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

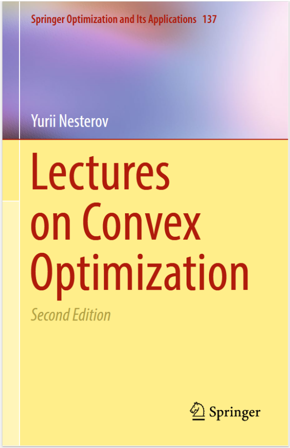
Hongda Li

UBC Okanagan

November 12, 2024

ToC

Nesterov's Book



- Yurri Nesterov's book: "Lectures on Convex Optimization" 2018, Springer [1].

Accelerated Proximal Point Method

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. [2]

Catalyst Acceleration

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 JLMR [3].
- Paquette et al. “Catalyst for gradient-based nonconvex optimization” 2018 JLMR [4].

Objectives of the Talk

List of objectives

1. Introduce the technique of Nesterov's estimating sequence for convergence proof of algorithms.
2. Understand the historical context for the inspirations of the Catalyst algorithm.
3. Understand the theories behind the Catalyst meta acceleration.
4. Understand key innovations for controlling the errors in Catalyst accelerations.
5. Introduce the Non-convex extension of the method.

A note on the scope

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

Objectives of the Talk

List of objectives

1. Introduce the technique of Nesterov's estimating sequence for convergence proof of algorithms.
2. Understand the historical context for the inspirations of the Catalyst algorithm.
3. Understand the theories behind the Catalyst meta acceleration.
4. Understand key innovations for controlling the errors in Catalyst accelerations.
5. Introduce the Non-convex extension of the method.

A note on the scope

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

Quick Notations

Nesterov's Estimating Sequence

Definition (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:

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

Nesterov's Estimating Sequence and Convergence

Observations

If we dsefine 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 $\forall k \geq 0$:

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

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

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

Example: accelerated proximal gradient

Quick Notations

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

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

Example: accelerated proximal gradient

Definition (Accelerated proximal gradient estimating sequence)

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

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

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

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

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

References

Y. Nesterov, *Lectures on Convex Optimization*, ser. Springer Optimization and Its Applications. Cham: Springer International Publishing, 2018, vol. 137. [Online]. Available: http://link.springer.com/10.1007/978-3-319-91578-4

O. Güler, "New Proximal Point Algorithms for Convex Minimization," *SIAM Journal on Optimization*, vol. 2, no. 4, pp. 649–664, Nov. 1992, publisher: Society for Industrial and Applied Mathematics. [Online]. Available: https://epubs.siam.org/doi/10.1137/0802032

H. Lin, J. Mairal, and Z. Harchaoui, "A Universal Catalyst for First-Order Optimization." MIT Press, Dec. 2015, p. 3384. [Online]. Available: https://inria.hal.science/hal-01160728

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: