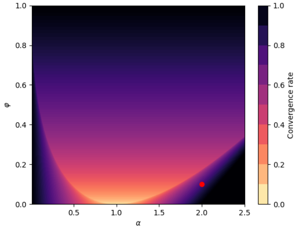
Optimization in Machine Learning

First order methods
Momentum on quadratic forms



Learning goals

- Momentum update in Eigenspace
- Effect of φ

RECAP: MOMENTUM UPDATE

$$
\begin{aligned}
\boldsymbol{\nu}^{[t+1]} &= \varphi \boldsymbol{\nu}^{[t]} + \alpha \nabla f(\mathbf{x}^{[t]}) \\
\mathbf{x}^{[t+1]} &= \mathbf{x}^{[t]} - \boldsymbol{\nu}^{[t+1]},
\end{aligned}
$$

which simplifies to

$$
\begin{aligned}
\boldsymbol{\nu}^{[t+1]} &= \varphi \boldsymbol{\nu}^{[t]} + \alpha (\mathbf{A}\mathbf{x}^{[t]} - b) \\
\mathbf{x}^{[t+1]} &= \mathbf{x}^{[t]} - \boldsymbol{\nu}^{[t+1]},
\end{aligned}
$$

for the quadratic form.

RECAP: DYNAMICS OF MOMENTUM

Changing the basis as before with $\boldsymbol{w}^{[t]} = \boldsymbol{V}^\top(\mathbf{x}^{[t]} - \mathbf{x}^*)$ and $\boldsymbol{u}^{[t]} = \boldsymbol{V}\boldsymbol{\nu}^{[t]}$, we get the following set of equations, where each component acts independently, although $w_i^{[t]}$ and $u_i^{[t]}$ are coupled:

$$
\begin{aligned}
u_i^{[t+1]} &= \varphi u_i^{[t]} + \alpha\lambda_i w_i^{[t]}, \\
w_i^{[t+1]} &= w_i^{[t]} - u_i^{[t+1]}
\end{aligned}
$$

We rewrite this:

$$
\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u_i^{[t+1]} \\ w_i^{[t+1]} \end{pmatrix} = \begin{pmatrix} \varphi & \alpha\lambda_i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u_i^{[t]} \\ w_i^{[t]} \end{pmatrix}
$$

and invert the matrix on the LHS:

$$
\begin{pmatrix} u_i^{[t+1]} \\ w_i^{[t+1]} \end{pmatrix} = \begin{pmatrix} \varphi & \alpha\lambda_i \\ -\varphi & 1 - \alpha\lambda_i \end{pmatrix} \begin{pmatrix} u_i^{[t]} \\ w_i^{[t]} \end{pmatrix} = R^{t+1} \begin{pmatrix} u_i^0 \\ w_i^0 \end{pmatrix}
$$

RECAP: DYNAMICS OF MOMENTUM / 2

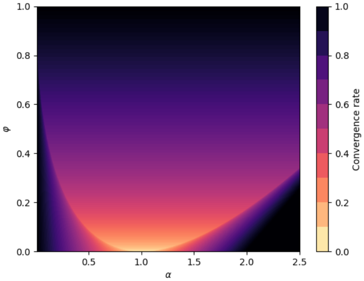
Taking a 2×2 matrix to the t^{th} power reduces to a formula involving the eigenvalues of R, σ_1 and σ_2:

$$R^t = \begin{cases} \sigma_1^t R_1 - \sigma_2^t R_2, & \text{if } \sigma_1 \neq \sigma_2 \\ \sigma_1^t (tR/\sigma_1 - (t-1)I), & \text{if } \sigma_1 = \sigma_2 \end{cases}$$

where $R_j = \frac{R - \sigma_j I}{\sigma_1 - \sigma_2}$.

In contrast to GD, where we got one geometric series, we have two coupled series with real or complex values.

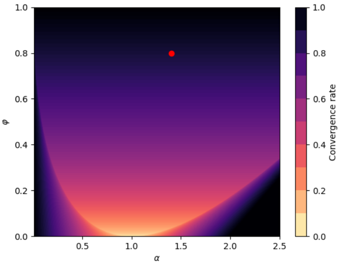
RECAP: DYNAMICS OF MOMENTUM / 3

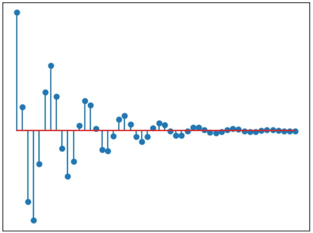


The achieved convergence rate is therefore the slowest of the two, $\max\{|\sigma_1|, |\sigma_2|\}$.

Each region shows a different convergence behavior.

