Nesterov Type Momentum Methods

Alto Legato *

May 29, 2024

Abstract

These are ntoes for Nesterov Type Acceleration Methods, in the convex case. They may get made into papers, proposal, and thesis in the future.

2010 Mathematics Subject Classification: Primary 47H05, 52A41, 90C25; Secondary 15A09, 26A51, 26B25, 26E60, 47H09, 47A63. **Keywords:**

1 Preliminaries

In this section we list fundational results that are important for proofs in coming sections. For this section, let the ambient space be $\mathbb{R}^n$ and $\|\cdot\|$ be the Euclidean 2 norm until it's specified in the context.

1.1 Lipschitz Smoothness

Definition 1.1 (Lipschitz Smooth) *Let f be differentiable. It has Lipschitz smoothness with constant L if for all x, y*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

Theorem 1.2 (Lipschitz Smoothness Equivalence) *With f convex, the following conditions are equivalent conditions for all x, y:*

*Subject type, Some Department of Some University, Location of the University, Country. E-mail: author.name@university.edu.

(i) $|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2$.

(ii) $|\langle \nabla f(y) - \nabla f(x), y - x \rangle| \leq L\|y - x\|^2$.

(iii) $x^+ \in \arg\min_x f(x) \implies \frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - f(x^+) \leq (L/2)\|x - x^+\|^2$

(iv) $1/(2L)\|x - y\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq (L/2)\|x - y\|^2$

Remark 1.3 In the convex of operator theorem, Lipschitz smoothness of the gradient of a convex function is an example of a Firmly Nonexpansive operator.

Definition 1.4 (Strong Convexity) *With $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$, it is strongly convex with constant α if and only if $f - (\alpha/2)\|\cdot\|^2$ is a convex function.*

Theorem 1.5 (Strongly Convex Equivalent Results) *With $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ α-strongly convex, the following conditions are equivalent conditions for all x, y:*

(i) $f(y) - f(x) - \langle \partial f(x), y - x \rangle \geq \frac{\alpha}{2}\|y - x\|^2$

(ii) $\langle \partial f(y) - \partial f(x), y - x \rangle \geq \alpha\|y - x\|^2$.

(iii) $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \alpha\frac{\lambda(1-\lambda)}{2}\|y - x\|^2, \forall \lambda \in [0, 1]$.

Theorem 1.6 (Strong Convexity Implications) *With $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ α-strongly convex, the following conditions are implied:*

Remark 1.7 In the context of operator theory, the subgradient of a strongly convex function is an example of a Strongly Monotone Operator.

1.2 Proximal Descent Inequality

Theorem 1.8 (Proximal Descent Inequality) *With $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}^n$ convex, fix any $x \in \mathbb{R}^n$, let $p = \operatorname{prox}_\alpha f(x)$, then for all y we have inequality*

$$\left(f(p) + \frac{1}{2\alpha}\|x - p\|^2\right) - \left(f(y) + \frac{1}{2\alpha}\|x - y\|^2\right) \leq -\frac{1}{2\alpha}\|y - p\|^2.$$

Recall $\operatorname{prox}_\alpha f(x) = \underset{u}{\operatorname{argmin}} \left\{f(u) + \frac{1}{2}\|u - x\|^2\right\}$.

We make use of this theorem in the proof of convergence of proximal point method.

Remark 1.9 This descent inequality can be generalized to bregman proximal mapping as well.

2 The Proximal Point Method in the Convex Case

In this section we quickly go over the analysis of Proximal point method (PPM) in the convex case and see how the theories can be generalized into the cases where PPM is approximated.

With $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ lsc proper and convex, given any x_0 the PPM generates sequence $(x_n)_{n\in\mathbb{N}}$ by $x_{k+1} = \text{prox}_{\eta_{k+1}} f(x_k)$ for all $k \in \mathbb{N}$ where the sequence $(\eta_k)_{k\in\mathbb{N}}$ is a nonegative sequence of real numbers.

2.1 The Lyapunov function of PPM

Theorem 2.1 *Define the quantity for all $u \in \mathbb{R}^n$*

$$\Phi_t = \left(\sum_{i=1}^{t} \eta_i\right)(f(x_t) - f(u)) + \frac{1}{2}\|u - x_t\|^2 \ \forall t \geq 1,$$
$$\Phi_0 = (1/2)\|x_0 - u\|^2,$$

then it is a Lyapunov function for the PPM algorithm. Meaning for all $(x_k)_{k\in\mathbb{N}}$ generated by PPM, it satisfies that $\Phi_{t+1} - \Phi_t \leq 0$. Additionally, by the definition we have

