On the stability of gradient descent with second order dynamics for time-varying cost functions

Travis E. Gibson[1,2,3,4] Sawal Acharya[4,5] Anjali Parashar[1] Joseph E. Gaudio[1] Anuradha M. Annaswamy[1]

1.Massachusetts Institute of Technology, 2.Harvard Medical School, 3.Broad Institute of MIT and Harvard, 4.Brigham and Women's Hospital, 5.Stanford

Introduction

Gradient based optimization algorithms deployed in Machine Learning (ML) applications are often analyzed and compared by their convergence rates or regret bounds. While these rates and bounds convey valuable information they don't always directly translate to stability guarantees. Stability and similar concepts, like robustness, will become ever more important as we move towards deploying models in real-time and safety critical systems. In this work we provide general stability guarantees for gradient descent with second order dynamics when applied to explicitly time varying cost functions. These more general results can aid in the design and certification of these optimization schemes so as to help ensure safe and reliable deployment for real-time learning applications. We also hope that the techniques provided here will stimulate and cross-fertilize the analysis that occurs on the same algorithms from the online learning and stochastic optimization communities.

Problem statement

We consider the following optimization problem, $\min_{\boldsymbol{x}\in\mathbb{R}} f_t(\boldsymbol{x})$ where for every time t, $f_t : \mathbb{R}^N \to \mathbb{R}$. The subscript t denotes the time variation in $f_t(\cdot)$ which could occur for a fixed cost function but with streaming data or in scenarios where the objective function itself changes over time. The algorithm we study in this report falls under the umbrella of what we are calling a High-order Tuner (HT) which are gradient descent algorithms with second order dynamics similar in structure to the popular momentum methods, ***but with explicitly time varying cost functions***.

The HT studied here is parameterized as follows

$$\boldsymbol{x}_t = \beta_t \boldsymbol{z}_t + (1-\beta_t)\boldsymbol{y}_t \tag{1a}$$
$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_t - \alpha_t \nabla f_t(\boldsymbol{x}_t) \tag{1b}$$
$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \eta_t \nabla f_t(\boldsymbol{x}_t) \tag{1c}$$

where α_t, β_t, and η_t are all (potentially) time varying hyper-parameters. This is almost identical to Nesterov with the significant difference being that the functions f_t are explicitly time varying.

For ease of discussion we will use the following definitions

$$\alpha_t := \frac{\mu_t}{N_t} \qquad \eta_t := \frac{\gamma_t}{N_t} \qquad \bar{\beta}_t := 1-\beta_t \qquad \bar{\mu}_t := 1-\mu_t \tag{2}$$

where γ_t, μ_t, β_t, and N_t are the free design parameters. We will refer to N_t as the normalization parameter.

Assumptions

Assumption 1: For each t the function f_t is an L_t-smooth convex function and furthermore the sequence L_t is bounded.
Assumption 2: There exists an $\boldsymbol{x}^*$ such that $f_t(\boldsymbol{x}^*) = \min f_t(\boldsymbol{x})$ for all t.
Assumption 3: For each t an upper bound on the smoothness of the function is obtained, $N_t \geq L_t$.

Why don't we talk about this with regret analysis

This notion of stability never enters one's mind when studying gradient algorithms with regret analysis because you a priori assume all the parameters and gradients are uniformly bounded before you even start the analysis.

$$\texttt{Regret}(T) := \sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - \sum_{t=1}^{T} f_t(\bar{\boldsymbol{x}}_T) \quad \text{where} \quad \bar{\boldsymbol{x}}_T = \arg\min_{\boldsymbol{x}} \sum_{t=1}^{T} f_t(\boldsymbol{x}). \tag{3}$$

ADAM paper assumptions - "Assume that the function f_t has bounded gradients, $\|\nabla f_t(\theta)\|_2 \leq G$, $\|\nabla f_t(\theta)\|_\infty \leq G_\infty$ for all $\theta \in R^d$ and distance between any θ_t generated by Adam is bounded, $\|\theta_n - \theta_m\|_2 \leq D$, $\|\theta_m - \theta_n\|_\infty \leq D_\infty$ for any $m, n \in \{1, ..., T\}$)"

