CS395T: Continuous Algorithms, Part V
Acceleration and high-order methods

Kevin Tian

1 Acceleration

In this lecture, we conclude our development on the basic theory of derivative-based optimization for structured functions. We begin by discussing the phenomenon of *acceleration*, a powerful tool introduced by [Nes83] (with an earlier version in [Nem82]). The main sales pitch for acceleration is that methods developed with this technique have resulted in provably optimal algorithms under a variety of oracle access models (e.g., gradient or high-order derivative queries). The general philosophy behind acceleration (maintaining history-dependent updates, or implementing *momentum*) has also been very empirically succesful in training deep neural networks [SMDH13].

Unfortunately, acceleration has garnered a reputation for being rather difficult to understand. There has been significant effort by the community to develop more intuition for how acceleration arises [Har13, BLS15, SBC16] and how to design accelerated algorithms [LRP16, ZO17, CST21].

We present a proof of acceleration which is built up to in several stages, patterned off Parts II and III of the notes. We begin from a continuous perspective (i.e., a *second-order* ODE), implicitly discretize the continuous dynamics to obtain an accelerated proximal point method, and finally fully discretize it to give an explicit gradient-based method. Each stage has relatively short proofs, which we hope grants intuition on how the different parts of the final accelerated method arise.

1.1 Acceleration in continuous time

The basic intuition behind acceleration is that it learns a trajectory over time, by accumulating previously-queried gradients. This can be naturally modeled as a *second-order* differential equation, which maintains an acceleration variable (i.e., the change in velocity) $\ddot{\mathbf{x}}_t := \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{x}_t$, in addition to a velocity variable $\dot{\mathbf{x}}_t$. Note that in the gradient flow dynamics, we only maintain a velocity variable.

The following accelerated gradient flow dynamics were introduced by [SBC16]:

$$\ddot{\mathbf{x}}_t + \frac{3}{t}\dot{\mathbf{x}}_t + \nabla f(\mathbf{x}_t) = 0. \tag{1}$$

An equivalent way to interpret the dynamics (1) is by decoupling the position and velocity variables:

$$\dot{\mathbf{v}}_t = -\frac{3}{t}\mathbf{v}_t - \nabla f(\mathbf{x}_t),\ \dot{\mathbf{x}}_t = \mathbf{v}_t.$$

We now give a proof that $f(\mathbf{x}_t)$ decays at an accelerated $\frac{1}{t^2}$ rate, following [SBC16].

Proposition 1 (Accelerated gradient flow). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth for $L > 0$,*[1] *and let $\mathbf{x}^\star \in \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. For $\mathbf{x}_t$ following the ODE (1) starting from $\mathbf{x}_0 \in \mathbb{R}^d$ and $\dot{\mathbf{x}}_0 = \mathbf{0}_d$,*

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \frac{2\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|_2^2}{t^2}.$$

Proof. Our strategy is to prove that $\dot{\Phi}_t \le 0$, where our potential function Φ_t is defined by

$$\Phi_t := t^2\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) + 2\left\|\mathbf{z}_t - \mathbf{x}^\star\right\|_2^2,\ \text{where } \mathbf{z}_t := \mathbf{x}_t + \frac{t}{2}\dot{\mathbf{x}}_t,$$

[1] This is enough to conclude that (1) has a unique solution, by the Picard-Lindelöf theorem.

from which the conclusion follows because $\Phi_t \geq f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ and $\mathbf{z}_0 = \mathbf{x}_0$. We first compute

$$\begin{aligned}\dot{\Phi}_t &= 2t\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) + t^2\left\langle \nabla f(\mathbf{x}_t), \dot{\mathbf{x}}_t\right\rangle + 2\left\langle \mathbf{z}_t - \mathbf{x}^\star, 3\dot{\mathbf{x}}_t + t\ddot{\mathbf{x}}_t\right\rangle \\ &= 2t\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) + t^2\left\langle \nabla f(\mathbf{x}_t), \dot{\mathbf{x}}_t\right\rangle - 2t\left\langle \nabla f(\mathbf{x}_t), \mathbf{z}_t - \mathbf{x}^\star\right\rangle \\ &= 2t\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star) - \left\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star\right\rangle\right),\end{aligned}$$

where we substituted (1) in the second line. Our claim $\dot{\Phi}_t \leq 0$ now follows by convexity. □

While Proposition 1 is a remarkably short proof-of-concept that acceleration is achievable, it is somewhat magical. One takeaway is that it is useful to design a "fast-forwarded" trajectory $\mathbf{z}_t$, which induces cancellations in potential functions due to the second-order nature of (1). A mystery that arises is the specific choice of the constant 3 in (1). There is discussion in Section 4, [SBC16] on a phase transition that arises around this constant based on the damping behavior of the ODE.

