# Nesterov Type Momentum Methods

Alto *

June 6, 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 $x, y$*

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

*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) $x^+ \in \operatorname{argmin}_x f(x) \implies \frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - f(x^+) \leq (L/2)\|x - x^+\|^2$, *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 $\alpha$ if and only if $f - (\alpha/2)\|\cdot\|^2$ is a convex function.*

**Theorem 1.5 (Stong convexity equivalences)** *With $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ $\alpha$-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 $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ $\alpha$-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 $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}^n$ $\beta$-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:* $\text{prox}_f(x) = \underset{u}{\text{argmin}} \left\{f(u) + \frac{1}{2}\|u - x\|^2\right\}$.

**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 $\omega$ induce a Bregman Divergence $D_\omega$ 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 $c \in dom(\omega), b \in dom(\partial\omega)$, If*

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

*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 $\omega$ 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 $\beta$ 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 $\beta \geq 0$ 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 $\Phi_t$ for all $u \in \mathbb{R}^n$:*

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

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

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

*And additionally, recovering the descent lemma:*

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

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

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

let $u = x_*$

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

let $u = x_t$

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

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

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

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

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

Therefore, $\Phi_t$ is a legitimate Lyapunov function for all $u, \beta \geq 0$. ■

**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:*

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

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

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

■

**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:*

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

*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. Do observe that, by the choice of $x_*$, the contraction property of the proximal operator is strictly stronger than necessary.

# 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 $x$:

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

The above characterization is generic enough to include the case where $l_f(x|\bar{x})$, 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.: $x_{t+1} = \text{prox}_{\eta_{t+1} l_f(\cdot|x_t)}(x_t)$.

## 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:*

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

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

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

*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 $\phi_t$ 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 $\Phi_t$ 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 $\forall y$:

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

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

Surprisingly, if $\eta_i \in (0, 2L^{-1})$, $\Phi_t$ 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

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

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

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

implies

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

therefore, we obtain the bound:

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

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

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

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

**Remark 3.2** Observe that inequality

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

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

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

where the algorithm generates the sequence. By smartly choosing the function $\phi_{t+1}$ 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.

## 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 $x$:

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

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 $f$. Let $x^+ = \text{prox}_\phi(x)$, then for all $u$:

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

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

# 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 of 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; 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 (Accelearted Gradient 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 (Accelerated Gradient Generic Triangular Form)**

**Definition 4.4 (Accelerated Gradient Generic PPM Form)**

### 4.2 Nesterov accelerated gradient via PPM

### 4.3 Convergence rate of Nesterov accelerated gradient via PPM

## 5 Classical analysis of Nesterov accelerated gradient

In this section, we reproduce some of the analysis for Nesterov accelerated gradient method with excruciating details.

## 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).

# Postponed Proofs