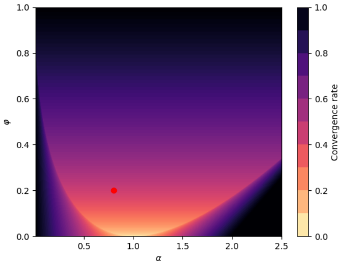
RECAP: DYNAMICS OF MOMENTUM / 4

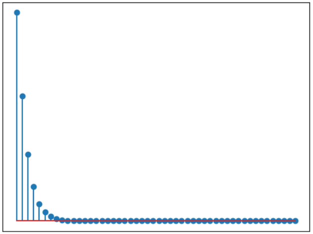




The eigenvalues of R are complex and we see low frequency ripples.

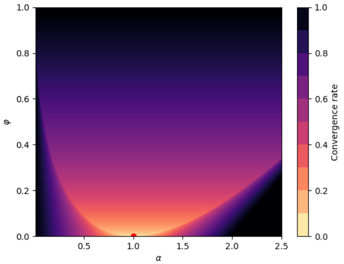
RECAP: DYNAMICS OF MOMENTUM / 5

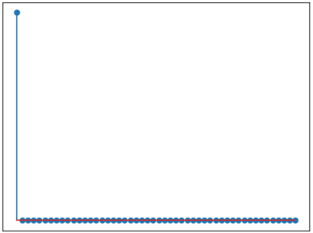




Here, both eigenvalues of R are positive with their norm being less than 1. This behavior resembles gradient descent.

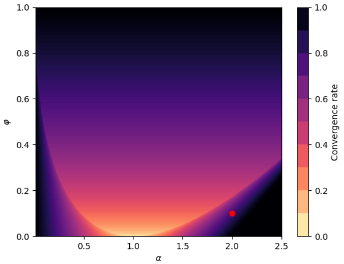
RECAP: DYNAMICS OF MOMENTUM / 6

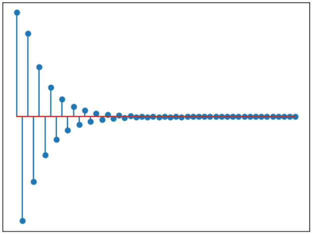




The step size is $\alpha = 1/\lambda_i$ and $\varphi = 0$ - we converge in one step.

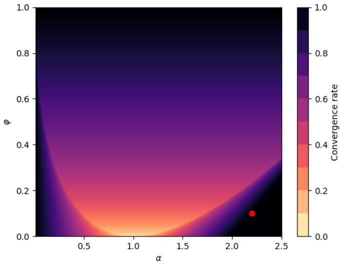
RECAP: DYNAMICS OF MOMENTUM / 7

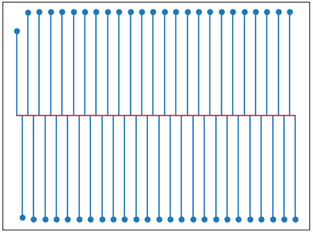




When $\alpha > 1/\lambda_i$, the iterates flip sign every iteration.

RECAP: DYNAMICS OF MOMENTUM / 8

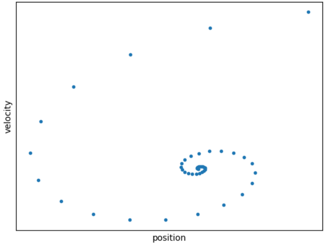




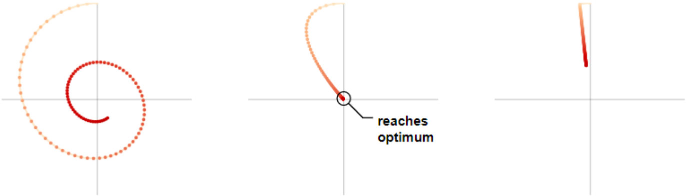
If $\max\{|\sigma_1|, |\sigma_2|\} > 1$, the iterates diverge.

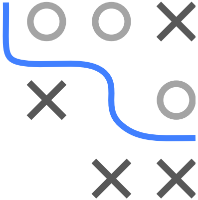
RECAP: DYNAMICS OF MOMENTUM / 9

- Finally, we investigate the role of φ.
- We can think of gradient descent with momentum as a damped harmonic oscillator: a weight on a spring. We pull the weight down and study the path back to the equilibrium in phase space (looking at the position and the velocity).
- Depending on the choice of φ, the rate of return to the equilibrium position is affected.



RECAP: DYNAMICS OF MOMENTUM / 10





Left: If φ is too large, we are underdamping. The spring oscillates back and forth and misses the optimum.
Middle: The best value of φ lies in the middle.

Right: If φ is too small, we are overdamping, meaning that the spring experiences too much friction and stops before reaching the equilibrium.