Nesterov Type Momentum Methods

Alto *

June 13, 2024

Abstract

These are notes for Nesterov Type Acceleration Methods in the convex case. They can be made into papers, proposals, and a thesis in the future.

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

1 Preliminaries

This section lists foundational results important for proof in the coming sections. For this section, let the ambient space be $\mathbb{R}^n$ and $\|\cdot\|$ be the 2-norm until specified in the context. For a general overview of smoothness and strong convexity in the Euclidean space, see [8, theorem 2.1.5, theorem 2.1.10] for a full exposition of the topic.

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*

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

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

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

(i) $L^{-1}\|\nabla f(y) - \nabla f(x)\|^2 \leq \langle \nabla f(y) - \nabla f(x), y - x\rangle \leq L\|y - x\|^2$.

(ii) $x^+ \in \underset{x}{\operatorname{argmin}} f(x) \implies \frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - f(x^+) \leq (L/2)\|x - x^+\|^2$, *co-coersiveness.*

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

Remark 1.3 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 (Stong convexity equivalences) *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:*

(i) $\frac{1}{2}\operatorname{dist}(\mathbf{0}; \partial f(x))^2 \geq \alpha(f(x) - f^+)$ *where f^+ is a minimum of the function, and this is called the Polyak-Lojasiewicz (PL) inequality.*

(ii) $\forall x, y \in \mathbb{E}, u \in \partial f(x), v \in \partial f(y) : \|u - v\| \geq \alpha\|x - y\|$.

(iii) $f(y) \leq f(x) + \langle \partial f(x), y - x\rangle + \frac{1}{2\alpha}\|u - v\|^2, \forall u \in \partial f(x), v \in \partial f(y)$.

(iv) $\langle \partial f(x) - \partial f(y), x - y\rangle \leq \frac{1}{\alpha}\|u - v\|^2, \forall u \in \partial f(x), v \in \partial f(y)$.

(v) *if $x^+ \in \arg\min_x f(x)$ then $f(x) - f(x^+) \geq \frac{\alpha}{2}\|x - x^+\|^2$ and x^+ is a unique minimizer.*

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

1.2 Proximal descent inequality

The proximal descent inequality below is a crucial piece of inequality for deriving the behaviours of algorithms.

Theorem 1.8 (Proximal Descent Inequality) *With $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}^n$ β-convex where $\beta \geq 0$, fix any $x \in \mathbb{R}^n$, let $p = \text{prox}_f(x)$, then for all y we have inequality*

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

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

Remark 1.9 We use this theorem to prove the convergence of the proximal point method. See the proof ([3], theorem 12.26). The additional strong convexity index is a consequence of theorem 1.6, item (v).

Theorem 1.10 (The Bregman proximal descent inequality) *Let ω induce a Bregman Divergence D_ω in $\mathbb{R}^n$ and assume that it satisfies Bregman Prox Admissibility conditions for the function $\varphi : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$. Then we claim that for all $c \in \text{dom}(\omega), b \in \text{dom}(\partial\omega)$, If*

$$a = \underset{x}{\operatorname{argmin}} \left\{\varphi(x) + D_\omega(x, b)\right\},$$

we have the inequality,

$$(\varphi(c) + D_\omega(c, b)) - (\varphi(a) + D_\omega(a, b)) \geq D_\omega(c, a).$$

Remark 1.11 For more information about what function ω can induce a Bregman divergence and the admissibility conditions for Bremgna proximal mapping, consult Heinz et.al [2].

2 The proximal point method with convexity

This section reviews the convex case's Proximal point method (PPM) analysis and generalizes the theories to approximated PPM.

2.1 Convex PPM literature reviews

Rockafellar [9] pioneered the analysis of the proximal point method in the convex case. He developed the analysis in the context of maximal monotone operators in Hilbert spaces.

Applications in convex optimizations are covered. Using his theorems appropriately requires some opportunities, realizations, and characterizations of assumptions (A), (B) in his paper in the context of the applications.

In this section, we will use the result from Rockafellar that, if a monotone operator A is β strongly convex, then the resolvent operator $\mathcal{J}_A = [I + A]^{-1}$ is a $(1+\beta)^{-1}$ Lipschitz operator, making $I - \mathcal{J}_A$ is a $1 - (1+\beta)^{-1}$ a strongly monotone operator.

2.2 The proximal point method

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.

2.3 The Lyapunov function of convex PPM

We present some theorems that illustrate the use of theorem 1.8. The readers can find similar analyses and techniques in Guler's work [5].

Theorem 2.1 (PPM Lyapunov Function) *With f being $\beta \ge 0$ convex (it's strongly convex if $\beta > 0$, else it's just convex) and $x_{t+1} = \text{prox}_{\eta_{t+1} f}$ generated by PPM. Define the Lyapunov function Φ_t for all $u \in \mathbb{R}^n$:*

$$\begin{aligned}
&\Phi_t := \left(\sum_{i=1}^{t} \eta_i\right)(f(x_t) - f(u)) + \frac{1}{2}\|u - x_t\|^2 \quad \forall t \ge 1, \\
&\Phi_0 := (1/2)\|x_0 - u\|^2,
\end{aligned}$$

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 $\Phi_{t+1} - \Phi_t \le 0$. Additionally, by definition, we have

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= \left(\sum_{i=1}^{t} \eta_i\right)(f(x_{t+1}) - f(x_t)) + \frac{1}{2}\|x_{t+1} - u\|^2 - \frac{1}{2}\|x_t - u\|^2 + \eta_{t+1}(f(x_{t+1}) - f(u)) \\
&\le - \left(\sum_{i=1}^{t} \eta_i\right)(1 + \beta\eta_{t+1}/2)\|x_{t+1} - x_t\|^2 + \left(-\frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{\beta\eta_{t+1}}{2}\|u - x_{t+1}\|^2\right) \\
&\le 0,
\end{aligned}$$

And additionally, recovering the descent lemma:

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

Proof. Let $\phi_{t+1} : \mathbb{R}^n \mapsto \overline{\mathbb{R}} = \eta_{t+1} f$ be convex, consider proximal point method $x_{t+1} = \operatorname{prox}_\phi(x_t)$, apply theorem 1.8, we have $\forall u \in \mathbb{R}^n$

$$\phi_{t+1}(x_{t+1}) + \frac{1}{2}\|x_{t+1} - x_t\|^2 - \phi_{t+1}(u) - \frac{1}{2}\|u - x_t\|^2 \le -\frac{1}{2}(1 + \beta\eta_{t+1})\|u - x_{t+1}\|^2$$

let $u = x_*$

$$\begin{aligned}
&\Longrightarrow \eta_{t+1}(f(x_{t+1}) - f(x_*)) + \frac{1}{2}\|x_* - x_{t+1}\|^2 + \frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \\
&\qquad \le -\frac{\beta\eta_{t+1}}{2}\|x_* - x_{t+1}\|^2 \\
&\Longleftrightarrow \eta_{t+1}(f(x_{t+1}) - f(x_*)) + \frac{1}{2}\|x_* - x_{t+1}\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \\
&\qquad \le -\frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{\beta\eta_{t+1}}{2}\|x_* - x_{t+1}\|^2 \le 0.
\end{aligned}$$

let $u = x_t$

$$\Longrightarrow f(x_{t+1}) - f(x_t) \le -\frac{1}{\eta_{t+1}}\|x_{t+1} - x_t\|^2 - \frac{\beta}{2}\|x_t - x_{t+1}\|^2 \le 0.$$

