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

Hongda Li

December 6, 2024

Abstract

Nesterov's accelerated gradient sparked numerous theoretical and practical advancements in Mathematics programming literatures since its conception back in 1983. 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 the Catalyst Acceleration method which is a modern practical realization of 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 also describe key aspects of the theoretical innovations involved to achieve the design of the algorithm in convex, and non-convex case.

1 Introduction

The optimal algorithm named accelerated gradient descent method is proposed in Nesterov's seminal work back in 1983 [7]. 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 [9] 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 this 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) [4] 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 [3] showed the convergence of PPM method in the convex case 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 [11] 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 ϵ_k has the following characterizations of inexactness as put forward by Guler [4]:

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

$\partial\mathcal{M}(\cdot;y)$ is the regular subgradient of the model function with respect to x. The difficulty comes from controlling the error ϵ 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. [5, 6] 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. [10]). 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 surveys historical major results and innovations in the conception and design of 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 by making connections to ideas in optimization. Three papers by Guler [4] and Lin [5] and Paquette et al. [6] together with Nesterov's [9] 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. If the readers are interested, consult Gower et al. [1] for more information about the usual candidates used to evaluate the inexact proximal point method in practice.

2 Preliminaries

Throughout the writing, let the ambient space be $\mathbb{R}^n$. 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. Unlike proofs using a Lyapunov argument which requires knowing exactly the algorithm in advance to verify the convergence rate, Nesterov's estimating sequence can produce an algorithm that generates iterates $(x_k)_{k \geq 0}$ such that it converges at some rate.

The technique is universal in deriving accelerated algorithm in the convex case. In addition to Catalyst and accelerated PPM, it's used to derive an accelerated mirror descent in Nesterov's book [9] (6.1.19); an accelerated regularized newton method for convex composite objectives by Geovani N. et al [2] and an accelerated regularized cubic newton by Nesterov [8].

2.1 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 ϕ_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 Estimating sequence for 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 [9]. 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

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

$\mathcal{G}_{L^{-1}}$ is commonly known as the gradient mapping in the literature. We define the Nesterov's estimating sequence used to derive the accelerated proximal gradient method next.

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

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

The Algorithm generates a sequence of vectors y_k, x_k, and scalars α_k satisfies the following:

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

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 ϕ_k let

$x = x^*$ then $\forall k \geq 0$:

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

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 [9]. Compare to Nesterov's work, we used proximal gradient operator instead of projected gradient.

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

4 Guler's estimating sequence

This section introduces the setup of the estimating sequence in Guler's accelerated Proximal Point method [4]. Guler showed the technique can accelerate proximal point method in the convex settings. Throughout the section, we assume that $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a convex function. Define:

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

For simplicity, we use $\mathcal{G}_k, \mathcal{J}_k$ as short for $\mathcal{G}_{\lambda_k}, \mathcal{J}_{\lambda_k}$ where $(\lambda_k)_{k\geq 0}$ is a fixed sequence used in the proximal point on the k iteration.

Definition 4.1 (Accelerated PPM estimating sequence)
The 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 $A \geq 0$:

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

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

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

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 comparing to Nesterov's result which is in his book. For a detailed proof of the estimating sequence with comparison to the accelerated proximal gradient method, see Appendix B.1. We note that the smoothness conditions here are not involved in the definition of Guler's estimating sequence, or the function F.

Guler cited Rockafellar [11] for condition (A') in his text for inexact proximal evaluation:

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

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:

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

Remark 4.4 For a proof of the theorem, see Appendix B.2.

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

Theorem 4.5 (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 [5] 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's estimating sequence

The section introduces the Nesterov's estimating sequence in Lin 2015 [5]. 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

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

Their algorithm is almost exactly the same as Nesterov's 2.2.20 [9] which we stated in the definition below:

Definition 5.1 (Lin's accelerated proximal point method) *Define the estimating sequence $(\phi_k)_{k\geq 0}$ recursively by:*

$$\begin{aligned} \phi_0(x) &:= F(x_0) + \frac{\gamma_0}{2}\|x - v_0\|^2, \\ \phi_k(x) &:= (1 - \alpha_{k-1})\phi_{k-1}(x) + \alpha_{k-1}\left(F(x_k) + \kappa\langle y_{k-1} - x_k, x - x_k\rangle + \frac{\mu}{2}\|x - x_k\|^2\right). \end{aligned}$$

Let the initial estimate be $x_0 \in \mathbb{R}^n$, fix parameters κ and α_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 $(x_k, y_k)_{k\geq 0}$ for all $k \geq 1$ such that:

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

Remark 5.2 The algorithm is similarity to Definition 3.1. In contrast, it has an inexact proximal point step controlled by ϵ_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 ϵ_k however is a whole new business. Lin 2015 [5] 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 ϵ_k to achieve accelerated convergence rate. In theory, ϵ_k decreases at a rate related to $F(x_0) - F^*$. It requires some knowledge about F^* in prior. Therefore, controlling ϵ_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 [5].

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 ϕ_k. One major innovation is Lemma A.7 in Lin's 2015 paper [5] which is stated below. The lemma allows the analysis of 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 strongly convex function with $\mu \geq 0$. Let κ be fixed. If x_k is an inexact proximal point evaluation of $x_k \approx \mathcal{J}_{\kappa^{-1}} y_{k-1}$ such that there exists ϵ_k where $\mathcal{M}^{\kappa^{-1}}(x_k; y_{k-1}) - \mathcal{M}^{\kappa^{-1}}(\mathcal{J}_{\kappa^{-1}} y_{k-1}, y_{k-1}) \leq \epsilon_k$. Denote $x_k^* = \mathcal{J}_{\kappa^{-1}} y_{k-1}$. Then it has 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 For a detailed proof of the lemma, see Appendix C.1. We note that it only uses the strong convexity of model function $\mathcal{M}^{\kappa^{-1}}(\cdot, y_{k-1})$, and the smoothness assumption of F is not involved in here.

Using this inexact proximal inequality, we showed the proof of Lemma A.8 in Lin's paper [5] in Appendix C.2 and Appendix C.3. These results construct the Nesterov's estimating sequence by rolling up the error in the recurrence while avoiding intractable quantities appearing in the algorithm.

6 Non-convex extension of Catalyst acceleration

The non-convex extension of Catalyst acceleration by Lin 2018 [6] is similar to the convex case in his 2015 paper [5]. The new algorithm handles function with unknown weak convexity constant ρ using a process called Auto Adapt subroutine. They only claimed convergence to stationarity for a weakly convex objective.

Fix κ. 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:*

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

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

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

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

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

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

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

For theorem follows, we note that a function is called ρ-weakly convex if and only if there exists $\rho \in \mathbb{R}$ such that $f + \frac{\rho}{2}\|\cdot\|^2$ is a convex function.

