}_{L^{-1}}\right) y.
\end{aligned}$$

**Theorem A.1 (Fundamental theorem of proximal gradient)** *Let $h = f+g$ and proximal gradient operator $T$ be given as in this section. Fix any $y$, we have for all $x \in \mathbb{R}^n$:*

$$h(x) - h(Ty) - \left\langle L(y - \widetilde{\mathcal{J}}_{L^{-1}} y), x - \widetilde{\mathcal{J}}_{L^{-1}} y \right\rangle \geq D_f(x,y).$$

*Proof.* By a direct observation:

$$\begin{aligned}
\widetilde{\mathcal{M}}^{L^{-1}}(x;y) &= g(x) + f(y) + \langle \nabla f(y), x-y\rangle + \frac{L}{2}\|x-y\|^2 \\
&= g(x) + f(x) - f(x) + f(y) + \langle \nabla f(y), x-y\rangle + \frac{L}{2}\|x-y\|^2 \\
&= h(x) - D_f(x,y) + \frac{L}{2}\|x-y\|^2 \\
&= \mathcal{M}^{L^{-1}}(x;y) - D_f(x,y).
\end{aligned}$$

Next, since $\widetilde{\mathcal{M}}^{L^{-1}}(\cdot, y)$ is strongly convex, it has quadratic growth conditions on its minimizer. Denote $y^+ = \widetilde{\mathcal{J}}_{L^{-1}} y$ then:

$$\begin{aligned}
&\widetilde{\mathcal{M}}^{L^{-1}}(x;y) - \widetilde{\mathcal{M}}^{L^{-1}}(y^+;y) - \frac{L}{2}\|x-y^+\|^2 \geq 0 \\
\implies &\left(\mathcal{M}^{L^{-1}}(x;y) - D_f(x,y)\right) - \mathcal{M}^{L^{-1}}(y^+;y) - \frac{L}{2}\|x-y^+\|^2 \geq 0 \\
\iff &\left(\mathcal{M}^{L^{-1}}(x;y) - \mathcal{M}^{L^{-1}}(y^+;y)\right) - D_f(x,y) - \frac{L}{2}\|x-y^+\|^2 \geq 0 \\
\iff &\left(F(x) - F(y^+) + \frac{L}{2}\|x-y\|^2 - \frac{L}{2}\|y^+-y\|^2\right) - D_f(x,y) - \frac{L}{2}\|x-y^+\|^2 \geq 0 \\
\iff &\left(F(x) - F(y^+) + \frac{L}{2}\left(\|x-y^+ + y^+ - y\|^2 - \|y-y^+\|^2\right)\right) - D_f(x,y) - \frac{L}{2}\|x-y^+\|^2 \geq 0 \\
\iff &\left(F(x) - F(y^+) + \frac{L}{2}\left(\|x-y^+\|^2 + 2\langle x-y^+, y^+-y\rangle\right)\right) - D_f(x,y) - \frac{L}{2}\|x-y^+\|^2 \geq 0 \\
\iff &\left(F(x) - F(y^+) + \frac{L}{2}\|x-y^+\|^2 - L\langle x-y^+, y-y^+\rangle\right) - D_f(x,y) - \frac{L}{2}\|x-y^+\|^2 \geq 0 \\
\iff &\ F(x) - F(y^+) - \langle L(y-y^+), x-y^+\rangle - D_f(x,y) \geq 0.
\end{aligned}$$

■

**Remark A.2** The quadratic growth with respect to minimizer of the Moreau Envelope is used to derive the inequality, please take caution that this condition is strictly weaker than strong convexity of the Moreau Envelope, which could be made weaker than the strong convexity of $F$. Compare the same theorems in older literatures, this proof doesn't use the subgradient inequality, making it appealing for generalizations outside convexity context.

**Theorem A.3 (Cannonical form of proximal gradient estimating sequence)** *Denote $\phi_k : \mathbb{R}^n \to \mathbb{R}$ as a sequence of functions such that it satisfies recursively for all $k \geq 0$ the following conditions*

$$\begin{aligned}
&g_k := L(y_k - \widetilde{\mathcal{J}}_{L^{-1}} y_k) \\
&l_F(x; y_k) := F\left(\widetilde{\mathcal{J}}_{L^{-1}} y_k\right) + \langle g_k, x - y_k\rangle + \frac{1}{2L}\|g_k\|^2, \\
&\alpha_k \in (0,1) \\
&\phi_{k+1}(x) := (1-\alpha_k)\phi_k(x) + \alpha_k(l_F(x;y_k) + \mu/2\|x-y_k\|^2).
\end{aligned}$$

*Where $(y_k)_{k\geq 0}$ is any auxiliary sequence. If we define the canonical form for $\phi_k$ as convex quadratic parameterized by positive sequence $(\gamma_k), \phi_k^*$ and*

$$\begin{aligned}
\phi_k^* &:= \min_x \phi_k(x) \\
\phi_k(x) &:= \phi_k^* + \frac{\gamma_k}{2}\|x - v_k\|^2.
\end{aligned}$$

*Then the auxiliary sequence $y_k, v_k$, parameters for the canonical form of estimating sequence must satisfy for all $k \geq 0$ these inequalities:*

$$\begin{aligned}
\gamma_{k+1} &= (1-\alpha_k)\gamma_k + \mu\alpha_k \\
v_{k+1} &= \gamma_{k+1}^{-1}(\gamma_k(1-\alpha_k)v_k - \alpha_k g_k + \mu\alpha_k y_k) \\
\phi_{k+1}^* &= (1-\alpha_k)\phi_k^* + \alpha_k\left(F\left(\widetilde{\mathcal{J}}_{L^{-1}} y_k\right) + \frac{1}{2L}\|g_k\|^2\right) \\
&\quad - \frac{\alpha_k^2}{2\gamma_{k+1}}\|g_k\|^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|v_k - y_k\|^2 + \langle v_k - y_k, g_k\rangle\right).
\end{aligned}$$

*Proof.* By the recursive definition of $\phi_k$:

$$\begin{aligned}
\phi_{k+1}(x) &= (1-\alpha_k)\phi_k(x) + \alpha_k(l_F(x;y_k) + \mu/2\|x-y_k\|^2) \\
&= (1-\alpha_k)\left(\phi_k^* + \gamma_k/2\|x-v_k\|^2\right) + \alpha_k\left(l_F(x;y_k) + \mu/2\|x-y_k\|^2\right) \to \text{(eqn1)}; \\
\nabla\phi_{k+1}(x) &= (1-\alpha_k)\gamma_k(x-v_k) + \alpha_k(g_k + \mu(x-y_k)); \\
\nabla^2\phi_{k+1}(x) &= \underbrace{((1-\alpha_k)\gamma_k + \alpha_k\mu)}_{=\gamma_{k+1}} I.
\end{aligned}$$

