IFT 6085 - Lecture 5
Accelerated Methods

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribes: Aldo Lamarre, Breandan Considine

Instructor: Ioannis Mitliagkas

1 Summary

Previously, we have seen some convergence guarantees, tightness bounds and topics in convex analysis. We established worst case convergence guarantees for various gradient descent algorithms on several families of convex functions:

Property of f	Rate of convergence
L-lipschitz	$\frac{D_1 L}{\sqrt{T}}$
β-smooth	$\frac{D_1^2 \beta}{T}$
α-strongly convex and L-lipschitz	$\frac{L^2}{\alpha T}$
α-strongly convex and β-smooth	$D_1^2 e^{\frac{-T}{\kappa}}$

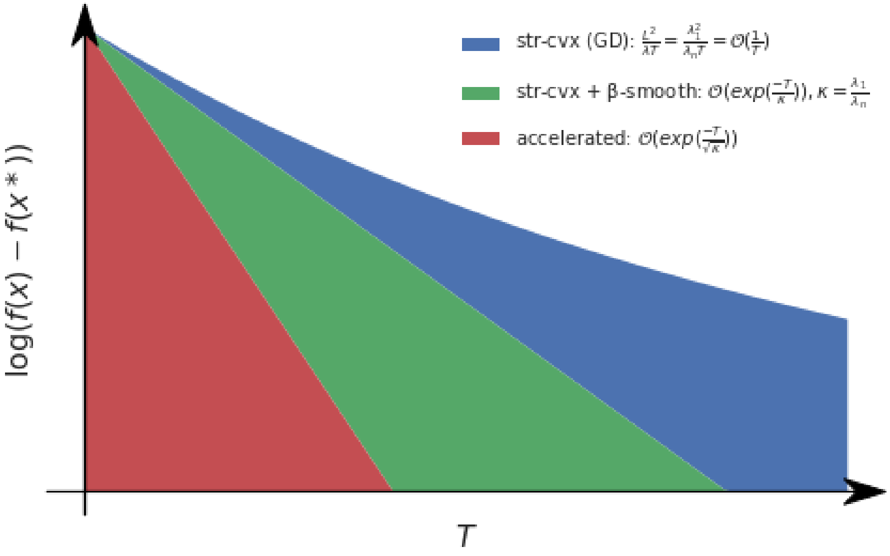
Weak Assumptions	Strong assumptions
$\lambda - Convexity$	Strong Convexity
Lipschitz	β-Smoothness
Continuity	

In this lecture we discuss the optimal learning rate for gradient descent, step size, Polyak's Momentum (a.k.a. the heavy-ball method) and some convergence guarantees on quadratic objectives.

$$f(x) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}, \quad \text{where } \mathbf{H} \succcurlyeq 0 \text{ positive semi-definite and } Dom(f) = \{x : \|x\|_2 \leq 1\} \tag{1}$$

Here, we assume a bounded domain, which implies the Lipschitz property. That is, the hessian has no eigenvalue less than 0. For differentiable functions, the norm of gradient descent does not exceed L.

- $\mathbf{H}$ is positive semi-definite $\implies \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0 \implies f$ is λ_n-strongly convex
- $\nabla f(x) = Hx \implies \|\nabla f(x)\|_2 \leq \|H\|_2 \|x\|_2 \leq \lambda_1 = L \implies f$ is λ_1-lipschitz



The notion of strong convexity is different from that of strict convexity.

- Strict convexity: $\nexists\lambda$ s.t. $\nabla^2 f(x) \succcurlyeq \lambda\mathbf{I}, \forall x$ (typically studied in unbounded domains)
- Strong convexity: $\exists\lambda \geq 0$ s.t. $\nabla^2 f(x) \succcurlyeq \lambda\mathbf{I}, \forall x \in Dom(f)$