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

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

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

Remark 6.3 Convergence to stationary is strictly weaker than convergence to any stationary point of F. Read Section D for a proof of this claim.

7 Acknowledgement

Thanks for Professor Shawn Wang's patiently reading through the manuscripts and giving constructive feedback.

References

[1] R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik, *Variance-reduced methods for machine learning*, Proceedings of the IEEE, 108 (2020), pp. 1968–1983.

[2] G. N. Grapiglia and Y. Nesterov, *Accelerated regularized newton methods for minimizing composite convex functions*, SIAM Journal on Optimization, 29 (2019), pp. 77–99.

[3] O. Guler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM Journal on Control and Optimization, 29 (1991), p. 17.

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

[5] H. Lin, J. Mairal, and Z. Harchaoui, *A universal catalyst for first-order optimization*, in Procedings of Advances in Neural Information Processing Systems, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds., vol. 28, Curran Associates, Inc., 2015.

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

[7] Y. Nesterov, *A method for solving the convex programming problem with convergence rate O(1/kˆ2)*, Proceedings of the USSR Academy of Sciences, (1983).

[8] Y. Nesterov, *Accelerating the cubic regularization of Newton's method on convex problems*, Mathematical Programming, 112 (2008), pp. 159–181.

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

[10] 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.

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

A 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

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

We also introduce Bregman Divergence. Let $h : Q \to \mathbb{R}$ be a differentiable smooth convex function on a closed convex domain $Q \subseteq \mathbb{R}^n$. Define

$$D_h : Q \times \operatorname{ri} Q \to \mathbb{R}^n := (x,y) \mapsto h(x) - h(y) - \langle \nabla h(y), x-y\rangle.$$

Theorem A.1 (Fundamental theorem of proximal gradient) *Let $F = f + g$, define the proximal gradient operator $\widetilde{\mathcal{J}}_{L^{-1}}$. For any fixed y, we have for all $x \in \mathbb{R}^n$:*

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

Proof. By a direct observation:

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

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

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

■

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 theorem in older literatures, this proof doesn't use the subgradient inequality, making it appealing for generalizations outside convexity.

Theorem A.3 (Canonical form of proximal gradient estimating sequence) *Fix any $x_0 \in \mathbb{R}^n$. Define $\phi_k : \mathbb{R}^n \to \mathbb{R}$ as a sequence of functions such that for all $k \geq 0$ it recursively satisfies 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 $(y_k)_{k\geq 0}$ any sequence. If we define the canonical form for ϕ_k as convex quadratic parameterized by positive sequence $(\gamma_k), \phi_k^*$:

$$\phi_k(x) := \phi_k^* + \frac{\gamma_k}{2}\|x - v_k\|^2,$$

where $\phi_k^* := \min_x \phi_k(x)$, $\gamma_0 > 0$ *and* $x_0 = v_0$. *Then the auxiliary sequence* y_k, v_k, *parameters for the canonical form of estimating sequence must satisfy for all* $k \geq 0$*:*

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

Proof. By the recursive definition of ϕ_k:

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

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 ϕ_{k+1} by definition of the canonical form, solving for x in $\nabla\phi_{k+1}(x) = \mathbf{0}$ yields v_{k+1} by:

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

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

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

Next, we simplify $\|v_{k+1} - y_k\|^2$ to get rid of v_{k+1}.

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

From the second to third inequality, we used:

$$\begin{aligned} & \gamma_{k+1} = (1-\alpha_k)\gamma_k + \mu\alpha_k \\ \iff & -(1-\alpha_k)\gamma_k = -\gamma_{k+1} + \mu\alpha_k. \end{aligned}$$

Therefore, it has:

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

Substitute it back to (eqn2) we have

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

From the second to third equality we added and then transformed the coefficient of $\|y_k - v_k\|$ using:

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

From the second to the third line, $\mu\alpha_k$ is brought in by the equation $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \mu\alpha_k$.

■

Remark A.4 The goal of having ϕ_k^* represented by iterates v_k, y_k, g_k is because it assists in the future when an inductive hypothesis $\phi_k^* \geq F(x_k) \geq l_F(x_k; y_k) + \frac{\mu}{2}\|x - y_k\|^2$ is made. We did things in advance here and grouped together a set of transforms by equations, nothing more.

Theorem A.5 (Verifying the conditions of implicit descent)
Let estimating sequence ϕ_k and auxiliary sequence $y_k, v_k, \gamma_k, \alpha_k$ be given by Theorem A.3. If for all $k \geq 0$:

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

then ϕ_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 ϕ_{k+1}:

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

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 setting; define $x_{k+1} := \widetilde{\mathcal{J}}_{L^{-1}} y_k$, we have $\forall z \in \mathbb{R}^n$:

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

Next, set $z = x_k$ and substitute it back to RHS of ϕ^*_{k+1} in Inequality A.1:

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

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

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

Re-arranging gives the equivalent representation:

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

On the second to third equality, we multiplied the numerator and denominator by γ_{k+1} and then simplified the numerator using equation $\gamma_{k+1} + \alpha_k\gamma_k = \gamma_k + \alpha_k\mu$. For α_k, γ_k, we have the equivalent representation of

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

■

Definition A.6 (Nesterov's accelerated proximal gradient raw form) *The accelerated proximal gradient algorithm generates vector iterates x_k, y_k, v_k using auxiliary sequence α_k, γ_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.7 (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 which is Definition A.6. Assume for all $k \geq 0$ it has $L\alpha_k^2 = \gamma_{k+1}$ Then Definition A.6 is algebraically equivalent to the following form which doesn't have γ_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:

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

We can eliminate the γ_k which defines the α_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.8 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.9 (Nesterov's accelerated proximal gradient momentum form) *Let the sequence α_k, and vectors y_k, x_k, v_k be given by the intermediate form of the Nesterov's*

accelerated proximal gradient without using v_k. The algorithm generates y_k, x_k, α_k such that it satisfies for all $k \geq 0$:

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

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 ξ_k, τ_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:

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

The above is true for all $k \geq 0$, so it's also true for $k+1$ as well. Therefore it gives:

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

We are closer than ever to proving it. This representation contains L, μ 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.10 This proof is very long because we took a detour to an intermediate form without the γ_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:

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

Observe ϕ_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 ϕ_{k+1}, it can be solved for by setting the gradient $\phi_{k+1}(v_{k+1}) = \mathbf{0}$ so for all $k \geq 0$:

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

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 α_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

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

Equality $A_{k+1} = (1-\alpha_k)A_k$ is used on the second last inequality for the above. 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

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

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):

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

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. For the Accelerated Proximal Gradient, we simplified ϕ_{k+1}^* prior to considering $F(x_{k+1}) \leq \phi_{k+1}^*$. In here, we substitute the inductive hypothesis $\phi_k^* \geq f(x_k)$ and skipped the canonical representation of ϕ_k^*.