We present an alternative convergence guarantee for a variant of accelerated gradient flow in the well-conditioned regime (cf. Section 4, Part II), i.e., assuming that the function of interest f is strongly convex. By the reduction in Lemma 11, Part II between the smooth and well-conditioned regimes, the rates achieved by Proposition 1 and Proposition 2 are analogous.

Proposition 2 (Accelerated gradient flow, well-conditioned regime)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex, let $\kappa := \frac{L}{\mu}$, and let $\mathbf{x}^\star \in \text{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. For $\mathbf{x}_t$ following the ODE*

$$\ddot{\mathbf{x}}_t + \frac{2}{\sqrt{\kappa}}\dot{\mathbf{x}}_t + \frac{1}{L}\nabla f(\mathbf{x}_t) = 0, \tag{2}$$

from $\mathbf{x}_0 \in \mathbb{R}^d$ and $\dot{\mathbf{x}}_0 = \mathbf{0}_d$,

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq 2\exp\left(-\frac{t}{\sqrt{\kappa}}\right)\left(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\right).$$

Proof. Similarly to the proof of Proposition 1, we rewrite (2) in the following way:

$$\dot{\mathbf{x}}_t + \sqrt{\kappa}\ddot{\mathbf{x}}_t = -\frac{1}{\mu\sqrt{\kappa}}\nabla f(\mathbf{x}_t) - \dot{\mathbf{x}}_t,$$

so that the "fast-forwarded" trajectory $\mathbf{z}_t := \mathbf{x}_t + \sqrt{\kappa}\dot{\mathbf{x}}_t$ satisfies

$$\dot{\mathbf{z}}_t = \dot{\mathbf{x}}_t + \sqrt{\kappa}\ddot{\mathbf{x}}_t = -\frac{1}{\mu\sqrt{\kappa}}\nabla f(\mathbf{x}_t) + \frac{1}{\sqrt{\kappa}}(\mathbf{x}_t - \mathbf{z}_t). \tag{3}$$

The potential we track in this proof is

$$\Phi_t := f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}\left\|\mathbf{z}_t - \mathbf{x}^\star\right\|_2^2.$$

We claim $\dot{\Phi}_t \leq -\frac{1}{\sqrt{\kappa}}\Phi_t$, from which the conclusion follows from Grönwall's inequality (Fact 1, Part II) and strong convexity, which implies $\Phi_0 = f(\mathbf{x}_0) - f(\mathbf{x}^\star) + \frac{\mu}{2}\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|_2^2 \leq 2\left(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\right)$.

To prove our claim, we derive

$$\begin{aligned}\dot{\Phi}_t &= \left\langle \nabla f(\mathbf{x}_t), \dot{\mathbf{x}}_t\right\rangle + \mu\left\langle \dot{\mathbf{z}}_t, \mathbf{z}_t - \mathbf{x}^\star\right\rangle \\ &= -\frac{1}{\sqrt{\kappa}}\left(\left\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{z}_t\right\rangle + \left\langle \nabla f(\mathbf{x}_t) + \mu(\mathbf{z}_t - \mathbf{x}_t), \mathbf{z}_t - \mathbf{x}^\star\right\rangle\right) \\ &= -\frac{1}{\sqrt{\kappa}}\left(\left\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star\right\rangle + \mu\left\langle \mathbf{z}_t - \mathbf{x}_t, \mathbf{z}_t - \mathbf{x}^\star\right\rangle\right) \\ &\leq -\frac{1}{\sqrt{\kappa}}\left(\left(\left\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star\right\rangle - \frac{\mu}{2}\left\|\mathbf{x}_t - \mathbf{x}^\star\right\|_2^2\right) + \frac{\mu}{2}\left\|\mathbf{z}_t - \mathbf{x}^\star\right\|_2^2\right) \leq -\frac{1}{\sqrt{\kappa}}\Phi_t.\end{aligned}$$

The second line used our definitions of $\mathbf{z}_t$ and $\dot{\mathbf{z}}_t$ in (3), the first inequality used Eq. (7), Part III and dropped a nonnegative term, and the second inequality used strong convexity of f. □

1.2 Accelerated proximal point

Acceleration is sometimes viewed as arising from a careful "linear coupling" between gradient descent and mirror descent, that benefits from the convergence guarantees of each [ZO17]. Indeed, the key analysis technique from mirror descent (i.e., the three-point equality in Eq. (7), Part III) has already suggestively made an appearance in the proof of Proposition 2.

