IFT 6085 - Lecture 5
Accelerated Methods - Polyak's Momentum (Heavy Ball Method)

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

Scribes
Winter 2020: Etienne Thuillier
Winter 2019: Veronica Chelu, Andre Cianflone
Winter 2018: Aldo Lamarre, Breandan Considine

Instructor: Ioannis Mitliagkas

1 Summary

In the previous lectures we summarized our analysis of upper bounds for the rate of convergence of gradient descent when applied to convex objective functions with various properties. We saw that using β-smoothness and α-strongly convex functions give exponential rate of convergence.
These upper bounds are summarized in Table 1

Property of the Convex Objective Function	Upper Bound on Convergence Rate
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	$D_1^2 \exp(\frac{-T}{\kappa})$

Table 1: Various upper bounds on convegence rates in gradient descent depending on the properties of the objective function. Note $D_1 = \|x_1 - x^*\|_2^2$, [3].

In the last lecture we derived a lower bound for the rate of convergence of any black box model on β-smoothness and α-strongly convex objectives by constructing the "hardest function" in that class of functions. Our analysis left a gap between the lower bound and the upper bound of the rate of convergence of gradient descent which we aim to fill in this lecture by showing that accelerated methods achieve this lower bound.

In this lecture we first analyze how the step size affects the convergence rate of gradient descent for quadratic objectives. Then we introduce an alternative convergence proof technique using eigenvalue analysis of operators. Subsequently, we introduce Polyak's momentum [5] (a.k.a. heavy ball method) and some convergence guarantees on the same objectives. This analysis allows us to close the gap between the upper bound and lower bound on the convergence rate of gradient descent for quadratic functions. These notes are based on [2, 6, 4].

2 Convergence of gradient descent

In this section we analyze how the choice of step size affects the rate of convergence of gradient descent and give some intuition behind the optimal choice using quadratic objectives.

Recall the update rule for standard gradient descent.

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

We start by considering the simplest scalar quadratic objective function:

$$f(x) = \frac{h}{2}x^2 \tag{2}$$

We seek to minimize this function using gradient descent, giving rise to the following update rule:

$$\begin{aligned} x_{t+1} &= x_t - \gamma \nabla f(x_t) \\ &= x_t - \gamma h x_t \\ &= (1 - \gamma h) x_t \\ &= (1 - \gamma h)(1 - \gamma h) x_{t-1} \quad \text{expanding } x_t \\ &= (1 - \gamma h)(1 - \gamma h)(1 - \gamma h) x_{t-2} \\ &\vdots \end{aligned}$$

We note that this is a linear system and repeated applications of the linear operator arrive at:

$$x_{t+1} = (1 - \gamma h)^t x_1, \tag{3}$$

The speed at which the sequence x_{t+1} converges to x^* is determined by the rate of convergence $\rho = |1 - \gamma h|$. This implies that if we set a suitable step size γ such that $\rho < 1$ then $\|x_t - x^*\| \to 0$.

We note the **relaxation property** of gradient descent:

$$f(x_{t+1}) \leq f(x_t), \tag{4}$$

which reflects the fact that the objective function does not increase, provided a small enough step size γ. This is a natural property to have for optimization, and it is crucial for the analysis of gradient descent. However, we lose it with certain accelerated methods like momentum.

Going back to the case of the scalar quadratic objective, below we show the convergence rate dependence on the learning rate and curvature respectively. We see that if $\gamma = \frac{2}{h}$, given the update rule $x_{t+1} = x_t - \frac{2}{h} h x_t = -x_t$, this means we will oscillate forever. Whenever $\gamma > \frac{1}{h}$ we will have some kind of oscillation. However, if $\gamma = 1$, nothing changes. The plots in Figure 1 look particularly identical because γ and h appear together in the rate of convergence equation.

