Δείξτε ότι εάν προσθέσουμε momentum στο SD, τότε πάντα υπάρχει συγκεκριμένος momentum που κάνει τον αλγόριθμο stable, ανεξάρτητα από το learning rate

Η steepest descent δίνει

$$\Delta x_k = -a \nabla F(x_k) = -a g_k$$

Εάν προσθέσω momentum, έχω

$$\Delta x_k = \gamma \Delta x_{k-1} - (1-\gamma) a g_k$$

Θυμίζω ότι η quadratic function είναι

$$F(x) = \frac{1}{2} x^T A x + d^T x + c$$

Η grad είναι $\nabla F(x) = Ax + d$

Αντικαθιστώντας έχω:

$$\Delta x_k = \gamma \Delta x_{k-1} - (1-\gamma) a (A x_k + d)$$

Αφού $\Delta x_k = x_{k+1} - x_k$ έχω

$$x_{k+1} - x_k = \gamma (x_k - x_{k-1}) - (1-\gamma) a (A x_k + d)$$

ή $$x_{k+1} = \left[(1+\gamma) I - (1-\gamma) a A \right] x_k - \gamma x_{k-1} - (1-\gamma) a d$$

Ορίζω το διάνυσμα:

$$\tilde{x}_k = \begin{bmatrix} x_{k-1} \\ x_k \end{bmatrix}$$

Άρα η SD με momentum γίνεται

$$\tilde{x}_{k+1} = \begin{bmatrix} 0 & I \\ -\gamma I & \left[(1+\gamma)I - (1-\gamma)\alpha A\right] \end{bmatrix} \tilde{x}_k + \begin{bmatrix} 0 \\ -(1-\gamma)\alpha d \end{bmatrix} = W\tilde{x}_k + v$$

Αυτό είναι linear dynamical σύστημα

Θα είναι stable, εάν οι eigenvalues του W είναι μικρότερες από 1

Θα βρούμε τις eigenvalues σε σειρά:

$$W = \begin{bmatrix} 0 & I \\ -\gamma I & T \end{bmatrix} \quad \text{όπου} \quad T = (1+\gamma)I - (1-\gamma)\alpha A$$

Οι eigenvalues ή eigenvectors του W θα ικανοποιούν

$$Wz^w = \lambda^w z^w \quad \text{or} \quad \begin{bmatrix} 0 & I \\ -\gamma I & T \end{bmatrix} \begin{bmatrix} z_1^w \\ z_2^w \end{bmatrix} = \lambda^w \begin{bmatrix} z_1^w \\ z_2^w \end{bmatrix}$$

Αυτό σημαίνει ότι

$$z_2^w = \lambda^w z_1^w \quad \text{και} \quad -\gamma z_1^w + T z_2^w = \lambda^w z_2^w$$

Έστω ότι επιλέγω την z_2^w να είναι eigenvector του T με αντίστοιχη eigenvalue την λ^t. Εάν δεν είναι σωστό, θα οδηγηθώ σε άτοπο (contradiction)

Άρα η προηγούμενη εξίσωση γίνεται

$$z_2^w = \lambda^w z_1^w \;\Rightarrow\; -\gamma z_1^w + \lambda^t z_2^w = \lambda^w z_2^w$$

Αντικαθιστώντας την πρώτη εξίσωση στην δεύτερη, βρίσκουμε

$$-\frac{\gamma}{\lambda^w} z_2^w + \lambda^t z_2^w = \lambda^w z_2^w \;\Rightarrow\; \left[(\lambda^w)^2 - \lambda^t(\lambda^w) + \gamma\right] z_2^w = \phi$$

Άρα για κάθε eigenvalue λ^t του T θα υπάρχουν δύο eigenvalues λ^w του W που είναι ρίζες της τετραγωνικής εξίσωσης

$$(\lambda^w)^2 - \lambda^t(\lambda^w) + \gamma = \phi$$

Η λύση αυτής $\lambda^w = \dfrac{\lambda^t \pm \sqrt{(\lambda^t)^2 - 4\gamma}}{2}$

Για να είναι stable ο αλγόριθμος το μέτρο της κάθε eigenvalue θα πρέπει να είναι < 1. Θα δείξουμε ότι πάντα υπάρχει εμβέλεια τιμών του γ που να ικανοποιεί αυτό.

Να σημειωθεί ότι εάν οι eigenvalues λ^w είναι complex, τότε το μέτρο τους είναι $|\lambda^w| = \sqrt{\dfrac{(\lambda^t)^2}{4} + \dfrac{4\gamma - (\lambda^t)^2}{4}} = \sqrt{\gamma}$

αυτό ισχύει για πραγματικό λ^t. Θα δείξουμε ότι είναι όντως real

Αφού το γ είναι $0 < \gamma < 1$, η τιμή της eigenvalue θα πρέπει να είναι μικρότερη από 1

Αναγένω να δείξουμε ότι υπάρχει εμβέλεια τιμών του γ για την οποία όλες οι eigenvalues είναι complex

Για να είναι complex η λ^w πρέπει

$$(\lambda^t)^2 - 4\gamma < \phi \Rightarrow |\lambda^t| < 2\sqrt{\gamma}$$

Ας εξετάσουμε τις eigenvalues λ^t του T
Αυτές μπορούν να γραφούν συναρτήσει των eigenvalues του A

Έστω $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ & $\{z_1, z_2, \dots, z_n\}$ eigenpairs του Hessian

Τότε

$$Tz_i = \left[(1+\gamma)I - (1-\gamma)\alpha A\right] z_i = (1+\gamma)z_i - (1-\gamma)\alpha A z_i$$

$$= (1+\gamma)z_i - (1-\gamma)\alpha\lambda_i z_i = \left\{(1+\gamma) - (1-\gamma)\alpha\lambda_i\right\} z_i = \lambda_i^t z_i$$

Άρα τα eigenvectors του T είναι τα ίδια με του A, και οι eigenvalues είναι

$$\lambda_i^t = \left\{(1+\gamma) - (1-\gamma)\alpha\lambda_i\right\}$$

Η λ_i^t είναι real, αφού γ, α και λ_i (για συμμετρικό A) είναι real

Άρα για να είναι η λ^w complex, πρέπει

$$|\lambda^t| < 2\sqrt{\gamma} \quad \text{ή} \quad \left|(1+\gamma) - (1-\gamma)\alpha\lambda_i\right| < 2\sqrt{\gamma}$$