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

Hongda Li

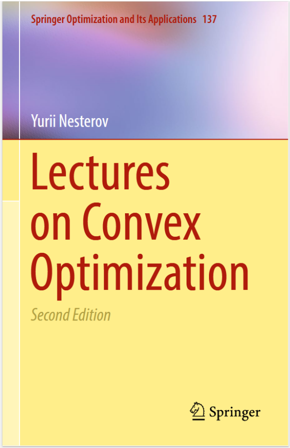
UBC Okanagan

November 14, 2024

ToC

1. Introduction
   - Major references
   - Nesterov's Estimating Sequence
   - Example: Accelerated Proximal Gradient
2. Guler 1992
   - Exact accelerated PPM
   - Inexact accelerated PPM
3. Lin 2015
   - The Catalyst Algorithm
   - The algorithm
   - Convergence and practical importance
   - Key theoretical innovations
4. Paquette 2018
   - Major contributions
   - The Basic 4WD Catalyst
   - Convergence claims
   - Convergence Proofs
5. Morals of the story
6. References

Nesterov's Book



- Yurri Nesterov's book: "Lectures on Convex Optimization" 2018, Springer [3].

Accelerated Proximal Point Method (PPM)

SIAM J. OPTIMIZATION
Vol. 2, No. 4, pp. 649–664, November 1992

NEW PROXIMAL POINT ALGORITHMS FOR CONVEX MINIMIZATION*

OSMAN GÜLER†

Abstract. This paper introduces two new proximal point algorithms for minimizing a proper, lower-semicontinuous convex function $f:\mathbf{R}^n \to R \cup \{\infty\}$. Under this minimal assumption on f, the first algorithm possesses the global convergence rate estimate $f(x_k) - \min_{x\in\mathbf{R}^n} f(x) = O(1/(\sum_{j=0}^{k-1}\sqrt{\lambda_j})^2)$, where $\{\lambda_k\}_{k=0}^{\infty}$ are the proximal parameters. It is shown that this algorithm converges, and global convergence rate estimates for it are provided, even if minimizations are performed inexactly at each iteration. Both algorithms converge even if f has no minimizers or is unbounded from below. These algorithms and results are valid in infinite-dimensional Hilbert spaces.

Key words. proximal point algorithms, global convergence rates, augmented Lagrangian algorithms, convex programming

AMS(MOS) subject classifications. primary 90C25; secondary 49D45, 49D37

- Osman Guler's: "New proximal point algorithm for convex optimization", SIAM J. Optimization 1992 [1].

A Universal Catalyst for First-Order Optimization

Hongzhou Lin[1], Julien Mairal[1] and Zaid Harchaoui[1,2]
[1]Inria [2]NYU
{hongzhou.lin,julien.mairal}@inria.fr
zaid.harchaoui@nyu.edu

Abstract

We introduce a generic scheme for accelerating first-order optimization methods in the sense of Nesterov, which builds upon a new analysis of the accelerated proximal point algorithm. Our approach consists of minimizing a convex objective by approximately solving a sequence of well-chosen auxiliary problems, leading to faster convergence. This strategy applies to a large class of algorithms, including gradient descent, block coordinate descent, SAG, SAGA, SDCA, SVRG, Finito/MISO, and their proximal variants. For all of these methods, we provide acceleration and explicit support for non-strongly convex objectives. In addition to theoretical speed-up, we also show that acceleration is useful in practice, especially for ill-conditioned problems where we measure significant improvements.

