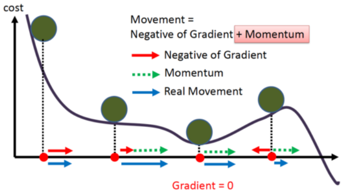
Optimization in Machine Learning

First order methods: GD on quadratic forms



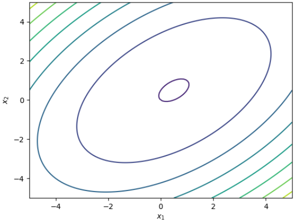
Learning goals

- Definition
- Max. Likelihood
- Normal regression
- Risk Minimization

GD ON QUADRATIC FORMS

- We consider the quadratic function $q(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{A}\mathbf{x} - \boldsymbol{b}^\top \mathbf{x}$.
- We assume that the Hessian $\boldsymbol{H} = \mathbf{A}$ is symmetric and invertible
- The optimal solution is $\mathbf{x}^* = \mathbf{A}^{-1} b$
- As $\nabla q(\mathbf{x}) = \mathbf{A}\mathbf{x} - \boldsymbol{b}$, the iterations of gradient descent are

$$\mathbf{x}^{[t+1]} = \mathbf{x}^{[t]} - \alpha(\mathbf{A}\mathbf{x}^{[t]} - \boldsymbol{b})$$



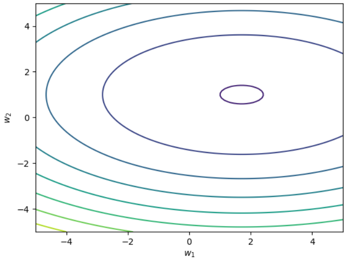
The following slides follow the blogpost by Goh, "Why Momentum Really Works", Distill, 2017. `http://doi.org/10.23915/distill.00006`

GD ON QUADRATIC FORMS

For $\mathbf{A}$, there exists an eigenvalue decomposition:

$$\mathbf{A} = \boldsymbol{V} \boldsymbol{\Lambda} \boldsymbol{V}^\top$$

where the columns of $\boldsymbol{V}$ contain the eigenvectors $\boldsymbol{v}_i$ and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_n)$ contains the eigenvalues λ_i sorted from smallest to biggest. We perform a change of basis $\boldsymbol{w}^{[t]} = \boldsymbol{V}^\top(\mathbf{x}^{[t]} - \mathbf{x}^*)$ to its eigenspace, where all dimensions act independently.



GD ON QUADRATIC FORMS

We get

$$\begin{aligned} \mathbf{x}^{[t+1]} &= \mathbf{x}^{[t]} - \alpha \left(\mathbf{A}\mathbf{x}^{[t]} - \boldsymbol{b} \right) \\ \boldsymbol{w}^{[t]} &= \boldsymbol{V}^\top (\mathbf{x}^{[t]} - \mathbf{x}^*) \\ \boldsymbol{V} \cdot \boldsymbol{w}^{[t]} + \mathbf{x}^* &= \mathbf{x}^{[t]} \end{aligned}$$

and with $\mathbf{x}^{[t]} = \boldsymbol{V} \cdot \boldsymbol{w}^{[t]} + \mathbf{x}^*$ we can write a GD step as:

GD ON QUADRATIC FORMS

$$
\begin{aligned}
\mathbf{x}^{[t+1]} &= \mathbf{x}^{[t]} - \alpha \left(\mathbf{A}\mathbf{x}^{[t]} - \boldsymbol{b} \right) \\
\boldsymbol{V} \cdot \boldsymbol{w}^{[t+1]} + \mathbf{x}^* &= \boldsymbol{V} \cdot \boldsymbol{w}^{[t]} + \mathbf{x}^* - \alpha \left(\mathbf{A} \cdot \boldsymbol{V} \cdot \boldsymbol{w}^{[t]} + \mathbf{A} \cdot \mathbf{x}^* - \boldsymbol{b} \right) && \Big| \; -\mathbf{x}^* \\
\boldsymbol{V} \cdot \boldsymbol{w}^{[t+1]} &= \boldsymbol{V} \cdot \boldsymbol{w}^{[t]} - \alpha \left(\mathbf{A} \cdot \boldsymbol{V} \cdot \boldsymbol{w}^{[t]} + \mathbf{A}\mathbf{x}^* - \boldsymbol{b} \right) && \Big| \; \boldsymbol{V}^\top (\text{NB: } \boldsymbol{V}^\top \boldsymbol{V} = \mathbf{1}) \\
\boldsymbol{w}^{[t+1]} &= \boldsymbol{w}^{[t]} - \alpha \boldsymbol{V}^\top \left(\mathbf{A}\boldsymbol{V} \cdot \boldsymbol{w}^{[t]} + \mathbf{A}\mathbf{x}^* - \boldsymbol{b} \right) && \Big| \; \mathbf{A}\mathbf{x}^* - \boldsymbol{b} = 0 \\
\boldsymbol{w}^{[t+1]} &= \boldsymbol{w}^{[t]} - \alpha \boldsymbol{V}^\top \mathbf{A}\boldsymbol{V} \cdot \boldsymbol{w}^{[t]} && \Big| \; \boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V} = \boldsymbol{\Sigma} \\
\boldsymbol{w}^{[t+1]} &= \boldsymbol{w}^{[t]} - \alpha \boldsymbol{\Sigma} \boldsymbol{w}^{[t]}
\end{aligned}
$$

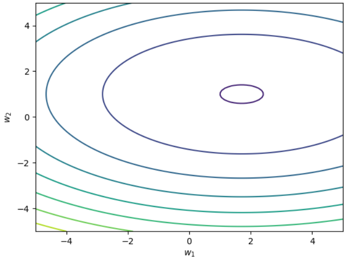
Which means:

$$
w_i^{[t+1]} = w_i^{[t]} - \alpha \lambda_i w_i^{[t]} = (1 - \alpha \lambda_i) w_i^{[t]} = (1 - \alpha \lambda_i)^{t+1} w_i^{[0]}
$$

GD ON QUADRATIC FORMS

If we now perform GD on $\boldsymbol{w}$, we get

$$\begin{aligned} w_i^{[t+1]} &= w_i^{[t]} - \alpha \lambda_i w_i^{[t]} \\ &= (1 - \alpha \lambda_i) w_i^{[t]} = (1 - \alpha \lambda_i)^{t+1} w_i^{[0]} \end{aligned}$$



GD ON QUADRATIC FORMS

Moving back to the original space, we get

$$\mathbf{x}^{[t]} - \mathbf{x}^* = \boldsymbol{V} \cdot \boldsymbol{w}^{[t]} = \sum_{i=1}^{d} w_i^{[0]} (1 - \alpha\lambda_i)^t \boldsymbol{v}_i$$

This allows a very intuitive interpretation: each element of $w^{[0]}$ is the component of the error in the initial guess in the eigenbasis and decays with a rate of $1 - \alpha\lambda_i$.

For most step sizes, the eigenvectors with the largest eigenvalues converge the fastest.

GD ON QUADRATIC FORMS

We now consider the contribution of each eigenvector to the total loss

$$q(\mathbf{x}^{[t]}) - q(\mathbf{x}^*) = \frac{1}{2}\sum_i^d (1 - \alpha\lambda_i)^{2t}\lambda_i(w_i^{[0]})^2$$