The first recurrence for is discovered as $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu$. Because $v_{k+1}$ is the minimizer of $\phi_{k+1}$ by definition of the canonical form, solving $\nabla\phi_{k+1}(x) = \mathbf{0}$ yields $v_{k+1}$. This is obtained by considering the following:

$$\begin{aligned}
\mathbf{0} &= \gamma_k(1-\alpha_k)(x-v_k) + \alpha_k g_k + \mu\alpha_k(x-y_k) \\
&= (\gamma_k(1-\alpha_k) + \mu\alpha_k)x - \gamma_k(1-\alpha_k)v_k + \alpha_k g_k - \mu\alpha_k y_k \\
\iff v_{k+1} := x &= \gamma_{k+1}^{-1}\left(\gamma_k(1-\alpha_k)v_k - \alpha_k g_k + \mu\alpha_k y_k\right).
\end{aligned}$$

From the second and third equality we used $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu$. Substituting the canonical form of $\phi_{k+1}$ back to (eqn1), choose $x = y_k$, it gives the following:

$$\begin{aligned}
\phi_{k+1}^* &= (1-\alpha_k)\phi_k^* + \frac{(1-\alpha_k)\gamma_k}{2}\|y_k - v_k\|^2 \\
&\quad - \frac{\gamma_{k+1}}{2}\|y_k - v_{k+1}\|^2 + \alpha_k\left(F\left(\widetilde{\mathcal{J}}_{L^{-1}} y_k\right) + \frac{1}{2L}\|g_k\|^2\right) \to \text{(eqn2)}.
\end{aligned}$$

Next move is to simplify the term $\|v_{k+1} - y_k\|^2$. We are doing this in advance so that later on the implicit descent conditions $F(x_{k+1}) \leq \phi_k^*$ will have easier algebra. With that it produces:

$$\begin{aligned}
v_{k+1} - y_k &= \gamma_{k+1}^{-1}\left(\gamma_k(1-\alpha_k)v_k - \alpha_k g_k + \mu\alpha_k y_k\right) - y_k \\
&= \gamma_{k+1}^{-1}\left(\alpha_k(1-\alpha_k)v_k - \alpha_k g_k + (-\gamma_{k+1} + \mu\alpha_k)y_k\right) \\
&\qquad \gamma_{k+1} = (1-\alpha_k)\gamma_k + \mu\alpha_k \\
&\qquad \gamma_{k+1} - \mu\alpha_k = (1-\alpha_k)\gamma_k \\
&= \gamma_{k+1}^{-1}\left(\alpha_k(1-\alpha_k)v_k - \alpha_k g_k(1-\alpha_k)\gamma_k y_k\right) \\
&= \gamma_{k+1}^{-1}(\alpha_k(1-\alpha_k)(v_k - y_k) - \alpha_k g_k).
\end{aligned}$$

Taking the norm of that we have:

$$\begin{aligned}
\|v_{k+1} - y_k\|^2 &= \|\gamma_{k+1}^{-1}(\alpha_k(1-\alpha_k)(v_k - y_k) - \alpha_k g_k)\|^2 \\
\frac{-\gamma_{k+1}}{2}\|v_{k+1} - y_k\|^2 &= -\frac{1}{2\gamma_{k+1}}\|\gamma_k(1-\alpha_k)(v_k - y_k) - \alpha_k g_k\|^2 \\
&= -\frac{\gamma_k^2(1-\alpha_k)^2}{2\gamma_{k+1}}\|v_k - y_k\|^2 - \frac{\alpha_k^2}{2\gamma_{k+1}}\|g_k\|^2 \\
&\quad + \gamma_k(1-\alpha_k)\gamma_{k+1}^{-1}\langle v_k - y_k, \alpha_k g_k\rangle.
\end{aligned}$$

Substitute it back to (eqn2) we have

$$\begin{aligned}
\phi_{k+1}^* &= (1-\alpha)\phi_k^* + \alpha_k\left(F\left(\widetilde{\mathcal{J}}_{L^{-1}}y_k\right) + \frac{1}{2L}\|g_k\|^2\right) \\
&\quad + \frac{(1-\alpha_k)\gamma_k}{2}\|y_k - v_k\|^2 - \frac{\gamma_k^2(1-\alpha_k)^2}{2\gamma_{k+1}}\|v_k - y_k\|^2 - \frac{\alpha_k^2}{2\gamma_{k+1}}\|g_k\|^2 \\
&\quad + \alpha_k\gamma_k(1-\alpha_k)\gamma_{k+1}^{-1}\langle v_k - y_k, g_k\rangle \\
&= (1-\alpha)\phi_k^* + \alpha_k\left(F\left(\widetilde{\mathcal{J}}_{L^{-1}}y_k\right) + \frac{1}{2L}\|g_k\|^2\right) \\
&\quad + \left(\frac{(1-\alpha_k)\gamma_k}{2} - \frac{\gamma_k^2(1-\alpha_k)^2}{2\gamma_{k+1}}\right)\|v_k - y_k\|^2 - \frac{\alpha_k^2}{2\gamma_{k+1}}\|g_k\|^2 \\
&\quad + \alpha_k\gamma_k(1-\alpha_k)\gamma_{k+1}^{-1}\langle v_k - y_k, g_k\rangle \\
&\quad \frac{(1-\alpha_k)\gamma_k}{2} - \frac{\gamma_k^2(1-\alpha_k)^2}{2\gamma_{k+1}} = \frac{(1-\alpha_k)\gamma_k}{2}\left(1 - \frac{\gamma_k(1-\alpha_k)}{\gamma_{k+1}}\right) \\
&\quad\qquad = \frac{(1-\alpha_k)\gamma_k}{2}\left(\frac{\gamma_{k+1} - \gamma_k(1-\alpha_k)}{\gamma_{k+1}}\right) \\
&\quad\qquad = \frac{(1-\alpha_k)\gamma_k}{2}\left(\frac{\mu\alpha_k}{\gamma_{k+1}}\right). \\
\iff &= (1-\alpha)\phi_k^* + \alpha_k\left(F\left(\widetilde{\mathcal{J}}_{L^{-1}}y_k\right) + \frac{1}{2L}\|g_k\|^2\right) \\
&\quad + \frac{(1-\alpha_k)\gamma_k}{2}\left(\frac{\mu\alpha_k}{\gamma_{k+1}}\right)\|v_k - y_k\|^2 - \frac{\alpha_k^2}{2\gamma_{k+1}}\|g_k\|^2 \\
&\quad + \alpha_k\gamma_k(1-\alpha_k)\gamma_{k+1}^{-1}\langle v_k - y_k, g_k\rangle \\
&= (1-\alpha)\phi_k^* + \alpha_k\left(F\left(\widetilde{\mathcal{J}}_{L^{-1}}y_k\right) + \frac{1}{2L}\|g_k\|^2\right) \\
&\quad - \frac{\alpha_k^2}{2\gamma_{k+1}}\|g_k\|^2 + \frac{(1-\alpha_k)\gamma_k\alpha_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|v_k - y_k\|^2 + \langle v_k - y_k, g_k\rangle\right).
\end{aligned}$$

