Understanding Nesterov's Acceleration via Proximal Point Method

Kwangjun Ahn* Suvrit Sra†

Abstract

The proximal point method (PPM) is a fundamental method in optimization that is often used as a building block for designing optimization algorithms. In this work, we use the PPM method to provide conceptually simple derivations along with convergence analyses of different versions of Nesterov's accelerated gradient method (AGM). The key observation is that AGM is a simple approximation of PPM, which results in an elementary derivation of the update equations and stepsizes of AGM. This view also leads to a transparent and conceptually simple analysis of AGM's convergence by using the analysis of PPM. The derivations also naturally extend to the strongly convex case. Ultimately, the results presented in this paper are of both didactic and conceptual value; they unify and explain existing variants of AGM while motivating other accelerated methods for practically relevant settings.

1 Introduction

In 1983, Nesterov introduced the *accelerated gradient method* (AGM) for minimizing a convex function $f : \mathbb{R}^d \to \mathbb{R}$ [Nes83]. The remarkable property of AGM is that AGM achieves a strictly faster convergence rate than the standard gradient descent (GD). Assuming that f has Lipschitz continuous gradients, T iterations of AGM are guaranteed to output a point x_T with the suboptimality gap $f(x_T) - \min_x f(x) \le O(1/T^2)$, whereas GD only ensures a suboptimality gap of $O(1/T)$. On top of being a landmark result of convex optimization, AGM is easy to implement and has found value in a myriad of applications such as sparse linear regression [BT09], compressed sensing [BBC11], the maximum flow problem [LRS13], and deep neural networks [SMDH13].

AGM's importance to both theory and practice has led to a flurry of works that seek to understand its scope and the principles that underlie it [SBC16, KBB15, WWJ16, LRP16, WRJ16, AZO17, DO19]; see §7 for details. However, one curious aspect of AGM that is not yet well-understood is the fact that it appears in various different forms. Below, we list the four most representative ones:

$$\begin{aligned} z_{t+1} &= y_t - \alpha_t^{(1)} \nabla f(y_t)\,, \\ y_{t+1} &= z_{t+1} + \beta_t^{(1)}(z_{t+1} - z_t)\,. \end{aligned}$$

Form I [Nes83, BT09].

$$\begin{aligned} y_t &= \alpha_t^{(2)} x_t + (1 - \alpha_t^{(2)}) z_t\,, \\ z_{t+1} &= y_t - \beta_t^{(2)} \nabla f(y_t)\,, \\ x_{t+1} &= x_t - \gamma_t^{(2)} \nabla f(y_t)\,. \end{aligned}$$

Form II [Nes18, AZO17].

$$\begin{aligned} y_t &= \alpha_t^{(3)} x_t + (1 - \alpha_t^{(3)}) z_t\,, \\ x_{t+1} &= x_t - \beta_t^{(3)} \nabla f(y_t)\,, \\ z_{t+1} &= \gamma_t^{(3)} x_{t+1} + (1 - \gamma_t^{(3)}) z_t\,. \end{aligned}$$

Form III [AT06, Tse08, GN18].

$$\begin{aligned} y_t &= \alpha_t^{(4)} x_t + (1 - \alpha_t^{(4)}) z_t\,, \\ x_{t+1} &= \beta_t^{(4)} x_t + (1 - \beta_t^{(4)}) y_t - \gamma_t^{(4)} \nabla f(y_t)\,, \\ z_{t+1} &= y_t - \delta_t^{(4)} \nabla f(y_t)\,. \end{aligned}$$

Form IV [Nes18].

The parameters $\alpha_t^{(\cdot)}, \beta_t^{(\cdot)}, \gamma_t^{(\cdot)}, \delta_t^{(\cdot)}$ are stepsizes that are carefully chosen to ensure an accelerated rate. An immediate question that one may ask is: *can we understand these variants of AGM in a unified manner?*

*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, kjahn@mit.edu. This work was done as the author's class project for 6.881 Optimization for Machine Learning at MIT, Spring 2020.

†Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, suvrit@mit.edu.

This paper answers this question by developing a transparent and unified analysis that captures all these variants of AGM by connecting them to the proximal point method (PPM). PPM is a well-known optimization algorithm that is often used as a conceptual building block for designing other optimization algorithms (see §2 for more background). The key insight (presented in §4) is that one can obtain AGM simply by viewing it as an approximation of PPM. This insight is inspired by the approach of Defazio [Def19], but now with more general acceleration settings and importantly, without any recourse to duality.

Contributions. In summary, we make the following contributions:

- We present an intuitive derivation of AGM by viewing it as an approximation of the proximal point method (PPM), a foundational, classical method in optimization.
- We present a unified method for deriving different versions of AGM, which may be of wider pedagogical interest. In particular, our approach readily extends to the strongly convex case and offers a short derivation of the most general version of AGM introduced by Nesterov in his textbook [Nes18, (2.2.7)].

We believe that the simple derivations presented in this paper are not only of pedagogical value but are also helpful for research because they clarify, unify, and deepen our understanding of the phenomenon of acceleration. The PPM view offers a transparent analysis of AGM based on the convergence analysis of PPM [Gül91]. Moreover, as we present in §5, the PPM view also motivates the key idea of the *method of similar triangles*, a version of AGM shown to have important extensions to practically relevant settings [Tse08, GN18]. Our approach also readily extends to the strongly convex case (§6). Finally, since PPM has been studied in settings much wider than convex optimization (see e.g., [Bac14]), we believe the connections exposed herein will help in advancing the development of accelerated methods in those settings.

Before presenting our derivations, let us first recall a brief background on the proximal point method.

2 Brief background on the proximal point method

The *proximal point method (PPM)* [Mor65, Mar70, Roc76] is a fundamental method in optimization which solves the minimization of the cost function $f : \mathbb{R}^d \to \mathbb{R}$ by iteratively solving the subproblem

$$x_{t+1} \leftarrow \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x) + \frac{1}{2\eta_{t+1}} \|x - x_t\|^2 \right\} \tag{2.1}$$

for a stepsize $\eta_{t+1} > 0$, where the norm is chosen as the ℓ_2 norm. Despite its simplicity, solving (2.1) is in general as difficult as solving the original optimization problem, and PPM is largely regarded as a "conceptual" guiding principle for accelerating optimization algorithms [Dru17].

The baseline of our discussion is the following convergence rate of PPM for convex costs proved in a seminal paper by Güler [Gül91] (here x_* denotes a global optimum point, i.e., $x_* \in \operatorname{argmin}_x f(x)$):

$$\boxed{f(x_T) - f(x_*) \le O\left(\left(\textstyle\sum_{t=1}^T \eta_t \right)^{-1} \right) \quad \text{for any } T \ge 1.} \tag{2.2}$$

In words, one can achieve an arbitrarily fast convergence rate by choosing stepsizes η_t's large. Below, we review a short Lyapunov function proof of (2.2), which will serve as a backbone to other analyses.