Let's define the following quantities for all $u, \beta \ge 0$:

$$\begin{aligned}
\Upsilon_{1,t+1}(u) &= \eta_{t+1}(f(x_{t+1}) - f(u)) + \frac{1}{2}(\|x_{t+1} - u\|^2 - \|x_t - u\|^2) \\
&\le -\frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{\beta\eta_{t+1}}{2}\|u - x_{t+1}\|^2, \\
\Upsilon_{2,t+1} &= \eta_{t+1}(f(x_{t+1}) - f(x_t)) \\
&\le -\|x_{t+1} - x_t\|^2 - \frac{\beta\eta_{t+1}}{2}\|x_{t+1} - x_t\|^2 \\
&= -(1 + \beta\eta_{t+1}/2)\|x_{t+1} - x_t\|^2 \le 0.
\end{aligned}$$

With Φ_t as defined in the theorem, observe the following demonstration for all u, $\beta \ge 0$:

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= \left(\sum_{i=1}^{t+1}\eta_i\right)(f(x_{t+1}) - f(u)) + \frac{1}{2}\|x_{t+1} - u\|^2 - \left(\sum_{i=1}^{t}\eta_i\right)(f(x_t) - f(u)) - \frac{1}{2}\|x_t - u\|^2 \\
&= \left(\sum_{i=1}^{t}\eta_i\right)(f(x_{t+1}) - f(x_t)) + \frac{1}{2}\|x_{t+1} - u\|^2 - \frac{1}{2}\|x_t - u\|^2 + \eta_{t+1}(f(x_{t+1}) - f(u)) \\
&= \left(\sum_{i=1}^{t}\eta_i\right)\Upsilon_{2,t+1} + \Upsilon_{1,t+1}(u) \\
&\le -\left(\sum_{i=1}^{t}\eta_i\right)(1 + \beta\eta_{t+1}/2)\|x_{t+1} - x_t\|^2 + \left(-\frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{\beta\eta_{t+1}}{2}\|u - x_{t+1}\|^2\right) \le 0.
\end{aligned}$$

Therefore, Φ_t is a legitimate Lyapunov function for all $u, \beta \ge 0$. ∎

Remark 2.2 The above Lyapunov is not unique, and it's not optimal for $\beta > 0$, strictly strongly convex functions.

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

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

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 consideration $u = x_*$, invoking previous theorem and do

$$\begin{aligned}
\Upsilon_T \Delta_T \le \Phi_T \le \Phi_0 &= \frac{1}{2}\|x_0 - x_*\|^2 \\
\implies \Delta_T &\le \frac{1}{2\Upsilon_T}\|x_0 - x_*\|^2.
\end{aligned}$$

■

Remark 2.4 With the same choice of the sequence $(\eta_t)_{t\in\mathbb{N}}$, convergence of the PPM method of a strongly convex function is faster. The above proof is the same for $\beta = 0$, or $\beta > 0$, because it didn't use the property that $\eta_{t+1} f$ is a $\eta_{t+1}\beta$ strongly convex function.

Theorem 2.5 (PPM Strongly Convex Lyapunov Function) *With f being $\beta > 0$ strogly convex, with $x_{t+1} = \operatorname{prox}_{\eta_{t+1} f}(x_t)$, then $\Phi_t = \|x_t - x_*\|$ is a Lyapunov function satisfying:*

$$\frac{\|x_{t+1} - x_*\|}{\|x_k - x_*\|} \le (1 + \eta_{t+1}\beta)^{-1}.$$

Proof. This is a direct application that $\operatorname{prox}_{\eta_{t+1} f}$ is a contraction with constant $(1+\beta\eta_{t+1})^{-1}$. ■

Remark 2.6 It's still a mystery on how to show $f(x_t) - f(x_*)$ is a Lyapunov function. The answer is not realized by us, nor it is in Rockafellar [9] or Guler's[5] writings. Do observe that, by the choice of x_*, the contraction property of the proximal operator is strictly stronger than necessary. This inequality at the end is tighter than what we derived for gradient descent.

3 Applying the analysis of PPM

The PPM method and the Lyaounov function derived above serve as the template for other algorithms. As an appetizer, we present an analysis of gradient descent using theorems related to the convergence of PPM

In optimizations, people use a lower or an upper approximation of the objective function to approximate the PPM. The methodology includes a diverse range of approaches. For example, it includes first-order optimization, such as gradient descents, and second-order algorithms, such as Newton's method. Its scope broadens to primal-dual optimization algorithms with creativities in the Lyapunov functions or theories in monotone operators.

To demonstrate, assume that f is a lsc convex function such that it can be approximated by a lower bounding function $l_f(x|\bar{x})$ at $\bar{x}$ such that it satisfies for all x:

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

The above characterization is generic enough to include the case where $l_f(x|\bar{x})$, the under-approximating function is nonsmooth. We assume that $l_f(x|\bar{x})$ is convex for all x, at all $\bar{x}$, so the previous theorems apply.

The approximated proximal point method applies PPM to the function $l_f(x|x_t)$ for each iteration, i.e.: $x_{t+1} = \operatorname{prox}_{\eta_{t+1} l_f(\cdot|x_t)}(x_t)$.

3.1 Generic gradient descent

We will consider deriving gradient descent via the PPM approach as a warm-up. Please pay attention to the remarks. They reveal parts of the proof that could inspire the idea of a non-monotone line search method in practical settings.