A more critical observation here is the Nesterov's estimating sequence from Section A is the same as Guler's Accelerated PPM if we choose $L^{-1} = \lambda_k$,$\mu = 0$ and $g \equiv 0$ because:

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

When $g \equiv 0$, $L^{-1} = \lambda_k$, the proximal gradient operator has $\widetilde{\mathcal{J}}_{L^{-1}} = \mathcal{J}_{\lambda_k}(y_k)$, making $\phi_k(x)$ here the same as accelerated proximal gradient.

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 λ_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.

Remark B.3 This is nothing new. First inequality is a direct consequence of strong convexity and the second inequality is the PL-inequality implied by strong convexity.

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 is the same as Section 5. 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.

C.1 Inexact proximal inequality

Proof. This is the proof for Lemma 5.3. Define $G_k := \mathcal{M}_F^{\kappa^{-1}}(\cdot, y_{k-1}), G_k^* = \mathcal{M}_F^{\kappa^{-1}}(x_k^*, y_{k-1})$. G_k is a $\mu+\kappa$ convex function which implies quadratic growth over minimizer x_k^*:

$$(\forall x) \quad G_k(x) \geq G_k^* + \frac{\kappa+\mu}{2}\|x - x_k^*\|^2.$$

Substituting definitions:

$$\begin{aligned}
F(x) &\geq G_k(x_k) + (G_k^* - G_k(x_k)) + \frac{\mu+\kappa}{2}\|x - x_k^*\|^2 - \frac{\kappa}{2}\|x - y_{k-1}\|^2 \\
&\geq G(x_k) - \epsilon_k + \frac{\kappa+\mu}{2}\|x - x_k^*\|^2 - \frac{\kappa}{2}\|x - y_{k-1}\|^2 \\
&= G_k(x_k) - \epsilon_k + \frac{\kappa+\mu}{2}\left(\|x - x_k - x_k + x_k^*\|^2\right) - \frac{\kappa}{2}\|x - y_{k-1}\|^2 \\
&= G_k(x_k) - \epsilon_k + \frac{\kappa+\mu}{2}\left(\|x - x_k\|^2 + \|x_k - x_k^*\|^2 + 2\langle x - x_k, x_k - x_k^*\rangle\right) - \frac{\kappa}{2}\|x - y_{k-1}\|^2 \\
&= \left(G_k(x_k) + \frac{\kappa}{2}\|x - x_k\|^2 - \frac{\kappa}{2}\|x - y_{k-1}\|^2\right) - \epsilon_k \\
&\quad + \frac{\mu}{2}\|x - x_k\|^2 + \frac{\kappa+\mu}{2}\|x_k - x_k^*\|^2 + (\kappa+\mu)\langle x - x_k, x_k - x_k^*\rangle.
\end{aligned}$$

Simplify terms inside the parenthesis:

$$\begin{aligned}
&G_k(x_k) + \frac{\kappa}{2}\|x - x_k\|^2 - \frac{\kappa}{2}\|x - y_{k-1}\|^2 \\
&= F(x_k) + \frac{\kappa}{2}\|x_k - y_{k-1}\|^2 + \frac{\kappa}{2}\|x - x_k\|^2 - \frac{\kappa}{2}\|x - y_{k-1}\|^2 \\
&= F(x_k) + \frac{\kappa}{2}\left(\|x_k - y_{k-1}\|^2 - \|x - y_{k-1}\|^2\right) + \frac{\kappa}{2}\|x - x_k\|^2 \\
&= F(x_k) + \frac{\kappa}{2}\left(\|x_k - x\|^2 + 2\langle x_k - x, x - y_{k-1}\rangle\right) + \frac{\kappa}{2}\|x - x_k\|^2 \\
&= F(x_k) + \kappa\|x_k - x\|^2 + \kappa\langle x_k - x, x - y_{k-1}\rangle \\
&= F(x_k) + \kappa\langle x_k - x, x - y_{k-1} + x_k - x\rangle \\
&= F(x_k) + \kappa\langle x_k - x, x_k - y_{k-1}\rangle.
\end{aligned}$$

Therefore, it has the inequality:

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

Re-arranging it to a form by the writer's preference:

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

Now we have the RHS to be exclusively about the inexact evaluation $x_k \approx \mathcal{J}_{\kappa^{-1}} y_{k-1}$, and $G_k(x_k) - G_k^* \leq \epsilon_k$. ■

Remark C.1 Two major things about this lemma:

(i) Set $\epsilon_k = 0$ so $x_k = x_k^*$, the lemma is the proximal inequality.

(ii) It's not entirely obvious on its relations to the proximal gradient inequality.

For point (ii), observe that the proximal gradient inequality given by Theorem A.1 has RHS that is completely different. We have made attempts at bridging the two inequalities. There are no obvious useful results and concrete claims. There are no choices of parameters that makes the two inequalities equivalent. We are not aware of any analogous results to this theorem in the literatures.

C.2 Approximated Nesterov's Estimating sequence

In this subsection, let ϕ_k be the Nesterov's estimating sequence for Catalyst. For all $k \geq 0$, ϕ_k satisfies

$$\begin{aligned}
\phi_0(x) &:= F(x_0) + \frac{\gamma_0}{2}\|x - v_0\|^2, \\
\phi_k(x) &:= (1-\alpha_{k-1})\phi_{k-1}(x) + \alpha_{k-1}\left(F(x_k) + \kappa\langle y_{k-1} - x_k, x - x_k\rangle + \frac{\mu}{2}\|x - x_k\|^2\right).
\end{aligned}$$

Observe that ϕ_k is a simple quadratic function. We prove following theorem which is Lemma A.6 in Lin's writings on Catalyst [5].

Theorem C.2 (Canonical form of Catalyst estimating sequence) *Define v_k, γ_k to be the parameters for the canonical form of ϕ_k as given by*

$$\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2}\|x - v_k\|^2,$$

where $\phi_k^* = \min_x \phi_k(x)$. Then the parameters of the canonical form $(\gamma_k)_{k\geq 0}, (v_k)_{k\geq 0}, (\phi_k^*)_{k\geq 0}$ satisfies for all $k \geq 1$:

$$\begin{aligned}
\gamma_k &= (1-\alpha_{k-1})\gamma_{k-1} + \alpha_{k-1}\mu, \\
v_k &= \gamma_k^{-1}((1-\alpha_{k-1})\gamma_{k-1}v_{k-1} + \alpha_{k-1}\mu x_k - \alpha_{k-1}\kappa(y_{k-1} - x_k)), \\
\phi_k^* &= (1-\alpha_{k-1})\phi_{k-1}^* + \alpha_{k-1}F(x_k) - \frac{\alpha_{k-1}^2}{2\gamma_k}\|\kappa(y_{k-1} - x_k)\|^2 \\
&\quad + \frac{\alpha_{k-1}(1-\alpha_{k-1})\gamma_{k-1}}{\gamma_k}\left(\frac{\mu}{2}\|x_k - v_{k-1}\|^2 + \langle \kappa(y_{k-1} - x_k), v_{k-1} - x_k\rangle\right).
\end{aligned}$$

Proof. For v_{k+1}, solve for $\nabla\phi_k(v_k) = \mathbf{0}$ using the recursive definition. For γ_k, consider the Hessian $\nabla^2\phi_k$ using the recursive definition. The process is similar to derivations in Section A.