Proof. [**Proof of** (2.2)] It turns out that the following Lyapunov function is suitable:

$$\boxed{\Phi_t := \left(\textstyle\sum_{i=1}^t \eta_i \right) \cdot \big(f(x_t) - f(x_*) \big) + \frac{1}{2} \|x_* - x_t\|^2 ,} \tag{2.3}$$

where $\Phi_0 := \frac{1}{2} \|x_* - x_0\|^2$ and here and below, $\|\cdot\|$ is the ℓ_2 norm unless stated otherwise. Now, it suffices to show that Φ_t is decreasing, i.e., $\Phi_{t+1} \le \Phi_t$ for all $t \ge 0$. Indeed, if Φ_t is decreasing, we have $\Phi_T \le \Phi_0$ for any $T \ge 1$, which precisely recovers (2.2). To that end, we use a standard result:

Proposition 2.1. (Proximal inequality (see e.g. [BC11, Proposition 12.26])) *For a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$, let x_{t+1} be the unique minimizer of the following proximal step: $x_{t+1} \leftarrow \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ \phi(x) + \frac{1}{2} \|x - x_t\|^2 \right\}$. Then, for any $u \in \mathbb{R}^d$,*

$$\phi(x_{t+1}) - \phi(u) + \frac{1}{2} \|u - x_{t+1}\|^2 + \frac{1}{2} \|x_{t+1} - x_t\|^2 - \frac{1}{2} \|u - x_t\|^2 \le 0 .$$

Now Proposition 2.1 completes the proof as follows: First, we apply Proposition 2.1 with $\phi = \eta_{t+1} f$ and $u = x_*$ and drop the term $\frac{1}{2}\|x_{t+1} - x_t\|^2$ to obtain:

$$(\mathsf{Ineq}_1) \qquad \eta_{t+1}\left[f(x_{t+1}) - f(x_*)\right] + \frac{1}{2}\|x_* - x_{t+1}\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \le 0\,.$$

Next, from the optimality of x_{t+1}, it readily follows that

$$(\mathsf{Ineq}_2) \qquad f(x_{t+1}) - f(x_t) \le 0\,.$$

Now, computing $(\mathsf{Ineq}_1) + (\sum_{i=1}^t \eta_i) \times (\mathsf{Ineq}_2)$ yields $\Phi_{t+1} \le \Phi_t$, which finishes the proof. ∎

2.1 Our conceptual question Although the convergence rate (2.2) seems powerful, it does not have any practical values as PPM is in general not implementable. Nevertheless, one can ask the following conceptual question:

"Can we efficiently approximate PPM for a large stepsize η_t?"

Perhaps, the most straightforward approximation would be to replace the cost function f in (2.1) with its lower-order approximations. We implement this idea in the next section.

3 Two simple approximations of the proximal point method

To analyze approximation errors, let us assume that the cost function f is L-smooth.

DEFINITION 3.1. (SMOOTHNESS) *For $L > 0$, we say a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if $f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|^2$ for any $x, y \in \mathbb{R}^d$.*

From the convexity and the L-smoothness of f, we have the following lower and upper bounds: for any $x, y \in \mathbb{R}^d$,

$$\underbrace{f(y) + \langle \nabla f(y), x - y\rangle}_{=:\,\mathsf{LOWER}(x;y)} \le f(x) \le \underbrace{f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|^2}_{=:\,\mathsf{UPPER}(x;y)}\,.$$

In this section, we use these bounds to approximate PPM.

3.1 First approach: using first-order approximation Let us first replace f in the objective (2.1) with its lower approximation:

$$x_{t+1} \leftarrow \operatorname*{argmin}_x \left\{ \mathsf{LOWER}(x; x_t) + \frac{1}{2\eta_{t+1}}\|x - x_t\|^2 \right\}. \tag{3.4}$$

Writing the optimality condition, one quickly notices that (3.4) actually leads to gradient descent:

$$x_{t+1} = x_t - \eta_{t+1}\nabla f(x_t)\,. \tag{3.5}$$

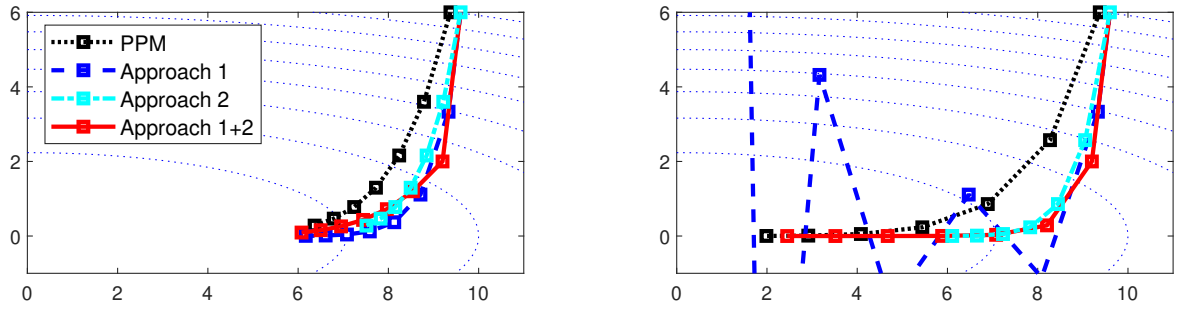
Let us see how well (3.4) approximates PPM:

Proof. [**Analysis of the first approach**] We first establish counterparts of (Ineq_1) and (Ineq_2). First, we apply Proposition 2.1 with $\phi(x) = \eta_{t+1}\mathsf{LOWER}(x; x_t)$ and $u = x_*$:

$$\phi(x_{t+1}) - \phi(x_*) + \frac{1}{2}\|x_* - x_{t+1}\|^2 + \frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \le 0\,.$$

Now using convexity and L-smoothness, we have

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



(a) $\eta_t \equiv 1/3$. (b) $\eta_t = t/3$.

Figure 1: Iterates comparison between PPM (2.1), the first approach (3.4), the second approach (3.6), and the combined approach (4.8). For the setting, we choose $f(x, y) = 0.1x^2 + y^2$ and $x_0 = (10, 10)$.

and hence the above inequality implies the following analogue of (Ineq_1):

$$(\mathsf{Ineq}_1^{\mathsf{GD}}) \qquad \eta_{t+1}\left[f(x_{t+1}) - f(x_*)\right] + \frac{1}{2}\left\|x_* - x_{t+1}\right\|^2 - \frac{1}{2}\left\|x_* - x_t\right\|^2 \le (\mathcal{E}_1^{\mathsf{GD}}),$$

where $(\mathcal{E}_1^{\mathsf{GD}}) := (\frac{L\eta_{t+1}}{2} - \frac{1}{2})\left\|x_{t+1} - x_t\right\|^2$. Next, we use the L-smoothness of f and the fact $\nabla f(x_t) = -1/\eta_{t+1}(x_{t+1} - x_t)$ (due to (3.5)), to obtain the following analogue of (Ineq_2):

$$(\mathsf{Ineq}_2^{\mathsf{GD}}) \qquad f(x_{t+1}) - f(x_t) \le \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\left\|x_{t+1} - x_t\right\|^2 = (\mathcal{E}_2^{\mathsf{GD}}),$$

where $(\mathcal{E}_2^{\mathsf{GD}}) := (\frac{L}{2} - \frac{1}{\eta_{t+1}})\left\|x_{t+1} - x_t\right\|^2$.

Now paralleling the proof of (2.2), to show that Φ_t (2.3) is a valid Lyapunov function, we need to find the stepsizes η_t's that satisfy the following relation: $(\mathcal{E}_1^{\mathsf{GD}}) + (\sum_{i=1}^t \eta_i) \times (\mathcal{E}_2^{\mathsf{GD}}) \le 0$. On the other hand, note that both $(\mathcal{E}_1^{\mathsf{GD}})$ and $(\mathcal{E}_2^{\mathsf{GD}})$ become positive numbers when $\eta_{t+1} > 2/L$. Hence, the admissible choices for η_t at each iteration are upper bounded by $2/L$, which together with the PPM convergence rate (2.2) implies that $O(1/\sum_{t=1}^T \eta_t) = O(1/T)$ is the best convergence rate one can prove. Indeed, choosing $\eta_t \equiv 1/L$, then we have $(\mathcal{E}_1^{\mathsf{GD}}) = 0$ and $(\mathcal{E}_2^{\mathsf{GD}}) < 0$, obtaining the well-known bound of $f(x_T) - f(x_*) \le \frac{L\|x_0 - x_*\|^2}{2T} = O(1/T)$. ∎

To summarize, the first approach only leads to a disappointing result: the approximation is valid only for the small stepsize regime of $\eta_t = O(1/L)$. We empirically verify this fact for a quadratic cost in Figure 1. As one can see from Figure 1, the lower approximation approach (3.4) overshoots for large stepsizes like $\eta_t = \Theta(t)$ and quickly steers away from the PPM iterates.

