A LYAPUNOV ANALYSIS OF ACCELERATED METHODS IN OPTIMIZATION

Paper Report
CHAU Dang Minh

1 Introduction

Convex optimization plays a central role in optimization, which contributes to the development of machine learning because of typical properties of convex functions. One form of the problem is

$$\inf_{x\in\mathcal{X}} f(x), \tag{$\mathcal{P}$}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a closed and convex, $f : \mathcal{H} \to \mathbb{R}$ is a continuously differentiable convex function. Iterative algorithms have been developed to solve this problems. They begin with an initial guess of the variable x and generate a sequence of improved estimates until they terminate, hopefully at a solution. The strategy used to move from one iterate to the next distinguishes one algorithm from another [3]. For many iterative algorithms, by continualization i.e. letting the step size tend to zero and taking appropriate limits, we arrive at an ODE. Convergent analysis of such ODEs provides novel insights to the original algorithm. For example, the Polyak's momentum algorithm takes the form

$$\begin{cases} x_0, x_1 \\ x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k-1}), k \geq 1. \end{cases} \tag{1}$$

where $\alpha, \beta > 0$. Set $h = \sqrt{\alpha}$, $t = kh$ and $\beta = 1 - \gamma h$, with $\gamma > 0$. We approximate $X(t) \approx x_k$. The update formula becomes

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} + \gamma\frac{x_k - x_{k-1}}{h} + \nabla f(x_k) = 0$$

Taking the limit when $h \downarrow 0$, we have

$$\begin{cases} X(0) = x_0, \dot{X}(0) = x_1 \\ \ddot{X}(t) + \gamma\dot{X}(t) + \nabla f(X(t)) = 0, \ \ t \geq 0. \end{cases} \tag{2}$$

The Lyapunov function argument is usually applied for convergence analysis of an ODE. Let $f^* = \inf_{\mathcal{X}} f(x)$ and $x^* = \operatorname{argmin}_{\mathcal{X}} f(x)$ and define the Lyapunov function

$$\mathcal{E}(t) = \frac{1}{2}\|\dot{X}\|^2 + f(X). \tag{3}$$

It is shown that the dynamic leads to convergence rate $O(1/t)$ [1].

2 Bregman Lagrangians

The Lagrangian is the core quantity in Lagrangian mechanics, which normally takes the form

$$\mathcal{L}(X, V, t) = \text{KE} - \text{PE}, \tag{4}$$

where

- X is the coordinate,

- V is the velocity,
- KE is the kinetic energy,
- PE is the potential energy.

Solving the Lagrange-Euler equation

$$\frac{\partial \mathcal{L}}{\partial X}\mathcal{L}(X,V,t) = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial V}\mathcal{L}(X,V,t). \tag{5}$$

yields the motion of the system [2]. In fact, a solution to (5) is a stationary point of the functional of action

$$\mathcal{A}(\mathcal{X}) = \int_{t_1}^{t_2} \mathcal{L}(X,V,t)\,\mathrm{d}t. \tag{6}$$

In a simple case where we let $\mathrm{KE} = \frac{1}{2}\|V\|^2$ and $\mathrm{PE} = f(X)$. The Lagrange-Euler equation becomes

$$-\nabla f(X) = \dot{V} = \ddot{X},$$

which can be discretized as $-\nabla f(x_k) = \dfrac{x_{k+1} - 2x_k + x_{k-1}}{\alpha}$ or

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + (x_k - x_{k-1}). \tag{7}$$

This is Polyak's momentum with $\beta = 1$. Two generalized Lagrangians are introduced [4] [5]

$$\mathcal{L}_1(X,V,t) = e^{\alpha_t + \gamma_t}(D_h(X + e^{-\alpha_t}V, X) - e^{\beta_t} f(X)), \tag{8}$$

$$\mathcal{L}_2(X,V,t) = e^{\alpha_t + \beta_t + \gamma_t}(\mu D_h(X + e^{-\alpha_t}V, X) - f(X)), \tag{9}$$

where

- $h : \mathcal{X} \to \mathbb{R}$ is a convex smooth function,
- $D_h(y,x) = h(y) - h(x) - \langle \nabla h(x), y - x\rangle$ is the Bregman divergence.