The second and third inequality used the equality $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \mu\alpha_k$. ■

**Theorem A.4 (Verifying the conditions of implicit descent)**
*Let estimating sequence $\phi_k$ and auxiliary sequence $y_k, v_k, \gamma_k, \alpha_k$ be given by Theorem A.3. If for all $k \geq 0$ they verify:*

$$\frac{1}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}} \geq 0,$$
$$\frac{\alpha_k \gamma_k}{\gamma_{k+1}}(v_k - y_k) + (\widetilde{\mathcal{J}}_{L^{-1}} y_k - y_k) = \mathbf{0},$$

*then $\phi_k$ is an estimating sequence that verifies $\forall x \in \mathbb{R}^n, k \geq 0$:*

$$F\left(\widetilde{\mathcal{J}}_{L^{-1}} y_{k-1}\right) \leq \phi_k^*$$
$$\phi_{k+1}(x) - \phi_k(x) \leq -\alpha(\phi_k(x) - F(x)).$$

*Proof.* Inductively assume that $x_k = \widetilde{\mathcal{J}}_{L^{-1}} y_{k-1}$ so $F(x_k) \leq \phi_k^*$. Substituting the $x_k$ into the equation for $\phi_{k+1}$:

$$\begin{aligned}
\phi_{k+1}^* &= (1-\alpha_k)\phi_k^* + \alpha_k \left(F(x_k) + \frac{1}{2L}\|g_k\|^2\right) \\
&\quad - \frac{\alpha_k^2}{2\gamma_{k+1}}\|g_k\|^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|v_k - y_k\|^2 + \langle v_k - y_k, g_k\rangle\right) \\
\implies &\geq (1-\alpha_k)F(x_k) + \alpha_k \left(F(x_k) + \frac{1}{2L}\|g_k\|^2\right) \\
&\quad - \frac{\alpha_k^2}{2\gamma_{k+1}}\|g_k\|^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|v_k - y_k\|^2 + \langle v_k - y_k, g_k\rangle\right) \\
\implies &\geq (1-\alpha_k)F(x_k) + \alpha_k \left(F(x_k) + \frac{1}{2L}\|g_k\|^2\right) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|g_k\|^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\langle v_k - y_k, g_k\rangle.
\end{aligned}$$

The first inequality comes from the inductive hypothesis. The second inequality comes from the non-negativity of the term $\frac{\mu}{2}\|v_k - y_k\|^2$. Now, recall from the fundamental proximal gradient inequality in the convex settings, we have $\forall z \in \mathbb{R}^n$:

$$\begin{aligned}
F(z) &\geq F\left(\widetilde{\mathcal{J}}_{L^{-1}} y_k\right) + \left\langle L(y - \widetilde{\mathcal{J}}_{L^{-1}} y_k), z - \widetilde{\mathcal{J}}_{L^{-1}} y_k\right\rangle + D_f(z, y) \\
&\quad \text{set: } x_{k+1} := \widetilde{\mathcal{J}}_{L^{-1}} y_k \\
&\geq F(x_{k+1}) + \langle g_k, z - x_k\rangle + \frac{\mu}{2}\|z - y\|^2 \\
&= F(x_{k+1}) + \langle g_k, z - y + y - x_k\rangle + \frac{\mu}{2}\|z - y\|^2 \\
&\geq F(x_{k+1}) + \langle g_k, z - y\rangle + \frac{1}{2L}\|g_k\|^2.
\end{aligned}$$

Now we set $z = x_k$ and substitute it back to RHS of $\phi_{k+1}$ which yields:

$$\begin{aligned}
\phi^*_{k+1} &\geq (1-\alpha_k)\left(F(x_{k+1}) + \langle g_k, x_k - y_k\rangle + \frac{1}{2L}\|g_k\|^2\right) \\
&\quad + \alpha_k\left(F(x_{k+1}) + \frac{1}{2L}\|g_k\|^2\right) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|g_k\|^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\langle v_k - y_k, g_k\rangle \\
&\geq F(x_{k+1}) + \left(\frac{1}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right)\|g_k\|^2 + (1-\alpha_k)\left\langle g_k, \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + (x_k - y_k)\right\rangle.
\end{aligned}$$

To assert $\phi^*_{k+1} \geq F(x_k)$, one set of sufficient conditions are

$$\begin{aligned}
\left(\frac{1}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right) &\geq 0 \\
\frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + (x_k - y_k) &= \mathbf{0}.
\end{aligned}$$

Before we finish it, re-arranging should give use the equivalent representations

$$\begin{aligned}
-(\alpha_k\gamma_k\alpha_{k+1}^{-1} + 1)y_k &= -\alpha_k\gamma_k\gamma_{k+1}^{-1}v_k - x_k \\
y_k &= \frac{\alpha_k\gamma_k\gamma_{k+1}^{-1}v_k + x_k}{1 + \alpha_k\gamma_k\gamma_{k+1}^{-1}} \\
&\qquad \gamma_{k+1} + \alpha_k\gamma_k = \gamma_k + \alpha_k\mu \\
&= \frac{\alpha_k\gamma_k v_k + \gamma_{k+1}x_k}{\gamma_k + \alpha_k\mu}.
\end{aligned}$$

And $\alpha_k, \gamma_k$, we have the equivalent representation of

$$\begin{aligned}
1 - \frac{L\alpha_k^2}{\gamma_{k+1}} &\geq 0 \\
1 &\geq L\alpha_k^2/\gamma_{k+1} \\
\gamma_{k+1} &\geq L\alpha_k^2 \\
L\alpha_k^2 &\leq \gamma_{k+1} = (1-\alpha_k)\gamma_k + \mu\alpha_k.
\end{aligned}$$