3.2 Second approach: using smoothness After seeing the disappointing outcome of the first approach, our second approach is to replace f with its upper approximation due to the L-smoothness:

$$x_{t+1} \leftarrow \operatorname*{argmin}_x \left\{ \mathsf{UPPER}(x; x_t) + \frac{1}{2\eta_{t+1}}\left\|x - x_t\right\|^2 \right\}. \tag{3.6}$$

Writing the optimality condition, (3.6) actually leads to a conservative update of gradient descent:

$$x_{t+1} = x_t - \frac{1}{L + \eta_{t+1}^{-1}}\nabla f(x_t)\,. \tag{3.7}$$

Note that regardless of how large η_{t+1} we choose, the actual update stepsize in (3.7) is always upper bounded by $1/L$. Although this conservative update prevents the overshooting phenomenon of the first approach, as we increase η_t, this conservative update becomes too tardy to be a good approximation of PPM; see Figure 1.

4 Nesterov's acceleration via alternating two approaches

In the previous section, we have seen that the two simple approximations of PPM both have limitations. Nonetheless, observe that their limitations are opposite to each other: while the first approach is too "reckless," the second approach is too "conservative." This observation motivates us to consider a *combination* of the two approaches which could mitigate each other's limitation.

REMARK 4.1. *A similar interpretation of Nesterov's acceleration as a combination of a reckless step and a conservative step also appeared in [AZO17, BG19]*

Let us implement this idea by alternating between the two approximations (3.4) and (3.6) of PPM. The key modification is that for both approximations, we introduce an additional sequence of points $\{y_t\}$ for cost function approximation; i.e., we use the following approximations for the t-th iteration:

$$f(y_t) + \langle \nabla f(y_t), x - y_t \rangle \le f(x) \le f(y_t) + \langle \nabla f(y_t), x - y_t \rangle + \frac{L}{2} \|x - y_t\|^2 .$$

Indeed, this modification is crucial. If we just use approximations at x_t, the resulting alternation merely concatenates (3.4) and (3.6) during each iteration, and the two limitations we discussed in §3 will remain in the combined approach. In particular, every other step that corresponds to the lower approximation would be still suffer from overshooting for large stepsizes.

Having introduced a separate sequence $\{y_t\}$ for cost approximations, we consider the following alternation where during each iteration, we update x_t with (3.4) and y_t with (3.6):

Approximate PPM with alternating two approaches. Given $x_0 \in \mathbb{R}^d$, let $y_0 = x_0$ and run:

$$x_{t+1} \leftarrow \operatorname{argmin}_x \left\{ \mathsf{LOWER}(x; y_t) + \tfrac{1}{2\eta_{t+1}} \|x - x_t\|^2 \right\}, \tag{4.8a}$$

$$y_{t+1} \leftarrow \operatorname{argmin}_x \left\{ \mathsf{UPPER}(x; y_t) + \tfrac{1}{2\eta_{t+1}} \|x - x_{t+1}\|^2 \right\}. \tag{4.8b}$$

In Figure 1, we empirically verify that (4.8) indeed gets the best of both worlds: this combined approach successfully approximates PPM even for the regime $\eta_t = \Theta(t)$. More remarkably, (4.8) is exactly equal to one version of AGM ("Form II" in the introduction). Turning (4.8) into the equational form by writing the optimality conditions, and introducing an auxiliary iterate $z_{t+1} := y_t - 1/L \nabla f(y_t)$ (only for simplicity), we obtain the following ($x_0 = y_0 = z_0$):

Equivalent representation of (4.8)**:**

$$y_t = \tfrac{1/L}{1/L+\eta_t} x_t + \tfrac{\eta_t}{1/L+\eta_t} z_t \,, \tag{4.9a}$$

$$x_{t+1} = x_t - \eta_{t+1} \nabla f(y_t) \,, \tag{4.9b}$$

$$z_{t+1} = y_t - \tfrac{1}{L} \cdot \nabla f(y_t) \,. \tag{4.9c}$$

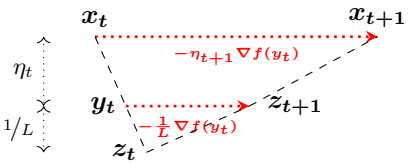


Figure 2: Illustration of (4.9).

Hence, we arrive at AGM without relying on any non-trivial derivations in the literature such as estimate sequence [Nes18] or linear coupling [AZO17]. To summarize, we have demonstrated:

Nesterov's AGM is an approximate instantiation of the proximal point method!

4.1 Understanding mysterious parameters of AGM It is often the case in the literature that the interpolation step (4.9a) is written as an abstract form $y_t = \tau_t x_t + (1 - \tau_t) z_t$ with a weight parameter $\tau_t > 0$ to be chosen [AZO17, LRP16, WRJ16, BG19]. That said, in the previous works, τ_t is carefully chosen according to the analysis without conveying much intuition. One important aspect of our PPM view is that it reveals a close relation between the weight parameter τ_t and the stepsize η_t. More specifically, τ_t is chosen so that the ratio of the distances $\|y_t - x_t\| : \|y_t - z_t\|$ is equal to $\eta_t : 1/L$ (see Figure 2).

4.2 Analysis based on PPM perspective In order to determine η_t's in (4.9), we revisit the analysis of PPM from §3. In turns out that following §3.1, one can derive the following analogues of (Ineq_1) and (Ineq_2) using Proposition 2.1 (we defer the derivations to §A.1):

$$(\mathsf{Ineq}_1^{\mathsf{AGM}}) \qquad \eta_{t+1}(f(z_{t+1}) - f(x_*)) + \frac{1}{2}\|x_* - x_{t+1}\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \le (\mathcal{E}_1^{\mathsf{AGM}}),$$

$$(\mathsf{Ineq}_2^{\mathsf{AGM}}) \qquad f(z_{t+1}) - f(z_t) \le (\mathcal{E}_2^{\mathsf{AGM}}),$$

where $(\mathcal{E}_1^{\mathsf{AGM}}) := (\frac{\eta_{t+1}^2}{2} - \frac{\eta_{t+1}}{2L})\|\nabla f(y_t)\|^2 + L\eta_t\eta_{t+1}\langle \nabla f(y_t), z_t - y_t\rangle$ and $(\mathcal{E}_2^{\mathsf{AGM}}) := -\frac{1}{2L}\|\nabla f(y_t)\|^2 - \langle \nabla f(y_t), z_t - y_t\rangle$. Given the above inequalities, consider the following modified Lyapunov function (2.3) which replaces the first x_t with z_t:

$$\Phi_t := (\textstyle\sum_{i=1}^t \eta_i)\cdot(f(z_t) - f(x_*)) + \frac{1}{2}\|x_* - x_t\|^2 . \tag{4.10}$$

We note that (4.10) is not new; it also appears in prior works [WRJ16, DO19, BG19], although with different motivations.