Additional ideal scaling conditions are applied

$$\dot{\gamma}_t = e^{\alpha_t}, \tag{10a}$$
$$\dot{\beta}^t \leq e^{\alpha_t}. \tag{10b}$$

Proposition 1. *Under condition (10a), the Bregman-Euler equation of $\mathcal{L}_1$ reduces to the ODE*

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla h(X + e^{-\alpha_t}V) = -e^{\alpha_t + \beta_t}\nabla f(X). \tag{11}$$

Proof. We have $D_h(X + e^{-\alpha_t}V, X) = h(X + e^{-\alpha_t}V) - h(X) - \langle \nabla h(X), e^{-\alpha_t}V\rangle$. Hence

$$\begin{aligned}\frac{\partial \mathcal{L}_1}{\partial X} &= e^{\alpha_t + \gamma_t}\left(\nabla h(X + e^{-\alpha_t}V) - \nabla h(X) - e^{-\alpha_t}\nabla^2 h(X)V - e^{\beta_t}\nabla f(X)\right) \\ &= e^{\alpha_t + \gamma_t}\left(\nabla h(X + e^{-\alpha_t}V) - \nabla h(X)\right) - e^{\gamma_t}\nabla^2 h(X)V - e^{\alpha_t + \beta_t + \gamma_t}\nabla f(X).\end{aligned}$$

$$\frac{\partial \mathcal{L}_1}{\partial V} = e^{\alpha_t + \gamma_t}\left(e^{-\alpha_t}\nabla h(X + e^{-\alpha_t}V) - e^{-\alpha_t}\nabla h(X)\right) = e^{\gamma_t}\left(\nabla h(X + e^{-\alpha_t}V) - \nabla h(X)\right),$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}_1}{\partial V} &= e^{\gamma_t}\left[\dot{\gamma}_t\left(\nabla h(X+e^{-\alpha_t}V)-\nabla h(X)\right)+\nabla^2 h(X+e^{-\alpha_t}V)\left(V+e^{\alpha_t}(-\dot{\alpha}_t V+\dot{V})\right)-\nabla^2 h(X)V\right]\\ &\overset{(10a)}{=} e^{\alpha_t+\gamma_t}\left(\nabla h(X+e^{-\alpha_t}V)-\nabla h(X)\right)-e^{\gamma_t}\left(\nabla^2 h(X+e^{-\alpha_t}V)\left(V+e^{\alpha_t}(-\dot{\alpha}_t V+\dot{V})\right)+\nabla^2 h(X)V\right)\end{aligned}$$

Hence the Bregman-Lagrange equation becomes

$$e^{\alpha_t+\beta_t+\gamma_t}\nabla f(X) = e^{\gamma_t}\nabla^2 h(X+e^{-\alpha_t}V)\left(V+e^{\alpha_t}(-\dot{\alpha}_t V+\dot{V})\right).$$

This is equivalent to (11). □

Proposition 2. *Under condition (10a), the Bregman-Euler equation of $\mathcal{L}_1$ reduces to the ODE*

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla h(X+e^{-\alpha_t}V) = \dot{\beta}_t\nabla h(X)-\dot{\beta}_t\nabla h(X+e^{-\alpha_t}V)-\frac{e^{\alpha_t}}{\mu}\nabla f(X). \tag{12}$$

Proof. Elaborate on the two sides as in Proposition 1. □

Theorem 1. *Under condition (10b), solutions to (11) satisfy*

$$f(X(t))-f(x^*)\le O\left(e^{-\beta_t}\right). \tag{13}$$

Proof. Consider the Lyapunov function

$$\mathcal{E}_t = D_h(x^*, X+e^{-\alpha_t}\dot{X})+e^{\beta_t}\left(f(X)-f(x^*)\right)\ge 0.$$

We have

$$\begin{aligned}\dot{\mathcal{E}}_t &= \left\langle -\frac{\mathrm{d}}{\mathrm{d}t}\nabla h(X+e^{-\alpha_t}V), x^*-X-e^{-\alpha_t}\dot{X}\right\rangle + e^{\beta_t}\left(\dot{\beta}_t(f(X)-f(x^*))+\langle\nabla f(X),\dot{X}\rangle\right)\\ &= e^{\alpha_t+\beta_t}\left\langle\nabla f(X), x^*-X-e^{-\alpha_t}\dot{X}\right\rangle + e^{\beta_t}\left(\dot{\beta}_t(f(X(t))-f(x^*))+\langle\nabla f(X),\dot{X}\rangle\right)\\ &= -e^{\alpha_t+\beta_t}D_f(x^*,X_t)+(\dot{\beta}_t-\alpha_t)(f(X)-f(x^*))\\ &\overset{(10b)}{\le} -e^{\alpha_t+\beta_t}D_f(x^*,X_t)\le 0.\end{aligned}$$