(a) Lin 2015

Catalyst Acceleration
for Gradient-Based Non-Convex Optimization

Courtney Paquette
University of Waterloo
c2paquette@uwaterloo.ca

Hongzhou Lin
MIT
hongzhou@mit.edu

Dmitriy Drusvyatskiy
University of Washington
ddrusv@uw.edu

Julien Mairal
Inria*
julien.mairal@inria.fr

Zaid Harchaoui
University of Washington
zaid@uw.edu

January 3, 2019

Abstract

We introduce a generic scheme to solve nonconvex optimization problems using gradient-based algorithms originally designed for minimizing convex functions. Even though these methods may originally require convexity to operate, the proposed approach allows one to use them on weakly convex objectives, which covers a large class of non-convex functions typically appearing in machine learning and signal processing. In general, the scheme is guaranteed to produce a stationary point with a worst-case efficiency typical of first-order methods, and when the objective turns out to be convex, it automatically accelerates in the sense of Nesterov and achieves near-optimal convergence rate in function values. These properties are achieved without assuming any knowledge about the convexity of the objective, by automatically adapting to the unknown weak convexity constant. We conclude the paper by showing promising experimental results obtained by applying our approach to incremental algorithms such as SVRG and SAGA for sparse matrix factorization and for learning neural networks.

(b) Paquette 2018

- Honzhou Lin et al. "Universal Catalyst for first order optimization", 2015 JMLR [2].
- Paquette et al. "Catalyst for gradient-based non-convex optimization", 2018 JMLR [4].

Objectives of the Talk

List of objectives

1. Introduce the technique of Nesterov's estimating sequence.
2. Survey the history of the Catalyst algorithm (Catalyst for short).
3. Survey the theories behind Catalyst.
4. Highlight key innovations.
5. Introduce the non-convex extension of Catalyst.

Objectives of the Talk

List of objectives

1. Introduce the technique of Nesterov's estimating sequence.
2. Survey the history of the Catalyst algorithm (Catalyst for short).
3. Survey the theories behind Catalyst.
4. Highlight key innovations.
5. Introduce the non-convex extension of Catalyst.

A note on the scope

Specific applications and algorithms are outside the scope because variance reduced stochastic method is itself a big topic.

Nesterov's Estimating Sequence

Definition (Nesterov's estimating sequence)

Let $(\phi_k : \mathbb{R}^n \to \mathbb{R})_{k \geq 0}$ be a sequence of functions. It's an estimating sequence when it satisfies the conditions:

1. $\exists (x_k)_{k \geq 0}$ such that for all $k \geq 0$ it has $F(x_k) \leq \phi_k^* := \min_x \phi_k(x)$.
2. $\exists (\alpha_k)_{k \geq 0}$ where $\alpha_k \in (0, 1)$ such that $\phi_{k+1}(x) - \phi_k(x) \leq -\alpha_k(\phi_k(x) - F(x))$ for all $k \geq 0, x \in \mathbb{R}^n$.

Nesterov's Estimating Sequence and Convergence

Observations

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

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

Unroll the recurrence, by setting $x = x^*$, $\Delta_k(x^*)$ is non-negative. By the properties of Nesterov's estimating:

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

Example: Accelerated Proximal Gradient

The following algorithm is proved in the report which it's similar to Nesterov's 2.2.20 in his book [3].

Quick Notations

Assume $F = f + g$ where f is L-Lipschitz smooth and $\mu \geq 0$ strongly convex and g is convex. Define

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

Example: Accelerated Proximal Gradient

Definition (estimating sequence for Accelerated Proximal Gradient)

$(\phi_k)_{k\geq 0}$ is the Nesterov's estimating sequence recursively given by:

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

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

The Algorithm generates a sequence of vectors y_k, x_k, and scalars 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 $\alpha_0 \in (0,1)$.

Guler 1992: Accelerated Proximal Point Method

Guler in 1992 discovered the following:

1. The method of proximal point can be accelerated via Nesterov's estimating sequence.
2. The accelerated convergence rate retains for certain magnitude of errors on inexact evaluation of proximal point method.

Quick notations

We use the following list of notations:

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

We use $\mathcal{G}_k, \mathcal{J}_k, \mathcal{M}_k$ as a short for $\mathcal{G}_{\lambda_k}, \mathcal{J}_{\lambda_k}, \mathcal{M}_{\lambda_k}$. $(\lambda_k)_{k\geq 0}$ is a sequence that controls proximal operator.

