Why Momentum Really Works

Summary Notes by Max Guo

July 10, 2022

- Note: this is a tutorial article, so the format may be a little different than previous articles.

1 Information

- **Year**: 2017
- **Journal**: Distill
- **Author**: Gabriel Goh

2 Main Idea

The main idea of this tutorial is understanding the mechanism of momentum in gradient descent through the example of the convex quadratic.

3 Gradient Descent

Consider the convex quadratic:

$$\begin{aligned} f(w) &= \frac{1}{2}w^T Aw - b^T w \\ \implies \nabla f(w) &= Aw - b \\ \implies w^{k+1} &= w^k - \alpha(Aw^k - b) \end{aligned}$$

where the last equation is the gradient descent update step. Assume A is symmetric and invertible, so the solution is $w^* = A^{-1}b$. Now decompose A:

$$\begin{aligned} A &= Q\Lambda Q^T, \\ Q &= [q_1, \dots, q_n] \\ \Lambda &= diag(\lambda_1, \dots, \lambda_n),\ \lambda_1 \leq \dots \leq \lambda_n \end{aligned}$$

Changing the basis, so $x^k = Q^T(w^k - w^*)$:

$$\begin{aligned} w^k - w^* &= \sum_{i=1}^{n} x_i^0(1 - \alpha\lambda_i)^k q_i \\ f(w^k) - f(w^*) &= \sum_{i=1}^{n} (1 - \alpha\lambda_i)^{2k}\lambda_i[x_i^0]^2 \end{aligned}$$

Interpretation: Initial error in the Q-basis decomposed into n errors, which each decrease exponentially at rate $1 - \alpha\lambda_i$.

Convergence: For convergence, we need $|1 - \alpha\lambda_i| < 1 \iff 0 < \alpha\lambda_i < 2$ for each i; it suffices to look at the smallest and largest is. The rate is determined by $\max\{|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n|\}$. The optimal (minimal) rate is when these are equal, e.g.

$$\begin{aligned} \alpha &= \frac{2}{\lambda_1 + \lambda_n} \\ \text{rate} &= \frac{\lambda_n/\lambda_1 - 1}{\lambda_n/\lambda_1 + 1} = \frac{\kappa - 1}{\kappa + 1} \end{aligned}$$

where κ is the condition number of A. The larger the condition number, the lower the optimal rate.

4 Momentum

The general momentum update:

$$\begin{aligned} z^{k+1} &= \beta z^k + \nabla f(w^k) \\ w^{k+1} &= w^k - \alpha z^{k+1} \end{aligned}$$

For $\nabla f(w^k) = Aw^k - b$ and change of basis $x^k = Q(w^k - w^*)$ and $y^k = Qz^k$, the update rule becomes:

$$\begin{aligned} y_i^{k+1} &= \beta y_i^k + \lambda_i x_i^k \\ x_i^{k+1} &= x_i^k - \alpha y_i^{k+1} \end{aligned}$$

or:

$$\begin{pmatrix} y_i^k \\ x_i^k \end{pmatrix} = R^k \begin{pmatrix} y_i^0 \\ x_i^0 \end{pmatrix}$$

where

$$R = \begin{pmatrix} \beta & \lambda_i \\ -\alpha\beta & 1 - \alpha\lambda_i \end{pmatrix}$$

For 2×2 matrices, there is an elegant formula for R^k in terms of the eigenvalues of σ_1 and σ_2. After some algebra, the convergence rate is $\max\{|\sigma_1|, |\sigma_2|\}$, and the convergence criterion is that this quantity is < 1. This gives us distinct regions for different convergence behavior, shown in fig. 1.

The range of step sizes is $0 < \alpha\lambda_i < 2 + 2\beta$, for $0 \leq \beta < 1$. The optimum parameters and convergence rate turn out to be:

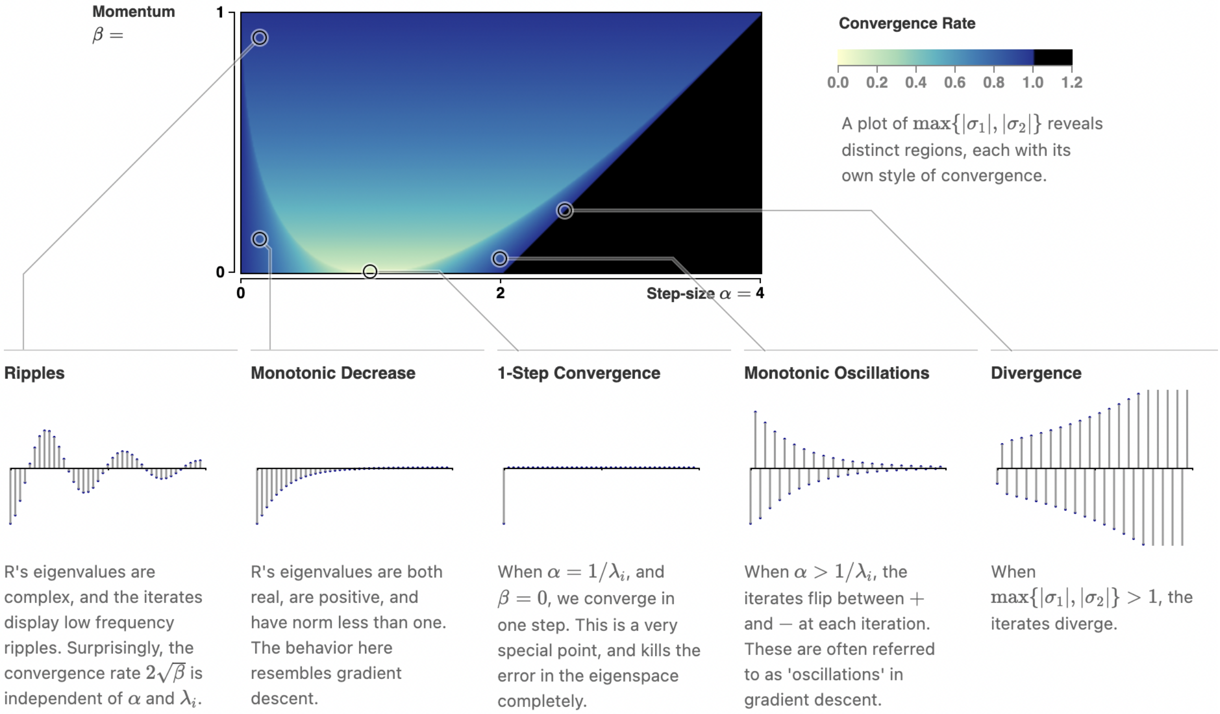
$$\begin{aligned} \alpha &= \left(\frac{2}{\sqrt{\lambda_1} + \sqrt{\lambda_n}}\right)^2 \\ \beta &= \left(\frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}}\right)^2 \\ \text{rate} &= \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \end{aligned}$$

Practical Notes: if the problem's conditioning is poor, optimal α is about twice in momentum than in gradient descent. Moreover, $\beta \approx 1$, so set β as high as possible then find the largest α that converges.

4.1 Example: Colorization Problem

Colorization Problem: On a graph G with edges E and distinguished set of vertices D, minimize

$$\frac{1}{2}\sum_{i \in D}(w_i - 1)^2 + \frac{1}{2}\sum_{i,j \in E}(w_i - w_j)^2 \tag{1}$$



A plot of $\max\{|\sigma_1|, |\sigma_2|\}$ reveals distinct regions, each with its own style of convergence.

R's eigenvalues are complex, and the iterates display low frequency ripples. Surprisingly, the convergence rate $2\sqrt{\beta}$ is independent of α and λ_i.

R's eigenvalues are both real, are positive, and have norm less than one. The behavior here resembles gradient descent.

When $\alpha = 1/\lambda_i$, and $\beta = 0$, we converge in one step. This is a very special point, and kills the error in the eigenspace completely.

When $\alpha > 1/\lambda_i$, the iterates flip between $+$ and $-$ at each iteration. These are often referred to as 'oscillations' in gradient descent.

When $\max\{|\sigma_1|, |\sigma_2|\} > 1$, the iterates diverge.

Figure 1: Momentum Dynamics

where the optimal solution is $\vec{w} = 1$. Gradient descent results in every value being updated as a weighted average of current value and its neighbors. In vectorized form, this problem is, minimize:

$$\frac{1}{2}x^T L_G x + \frac{1}{2}\sum_{i \in D} x^T e_i e_i^T x - e_i^T x \tag{2}$$

where L_G is the Laplacian matrix. The condition number of L_G is dependent on the connectivity of the graph. (e.g. long wiry graphs have poor conditioning).

4.2 Limitations

We can unroll the gradient descent loop and write the algorithm as:

$$w^{k+1} = w^0 + \sum_{i}^{k} \Gamma_i^k \nabla f(w^i)$$

where Γ_i^k are diagonal matrices. This describes gradient descent, gradient descent with momentum, Adam, Adagrad, etc. If we consider the colorizable problem with the (really badly conditioned) graph that is just a single path. Then it turns out momentum achieves the best we can do on this problem as $n \to \infty$.

4.3 Stochastic Gradient Descent

Stochastic gradient descent with momentum has tradeoffs (e.g. increasing step size results in compounding errors vs. increasinag rate of convergence), but shown to be competitive on NNs. Noise could be an implicit regularizer?