■

**Definition A.5 (Nesterov's accelerated proximal gradient raw form)** *The accelerated proximal gradient algorithm generates vector iterates $x_k, y_k, v_k$ using auxiliary sequence $\alpha_k, \gamma_k$ such that for all $k \geq 0$ they satisfy conditions:*

$$\begin{aligned}
L\alpha_k^2 &\leq (1-\alpha_k)\gamma_k + \alpha_k\mu = \gamma_{k+1}; \alpha_k \in (0,1), \\
y_k &= (\gamma_k + \alpha_k\mu)^{-1}(\alpha_k\gamma_k v_k + \gamma_{k+1}x_k), \\
x_{k+1} &= \widetilde{\mathcal{J}}_{L^{-1}}y_k \\
v_{k+1} &= \gamma_{k+1}^{-1}\left((1-\alpha_k)\gamma_k v_k + \alpha_k\mu y_k - \alpha_k g_k\right).
\end{aligned}$$

**Theorem A.6 (Intermediate form of accelerated proximal gradient)**
*Let iterates $(x_k, y_k, v_k)$ be given by the raw form of Nesterov's accelerated proximal gradient. If we assume that $L\alpha_k^2 = \gamma_{k+1}$, then it simplifies into the following representation without parameter $\gamma_k$:*

$$\begin{aligned}
y_k &= \left(1 + \frac{L - L\alpha_k}{L\alpha_k - \mu}\right)^{-1} \left(v_k + \left(\frac{L - L\alpha_k}{L\alpha_k - \mu}\right) x_k\right) \\
x_{k+1} &= y_k - L^{-1} g_k \\
v_{k+1} &= \left(1 + \frac{\mu}{L\alpha_k - \mu}\right)^{-1} \left(v_k + \left(\frac{\mu}{L\alpha_k - \mu}\right) y_k\right) - \frac{1}{L\alpha_k} g_k \\
0 &= \alpha_k^2 - \left(\mu/L - \alpha_{k-1}^2\right) \alpha_k - \alpha_{k-1}^2.
\end{aligned}$$

*Here we have $g_k = \widetilde{\mathcal{G}}_{L^{-1}} y_k$.*

*Proof.*

From definition, we have equality: $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k \mu$, so $\gamma_{k+1} + \alpha_k \gamma_k = \gamma_k + \alpha_k \mu$, with that in mind we can simplify the expression for $y_k$ by

$$\begin{aligned}
y_k &= (\gamma_k + \alpha_k \mu)^{-1} (\alpha_k \gamma_k v_k + \gamma_{k+1} x_k) \\
&= (\gamma_{k+1} + \alpha_k \gamma_k)^{-1} (\alpha_k \gamma_k v_k + \gamma_{k+1} x_k) \\
&= \left(\frac{\gamma_{k+1}}{\alpha_k \gamma_k} + 1\right)^{-1} \left(v_k + \frac{\gamma_{k+1}}{\alpha_k \gamma_k} x_k\right) \\
&= \left(1 + \frac{L\alpha_k^2}{\alpha_k L \alpha_{k-1}^2}\right)^{-1} \left(v_k + \frac{L\alpha_k^2}{\alpha_k L \alpha_{k-1}^2} x_k\right) \\
&= \left(1 + \frac{\alpha_k}{\alpha_{k-1}^2}\right)^{-1} \left(v_k + \frac{\alpha_k}{\alpha_{k-1}^2} x_k\right).
\end{aligned}$$

For $v_{k+1}$ we use $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \mu\alpha_k$ which gives us:

$$\begin{aligned}
v_{k+1} &= \gamma_{k+1}^{-1} ((1 - \alpha_k)\gamma_k v_k + \mu \alpha_k y_k) - \alpha_k \gamma_{k+1}^{-1} g_k \\
&= \left((1 - \alpha_k)\gamma_k + \alpha_k \mu\right)^{-1} \left((1 - \alpha_k)\gamma_k v_k + \mu \alpha_k y_k\right) - \alpha_k \gamma_{k+1}^{-1} g_k \\
&= \left(1 + \frac{\alpha_k \mu}{(1 - \alpha_k)\gamma_k}\right)^{-1} \left(v_k + \frac{\alpha_k \mu}{(1 - \alpha_k)\gamma_k} y_k\right) - \alpha_k \gamma_{k+1}^{-1} g_k \\
&= \left(1 + \frac{\alpha_k \mu}{(1 - \alpha_k) L \alpha_{k-1}^2}\right)^{-1} \left(v_k + \frac{\alpha_k \mu}{(1 - \alpha_k) L \alpha_{k-1}^2} y_k\right) - \frac{1}{L\alpha_k} g_k
\end{aligned}$$

We can eliminate the $\gamma_k$ which defines the $\alpha_k$ by considering

$$\begin{aligned} L\alpha_k^2 &= (1-\alpha_k)\gamma_k + \alpha_k\mu \\ &= (1-\alpha_k)L\alpha_{k-1}^2 + \alpha_k\mu \\ L\alpha_k^2 &= L\alpha_{k-1}^2 + (\mu - L\alpha_{k-1}^2)\alpha_k \\ \iff 0 &= L\alpha_k^2 - (\mu - L\alpha_{k-1}^2)\alpha_k - L\alpha_{k-1}^2. \end{aligned}$$

Next, we simplify the coefficients using the above relations further. From the above results we have the relation $(1-\alpha_k)L\alpha_{k-1}^2 = L\alpha_k^2 - \alpha_k\mu$. Therefore, it gives

$$\frac{\alpha_k\mu}{(1-\alpha_k)L\alpha_{k-1}^2} = \frac{\alpha_k\mu}{L\alpha_k^2 - \alpha_k\mu} = \frac{\mu}{L\alpha_k - \mu}.$$

Next we have:

$$\begin{aligned} L\alpha_k^2 &= (1-\alpha_k)L\alpha_{k-1}^2 + \alpha_k\mu \\ L\alpha_k^2 - \alpha_k\mu &= (1-\alpha_k)L\alpha_{k-1}^2 \\ \alpha_{k-1}^2 &= \frac{L\alpha_k^2 - \alpha_k\mu}{L(1-\alpha_k)} \\ \frac{1}{\alpha_{k-1}^2} &= \frac{L(1-\alpha_k)}{L\alpha_k^2 - \alpha_k\mu} \\ \frac{\alpha_k}{\alpha_{k-1}^2} &= \frac{L - L\alpha_k}{L\alpha_k - \mu}. \end{aligned}$$

Substitute these results back to the expression for $y_k, v_{k+1}$, it gives what we want.

■

**Remark A.7** This intermediate form representation of the algorithm eliminated the sequence $(\gamma_k)_{k\geq 0}$ which were used for the Nesterov's estimating sequence.