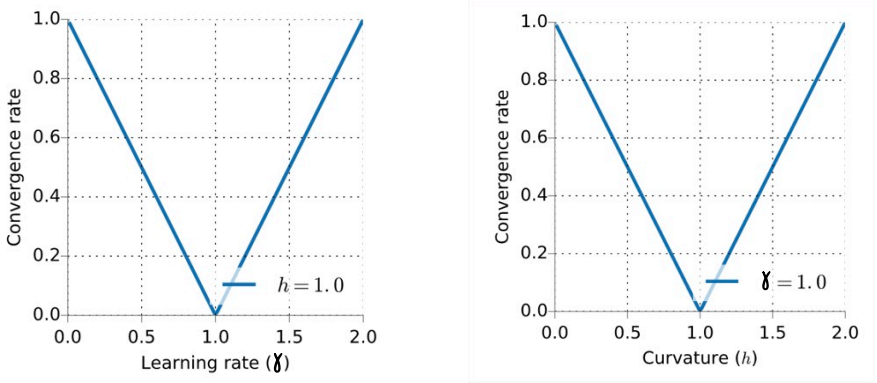

Figure 1: Convergence rate as a function of the learning rate (left plot) and curvature (right plot), for the function $\frac{h}{2}x^2$, [3].

Moving on to a multivariate quadratic objective function, we assume without loss of generality that the Hessian H is diagonal. We can make this assumption because if H is general non-diagonal, and symmetric by definition, we can always do a decomposition into $H = U\Lambda U^T$ and a change of basis to get a diagonal Hessian:

$$f(x) = \frac{1}{2}x^T U\Lambda U^T x, \text{ with } w = U^T x \text{ giving } f(w) = \frac{1}{2}w^T \Lambda w \tag{5}$$

Suppose then that we have the following quadratic:

$$f(x) = \frac{1}{2}x^T Hx \quad \text{where } H = \begin{pmatrix} h_1 & 0 & 0 \\ 0 & h_2 & 0 \\ 0 & 0 & h_3 \end{pmatrix} \tag{6}$$

We can further decompose the vector dynamics into scalar dynamics dependent on eigendirections, where we denote with i the component of the vector x and the correspondent curvature in the Hessian h.

$$\begin{aligned} x_{t+1}(i) &= x_t(i) - \gamma h(i) x_t(i) \\ &= (1 - \gamma h(i)) x_t(i) \\ &= (1 - \gamma h(i))^t x_1(i) \end{aligned}$$

Taking the following concrete example, where

$$H = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \tag{7}$$

we observe that we get contrasting convergence rates for different directions that have distinct curvatures. The convergence rate thus decomposes into a sum where the largest convergence rate dominates - indicated by the lowest eigenvalue:

$$||x_t - x^*||^2 \leq c1\rho_1^{2t} + c2\rho_2^{2t} + c3\rho_3^{2t}. \tag{8}$$

Note the convergence rate for a given curvature will vary depending on the learning rate. In the left plot of Figure 2, convergence rate continues to decrease with an increasing curvature and a learning of $\gamma = 0.25$. In the right plot of Figure 2, convergence diverges with increasing curvature, for a learning rate of $\gamma = 0.75$.
We can establish then that the goal is to find:

$$\min_{\gamma} \max\{\rho_1, \rho_2, \rho_3\} \tag{9}$$

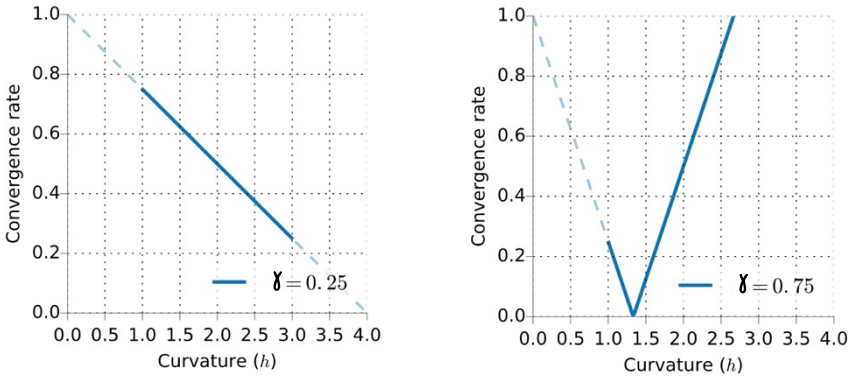

Figure 2: Convergence rate as a function of curvature. The left plot shows convergence with learning rate $\gamma = 0.25$, and $\gamma = 0.75$ for the right plot, [3].