Before fully discretizing the ODE (1), in this section we present an implicit variant known as the accelerated proximal point algorithm (APPA) [Gü92], in analogy to the proximal point method of Section 2, Part III. We choose this presentation format for a few reasons, e.g., it leads to a slightly simpler proof, and is more consistent with our later development in Section 2 for accelerated high-order methods. Moreover, by comparing this section with our final method in Section 1.3, the role of gradient descent under the linear coupling perspective becomes more clear.

The APPA is initialized with $(A_0, \mathbf{x}_0) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^d$, and is driven by step sizes $\{\lambda_k\}_{k\geq 1} \subset \mathbb{R}_{\geq 0}$. It evolves three sequences $\{\mathbf{x}_k\}_{k\geq 0}$, $\{\mathbf{y}_k\}_{k\geq 0}$, $\{\mathbf{z}_k\}_{k\geq 0}$, initialized at $\mathbf{z}_0 = \mathbf{y}_0 \leftarrow \mathbf{x}_0$, as follows:

$$\begin{aligned} \mathbf{y}_k &\leftarrow \frac{A_k}{A_{k+1}}\mathbf{x}_k + \frac{a_{k+1}}{A_{k+1}}\mathbf{z}_k, \\ \mathbf{z}_{k+1} &\leftarrow \mathbf{z}_k - a_{k+1}\nabla f(\mathbf{x}_{k+1}), \\ \mathbf{x}_{k+1} &\leftarrow \mathbf{y}_k - \lambda_{k+1}\nabla f(\mathbf{x}_{k+1}), \end{aligned} \tag{4}$$

for all $k \geq 0$, where

$$A_{k+1} \leftarrow A_k + a_{k+1},\ a_{k+1} \leftarrow \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4\lambda_{k+1}A_k}}{2}. \tag{5}$$

At this point, it is helpful to provide some intuition on the form of (4), (5), so they do not seem to appear out of nowhere. First, observe that if we fix A_0 and the step size sequence $\{\lambda_k\}_{k\geq 1}$, the sequences $\{A_k, a_k\}_{k\geq 1}$ are uniquely induced via (5). Moreover, by the quadratic formula, we have from (5) that the following recurrence relation holds for all $k \geq 0$:

$$a_{k+1}^2 - \lambda_{k+1}a_{k+1} - \lambda_{k+1}A_k = a_{k+1}^2 - \lambda_{k+1}A_{k+1} = 0 \implies a_{k+1}^2 = \lambda_{k+1}A_{k+1}. \tag{6}$$

The form of the sequences in (4) has a simple intuition. The sequence of $\{\mathbf{x}_k\}_{k\geq 1}$ evolve via the proximal point method (Theorem 1, Part III with $\mathcal{X} = \mathbb{R}^d$ and $\varphi = \frac{1}{2}\|\cdot\|_2^2$) using step sizes $\{\lambda_k\}_{k\geq 1}$. Conversely, the sequence of $\{\mathbf{z}_k\}_{k\geq 1}$ evolve via mirror descent (Theorem 2, Part III), using step sizes $\{a_k\}_{k\geq 1}$ and linear functions induced via the $\{\nabla f(\mathbf{x}_k)\}_{k\geq 1}$.

Finally, the sequence of $\{\mathbf{y}_k\}_{k\geq 1}$ are formed by convex combinations of the other two sequences. For this reason, APPA can be viewed as a "linear coupling" of the proximal point method and mirror descent, foreshadowing a similar linear coupling interpretation of our final accelerated method.

The main invariant satisfied by the iterations (4), (5) is given by the following technical lemma.