Proposition 1 - Warm-up to control theory style analysis

Let f_t satisfy Assumptions 1-3. For the HT defined in Equation (1) with arbitrary initial conditions $\{\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0\}$, $N_t \geq L_t$ and

$$\beta_t \in [0,1] \tag{4a}$$
$$\mu_t \in [\epsilon, 1], \quad \epsilon > 0 \tag{4b}$$
$$\gamma_t = \tfrac{1}{2}\mu_t. \tag{4c}$$

then

$$V_t = \|\boldsymbol{z}_t - \boldsymbol{x}^*\|^2 + \|\boldsymbol{y}_t - \boldsymbol{z}_t\|^2 \tag{5}$$

is a Lyapunov candidate and $V_t \in \ell_\infty$. Furthermore, $\lim_{t\to\infty}\left[f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*)\right] = 0$.

proof:

$$\Delta V_{t+1} \leq c_1 \left\| \boldsymbol{a}_t + \frac{c_2}{2c_1}\boldsymbol{b}_t \right\|^2 + \left(c_3 - \frac{c_2^2}{4c_1}\right)\|\boldsymbol{b}_t\|^2 + c_4. \tag{6}$$

where

$$\boldsymbol{a}_t := \frac{1}{N_t}\nabla f_t(\boldsymbol{x}_t) \qquad c_1 := \gamma_t^2 + (\gamma_t - \mu_t)^2 - \gamma_t \qquad c_3 := 1 - \bar{\beta}_t^{-2}$$
$$\boldsymbol{b}_t := \boldsymbol{x}_t - \boldsymbol{z}_t \qquad c_2 := 2(2\gamma_t - \mu_t) \qquad c_4 := 2\frac{\gamma_t}{N_t}(f_t(\boldsymbol{x}^*) - f_t(\boldsymbol{x}_t))$$

- For $\Delta V_{t+1} \leq 0$ we need $c_1, c_4 \leq 0$ and $c_2^2 - 4c_1c_3 \leq 0$

$$\Delta V_{t+1} \leq 2\frac{\gamma_t}{N_t}(f_t(\boldsymbol{x}^*) - f_t(\boldsymbol{x}_t)) < 0 \; \forall \boldsymbol{x}_t \neq \boldsymbol{x}^*. \tag{7}$$

- $(\boldsymbol{z}_t - \boldsymbol{x}^*)$ and $(\boldsymbol{y}_t - \boldsymbol{z}_t)$ are bounded
- $\boldsymbol{x}^*$ is fixed and clearly bounded $\to \boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{z}_t \in \ell_\infty$.
- From smoothness and assumptions $\|\nabla f_t(\boldsymbol{x}_t)\| \in \ell_\infty$
- smooth functions on compact sets are bounded and thus $f_t(\boldsymbol{x}_t) \in \ell_\infty$.
- all signals in the model are bounded $\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{z}_t, f_t(\boldsymbol{x}_t), \nabla f_t(\boldsymbol{x}_t) \in \ell_\infty$.

To prove convergence we analyze the sequence

$$a_\tau := \sum_{t=0}^{\tau}[f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*)] \leq -\frac{\bar{N}}{\epsilon}\sum_{t=0}^{\tau}\Delta V_{t+1} = \frac{\bar{N}}{\epsilon}[V_0 - V_{\tau+1}] \leq \frac{\bar{N}}{\epsilon}V_0 < \infty \tag{8}$$

With a_τ a bounded monotonic sequence it follows that $\lim_{\tau\to\infty} a_\tau$ exists and therefore $\lim_{t\to\infty}\left[f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*)\right] = 0$.

Theorem 2 - Main result

Let f_t satisfy Assumptions 1-3. For the HT defined in (1) with arbitrary initial conditions $\{\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0\}$, $N_t \geq L_t$ and with algorithm hyper-parameters $\{\gamma_t, \mu_t, \beta_t\}$ along with the analysis parameters $\lambda_t \in [0, 1-\epsilon], \xi_t \in [\epsilon, \infty), \epsilon > 0$ chosen as

$$c_5 < 0, \quad \text{and} \quad c_7 - \frac{c_6^2}{4c_5} \leq 0 \tag{9}$$

where

$$c_5 := \gamma_t^2 + \xi_{t+1}(\gamma_t - \mu_t)^2 - (1+\lambda_t)\gamma_t \qquad c_6 := 2[\xi_{t+1}(\gamma_t - \mu_t) + \gamma_t] \tag{10a}$$
$$c_7 := \xi_{t+1} - \xi_t\bar{\beta}_t^{-2} \tag{10b}$$

then

$$W_t = \|\boldsymbol{z}_t - \boldsymbol{x}^*\|^2 + \xi_t\|\boldsymbol{y}_t - \boldsymbol{z}_t\|^2 \tag{11}$$

is a Lyapunov candidate and it follows that $W_t \in \ell_\infty$. Furthermore, $\lim_{t\to\infty}\left[f_t(\boldsymbol{x}_{t+1}) - f_t(\boldsymbol{x}^*)\right] = 0$.

