A LYAPUNOV ANALYSIS OF ACCELERATED METHODS IN OPTIMIZATION

Paper Representation

Supervisor: Dr. Samir Adly
Student: CHAU Dang Minh

University of Limoges

Outline

Introduction

Scenerio

$$\min_{x \in \mathcal{X}} f(x), \tag{1}$$

where

- $\mathcal{X} \in \mathbb{R}^d$ is a closed convex set,
- $f : \mathcal{X} \to \mathbb{R}$ is a continuously differentiable convex.

We use the standard Euclidean norm $\|x\| = \frac{1}{2}\langle x, x\rangle$. function

Discreteness and Continuality

- Early iterative optimization algorithms (Gradient Descent and Polyak's Momentum Acceleration) are intuitively interpretable.
- Nesterov's Acceleration is less intuitive

$$\begin{cases} x_k = y_{k-1} - s\nabla f(y_{k-1}) \\ y_k = x_k + \dfrac{k-1}{k+2}(x_k - x_{k-1}). \end{cases} \tag{2}$$

- Continualized versions as ODEs are available.
- Current orientation: starting from an ODE and derive a family of discrete algorithms using Euler's methods.

Lagrangian mechanics and the Lagrangian

The Lagrangian $\mathcal{L}(X, V, t)$ is introduced [1] as a framework to derive ODEs, where

- $X = X(t)$ is the coordinate
- $V = \dot{X}(t)$ is the velocity
- $t \in \mathbb{R}$ is the time

The action in $[t_1, t_2]$ is $\mathcal{A}(X) = \int_{t_1}^{t_2} \mathcal{L}(X, V, t)\,\mathrm{d}t$. A trajectory X being a stationary function of $\mathcal{A}$ solves the Euler-Lagrange equation

$$\frac{\partial \mathcal{L}}{\partial X}\mathcal{L}(X, V, t) = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial V}\mathcal{L}(X, V, t). \tag{3}$$

[1] Wibisono, Andre, Ashia C. Wilson, and Michael I. Jordan. "A variational perspective on accelerated methods in optimization." proceedings of the National Academy of Sciences 113.47 (2016): E7351-E7358.

NAG's Lagrangian

The ODE associated to NAG [2]

$$\ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X) = 0 \tag{4}$$

has corresponding Lagrangian

$$\mathcal{L}(X, V, t) = t^3 \left(\frac{1}{2}\|V\|^2 - f(X) \right). \tag{5}$$

Indeed, $\frac{\partial \mathcal{L}}{\partial X}\mathcal{L} = -t^3 \nabla f(X)$, $\frac{\partial \mathcal{L}}{\partial V}\mathcal{L} = t^3 V$ and $\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial V}\mathcal{L} = 3t^2 V + t^3 \dot{V}$.
Hence $t^3 \dot{X} = 3t^2 \dot{X} + t^3 \ddot{X}$. Divide by t^3 and rearrange to get (4).

[2] Su, Weijie, Stephen Boyd, and Emmanuel J. Candes. "A differential equation for modeling Nesterov's accelerated gradient method: theory and insights." arXiv preprint arXiv:1503.01243 (2015).

Lagrangian

- We can also see that the standard Lagrangian

$$\mathcal{L}(X, V, t) = \frac{1}{2}\|V\|^2 - f(X) \tag{6}$$

  derives Polyak's acceleration with momentum $\beta = 1$.
- In (6), $\frac{1}{2}\|V\|^2$ is the kinetic energy and $f(X)$ is the potential energy.
- Idea: generalizing this difference, proving convergence of the derived ODE, then discretizing and proving the convergence of iterative algorithms.

Bregman Lagrangians

Bregman Lagrangians

- The kinetic energy is generalized by the Bregman divergence

$$D_y(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle, \tag{7}$$

where $h : \mathcal{X} \to \mathbb{R}$ is convex and smooth.

- Rescaling factors are added.
- The first Bregman Lagrangian is defined [3] by

$$\mathcal{L}(X, V, t) = e^{\alpha_t + \gamma_t}(D_h(X + e^{-\alpha_t} V, X) - e^{\beta_t} f(X)). \tag{8}$$

- When $h(x) = \frac{1}{2}\|x\|^2$ and $Y = X + e^{-\alpha_t} V$ is near X, we recover a scaled kinetic energy

$$e^{\alpha_t + \gamma_t} D_h(X + e^{-\alpha_t} V, X) \approx e^{\alpha_t + \gamma_t} \frac{1}{2}\|Y - X\|^2 = e^{\gamma_t - \alpha_t} \frac{1}{2}\|V\|^2.$$

[3] Wibisono, Andre, Ashia C. Wilson, and Michael I. Jordan. "A variational perspective on accelerated methods in optimization." proceedings of the National Academy of Sciences 113.47 (2016): E7351-E7358.

Bregman Lagrangians

Under ideal rescaling conditions

$$\dot{\gamma}_t = e^{\alpha_t} \text{ and } \dot{\beta}^t \leq e^{\alpha_t}, \tag{9}$$

the first Bregman Lagrangian reduced to the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} \nabla h(X + e^{-\alpha_t} V) = -e^{\alpha_t + \beta_t} \nabla f(X). \tag{10}$$

Using a Lyapunov function, it is proven that for some $x^* \in \operatorname*{argmin}_{x \in \mathcal{X}} f(x)$,

$$f(X(t)) - f(x^*) \leq O(e^{-\beta_t}). \tag{11}$$

Bregman Lagrangians

The second Bregman Lagrangian is introduced [4] as

$$\mathcal{L}(X, V, t) = e^{\alpha_t + \beta_t + \gamma_t}(\mu D_h(X + e^{-\alpha_t} V, X) - f(X)). \tag{12}$$

The derived ODE under rescaling conditions (9) is more general than that of the first Bregman Lagrangian.

$$\frac{\mathrm{d}}{\mathrm{d}t} \nabla h(X + e^{-\alpha_t} V) = \dot{\beta}_t \nabla h(X) - \dot{\beta}_t \nabla h(X + e^{-\alpha_t} V) - \frac{e^{\alpha_t}}{\mu} \nabla f(X). \tag{13}$$

When $h(x) = \frac{1}{2}\|x\|^2$ and $\beta_t = \sqrt{\mu} t$, (13) is reduced to Polyak's momentum

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X) = 0.$$

[4] Wilson, Ashia C., Ben Recht, and Michael I. Jordan. "A Lyapunov analysis of accelerated methods in optimization." Journal of Machine Learning Research 22.113 (2021): 1-34.