Next, we work on ϕ_k^*. The goal of the canonical form is to simplify the process of solving for an implicit sequence x_k such that $f(x_k) \leq \phi_k^*$. To start consider for all $k \geq 1$:

$$\begin{aligned}
\phi_k(x_k) &= \phi_k^* + \frac{\gamma_k}{2}\|x_k - v_k\|^2 \\
&= (1-\alpha_{k-1})\left(\phi_{k-1}^* + \frac{\gamma_{k-1}}{2}\|x_k - v_{k-1}\|^2\right) + \alpha_{k-1}F(x_k) \\
\iff \phi_k^* &= (1-\alpha_{k-1})\left(\phi_{k-1}^* + \frac{\gamma_{k-1}}{2}\|x_k - v_{k-1}\|^2\right) + \alpha_{k-1}F(x_k) - \frac{\gamma_k}{2}\|x_k - v_k\|^2 \\
&= (1-\alpha_{k-1})\phi_{k-1} + \alpha_{k-1}F(x_k) + \frac{(1-\alpha_{k-1})\gamma_{k-1}}{2}\|x_k - v_{k-1}\|^2 - \frac{\gamma_k}{2}\|x_k - v_k\|^2.
\end{aligned}$$

Now, it would be great to express $\|x_k - v_k\|^2$, so it depends on the iterates from previous iteration. From the definition of v_k we have

$$\begin{aligned}
v_k - x_k &= \gamma_k^{-1}((1-\alpha_{k-1})\gamma_{k-1}v_{k-1} + \alpha_{k-1}\mu x_k - \alpha_{k-1}\kappa(y_{k-1} - x_k)) - x_k \\
&= \gamma_k^{-1}((1-\alpha_{k-1})\gamma_{k-1}v_{k-1} + (\alpha_{k-1}\mu - \gamma_k)x_k - \alpha_{k-1}\kappa(y_{k-1} - x_k)) \\
&= \gamma_k^{-1}((1-\alpha_{k-1})\gamma_{k-1}v_{k-1} - (1-\alpha_{k-1})\gamma_{k-1}x_k - \alpha_{k-1}\kappa(y_{k-1} - x_k)) \\
&= \gamma_k^{-1}((1-\alpha_{k-1})\gamma_{k-1}(v_{k-1} - x_k) - \alpha_{k-1}\kappa(y_{k-1} - x_k)).
\end{aligned}$$

Taking the norm and multiplying by $\gamma_k/2$ to match the terms in (eqn1) then:

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

Substituting it back we have

$$\begin{aligned}
\phi_k^* &= (1-\alpha_{k-1})\phi_{k-1} + \alpha_{k-1}F(x_k) + \frac{(1-\alpha_{k-1})\gamma_{k-1}}{2}\|x - v_{k-1}\|^2 - \frac{(1-\alpha_{k-1})^2\gamma_{k-1}^2}{2\gamma_k}\|x_k - v_{k-1}\|^2 \\
&\quad + \left(-\frac{\alpha_{k-1}^2}{2\gamma_k}\|\kappa(y_{k-1} - x_k)\|^2 + \frac{\gamma_{k-1}(1-\alpha_{k-1})\alpha_{k-1}}{\gamma_k}\langle v_{k-1} - x_k, \kappa(y_{k-1} - x_k)\rangle\right) \\
&= (1-\alpha_{k-1})\phi_{k-1} + \alpha_{k-1}F(x_k) + \frac{\alpha_{k-1}(1-\alpha_{k-1})\gamma_{k-1}\mu}{2\gamma_k}\|x_k - v_{k-1}\|^2 - \frac{\alpha_{k-1}^2}{2\gamma_k}\|\kappa(y_{k-1} - x_k)\|^2 \\
&\quad + \frac{\gamma_{k-1}(1-\alpha_{k-1})\alpha_{k-1}}{\gamma_k}\langle v_{k-1} - x_k, \kappa(y_{k-1} - x_k)\rangle \\
&= (1-\alpha_{k-1})\phi_{k-1} + \alpha_{k-1}F(x_k) \\
&\quad + \frac{\alpha_{k-1}(1-\alpha_{k-1})\gamma_{k-1}}{\gamma_k}\left(\frac{\mu}{2}\|x_k - v_{k-1}\|^2 + \langle v_{k-1} - x_k, \kappa(y_{k-1} - x_k)\rangle.\right)
\end{aligned}$$

On the second equality, we made use of the following to simplify the coefficients for $\|x_k - v_{k-1}\|^2$:

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

■

Remark C.3 The differences of the Canonical form is cosmetic compare results from Guler, and Acceleration proximal gradient. The arrangement is different because the inexact proximal inequality from Lemma 5.3 are anchored on different iterates.

C.3 Controlling the errors

The following theorem is Theorem A.8 from Lin's writing of Catalyst [5]. It stated the propagation of errors ϵ_k from the inexact proximal inequality to the descent condition $F(x_k) \leq \phi_k^*$.

Theorem C.4 (Controlling the error in Nesterov's estimating sequence) *If the auxiliary sequences $v_k, y_k, \gamma_k, \alpha_k$ satisfies the conditions:*

$$\begin{aligned}
\gamma_k - (\kappa+\mu)\alpha_{k-1}^2 &= 0, \\
(1-\alpha_{k-1})\gamma_{k-1} + \alpha_{k-1}\mu &= (\kappa+\mu)\alpha_{k-1}^2.
\end{aligned}$$

Then the canonical representation of estimating sequence ϕ_k^* and the function value the inexact proximal point iterates $F(x_k)$ satisfy for all $k \geq 1$

$$\begin{aligned} F(x_k) &\leq \phi_k^* + \xi_k, \\ \xi_k &= (1-\alpha_{k-1})(\xi_{k-1} + \epsilon_k - (\kappa+\mu)\langle x_k - x_k^*, x_{k-1} - x_k\rangle). \end{aligned}$$