**Theorem A.8 (Nesterov's accelerated proximal gradient momentum form)** *Let the sequence $\alpha_k$, and vectors $y_k, x_k, v_k$ be given by the intermediate form of the Nesterov's accelerated proximal gradient, then it can be simplified to void of $v_k$. The algorithm generates $y_k, x_k, \alpha_k$ such that it satisfies for all $k \geq 0$:*

$$\begin{aligned} &\textit{find } \alpha_{k+1} \textit{ such that: } L\alpha_{k+1}^2 = (1-\alpha_{k+1})L\alpha_{k-1} + \mu\alpha_{k+1} \\ &x_{k+1} = \widetilde{\mathcal{J}}_{L^{-1}} y_k \\ &y_{k+1} = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}(x_{k+1} - x_k). \end{aligned}$$

*Initially we choose $x_0 = y_0, \alpha_0 \in (0,1)$.*

*Proof.* To show that, we write down the intermediate form with new symbols to make it easier to read:

$$\begin{aligned} y_k &= (1+\tau_k)^{-1}(v_k + \tau_k x_k), \\ v_{k+1} &= (1+\xi_k)^{-1}(v_k + \xi_k y_k) - (1+\xi_k)^{-1}\delta_k g_k, \\ x_{k+1} &= y_k - L^{-1} g_k. \end{aligned}$$

Where for all $k \geq 0$:

$$\begin{aligned} \tau_k = \frac{L(1-\alpha_k)}{L\alpha_k - \mu}, \xi_k = \frac{\mu}{L\alpha_k - \mu}, \\ (1+\xi_k)^{-1}\delta_k = \frac{1}{L\alpha_k} \iff L\delta_k = \frac{1+\xi_k}{\alpha_k}, \\ L\alpha_k^2 = (1-\alpha_k)L\alpha_{k-1}^2 + \mu\alpha_k. \end{aligned}$$

Next, we show that if $L\delta_k = 1 + \xi_k + \tau_k$ then $v_{k+1} - x_{k+1} = (1+\xi_t)^{-1}\tau_t(x_{k+1} - x_k)$.

$$\begin{aligned} v_{k+1} &= (1+\xi_k)^{-1}(v_k + \xi_k y_k) - (1+\xi_k)^{-1}\delta_k g_k \\ &= (1+\xi_k)^{-1}((1+\tau_k)y_k - \tau_k x_k + \xi_k y_k) - (1+\xi_k)^{-1}\delta_k g_k \\ &= (1+\xi_k)^{-1}((1+\tau_k+\xi_k)y_k - \tau_k x_k) - (1+\xi_k)^{-1}\delta_k g_k \\ \iff v_{k+1} - x_{k+1} &= (1+\xi_k)^{-1}((1+\tau_k+\xi_k)y_k - \tau_k x_k - \delta_k g_k) - y_k + L^{-1} g_k \\ &= (1+\xi_k)^{-1}(\tau_k y_k - \tau_k x_k - \delta_k g_k) + L^{-1} g_k \\ &= (1+\xi_k)^{-1}(\tau_k y_k - \tau_k x_k + (L^{-1}(1+\xi_k) - \delta_k) g_k) \\ &= (1+\xi_k)^{-1}\tau_k(y_k - x_k + \tau_k^{-1}(L^{-1} + L^{-1}\xi_k - \delta_k) g_k) \end{aligned}$$

Next, consider $x_{k+1} - x_k$:

$$x_{k+1} - x_k = y_k - x_k - L^{-1} g_k.$$

Observe that, if we substitute $\delta_k = L^{-1}(1+\xi_k) + L^{-1}\tau_k$, then

$$\begin{aligned} v_{k+1} - x_{k+1} &= (1+\xi_k)^{-1}\tau_k(y_k - x_k + \tau_k^{-1}(-L^{-1}\tau_k) g_k) \\ &= (1+\xi_k)^{-1}\tau_k(y_k - x_k - L^{-1} g_k) \\ &= (1+\xi_k)^{-1}\tau_k(x_{k+1} - x_k). \end{aligned}$$

Next, it remains to verify that $L\delta_k = 1 + \xi_k + \tau_k$ is true. This is true because by definitions

the RHS:

$$\begin{aligned}
1+\tau_k+\xi_k &= 1+\frac{L(1-\alpha_k)}{L\alpha_k-\mu}+\frac{\mu}{L\alpha_k-\mu} \\
&= 1+\frac{L-L\alpha_k+\mu}{L\alpha_k-\mu} \\
&= \frac{L-L\alpha_k+\mu+L\alpha_k-\mu}{L\alpha_k-\mu} \\
&= \frac{L}{L\alpha_k-\mu}.
\end{aligned}$$

And the LHS:

$$\frac{1+\xi_k}{\alpha_k} = \frac{1+\frac{\mu}{L\alpha_k-\mu}}{\alpha_k} = \frac{\frac{L\alpha_k-\mu+\mu}{L\alpha_k-\mu}}{\alpha_k} = \frac{L}{L\alpha_k-\mu}.$$

They are matched. Substitute $\xi_k, \tau_k$ into $v_{k+1} = x_{k+1} + (1+\xi_k)^{-1}\gamma_k(x_{k+1}-x_k)$:

$$\begin{aligned}
v_{k+1} &= x_{k+1} + \left(1+\frac{\mu}{L\alpha_k-\mu}\right)^{-1}\left(\frac{L(1-\alpha_k)}{L\alpha_k-\mu}\right)(x_{k+1}-x_k) \\
&= x_{k+1} + \left(\frac{L\alpha_k}{L\alpha_k-\mu}\right)^{-1}\left(\frac{L(1-\alpha_k)}{L\alpha_k-\mu}\right)(x_{k+1}-x_k) \\
&= x_{k+1} + \left(\frac{L\alpha_k-\mu}{L\alpha_k}\right)\left(\frac{L-L\alpha_k}{L\alpha_k-\mu}\right)(x_{k+1}-x_k) \\
&= x_{k+1} + \left(\alpha_k^{-1}-1\right)(x_{k+1}-x_k).
\end{aligned}$$

With $v_{k+1}$ rid of $v_k$, the next step is to make $y_{k+1}$ only using $x_k, x_{k+1}$. By definition $y_{k+1}$ is produced by:

$$\begin{aligned}
y_k &= \left(1+\frac{L-L\alpha_k}{L\alpha_k-\mu}\right)^{-1}\left(v_k+\left(\frac{L-L\alpha_k}{L\alpha_k-\mu}\right)x_k\right) \\
&= \left(\frac{L-\mu}{L\alpha_k-\mu}\right)^{-1}\left(v_k+\left(\frac{L-L\alpha_k}{L\alpha_k-\mu}\right)x_k\right) \\
&= \frac{L\alpha_k-\mu}{L-\mu}v_k + \frac{L-L\alpha_k}{L-\mu}x_k.
\end{aligned}$$