Estimating sequence of accelerated PPM

Definition (Accelerated PPM estimating sequence)

$(\phi_k)_{k\geq 0}$ has for all $k \geq 0$, any $A \geq 0$:

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

$(\lambda_k)_{k\geq 0}$, $x_k = \mathcal{J}_\lambda y_k$. Auxiliary vectors (y_k, v_k), and $(\alpha_k, A_k)_{k\geq 0}$ satisfies $k \geq 0$:

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

Convergence of accelerated PPM

We now state a major result of Guler 1992 [1] on accelerated PPM.

Accelerated rate of convergence from Guler

The accelerated PPM generate $(x_k)_{k\geq 0}$ such that $F(x_k) - F^*$ converges at a rate of:

$$\mathcal{O}\left(\frac{1}{\left(\sum_{i=1}^{k}\sqrt{\lambda_i}\right)^2}\right).$$

Note, PPM without accelerate converges at a rate of $\mathcal{O}((\sum_{i=1}^{k}\lambda_i)^{-1})$.

Accelerated Inexact PPM

Guler cited Rockafellar 1976 [5] for condition (A'):

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

Putting things into the context of accelerated PPM, the theorem follows is pivotal:

Theorem (Guler's inexact proximal point error bound (Lemma 3.1))

Define Moreau Envelope at y_k: $\mathcal{M}_k^* := \min_z \mathcal{M}^{\lambda_k}(z; y_k)$. If x_{k+1} is an inexact evaluation under condition (A'), then:

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

Guler's Major Results

The following is a major result from Guler [1] on inexact accelerated PPM.

Theorem (Guler's accelerated inexact PPM convergence (Theorem 3.3))

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

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

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

Guler's Major Results

The following is a major result from Guler [1] on inexact accelerated PPM.

Theorem (Guler's accelerated inexact PPM convergence (Theorem 3.3))

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

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

If $\sigma \geq 3/2$, the method converges at a rate of $\mathcal{O}(1/k^2)$. It looks exciting, but it's not exciting for practical purposes because:

1. Determining $(\epsilon_k)_{k\geq 0}$ requires knowledge on ϕ_k^*.
2. ϕ_k^* is expressed with intractable quantity: $F(\mathcal{J}_k y_k)$.

So the algorithm contains intractable quantities: $F(\mathcal{J}_k y_k)$. It's not yet ready to be formulated into a concrete algorithm.

Lin 2015

Hongzhou Lin 2015 [2] did the following:

1. Improved the proof from Guler 1992 to include strongly convex objectives.
2. Showed that $(\epsilon_k)_{k\geq 0}$ can be determined algorithmically and an accelerated rate can be achieved.
3. Invented his own accelerated variance reduced incremental method called: "Accelerated MISO-Prox" to demonstrate the Catalyst Framework.
4. First time in history he obtained an accelerated rate for incremental methods in general.

Quick notations

Assume F is a $\mu \geq 0$ strongly convex function. Fix κ and the notations are:

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

Lin's accelerated proximal point method

Definition (Lin's accelerated proximal point method)

Initialize any $y_0 = x_0 \in \mathbb{R}^n$, fix parameters κ and α_0. Let $(\epsilon_k)_{k\geq 0}$ be the error of inexact proximal point evaluations. Then the algorithm generates (x_k, y_k) satisfies for all $k \geq 1$:

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

Lin's accelerated proximal point method

Definition (Lin's accelerated proximal point method)

Initialize any $y_0 = x_0 \in \mathbb{R}^n$, fix parameters κ and α_0. Let $(\epsilon_k)_{k\geq 0}$ be the error of inexact proximal point evaluations. Then the algorithm generates (x_k, y_k) satisfies for all $k \geq 1$:

$$\text{find } x_k \approx \mathcal{J}_{\kappa^{-1}} y_{k-1} \text{ such that } \mathcal{M}^{\kappa^{-1}}(x_k, y_{k-1}) - \mathcal{M}^{\kappa^{-1}}(\mathcal{J}_{\kappa^{-1}} y_{k-1}, y_{k-1}) \leq \epsilon_k$$

$$\text{find } \alpha_k \in (0,1) \text{ such that } \alpha_k^2 = (1-\alpha_k)\alpha_{k-1}^2 + (\mu/(\mu+\kappa))$$

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

1. It's similar compared to accelerated proximal gradient!!!
2. Determining $(\epsilon_k)_{k\geq 1}$ requires knowledge about $F(x_0) - F^*$.
3. Strong convexity of $\mathcal{M}_k(\cdot; y_{k-1})$ helps with approximating x_k up to ϵ_k (i.e: PL Inequality).