Where we have the base case that $\xi_0 = 0$

Proof. We prove it via induction. Base case is trivially satisfied via $\phi_0^* = F(x_0)$ and $\xi_0 = 0$. Inductively we assume that $F(x_{k-1}) \leq \phi_{k-1}^* + \xi_k$. By definition, it means

$$\begin{aligned} \phi_{k-1}^* &\geq F(x_{k-1}) - \xi_{k-1} \\ &\geq F(x_k) + \langle \kappa(y_{k-1} - x_k), x_{k-1} - x_k\rangle + (\kappa+\mu)\langle x_k - x_k^*, x_{k-1} - x_k\rangle - \epsilon_k - \xi_{k-1} \\ &= F(x_k) + \langle \kappa(y_{k-1} - x_k), x_{k-1} - x_k\rangle - (1-\alpha_{k-1})^{-1}\xi_k. \end{aligned} \tag{C.1}$$

Substituting it into the canonical form representation of the estimating sequence derived from the previous section, it has

$$\begin{aligned} \phi_k^* &= (1-\alpha_{k-1})\phi_{k-1}^* + \alpha_{k-1}F(x_k) - \frac{\alpha_{k-1}}{2\gamma_k}\|\kappa(y_{k-1} - x_k)\|^2 \\ &\quad + \frac{\alpha_{k-1}(1-\alpha_{k-1})\gamma_{k-1}}{\gamma_k}\left(\frac{\mu}{2}\|x_k - v_{k-1}\|^2 + \langle \kappa(y_{k-1} - x_k), v_{k-1} - x_k\rangle\right) \\ &\geq (1-\alpha_{k-1})\left(F(x_k) + \langle \kappa(y_{k-1} - x_k), x_{k-1} - x_k\rangle - (1-a_{k-1})^{-1}\xi_k\right) + \alpha_{k-1}F(x_k) \\ &\quad + \frac{\alpha_{k-1}(1-\alpha_{k-1})\gamma_{k-1}}{\gamma_k}\left(\frac{\mu}{2}\|x_k - v_{k-1}\|^2 + \langle \kappa(y_{k-1} - x_k), v_{k-1} - x_k\rangle\right) - \frac{\alpha_{k-1}}{2\gamma_k}\|\kappa(y_{k-1} - x_k)\|^2 \\ &= (1-\alpha_{k-1})\langle \kappa(y_{k-1} - x_k), x_{k-1} - x_k\rangle - \frac{\alpha_{k-1}}{2\gamma_k}\|\kappa(y_{k-1} - x_k)\|^2 \\ &\quad + \frac{\alpha_{k-1}(1-\alpha_{k-1})\gamma_{k-1}}{\gamma_k}\left(\frac{\mu}{2}\|x_k - v_{k-1}\|^2 + \langle \kappa(y_{k-1} - x_k), v_{k-1} - x_k\rangle\right) + F(x_k) - \xi_k \\ &= (1-\alpha_{k-1})\left\langle \kappa(y_{k-1} - x_k), \frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}(v_{k-1} - x_k) + x_{k-1} - x_k\right\rangle - \frac{\alpha_{k-1}}{2\gamma_k}\|\kappa(y_{k-1} - x_k)\|^2 \\ &\quad + \frac{\mu\alpha_{k-1}(1-\alpha_{k-1})\gamma_{k-1}}{2\gamma_k}\|x_k - v_{k-1}\|^2 + F(x_k) - \xi_k. \end{aligned} \tag{C.2}$$

Next, we need to focus on the first 2 terms on the RHS

$$(1-\alpha_{k-1})\left\langle \kappa(y_{k-1} - x_k), \frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}(v_{k-1} - x_k) + x_{k-1} - x_k\right\rangle - \frac{\alpha_{k-1}}{2\gamma_k}\|\kappa(y_{k-1} - x_k)\|^2. \tag{C.3}$$

The first inner product term in C.3 has

$$\begin{aligned}
&\left\langle \kappa(y_{k-1}-x_k), x_{k-1}-y_{k-1}+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}(v_{k-1}-y_{k-1}+y_{k-1}-x_k)\right\rangle \\
&=\left\langle \kappa(y_{k-1}-x_k), x_{k-1}-y_{k-1}+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}(v_{k-1}-y_{k-1})\right\rangle \\
&\quad+\left\langle \kappa(y_{k-1}-x_k), y_{k-1}-x_k+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}(y_{k-1}-x_k)\right\rangle \\
&=\left\langle \kappa(y_{k-1}-x_k), x_{k-1}-y_{k-1}+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}(v_{k-1}-y_{k-1})\right\rangle \\
&\quad+\kappa\left(1+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}\right)\|v_{k-1}-y_{k-1}\|^2.
\end{aligned}$$

With the above C.3 simplifies to

$$\begin{aligned}
&(1-\alpha_{k-1})\left\langle \kappa(y_{k-1}-x_k), x_{k-1}-y_{k-1}+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}(v_{k-1}-y_{k-1})\right\rangle \\
&-\frac{\alpha_{k-1}}{2\gamma_k}\|\kappa(y_{k-1}-x_k)\|^2+\kappa(1-\alpha_{k-1})\left(1+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}\right)\|v_{k-1}-y_{k-1}\|^2 \\
&=(1-\alpha_{k-1})\left\langle \kappa(y_{k-1}-x_k), x_{k-1}-y_{k-1}+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}(v_{k-1}-y_{k-1})\right\rangle \\
&\quad+(1-\alpha_{k-1})\kappa\left(1+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}-\frac{\kappa\alpha_{k-1}}{2\gamma_k}\right)\|y_{k-1}-x_k\|^2.
\end{aligned}$$

We can simplify the coefficient of $\|y_{k-1} - x_k\|^2$ in the above expression by the recurrence of parameter γ_k from the canonical form of estimating sequence.

$$
\begin{aligned}
&(1-\alpha_k)\kappa\left(1+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}-\frac{\kappa\alpha_{k-1}}{2\gamma_k}\right)\\
&=\kappa\left(1-\alpha_{k-1}+\frac{(1-\alpha_{k-1})\alpha_{k-1}\gamma_{k-1}}{\gamma_k}-\frac{\alpha_{k-1}^2\kappa}{2\gamma_k}\right)\\
&\text{Use: } \gamma_k-\alpha_{k-1}\mu=(1-\alpha_{k-1})\gamma_{k-1}\\
&=\kappa\left(1-\alpha_{k-1}+\frac{(\gamma_k-\alpha_{k-1}\mu)\alpha_{k-1}}{\gamma_k}-\frac{\alpha_{k-1}^2\kappa}{2\gamma_k}\right)\\
&=\kappa\left(1+\frac{-2\gamma_k\alpha_{k-1}+2(\gamma_k-\alpha_{k-1}\mu)\alpha_{k-1}-\alpha_{k-1}^2\kappa}{2\gamma_k}\right)\\
&=\kappa\left(1+\frac{-2\alpha_{k+1}^2\mu-\alpha_{k-1}^2\kappa}{2\gamma_k}\right)\\
&=\kappa\left(1-\frac{(2\mu+\kappa)\alpha_{k-1}^2}{2\gamma_k}\right)\\
&=\kappa\left(1-\frac{(\mu+\kappa/2)\alpha_{k-1}^2}{\gamma_k}\right).
\end{aligned}
$$