$$\Phi_{t+1} - \Phi_t = \eta_{t+1}(f(x_{t+1}) - f(u)) + \frac{1}{2}\|u - x_{t+1}\|^2 - \frac{1}{2}\|u - x_t\|^2 \leq -\frac{1}{2}\|x_{t+1} - x_t\|^2,$$

as a consequence of choosing $u = x_{t+1}$

$$f(x_{t+1}) - f(x_t) \leq \frac{1}{\eta_{t+1}}\|x_{t+1} - x_t\|^2.$$

Proof. With p, α, y in theorem 1.8 as x_{t+1}, η_{t+1}, u we have for all u

$$\eta_{t+1}(f(x_{t+1}) - f(u)) + \frac{1}{2}(\|x_t - x_{t+1}\|^2 - \|x_t - u\|^2) \leq -\frac{1}{2}\|u - x_{t+1}\|^2$$
$$\eta_{t+1}(f(x_{t+1}) - f(u)) + \frac{1}{2}\|u - x_{t+1}\|^2 - \frac{1}{2}\|u - x_t\|^2 \leq -\frac{1}{2}\|x_{t+1} - x_t\|^2,$$

next it's not hard to check that $\Phi_{t+1} - \Phi_t$ equals to the above quantity. ■

Remark 2.2 This theorem act as a descent inequality and it can be generalized under a diverse context, such as all algorithms that are approximation to the PPM method.

Theorem 2.3 (Convergence Rate of PPM) *The convergence rate of PPM applied to f, closed, convex proper, we have convergence rate of the function value:*

$$f(x_T) - f(x_*) \leq O\left(\left(\sum_{i=1}^{T} \eta_t\right)^{-1}\right).$$

Where x_* is the minimizer of f.

Proof. With $\Delta_t = f(x_t) - f(x_*), \Upsilon_t = \sum_{i=1}^{t} \eta_i$ so $\Phi_t = \Upsilon_t \Delta_t + \frac{1}{2}\|x_t - x_*\|^2$ by consider $u = x_*$, invoking previous theorem and do

$$\Upsilon_T \Delta_T \leq \Phi_T \leq \Phi_0 = \frac{1}{2}\|x_0 - x_*\|^2$$
$$\implies \Delta_T \leq \frac{1}{2\Upsilon_T}\|x_0 - x_*\|^2.$$

Hence, the convergence rate of PPM for all convex function is the above. ■

Remark 2.4 The analysis of the above is taken from (REFERENCE NEEDED).

However, the convergence rate of the algorithm can be faster for the same choice of the sequence $(\eta_t)_{t\in\mathbb{N}}$.

Theorem 2.5 (Convergence Rate of PPM (Strongly Convex))

3 Application of the Proximal Point Method

The PPM method and the Lyaounov function derived above serves as tamplate for other algorithms. In optimizations, people uses lower and upper approximation of the objective function to approximate the PPM. Such an approach involves a diverse range of algorithms, including second order algorithms such as Newton's method. To demonstrate, assume that f is a lsc convex function such that it can be approximated by an lower bounding function $l_f(x|\bar{x})$ at $\bar{x}$ such that it satisfies for all x:

$$l_f(x|\bar{x}) \leq f(x) \leq l_f(x|\bar{x}) + \frac{L}{2}\|x - \bar{x}\|^2. \tag{1}$$

The above characterization is generic enough to include the case where $l_f(x|\bar{x})$ under approximates function that is non-smooth. To make use of the theorems discussed previously, we assume that $l_f(x|\bar{x})$ is convex for all x, at all $\bar{x}$.

The approximated proximal point method is applying PPM to the function $l_f(x|x_t)$ for each iteration, i.e: $x_{t+1} = \text{prox}_{\eta_{t+1}}[u \mapsto l_f(u|x_t)](x_t)$.

Theorem 3.1 (Convergence of Approximated PPM) *With f be a function that has minimizer: x_*; $l_f(x; x_t)$ be a convex, lsc, proper lower-bounding function then the lower-bounding φ function satisfies inequality:*

$$\varphi_t(x) \leq \eta_{t+1} f(x) \leq \varphi_t(x) + \frac{L\eta_{n+1}}{2}\|x - x_t\|^2 \quad \forall x \in \mathbb{R}^n,$$

Assume an algorithm the makes:

$$x_{t+1} = \operatorname*{argmin}_x \left\{ l_f(x; x_t) + \frac{1}{2\eta_{t+1}}\|x - x_t\|^2 \right\},$$

Then the iterates satisfies:

$$\eta_{t+1}(f(x_{t+1}) - f(x_*)) + \frac{1}{2}\|x_* - x_{t+1}\|^2 - \frac{1}{2}\|x_* - x_t\| \leq \left(\frac{L\eta_{n+1}}{2} - \frac{1}{2}\right)\|x_{t+1} - x_t\|^2.$$

Proof.

■