Integration methods and Accelerated Optimization Algortihms

Damien Scieur, Vincent Roulet, Francis Bach and Alexandre d'Aspremont

INRIA, Ecole normale superieure, PSL Research University, CNRS

June 6, 2018

Motivation

- Intuition to build convex optimization algorithms sometimes mysterious, e.g. accelerated algorithms
- Continuous time interpretation may help
- Here start from basic gradient flow

$$\dot{x}(t) = -\nabla f(x)$$

- Other interpretations possible through second order derivative equations [Wibosono et al. 2016] but
  - less straightforward
  - not proven to be linked to proper integration methods

Plan

Optimization setting

Problem is

$$\text{minimize} \quad f(x)$$

on variable x where $f \in \mathcal{C}^1(\mathbb{R}^d)$ is

- L-smooth, i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|, \quad \text{for every} \quad x, y \in \mathbb{R}^d$$

- μ-strongly convex, i.e.

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu\|x - y\|^2, \quad \text{for every} \quad x, y \in \mathbb{R}^d$$

Continuous time translation

Study of curves $x(t)$ satisfying Ordinary Differential Equation (ODE)

$$\begin{aligned} x(0) &= x_0 \\ \dot{x}(t) &= g(x(t)) \end{aligned}$$

where

- g comes from a potential $-f$, i.e. $g = -\nabla f(x)$
- g is L-Lipschitz, i.e.

$$\|g(x) - g(y)\|_2 \leq L\|x - y\|, \quad \text{for every} \quad x, y \in \mathbb{R}^d$$

- $-g$ is μ-strongly monotone, i.e.

$$-\langle g(x) - g(y), x - y\rangle \geq \mu\|x - y\|^2, \quad \text{for every} \quad x, y \in \mathbb{R}^d$$

Properties of the gradient flow

- g Lipschitz $\Rightarrow$ existence and uniqueness of $x(t)$
- $-g$ monotone $\Rightarrow$ uniqueness of equilibrium x^*, s.t.
$$g(x^*) = 0 \text{ and } x(\infty) = x^*$$
- Continuous time rates

$$f(x(t)) - f^* \leq (f(x_0) - f^*)e^{-2\mu t}$$
$$\|x(t) - x^*\| \leq \|x_0 - x^*\|e^{-\mu t}$$

Plan

Goal of integration methods

Generally no analytical form for $x(t)$...

Goal of integration methods:

- Approximate curve $x(t)$ on a finite time interval $[0, t_{max}]$
- Done on a time grid t_k by building sequence x_k s.t. $x_k \approx x(t_k)$
- Here regular grid, $t_k = kh$ with h, the stepsize

Euler's explicit scheme

- **Idea:** Use Taylor expansion at time t

$$x(t+h) = x(t) + h\dot{x}(t) + O(h^2).$$

  Neglects second order term, you get Euler's explicit scheme

$$x_{k+1} = x_k + hg(x_k).$$

- For $g(x) = -\nabla f(x)$, corresponds to gradient descent

Euler's implicit scheme

- **Idea:** Use Taylor expansion at time $t + h$

  $$x(t) = x(t + h) - h\dot{x}(t) + O(h^2).$$

  Neglects second order term, you get Euler's implicit scheme

  $$x_{k+1} = x_k + hg(x_{k+1}).$$

- Requires solution of an implicit problem
  $\rightarrow$ costly but potentially more precise
- For $g(x) = -\nabla f(x)$, corresponds to proximal point algorithm

  $$x_{k+1} = \arg\min_z \frac{1}{2}\|z - x_k\|^2 + hf(z)$$

Multistep schemes

- **Idea:** Use s previous points to build next one

$$x_{k+s} = -\sum_{i=0}^{s-1} \rho_i x_{k+i} + h \sum_{i=0}^{s} \sigma_i g(x_{k+i}), \quad \text{for } k \geq 0,$$

- If $\sigma_s = 0$ the method is *explicit* otherwise it is *implicit*
- Compactly defined by s initial points and

$$\rho(E) x_k = h\sigma(E) g_k, \quad \text{for every } k \geq 0,$$

where $E : x_k \to x_{k+1}$ is the shift operator, ρ and σ are polynomials of degree s and $\rho_s = 1$.

Plan

Gradient flow

Integration methods

Proper integration on finite time

Stability in infinite time

Integration view for optimization

Proper integration

- An integration method effectively integrates the ODE on a finite time interval if

$$\lim_{h \to 0} \|x_k - x(t_k)\| = 0 \quad \text{for any } k \in [\![0, t_{max}/h]\!]$$

- Error can be decomposed as

$$\|x_k - x_(t_k)\| \approx \text{error in initial points} + \text{accumulated local error}$$

$\rightarrow$ First term controlled by *zero-stability*
$\rightarrow$ Second term controlled by *consistency*

Zero-stability

- Sensitivity to initial conditions controlled by capacity to produce bounded solutions in the case $g = 0$
- Reduce to study homogeneous differential equation $\rho(E)x_k = 0$

Proposition

A multistep method is *zero-stable* iff

$$\begin{array}{ll} \text{roots}(\rho(z)) & \text{lie in the unit disk} \\ \text{roots}(\rho(z)) & \text{in the unit circle have multiplicity one} \end{array}$$

Consistency

- Define a measure of local error, called truncation error

$$T(h) = \frac{x(t_{k+s}) - x_{k+s}}{h} \qquad \text{assuming } x_{k+i} = x(t_{k+i}),\ i \in [\![0, s-1]\!]$$