Now, substitute the above back to C.3 and then substitute C.3 back to C.2 to get:

$$
\begin{aligned}
\phi_k^* &\geq (1-\alpha_{k-1})\left\langle \kappa(y_{k-1}-x_k), x_{k-1}-y_{k-1}+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}(v_{k-1}-y_{k-1})\right\rangle\\
&\quad+\kappa\left(1-\frac{(\mu+\kappa/2)\alpha_{k-1}^2}{\gamma_k}\right)\|y_{k-1}-x_k\|^2+\frac{\mu\alpha_{k-1}(1-\alpha_{k-1})\gamma_{k-1}}{2\gamma_k}\|x_k-v_{k-1}\|^2+F(x_k)-\xi_k\\
&\geq (1-\alpha_{k-1})\left\langle \kappa(y_{k-1}-x_k), x_{k-1}-y_{k-1}+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}(v_{k-1}-y_{k-1})\right\rangle\\
&\quad+\kappa\left(1-\frac{(\mu+\kappa/2)\alpha_{k-1}^2}{\gamma_k}\right)\|y_{k-1}-x_k\|^2+\|x_k-v_{k-1}\|^2+F(x_k)-\xi_k.
\end{aligned}
$$

To assert the inductive hypothesis $F(x_k) \leq \phi_k^*$ it's sufficient to have the inner product term equals to zero, and the coefficient of $\|y_k - x_k\|$ to be non-negative. Therefore, it suffices to consider

$$
\begin{aligned}
x_{k-1}-y_{k-1}+\frac{\alpha_{k-1}\gamma_{k-1}}{\gamma_k}(v_{k-1}-y_{k-1})&=\mathbf{0},\\
1-\frac{(\kappa/2+\mu)\alpha_{k-1}^2}{\gamma_k}\leq 1-(\kappa+\mu)\frac{\alpha_{k-1}^2}{\gamma_k}&\leq 0.
\end{aligned}
$$

In addition, if we assume equality holds then it gives:

$$\gamma_k - (\kappa+\mu)\alpha_{k-1}^2 = 0,$$
$$(1-\alpha_{k-1})\gamma_{k-1} + \alpha_{k-1}\mu = (\kappa+\mu)\alpha_{k-1}^2.$$

■

Remark C.5 For all $k \geq 1$, with vector x_k, v_k and scalar γ_k, α_k given, the updates for $x_{k+1}, y_{k+1}, v_{k+1}$ and γ_{k+1} are produced by:

$$\begin{aligned}
\gamma_{k+1} &= (1-\alpha_k)\gamma_k + \alpha_k\mu = (\kappa+\mu)\alpha_{k-1}^2,\\
y_k &= (\gamma_k + \alpha_k\mu)^{-1}(\alpha_k\gamma_k v_k + \gamma_{k+1}x_k),\\
\tilde{x}_{k+1} &\approx \mathcal{J}_{\kappa^{-1}}y_k \text{ s.t: } \mathcal{M}^{\kappa^{-1}}(x_{k+1}, y_k) - \mathcal{M}^{\kappa^{-1}}(\mathcal{J}_{\kappa^{-1}}y_k, y_k) \leq \epsilon_k,\\
\widetilde{\mathcal{G}}_{\kappa^{-1}}y_k &= \kappa(y_k - \tilde{x}_{k+1}),\\
x_{k+1} &= y_k - \widetilde{\mathcal{G}}_{\kappa^{-1}}y_k,\\
v_{k+1} &= \gamma_{k+1}^{-1}\left((1-\alpha_k)\gamma_k v_k + \alpha_k(\mu+\kappa)x_k - \alpha_k\kappa y_k\right)\\
&= \gamma_{k+1}^{-1}((1-\alpha_k)\gamma_k v_k + \alpha_k\mu x_k + \alpha_k\kappa(x_k - y_k))\\
&= \gamma_{k+1}^{-1}\left((1-\alpha_k)\gamma_k v_k + \alpha_k\mu x_k - \alpha_k\widetilde{\mathcal{G}}_{\kappa^{-1}}y_k\right).
\end{aligned}$$

In terms of just the formatting of updates on (v_k, y_k, x_k), α_k, γ_k are conducted, it's exactly the same as the Accelerated proximal gradient algorithm proved back in Section A, but with $L = \kappa + \mu$ and $\tilde{x}_{k+1}$ produced by the inexact proximal point instead of proximal gradient. The results on the equivalent representations of the accelerated proximal gradient algorithm from the previous section will apply for this the Catalyst as well. Therefore, by analysis from previous section, we also have for all $k \geq 0$:

$$y_k = x_k + \frac{\alpha_{k-1}(1-\alpha_{k-1})}{\alpha_{k-1}^2 + \alpha_k}(x_k - x_{k-1}).$$

Lin's innovation of the inexact proximal inequality allows us to roll the error from the inexact proximal point into ξ_k to characterize the descent of the sequence $x_k \approx \mathcal{J}_{\kappa^{-1}}y_{k-1}$. It undoubtedly improved the results from Lemma 3.2 of Guler's paper in 1992 [4].

D Proofs for 4WD Catalyst Acceleration

In this section we adopt the same notations as in Section 6. 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 (Bounds of the inverted FISTA sequence) *With base case $\alpha_1 = 1$ If the sequence α_k has for all $k \geq 1$:*

$$\alpha_{k+1} = \frac{\sqrt{\alpha_k^4 + 4\alpha_k^2} - \alpha_k^2}{2}, \alpha_1 = 1$$

then for all $k \geq 0$:

$$\frac{\sqrt{2}}{k+1} \leq \alpha_k \leq \frac{2}{k+1}.$$

Proof. The proof of the upper bound is by inductive hypothesis, assume $\alpha_k \geq 2/(k+1)$ then consider

$$\begin{aligned}
\alpha_{k+1} &= \frac{\sqrt{\alpha_k^4 + 4\alpha_k^2} - \alpha_k^2}{2} = \frac{\sqrt{\alpha_k^4 + 4\alpha_k^2} - \alpha_k^2}{2} \frac{\sqrt{\alpha_k^4 + 4\alpha_k^2} + \alpha_k^2}{\sqrt{\alpha_k^4 + 4\alpha_k^2} + \alpha_k^2} \\
&= \frac{\alpha_k^4 + 4\alpha_k^2 - \alpha_k^4}{2(\sqrt{\alpha_k^4 + 4\alpha_k^2}) + \alpha_k^2)} \\
&= \frac{2}{\sqrt{1 + 4\alpha_k^{-2}} + 1} \leq \frac{2}{\sqrt{1 + (k+1)^2} + 1} \leq \frac{2}{k+2}.
\end{aligned}$$

