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

Hongda Li

November 11, 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 in the convex case 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

Nesterov first proposed the idea of an optimal algorithm named accelerated gradient descent method in his seminal work back in 1983 [7]. It was seminal at the time because the algorithm's upper bound on the iteration complexity sealed the gap between the lower bound for all first order Lipschitz smooth convex function and the upper bound for this class of functions. For a specific definition of the class of algorithms that are considered "First Order", we refer reader to Chapter 2 of Nesterov's new book [8] for more information. In brief the method of gradient descent has an upper bound of $\mathcal{O}(1/k)$ in iteration complexity. It doesn't achieve the $\mathcal{O}(1/k^2)$ lower iteration complexity bound for first order optimization algorithms. The method of accelerated gradient descent has an upper bound of $\mathcal{O}(1/k^2)$, making it optimal.

On first judgement, it's tempting to think that the existence of this optimal algorithm sealed the ceiling for the theoretical development for the entire class of convex first-order smooth optimization. The judgement is correct but lacks the nuance in understanding. The missing piece here is the fact that Nesterov's accelerated gradient is a system of analysis technique instead of any specific design patterns in algorithms.

To demonstrate, the introduction of Guler's works in 1993 [4] proposed an accelerated scheme using the technique of Nesterov's estimating sequence for Proximal Point Method (PPM) in the convex case. Let $(\lambda_k)_{k\geq 0}$ be the sequence of scalars used for regularizing the proximal point method which generates sequence $(x_k)_{k\geq 0}$ given any initial guess x_0. Guler's prior work [3] showed that convergence of PPM method in the convex case has $\mathcal{O}\left(1/\sum_{i=1}^{n}\lambda_i\right)$. His new algorithm using the technique introduced in Nesterov's accelerated gradient achieves a convergence rate of $\mathcal{O}\left(1/(\sum_{i=1}^{n}\sqrt{\lambda_i})^2\right)$. In addition, he also proposed together an inexact Accelerated PPM method using conditions described in Rockafellar's works in 1976 [10].

One would be tempting to conclude that this has sealed the ceiling for research on the topic of extending Nesterov's acceleration. That is indeed correct, but not from a practical point of view. Let $F : \mathbb{R}^n \mapsto \mathbb{R}$ 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]:

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

However, this is troublesome because if we need to approximate the resolvent operator $\mathcal{J}_\lambda$, then it's probably difficult to compute the subgradient $\partial\mathcal{M}(\cdot; y)$, which make it difficult to know when we achieved the required exactness for a PPM evaluation. Otherwise, if we already know the subgradient well, then why approximate it in the first place?

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. [9]). It's called a meta algorithm because it uses other first order algorithm to evaluate inexact proximal point method and then performs the accelerated PPM using Nesterov's acceleration. Their innovations are tracking and controlling the errors made in the inexact PPM throughout the algorithm and some original example usages of the Catalyst framework.

One would be tempting to assert that this has sealed the ceiling for both theories and practice of Nesterov's acceleration hence it must be the center of discussion in this report. The conclusion is indeed correct which it will happen in the sections that follow while the assertion remains open.

1.1 Contributions

The writing is expository and comprehensive. We reviewed the literatures and faithfully reproduced some claims, in addition we give insights into understanding the claim in relations to other papers and foundational ideas in optimization. Three papers by Guler [4] and Lin [5] and Paquette et al. [6] together with Nesterov's [8] method of estimating sequence which is introduced in his book will be covered in detail for this report.

We only cover innovations in the theoretical aspect of Catalyst Acceleration. Detailed applications and specific examples will be out of the scope of this paper simply because there will be too many and requires a comprehensive understanding on a different collection of papers.

2 Preliminaries

Throughout the entire writing, let our ambient space is $\mathbb{R}^n$. We assume the optimization problem of:

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

In this section we introduce the idea of Nesterov's estimating sequence. Nesterov's estimating sequence is fundamental to works in Guler's accelerated PPM method, and Catalyst meta acceleration as a whole.

2.1 Method of Nesterov's Estimating Sequence

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

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

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

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 specifically at the points: x_*, x_k. Additionally, 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:

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

Therefore, it yields a convergence of the sequence $f(x_k) \to f(x^*)$ with a rate relates to sequence $(\alpha_k)_{k \in \mathbb{N}}$.

Much of the analysis of convergence Nesterov's type accelerated gradient method inherit the idea of Nesterov's estimating sequence. Such a proof won't result in simple proof because the construction of ϕ_k is non-trivial, but it comes with the advantage too because we can put creativity into 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 the FISTA algorithm by Beck[2], which is specific case of Algorithm (2.2.63), in Nesterov's book [8]. Discussion on these algorithms are relevant because they share the same format as the Catalyst Acceleration framework and accelerated PPM.

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:*

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

And the sequence of vector 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 *One key component of the Nesterov's estimating sequence is the use of the proximal gradient inequality: $l_F(x; y_k) + \mu/2\|x - y_k\|^2$. In the convex case the function has the property $l_F(\cdot, y) \leq F(\cdot)$ for all y. More precisely, 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. And we have*

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

This is the proximal inequality. Observe that the inequality with proximal gradient term can be interpreted as an example of inexact evaluation of the PPM and the inequality.

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 [8]. 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 [1], Theorem 10.16.