Lemma 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and convex. Following the notation (4), (5), define*

$$\Phi_k := A_k\epsilon_k + r_k,\ \text{where } \epsilon_k := f(\mathbf{x}_k) - f(\mathbf{x}^\star),\ \text{and } r_k := \frac{1}{2}\|\mathbf{z}_k - \mathbf{x}^\star\|_2^2, \tag{7}$$

for all $k \geq 0$, where $\mathbf{x}^\star := \text{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. Then for all $k \geq 0$, $\Phi_{k+1} \leq \Phi_k$.

Proof. By the definition of $\mathbf{y}_k$ in (4), we have

$$\begin{aligned} a_{k+1}\langle\nabla f(\mathbf{x}_{k+1}), \mathbf{x}^\star - \mathbf{z}_k\rangle &= a_{k+1}\langle\nabla f(\mathbf{x}_{k+1}), \mathbf{x}^\star - \mathbf{x}_{k+1}\rangle + a_{k+1}\langle\nabla f(\mathbf{x}_{k+1}), \mathbf{x}_{k+1} - \mathbf{z}_k\rangle \\ &= a_{k+1}\langle\nabla f(\mathbf{x}_{k+1}), \mathbf{x}^\star - \mathbf{x}_{k+1}\rangle \\ &\quad + A_k\langle\nabla f(\mathbf{x}_{k+1}), \mathbf{x}_k - \mathbf{x}_{k+1}\rangle + A_{k+1}\langle\nabla f(\mathbf{x}_{k+1}), \mathbf{x}_{k+1} - \mathbf{y}_k\rangle \\ &\leq a_{k+1}(f(\mathbf{x}^\star) - f(\mathbf{x}_{k+1})) + A_k(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})) \\ &\quad + A_{k+1}\langle\nabla f(\mathbf{x}_{k+1}), \mathbf{x}_{k+1} - \mathbf{y}_k\rangle, \end{aligned}$$

where in the only inequality, we applied convexity twice. By substituting the definition of $\mathbf{x}_{k+1}$ from (4), and using the equality (6), we have

$$\begin{aligned} a_{k+1}\left\langle \nabla f(\mathbf{x}_{k+1}), \mathbf{x}^\star - \mathbf{z}_k\right\rangle &\le A_k\left(f(\mathbf{x}_k) - f(\mathbf{x}^\star)\right) - A_{k+1}\left(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^\star)\right) \\ &\quad - a_k^2\left\|\nabla f(\mathbf{x}_{k+1})\right\|_2^2 \\ &= A_k\epsilon_k - A_{k+1}\epsilon_{k+1} - a_{k+1}^2\left\|\nabla f(\mathbf{x}_{k+1})\right\|_2^2. \end{aligned} \tag{8}$$

Moreover, by using the standard mirror descent analysis (see, e.g., Eq. (11), Part III) we have

$$\begin{aligned} a_{k+1}\left\langle \nabla f(\mathbf{x}_{k+1}), \mathbf{z}_k - \mathbf{x}^\star\right\rangle &\le \frac{1}{2}\left\|\mathbf{z}_k - \mathbf{x}^\star\right\|_2^2 - \frac{1}{2}\left\|\mathbf{z}_{k+1} - \mathbf{x}^\star\right\|_2^2 + \frac{a_{k+1}^2}{2}\left\|\nabla f(\mathbf{x}_{k+1})\right\|_2^2 \\ &= r_k - r_{k+1} + \frac{a_{k+1}^2}{2}\left\|\nabla f(\mathbf{x}_{k+1})\right\|_2^2. \end{aligned} \tag{9}$$

Combining (8) and (9), we have the conclusion:

$$0 \le \Phi_k - \Phi_{k+1} - \frac{a_{k+1}^2}{2}\left\|\nabla f(\mathbf{x}_{k+1})\right\|_2^2 \le \Phi_k - \Phi_{k+1}. \tag{10}$$

□

At this point, by applying $\Phi_k \le \Phi_0$, it is evident that the function error ϵ_k decreases at a rate proportional to $\frac{1}{A_k}$, so our goal is to choose $\{\lambda_k\}_{k\ge1}$ so that A_k grows as quickly as possible. In principle, we could simply choose $\lambda_k \to \infty$; however, this will pose issues when we discretize APPA in the following Section 1.3. We provide a simple bound when $\lambda_k \equiv \lambda$ uniformly.