The lower bound is given by the hidden recursive definition of α_k:

$$\begin{aligned}
\alpha_{k+1}^2 &= (1 - \alpha_{k+1})\alpha_k^2 \\
&= \alpha_1^2 \prod_{i=2}^{k+1} (1 - \alpha_i) \\
&\geq \prod_{i=2}^{k+1} \left(1 - \frac{2}{i+1}\right) = \prod_{i=2}^{k+1} \left(\frac{i-1}{i+1}\right) \\
&= \frac{2}{(k+1)(k+2)} \geq \frac{2}{(k+2)^2}.
\end{aligned}$$

On the third step of the inequality, we used $\alpha_k \leq 2/(k+1)$ and $\alpha_1 = 1$. ■

Remark D.2 If $\lambda_k = a_k^{-1}$, then λ_k would be the FISTA sequence.

Lemma D.3 (Stationarity condition) *Assume F is ρ weakly convex. Fix any y, suppose that y^+ satisfies* $\operatorname{dist}(\mathbf{0}, \partial \mathcal{M}^{k^{-1}}(y^+; y)) \leq \epsilon$ *then the following inequality holds:*

$$\operatorname{dist}(\mathbf{0}; \partial F(y^+)) \leq \epsilon + \kappa \|y^+ - y\|.$$

Proof. Take it as a fact that the limiting subgradient of a weakly convex function is a closed set. Fix any y, there exists $w \in \partial \mathcal{M}(y^+; y)$ such that $\text{dist}(\mathbf{0}; \partial \mathcal{M}(y^+; y)) = \|w\|$ because $\partial M(\cdot; y)$ is a $\rho - \kappa$ weakly convex function. Next, by definition we have

$$\begin{aligned} & w \in \partial F(y^+) + \kappa(y^+ - y) \\ \iff & \exists v \in \partial F(y^+) : w = v + \kappa(y^+ - y) \\ \implies & \epsilon = \|w\| = \|v + \kappa(y^+ - y)\| \geq \|v\| - \|\kappa(y^+ - y)\| \\ \implies & \text{dist}(\mathbf{0}, \partial F(y^+)) \leq \|v\| \leq \epsilon + \|\kappa(y^+ - y)\|. \end{aligned}$$

■

D.1 Convergence proof of the basic 4WD-Catalyst

This subsection proves Theorem 6.2 which is about the convergence of Definition 6.1.

For any $k \geq 1$, the algorithm asserts

$$F(x_{k-1}) \geq \mathcal{M}(\bar{x}_k, x_{k-1}) \geq F(x_k) + \frac{\kappa}{2}\|\bar{x}_k - x_{k-1}\|^2.$$

Use Lemma D.3 with $\epsilon = \kappa\|\bar{x}_k - x_{k-1}\|$, $y = x_{k-1}$, $y^+ = \bar{x}_k$ then

$$\text{dist}(\mathbf{0}, \partial F(\bar{x}_k)) \leq 2\kappa\|\bar{x}_k - x_{k-1}\|.$$

With F bounded below, denote F^* to be the minimum then using the two results above:

$$\begin{aligned} F(x_{k-1}) - F(x_k) &\geq \frac{\kappa}{2}\|\bar{x}_k - x_{k-1}\|^2 \\ 8\kappa(F(x_{k-1}) - F(x_k)) &\geq 4\|\kappa(\bar{x}_k - x_{k-1})\|^2 \geq \text{dist}^2(\mathbf{0}, \partial F(\bar{x}_k)) \\ \implies \text{dist}^2(\mathbf{0}, \partial F(\bar{x}_k)) &\leq 8\kappa(F(x_{k-1}) - F(x_k)) \\ \implies \min_{j=1,\cdots,N} \text{dist}^2(\mathbf{0}, \partial F(\bar{x}_j)) &\leq \frac{8\kappa}{N}\sum_{j=1}^{N} F(x_{j-1}) - F(x_j) \\ &\leq \frac{8\kappa}{N}(F(x_0) - F(x_N)) \leq \frac{8\kappa}{N}(F(x_0) - F^*). \end{aligned}$$

Therefore, the set limits of $\partial F(\bar{x}_j)$ contains $\mathbf{0}$. The second part of the claim about the convergence to optimality requires additional assumptions that

(i) F is convex.

(ii) F is bounded below and has a minimizer x^*.

From the algorithm we have $\xi_k \in \partial\mathcal{M}(\tilde{x}_k, y_k)$ such that $\|\xi_k\| \leq \frac{\kappa}{k+1}\|\tilde{x}_k - y_k\|$. Then for any $x \in \mathbb{R}^n$, κ strong convexity of $\mathcal{M}(\cdot, y_k)$ yields inequality:

$$\begin{aligned}
0 &\leq F(x) + \frac{\kappa}{2}\|x - y_k\|^2 - \left(F(\tilde{x}) + \frac{\kappa}{2}\|\tilde{x}_k - y_k\|^2\right) - \frac{\kappa}{2}\|x - \tilde{x}_k\|^2 - \langle \xi_k, x - \tilde{x}_k\rangle, \\
F(x_k) \leq F(\tilde{x}_k) &\leq F(x) + \frac{\kappa}{2}\left(\|x - y_k\|^2 - \|x - \tilde{x}_k\|^2 - \|\tilde{x}_k - y_k\|^2\right) + \langle \xi_k, \tilde{x}_k - x\rangle \\
&\leq F(x) + \frac{\kappa}{2}\left(\|x - y_k\|^2 - \|x - \tilde{x}_k\|^2 - \|\tilde{x}_k - y_k\|^2\right) + \frac{\kappa}{k+1}\|\tilde{x}_k - y_k\|\|x - \tilde{x}_k\|.
\end{aligned}$$

