Nesterov Type Momentum Methods

Alto *

June 2, 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*

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

Theorem 1.2 (Lipschitz Smoothness Equivalence) *With f convex and L-Lipschitz smooth, 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) $L^{-1}\|\nabla f(y) - \nabla f(x)\|^2 \le \langle \nabla f(y) - \nabla f(x), y - x\rangle \le L\|y - x\|^2$.

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

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

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

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

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

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

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

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

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

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

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

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 $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}^n$ β-convex where $\beta \geq 0$, fix any $x \in \mathbb{R}^n$, let $p = \operatorname{prox}_f(x)$, 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 $\operatorname{prox}_\alpha f(x) = \underset{u}{\operatorname{argmin}} \left\{f(u) + \frac{1}{2}\|u - x\|^2\right\}$.

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 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.

2.1 Literature reviews

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 $x_{k+1} = \operatorname{prox}_{\eta_{k+1}} f(x_k)$ for all $k \in \mathbb{N}$ where the sequence $(\eta_k)_{k\in\mathbb{N}}$ is a nonegative sequence of real numbers.

2.2 The Lyapunov function of Convex PPM

Theorem 2.1 *With f being $\beta \geq 0$ strongly convex and $x_{t+1} = \operatorname{prox}_{\eta_{t+1} f}$ generated by PPM. Define the Lyapunov function Φ_t for all $u \in \mathbb{R}^n$:*

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

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 $\Phi_{t+1}-\Phi_t\le 0$. Additionally, by the 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 additionarlly, 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 $x_{t+1}=\text{prox}_\phi(x_t)$, apply theorem 1.8, we have $\forall u\in\mathbb{R}^n$

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

let $u=x_*$

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

let $u=x_t$

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

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, $\beta \geq 0$:

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

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

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

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

With the same choice of the sequence $(\eta_t)_{t\in\mathbb{N}}$, convergence of PPM method of a strongly convex function is faster.

3 Applying the analysis of PPM

The PPM method and the Lyaounov function derived above serves as the tamplate for other algorithms. In optimizations, people use a lower, or an upper approximation of the objective function to approximate the PPM. The approaches are a diverse, 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:

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

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 that the previous theorems are applicable.

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

3.1 Generic gradient descent

As a warm up, we consider deriving gradient descent via the PPM approach. Please pay attention to the remarks, it reveals parts of the proof that could inspirate the idea of non-monotone line search method in a practical settings.

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

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

then it has

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

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 $\alpha = 1$ and $f = \phi_t$, $x = x_t$ making $x_{t+1} = p$, 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 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:

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

which is previously proposed Lyapunov function for PPM, we define the basecase $\Phi_0 = \frac{1}{2}\|y - x_0\|^2$. 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 Φ_t is applicable.

Surprisingly, if $\eta_i \in (0, 2L^{-1})$, Φ_t still convergeces. 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:

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

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 sequence is generated by the algorithm. This can be achieved by choosing the function $\phi_t t + 1$ at each iteration smartly, then it's possible to still have the same convergenace rate. In a practical setting, when $L_t = L$, and $\phi_{t+1}(x) = \eta_{t+1} f$, this is called a line search.

The convergence rate is loose and when function f exhibits additional favorable properties, such as being strongly convex, the convergence rate can be faster.

3.2 Examples

Example 3.3 (Proximal Gradient) In this section, we illustrate algorithms that satisfies the lower and uppwer bound estimate used in the above proof. Consider $f = g + h$ with h nonsmooth convex, and g being L-Lipschitz smooth convex and differentiable. Define

$D_g(x, y) = g(x) - g(y) - \langle \nabla f(x), y - x\rangle$, 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 $\phi_t(x) = \eta_{t+1}(h(x) + l_g(x; x_t))$, then results from previous theorems apply.

Remark 3.4 The envelope interpretation restricts the use of the theorem, since it requires that the proixmal operator is applied to the gradient of a function. Extending the usage of the PPM descent inequality to other context requires operator theories and creativities.

4 Accelerated gradient descent

By recent works from (CITATION NEEDED), Nesterov accelerated gradient (CITATION NEEDED) can be interpreted as an approximation to the PPM method. Additionally, works on accelerating PPM had bey done by Guler (CITATION NEEDED) decades ago. The recent interpretations of Nesterov acceleration method via PPM focuses on the idea of a similar triangle, and unifying all varieties of Nesterov acceleration.

4.1 The varieties of Nesterov accelerated gradient

In this section, we list different varities of Nesterov accelerated method. Each of the varieties will be presented generically because we are only interested in the forms they take.

4.2 Interpreting Nesterov accelerated gradient via PPM

Postponed Proofs