Hence $e^{\beta_t}(f(X)-f(x^*))=\mathcal{E}_t\le\mathcal{E}_0$ implying (13). □

Theorem 2. *Assume that f is μ-uniformly convex with respect to h, i.e.*

$$D_f(x,y)\ge\mu D_h(x,y), \tag{14}$$

solutions to ODE (12) also satisfy

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

Proof. Elaborate on the Lyapunov function

$$\mathcal{E}_t = e^{\beta_t}\left(\mu D_h(X+e^{-\alpha_t}V, X)+f(X)-f(x)\right). \tag{15}$$

□

3 Discretization

Recall that given the dynamic $\dot{X} = v(X)$,

- The explicit Euler method uses the update formula $x_{k+1} = v(x_k) + sx_k$.
- The implicit Euler method uses the update formula $x_{k+1} = v(x_{k+1}) + sx_k$.

Firstly, rewrite ODE (11) as

$$\begin{cases} Z = X + \dfrac{e^{\beta_t}}{\frac{\mathrm{d}}{\mathrm{d}t}e^{\beta_t}}\dot{X} \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\nabla h(Z) = -\frac{\mathrm{d}}{\mathrm{d}t}e^{\beta_t}\nabla f(X). \end{cases} \tag{16}$$

The family of ODEs where $\alpha_t = \log p - \log t, \beta_t = p\log t + \log C$ and $\gamma_t = p\log t$ is concerned [4], which reduces the system of ODEs to

$$\begin{cases} Z = X + \dfrac{t}{p}\dot{X} \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\nabla h(Z) = -Cpt^{p-1}\nabla f(X) \end{cases} \tag{17}$$

It is shown that the approximation $t = \delta k$, $x_k = X(t)$, $x_{k+1} = X(t+\delta) \approx X(t) + \delta\dot{X}_t$, $z_k = Z(t)$, $z_{k+1} \approx Z(t) + \delta\dot{Z}_t$ does not lead to a stable algorithm [4]. Instead, let $k^{(p)} = k(k+1)\dots(k+p-1)$ and approximate $e^{\beta_t} \approx A_k = C\delta^p k^{(p)}$, $\frac{\mathrm{d}}{\mathrm{d}t}e^{\beta_t} \approx \dfrac{A_{k+1} - A_k}{\delta} = Cp\delta^{p-1}k^{(p-1)}$. Also denote $\tau_k = \dfrac{\alpha_k}{A_k}$, we have the iterative algorithm

$$\begin{cases} z_{k+1} = \underset{z}{\operatorname{argmin}}\left\{A_k f\left(\dfrac{\delta\tau_k}{1+\delta\tau_k}z + \dfrac{1}{1+\delta\tau_k}x_k\right) + \dfrac{1}{\delta\tau_k}D_h(z, z_k)\right\} \\ x_{k+1} = \dfrac{\delta\tau_k}{1+\delta\tau_k}z_{k+1} + \dfrac{1}{1+\delta\tau_k}x_k \end{cases} \tag{18}$$

Proposition 3. *The algorithm (18) yields*

$$f(x_k) - f(x^*) \le O\left(1/A_k\right). \tag{19}$$

It is worth noting that the update of (z_k) is as difficult to carry as solving the original problem. Therefore, the explicit method is implied with an extrapolating sequence (y_k) and under some conditions to guarantee the same convergence rate as that of the implicit method. In particular, consider two possible explicit algorithm

$$\begin{cases} x_{k+1} = \delta\tau_k z_k + (1-\delta\tau_k)y_k \\ \nabla h(z_{k+1}) = \nabla h(z_k) - \delta\alpha_k\nabla f(y_{k+1}), \end{cases} \tag{20}$$

$$\begin{cases} x_{k+1} = \delta\tau_k z_k + (1-\delta\tau_k)y_k \\ \nabla h(z_{k+1}) = \nabla h(z_k) - \delta\alpha_k\nabla f(x_{k+1}). \end{cases} \tag{21}$$

Proposition 4. *Assume that h is σ-uniformly convex with respect to $\|\cdot\|^p$, both algorithms (20) and (21) has*

$$f(x_k) - f(x^*) \le O\left(1/A_k\right).$$

Different applications of (y_k) reveal published algorithms [5]. The second Bregman Lagrangian (9) is also studied. We give a summary in Table 1 and 2.

4 Conclusion

Generalized Lagrangians serve as a new framework for developing novel optimization algorithms. Further studies can focus on the choice of generalization for the energies and scaling factors, or the choice of the parameter in existing Bregman Lagrangians.