2 Optimal learning rate for standard Gradient Descent

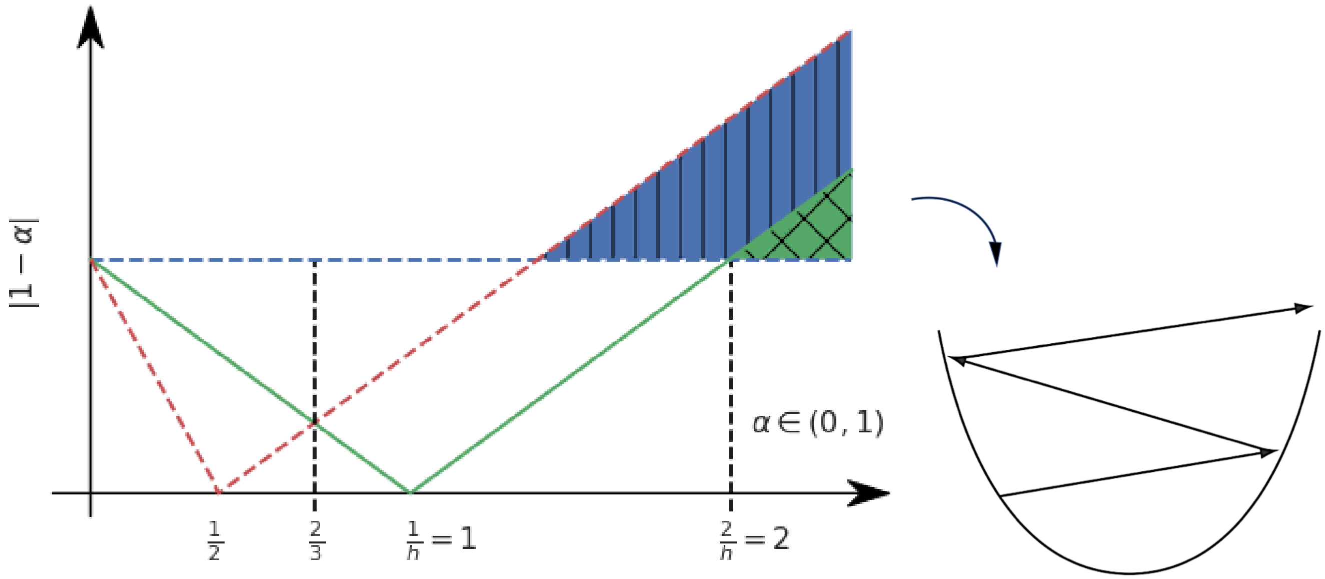
In this section, we give an intuition behind the optimal choice of learning rate for G.D. by using a quadratic example.

$$\alpha = \frac{2}{\beta + \lambda} \tag{2}$$

Recall the update rule for standard gradient descent.

$$x_{t+1} = x_t - \alpha\nabla f(x_t) \tag{3}$$

We have $|x_t - x^*| = |1 - \alpha||x_0|$.



Here, we want to minimize the maximum of the learning rates along each dimension ($max\{|1-\lambda_1\alpha|, |1-\lambda_n\alpha|\}$). If one rate is strictly greater than the other, we want to minimize this rate, since it dominates the other asymptotically.

The shaded section to the upper right is known as ”divergence”. The optimal choice of learning rate is where the two rates intersect (which in the following example is $\frac{2}{3}$). When α is too big, we start to bounce upward. We will now proceed to show the optimal step size for an example quadratic function of our choice.

Example

- $f(x) = \frac{1}{2}x^2$
- $g(x) = x^2$

Consider the following three-dimensional quadratic function:

$$f(x, y) = \frac{1}{2}x^2 + y^2 = f(z) = \frac{1}{2}\mathbf{z}^T\mathbf{H}\mathbf{z}, \text{ where } \mathbf{H} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \tag{4}$$

Where $z = (x, y)$. Taking the gradient, we get $\begin{bmatrix} x \\ 2y \end{bmatrix}$. The recurrence then becomes:

$$\nabla f(x,y) = \begin{bmatrix} x \\ 2y \end{bmatrix} \implies \begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} x_t \\ y_t \end{bmatrix} - \alpha \begin{bmatrix} x_t \\ 2y_t \end{bmatrix} \tag{5}$$