Then as before, to prove the validity of the chosen Lyapunov function, it suffices to verify $(\mathcal{E}_1^{\mathsf{AGM}}) + (\sum_{i=1}^t \eta_i)\cdot(\mathcal{E}_2^{\mathsf{AGM}}) \le 0$, which is equivalent to

$$\tfrac{1}{2L}\left(L\eta_{t+1}^2 - \textstyle\sum_{i=1}^{t+1}\eta_i\right)\|\nabla f(y_t)\|^2 + \left(L\eta_t\eta_{t+1} - \sum_{i=1}^t \eta_i\right)\langle \nabla f(y_t), z_t - y_t\rangle \le 0 \tag{4.11}$$

From (4.11), it suffices to choose $\{\eta_t\}$ so that $L\eta_t\eta_{t+1} = \sum_{i=1}^t \eta_i$. Indeed, with such a choice, the coefficient of the inner product term in (4.11) becomes zero and the coefficient of the squared norm term becomes $1/2L(L\eta_{t+1}^2 - L\eta_{t+1}\eta_{t+2}) \le 0$ (if $\{\eta_t\}$ is increasing). Indeed, one can quickly notice that choosing $\eta_t = t/2L$ satisfies the desired relation. Therefore, we obtain the well known accelerated convergence rate of $f(z_T) - f(x_*) \le \frac{2L\|x_0 - x_*\|^2}{T(T+1)} = O(1/T^2)$.

5 Similar triangle approximations and other variants

In §4, we have demonstrated that AGM is nothing but an approximation of PPM. This view point has not only provided simple derivations of versions of AGM, but also offered clear explanations of the stepsizes. In this section, we demonstrate that these interpretations offered by PPM actually lead to a great simplification of Nesterov's AGM in the form of the *method of similar triangles* [Nes18, GN18].

Our starting point is the observations made in the previous section: (i) from §4.1, we have seen $\|y_t - x_t\| : \|y_t - z_t\| = \eta_t : 1/L$; (ii) from §4.2, we have seen that we need to choose $\eta_t = \Theta(t)$, in which case $\eta_{t+1} \approx \eta_t \gg 1$. From these observations, one can readily see that the triangle $\triangle x_t x_{t+1} z_t$ is approximately similar to $\triangle y_t z_{t+1} z_t$. Therefore, one can simplify AGM by further exploiting this fact: we modify the updates so that the two triangles are indeed *similar*. There are two different ways one can keep the two triangles similar:

1. We modify the update of x_{t+1} so that the two triangles are similar.
2. We modify the update of z_{t+1} so that the two triangles are similar.

We discuss the above two ways in turn.

5.1 First similar triangles approximation: momentum form of AGM We first adopt the first way to keep the two triangles similar. We have the following update.

First similar triangle approximation:

$$y_t = \tfrac{1/L}{1/L+\eta_t} x_t + \tfrac{\eta_t}{1/L+\eta_t} z_t\,, \tag{5.12a}$$

$$z_{t+1} = y_t - \tfrac{1}{L}\nabla f(y_t)\,, \tag{5.12b}$$

$$x_{t+1} = z_{t+1} + L\eta_t(z_{t+1} - z_t)\,. \tag{5.12c}$$

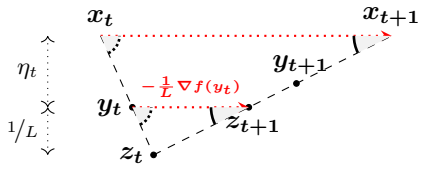


Figure 3: The updates of (5.12).

In fact, (5.12) can be equivalently expressed without $\{x_t\}$, as illustrated with dots in Figure 3. More specifically, during the t-th iteration, once we compute (5.12b), one can directly update y_{t+1} via $y_{t+1} = z_{t+1} + \frac{L\eta_t}{L\eta_{t+1}+1}(z_{t+1} - z_t)$. In other words,

$$(5.12) \quad \Longleftrightarrow \quad \begin{cases} z_{t+1} = y_t - \frac{1}{L}\nabla f(y_t)\,, \\ y_{t+1} = z_{t+1} + \frac{L\eta_t}{L\eta_{t+1}+1}(z_{t+1} - z_t)\,. \end{cases}$$

Hence, (5.12) is equivalent to the well-known momentum form of AGM ("Form I" in the introduction).

Recovering popular stepsize choices. Notably, our PPM-based analysis suggests the choice of $\{\eta_t\}$ as per the recursive relation $(L\eta_{t+1} + \frac{1}{2})^2 = (L\eta_t + 1)^2 + \frac{1}{4}$, which after substitution $L\eta_t + 1 \leftarrow a_t$ exactly recovers the popular recursive relation $a_{t+1} = \frac{1}{2}(1 + \sqrt{1 + 4a_t^2})$ in [Nes83, BT09]. The analysis is similar to the one given in §4.2. Below we provide the details.

Following §4.2, we again derive the following counterparts of (Ineq_1) and (Ineq_2) with straightforward arguments (see §A.2 for details):

$$(\mathsf{Ineq}_1^{\mathsf{SIM}}) \qquad \widetilde{\eta}_{t+1}[f(z_{t+1}) - f(x_*)] + \frac{1}{2}\|x_* - x_{t+1}\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \le (\mathcal{E}_1^{\mathsf{SIM}})\,,$$

$$(\mathsf{Ineq}_2^{\mathsf{SIM}}) \qquad f(z_{t+1}) - f(z_t) \le (\mathcal{E}_2^{\mathsf{SIM}})\,,$$

where using the notation $\widetilde{\eta}_{t+1} := \eta_t + \frac{1}{L}$, the right hand side of the above inequalites are defined as $(\mathcal{E}_1^{\mathsf{SIM}}) := \frac{1}{2}\left(-(L\eta_t + 1)^2 + L\widetilde{\eta}_{t+1}\right) \cdot \|z_{t+1} - y_t\|^2 + \widetilde{\eta}_{t+1} \cdot \langle \nabla f(y_t), z_{t+1} - x_{t+1}\rangle$ and $(\mathcal{E}_2^{\mathsf{SIM}}) := \frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), z_{t+1} - z_t\rangle$.

Having established counterparts of (Ineq_1) and (Ineq_2), following §4.2, we choose

$$\Phi_t := \left(\textstyle\sum_{i=1}^t \widetilde{\eta}_i\right) \cdot (f(z_t) - f(x_*)) + \frac{1}{2}\|x_* - x_t\|^2\,. \tag{5.13}$$

To prove the validity of the chosen Lyapunov function, it suffices to verify

$$(\mathcal{E}_1^{\mathsf{SIM}}) + \left(\textstyle\sum_{i=1}^t \widetilde{\eta}_i\right) \cdot (\mathcal{E}_2^{\mathsf{SIM}}) \le 0 \tag{5.14}$$

which is equivalent to showing (because $z_{t+1} - x_{t+1} = -L\eta_t(z_{t+1} - z_t)$):

$$\tfrac{1}{2}\left(-(L\eta_t + 1)^2 + \textstyle\sum_{i=1}^{t+1} L\widetilde{\eta}_i\right) \cdot \|z_{t+1} - y_t\|^2 + \left(L\eta_t\widetilde{\eta}_{t+1} - \textstyle\sum_{i=1}^t \widetilde{\eta}_i\right)\langle \nabla f(y_t), z_{t+1} - z_t\rangle \le 0\,. \tag{5.15}$$