Lemma 2. *Following the notation in (4), (5), suppose that $\lambda_k = \lambda$ for all $k \ge 1$. Then,*

$$A_k \ge \frac{\lambda k^2}{4}.$$

Proof. This follows from the sequence of bounds:

$$\begin{aligned} \sqrt{A_k} = \sqrt{A_k} - \sqrt{A_0} &= \sum_{i\in[k]} \sqrt{A_i} - \sqrt{A_{i-1}} \\ &= \sum_{i\in[k]} \frac{a_i}{\sqrt{A_i} + \sqrt{A_{i-1}}} = \sum_{i\in[k]} \frac{\sqrt{\lambda_i A_i}}{\sqrt{A_i} + \sqrt{A_{i-1}}} \\ &\ge \frac{1}{2}\sum_{i\in[k]} \sqrt{\lambda_i} = \frac{k\sqrt{\lambda}}{2}, \end{aligned}$$

where the second line used (6), and the third line used that the $\{A_k\}_{k\ge1}$ are nondecreasing. □

By combining Lemma 1 with Lemma 2, we obtain the main result of this section.

Theorem 1 (Accelerated proximal point)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, and suppose for $\mathbf{x}_0 \in \mathbb{R}^d$ we have $\|\mathbf{x}_0 - \mathbf{x}^\star\|_2 \le R$ for $\mathbf{x}^\star \in \text{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. Further, let $A_0 = 0$ and $\lambda_k = \lambda > 0$ for all $k \ge 1$. Then iterating (4), (5) for $0 \le k < T$,*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{2R^2}{\lambda T^2}.$$

Proof. By Lemma 1, we have

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) = \epsilon_T \le \frac{\Phi_T}{A_T} \le \frac{\Phi_0}{A_T} \le \frac{R^2}{2A_T}.$$

The conclusion follows from Lemma 2. □

Theorem 1 generically improves upon our analysis of smooth gradient descent (Theorem 3, Part II) by a quadratic factor in the dependence on T, as long as $\lambda \ge \frac{1}{L}$. The catch, of course, is that the iteration (4) is implicit: we cannot exactly compute the proximal point sequence $\{\mathbf{x}_k\}_{k\ge0}$ in general (cf. discussion in Remark 2, Part III). This motivates our development in the following Section 1.3, where for smooth functions f, we show how to match the convergence rate of Theorem 1 (for an appropriate choice of λ) using an explicit first-order method.

1.3 Accelerated gradient descent

In this section we finally show how to fully discretize the accelerated gradient flow (1) to obtain Nesterov's accelerated gradient descent (AGD) algorithm. In fact, we have already seen the main pieces that we need to put together. Roughly speaking, the key observation is that there is slack in the proof of Lemma 2, in the form of a squared gradient norm (see (10)). We use this slack to compensate for the discretization error that occurs when we take a gradient descent step in place of the proximal point iteration used to define (4), using Corollary 2, Part II.

We again initialize AGD with $(A_0, \mathbf{x}_0) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^d$. In this section, we uniformly set $\lambda_k = \frac{1}{L}$ for all $k \geq 1$, where L is the smoothness of f that we wish to minimize. We again initialize our sequences $\{\mathbf{x}_k\}_{k\geq 0}$, $\{\mathbf{y}_k\}_{k\geq 0}$, $\{\mathbf{z}_k\}_{k\geq 0}$ from $\mathbf{z}_0 = \mathbf{y}_0 \leftarrow \mathbf{x}_0$, and evolve them very similarly to (4):

$$\begin{aligned} \mathbf{y}_k &\leftarrow \frac{A_k}{A_{k+1}}\mathbf{x}_k + \frac{a_{k+1}}{A_{k+1}}\mathbf{z}_k, \\ \mathbf{z}_{k+1} &\leftarrow \mathbf{z}_k - a_{k+1}\nabla f(\mathbf{y}_k), \\ \mathbf{x}_{k+1} &\leftarrow \mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k), \end{aligned} \tag{11}$$

for all $k \geq 0$, where $\{A_k, a_k\}_{k\geq 1}$ again follow the recursion (5). That is, compared to (4), (11) is exactly the same except it uses the (explicit) gradients computed at the previous point $\mathbf{y}_k$, rather than the (implicit) gradient computed via the proximal point iterate $\mathbf{x}_{k+1}$.