Corollary 2.1 - Simplified conditions

The following simplified conditions satisfy Theorem 2

$$\gamma_t \in [1,\ 1.5] \tag{12a}$$
$$\mu_t = 1 \tag{12b}$$
$$\beta_t = 1/\gamma_t \tag{12c}$$

What about time varying optimal points?

For any piecewise constant x_t^* that changes in value only a finite number of times our algorithm remains stable. Given that our analysis holds for arbitrary initial conditions we remain bounded even if the optimal point is changed.

Consider our problem setting with arbitrary initial conditions $\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0$ and an optimal $\boldsymbol{x}^*_{[1]}$ to begin with. Our proofs say that we are stable and $f_t(\boldsymbol{x}_t)$ will asymptotically converge to $f_t(\boldsymbol{x}^*_{[1]})$. Now at some time T we get a new optimal $\boldsymbol{x}^*_{[2]}$ (with the subscript [2]). Importantly we do nothing to the algorithm and it just continues to run as originally defined. For analysis we now say $\boldsymbol{x}_T, \boldsymbol{y}_T, \boldsymbol{z}_T$ are our new initial conditions. We are stable for arbitrary initial conditions and so we will still be bounded for all time with the new $\boldsymbol{x}^*_{[2]}$ and $f_t(\boldsymbol{x}_t)$ will now asymptotically converge to the new $f_t(\boldsymbol{x}^*_{[2]})$ as well.

Experimental setup

Cost function for experiments

$$f_t(x) = \log\left(a_t \exp(-b_t(x - c_t)) + a_t \exp(b_t(x - c_t))\right). \tag{13}$$

- GD and NAGD (Figures 1 and 2A)
  - GD : $x_{t+1} = \boldsymbol{x}_t - \frac{1}{N}\nabla f_t(\boldsymbol{x}_t)$, (normalization parameter N is constant).
  - NAGD: Dynamics in (1) with $\mu_t = 1$, $\gamma_t = \frac{t}{2}$, $\beta_t = \frac{2}{t+1}$ (normalization parameter $N_t = N$ is constant).
- TN-GD and TN-NAGD (Figure 2B)
  - TN-GD : $x_{t+1} = \boldsymbol{x}_t - \frac{1}{N_t}\nabla f_t(\boldsymbol{x}_t)$, (normalization parameter N_t changes over time).
  - TN-NAGD: Dynamics in (1) with $\mu_t = 1$, $\gamma_t = \frac{t}{2}$, $\beta_t = \frac{2}{t+1}$ (normalization parameter N_t can now vary over time).
- Adagrad and Adam (Figure 3)
  - Adagrad: $x_{t+1} = \boldsymbol{x}_t - \alpha\frac{\nabla f_t(\boldsymbol{x}_t)}{\sqrt{\sum \nabla f_\tau(x_\tau)^2 + \epsilon}}$, α is a constant
  - Adam: $x_{t+1} = \boldsymbol{x}_t - \alpha_t\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$, $\alpha_t = \frac{\alpha}{\sqrt{t}}$