From (5.15), it suffices to choose $\{\eta_t\}$ so that $L\eta_t\widetilde{\eta}_{t+1} = \sum_{i=1}^t \widetilde{\eta}_i$. Indeed, with such a choice, the coefficient of the inner product term in (5.15) becomes zero and the coefficient of the squared norm term becomes

$$\begin{aligned} \tfrac{1}{2}\left(-(L\eta_t + 1)^2 + \textstyle\sum_{i=1}^{t+1} L\widetilde{\eta}_i\right) &= \frac{1}{2}\left(-(L\eta_t + 1)^2 + L\widetilde{\eta}_{t+1} + L\widetilde{\eta}_{t+1} \cdot L\eta_t\right) \\ &= \tfrac{1}{2}\left(-(L\eta_t + 1)^2 + L\widetilde{\eta}_{t+1}(L\eta_t + 1)\right) = 0 \end{aligned}$$

since $L\widetilde{\eta}_{t+1} = L\eta_t + 1$. Indeed, one can actually simplify the relation $L\eta_t\widetilde{\eta}_{t+1} = \sum_{i=1}^t \widetilde{\eta}_i$:

$$L\eta_{t+1} \cdot (L\eta_{t+1} + 1) = L\eta_{t+1} \cdot L\widetilde{\eta}_{t+2} = \textstyle\sum_{i=1}^{t+1} L\widetilde{\eta}_i = L\widetilde{\eta}_{t+1} + L\eta_t \cdot L\widetilde{\eta}_{t+1} = (L\eta_t + 1)^2\,.$$

After rearranging, we obtain the recursive relation: $(L\eta_{t+1} + \frac{1}{2})^2 = (L\eta_t + 1)^2 + \frac{1}{4}$, which after the substitution $L\eta_t + 1 = a_t$ exactly recovers the popular recursive relation $a_{t+1} = \frac{1+\sqrt{1+4a_t^2}}{2}$ in [Nes83, BT09].

5.2 Second similar triangles approximation: acceleration for composite costs We now adopt the second way to keep the two triangles similar. We have the following update.

Second similar triangle approximation:

$$y_t = \tfrac{1/L}{1/L+\eta_t} x_t + \tfrac{\eta_t}{1/L+\eta_t} z_t \,, \tag{5.16a}$$

$$x_{t+1} = x_t - \eta_{t+1} \nabla f(y_t) \,, \tag{5.16b}$$

$$z_{t+1} = \tfrac{1/L}{1/L+\eta_t} x_{t+1} + \tfrac{\eta_t}{1/L+\eta_t} z_t \,. \tag{5.16c}$$

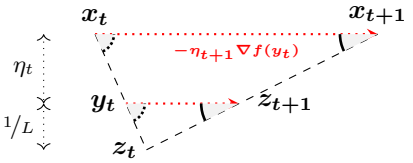


Figure 4: Illustration of (5.16).

This is "Form III" in the introduction. Below, we provide a PPM-based analysis for a more general setting.

One advantage of (5.16) is that it admits a simple extension to the practical setting of constrained optimization on composite costs (see e.g. [Nes18, §6.1.3] for applications). More specifically, for a closed convex set $Q \subseteq \mathbb{R}^d$ and a closed[1] convex function $\Psi : Q \to \mathbb{R}$, consider

$$\min_{x \in Q} f^{\Psi}(x) := f(x) + \Psi(x) \,,$$

where $f : Q \to \mathbb{R}$ is a differentiable convex function which is L-smooth with respect to a norm $\|\cdot\|$ that is not necessarily the ℓ_2 norm (i.e., we regard the norm in Definition 3.1 to be our chosen norm). For the general norm case, we use the Bregman divergence.

DEFINITION 5.1. *Given a 1-strongly convex (w.r.t the chosen norm $\|\cdot\|$) function $h : Q \to \mathbb{R} \cup \{\infty\}$ that is differentiable on the interior of Q, $D_h(u,v) := h(u) - h(v) - \langle \nabla h(v), u - v \rangle$ for all $u, v \in Q$.*

Under the above setting and assumption, (5.16) admits a simple generalization:

Generalization of (5.16) **to composite costs:**

$$y_t = \tfrac{1/L}{1/L+\eta_t} x_t + \tfrac{\eta_t}{1/L+\eta_t} z_t \,, \tag{5.17a}$$

$$x_{t+1} \leftarrow \operatorname{argmin}_{x \in Q} \left\{ \mathsf{LOWER}(x; y_t) + \tfrac{1}{\eta_{t+1}} D_h(x, x_t) + \Psi(x) \right\} , \tag{5.17b}$$

$$z_{t+1} = \tfrac{1/L}{1/L+\eta_t} x_{t+1} + \tfrac{\eta_t}{1/L+\eta_t} z_t \,. \tag{5.17c}$$

Now we provide a simple PPM-based analysis of (5.17):

Proof. [**PPM-based analysis of** (5.17)] To obtain counterparts of (lneq_1) and (lneq_2), we now use a generalization of Proposition 2.1 to the Bregman divergence ([Teb18, Lemma 3.1]). With such a generalization, we obtain the following inequality for $\phi^{\Psi}(x) := \eta_{t+1}[f(y_t) + \langle \nabla f(y_t), x - y_t \rangle + \Psi(x)]$:

$$\phi^{\Psi}(x_{t+1}) - \phi^{\Psi}(x_*) + D_h(x_*, x_{t+1}) + D_h(x_{t+1}, x_t) - D_h(x_*, x_t) \le 0 \,, \tag{5.18}$$

where $x_* \in \operatorname{argmin}_{x \in Q} f^{\Psi}(x)$. Now using (5.18), one can derive from first principles the following inequalities (we defer the derivations to §A.3):

$$(\mathsf{lneq}_1^{\mathsf{SIM}'}) \qquad \eta_{t+1}(f^{\Psi}(z_{t+1}) - f^{\Psi}(x_*)) + D_h(x_*, x_{t+1}) - D_h(x_*, x_t) \le (\mathcal{E}_1^{\mathsf{SIM}'}) \,,$$

$$(\mathsf{lneq}_2^{\mathsf{SIM}'}) \qquad f^{\Psi}(z_{t+1}) - f^{\Psi}(z_t) \le (\mathcal{E}_2^{\mathsf{SIM}'}) \,.$$

where $(\mathcal{E}_1^{\mathsf{SIM}'}) := -\frac{1}{2}\|x_{t+1} - x_t\|^2 + \eta_{t+1}[\frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), z_{t+1} - x_{t+1} \rangle + \Psi(z_{t+1}) - \Psi(x_{t+1})]$ and $(\mathcal{E}_2^{\mathsf{SIM}'}) := \frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), z_{t+1} - z_t \rangle + \Psi(z_{t+1}) - \Psi(z_t)$. Similar to §4.2, yet replacing the norm squared term with the Bregman divergence, we choose

$$\Phi_t := (\textstyle\sum_{i=1}^{t} \eta_i) \cdot (f^{\Psi}(z_t) - f^{\Psi}(x_*)) + D_h(x_*, x_t) \,.$$

[1] This means that the epigraph of the function is closed. See [Nes18, Definition 3.1.2].