Theorem 3.1 (Generic Approximated PPM) *With f convex having minimizer: x_*; $l_f(\cdot; x_t)$ convex, lsc and proper, define $\phi_t(x) = \eta_{t+1} l_f(x; x_t)$. Assume the following estimates hold:*

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

Fix any x_0, let the iterates x_t defined for $t \in \mathbb{N}$ satisfies

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

then it has

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

Additionally if $\exists \epsilon > 0 : \eta_t \in (\epsilon, 2L^{-1} - \epsilon)$, *for all* $t \in \mathbb{N}$, *the algorithm has sublinear convergence rates of*

$$\begin{aligned} f(x_T) - f(x_*) &\leq \frac{L - \epsilon^{-1}}{TL\epsilon}(f(x_0) - f(x_T)) \\ &\leq \frac{L - \epsilon^{-1}}{TL\epsilon}(f(x_0) - f(x_*)) \end{aligned}$$

Proof. By ϕ_t convex, apply theorem 1.8 with $f = \phi_t$, $x = x_t$, $x_{t+1} = p$, yielding $\forall y$

$$\begin{aligned} \phi_t(x_{t+1}) + \frac{1}{2}\|x_t - x_{t+1}\|^2 - \phi_t(y) - \frac{1}{2}\|x_t - y\|^2 &\leq -\frac{1}{2}\|y - x_{t+1}\|^2 \\ \phi_t(x_{t+1}) - \phi_t(y) + \frac{1}{2}(\|y - x_{t+1}\|^2 - \|x_t - y\|^2) &\leq -\frac{1}{2}\|x_t - x_{t+1}\|^2 \\ \left(\phi_t(x_{t+1}) + \frac{L\eta_{t+1}}{2}\|x_{t+1} - x_t\|\right) - \phi_t(y) + \frac{1}{2}(\|y - x_{t+1}\|^2 - \|x_t - y\|^2) &\leq \left(\frac{L\eta_{t+1}}{2} - \frac{1}{2}\right)\|x_t - x_{t+1}\|^2 \\ \implies \eta_{t+1}f(x_{t+1}) - \eta_{t+1}f(y) + \frac{1}{2}(\|y - x_{t+1}\|^2 - \|x_t - y\|^2) &\leq \left(\frac{L\eta_{t+1}}{2} - \frac{1}{2}\right)\|x_t - x_{t+1}\|^2. \end{aligned}$$

Setting $y = x_t$ yields

$$\begin{aligned} \eta_{t+1}(f(x_{t+1}) - f(x_t)) + \frac{1}{2}\|x_t - x_{t+1}\|^2 &\leq \left(\frac{L\eta_{t+1}}{2} - \frac{1}{2}\right)\|x_t - x_{t+1}\|^2 \\ \iff \eta_{t+1}(f(x_{t+1}) - f(x_t)) &\leq \left(\frac{L\eta_{t+1}}{2} - 1\right)\|x_t - x_{t+1}\|^2. \end{aligned}$$

In a similar manner to the derivation of the Lyapunov function for PPM, we make for all y:

$$\begin{aligned} \Upsilon_{1,t+1}(y) &= \eta_{t+1}(f(x_{t+1}) - f(y)) + \frac{1}{2}(\|x_{t+1} - y\|^2 - \|x_t - y\|^2) \\ &\leq \left(\frac{L\eta_{t+1}}{2} - \frac{1}{2}\right)\|x_t - x_{t+1}\|^2, \\ \Upsilon_{2,t+1} &= \eta_{t+1}(f(x_{t+1}) - f(x_t)) \\ &\leq \left(\frac{L\eta_{t+1}}{2} - 1\right)\|x_t - x_{t+1}\|^2. \end{aligned}$$

Now, consider defining Φ_t for all y:

$$\Phi_t = \left(\sum_{i=1}^{t} \eta_i\right)(f(x_t) - f(y)) + \frac{1}{2}\|y - x_t\|^2,$$

it is the proposed Lyapunov function for PPM; we define the base case $\Phi_0 = \frac{1}{2}\|y - x_0\|^2$. Consider the difference $\forall y$:

$$\begin{aligned}\Phi_{t+1} - \Phi_t &= \left(\sum_{i=1}^{t} \eta_i\right) \Upsilon_{2,t+1} + \Upsilon_{1,t+1}(y) \\ &\le \left(\sum_{i=1}^{t} \eta_i\right) \left(\frac{L\eta_{t+1}}{2} - 1\right) \|x_t - x_{t+1}\|^2 + \left(\frac{L\eta_{t+1}}{2} - \frac{1}{2}\right) \|x_t - x_{t+1}\|^2.\end{aligned}$$

Observe that if $\eta_i \le L^{-1}$, then $\Phi_{t+1} - \Phi_t \le 0$, hence the convergence rate of $\mathcal{O}\left((\sum_{i=1}^{t} \eta_i)^{-1}\right)$ of PPM for Φ_t is applicable.

Surprisingly, if $\eta_i \in (0, 2L^{-1})$, Φ_t still converges under mild conditions. For simplicity we set $\sigma_t := \sum_{i=1}^{t} \eta_i$. It starts with considerations that $(L\eta_{t+1}/2 - 1) < 0$, so that

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\le \left(\frac{L\eta_{t+1}}{2} - 1\right) \|x_{t+1} - x_t\|^2 \\ f(x_T) - f(x_0) &\le \underbrace{\left(\frac{L\sigma_T}{2} - T\right)}_{<0} \sum_{t=0}^{T-1} \|x_{t+1} - x_t\|^2 \\ \Longrightarrow \sum_{t=0}^{T-1} \|x_t - x_{t+1}\|^2 &\le \left(\frac{L}{2}\sigma_T - T\right)^{-1} (f(x_T) - f(x_0)) \end{aligned}$$

Continue on the RHS of $\Phi_{t+1} - \Phi_t$ so

$$\begin{aligned} \sum_{t=0}^{T-1} \Phi_{t+1} - \Phi_t &\le \left(\frac{L}{2}\sigma_T - \frac{T}{2}\right) \sum_{t=0}^{T-1} \|x_{t+1} - x_t\|^2 \\ \Phi_T - \Phi_0 &\le \left(\frac{\frac{L}{2}\sigma_T - \frac{T}{2}}{\frac{L}{2}\sigma_T - T}\right) (f(x_T) - f(x_0)) \\ &= \left(\frac{L\sigma_T - T}{L\sigma_T - 2T}\right) (f(x_T) - f(x_0)), \end{aligned}$$

implies

$$\begin{aligned} \sigma_T(f(x_T) - f(y)) + \frac{1}{2}\|y - x_t\|^2 - \frac{1}{2}\|y - x_0\|^2 &\le \left(\frac{L\sigma_T - T}{L\sigma_T - 2T}\right) (f(x_T) - f(x_0)) \\ \Longleftrightarrow f(x_T) - f(y) + \frac{1}{2\sigma_T}(\|y - x_t\|^2 - \|y - x_0\|^2) &\le \left(\frac{L - T\sigma_T^{-1}}{2T - L\sigma_T}\right) (f(x_0) - f(x_T)), \end{aligned}$$

therefore, we obtain the bound:

$$f(x_T) - f(y) \le \left(\frac{L - T\sigma_T^{-1}}{2T - L\sigma_T}\right) (f(x_0) - f(x_T)) - \frac{1}{2\sigma_T}(\|y - x_t\|^2 - \|y - x_0\|^2)$$

In the case where $\sup_{i\in\mathbb{N}}\eta_i \le 2L^{-1} - \epsilon$, and $\inf_{i\in\mathbb{N}}\eta_i \ge \epsilon$ with $\epsilon > 0$. Then we have

$$\begin{aligned}\frac{L - T\sigma_T^{-1}}{2T - L\sigma_T} &\le \frac{L - \epsilon^{-1}}{2T - LT(2L^{-1} - \epsilon)}\\ &= \frac{L - \epsilon^{-1}}{2T - T(2 - L\epsilon)}\\ &= \frac{L - \epsilon^{-1}}{TL\epsilon}.\end{aligned}$$

With $y = x_*$, we get the claimed convergence rate because $f(x_t)$ is strictly monotone decreasing. ■

Remark 3.2 Observe that inequality

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

was invoked with $x = x_{t+1}$ for the PPM descent inequality in the above proof, meaning that if $\forall(x_t)_{t\in\mathbb{N}}$ generated by the algorithm, $\exists(L_t)_{t\in\mathbb{N}}$ such that

$$\phi_t(x) \le \eta_{t+1}f(x) \le \phi_t(x) + \frac{L_t\eta_{t+1}}{2}\|x - x_t\|^2,$$

where the algorithm generates the sequence. By smartly choosing the function ϕ_{t+1} at each iteration, we can increase the stepsize while retaining a similar convergence proof. In a practical setting, when $L_t = L$, and $\phi_t(x) = \eta_{t+1}f$, this is called a line search.