Increment $k$ into $k+1$ then

$$\begin{aligned}
v_{k+1} &= x_{k+1} + (\alpha_k^{-1} - 1)(x_{k+1} - x_k) \\
(L\alpha_{k+1} - \mu)v_{k+1} &= (L\alpha_{k+1} - \mu)x_{k+1} + (L\alpha_{k+1} - \mu)(\alpha_k^{-1} - 1)(x_{k+1} - x_k), \\
y_{k+1} &= (L-\mu)^{-1}((L\alpha_{k+1} - \mu)v_{k+1} + (L - L\alpha_{k+1})x_{k+1}) \\
&= (L-\mu)^{-1}\left((L\alpha_{k+1} - \mu)x_{k+1} + (L\alpha_{k+1} - \mu)(\alpha_k^{-1} - 1)(x_{k+1} - x_k) + (L - L\alpha_{k+1})x_{k+1}\right) \\
&= (L-\mu)^{-1}\left((L-\mu)x_{k+1} + (L\alpha_{k+1} - \mu)(\alpha_k^{-1} - 1)(x_{k+1} - x_k)\right) \\
&= x_{k+1} + \frac{(L\alpha_{k+1} - \mu)(\alpha_k^{-1} - 1)}{L-\mu}(x_{k+1} - x_k).
\end{aligned}$$

We are closer than ever to proving it. This representation contains $L, \mu$ on the momentum coefficients, to get rid of that consider:

$$\begin{aligned}
\frac{(L\alpha_{k+1} - \mu)(\alpha_k^{-1} - 1)}{L-\mu} &= \frac{(L\alpha_{k+1} - \mu)\alpha_k(1-\alpha_k)}{\alpha_k^2(L-\mu)} \\
&= \alpha_k(1-\alpha_k)\left(\frac{\alpha_k^2(L-\mu)}{L\alpha_{k+1} - \mu}\right)^{-1} \\
&= \alpha_k(1-\alpha_k)\left(\frac{L\alpha_k^2 - \mu\alpha_k^2}{L\alpha_{k+1} - \mu}\right)^{-1} \\
&= \alpha_k(1-\alpha_k)\left(\frac{(L\alpha_{k+1} - \mu)(\alpha_k^2 + \alpha_{k+1})}{L\alpha_{k+1} - \mu}\right)^{-1} \\
&= \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}.
\end{aligned}$$

Here, on the third to the 4th equality, we used $L\alpha_{k+1}^2 = (1-\alpha_{k+1})L\alpha_k^2 + \mu\alpha_{k+1}$ in this way:

$$\begin{aligned}
(L\alpha_{k+1} - \mu)(\alpha_k^2 + \alpha_{k+1}) &= L\alpha_{k+1}\alpha_k^2 - \mu\alpha_k^2 + L\alpha_{k+1}^2 + \mu\alpha_{k+1} \\
&= L\alpha_{k+1}\alpha_k^2 - \mu\alpha_k^2 + ((1-\alpha_{k+1})L\alpha_k^2 - \mu\alpha_{k+1}) - \mu\alpha_{k+1} \\
&= L\alpha_k^2 - \mu\alpha_k^2.
\end{aligned}$$

■

**Remark A.9** This proof is very long because we took a detour to an intermediate form without the $\gamma_k$ that the writer personally prefer more than a direct path.

# B Proofs for accelerated PPM

## B.1 Exact accelerated PPM

This section contains all the proofs for the exact accelerated PPM method. It follows the same notation from the accelerated PPM section that $F$ is a convex function and the notations are $\mathcal{J}_k, \mathcal{G}_k$ for the proximal point evaluation and gradient mapping. Recall the definition of the estimating sequence is:

$$\begin{aligned}\phi_0 &:= f(x_0) + \frac{A}{2}\|x - x_0\|^2, \\ \phi_{k+1}(x) &:= (1-\alpha_k)\phi_k(x) + \alpha_k(F(\mathcal{J}_k y_k) + \langle \mathcal{G}_k y_k, x - \mathcal{J}_k y_k\rangle).\end{aligned}$$

Observe $\phi_k$ is a sequence of simple quadratic functions. We define the canonical representation to be:

$$(\forall k \geq 0) \quad \phi_k(x) = \phi_k^* + \frac{A_k}{2}\|x - v_k\|^2.$$

Substituting the canonical form, we obtained a recursive definition of the Hessian and gradient of the estimating sequence:

$$\begin{aligned}\phi_{k+1}^* + \frac{A_{k+1}}{2}\|x - v_{k+1}\|^2 &= (1-\alpha_k)\left(\phi_k^* + \frac{A_k}{2}\|x - v_k\|^2\right) \\ &\quad + \alpha_k(F(\mathcal{J}_k y_k) + \langle \mathcal{G}_k y_k, x - \mathcal{J}_k y_k\rangle) \\ \implies \begin{cases} A_{k+1} = (1-\alpha_k)A_k, \\ \nabla\phi(x) = (1-\alpha_k)A_k(x - v_k) + \alpha_k \mathcal{G}_k y_k. \end{cases}\end{aligned}$$

In the canonical form, $v_{k+1}$ is the minimizer of $\phi_{k+1}$, it can be solved for by setting the gradient $\phi_{k+1}(x) = \mathbf{0}$ so for all $k \geq 0$:

$$\begin{aligned}\mathbf{0} &= (1-\alpha_k)A_k(v_{k+1} - v_k) + \alpha_k \mathcal{G}_k y_k \\ v_{k+1} - v_k &= \frac{\alpha_k}{\lambda_k(1-\alpha_k)A_k}(y_k - \mathcal{J}_k y_k) \\ &= \frac{\alpha_k}{\lambda_k A_{k+1}}(y_k - \mathcal{J}_k y_k).\end{aligned}$$

**Theorem B.1 (Estimating sequence for accelerated PPM)** *The parameters for the*

*estimating sequence:* $(\phi_k^*)_{k\geq 0}, (v_k)_{k\geq 0}, (A_k)_{k\geq 0}$ *satisfies for all* $k \geq 0$ *the following conditions:*

$$\begin{aligned} A_{k+1} &= (1-\alpha_k)A_k, \\ v_{k+1} - v_k &= -\frac{\alpha_k}{A_{k+1}\lambda_k}(y_k - \mathcal{J}_k y_k), \\ \phi_{k+1}^* &\geq F(\mathcal{J}_k y_k) + \frac{1}{2\lambda_k}\left(2 - \frac{\alpha_k^2}{A_{k+1}\lambda_k}\right)\|y_k - \mathcal{J}_k y_k\|^2 \\ &\quad + \lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, (1-\alpha_k)x_k + \alpha_k v_k - y_k\rangle. \end{aligned}$$