Then, it suffices to show $(\mathcal{E}_1^{\mathsf{SIM}'}) + (\sum_{i=1}^t \eta_i) \cdot (\mathcal{E}_2^{\mathsf{SIM}'}) \le 0$. Using the facts (i) $z_{t+1} - x_{t+1} = L\eta_t(z_t - z_{t+1})$ and (ii) $\|x_{t+1} - x_t\| = (L\eta_t + 1)\|z_{t+1} - y_t\|$ (both are immediate consequences of the similar triangles) and rearranging, one can easily check that $(\mathcal{E}_1^{\mathsf{SIM}'}) + (\sum_{i=1}^t \eta_i) \cdot (\mathcal{E}_2^{\mathsf{SIM}'})$ is equal to

$$\tfrac{1}{2}\left(-(L\eta_t+1)^2 + L\eta_{t+1} + L\textstyle\sum_{i=1}^t \eta_i\right)\|z_{t+1} - y_t\|^2 \tag{5.19}$$

$$+\left(L\eta_t\eta_{t+1} - \textstyle\sum_{i=1}^t \eta_i\right)\langle \nabla f(y_t), z_t - z_{t+1}\rangle \tag{5.20}$$

$$+\eta_{t+1}[\Psi(z_{t+1}) - \Psi(x_{t+1})] + \left(\textstyle\sum_{i=1}^t \eta_i\right)\cdot[\Psi(z_{t+1}) - \Psi(z_t)]. \tag{5.21}$$

Now choosing $\eta_t = t/2L$ analogously to §4.2, one can easily verify (5.19) + (5.20) + (5.21) ≤ 0. Indeed, for (5.19), since $L\eta_t\eta_{t+1} = \sum_{i=1}^t \eta_i$, the coefficient becomes $1/2(L\eta_t+1)(L\eta_{t+1} - L\eta_t - 1)$ which is a negative number since $L\eta_{t+1} - L\eta_t - 1 = -1/2$; for (5.20), the coefficient becomes zero due to the relation $L\eta_t\eta_{t+1} = \sum_{i=1}^t \eta_i$; lastly, for (5.21), we have

$$(5.21) = \eta_{t+1}\left[(1+L\eta_t)\Psi(z_{t+1}) - \Psi(x_{t+1}) - L\eta_t\Psi(z_t)\right] \le 0\,, \tag{5.22}$$

where the equality is due to the relation $L\eta_t\eta_{t+1} = \sum_{i=1}^t \eta_i$, and the inequality is due to the update (5.17c) (which can be equivalently written as $(1+L\eta_t)z_{t+1} = x_{t+1} + L\eta_t z_t$) and the convexity of Ψ. Hence, we obtain the accelerated rate of $f^\Psi(z_T) - f^\Psi(x_*) \le \frac{4LD_h(x_*,x_0)}{T(T+1)} = O(1/T^2)$. ∎

6 Extension to strongly convex costs

In this section, we extend our PPM framework to the case of strongly convex costs. As we shall see, our framework gives rise to a simple derivation of the most general version of AGM called "*General Scheme for Optimal Method*" [Nes18, (2.2.7)]. We first make the approximate PPM (4.8) more flexible by considering two separate stepsizes.

Approximate PPM with two separate stepsizes $\{\eta_t\}$ and $\{\widetilde{\eta}_t\}$. Given $x_0 = y_0 \in \mathbb{R}^d$,

$$x_{t+1} \leftarrow \operatorname{argmin}_x \left\{\mathsf{LOWER}(x; y_t) + \tfrac{1}{2\eta_{t+1}}\|x - x_t\|^2\right\}, \tag{6.23a}$$

$$y_{t+1} \leftarrow \operatorname{argmin}_x \left\{\mathsf{UPPER}(x; y_t) + \tfrac{1}{2\widetilde{\eta}_{t+1}}\|x - x_{t+1}\|^2\right\}. \tag{6.23b}$$

Now let us apply our PPM view to the strongly convex cost case.

Definition 6.1. (Strong convexity) *For $\mu > 0$, we say a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if $f(x) \ge f(y) + \langle \nabla f(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2$ for any $x, y \in \mathbb{R}^d$.*

Since f is additionally assumed to be strongly convex, one can now strengthen the lower approximation $\mathsf{LOWER}(x; y_t)$ in (6.23a) to $\mathsf{LOWER}(x; y_t) + \frac{\mu}{2}\|x - y_t\|^2$. In other words, we obtain

Approximate PPM for strongly-convex costs. Given $x_0 = y_0 \in \mathbb{R}^d$,

$$x_{t+1} \leftarrow \operatorname*{argmin}_{x\in\mathbb{R}^d} \left\{\mathsf{LOWER}(x; y_t) + \underbrace{\frac{\mu}{2}\|x - y_t\|^2}_{\substack{\text{additional term due to}\\ \text{strong convexity}}} + \frac{1}{2\eta_{t+1}}\|x - x_t\|^2\right\}, \tag{6.24a}$$

$$y_{t+1} \leftarrow \operatorname{argmin}_x \left\{\mathsf{UPPER}(x; y_t) + \tfrac{1}{2\widetilde{\eta}_{t+1}}\|x - x_{t+1}\|^2\right\}. \tag{6.24b}$$

Writing the optimality condition of (6.24), it is straightforward to check that the approximate PPM (6.23) is equivalent to the following updates ($x_0 = y_0 = z_0$):

Equivalent representation of (6.24)**:**

$$y_t = \tfrac{1/L}{1/L+\widetilde{\eta}_t} x_t + \tfrac{\widetilde{\eta}_t}{1/L+\widetilde{\eta}_t} z_t \,, \tag{6.25a}$$

$$x_{t+1} = \tfrac{1/\mu}{1/\mu+\eta_{t+1}} x_t + \tfrac{\eta_{t+1}}{1/\mu+\eta_{t+1}} y_t - \tfrac{1/\mu \cdot \eta_{t+1}}{1/\mu+\eta_{t+1}} \nabla f(y_t) \,, \tag{6.25b}$$

$$z_{t+1} = y_t - \tfrac{1}{L}\nabla f(y_t) \,. \tag{6.25c}$$

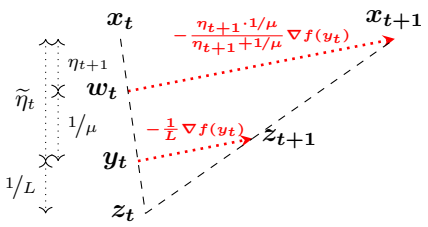


Figure 5: Illustration of (6.25).

Note that (6.25) is the most general version of AGM due to Nesterov called "*General Scheme for Optimal Method*" [Nes18, (2.2.7)] ("Form IV" in the introduction). Again, our derivation provides new insights into the choices of the AGM stepsizes by expressing them in terms of the PPM stepsizes η_t's and $\widetilde{\eta}_t$'s.

6.1 Relation to well known momentum version Perhaps, the most well known version of AGM for strongly convex costs is the momentum version due to Nesterov (see, e.g., [Nes18, (2.2.22)])

$$\begin{aligned} z_{t+1} &= y_t - \tfrac{1}{L}\nabla f(y_t) \,, \\ y_{t+1} &= z_{t+1} + \tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(z_{t+1} - z_t) \,. \end{aligned} \tag{6.26}$$

One might wonder whether one can better understand the stepsizes in (6.26) from (6.25).

