On the stability of optimization algorithms given by discretizations of the Euler-Lagrange ODE

Rachel Walker
Central Washington University

Emily Zhang
Massachusetts Institute of Technology

Problem Setting

We consider the optimization problem

$$x^* = \arg\min_{x \in \mathbb{R}^d} f(x), \tag{1}$$

where $f(x) = \frac{1}{2}(x - x^*)^T A(x - x^*)$ is a convex function $f : \mathbb{R}^d \to \mathbb{R}$ with some unique minimizer $x^* \in \mathbb{R}^d$ that satisfies the optimality condition $\nabla f(x^*) = Ax^* = \vec{0}$, and A is a positive definite, symmetric $d \times d$ matrix.

Convergence and Discretization of Euler-Langrange ODE

Theorem 2.1 from [1]. Let p and C be constants such that $p \geq 2$ and $C \geq 0$. Then the Euler-Lagrange ODE

$$\ddot{X}_t + \frac{p+1}{t}\dot{X}_t + Cp^2 t^{p-2} \nabla f(X_t) = 0. \tag{2}$$

has the convergence rate

$$f(X_t) - f(x^*) \leq O\left(\frac{1}{t^p}\right). \tag{3}$$

Naive Discretization (Algorithm 1). Let the identification between continuous and discrete time be defined by $t = k\delta$. The forward-backward Euler Discretization of the Euler-Lagrange (2) is given by the update equations

$$z_k = z_{k-1} - Cp(\delta k)^{p-1} \nabla f(x_k) \tag{4}$$

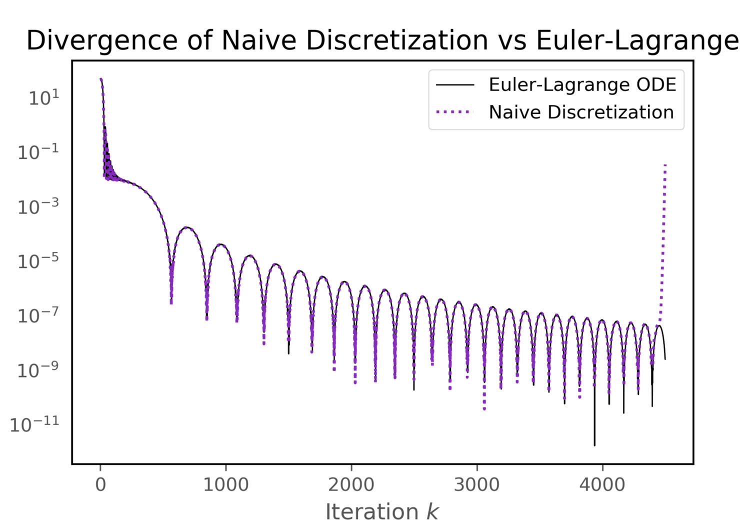
$$x_{k+1} = \frac{p}{k} z_k + \frac{k-p}{k} x_k. \tag{5}$$

Research Goals and Approach

Motivation. In [1], it was noted that Algorithm 1 eventually diverges after approaching and oscillating around the minimizer, yet it is unknown why this occurs. (See figure below)

Research Goals

1. Understand in what cases the Naive Discretization converges, and on what iteration it shoots off to infinity in cases where it diverges.
2. Develop methods of analysis that allow us to determine where divergence occurs in a given optimization algorithm.



Approach. To analyze convergence, we rewrite the update equations from Algorithm 1 in matrix form.

$$\begin{bmatrix} x_{k+1} \\ z_{k+1} \end{bmatrix} = \underbrace{\begin{bmatrix} (1-\frac{p}{k})I & \frac{p}{k}I \\ -Cp\epsilon(k+1)^{p-1}(\frac{k-p}{k})A & I - Cp\epsilon(k+1)^{p-1}(\frac{p}{k})A \end{bmatrix}}_{M_k} \begin{bmatrix} x_k \\ z_k \end{bmatrix}. \tag{6}$$

Main Theorem

Theorem. Let $f(x) : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth function defined as

$$f(x) = \frac{1}{2}(x - x^*)^T A(x - x^*) \tag{7}$$

where $x^* \in \mathbb{R}^d$ is the unique minimizer with $\nabla f(x^*) = \vec{0}$ and A is a positive definite, symmetric $d \times d$ matrix. Let $\delta < \frac{1}{L}$ and $\epsilon = \delta^p$. Then, after we go out enough iterations in the system of update equations given by Algorithm 1 such that $k > p$ and take $C < \frac{1}{\epsilon L}$, we have the following properties:

1. If $p = 2$, the naive method exhibits stable end behavior.
2. If $p > 2$, the naive method will exhibit stable behavior when

$$k < \left(\frac{4}{CLp^2\epsilon}\right)^{\frac{1}{p-2}}.$$

Sketch of Proof

Reducing The Problem to a One-Dimensional Problem
We rewrite $f(x)$ as follows:

$$f(x) = \frac{1}{2}(x - x^*)^T A(x - x^*) = \frac{1}{2}\tilde{x}^T \Sigma \tilde{x} \tag{8}$$

where $\tilde{x} = U^T(x - x^*)$, U is the matrix of eigenvectors of A, and Σ is the diagonal matrix of eigenvalues of A.

Without loss of generality, we study the case where x is one-dimensional since all dimensions of $\tilde{x}$ update independently of each other. In particular, we focus on the dimension associated with the largest eigenvalue, which is equal to L.