Definition 3.4 (Accelerated proixmal gradient algorithm) *The algorithm of accelerated proximal gradient generates sequence of iterates $(x_k, y_k)_{k\geq 0}$ which satisfies for all $k \geq 0$:*

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

Remark 3.5 The simple case of accelerated gradient descent is stated as (2.2.63) in Nesterov's book [8].

For a proof that proves Definition 3.1 is an estimating sequence, and Definition 3.4 is the accelerated proximal gradient algorithm, please visit 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. In addition, this section will highlight some observations and theoretical results accordingly.

Throughout this section, we assume that $F : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ 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.

Next, we discuss the procedures Guler did to allow the inexact evaluation of proximal method for the accelerated proximal point method. Right out of the Batch he cited Rockafellar[10] on the following conditions for exact evaluations of resolvent on the subgradient of a convex function (Condition (A') in Rockafellar's text):

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

Condition A' also characterize the property of sequence ϵ_k for convergence of the inexact proximal point method. Guler strengthens it in his context and proved the following theorem, which we state it using our notations:

Theorem 4.3 (Guler's inexact proximal point error bound) *Consider defining minimum for the envelope function given by* $\mathcal{M}_k^* := \min_z \mathcal{M}^{\lambda_k}(z; y_k)$. *If* x_{k+1} *is an inexact evaluation under condition (A'), then the estimating sequence admits the conditions that:*

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

We now state the major results (Theorem 3.3) of Guler's 1993 papers on inexact accelerated proximal point 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* $x \in \mathbb{R}^n$*:*

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

The theorem may sound exciting, but as pointed out in Lin 2015 [5] page 11 that the quantities $\mathcal{G}_k^*, \mathcal{J}_{\lambda_k} y_k$ are both intractable. In Guler's work, these intractable quantities were built into the Nesterov's estimating sequence which means it impossible to control how $\epsilon_k \to 0$. If we use the inexact formulation from Guler, it would inevitably result in any algorithm whose definition contains the vector $\mathcal{J}_{\lambda_k} y_k$ exact evaluations of proximal point method.

5 Lin 2015

This section introduced the setup of Nesterov's estimating sequence in Lin 2015 [5]. Right at the beginning we warn the readers about the following:

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

(ii) It's context specific for controlling the errors of inexact proximal point evaluations are being controlled. He hinted at the way of controlling and tracking the errors from inexact proximal point evaluations, but it's context specific. He only illustrated the use of the meta acceleration on their own method called: "Proximal MISO", but in general the problem still remains open.

(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 [8] 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 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) 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 most exciting thing about this algorithm is the similarity it has compares to Definition 3.4. The only difference here is the inclusion of an inexact proximal point step, 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]. The lemma allows the analysis Nesterov's estimating sequence to be carried through without using intractable quantities: $\mathcal{M}^{\kappa^{-1}}(\mathcal{J}_{k^{-1}}y_{k-1}, y_{k-1}), \mathcal{J}_{\kappa^{-1}}y_{k-1}$.

6 Non-convex Extension of Catalyst Acceleration

The non-convex extension of Catalyst acceleration by Lin 2018 [6] remains similar to the algorithm in the convex case in his 2015 paper [5]. The algorithm is now adapted to handle function with unknown weak convexity constant ρ using the process called Auto Adapt subroutine. There is no accelerated rate for the non-convex case since it's impossible, but the theoretical convergence to stationary point is claimed for the non-convex case.

References

[1] A. Beck, *First-order Methods in Optimization*, MOS-SIAM Series in Optimization, SIAM, israel, 2017.

[2] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.

[3] O. Guler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM Journal on Control and Optimization, 29 (1991), p. 17. Num Pages: 17 Place: Philadelphia, United States Publisher: Society for Industrial and Applied Mathematics.

[4] O. Güler, *New Proximal Point Algorithms for Convex Minimization*, SIAM Journal on Optimization, 2 (1992), pp. 649–664. Publisher: Society for Industrial and Applied Mathematics.

[5] H. Lin, J. Mairal, and Z. Harchaoui, *A Universal Catalyst for First-Order Optimization*, MIT Press, Dec. 2015, p. 3384.

[6] ——, *Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice*, in Journal of Machine Learning Research, vol. 18, 2018, pp. 1–54.

[7] Y. Nesterov, *A method for solving the convex programming problem with convergence rate O(1/k^2)*, Proceedings of the USSR Academy of Sciences, (1983).

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

[9] 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. ISSN: 2640-3498.

[10] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, 14, pp. 877–898.

A Postponed proofs

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 $\phi_k : \mathbb{R}^n \mapsto \mathbb{R}$ as a sequence of functions such that it satisfies recursively for all $k \geq 0$ the following conditions*

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

Where $(y_k)_{k\geq 0}$ is any auxiliary sequence. If we define the canonical form for ϕ_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 ϕ_k:

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

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 ϕ_{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 ϕ_{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 $\|v_{k+1} - y_k\|^2$. 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 ϕ_k and auxiliary sequence $y_k, v_k, \gamma_k, \alpha_k$ be given by Theorem A.3. If for all $k \geq 0$ they verify:

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

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}:

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

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 ϕ_{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 α_k, γ_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 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.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 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:

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

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.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 α_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, 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.

■

B Proofs for accelerated PPM

C Proofs for Catalyst Meta Acceleration

D Proofs for 4WD Catalyst Acceleration