Bregman Lagrangians

Using the Lyapunov function

$$\mathcal{E}_t = e^{\beta_t}\left(\mu D_h(X + e^{-\alpha_t}V, X) + f(X) - f(x)\right) \tag{14}$$

with $x = x^*$, the inequality $\mathcal{E}_t \leq \mathcal{E}_0$ leads to the same convergence rate as that of the first Bregman Lagrangian

$$f(X(t)) - f(x^*) \leq O(e^{-\beta_t}).$$

Discretization

Discretization

To this point, we have two ODEs. Recalling (10)

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla h(X + e^{-\alpha_t}V) = -e^{\alpha_t+\beta_t}\nabla f(X).$$

Using $\alpha_t = \log p - \log t, \beta_t = p \log t + \log C$ and $\gamma_t = p \log t$ for some $p > 0$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla h(X + \frac{t}{p}V) = -Cpt^{p-1}\nabla f(X).$$

Or

$$\begin{cases} Z = X + \frac{t}{p}\dot{X} \\ \frac{\mathrm{d}}{\mathrm{d}t}\nabla h(Z) = -Cpt^{p-1}\nabla f(X) \end{cases} \tag{15}$$

Let $t = \delta k$, now we discretize $x_k = X(t)$, $x_{k+1} = X(t+\delta) \approx X(t) + \delta \dot{X}_t$ and similarly for $Z(t)$.

Discretization

The first equation becomes

$$x_{k+1} = \frac{p}{k} z_k + \frac{k-p}{k} x_k$$

The second equations becomes

$$\nabla h(z_k) - \nabla h(z_{k-1}) = -Cp\delta^p k^{p-1} \nabla f(x_k).$$

Equivalently,

$$\nabla_z (Cpk^{p-1} \langle \nabla f(x_k), z \rangle + \frac{1}{\delta^p} D_h(z, z_{k-1})) = 0.$$

Proving that the function taken gradient is convex, we can update

$$z_k = \operatorname*{argmin}_z \left\{ Cpk^{p-1} \langle \nabla f(x_k), z \rangle + \frac{1}{\delta^p} D_h(z, z_{k-1}) \right\}.$$

Unfortunately, it is proven that this algorithm is not stable.

Discretization

Using the combination instead of the exponent

$$\begin{cases} x_{k+1} = \frac{p}{k} z_k + \frac{k-p}{k} x_k \\ z_k = \underset{z}{\operatorname{argmin}} \left\{ Cp \binom{p+k-2}{p} \langle \nabla f(x_k), z \rangle + \frac{1}{\delta^p} D_h(z, z_{k-1}) \right\}, \end{cases} \tag{16}$$

the algorithm is proven to converge with rate $O(1/(\delta k)^p)$.

The same implicit method is applied for ODE (13) and convergence is also guaranteed. However, solving for z_k is as difficult as the original problem. Hence we consider cases where explicit discretization arrives at reasonable convergence rate.

Discretization

Use an extrapolating sequence (y_k), there are two possible updates. For example with ODE (10)

$$\begin{cases} x_{k+1} = \beta_k z_k + (1-\beta_k) y_k \\ \nabla h(z_{k+1}) = \nabla h(z_k) - \delta\alpha_k \nabla f(y_{k+1}), \end{cases} \tag{17}$$

$$\begin{cases} x_{k+1} = \beta_k z_k + (1-\beta_k) y_k \\ \nabla h(z_{k+1}) = \nabla h(z_k) - \delta\alpha_k \nabla f(x_{k+1}). \end{cases} \tag{18}$$

Using Lyapunov's method with appropriate conditions, convergence rate $O(1/(\delta k)^p)$ is guaranteed.

Derivation

Derivations

Difference choices of (y_k) reveal published algorithms. For example

- Acceleration of gradient descent (plugged to (17)) [5]

$$y_{k+1} = \underset{y}{\operatorname{argmin}} \left\{ f(x_{k+1}) + \langle \nabla f(x_{k+1}), y - x_{k+1} + \frac{1}{2\nu} \|y - x_{k+1}\|^2 \rangle \right\}.$$

- Acceleration of tensor methods [6]

$$y_{k+1} = \underset{y}{\operatorname{argmin}} \left\{ \sum_{t=0}^{p-1} \frac{1}{i!} \nabla^i f(x)(y - x)^i + \frac{1}{p\nu} \|x - y\|^p \right\}.$$

[5]Nesterov, Yurii. Introductory lectures on convex optimization: A basic course. Vol. 87. Springer Science & Business Media, 2013.
[6]Nesterov, Yu. "Accelerating the cubic regularization of Newton's method on convex problems." Mathematical Programming 112.1 (2008): 159-181.

Conclusion and Future Work

- The concerned paper revisits Lagrangians, with an aim of unifying some know algorithms and providing a framework for algorithm design.
- A new Lagrangian is introduced.
- It may require the elaboratively derivations to specific algorithms.

Thank you for listening !