*Additionally, the sequence* $\alpha_k$ *has* $\forall k \geq 0$*:*

$$\alpha_k = \frac{1}{2}\left(\sqrt{(A_k\lambda_k)^2 + 4A_k\lambda_k} - A_k\lambda_k\right).$$

*Proof.* Using induction, we assume the inductive hypothesis: $\phi_k^* \geq f(x_k)$ for the sequence $(x_i)_{i\geq 0}$ up to and including $i = k$. Proceed with the inductive hypothesis we have

$$\phi_k^* \geq F(x_k) \geq F(\mathcal{J}_k y_k) + \langle \mathcal{G}_k y_k, x_k - \mathcal{J}_k y_k\rangle.$$

We used the proximal inequality of $F$. Inductively using the definition of the estimating sequence and substitute the canonical form we will have

$$\begin{aligned} \phi_{k+1}^* &= \phi_{k+1}(v_{k+1}) \\ &= (1-\alpha_k)\phi_k(v_{k+1}) + \alpha_k F(\mathcal{J}_k y_k) + \alpha_k\lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, v_{k+1} - y_{k+1}\rangle \\ &= (1-\alpha_k)\left(\phi_k^* + \frac{A_k}{2}\|v_{k+1} - v_k\|^2\right) + \alpha_k F(\mathcal{J}_k y_k) + \alpha_k\lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, v_{k+1} - y_{k+1}\rangle \\ A_{k+1} &= (1-\alpha_k)A_k \\ &= (1-\alpha_k)\phi_k^* + \frac{A_{k+1}}{2}\|v_{k+1} - v_k\|^2 + \alpha_k F(\mathcal{J}_k y_k) + \alpha_k\lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, v_{k+1} - y_{k+1}\rangle \end{aligned}$$

Next, we substitute the inequality by the inductive hypothesis which gives:

$$\begin{aligned} \phi_{k+1}^* &\geq (1-\alpha_k)\left(F(\mathcal{J}_k y_k) + \langle \mathcal{G}_k y_k, x_k - \mathcal{J}_k y_k\rangle\right) \\ &\quad + \frac{A_{k+1}}{2}\|v_{k+1} - v_k\|^2 + \alpha_k F(\mathcal{J}_k y_k) + \alpha_k\lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, v_{k+1} - y_{k+1}\rangle \\ &= F(\mathcal{J}_k y_k) + \frac{A_{k+1}}{2}\|v_{k+1} - v_k\|^2 + \lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, (1-\alpha_k)(x_k - \mathcal{J}_k y_k) + \alpha_k(v_{k+1} - \mathcal{J}_k y_k)\rangle \\ &= F(\mathcal{J}_k y_k) + \frac{A_{k+1}}{2}\|v_{k+1} - v_k\|^2 + \lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, (1-\alpha_k)x_k + \alpha_k v_{k+1} - \mathcal{J}_k y_k\rangle. \end{aligned}$$

This requires further simplifications. Rearranging the elements in the inner product we have:

$$(1-\alpha_k)x_k + \alpha_k v_{k+1} - \mathcal{J}_k y_k = ((1-\alpha_k)x_k + \alpha_k v_k - y_k) + \alpha_k(v_{k+1} - v_k) + (y_k - \mathcal{J}_k y_k).$$

We also use the equality:

$$\begin{aligned} v_{k+1} - v_k &= -\frac{\alpha_k}{A_{k+1}\lambda_k}(y_k - \mathcal{J}_k y_k) \\ \implies \|v_{k+1} - v_k\|^2 &= \left\| -\frac{\alpha_k}{A_{k+1}\lambda_k}\left(y_k - \mathcal{J}_k y_k\right) \right\|^2 \\ \|v_{k+1} - v_k\|^2 &= \left(\frac{\alpha_k}{A_{k+1}\lambda_k}\right)^2 \|y_k - \mathcal{J}_k y_k\|^2 \\ \frac{A_{k+1}}{2}\|v_{k+1} - v_k\|^2 &= \frac{\alpha_k^2}{2A_{k+1}\lambda_k^2}\|y_k - \mathcal{J}_k y_k\|^2 . \end{aligned}$$

Substituting both equality we simplify the inequality into

$$\begin{aligned} \phi_{k+1}^* &\geq F(\mathcal{J}_k y_k) + \frac{\alpha_k^2}{2A_{k+1}\lambda_k^2}\|y_k - \mathcal{J}_k y_k\|^2 \\ &\quad + \lambda_k^{-1}\left\langle y_k - \mathcal{J}_k y_k, (1-\alpha_k)x_k + \alpha_k v_k - y_k + \alpha_k(v_{k+1} - v_k) + (y_k - \mathcal{J}_k y_k)\right\rangle \\ &= F(\mathcal{J}_k y_k) + \frac{\alpha_k^2}{2A_{k+1}\lambda_k^2}\|y_k - \mathcal{J}_k y_k\|^2 \\ &\quad + \lambda_k^{-1}\left\langle y_k - \mathcal{J}_k y_k, (1-\alpha_k)x_k + \alpha_k v_k - y_k\right\rangle \\ &\quad + \lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, \alpha_k(v_{k+1} - v_k) + (y_k - \mathcal{J}_k y_k)\rangle \end{aligned} \tag{1}$$

Simplifying the second cross term on the RHS of (1):

$$\begin{aligned} &\lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, \alpha_k(v_{k+1} - v_k) + (y_k - \mathcal{J}_k y_k)\rangle \\ &= \lambda_k^{-1}\left\langle y_k - \mathcal{J}_k y_k, -\frac{\alpha_k^2}{A_{k+1}\lambda_k}(y_k - \mathcal{J}_k y_k) + (y_k - \mathcal{J}_k y_k)\right\rangle \\ &= \lambda_k^{-1}\left\langle y_k - \mathcal{J}_k y_k, -\frac{\alpha_k^2}{A_{k+1}\lambda_k}(y_k - \mathcal{J}_k y_k) + (y_k - \mathcal{J}_k y_k)\right\rangle \\ &= \lambda_k^{-1}\left(1 - \frac{\alpha_k^2}{A_{k+1}\lambda_k}\right)\|y_k - \mathcal{J}_k y_k\|^2. \end{aligned}$$