Figure 3 shows that the optimal learning rate is that which balances the convergence rate between extreme curvatures. Setting the convergence rate to be equal for the smallest and largest eigenvalues, we can solve for the optimal step size and rate:

$$|1 - \gamma h_{\min}| = |1 - \gamma h_{\max}| \tag{10}$$

$$\gamma^* = \frac{2}{h_{\min} + h_{\max}} = 0.5 \tag{11}$$

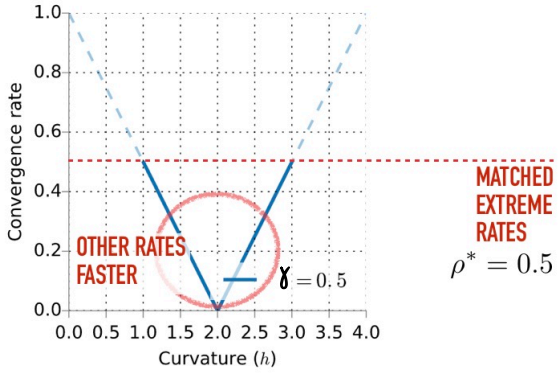

Figure 3: Convergence rate as a function of curvature given the optimal learning rate, [3].

It may seem strange that the dimension with highest curvature would converge at the same rate as the dimension with lowest curvature, as shown in Figure 3. The intuition is that the sequence of updates oscillates in the dimension with highest curvature, while in the other one convergence is very slow.

3 Polyak's momentum

Momentum gradient descent, or the heavy ball algorithm was first proposed in the 60s. It combines the current gradient with a history of the previous step to accelerate the convergence of the algorithm. For example, the images below show a valley like landscape, where the algorithm wants to reach the optimal point.

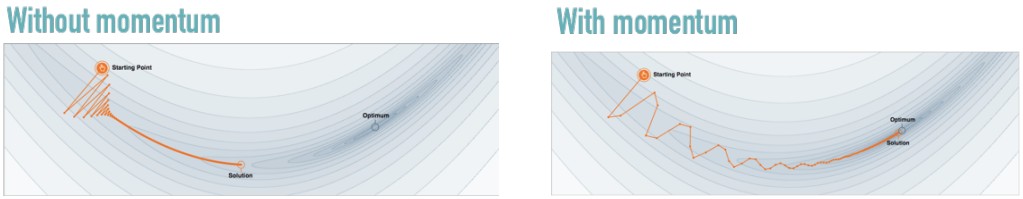

Figure 4: Without momentum, gradient descent oscillates, whereas with momentum, we find that it converges much closer to the optimal point in the same number of iterations, [2].

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

$$x_{t+1} = x_t - \gamma \nabla f(x_t) + \mu(x_t - x_{t-1}) \tag{12}$$

where μ is a hyperparameter (typically $\mu \in [0, 1]$, although not limited to it), which scales down the previous step. Adding this scaled previous step controls oscillation and in low curvatures causes acceleration along the same direction. The overall effect is that it allows the step size γ to be larger and decreases the number of steps to convergence,

which is illustrated in the change of convergence rate.

Figure 5 illustrates how the convergence rate dynamics change in this case, i.e. we achieve the same rate of convergence for all curvature - a property unique to Polyak's momentum.

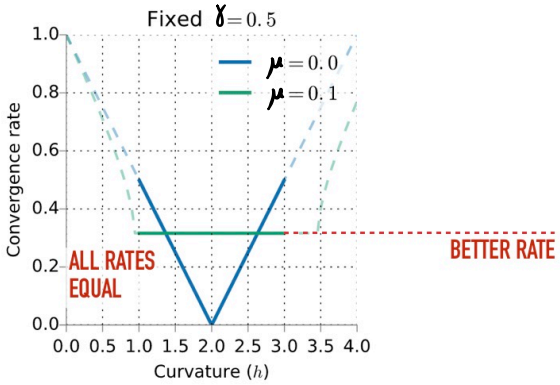

Figure 5: Convergence rate as a function of curvature. The blue curve is standard gradient descent without momentum, whereas the green curve includes momentum with $\mu = 0.1$, [3].