As we can see, (11) performs a linear coupling of mirror and gradient descent, just as (4) linearly coupled mirror descent and the proximal point method. The intuition provided in [ZO17] is that mirror descent works well when gradients are small (as reflected in the Lipschitz parameter arising in Theorem 2, Part III), and gradient descent works well when gradients are large (as seen in the progress bound in Corollary 2, Part II). Thus, (11) balances the benefits of each method.

We produce the analog of Lemma 1 in the setting of AGD.

Lemma 3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex. Following the notation in (5), (7), (11), and letting $\lambda_k = \frac{1}{L}$ for all $k \geq 1$, we have for all $k \geq 0$ that $\Phi_{k+1} \leq \Phi_k$.*

Proof. Replicating the proof of Lemma 1, we first derive

$$\begin{aligned} a_{k+1}\langle \nabla f(\mathbf{y}_k), \mathbf{x}^\star - \mathbf{z}_k\rangle &= a_{k+1}\langle \nabla f(\mathbf{y}_k), \mathbf{x}^\star - \mathbf{y}_k\rangle + a_{k+1}\langle \nabla f(\mathbf{y}_k), \mathbf{y}_k - \mathbf{z}_k\rangle \\ &= a_{k+1}\langle \nabla f(\mathbf{y}_k), \mathbf{x}^\star - \mathbf{y}_k\rangle + A_k\langle \nabla f(\mathbf{y}_k), \mathbf{x}_k - \mathbf{y}_k\rangle \\ &\leq a_{k+1}\left(f(\mathbf{x}^\star) - f(\mathbf{y}_k)\right) + A_k\left(f(\mathbf{x}_k) - f(\mathbf{y}_k)\right), \end{aligned}$$

where again we used the definition of $\mathbf{y}_k$ and convexity twice. Next, analogously to (8), we have

$$\begin{aligned} a_{k+1}\langle \nabla f(\mathbf{y}_k), \mathbf{z}_k - \mathbf{x}^\star\rangle &\leq r_k - r_{k+1} + \frac{a_{k+1}^2}{2}\left\|\nabla f(\mathbf{y}_k)\right\|_2^2 \\ &\leq r_k - r_{k+1} + a_{k+1}^2 L\left(f(\mathbf{y}_k) - f(\mathbf{x}_{k+1})\right) \\ &= r_k - r_{k+1} + A_{k+1}\left(f(\mathbf{y}_k) - f(\mathbf{x}_{k+1})\right), \end{aligned}$$

where in the second line, we used the progress of smooth gradient descent (Corollary 2, Part II), and in the last line, we used (6) and $\lambda_{k+1} = \frac{1}{L}$. □

We have arrived at our main AGD convergence result, whose proof is identical to Theorem 1.

Theorem 2 (Accelerated gradient descent)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex, and suppose for $\mathbf{x}_0 \in \mathbb{R}^d$ we have $\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|_2 \leq R$ for $\mathbf{x}^\star \in \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. Further, let $A_0 = 0$ and $\lambda_k = \frac{1}{L} > 0$ for all $k \geq 1$. Then iterating (5), (11) for $0 \leq k < T$,*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{2LR^2}{T^2}.$$

Theorem 2 generically improves Theorem 3, Part II, and (via the reduction in Lemma 11, Part II) achieves the tight rate for smooth and well-conditioned convex optimization via gradient queries.

We note that there is a rewriting of the updates (11) commonly seen in derivations of AGD that has a rather intuitive interpretation. In particular, observe that by using (6),

$$\begin{aligned}\mathbf{z}_{k+1} &= \mathbf{z}_k + a_{k+1}L\left(\mathbf{x}_{k+1} - \mathbf{y}_k\right) \\ &= \left(\mathbf{y}_k + \frac{A_k}{a_{k+1}}\left(\mathbf{y}_k - \mathbf{x}_k\right)\right) + \frac{A_{k+1}}{a_{k+1}}\left(\mathbf{x}_{k+1} - \mathbf{y}_k\right) = \mathbf{x}_{k+1} + \frac{A_k}{a_{k+1}}\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right),\end{aligned}$$

so that the update (11) is equivalent to iterating

$$\mathbf{x}_{k+1} \leftarrow \mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k),\ \mathbf{y}_{k+1} \leftarrow \mathbf{x}_{k+1} + \frac{A_k}{A_{k+1}}\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right). \tag{12}$$