Relationship Between the Eigenvalues of M_k and the Iterates
We define $u_i := \begin{pmatrix} \tilde{x}_i \\ z_i \end{pmatrix}$. Computing u_k from u_0, we have

$$u_k = M_k M_{k-1} \ldots M_2 M_1 u_0 \tag{9}$$

When the all eigenvalues of M_i have magnitude less than 1, then $\|u_i\| < \|u_{i-1}\|$. Since $\|\tilde{x}_i\| \leq \|u_i\|$, the upper bound on $\|\tilde{x}_i\|$ is also strictly decreasing when all eigenvalues' magnitudes are less than 1.

Eigenvalue Analysis
To determine eventual convergence or divergence, we analyze the eigenvalues of $M_\infty = \lim_{k\to\infty} M_k$.

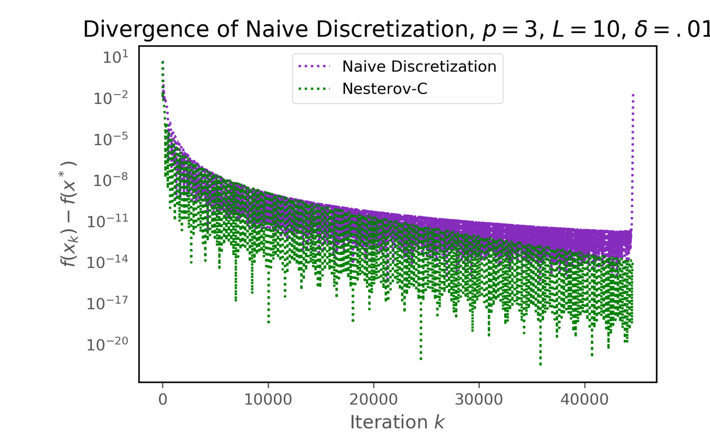
For the $p = 2$ case, we find that the magnitudes of the eigenvalues of M_k start out less than 1, and converge to 1 as $k \to \infty$.

For the $p > 2$ case, the magnitudes of the eigenvalues go to infinity as $k \to \infty$. The magnitude of the eigenvalue with larger magnitude becomes greater than 1 soon after the Kth iteration, where $K := \sqrt[p-2]{\frac{4}{Lp^2\epsilon}}$. That is where the iterations stop converging and start diverging.

Numerical Results

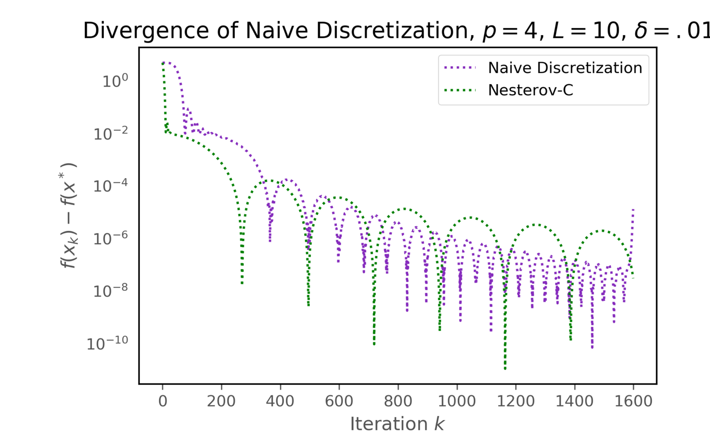
L	δ	Iteration k of divergence
10	.01	44,445
10	.001	44,444,445
100	.01	4,445
100	.001	4,444,445

For $p = 3$



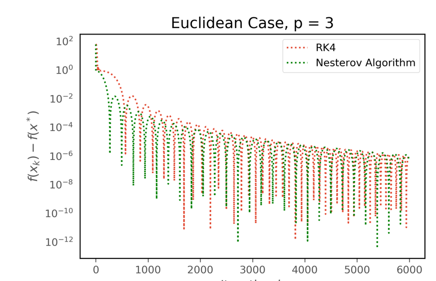
L	δ	Iteration k of divergence
10	.01	1,582
10	.001	158,113
100	.01	500
100	.001	50,000

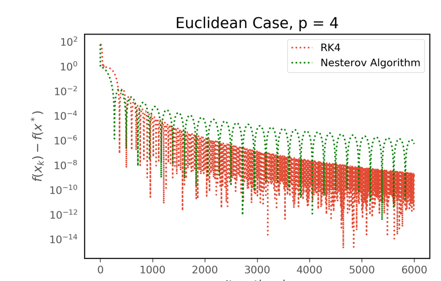
For $p = 4$

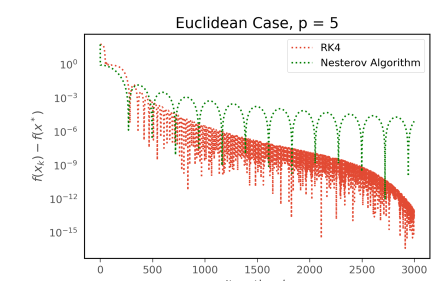


Discussion and Future Direction

1. The Euler-Lagrange ODE performs better than Nesterov's when $p > 3$. However, a higher p allows for less iterations.
2. In [1], it is shown that there is a convergent rate-matching discretization of the Euler-Lagrange by adding a third sequence. It is of interest to study why adding the third update sequence removes the problem seen with the Naive Discretization.
3. Below, we see that a fourth order Runge Kutta discretization of the Euler-Lagrange performs better than Nesterov's for $p \geq 3$ but eventually diverges. It would be of interest to do a similar analysis on this discretization scheme in future research.







Acknowledgements

This research was conducted as part of the 2019 REU program at Georgia Tech and was supported by NSF grant DMS1851843. We would like to thank our adviser Professor Rachel A. Kuske for her guidance and Dr. Andre Wibisono for his help during the research.

References

[1] Andre Wibisono, 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.

[2] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177-205, 2006

[3] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct runge-kutta discretization achieves acceleration. In *Advances in Neural Information Processing Systems*, pp. 3900-3909, 2018