3.1 Convergence rate of Polyak's Momentum for a quadratic loss

Going back to our previous scalar quadratic objective:

$$f(x) = \frac{h}{2}x^2 \tag{13}$$

Let's write the momentum update rule for this function and do some simple algebraic manipulations to obtain a more convenient form:

$$\begin{aligned} x_{t+1} &= x_t - \gamma \nabla f(x_t) + \mu(x_t - x_{t-1}) \\ &= x_t - \gamma h x_t + \mu(x_t - x_{t-1}) \\ &= (1 + \mu - \gamma h)x_t - \mu x_{t-1} \end{aligned}$$

We can further write the above equation as a linear system:

$$\begin{bmatrix} x_{t+1} \\ x_t \end{bmatrix} = \begin{bmatrix} 1 - \gamma h + \mu & -\mu \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix} \tag{14}$$

Note the recursive property of the above. By denoting the linear operator with A, we can recurse A for t steps and express x_{t+1}, x_t as a function of starting values x_1, x_0:

$$\begin{bmatrix} x_{t+1} \\ x_t \end{bmatrix} = A^t \begin{bmatrix} x_1 \\ x_0 \end{bmatrix} \tag{15}$$

Consider comparing our iterated x_t with optimal x^*:

$$\begin{aligned} \begin{bmatrix} x_{t+1} - x^* \\ x_t - x^* \end{bmatrix} &= A^t \begin{bmatrix} x_1 - x^* \\ x_0 - x^* \end{bmatrix}, \\ \left\| \begin{bmatrix} x_{t+1} - x^* \\ x_t - x^* \end{bmatrix} \right\|_2 &= \left\| A^t \begin{bmatrix} x_1 - x^* \\ x_0 - x^* \end{bmatrix} \right\|_2, \\ &\leq \left\| A^t \right\|_2 \left\| \begin{bmatrix} x_1 - x^* \\ x_0 - x^* \end{bmatrix} \right\|_2. \end{aligned} \tag{16}$$

One could be tempted to note that:

$$\begin{aligned} \|A^t\|_2 &\leq \|A\|_2^t, && \text{(submultiplocativity of norms, see [1], p. 4),} \\ &= \sqrt{\rho(A^T A)}^t, && \text{(Theorem 1, Appendix A),} \end{aligned}$$

where $\rho(\cdot)$ denotes the spectral radius.

Unfortunately, this result produces a diverging upper bound if combined into (16), since $\sqrt{\rho(A^T A)} > 1$ (we do not demonstrate this inequality here).

Thankfully, Theorem 4 (Appendix A) provides the means for obtaining the converging bound we seek.

This Theorem, when applied to matrix A and for a chosen constant $\epsilon > 0$ (more on this later), provides that a norm, denoted below $|||\cdot|||$, exists that verifies:

$$|||A||| \leq \rho(A) + \epsilon. \tag{17}$$

Furthermore, following an argument in [1], page 7, the equivalence of the norms provides that there exists a constant $\mathrm{C} > 0$ for which:

$$\|M\|_2 \leq \mathrm{C}|||M|||, \qquad \forall\, M \in \mathbb{C}^{n\times n}.$$

Hence

$$\begin{aligned} \|A^t\|_2 &\leq \mathrm{C}|||A^t|||, \\ &\leq \mathrm{C}|||A|||^t, && \text{(submultiplicativity of norms, see [1], p. 4).} \end{aligned}$$

Use of (17) now yields

$$\|A^t\|_2 \leq \mathrm{C}|||A|||^t \leq \mathrm{C}\left(\rho\left(A\right) + \epsilon\right)^t,$$

which by combining back into (16) yields

$$\begin{aligned} \left\| \begin{bmatrix} x_{t+1} - x^* \\ x_t - x^* \end{bmatrix} \right\|_2 &\leq \mathrm{C}\left(\rho\left(A\right) + \epsilon\right)^t \left\| \begin{bmatrix} x_1 - x^* \\ x_0 - x^* \end{bmatrix} \right\|_2, \\ \left\| \begin{bmatrix} x_{t+1} - x^* \\ x_t - x^* \end{bmatrix} \right\|_2 &= \mathcal{O}\left(\rho\left(A\right)^t + \epsilon^t\right), \qquad \epsilon > 0. \end{aligned}$$

Since we can set $\epsilon \ll \rho\left(A\right)$, in practice we discard the term in ϵ for conciseness (although this is not absolutely rigorous) and we obtain the following upper bound on the convergence rate:

$$\left\| \begin{bmatrix} x_{t+1} - x^* \\ x_t - x^* \end{bmatrix} \right\|_2 = \mathcal{O}\left(\rho\left(A\right)^t\right).$$

3.2 Robust region and optimal step size

If we write down the algebraic form of the determinant:

$$det(A) = \lambda_1 \lambda_2 = \mu \tag{18}$$

we can observe that it is not dependent on γ and this gives an intuitive reason for the flatness of the convergence rate.