The convergence rate is loose, and when f exhibits additional favourable properties, such as being strongly convex, the convergence rate can be faster. Furthermore, if the choice of y remains arbitrary, then the theorem is applicable for function without minimizers.

3.2 Examples

Example 3.3 (Convergence of the proximal gradient method) This section illustrates algorithms that satisfy the above proof's lower and upper bound estimates. Consider $f = g + h$ with h nonsmooth convex, and g being L-Lipschitz smooth convex and differentiable. Define $D_g(x, y) = g(x) - g(y) - \langle \nabla f(x), y - x\rangle$, $l_g(x; y) = g(y) + \langle \nabla g(y), y - x\rangle$, which is the Bregman divergence of the function g. Consider for all x:

$$\begin{aligned}0 \le D_g(x, y) &\le \frac{L}{2}\|x - y\|^2\\ l_g(x; y) \le g(x) &\le l_g(x; y) + \frac{L}{2}\|x - y\|^2\\ h(x) + l_g(x; y) \le f(x) = g(x) + h(x) &\le l_g(x; y) + h(x) + \frac{L}{2}\|x - y\|^2.\end{aligned}$$

Define $\phi_{t+1}(x) = \eta_{t+1}(h(x) + l_g(x; x_t))$, then results from previous theorems apply.

Remark 3.4 The envelope interpretation restricts the use of the theorem since it requires that the proximal operator be a resolvent of a gradient. Extending the usage of the PPM descent inequality to other contexts requires operator theories and creativities.

Example 3.5 (The fundamental proximal gradient lemma) The fundamental proximal gradient lemma was used heavily in the literature to derive convergence results in the convex case. The "fundamental proximal gradient lemma" originates from Beck's writings [4, theorem 10.16]. We demonstrate in this example that it's a consequence of theorem 1.10.

With $f = g + h$, h convex, lsc, g be L-Lipschitz smooth, then for all $y \in \mathbb{R}^n$, $x \in \mathbb{R}^n$, $y^+ := \text{prox}_{L^{-1}h}(y - L^{-1}\nabla(y))$ satisfies:

$$f(x) - f(y^+) \geq \frac{L}{2}\|x - y^+\|^2 - \frac{L}{2}\|x - y\|^2 + D_g(x, y).$$

A similar analysis as theorem 3.1 with theorem 1.10 obtains the same inquality. With $\phi(y) = \eta(h(y) + g(x) + \langle \nabla g(x), y - x\rangle)$ as an lower bounding function of f. Choose any x, let $x^+ = \text{prox}_\phi(x)$, then for all u:

$$\begin{aligned}
&\phi(u) + \frac{1}{2}\|u - x\|^2 - \phi(x^+) - \frac{1}{2}\|x^+ - x\|^2 \geq \frac{1}{2}\|x^+ - u\|^2 \\
\implies &\eta\underbrace{(h(u) + g(x) + \langle \nabla g(x), u - x\rangle)}_{=\phi(u)} - \eta\underbrace{f(x^+)}_{\leq \phi(x^+)} - \frac{1}{2}\|x - x^+\|^2 + \frac{1}{2}\|u - x\|^2 \geq \frac{1}{2}\|x^+ - u\|^2 \\
\iff &f(u) + (g(x) - g(u) + \langle \nabla g(x), u - x\rangle) - f(x^+) - \frac{1}{2\eta}\|x - x^+\|^2 + \frac{1}{2}\|u - x\|^2 \geq \frac{1}{2\eta}\|x^+ - u\|^2 \\
\iff &f(u) - f(x^+) - D_g(u, x) + \frac{1}{2\eta}\|u - x\|^2 - \frac{1}{2\eta}\|x^+ - x\|^2 \geq \frac{1}{2\eta}\|x^+ - u\|^2.
\end{aligned}$$

Removing the negative term $-1/2\eta\|x - x^+\|^2$ makes LHS larger, establishing the fundamental proximal gradient lemma.

Remark 3.6 Linking the PPM descent inequality to the Bregman divergence of the smooth part of the function on parameters u, x is a clever move.

4 Accelerated gradient descent and PPM

Recent works from Ahn [1] and Nesterov [8] inspired content in this section. In his works, Ahn explored the interpretation of Nesterov acceleration via PPM. They proposed the idea of "similar triangle" for unifying all varieties of Nesterov accelerated gradient. They used

PPM to derive several variations of the Nesterov accelerated gradient algorithms. Finally, they refurnished theorem 3.1 for the proof of convergence rate for the accelerated gradient. Their analysis results in relatively simple arguments that exhibits powerful extensions to several variants of the Nesterov accelerated gradient.

Interestingly, the Nesterov accelerated gradient applies to PPM too; Guler [6] did it two decades ago. He uses the idea of a Nesterov acceleration sequence faithfully. One recent development of the accelerated PPM is an algorithmic framework named: "Universal Catalyst acceleration", proposed by Lin et al [7]. It is an application of Guler's work in the context of variance-reduction stochastic gradient algorithms for machine learning.

In this section, we

(i) State Nesterov accelerated gradient and their varieties, and point to the literature discussing them.

(ii) Derive the Nesterov accelerated gradient.

(iii) Derive the popular step size choices along with the convergence rate.

Some content will differ from Ahn's works because we hope to generalize these ideas for our own use.

4.1 Varieties of Nesterov accelerated gradient

In this section, we list different varieties of the Nesterov accelerated method. We present these varieties generically because these algorithms' forms are of interest.