Practical importance

A major result on page 6 of Lin's 2015 [2].

Accelerated convergence

Assume F is $\mu \geq 0$ strongly convex and L-Lipschitz smooth. Using full gradient, or randomized coordinate descent to evaluation $x_k \approx \mathcal{J}_{\kappa^{-1}} y_{k-1}$ up to ϵ_k, then the overall complexity is:

$$\widetilde{\mathcal{O}}\left(n\sqrt{L/\mu}\log(1/\epsilon)\right).$$

It's the same as accelerated gradient method up to a log term. In absent of strong convexity, acceleration for Variance reduced stochastic method such as: SAG, SAGA, Finito/MISO-Prox, SDCA, SVRG becomes $\widetilde{\mathcal{O}}(nL/\sqrt{\epsilon})$, strictly faster $\mathcal{O}(nL/\epsilon)$ without acceleration.

Note: SAG, SAGA, Finito/MISO-Prox, SDCA, SVRG are examples of variance reduced incremental methods. $\widetilde{\mathcal{O}}$ hides log factor in complexity.

Inexact proximal inequality in Lin 2015

Lemma A.7 in Lin 2015 [2] stated the following:

Lemma (Inexact proximal inequality)

Let F be a $\mu \geq 0$ strongly convex and $\kappa > 0$ fixed. Suppose $x_k \approx \mathcal{J}_{\kappa^{-1}} y_{k-1}$ satisfies $\mathcal{M}^{\kappa^{-1}}(x_k; y_{k-1}) - \mathcal{M}^{\kappa^{-1}}(\mathcal{J}_{\kappa^{-1}} y_{k-1}, y_{k-1}) \leq \epsilon_k$. Denote $x_k^* = \mathcal{J}_{\kappa^{-1}} y_{k-1}$ to be the exact evaluation of the proximal point then for all x:

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

1. If $x_k = x_k^*$, we get the red part only which is the proximal inequality.
2. This inequality helps to define an estimating sequence $(\phi_k)_{k\geq 0}$ without the term: $\mathcal{J}_{\kappa^{-1}} y_{k-1}$, which is x_k^* in this case.

Major contribution in Paquette 2018

In Paquette 2018, these major improvements for Lin's Universal Catalyst had been made:

1. It supports weakly convex function with an unknown weak convexity constant through a procedures call Auto Adapt.
2. The convergence to stationary point under weak convexity is claimed.
3. The method retains accelerated convergence rate if the function is convex.

Note: F is ρ-weakly convex if and only if $f + \rho/2\|\cdot\|^2$ is convex.

Quick notations

Fix κ we use the following notations:

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

Basic 4WD

Definition (Basic 4WD Catalyst Algorithm)

Find any $x_0 \in \operatorname{dom}(F)$. 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:

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

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

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

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

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

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

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

Convergence claim

We state the convergence results for Basic 4WD Catalyst from Paquette 2018 [4].

Theorem (Basic 4WD Catalyst Convergence)

Let (x_k, v_k, y_k) be generated by the basic 4wd Catalyst algorithm. If F is weakly convex and bounded below, then x_k converges to a stationary point where

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

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

Convergence claim

We state the convergence results for Basic 4WD Catalyst from Paquette 2018 [4].

Theorem (Basic 4WD Catalyst Convergence)

Let (x_k, v_k, y_k) be generated by the basic 4wd Catalyst algorithm. If F is weakly convex and bounded below, then x_k converges to a stationary point where

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

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

Let's prove this.

- The convergence to stationary point is true for any $\kappa > 0$ that is not necessarily lager than the weak convexity constant.