Sure, now observe that with the substitutions $x = \alpha_k x^* + (1 - \alpha_k)x_{k-1}$ where x^* is the minimizer then

$$\begin{aligned}
x - y_k &= \alpha_k x^* + (1 - \alpha_k)x_{k-1} - y_k \\
&= \alpha_k x^* + (1 - \alpha_k)x_{k-1} - (\alpha_k v_{k-1} + (1 - \alpha_k)x_{k-1}) \\
&= \alpha_k(x^* - v_{k-1}), \\
x - \tilde{x}_k &= \alpha_k x^* + (1 - \alpha_k)x_{k-1} - \tilde{x}_k \\
&\qquad\qquad v_k = x_{k-1} + \alpha_k^{-1}(\tilde{x}_k - x_{k-1}) \\
&\qquad\quad \tilde{x}_k - x_{k-1} = \alpha_k(v_k - x_{k-1}) \\
&\qquad\qquad \tilde{x}_k = x_{k-1} + \alpha_k(v_k - x_{k-1}) \\
&= \alpha_k x^* + (1 - \alpha_k)x_{k-1} - (x_{k-1} + \alpha_k(v_k - x_{k-1})) \\
&= \alpha_k x^* - \alpha_k x_{k-1} - \alpha_k(v_k - x_{k-1}) \\
&= \alpha_k(x^* - v_k).
\end{aligned}$$

Using convexity, it transforms the inequality into

$$\begin{aligned}
F(x_k) &\leq \alpha_k F(x^*) + (1 - \alpha_k)F(x_{k-1}) + \frac{\alpha_k^2\kappa}{2}\left(\|x^* - v_{k-1}\|^2 - \|v_k - x^*\|^2\right) \\
&\quad - \frac{\kappa}{2}\|\tilde{x}_k - y_k\|^2 + \frac{\kappa\alpha_k}{k+1}\|\tilde{x} - y_k\|\|v_k - x^*\| \\
&= \alpha_k F(x^*) + (1 - \alpha_k)F(x_{k-1}) + \frac{\alpha_k^2\kappa}{2}\left(\|x^* - v_{k-1}\|^2 - \|v_k - x^*\|^2\right) \\
&\quad - \frac{\kappa}{2}\left(\|\tilde{x}_k - y_k\| - \frac{\alpha_k}{k+1}\|v_k - x^*\|\right)^2 + \frac{\kappa}{2}\left(\frac{\alpha_k}{k+1}\right)^2\|v_k - x^*\|^2 \\
&\leq \alpha_k F(x^*) + (1 - \alpha_k)F(x_{k-1}) + \frac{\alpha_k^2\kappa}{2}\left(\|x^* - v_{k-1}\|^2 - \|v_k - x^*\|^2\right) \\
&\quad + \frac{\kappa\alpha_k^2}{2}\left(\frac{1}{k+1}\right)^2\|v_k - x^*\|^2 \\
\iff F(x_k) - F^* &\leq (1 - \alpha_k)(F(x_{k-1}) - F^*) \\
&\quad + \frac{\alpha_k^2\kappa}{2}\left(\|x^* - v_{k-1}\|^2 - \left(1 - \frac{1}{(k+1)^2}\right)\|v_k - x^*\|^2\right)
\end{aligned}$$

Denote $A_k := 1 - 1/(1+k)^2$ to simplify the notations. Continue the simplifications of the above inequality

$$\begin{aligned}
F(x_k) - F^* + \frac{\alpha_k^2 \kappa}{2}\left(1 - \frac{1}{(k+1)^2}\right)\|v_k - x^*\|^2 &\leq (1-\alpha_k)(F(x_{k-1}) - F^*) + \frac{\alpha_k^2\kappa}{2}\|x^* - v_{k-1}\|^2 \\
\iff \alpha_k^{-2}(F(x_k) - F^*) + \frac{\kappa A_k}{2}\|v_k - x^*\|^2 &\leq \alpha_k^{-2}(1-\alpha_k)(F(x_{k-1}) - F^*) + \frac{\kappa}{2}\|x^* - v_{k-1}\|^2 \\
\iff \alpha_k^{-2}(F(x_k) - F^*) + \frac{\kappa A_k}{2}\|v_k - x^*\|^2 &\leq \alpha_{k-1}^{-2}(F(x_{k-1}) - F^*) + \frac{\kappa}{2}\|x^* - v_{k-1}\|^2 \\
&\leq \frac{1}{A_{k-1}}\left(\alpha_{k-1}^{-2}(F(x_{k-1}) - F^*) + \frac{\kappa A_{k-1}}{2}\|x^* - v_{k-1}\|^2\right).
\end{aligned}$$

The second last inequality uses the fact that $(1-\alpha_k)/\alpha_k^2 = \alpha_{k-1}^{-2}$ and $\alpha_1 = 1$. The last inequality used the fact that $A_{k-1} \in (0, 1]$. Simplifying a bit the above is the same as for all $k \geq 1$:

$$\begin{aligned}
\alpha_{k+1}^{-2}(F(x_{k+1}) - F^*) + \frac{\kappa A_k}{2}\|v_k - x^*\|^2 &\leq \frac{1}{A_k}\left(\alpha_k^{-2}(F(x_k) - F^*) + \frac{\kappa A_k}{2}\|v_k - x^*\|^2\right) \\
&\leq \left(\prod_{i=1}^k A_i^{-1}\right)\left(\underbrace{\alpha_1^2(F(x_1) - F^*) + \frac{\kappa A_1}{2}\|v_1 - x^*\|^2}_{=:C}\right) \\
\implies \alpha_{k+1}^{-2}(F(x_{k+1}) - F^*) &\leq \left(\prod_{i=1}^k A_i^{-1}\right) C \\
F(x_{k+1}) - F^* &\leq \alpha_{k+1}^2\left(\prod_{i=1}^k A_i^{-1}\right) C.
\end{aligned}$$

Fortunately we have the big product bounded because

$$\begin{aligned}
\prod_{i=1}^k A_j^{-1} &= \prod_{i=1}^k \left(1 - \frac{1}{(i+1)^2}\right)^{-1} \\
&= \left(\prod_{i=1}^k \left(\frac{(i+1)^2 - 1}{(i+1)^2}\right)\right)^{-1} = \left(\prod_{j=2}^k \left(\frac{j^2-1}{j^2}\right)\right)^{-1} \\
&= \exp\left(\sum_{j=2}^{k+1} \log\left(\frac{j+1}{j}\right) - \log\left(\frac{j}{j-1}\right)\right)^{-1} \\
&= \left(\exp \circ \log\left(\frac{k+3}{k+2}\frac{1}{2}\right)\right)^{-1} \leq \left(\frac{1}{2}\right)^{-1} = 2.
\end{aligned}$$

Therefore,

$$F(x_{k+1}) - F^* \leq \alpha_{k+1}^2 2C \leq \frac{4C}{(k+1)^2}.$$

Which indicates that under controlled inexact evaluations of the inexact proximal step for $\tilde{x}_k$ and $\bar{x}_k$, the basic 4WD Catalyst achieves optimal convergence when F is convex, and it can minimize the limiting subgradient when the function is weakly convex.

Remark D.4 It's tempting to think whether KL conditions assists with convergence to a stationary point for Nesterov's accelerated gradient method. Surveys conducted by the writer in the literatures showed that this remains a huge mystery in general.