Integration Methods and Optimization Algorithms

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

SIERRA Team - École Normale Supérieure/CNRS/INRIA/PSL Research University, 75005, Paris, France

Overview

Interpret convex optimization algorithms as integration methods of the gradient flow equation

$$\dot{x}(t) = -\nabla f(x(t)), \quad x(0) = x_0$$

for f L-**smooth**, μ-**strongly convex**.

Shows that acceleration methods can be seen as integration methods with big step-sizes.

Continuous time solution

Gradient Flow is an Ordinary Differential Equation ODE

$$\dot{x}(t) = g(x(t)), \quad x(0) = x_0$$

where g is

- Lipschitz $\Rightarrow$ Existence and uniqueness of solution $x(t)$
- monotone $\Rightarrow$ Existence and uniqueness of equilibrium x^* s.t. $g(x^*) = 0$ and $x(\infty) = x^*$

Convergence of $x(t)$ to x^* can be measured as

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

Integration methods

Approximate curve $x(t)$ on $[0, t_{\max}]$ by sequence x_k s.t.

$$x_k \approx x(kh) \quad \text{for } k \in \{0, \ldots, t_{\max}/h\}$$

where h is the step size of integration.

Euler's method

First-order approximation of $x(t+h)$ around t,

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

From $x_k \approx x(kh)$, $x_{k+1} \approx x(kh+h)$ is built as

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

For $g(x) = -\nabla f(x)$, one recognizes **gradient descent**.

Linear Multistep Methods (LMM)

Linear (explicit) s-Multistep Methods generate next point from s previous ones as

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

where $\rho_i, \sigma_i \in \mathbb{R}$ and $x_0, \ldots, x_{s-1}$ are initialized beforehand.

Using $E : x_k \to x_{k+1}$, $g_k = g(x_k)$, an LMM reads

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

where ρ, σ are polynomials (with $\rho_s = 1$).
An integration method *must satisfy*

- **Zero Stability:** Not sensitive to initializations
- **Consistency:** Local error decreases with step size s.t.

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

Stability

Optimization focuses on *infinite time* horizon.
For integration methods, this requires **Stability**, i.e.

If $x(t)$ bounded, then x_k must be bounded.
$\rightarrow$ depends on g and step-size h.

Complex in the general case, but reduces to simple considerations for linear ODE

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

There the LMM builds a sequence satisfying

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

Stability of an LMM is then ensured if

$$|\text{roots}(\rho + h\lambda\sigma)| < 1$$

Convergence to equilibrium is then measured by

$$\|x_k - x^*\| \leq O\Big(\max(|\text{roots}(\rho + h\lambda\sigma)|)^k\Big)$$

Analysis of LMM for linear ODE

For linear ODE (quadratic optimization),

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

with $\text{Sp}(A) = [\mu, L]$, long-term behavior of LMM of order s with step-size h controlled by

$$\max_{\lambda \in [\mu, L]} |\text{roots}(\rho + h\lambda\sigma)|$$

where $|\text{roots}(\rho)| \leq 1$ (Zero Stability)
$\rho(1) = 1, \ \rho'(1) = \sigma(1)$ (Consistency)

Optimal methods in ρ, σ, h are

- when $s = 1$: Gradient Descend (with optimal step size)
- when $s = 2$: Polyak's Heavy Ball method
- when $s = 2$ for given h: Nesterov's Fast Gradient

Acceleration interpretation

For step-size h and integration method s.t. $x_k \approx x(kh)$,

$$\|x_k - x^*\| \approx \|x(kh) - x^*\| \leq e^{-\mu kh}\|x_0 - x^*\|$$

The bigger the step-size, the faster the convergence

As an integration method, Nesterov's fast gradient method has step-size

$$h_{\text{Nest}} \approx \frac{1}{L}\sqrt{\frac{L}{\mu}} = \sqrt{\frac{L}{\mu}} h_{\text{grad}}$$

$\Rightarrow$ Nesterov's gradient is a stable integration method with bigger step size

Generalizations

Composite, non-Euclidean settings can be cast by corresponding gradient flows and adequate LMM

In the convex case, Nesterov's method shows also bigger step size of integration