Auxiliary Sequence Bounds

Lemma (Nesterov's Estimating sequence auxiliary sequence bounds)

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 $k \geq 0$:

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

Note: $1/\alpha_k$ would be the FISTA sequence.

Proof.

Skipped, see report. □

Proximal Stationary Point

Lemma (Lemma B.2)

Assume that F is weakly convex. Fix any y, suppose that y^+ satisfies $\operatorname{dist}(\mathbf{0}, \partial\mathcal{M}(y^+; y)) \le \epsilon$. *Then the following inequality holds:*

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

Proof.

Skipped, see report. □

Basic 4WD Catalyst Convergence

A short proof

Algorithm has:

$$F(x_{k-1}) \geq \mathcal{M}(\bar{x}_k, x_{k-1}) \geq F(x_k) + \frac{\kappa}{2}\|\bar{x}_k - x_{k-1}\|^2. \tag{ineq1}$$

By $F(x_k) = \min(F(\bar{x}_k), F(\tilde{x}_k))$. Using Lemma B.2, set $\epsilon = \kappa\|\bar{x}_k - x_{k-1}\|$, $y = x_{k-1}$, $y^+ = \bar{x}_k$ then

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

Basic 4WD Catalyst Convergence

A short proof

Using (ineq1):

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

Basic 4WD Catalyst Convergence

A short proof

Now assume F is convex with minimum F^* and minimizer x^*. By convexity $\mathcal{M}(\cdot, y_k)$ is κ strong convex. By the algorithm there exists $\xi_k \in \partial\mathcal{M}(\tilde{x}_k, y_k)$ such that $\|\xi_k\| \leq \frac{\kappa}{k+1}\|\tilde{x}_k - y_k\|$. Therefore, it has for all x:

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

Basic 4WD Catalyst Convergence

A short proof

Set $x = \alpha_k x^* + (1-\alpha_k)x_{k-1}$ where x^* is the minimizer. Then by the algorithm:

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

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

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

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

$$\begin{aligned}
&= \alpha_k x^* + (1-\alpha_k)x_{k-1} - (x_{k-1} + \alpha_k(v_k - x_{k-1})) \\
&= \alpha_k x^* - \alpha_k x_{k-1} - \alpha_k(v_k - x_{k-1}) \\
&= \alpha_k (x^* - v_k).
\end{aligned}$$

Basic 4WD Catalyst Convergence

A short proof

Substituting back, use F convex and $\alpha_k \in (0,1), k \geq 1$:

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

Basic 4WD Catalyst Convergence

A short proof

Denote $A_k := 1 - 1/(1+k)^2$ to simplify the notations. Rearranging and use $(1-\alpha_k)/\alpha_k^2 = \alpha_{k-1}^{-2}$ it has for all $k \geq 2$:

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

Basic 4WD Catalyst Convergence

A short proof

Make k into $k+1$ so for all $k \geq 1$:

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

Basic 4WD Catalyst Convergence

A short proof

Finally,

$$\prod_{i=1}^{k} A_j^{-1} = \prod_{i=1}^{k} \left(1 - \frac{1}{(i+1)^2}\right)^{-1}$$
$$\leq \left(1 - \frac{1}{4}\right)^{-1} \leq 2.$$

So by the lemma for α_k:

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

The morals of the story

Morals of the story

Besides all the gory details of the theories, the historical development reveals crucial aspects and general patterns in developing theories for optimization algorithms.

1. Identify existing theoretical frameworks relevant to optimization algorithms but still leave rooms for creativity, i.e: Nesterov's acceleration and inexact proximal point.
2. Always assume inexact evaluations of proximal point for flexibility and:
    1. "outsource" it to existing algorithms;
    2. control the errors and keep track of the complexity;
    3. show that it still has a favorable complexity at the end.
3. Finally, and perhaps most importantly: identify the demands for applications, in this case it was: Machine Learning and accelerating existing incremental methods.

References

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

H. Lin, J. Mairal, and Z. Harchaoui, *A universal catalyst for first-order optimization*, MIT Press, Dec. 2015, p. 3384.

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

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.

R. T. Rockafellar, *Monotone Operators and the Proximal Point Algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.