The iteration (12) explains why acceleration is also often referred to as "momentum," as it can be more concisely described by two update sequences, one of which is advanced via gradient descent, and the other of which is advanced via a history-dependent difference sequence. We also see how this momentum update reflects a discretization of our accelerated gradient flow ODE (1), i.e.,

$$\mathbf{y}_{k+1} \leftarrow \mathbf{y}_k + \frac{A_k}{A_{k+1}}(\mathbf{x}_{k+1} - \mathbf{x}_k) - \frac{1}{L}\nabla f(\mathbf{y}_k),$$

where the two update terms correspond to a momentum term (i.e., accumulating part of the current velocity), and a gradient term, just as in (1). We remark that we are not aware of a "one-sequence" iteration that achieves the tight accelerated rate, so advancing two different sequences of iterates in the discretization of (1) may be inherent. However, there has been interesting recent work on beating the $\frac{1}{T}$ rate of standard gradient descent for smooth, convex functions using alternative discretization techniques such as choosing a careful sequence of step sizes [Gri24, AP23].

Finally, we briefly describe when acceleration is possible in non-Euclidean settings. The conventional wisdom in this regard is that there are two criteria that must be met: to achieve a $\frac{1}{T^2}$ rate for minimizing convex $f : \mathcal{X} \to \mathbb{R}$, we should require that f is L-smooth with respect to some norm $\|\cdot\|$, and that there exists a "small" regularizer $\varphi : \mathcal{X} \to \mathbb{R}$ that is 1-strongly convex with respect to the same norm. For an upper bound achieving this rate, see Theorem 4.1 of [ZO17]. More formally, by small we mean that if φ has an additive range of Θ over $\mathcal{X}$, then the convergence rate scales as $\frac{L\Theta}{T^2}$. Theorem 2 reflects this, where $\Theta = \frac{1}{2}R^2$ for the regularizer $\varphi = \frac{1}{2}\|\cdot\|_2^2$ over $\mathbb{B}(R)$. This poses an issue in norms for which there provably do not exist strongly convex regularizers with additive ranges growing slowly with the dimension d, e.g., the ℓ_∞ norm (cf. Appendix A.1, [ST18]).

One could hope that weaker conditions suffice for acceleration that bypass this strongly convex additive range issue, e.g., that f is relatively smooth in φ (Definition 2, Part III), which in principle could let us design smaller regularizers φ more directly tailored to the geometry of f. Unfortunately, there are lower bounds precluding acceleration in this setting [DTdB22], highlighting that this phenomenon is brittle when extending to non-Euclidean applications.

2 High-order methods

kjtian: TODO: will complete by next week.

2.1 Stationary points

Source material

Portions of this lecture are based on reference material in [], as well as the author's own experience working in the field.

References

[AP23] Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging I: multi-step descent and the silver stepsize schedule. *CoRR*, abs/2309.07879, 2023.

[BLS15] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to nesterov's accelerated gradient descent. *CoRR*, abs/1506.08187, 2015.

[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, January 6-8, 2021, Virtual Conference*, volume 185 of *LIPIcs*, pages 62:1–62:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[DTdB22] Radu-Alexandru Dragomir, Adrien B. Taylor, Alexandre d'Aspremont, and Jérôme Bolte. Optimal complexity and certification of bregman first-order methods. *Math. Program.*, 194(1):41–83, 2022.

[Gri24] Benjamin Grimmer. Provably faster gradient descent via long steps. *SIAM J. Optim.*, 34(3):2588–2608, 2024.

[Gü92] Osman Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

[Har13] Moritz Hardt. The zen of gradient descent. http://blog.mrtz.org/2013/09/07/the-zen-of-gradient-descent.html, 2013. Accessed: 2025-02-09.

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

[Nem82] Arkadi Nemirovski. Orth-method for smooth convex optimization. *Izvestia AN SSSR, Ser. Tekhnicheskaya Kibernetika*, 2, 1982.

[Nes83] Yurii Nesterov. A method for solving a convex programming problem with convergence rate $o(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.

[SBC16] Weijie Su, Stephen P. Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *J. Mach. Learn. Res.*, 17:153:1–153:43, 2016.

[SMDH13] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1139–1147. JMLR.org, 2013.

[ST18] Aaron Sidford and Kevin Tian. Coordinate methods for accelerating ℓ_∞ regression and faster approximate maximum flow. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018*, pages 922–933. IEEE Computer Society, 2018.

[ZO17] Zeyuan Allen Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 3:1–3:22. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.