Bregman Lagrangian	$\mathcal{L}_1(X,V,t) = e^{\alpha_t+\gamma_t}(D_h(X+e^{-\alpha_t}V,X) - e^{\beta_t}f(X))$
Dynamic	$\frac{\mathrm{d}}{\mathrm{d}t}\nabla h(X+e^{-\alpha_t}V) = -e^{\alpha_t+\beta_t}\nabla f(X)$
Family	$\beta_t = p\log t + \log C$
Lyapunov function	$\mathcal{E}_t = D_h(x^*, X+e^{-\alpha_t}\dot{X}) + e^{\beta_t}\left(f(X)-f(x^*)\right)$
Continuous convergence rate	$O\left(e^{-\beta_t}\right)$
Implicit discretization	$\begin{cases} z_{k+1} = \underset{z}{\operatorname{argmin}}\left\{A_k f\left(\frac{\delta\tau_k}{1+\delta\tau_k}z + \frac{1}{1+\delta\tau_k}x_k\right) + \frac{1}{\delta\tau_k}D_h(z,z_k)\right\} \\ x_{k+1} = \frac{\delta\tau_k}{1+\delta\tau_k}z_{k+1} + \frac{1}{1+\delta\tau_k}x_k \end{cases}$ $(\delta>0, A_k = C\delta^p k^{(p)}, \tau_k = \frac{\alpha_k}{A_k})$
Implicit convergence rate	$O\left(1/A_k\right)$
Explicit discretization	$\begin{cases} x_{k+1} = \delta\tau_k z_k + (1-\delta\tau_k)y_k \\ \nabla h(z_{k+1}) = \nabla h(z_k) - \delta\alpha_k\nabla f(y_{k+1}) \end{cases}$ or $\begin{cases} x_{k+1} = \delta\tau_k z_k + (1-\delta\tau_k)y_k \\ \nabla h(z_{k+1}) = \nabla h(z_k) - \delta\alpha_k\nabla f(x_{k+1}) \end{cases}$ (h is σ-uniformly convex with respect to $\|\cdot\|^p$ and $0 < C < 1/\sigma p^p$)
Explicit convergence rate	$O\left(1/A_k\right)$

Table 1: Summary on the first Bregman Lagrangian derivations

References

[1] Felipe Alvarez. "On the minimizing property of a second order dissipative system in Hilbert spaces". In: *SIAM Journal on Control and Optimization* 38.4 (2000), pp. 1102–1119.

[2] Anatolii Isakovich Lurie. *Analytical mechanics*. Springer Science & Business Media, 2013.

[3] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

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

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

Bregman Lagrangian	$\mathcal{L}_2(X,V,t) = e^{\alpha_t+\beta_t+\gamma_t}(\mu D_h(X+e^{-\alpha_t}V,X) - f(X))$
Dynamic	$\frac{\mathrm{d}}{\mathrm{d}t}\nabla h(X+e^{-\alpha_t}V) = \dot{\beta}_t\nabla h(X) - \dot{\beta}_t\nabla h(X+e^{-\alpha_t}V) - \frac{e^{\alpha_t}}{\mu}\nabla f(X)$
Family	$\beta_t = \sqrt{\mu}t$
Lyapunov function	$\mathcal{E}_t = e^{\beta_t}\left(\mu D_h(X+e^{-\alpha_t}V,X) + f(X) - f(x)\right)$
Implicit discretization	$\begin{cases} z_{k+1} = \underset{z}{\operatorname{argmin}}\left\{f(x) + \mu D_h(z,x) + \frac{\mu}{\delta\tau_k}D_h(z,z_k)\right\} \\ x_{k+1} = \frac{\delta\tau_k}{1+\delta\tau_k}z_{k+1} + \frac{1}{1+\delta\tau_k}x_k \end{cases}$ $(x = \frac{\delta\tau_k}{1+\delta\tau_k}z + \frac{1}{1+\delta\tau_k}x_k, \delta > 0, A_k = (1+\sqrt{\mu}\delta)^k, \tau_k = \frac{\alpha_k}{A_k})$
Implicit convergence rate	$O\left(1/A_k\right)$
Explicit discretization	$\begin{cases} x_k = \frac{\delta\tau_k}{1+\delta\tau_k}z_k + \frac{1}{1+\delta\tau_k}y_k \\ \nabla h(z_{k+1}) - \nabla h(z_k) = \delta\tau_k\left(\nabla h(x_k) - \nabla h(z_k) - \frac{1}{\mu}\nabla f(x_k)\right) \end{cases}$
Explicit convergence rate	$O\left(e^{-\mu k}\right)$

Table 2: Summary on the second Bregman Lagrangian derivations