Definition 4.1 (AG Generic Original Form) *Let f be a L Lipschitz smooth and $\mu \geq 0$ strongly convex function. Choose x_0, $\gamma_0 > 0$, set $v_0 = x_0$, for iteration $k \geq 0$, it*

1. *computes $\alpha_k \in (0,1)$ by solving $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu$;*
2. *sets $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu$;*
3. *chooses $y_k = (\gamma_k + \alpha_k\mu)(\alpha_k\gamma_k v_k + \gamma_{k+1}x_k)$. Compute $f(y_k)$ and $\nabla f(y_k)$;*
4. *finds x_{k+1} such that $f(x_{k+1}) \leq f(y_k) - (2L)^{-1}\|\nabla f(y_k)\|^2$;*
5. *sets $v_{k+1} = \gamma_{k+1}^{-1}((1-\alpha_k)\gamma_k v_k + \alpha_k\mu y_k - \alpha_k\nabla f(y_k))$.*

Remark 4.2 This is in Nesterov's book [8, (2.2.7)]. It is the most generic algorithm in his book about accelerated gradient method. The genericity of the algorithm is provided by item 4., which is the a special case of the smooth descent lemma.

Definition 4.3 (AG Generic PPM Form) *With f convex, the generic PPM form is formulated for strictly positive stepsizes $\tilde{\eta}_i, \eta_i$ such that:*

$$\begin{aligned} x_{t+1} &= \operatorname*{argmin}_x \left\{ l_f(x; y_t) + \frac{1}{2\tilde{\eta}_{t+1}} \|x - x_t\|^2 \right\}, \\ y_{t+1} &= \operatorname*{argmin}_x \left\{ l_f(x; y_t) + \frac{L}{2}\|x - y_t\|^2 + \frac{1}{2\eta_{t+1}} \|x - x_{t+1}\|^2 \right\}. \end{aligned}$$

where $l_f(x, \bar{x})$ is a convex function satisfying

$$l_f(x, \bar{x}) \le f(x) \le f(\bar{x}) + \frac{L}{2}\|x - \bar{x}\|^2.$$

Remark 4.4 With $l_f(x, \bar{x}) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle$, then the above is equivalent to AG Generic Triangular form.

Definition 4.5 (AG Generic Triangular Form) *With f be L-Lipschitz smooth and $\mu \ge 0$ strongly convex, choose any $y_0 = x_0 = z_0$, the algorithm admits form:*

$$\begin{aligned} x_{t+1} &= x_t - \tilde{\eta}_{t+1} \nabla f(y_t) \\ z_{t+1} &= y_t - L^{-1} \nabla f(y_t) \\ y_{t+1} &= (1 + L\eta_{t+1})^{-1} (x_{t+1} + L\eta_{t+1} z_{t+1}). \end{aligned}$$

Remark 4.6 The parameter η_t, η_{t+1} are coming from the previous generic PPM. They are the same parameters for both algorithms.

Definition 4.7 (AG Generic PPM Form with Strong Convexity) *Let f be convex and differentiable with Lipschitz gradient and $\mu \ge 0$-strongly convex, define $l_f(x; y)$ to be a function such that for all $x, \bar{x} \in \mathbb{R}^n$:*

$$l_f(x; \bar{x}) + \frac{\mu}{2}\|x - \bar{x}\|^2 \le f(x; \bar{x}) \le l_f(x; \bar{x}) + \frac{L}{2}\|x - \bar{x}\|^2,$$

then for all $x_0 \in \mathbb{R}^n$, and let $y_0 = x_0$, define the following variants of PPM for function f.

$$\begin{aligned} x_{t+1} &= \operatorname*{argmin}_x \left\{ l_f(x; y_t) + \frac{\mu}{2}\|x - y_t\|^2 + \frac{1}{\tilde{\eta}_{t+1}} \|x - x_t\|^2 \right\}, \\ y_{t+1} &= \operatorname*{argmin}_x \left\{ l_f(x; y_t) + \frac{L}{2}\|x - y_t\|^2 + \frac{1}{2\eta_{t+1}} \|x - x_{t+1}\|^2 \right\}. \end{aligned}$$

Here we assume that $\eta_{t+1} > 0$ for all $t \in \mathbb{N}$.

Remark 4.8 Set $\mu = 0$, the above is the same as AG Generid Triangular Form.

Proposition 4.9 (Triangular Form via PPM Form) *with f being L-Lipschitz continuous, let $l_f(x, \bar{x}) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle$, we can show that the AG generic triangular form is a consequence of th AG generic PPM form.*

Proof. Solving the optimality on the first PPM yields:

$$\begin{aligned}\mathbf{0} &= \nabla f(y_t) + \frac{1}{\tilde{\eta}_{t+1}}(x - x_t) \\ x &= x_t - \tilde{\eta}_{t+1} \nabla f(y_t).\end{aligned}$$

Therefore, $x_{t+1} = x_t - \tilde{\eta}_{t+1} \nabla f(y_y)$. Similarly, for the updates of y_{t+1}, we have optimality condition of

$$\begin{aligned}\mathbf{0} &= \nabla f(y_t) + L(x - y_t) + \eta_{t+1}^{-1}(x - x_{t+1}) \\ \mathbf{0} &= \eta_{t+1} \nabla f(y_t) + \eta_{t+1} L(x - y_t) + x - x_{t+1} \\ \mathbf{0} &= \eta_{t+1} \nabla f(y_t) - \eta_{t+1} L y_t + (\eta_{t+1} L + 1)x - x_{t+1} \\ (1 + \eta_{t+1} L)x &= x_{t+1} - \eta_{t+1} \nabla f(y_t) + \eta_{t+1} L y_t \\ \text{define: } y_{t+1} &:= x.\end{aligned}$$

In the above expression, it hides a step of gradient descent, continuing it we have

$$\begin{aligned}(1 + \eta_{t+1} L) y_{t+1} &= x_{t+1} + \eta_{t+1} L(-L^{-1} \nabla f(y_t) + y_t) \\ \text{let: } z_{t+1} &= y_t - L^{-1} \nabla f(y_t), \text{ so,} \\ (1 + \eta_{t+1} L) y_{t+1} &= x_{t+1} + L \eta_{t+1} z_{t+1}.\end{aligned}$$

Combining it yields the tree points update format

$$\begin{aligned}x_{t+1} &= x_t - \tilde{\eta}_{t+1} \nabla f(y_t) \\ z_{t+1} &= y_t - L^{-1} \nabla f(y_t) \\ y_{t+1} &= (1 + L\eta_{t+1})^{-1}(x_{t+1} + L\eta_{t+1} z_{t+1}),\end{aligned}$$

the ordering of x_{t+1}, z_{t+1} can be permuted. The base case is when $t = 0$, so then $x_0 = y_0$ for the initial guess. ■

4.2 Generic Lyapunov analysis for Accelerated gradient via PPM

The first theorem states the generic convergence results when f is differentiable with L-Lipschitz gradient, using 4.3.