Conditions in Corollary 2.1 are nonvacuous

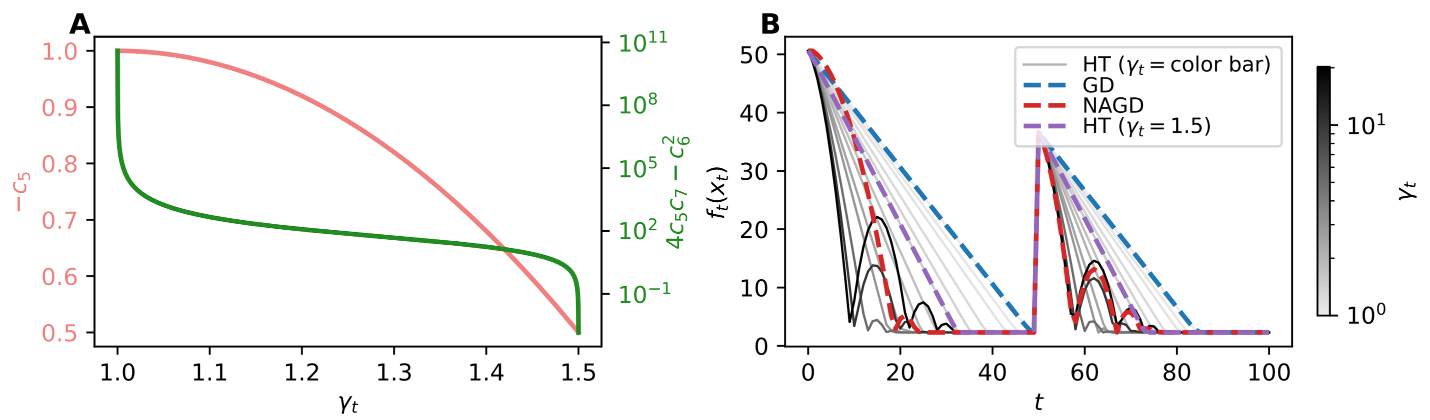

Figure 1. (A) Visualization of the sufficient conditions outlined in Equation (9) under the parameter settings in Equation (12). There is an asymptote at $\gamma_t = 1.5$ for the value of $4c_5c_7 - c_6^2$ (green line) (B) Algorithms are optimizing the function in (13) where $a_t = 1$ and $b_t = 7$. For $t < 50$ $c_t = 0$ and for $t > 50$ $c_t = 5$. The optimal solution x^* changes from 0 to 5 at $t = 50$. For the HT with γ_t varied from 1 to 10 see the solid monochrome (shades of grey) lines. With $\gamma_t = 1$ the optimizer coincides with GD and as γ_t is increased the HT begins to exhibit acceleration more closely resembling NAGD (light grey to black solid line).

An abrupt change in smoothness can cause instability

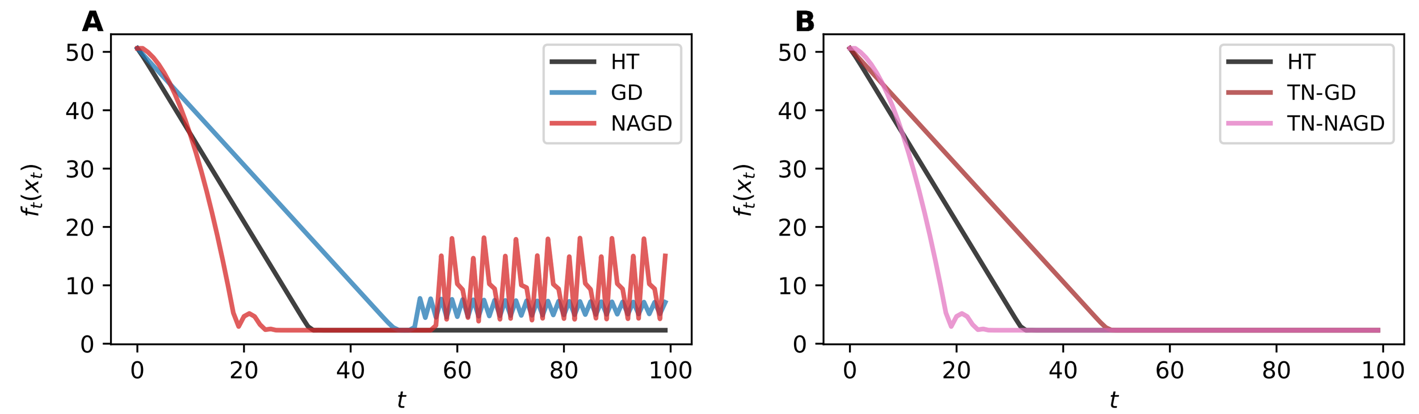

Figure 2. At $t = 50$ the parameter b_t in (13) is changed from 7 to 14. (A) The normalizer for the learning rate, N, is designed for $b_t = 7$ and when $b_t = 14$ the normalizer is too small and GD and NAGD become unstabe. The HT is provably stable with time varying N_t. (B) When the classic algorithms are given a time-varying normalizer they do not demonstrate the same poor performance. TN-GD is provably stable, but there is not such theory for TN-NAGD.