Let us first recall the well known convergence rate of PPM for strongly convex costs due to Rockafellar [Roc76, (1.14)]:

$$f(x_T) - f(x_*) \le O\left(\textstyle\prod_{t=1}^{T}(1+\mu\eta_t)^{-1}\right) \quad \text{for any } T \ge 1. \tag{6.27}$$

From (6.27), one can see that in order to achieve the accelerated convergence rate $O(\exp(-T/\sqrt{\kappa}))$ where κ is the condition number L/μ, the stepsizes η_t must be chosen so that $\eta_t \approx \mu^{-1}(\sqrt{\kappa})^{-1}$. In fact, the well known version (6.26) corresponds to choosing the following stepsizes for (6.25):

$$\eta_t \equiv \eta := \mu^{-1}(\sqrt{\kappa}-1)^{-1} \quad \text{and} \quad \widetilde{\eta}_t \equiv \widetilde{\eta} := \mu^{-1}(\sqrt{\kappa})^{-1}. \tag{6.28}$$

To see this, note that with such choice of η and $\widetilde{\eta}$, (6.25) becomes:

(6.25) with stepsize chosen as (6.28)**:**

$$y_t = \tfrac{1}{1+\sqrt{\kappa}} x_t + \tfrac{\kappa}{1+\sqrt{\kappa}} z_t \,, \tag{6.29a}$$

$$x_{t+1} = \tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}} x_t + \tfrac{1}{\sqrt{\kappa}} y_t - \tfrac{\sqrt{\kappa}}{L}\nabla f(y_t) \,, \tag{6.29b}$$

$$z_{t+1} = y_t - \tfrac{1}{L}\nabla f(y_t) \,. \tag{6.29c}$$

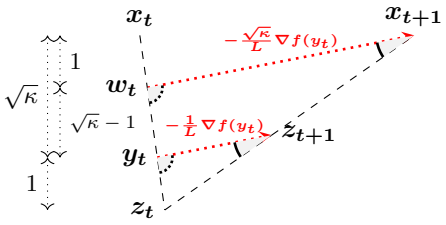


Figure 6: Illustration of (6.29).

As shown in Figure 6, $\triangle w_t x_{t+1} z_t$ is similar to $\triangle y_t z_{t+1} z_t$, so one can write the updates (6.29) without $\{x_t\}$ and $\{w_t\}$, which precisely recovers (6.26):

$$(6.29) \iff (6.26) = \begin{cases} z_{t+1} = y_t - \tfrac{1}{L}\nabla f(y_t) \,, \\ y_{t+1} = z_{t+1} + \tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(z_{t+1} - z_t) \,. \end{cases}$$

7 Related work

Our approach is inspired by that of Defazio [Def19] that establishes an inspiring connection between AGM and PPM. The main observation in that paper is that for strongly convex costs, one can derive a version of AGM from

the primal-dual form of PPM with a tweak of geometry. Compared with [Def19], our approach strengthens the connection between AGM and PPM by considering more versions of AGM and their analyses. Another advantage of our approach is that it does not require duality.

We now summarize previous works on developing alternative approaches to Nesterov's acceleration. Most works have studied the continuous limit dynamics of Nesterov's AGM [SBC16, KBB15, WWJ16]. These continuous dynamics approaches have brought about new intuitions about Nesterov's acceleration, and follow-up works have developed analytical techniques for such dynamics [WRJ16, DO19]. Another notable contribution is made based on the linear coupling framework [AZO17]. The main observation is that the two most popular first-order methods, namely gradient descent and mirror descent, have complementary performances, and hence, one can come up with a faster method by linearly coupling the two methods. Lastly, Nesterov's acceleration has been explained from the perspective of computing the equilibrium in a primal-dual game [WA18, CST21].

PPM has been used to design or interpret other optimization methods [Dru17]. To list few instances, PPM has given rise to fast methods for weakly convex problems [DG19], the prox-linear methods for composite optimizations [BF95, Nes07, LW16], accelerated methods for stochastic optimizations [LMH15], and methods for saddle-point problems [MOP19].

8 Conclusion

This work provides a way to understand Nesterov's acceleration based on the proximal point method. The framework presented in this paper motivates a simplification of AGM using similar triangles and readily extends to the strongly convex case and recovers the most general accelerated method due to Nesterov.

We believe that the simple derivations presented in this paper clarify and deepen our understanding of Nesterov's acceleration. Our framework is therefore not only of pedagogical value but also helpful for research. For future directions, it would be interesting to connect our PPM view to accelerated stochastic methods [LMH15, LZ18] and other accelerated methods, including geometric descent [BLS15]. Furthermore, we hope the connections presented in this work will help advance the development of accelerated methods in settings much wider than convex optimization (see e.g., [Bac14]).

Acknowledgement

We thank Alp Yurtsever and Jingzhao Zhang for detailed comments and stimulating discussions, Aaron Defazio for clarifications that help the author develop §5.2, and Heinz Bauschke for constructive suggestions on the presentation of §4 and §6.1. Kwangjun Ahn and Suvrit Sra acknowledge support from the NSF Grant (CAREER: 1846088). Kwangjun Ahn also acknowledge support from Kwanjeong Educational Foundation.

References

[AT06] Alfred Auslender and Marc Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.

[AZO17] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *ITCS 2017*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[Bac14] Miroslav Bacák. *Convex analysis and optimization in Hadamard spaces*, volume 22. Walter de Gruyter GmbH & Co KG, 2014.

[BBC11] Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[BC11] Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[BF95] James V Burke and Michael C Ferris. A Gauss-Newton method for convex composite optimization. *Mathematical Programming*, 71(2):179–194, 1995.

[BG19] Nikhil Bansal and Anupam Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(4):1–32, 2019.

[BLS15] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint: 1506.08187*, 2015.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[CST21] Michael B Cohen, Aaron Sidford, and Kevin Tian. Relative lipschitzness in extragradient methods and a direct recipe for acceleration. In *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

[Def19] Aaron Defazio. On the curved geometry of accelerated optimization. In *Advances in Neural Information Processing Systems*, pages 1764–1773, 2019.

[DG19] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

[DO19] Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019.

[Dru17] Dmitriy Drusvyatskiy. The proximal point method revisited. *arXiv preprint: 1712.06038*, 2017.

[GN18] Alexander Vladimirovich Gasnikov and Yu E Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58(1):48–64, 2018.

[Gül91] Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991.

[KBB15] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems*, pages 2845–2853, 2015.

[LMH15] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

[LRP16] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[LRS13] Yin Tat Lee, Satish Rao, and Nikhil Srivastava. A new approach to computing maximum flows using electrical flows. In *Proceedings of ACM STOC*, pages 755–764, 2013.

[LW16] Adrian S Lewis and Stephen J Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1-2):501–546, 2016.

[LZ18] Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, 171(1-2):167–215, 2018.

[Mar70] Bernard Martinet. Régularisation d'inéquations variationnelles par approximations successives. rev. française informat. *Recherche Opérationnelle*, 4:154–158, 1970.

[MOP19] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: proximal point approach. *arXiv preprint: 1901.08511*, 2019.

[Mor65] Jean-Jacques Moreau. Proximité et dualité dans un espace Hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[Nes83] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

[Nes07] Yu Nesterov. Modified Gauss-Newton scheme with worst case guarantees for global performance. *Optimisation methods and software*, 22(3):469–483, 2007.

[Nes18] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[Roc76] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[SBC16] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *JMLR*, 17(1):5312–5354, 2016.

[SMDH13] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, pages 1139–1147, 2013.

[Teb18] Marc Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, 2018.

[Tse08] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2008.

[WA18] Jun-Kun Wang and Jacob D Abernethy. Acceleration through optimistic no-regret dynamics. *Advances in Neural Information Processing Systems*, 31, 2018.

[WRJ16] Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint: 1611.02635*, 2016.

[WWJ16] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *PNAS*, 113(47):E7351–E7358, 2016.