Lemma 4.10 (Smooth Generic AG Lyapunov Analysis) *With f having L-Lipschitz gradient and minimizer x_*, using definition 4.3 and theorem 1.8 then these are the upper bounds for the RHS of the PPM descent inequality:*

$$
\begin{aligned}
\Upsilon_{1,t+1}^{AG} &= \tilde{\eta}_{t+1}(f(z_{t+1}) - f(x_*)) + \frac{1}{2}(\|x_{t+1} - x_*\|^2 - \|x_t - x_*\|^2) \\
&\le -\frac{1}{2}\|x_{t+1} - x_t\|^2 + \frac{\tilde{\eta}_{t+1}L}{2}\|z_{t+1} - y_t\|^2 - \langle \tilde{\eta}_{t+1}\nabla f(y_t), x_{t+1} - z_{t+1}\rangle, \\
\Upsilon_{2,t+1}^{AG} &= f(z_{t+1}) - f(z_t) \le \langle \nabla f(y_t), z_{t+1} - z_t\rangle + \frac{L}{2}\|z_{t+1} - y_t\|^2.
\end{aligned}
$$

Proof. Define $\phi_t(x) = \tilde{\eta}_{t+1}(f(y_t) + \langle \nabla f(y_t), x - y_t\rangle)$. With L-smoothness of f in mind, consider the following sequence of inequalities:

$$
\begin{aligned}
\phi_t(x_{t+1}) &= \tilde{\eta}_{t+1}(f(y_t) + \langle \nabla f(y_t), x_{t+1} - y_t\rangle) \\
\phi_t(x_{t+1}) &= \tilde{\eta}_{t+1}(f(y_t) + \langle \nabla f(y_t), (x_{t+1} - z_{t+1}) + (z_{t+1} - y_t)\rangle) \\
&\ge \tilde{\eta}_{t+1}\left(f(z_{t+1}) - \frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), x_{t+1} - z_{t+1}\rangle\right),
\end{aligned}
$$

performing PPM on the function produces the PPM Lyapunov inequality, substituting yields equivalences for all x_*:

$$
\begin{aligned}
&\phi_t(x_{t+1}) - \phi_t(x_*) + \frac{1}{2}\|x_* - x_{t+1}\|^2 - \frac{1}{2}\|x_* - x_t\|^2 =: \Upsilon_{1,t+1}^{\mathrm{AG}} \\
&\quad \le -\frac{1}{2}\|x_{t+1} - x_t\|^2 \\
\implies &\tilde{\eta}_{t+1}\left(f(z_{t+1}) - \frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), x_{t+1} - z_{t+1}\rangle\right) - \tilde{\eta}_{t+1}f(x_*) + \frac{1}{2}\left(\|x_{t+1} - x_*\|^2 - \|x_t - x_*\|^2\right) \\
&\quad \le -\frac{1}{2}\|x_{t+1} - x_t\|^2 \\
\iff &\tilde{\eta}_{t+1}\left(f(z_{t+1}) - f(x_*)\right) + \frac{1}{2}\|x_{t+1} - x_*\|^2 - \frac{1}{2}\|x_t - x_*\|^2 \\
&\quad \le -\frac{1}{2}\|x_{t+1} - x_t\|^2 + \frac{\tilde{\eta}_{t+1}}{2}\|z_{t+1} - y_t\|^2 - \langle \tilde{\eta}_{t+1}\nabla f(y_t), x_{t+1} - z_{t+1}\rangle.
\end{aligned}
$$

Observe that, the rhs and lhs of the Lyapunov inequality are anchored at z_{t+1}. Similarly for the descent inequality we wish to obtain:

$$
\begin{aligned}
f(z_{t+1}) - f(z_t) &= f(z_{t+1}) - f(y_t) + f(y_t) - f(z_t) \\
&\le \langle \nabla f(y_t), z_{t+1} - y_t\rangle + \frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), y_t - z_t\rangle \\
&= \langle \nabla f(y_t), z_{t+1} - z_t\rangle + \frac{L}{2}\|z_{t+1} - y_t\|^2 \\
\text{let } \Upsilon_{2,t+1}^{\mathrm{AG}} &:= f(z_{t+1}) - f(z_t).
\end{aligned}
$$

Which is the descent inequality anchored on z_{t+1}. Merging the $(z_{t+1} - y_t)$ with $y_t - z_t$ together yield the desired results. ∎

Theorem 4.11 (Generic AG Convergence) *If there is a choice of $\eta_i, \tilde{\eta}_i$ in definition 4.3 such that we have*

$$\left(\sum_{i=1}^{t} \tilde{\eta}_i\right) \Upsilon_{2,t+1}^{AG} + \Upsilon_{2,t+1}^{AG} \leq 0,$$

then with

$$\Phi_t = \left(\sum_{i=1}^{t} \tilde{\eta}_{t+1}\right) (f(z_t) - f(x_*)) + \frac{1}{2}\|x_t - x_*\|^2 \quad \forall t \in \mathbb{N}$$
$$\Phi_0 = \frac{1}{2}\|x_t - x_*\|^2,$$

we have $\Phi_{t+1} - \Phi_t \leq 0$, which allows for a convergence rate of $\mathcal{O}\left(\sum_{i=1}^{T} \eta_i^{-1}\right)$ for $f(z_T) - f(x_*)$.

Proof. By definition we have

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= \left(\sum_{i=1}^{t+1} \tilde{\eta}_i\right) (f(z_{t+1}) - f(x_*)) - \left(\sum_{i=1}^{t} \tilde{\eta}_i\right) (f(z_t) - f(x_*)) + \frac{1}{2}\|x_t - x_*\|^2 - \frac{1}{2}\|x_{t+1} - x_*\|^2 \\
&= \tilde{\eta}_{t+1}(f(z_{t+1}) - f(z_*)) + \left(\sum_{i=1}^{t} \tilde{\eta}_i\right) (f(z_{t+1}) - f(z_t)) + \frac{1}{2}\|x_t - x_*\|^2 - \frac{1}{2}\|x_{t+1} - x_*\|^2 \\
&= \left(\sum_{i=1}^{t} \tilde{\eta}_i\right) \Upsilon_{2,t+1}^{\mathrm{AG}} + \Upsilon_{2,t+1}^{\mathrm{AG}} \leq 0.
\end{aligned}$$

The derivation for the convergence raet is direct and similar to theorem 2.3 Therefore, if we can identify parameter for the generic algorithm that asserts the condition above, then we have convergence for the algorithm. ■

4.3 Scenario I

In this scenario, we aim to recover the following variant of the accelerated gradient algorithm from 12.1 in Ryu's writing [10], it is:

$$\begin{aligned}
z_{t+1} &= y_k + L^{-1}\nabla f(y_t), \\
y_{t+1} &= z_{t+1} + \frac{t-1}{t+2}\left(z_{k+1} - z_k\right).
\end{aligned}$$

An equivalent form of the above is also provided by Ryu:

$$\begin{aligned}
z_{t+1} &= y_t + L^{-1}\nabla f(y_t), \\
x_{t+1} &= x_t + (t+1)(2L)^{-1}\nabla f(y_t), \\
y_{t+1} &= \left(1 - \frac{2}{t+2}\right) z_{t+1} + \left(\frac{1}{t+2}\right) x_{t+1}.
\end{aligned}$$

The base case requires cares since it's not unique, and depending on how the base case is define, there even more representations of it.