This now breaks up into the following scalar recurrences:

$$\begin{aligned} x_{t+1} &= (1-\alpha)x_t \\ y_{t+1} &= (1-2\alpha)y_t \end{aligned} \tag{6}$$

Starting with the $L2$ norm squared, we proceed to solve for α,

$$\implies \left\| \begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} \right\|_2^2 = \|x_{t+1}\|_2^2 + \|y_{t+1}\|_2^2 = (1-\alpha)^{2(t+1)}|x_0|^2 + (1-2\alpha)^{2(t+1)}|y_0|^2 \tag{7}$$

Since α is always positive by equation 2, we have the following inequality:

$$1 - \alpha > 1 - 2\alpha \tag{8}$$

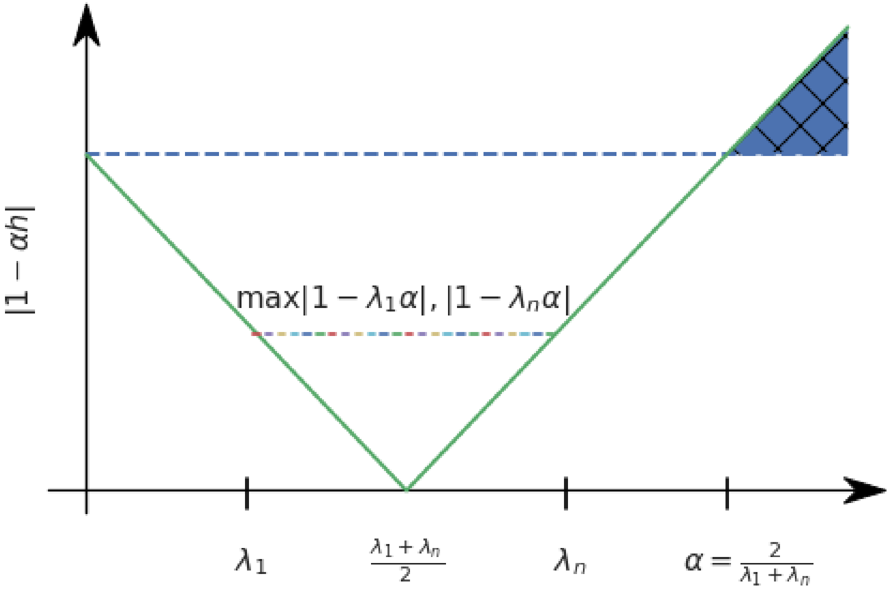
Now, from the previous inequality and system of equations:

$$\text{Equation } 8 \wedge \text{equation } 7 \implies \mathcal{O}\left(\left\| \begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} \right\|_2^2 \right) = \mathcal{O}\left((1-\alpha)^{2(t+1)} \right) \tag{9}$$

$$\implies \alpha = \frac{2}{\lambda_1 + \lambda_n} = \frac{2}{1+2} = \frac{2}{3} \tag{10}$$

Over the range of curvatures present, we would like to have a uniform rate of convergence along all directions. That is, we do not want our gradient descent to converge to the optimum in any one direction more quickly than others.

We would like to find the inverse of the Hessian. If we could do so, we would converge to the global optimum (in one step!) for quadratic functions. Due to the complexity of this task, we must approximate. One way to do so is calculating the inverse average of the eigenvalues. The following figure depicts the inverse average of the eigenvalues.


$|1 - \alpha h|$
$\max|1-\lambda_1\alpha|, |1-\lambda_n\alpha|$
λ_1
$\frac{\lambda_1+\lambda_n}{2}$
λ_n
$\alpha = \frac{2}{\lambda_1+\lambda_n}$