The above term repeats with one of the term in (1), merging their coefficient it yields

$$\begin{aligned} &\lambda_k^{-1}\left(1 - \frac{\alpha_k^2}{A_{k+1}\lambda_k}\right) + \frac{\alpha_k^2}{2A_{k+1}\lambda_k^2} \\ &= \lambda_k^{-1} - \frac{\alpha_k^2}{A_{k+1}\lambda_k^2} + \frac{\alpha_k^2}{2A_{k+1}\lambda_k^2} \\ &= \frac{1}{2\lambda_k}\left(2 - \frac{\alpha_k^2}{2A_{k+1}\lambda_k}\right). \end{aligned}$$

Substituting back to (1):

$$\begin{aligned}\phi_{k+1}^* &\geq F(\mathcal{J}_k y_k) + \frac{\alpha_k^2}{2A_{k+1}\lambda_k^2}\|y_k - \mathcal{J}_k y_k\|^2 \\ &\quad + \lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, (1-\alpha_k)x_k + \alpha_k v_k - y_k\rangle \\ &\quad + \lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, \alpha_k(v_{k+1} - v_k) + (y_k - \mathcal{J}_k y_k)\rangle \\ &= F(\mathcal{J}_k y_k) + \frac{1}{2\lambda_k}\left(2 - \frac{\alpha_k^2}{A_{k+1}\lambda_k}\right)\|y_k - \mathcal{J}_k y_k\|^2 \\ &\quad + \lambda_k^{-1}\langle y_k - \mathcal{J}_k y_k, (1-\alpha_k)x_k + \alpha_k v_k - y_k\rangle .\end{aligned}$$

Next, if induction hypothesis $\phi_{k+1} \geq F(\mathcal{J}_k y_k) = F(x_{k+1})$ is true, it's sufficient to take the coefficient of $\|y_k - \mathcal{J}_k y_k\|^2$ to be greater than zero and make the inner product term zero which produces:

$$\begin{aligned}y_k &= (1-\alpha_k)x_k + \alpha_k v_k \\ \frac{1}{2\lambda_k}&\left(2 - \frac{\alpha_k^2}{A_{k+1}\lambda_k}\right) \geq 0.\end{aligned}$$

Solving, the second inequality is equivalent to

$$\begin{aligned}\frac{\alpha_k^2}{A_{k+1}\lambda_k} &\leq 2 \\ \alpha_k &\leq \sqrt{2A_{k+1}\lambda_k} = \sqrt{2A_k(1-\alpha_k)\lambda_k}.\end{aligned}$$

Using $\alpha_k^2 = A_k(1-\alpha_k)\lambda_k$ to solving the quadratic:

$$\alpha_k = \frac{1}{2}\left(\sqrt{(A_k\lambda_k)^2 + 4A_k\lambda_k} - A_k\lambda_k\right).$$

■

**Remark B.2** The approach by Guler is slightly different compare to Accelerated Proximal Gradient method completed in Section A.1. For the Accelerated Proximal Gradient, we simplified $\phi_{k+1}^*$ prior to considering $F(x_{k+1}) \leq \phi_{k+1}^*$.

## B.2 Inexact accelerated PPM

This section proofs the lemma that characterizes the inexact PPM errors for Guler's accelerated inexact PPM. Now, we prove Theorem 4.3.

Denote $\mathcal{M}_k(x) = \mathcal{M}_k(x; y_k)$ for short. The proof is direct by considering the strong convexity of $\mathcal{M}_k(\cdot, y_k)$ together with the subgradient inequality. Choose any $w_k \in \partial\mathcal{M}_k(\mathcal{J}_k y_k)$ it has

$$\begin{aligned}\mathcal{M}_k(x_{k+1}) - \mathcal{M}_k^* &= \mathcal{M}_k(x_{k+1}) - \mathcal{M}_k(\mathcal{J}_k y_k) \\ &\geq \left(\langle w_k, x_{k+1} - \mathcal{J}_k y_k\rangle + \mathcal{M}_k(\mathcal{J}_k y_k) + \frac{1}{2\lambda_k}\|x_{k+1} - \mathcal{J}_k y_k\|^2\right) - \mathcal{M}_k(\mathcal{J}_{\lambda_k} y_k) \\ &= \frac{1}{2\lambda_k}\|x_{k+1} - \mathcal{J}_k y_k\|^2.\end{aligned}$$

We used $\mathbf{0} \in \partial M_k(\mathcal{J}_{\lambda_k} y_k)$ to get rid of the inner product. Consider inexact evaluation of $x_{k+1}$ which results in $w_k \in \partial\mathcal{M}_k(x_{k+1})$ with $\|w_k\| \leq \epsilon_k/\lambda_k$. By $\lambda_k^{-1}$ strong convexity of $\mathcal{M}_k$, we have

$$\begin{aligned}\mathcal{M}_k(\mathcal{J}_k y_k) - \mathcal{M}_k(x_{k+1}) &\geq \langle w_k, \mathcal{J}_k y_k - x_{k+1}\rangle + \frac{1}{2\lambda_k}\|\mathcal{J}_k y_k - x_{k+1}\|^2 \\ &\geq -\|w_k\|\|\mathcal{J}_k y_k - x_{k+1}\|^2 + \frac{1}{2\lambda_k}\|\mathcal{J}_k y_k - x_{k+1}\|^2 \\ &\geq -\frac{\epsilon_k}{\lambda_k}\|\mathcal{J}_k y_k - x_{k+1}\|^2 + \frac{1}{2\lambda_k}\|\mathcal{J}_k y_k - x_{k+1}\|^2 \\ &\geq \frac{1}{\lambda_k}\min_{t\in\mathbb{R}}\left\{\frac{1}{2}t^2 - \epsilon_k t\right\} = -\frac{\epsilon_k}{2\lambda_k}.\end{aligned}$$

The upper bound is proved.

# C Proofs for Catalyst Meta Acceleration

In this section, we prove the inexact proximal inequality and give the Nesterov's estimating sequence void of intractable quantities. The notation follows from Section 5. Let's now prove the inexact proximal inequality. Recall inexact evaluation $x_k \approx \mathcal{J}_{\kappa^{-1}} y_{k-1}$ such that $\mathcal{M}^{\kappa^{-1}}(x_k; y_{k-1}) - \mathcal{M}^{\kappa^{-1}}(\mathcal{J}_{\kappa^{-1}} y_{k-1}; y_{k-1}) \leq \epsilon_k$; $x_k^* = \mathcal{J}_{\kappa^{-1}} y_{k-1}$ to be the exact evaluation.

# D Proofs for 4WD Catalyst Acceleration

We prove Theorem 6.2 in this section. The lemma below characterize the lower and upper bound on the $(\alpha_k)_{k\geq 0}$ which ultimately control the convergence rate.

**Lemma D.1** *app:lemma:momentum-sequence-bounds*