Catalyst Meta Acceleration Framework: The Gist of its Theories

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.

2010 Mathematics Subject Classification: Primary 65K10, 90c25, 90C30; Secondary 65Y20. **Keywords:** Nesterov acceleration, Proximal point method.

1 Introduction

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 won't contain major results. 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.

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* $\forall x \in \mathbb{R}^n$*:*

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

Unrolling the above recursion it yields:

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

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

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

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 $(\phi_k)_{k\ge 0}$.

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 $\mu \ge 0$ 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:

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

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.

5 Lin 2015

6 Non-convex Extension of Catalyst Acceleration

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/kˆ2)*, 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

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

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:

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

■

Theorem A.2 (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:

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

Next move is to simplify the term $\|v_{k+1} - y_k\|^2$. With that it produces:

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

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 \phantom{\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(\frac{\gamma_{k+1} - \gamma_k(1-\alpha_k)}{\gamma_{k+1}}\right) \\
&\quad \phantom{\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(\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.3 (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.2. If for all $k \geq 0$ they verify:

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

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

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

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 $\forall z \in \mathbb{R}^n$:

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

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

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

■