- An integration method is said *consistent* if

$$\lim_{h \to 0} \| T(h) \| = 0.$$

Normalization by h because number of errors grows as t_{max}/h

- Looking at Taylor expansion, this simplifies

Proposition

A multistep method is *consistent* iff

$$\rho(1) = 0 \qquad \text{and} \qquad \rho'(1) = \sigma(1)$$

Dahlquist theorem

Dahlquist's theorem

Given a multistep method whose starting values $x_i \to x(t_i)$ for $i \in [\![1, s-1]\!]$, *zero-stability* and *consistency* are necessary and sufficient to ensure on a finite time interval $[0, t_{\max}]$ that $\|x_k - x(t_k)\| \to 0$ for any k when $h \to 0$

Plan

Infinite time horizon

- Proper integration is traditionally studied on *finite* time intervals
- Optimization focuses on infinite time horizon $x(\infty) = x^*$
- Needs condition of stability for infinite time horizon
  - Here study in case fo linear gradient flows (quadratic optimization)
  - Gives necessary condition to integrate smooth strongly convex functions

Absolute stability

- Linear ODE with $\mu I \preceq A \preceq LI$ reads

$$\dot{x}(t) = -Ax(t)$$

so $x(\infty) = 0$

- For fixed A, h, a method is *absolutely stable* if it produces bounded sequences x_k when applied to the linear ODE
- After diagonalization of A, reduces to study homogeneous differential equation

$$(\rho + \lambda h \sigma)(E)x_k = 0$$

where $\lambda \in Sp(A)$

Proposition

Region of absolute stability of a multistep method given by ρ, σ is

$$\{h\lambda : \text{roots}(\rho(z) + \lambda h \sigma(z))\text{lie in the unit disk}\}$$

Convergence rates for linear ODE

- By construction, absolute stability gives also rates of convergence to equilibrium x^* for linear ODE given by $\mu I \preceq A \preceq LI$
- For a multistep method (ρ, σ) and a stepsize h, define

$$r_{\max} = \max_{\lambda \in [\mu, L]} \max\{|r| : r \in \text{roots}(\rho(z) + \lambda h \sigma(z))\}$$

then, if $r_{\max} < 1$, built sequence x_k satisfies

$$\|x_k - x^*\| = O(r_{max}^k)$$

Plan

Gradient flow

Integration methods

Proper integration on finite time

Stability in infinite time

Integration view for optimization

Analysis of multi-step methods

- Analyze one and two step explicit methods through their
  - consistency
  - zero-stability
  - region of absolute stability
  - rate of convergence for linear ODE
- **Intuition:**

The larger h, the faster the algorithm

$$f(x_k) - f^* \approx f(x(t_k)) - f^* \leq e^{-2\mu k h}(f(x_0) - f^*)$$

One-step explicit method

- Euler's explicit scheme

$$x_{k+1} = x_k + hg(x_k).$$

- Zero-stable ✓
- Consistent ✓
- Optimal step-size for convergence on linear ODE

$$h = \frac{2}{L + \mu}$$

and corresponding rate

$$\|x_k - x^*\| = O\left(\left(\frac{1 - \mu/L}{1 + \mu/L}\right)^k\right)$$

Two steps methods

- Complete analysis gives a family of two step methods parametrized by one parameter
- Polyak, 1964 heavy ball method and Nesterov, 1983 accelerated method, seen as integration methods, belong to this class
- Polyak's method is optimal among this class (bigger step size and better convergence rates)
- But Polyak do not optimize general smooth strongly convex functions [Lessard et al. 2016]

Stepsize fo acceleratd method

- Nesterov, 1983 accelerated gradient reads

$$y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k),$$
$$x_{k+1} = y_{k+1} + \beta(y_{k+1} - y_k).$$

i.e.

$$\beta x_k - (1+\beta) x_{k+1} + x_{k+2} = \frac{1}{L}\left(-\beta(-\nabla f(x_k)) + (1+\beta)(-\nabla f(x_{k+1}))\right)$$

- Zero-stable ✓
- Consistency conditions ($\rho(1) = 0$, $\rho'(1) = \sigma(1)$) give

$$h = \frac{1}{L(1-\beta)}$$

Acceleration explanation

- Gradient descent step size $h = 1/L$
- Gradient descent approximative rate

$$f(x_k^{\text{grad}}) - f(x^*) \approx f(x(k/L)) - f(x^*) \leq (f(x_0) - f(x^*))e^{-2k\frac{\mu}{L}}$$

- Nesterov's stepsize

$$h_{\text{nest}} = \frac{1}{L(1-\beta)} = \frac{1+\sqrt{\mu/L}}{2\sqrt{\mu L}} \approx \frac{1}{\sqrt{4\mu L}}$$

- Nesterov's approximative rate

$$f(x_k^{\text{nest}}) - f(x^*) \approx f(x(k/\sqrt{4\mu L})) - f(x^*) \leq (f(x_0) - f(x^*))e^{-k\sqrt{\mu/L}}$$

Illustration

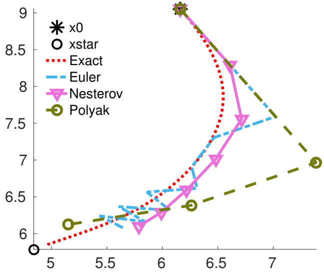

Figure: Integration of a linear ODE with optimal (left) and small (right) step sizes.

Conclusion

- Accelerated optimization methods can be seen as multistep integration schemes applied to the basic gradient flow equation
- Natural interpretation of acceleration:

  Larger steps speed up convergence

- Further links btw integration methods and other well-known optimization algorithms:
  - proximal gradient descent,
  - mirror gradient decent,
  - extra-gradient algorithm
  - ...

Future work

- Analyze smooth and strongly convex case (not only quadratics)
- Extend to weakly convex case