We can now do an analysis of the spectral radius of the momentum operator A:

$$A = \begin{bmatrix} 1 - \gamma h + \mu & -\mu \\ 1 & 0 \end{bmatrix} \tag{19}$$

Figure 6 shows the two distinct dynamics region induced by the discriminant:

$$\Delta = tr(A)^2 - 4det(A) \tag{20}$$

We can observe that when $\Delta > 0$ we get two real eigenvalues and when $\Delta < 0$ we have two complex conjugate eigenvalues $\frac{(1-\sqrt{\mu})^2}{h} \leq \gamma \leq \frac{(1+\sqrt{\mu})^2}{h}$. We can visualize this process in the figure 6, where we can see that the two eigenvalues split at some point and become conjugate of each other and as such we enter a robust region where the rate of convergence is $\rho(A) = \sqrt{\mu}$. Also note that the width of the robust region is given by μ.
We can demonstrate this by explicitly computing the eigenvalues of our matrix:

$$\lambda_1 = \frac{1}{2}\left(1 - \gamma h + \mu + \sqrt{(-\gamma h + \mu + 1)^2 - 4\mu}\right) \tag{21}$$

$$\lambda_2 = \frac{1}{2}\left(1 - \gamma h + \mu - \sqrt{(-\gamma h + \mu + 1)^2 - 4\mu}\right) \tag{22}$$

When $(-\gamma h + \mu + 1)^2 - 4\mu < 0$, then the roots are complex conjugates, which implies the absolute values of the eigenvalues are identical. Therefore:

$$|\lambda_1| = |\lambda_2| = \sqrt{(1 - \gamma h + \mu)^2 + |(-\gamma h + \mu + 1)^2 - 4\mu|} = \sqrt{\mu} \tag{23}$$

Which implies the converge rate is the same and solely dependent on μ. This leads to the following lemma.

Lemma 1 (Robust Region). *For some choice of γ, the absolute eigenvalues are identical and*

$$\rho(A) = \sqrt{\mu}$$

holds for a "robust region", that is to say the convergence rate is constant for the range of γ where the discriminant of the momentum operator A is less than 0.

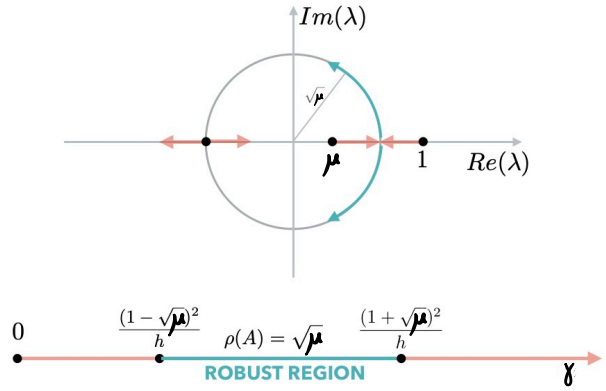

Figure 6: Real and imaginary eigenvalues of the momentum operator A with varying learning rates, [3].

The images below depict the rate of convergence with respect to the curvature and the step size. We can note that larger values of μ give larger widths for the robust region, but also larger rate of convergence.

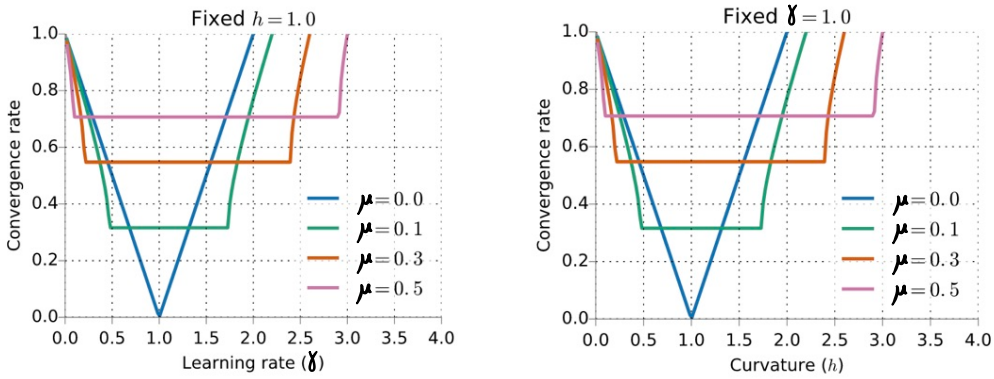

Figure 7: Convergence rates displaying robust regions as a function of learning rate and curvature for various momentum sizes, [3].