A Deferred derivations

A.1 Deferred derivations from §4.2 Let us first derive ($\mathsf{Ineq}_1^{\mathsf{AGM}}$). Applying Proposition 2.1 with $\phi(x) = \eta_{t+1}[f(y_t) + \langle \nabla f(y_t), x - y_t\rangle]$ to (4.8a), we obtain:

$$\phi(x_{t+1}) - \phi(x_*) + \frac{1}{2}\|x_* - x_{t+1}\|^2 + \frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \le 0\,. \tag{A.1}$$

Now from the convexity of f, it holds that $\phi(x_*) \le \eta_{t+1} f(x_*)$. This together with the L-smoothness of f, it follows that

$$\begin{aligned}\phi(x_{t+1}) &= \eta_{t+1}[f(y_t) + \langle \nabla f(y_t), z_{t+1} - y_t\rangle + \langle \nabla f(y_t), x_{t+1} - z_{t+1}\rangle] \\ &\ge \eta_{t+1}\left[f(z_{t+1}) - \frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), x_{t+1} - z_{t+1}\rangle\right].\end{aligned}$$

Plugging these inequalities back to (A.1) and rearranging, we obtain the following inequality:

$$\begin{aligned}&\eta_{t+1}[f(z_{t+1}) - f(x_*)] + \frac{1}{2}\|x_* - x_{t+1}\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \\ &\le -\frac{1}{2}\|x_{t+1} - x_t\|^2 + \eta_{t+1}\left[\frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), z_{t+1} - x_{t+1}\rangle\right].\end{aligned} \tag{A.2}$$

Now decomposing the inner product term in (A.2) into

$$\eta_{t+1}\langle \nabla f(y_t), z_{t+1} - y_t\rangle + \eta_{t+1}\langle \nabla f(y_t), y_t - x_t\rangle + \eta_{t+1}\langle \nabla f(y_t), x_t - x_{t+1}\rangle,$$

and using $x_{t+1} - x_t = -\eta_{t+1}\nabla f(y_t)$ and $z_{t+1} - y_t = -1/L\nabla f(y_t)$ (which are (4.9b) and (4.9c), respectively), (A.2) becomes $\left(\frac{\eta_{t+1}^2}{2} - \frac{\eta_{t+1}}{2L}\right)\|\nabla f(y_t)\|^2 + \eta_{t+1}\langle \nabla f(y_t), y_t - x_t\rangle$. Now, using the relation $y_t - x_t = L\eta_t(z_t - y_t)$ (which is (4.9a)), we obtain ($\mathcal{E}_1^{\mathsf{AGM}}$). Thus, ($\mathsf{Ineq}_1^{\mathsf{AGM}}$) follows.

Next, ($\mathsf{Ineq}_2^{\mathsf{AGM}}$) readily follows from the L-smoothness and the convexity of f:

$$\begin{aligned}f(z_{t+1}) - f(z_t) &= f(z_{t+1}) - f(y_t) + f(y_t) - f(z_t) \\ &\le \langle \nabla f(y_t), z_{t+1} - y_t\rangle + \frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), y_t - z_t\rangle \\ &\overset{(a)}{=} -\frac{1}{2L}\|\nabla f(y_t)\|^2 + \langle \nabla f(y_t), y_t - z_t\rangle = (\mathcal{E}_2^{\mathsf{AGM}}),\end{aligned}$$

where (a) is due to $z_{t+1} - y_t = -1/L\nabla f(y_t)$.

A.2 Deferred derivations from §5.1 We first derive ($\mathsf{Ineq}_1^{\mathsf{SIM}}$). By the updates (5.12), we have $x_{t+1} = x_t - (\eta_t + \frac{1}{L})\nabla f(y_t)$. Letting $\widetilde{\eta}_{t+1} := \eta_t + \frac{1}{L}$, this relation can be equivalently written as:

$$x_{t+1} \leftarrow \mathrm{argmin}_x \left\{f(y_t) + \langle \nabla f(y_t), x - y_t\rangle + \tfrac{1}{2\widetilde{\eta}_{t+1}}\|x - x_t\|^2\right\} \tag{A.3}$$

The rest is similar to §A.1: we apply Proposition 2.1 with $\phi(x) = \widetilde{\eta}_{t+1}[f(y_t) + \langle \nabla f(y_t), x - y_t\rangle]$:

$$\phi(x_{t+1}) - \phi(x_*) + \frac{1}{2}\|x_* - x_{t+1}\|^2 + \frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \le 0\,. \tag{A.4}$$

Now from the convexity, we have $\phi(x_*) \le \widetilde{\eta}_{t+1} f(x_*)$, and from the L-smoothness, we have

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

Plugging these inequalities back to (A.4) and rearranging, we obtain the following inequality:

$$\begin{aligned}
&\widetilde{\eta}_{t+1}[f(z_{t+1}) - f(x_*)] + \frac{1}{2}\left\|x_* - x_{t+1}\right\|^2 - \frac{1}{2}\left\|x_* - x_t\right\|^2 \\
&\quad \le -\frac{1}{2}\left\|x_{t+1} - x_t\right\|^2 + \widetilde{\eta}_{t+1}\left[\frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), z_{t+1} - x_{t+1}\rangle\right] \\
&\quad = \frac{1}{2}\left(-(L\eta_t + 1)^2 + L\widetilde{\eta}_{t+1}\right) \cdot \left\|z_{t+1} - y_t\right\|^2 + \widetilde{\eta}_{t+1} \cdot \langle \nabla f(y_t), z_{t+1} - x_{t+1}\rangle = (\mathcal{E}_1^{\mathsf{SIM}}),
\end{aligned}$$

where the last line follows since $\|x_{t+1} - x_t\| = (L\eta_t + 1) \cdot \|z_{t+1} - z_t\|$ (see Figure 3).

Next we derive ($\mathsf{Ineq}_1^{\mathsf{SIM}}$). From the L-smoothness and the convexity of f:

$$\begin{aligned}
f(z_{t+1}) - f(z_t) &= f(z_{t+1}) - f(y_t) + f(y_t) - f(z_t) \\
&\le \langle \nabla f(y_t), z_{t+1} - y_t\rangle + \frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), y_t - z_t\rangle \\
&= \frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), z_{t+1} - z_t\rangle = (\mathcal{E}_2^{\mathsf{SIM}}).
\end{aligned}$$

A.3 Deferred derviations from §5.2 Let us first derive ($\mathsf{Ineq}_1^{\mathsf{SIM}'}$). From convexity, we have $\phi^\Psi(x_*) \le \eta_{t+1} f^\Psi(x_*)$, and from the L-smoothness, we have the following lower bound:

$$\begin{aligned}
\phi^\Psi(x_{t+1}) &= \eta_{t+1}[f(y_t) + \langle \nabla f(y_t), z_{t+1} - y_t\rangle + \langle \nabla f(y_t), x_{t+1} - z_{t+1}\rangle + \Psi(x_{t+1})] \\
&\ge \eta_{t+1}\left[f^\Psi(z_{t+1}) - \frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), x_{t+1} - z_{t+1}\rangle + \Psi(x_{t+1}) - \Psi(z_{t+1})\right].
\end{aligned}$$

Plugging these back to (5.18), and using the bound $-D_h(x_{t+1}, x_t) \le -\frac{1}{2}\|x_{t+1} - x_t\|^2$, ($\mathsf{Ineq}_1^{\mathsf{SIM}'}$) follows.

Next, to derive ($\mathsf{Ineq}_2^{\mathsf{SIM}'}$), we use L-smoothness and the convexity of f to obtain the following:

$$\begin{aligned}
f^\Psi(z_{t+1}) - f^\Psi(z_t) &\le f(z_{t+1}) - f(y_t) + f(y_t) - f(z_t) + \Psi(z_{t+1}) - \Psi(z_t) \\
&\le \frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), z_{t+1} - z_t\rangle + \Psi(z_{t+1}) - \Psi(z_t),
\end{aligned}$$

which is precisely equal to ($\mathcal{E}_2^{\mathsf{SIM}'}$).