Regret optimal strategies have degraded performance over time

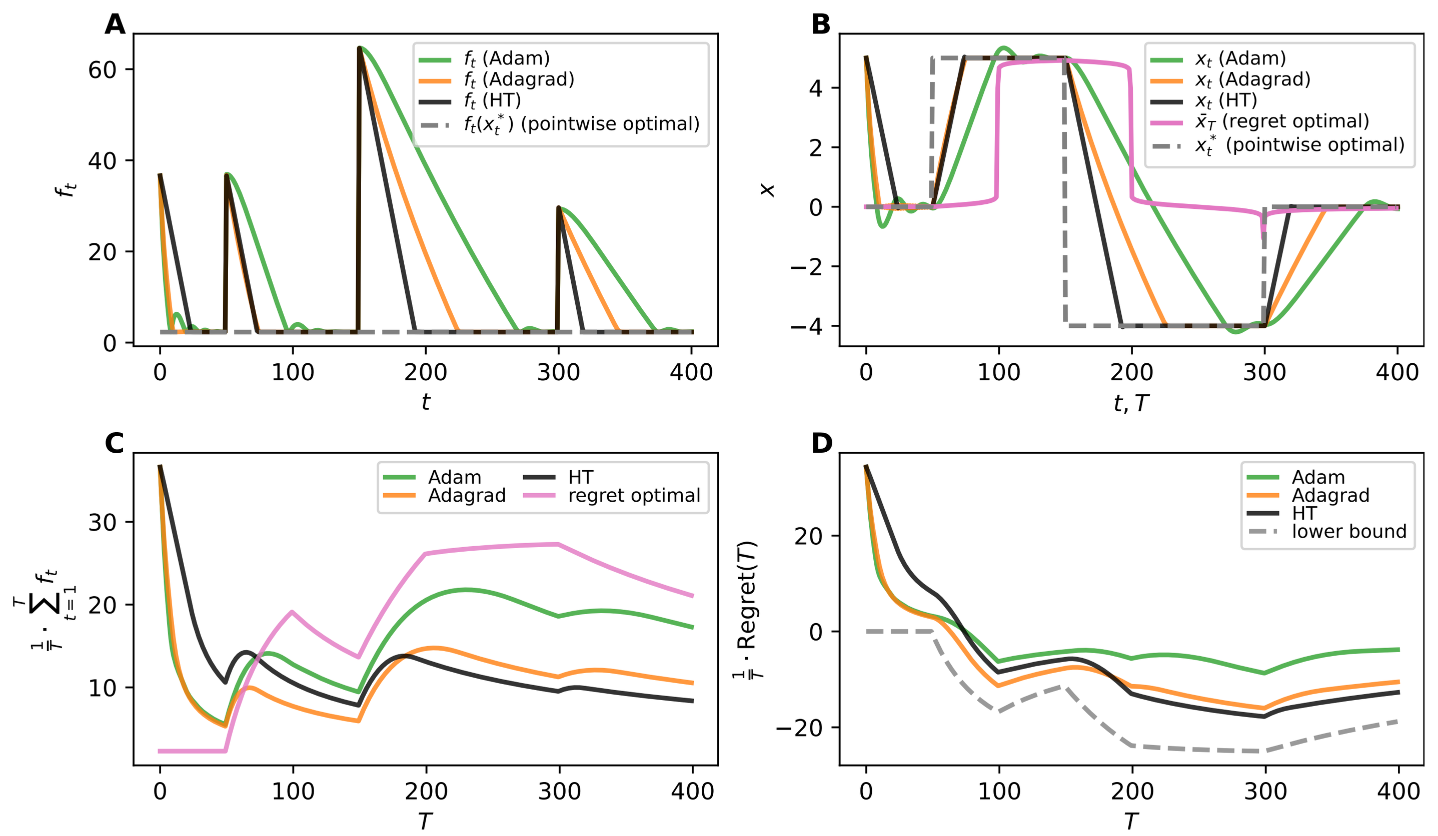

Figure 3. The parameter that controls the optimal solution c_t in (13) is initially at 0, is changed to 5 at $t = 50$, is changed to -4 at $t = 150$, and then is returned to 0 at $t = 300$. (A) The objective function f_t over time. One can see that the performance of Adam and Adagrad degrades over time. (B) The estimates x_t from the optimizers along with the regret optimal and pointwise optimal solutions. (C) Average cost over time. (D) Average regret over time.

No such thing as acceleration for time varying cost?

$$f_t(\boldsymbol{x}_t) = (1 - \boldsymbol{D}_t^\intercal \boldsymbol{x}_t)^2 \tag{14}$$

where $\boldsymbol{D}_t \in \mathbb{R}^2$ is the time varying data vector that only takes on two values (depending on the time t)

$$\boldsymbol{D}_t = \begin{cases} [1\ \ 0]^\intercal, & \text{for } t < \tau \\ [0\ \ 1]^\intercal, & \text{otherwise.} \end{cases} \tag{15}$$

The only fixed value that is optimal for all time is $\boldsymbol{x}^* = [1\ \ 1]^\intercal$, but the data stream is orthogonal before and after τ and thus gradients only update part of $\boldsymbol{x}$.