Observation 4.12 *Observe that the second equivalent form is an example of definition 4.5.*

It is stated in Ryu's writing that the base case is $x_0 = y_0 = z_0$, we think this is not the only options. We remind the reader that there is not an obvious choice of $t \in \mathbb{Z}$ such that $x_t = y_t = z_t$ where the above algorithm remains consistent for all $t \geq -1$. The choice of base case presented in Ryu's writing is a choice and not a necessity. We state that one of the base case scenario only requires knowing y_1.

We need to observe the case where $t = 0$ to understand the base case of the algorithm. At each iteration t, Given (x_t, y_t), the above formula maps to $(x_{t+1}, y_{t+1}, z_{t+1})$. When $t = 0$, it has $y_1 = x_1$. Therefore, knowing only either y_1, or x_1 can initiate the algorithm. This is an alternative scenario of the base case. If, we wish to write $y_0 = x_0$ at $t = 0$ for the base case instead, then it creates the following algorithm

$$\begin{aligned} y_t &= \left(1 - \frac{2}{t+2}\right) z_t + \left(\frac{1}{t+2}\right) x_t, \\ z_{t+1} &= y_t + L^{-1}\nabla f(y_t), \\ x_{t+1} &= x_t + (t+1)(2L)^{-1}\nabla f(y_t). \end{aligned}$$

It has a different representations due to a specific choice of the base case.

Finally, by the observation that given (z_t, y_t) for any $t \geq -1$, we can solve for x_t via $y_t = (1-2/(1+t))z_t + 1/(t+1)x_t$, obtaining (x_t, y_t, z_t), hence the algorithm can be reduced to a representations with only (z_t, y_t) and an updates using only (z_t, y_t) to (z_{t+1}, y_{t+1}). That is the momentum form presented above.

We advise the reader to read the Biliography notes by Ryu [10, chapter 12] of his book. It goes over the full context and history for this particular variant of gradient acceleration method. As the time of composing this notes, the writer is still reading on the topic of accelerated gradient.

Proposition 4.13 *With $\eta_t, \tilde{\eta}_t$ in definition 4.3 where $\eta_{t+1} = \tilde{\eta}_{t+1} = (t-1)/L$, we derive the following algorithm:*

Proof. Consider the upper bounds of $\Upsilon^{\text{AG}}_{1,t+1}, \Upsilon^{\text{AG}}_{2,t+1}$, together with the updates of the Generic Triangular Form of the algorithm:

$$\begin{aligned} x_{t+1} - x_t &= \tilde{\eta}_{t+1}\nabla f(y_t) \\ z_{t+1} - y_t &= L^{-1}\nabla f(y_t), \end{aligned}$$

we consider the upper bound of $\Upsilon^{\mathrm{AG}}_{1,t+1}$, with that we can simplify it:

$$\begin{aligned}
\Upsilon^{\mathrm{AG}}_{1,t+1} &\leq -\frac{1}{2}\|x_{t+1}-x_t\|^2 + \frac{\tilde{\eta}_{t+1}L}{2}\|z_{t+1}-y_t\|^2 - \langle \tilde{\eta}_{t+1}\nabla f(y_t), x_{t+1}-z_{t+1}\rangle \\
&\leq -\frac{1}{2}\|\tilde{\eta}_{t+1}\nabla f(y_t)\|^2 + \frac{\tilde{\eta}_{t+1}L}{2}\|L^{-1}\nabla f(y_t)\|^2 + \langle -\tilde{\eta}_{t+1}\nabla f(y_t), -L^{-1}\nabla f(y_t) + y_t - x_t + \tilde{\eta}_{t+1}\nabla f(y_t)\rangle \\
&= \left(\frac{\tilde{\eta}_{t+1}^2}{2} + \frac{\tilde{\eta}_{t+1}}{2L}\right)\|\nabla f(y_t)\|^2 + \tilde{\eta}_{t+1}\left\langle \nabla f(y_t), (\tilde{\eta}_{t+1} - L^{-1})\nabla f(y_t) + y_t - x_t\right\rangle \\
&= (1/2)(L^{-1}\tilde{\eta}_{t+1} - \tilde{\eta}_{t+1}^2)\|\nabla f(y_t)\|^2 + \tilde{\eta}_{t+1}(\tilde{\eta}_{t+1} - L^{-1})\|\nabla f(y_t)\|^2 + \tilde{\eta}_{t+1}\langle \nabla f(y_t), y_t - x_t\rangle \\
&= \frac{1}{2}\left(-L^{-1}\tilde{\eta}_{t+1} + \tilde{\eta}_{t+1}^2\right)\|\nabla f(y_t)\|^2 + \tilde{\eta}_{t+1}\langle \nabla f(y_t), y_t - x_t\rangle.
\end{aligned}$$

For $\Upsilon^{\mathrm{AG}}_{2,t+1}$, we simplify it with the same updates relation in mind:

$$\begin{aligned}
\Upsilon^{\mathrm{AG}}_{2,t+1} &\leq \langle \nabla f(y_t), z_{t+1} - y_t\rangle + \frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), y_t - z_t\rangle \\
&= \langle \nabla f(y_t), -L^{-1}\nabla f(y_t)\rangle + \frac{L}{2}\|L^{-1}\nabla f(y_t)\|^2 + \langle \nabla f(y_t), y_t - z_t\rangle \\
&= -\frac{1}{2L}\|\nabla f(y_t)\|^2 + \langle \nabla f(y_t), y_t - z_t\rangle.
\end{aligned}$$

Observe that the cross product term for $\Upsilon^{\mathrm{AG}}_{1,t+1}, \Upsilon^{\mathrm{AG}}_{2,t+1}$ doesn't match. Hence let's consider $y_t - x_t = L\eta_t(z_t - y_t)$ from the algorithm, we make the choice to do surgery on upper bound of $\Upsilon^{\mathrm{AG}}_{2,t+1}$, so $\langle \nabla f(y_t), y_t - x_t\rangle = \langle \nabla f(y_t), L\eta_t(z_t - y_t)\rangle$. With this in mind, applying theorem 4.11 and upper bound from lemma 4.10 yields inequality:

$$\begin{aligned}
&\Upsilon^{\mathrm{AG}}_{1,t+1} + \left(\sum_{i=1}^{t}\tilde{\eta}_i\right)\Upsilon^{\mathrm{AG}}_{2,t+1} \\
&\leq \frac{1}{2}\left(-L^{-1}\tilde{\eta}_{t+1} + \tilde{\eta}_{t+1}^2\right)\|\nabla f(y_t)\|^2 + \tilde{\eta}_{t+1}\langle \nabla f(y_t), y_t - x_t\rangle + \left(\sum_{i=1}^{t}\tilde{\eta}_i\right)\left(-\frac{1}{2L}\|\nabla f(y_t)\|^2 + \langle \nabla f(y_t), y_t - z_t\rangle\right) \\
&= \left(\frac{1}{2}\tilde{\eta}_{t+1}\left(\tilde{\eta}_{t+1} - L^{-1}\right) - \frac{1}{2L}\sum_{i=1}^{t}\tilde{\eta}_i\right)\|\nabla f(y_t)\|^2 + \tilde{\eta}_{t+1}\langle \nabla f(y_t), L\eta_t(z_t - y_t)\rangle + \left(\sum_{i=1}^{t}\tilde{\eta}_i\right)\langle \nabla f(y_t), y_t - z_t\rangle \\
&= \left(\frac{1}{2}\tilde{\eta}_{t+1}\left(\tilde{\eta}_{t+1} - L^{-1}\right) - \frac{1}{2L}\sum_{i=1}^{t}\tilde{\eta}_i\right)\|\nabla f(y_t)\|^2 + \left(L\eta_t\tilde{\eta}_{t+1} - \sum_{i=1}^{t}\tilde{\eta}_i\right)\langle \nabla f(y_t), z_t - y_t\rangle.
\end{aligned}$$