3 Polyak's momentum

Polyak's momentum introduces a "momentum" term $\mu|x_t - x_{t-1}|$, inspired by physics. If we imagine the current iterate as a object with mass, then our gradient descent update should be proportional to the previous step size. The full momentum update is:

$$x_{t+1} = x_t - \alpha\nabla f(x_t) + \mu|x_t - x_{t-1}| \tag{11}$$

Where $\mu \in [0, 1)$. Using an augmented state space, we get the linear recurrence:

$$\begin{bmatrix} x_{t+1} \\ x_t \end{bmatrix} = \begin{bmatrix} 1 - \alpha h_i + \mu & -\mu \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_t(i) \\ x_{t-1}(i) \end{bmatrix} \tag{12}$$

$$x_{t+1} = x_t - \alpha\mathbf{H}x_t + \mu(x_t - x_{t-1}) \tag{13}$$

Without loss of generality, $x^* = 0$. $f(x) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$. If we unroll the recursion above, we end up with an exponentially weighted average of past gradients.

$$x_{t+1} = x_t - \alpha\mathbf{H}x_t + \mu(x_t - x_{t-1}) = x_{t+1}(i) = x_t(i) - \alpha h_i x_t(i) + \mu(x_t(i) - x_{t-1}(i)) \tag{14}$$

We could try to use the determinant to solve this set of linear equations. But for simplicity's sake, we will focus on the asymptotic behavior. In this case, the spectral radius will tell us the actual rate. To get this, we use the following linear operator:

$$\left\| \begin{bmatrix} x_{t+1} \\ x_t \end{bmatrix} \right\|_2 = \left\| \mathbf{A}^{t-1} \begin{bmatrix} x_1(i) \\ x_0(i) \end{bmatrix} \right\|_2 \tag{15}$$

Where $\mathbf{A}$ is taken from 12.

$$\mathbf{A} = \begin{bmatrix} 1 - \alpha h_i + \mu & -\mu \\ 1 & 0 \end{bmatrix} \tag{16}$$

By the induced norm we have:

$$\left\|\mathbf{A}^{t-1}\right\| \left\| \begin{bmatrix} x_1(i) \\ x_0(i) \end{bmatrix} \right\|_2 \leq \left\|\mathbf{A}^{t-1}\right\| \left\| \begin{bmatrix} x_1(i) \\ x_0(i) \end{bmatrix} \right\|_2 \tag{17}$$

With lemma 11 from [4] there must exist a matrix norm such that:

$$\left\|\mathbf{A}^{t-1}\right\| \leq (\rho(A) + \epsilon)^{t-1} \tag{18}$$

with $\rho(A) = \max\{|\lambda_1|, |\lambda_2|\}$ (spectral radius, or the max of the two eigenvalue magnitudes). Therefore,

$$\left\|\mathbf{A}^{t-1}\right\| \left\| \begin{bmatrix} x_1(i) \\ x_0(i) \end{bmatrix} \right\|_2 \leq (\rho(A) + \epsilon)^{t-1} \left\| \begin{bmatrix} x_1(i) \\ x_0(i) \end{bmatrix} \right\|_2 \tag{19}$$

Assumption 1. *Assume that,* $\forall i, (1 - \sqrt{\mu})^2 \leq \alpha h_i \leq (1 + \sqrt{\mu})^2$. *This implies:*
$\rho(A) = \sqrt{\mu}$

As long as the previous assumption holds, the convergence rate of Polyak's Momentum does not depend on the step size or curvature, only on the momentum term $\mu|x_t - x_{t-1}|$. There is a large range of step sizes and curvatures over which the convergence rate of Polyak's Momentum stays constant.

Lemma 2 (Result). $\mu^* = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$
rate $= \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$ *cf. Gradient Descent:* $\left(\frac{\kappa-1}{\kappa+1}\right) \leq e^{\frac{-1}{\kappa}}$

References

[1] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.

[2] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[3] S. Bubeck. Nesterovs accelerated gradient descent for smooth and strongly convex optimization, 2014.

[4] S. Foucart. University Lecture, 2012. URL `http://www.math.drexel.edu/~foucart/TeachingFiles/F12/M504Lect6.pdf`.

[5] G. Goh. Why momentum really works. *Distill*, 2(4):e6, 2017.

[6] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[7] J. Zhang, I. Mitliagkas, and C. Ré. Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471*, 2017.