Note that as we increase the width of the robust regions for the extreme curvatures, the point where the two regions meet corresponds to the optimal step size. Consequently, this leads to the following optimal step size lemma.

Lemma 2 (Optimal γ, μ)**.** *The optimal step size γ is given by:*

$$\gamma^* = \frac{(1+\sqrt{\mu})^2}{h_{\max}} = \frac{(1-\sqrt{\mu})^2}{h_{\min}}, \tag{24}$$

from which the value of the optimal rate of convergence ρ can be derived as:

$$\rho^* = \sqrt{\mu^*} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \underset{\text{for large } k}{\approx} \exp(-\frac{C}{\sqrt{\kappa}}), \tag{25}$$

where κ denotes the condition number $\kappa = \frac{h_{\max}}{h_{\min}}$, and μ is the momentum coefficient

A Spectral radius $\rho(\cdot)$ and operator norm $\|\cdot\|$

Definitions

Definition 1 (Operator Norm, [1], Def. 8)**.** *If $\|\cdot\|$ is a vector norm on $\mathbb{C}^n$, then the induced norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$ defined by*

$$\|A\| \triangleq \max_{\|x\|=1} \|Ax\|$$

is a matrix norm on $\mathbb{C}^{n\times n}$.

Note that in that the notation $\|\cdot\|$ is overloaded in the above definition: it designates the vector norm for a vector argument, and an operator norm for a matrix argument. In particular, we note $\|\cdot\|_q$ the operator norm induced by the vector norm $\|\cdot\|_q$, e.g.:

$$\|A\|_1 \triangleq \max_{\|x\|_1=1} \|Ax\|_1,$$
$$\|A\|_2 \triangleq \max_{\|x\|_2=1} \|Ax\|_2.$$

Definition 2 (Spectral Radius, [1], Sec. 3)**.** *The spectral radius of a matrix $M \in \mathbb{C}^{n\times n}$ is given by:*

$$\rho(A) \triangleq \max\{|\lambda|, \lambda \textit{ eigenvalue of } A\}$$

Equality relations (for $\|\cdot\|_2$ only)

Equality relations exist between the spectral radius and the matrix 2-norm specifically. These are cited here for completeness as they are not used in the development of the convergence rate bound.

Theorem 1 ([1], Proposition 9)**.**

$$\|A\|_2 = \sqrt{\rho(A^*A)}, \qquad \forall\, A \in \mathbb{C}^{n\times n}.$$

This relation simplifies further if the matrix is hermitian symmetric.

Theorem 2 ([1], page 5)**.**

$$\|A\|_2 = \rho(A), \qquad \forall\, A \in \mathbb{C}^{n\times n} \textit{ and hermitian, i.e. for which } A^* = A.$$

Inequality relations

The following inequalities hold more generally. Note that the first one is cited for completeness: it is not used in the development of the convergence rate bound. However the second one is.

Theorem 3 (Lower bound on operator norm, [1], Lemma 10)**.**

$$\rho(A) \leq \|A\|, \qquad \forall\, A \in \mathbb{C}^{n\times n}.$$

Theorem 4 (Upper bound on operator norm, [1], Lemma 11)**.** *Given $A \in \mathbb{C}^{n\times n}$ and $\epsilon > 0$, there exists a norm $\|\cdot\|$ such that*

$$\|A\| \leq \rho(A) + \epsilon.$$

References

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

[2] G. Goh. Why momentum really works. *Distill*, 2017. doi: 10.23915/distill.00006. URL http://distill.pub/2017/momentum.

[3] I. Mitliagkas. An interesting property of polyak's momentum. Course slides, 2019. URL http://mitliagkas.github.io/ift6085-2019/ift-6085-lecture-5-slides.pdf.

[4] Y. Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 1998.

[5] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

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