Next, we select a choice for stepsize parameter $\eta_t, \tilde{\eta}_{t+1}$ such that the coefficient for the $\|\nabla f(y_t)\|^2$ is negative, and is zero for the cross product term. In this scenario, we makes the choice of $\tilde{\eta}_t = \eta_t$. Continuting will simplify the upper bound so that it is:

$$\frac{1}{2}\left(\eta_{t+1}^2 - L^{-1}\eta_{t+1} - L^{-1}\sum_{i=1}^{t}\eta_i\right)\|\nabla f(y_t)\|^2 + \left(L\eta_t\eta_{t+1} - \sum_{i=1}^{t}\eta_i\right)\langle \nabla f(y_t), z_t - y_t\rangle \leq 0.$$

Assuming the above inequality holds then one of the sufficient conditions would happen when the coefficient of the cross product term is zero and the coefficient for the normed term is negative, yielding the condition for all $t \in \mathbb{N}$:

$$\begin{cases} L\eta_{t+1}^2 + \eta_{t+1} - \sum_{i=1}^t \eta_i \leq 0, \\ L\eta_t\eta_{t+1} - \sum_{i=1}^t \eta_i = 0. \end{cases}$$

Substituting the sequence equality back to the first one yield:

$$\begin{aligned} L\eta_{t+1}^2 - (\eta_{t+1} + L\eta_t\eta_{t+1}) &\leq 0 \\ L\eta_{t+1}^2 - \eta_{t+1} &\leq L\eta_t\eta_{t+1} \\ \eta_{t+1}(L\eta_{t+1} - 1) &\leq L\eta_t\eta_{t+1} \\ \eta_t > 0 \implies L\eta_{t+1} - 1 &\leq L\eta_t \\ \eta_{t+1} &\leq \eta_t + L^{-1}. \end{aligned}$$

To satisfy the equality, reader should verify that $\eta_{t+1} = (t-1)/L$ is one of the options. And there are not many other options for the choice of the stepszies for the equality to be satisfied. ■

4.4 Scenario II

With the choice of $\tilde{\eta}_{t+1} = \eta_t + L^{-1}$, we can recover the seminal Nesterov accelerated gradient algorithm from 1983.

4.5 Scenario III

5 Classical analysis of Nesterov accelerated gradient

In this section, we reproduce some of the analysis for Nesterov accelerated gradient method with excruciating details. We will also go over some details about the history, development and motivations behind the accelerated gradient algorithm.

6 Modern techniques in analysis of accelerated gradient algorithms

References

[1] Kwangjun Ahn and Suvrit Sra. *Understanding nesterov's acceleration via proximal point method*. June 2, 2022. DOI: 10.48550/arXiv.2005.08304. arXiv: 2005.08304[cs,math]. URL: http://arxiv.org/abs/2005.08304 (visited on 11/04/2023).

[2] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. "A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications". In: *Mathematics of Operations Research* 42.2 (May 2017), pp. 330–348. ISSN: 0364-765X, 1526-5471. DOI: 10.1287/moor.2016.0817. URL: https://pubsonline.informs.org/doi/10.1287/moor.2016.0817 (visited on 12/11/2023).

[3] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Cham: Springer International Publishing, 2017. ISBN: 978-3-319-48310-8 978-3-319-48311-5. DOI: 10.1007/978-3-319-48311-5. URL: https://link.springer.com/10.1007/978-3-319-48311-5 (visited on 11/29/2023).

[4] Amir Beck. *First-Order Methods in Optimization — SIAM Publications Library*. MOS-SIAM Series in Optimization. SIAM. ISBN: 978-1-61197-498-0. URL: https://epubs.siam.org/doi/book/10.1137/1.9781611974997 (visited on 10/19/2023).

[5] Osman Guler. "On the Convergence of the Proximal Point Algorithm for Convex Minimization". In: *SIAM Journal on Control and Optimization* 29.2 (Mar. 1991). Num Pages: 17 Place: Philadelphia, United States Publisher: Society for Industrial and Applied Mathematics, p. 17. ISSN: 03630129. DOI: 10.1137/0329022. URL: https://www.proquest.com/docview/925962166/abstract/A60B4BA7798A45D1PQ/1 (visited on 05/18/2024).

[6] Osman Güler. "New Proximal Point Algorithms for Convex Minimization". In: *SIAM Journal on Optimization* 2.4 (Nov. 1992). Publisher: Society for Industrial and Applied Mathematics, pp. 649–664. ISSN: 1052-6234. DOI: 10.1137/0802032. URL: https://epubs.siam.org/doi/10.1137/0802032 (visited on 11/30/2023).

[7] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. "A Universal Catalyst for First-Order Optimization". In: NIPS - Advances in Neural Information Processing Systems. MIT Press, Dec. 7, 2015, p. 3384. URL: https://inria.hal.science/hal-01160728 (visited on 05/31/2024).

[8] Yurii Nesterov. *Lectures on Convex Optimization*. Vol. 137. Springer Optimization and Its Applications. Cham: Springer International Publishing, 2018. ISBN: 978-3-319-91577-7 978-3-319-91578-4. DOI: 10.1007/978-3-319-91578-4. URL: http://link.springer.com/10.1007/978-3-319-91578-4 (visited on 10/11/2023).

[9] R. Tyrrell Rockafellar. "Monotone Operators and the Proximal Point Algorithm". In: *SIAM Journal on Control and Optimization* 14.5 (Aug. 1976), pp. 877–898. ISSN: 0363-0129, 1095-7138. DOI: 10.1137/0314056. URL: http://epubs.siam.org/doi/10.1137/0314056 (visited on 11/06/2023).

[10] Ernest K. Ryu and Wotao Yin. *Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators*. Cambridge: Cambridge University Press, 2022. ISBN: 978-1-00-916085-8. DOI: 10.1017/9781009160865. URL: https://large-scale-book.mathopt.com/ (visited on 01/22/2024).

Postponed Proofs