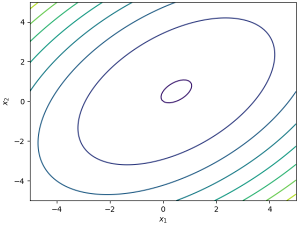
Optimization in Machine Learning

First order methods: GD on quadratic forms



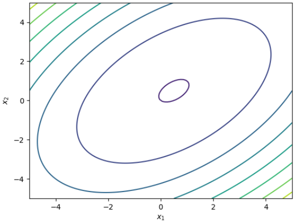
Learning goals

- Eigendecomposition of quadratic forms
- GD steps in eigenspace

QUADRATIC FORMS & GD

- We consider the quadratic function $q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$.
- We assume that Hessian $\mathbf{H} = 2\mathbf{A}$ has full rank
- Optimal solution is $\mathbf{x}^* = \frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$
- As $\nabla q(\mathbf{x}) = 2\mathbf{A}\mathbf{x} - \mathbf{b}$, iterations of gradient descent are

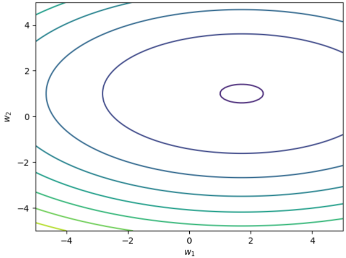
$$\mathbf{x}^{[t+1]} = \mathbf{x}^{[t]} - \alpha(2\mathbf{A}\mathbf{x}^{[t]} - \mathbf{b})$$



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

EIGENDECOMPOSITION OF QUADRATIC FORMS

- We want to work in the coordinate system given by q
- **Recall:** Coordinate system is given by the eigenvectors of $\mathbf{H} = 2\mathbf{A}$
- Eigendecomposition of $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$
- $\mathbf{V}$ contains eigenvectors $\mathbf{v}_i$ and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_n)$ eigenvalues
- Change of basis: $\mathbf{w}^{[t]} = \mathbf{V}^\top(\mathbf{x}^{[t]} - \mathbf{x}^*)$



GD STEPS IN EIGENSPACE

With $\mathbf{w}^{[t]} = \mathbf{V}^\top(\mathbf{x}^{[t]} - \mathbf{x}^*)$, a single GD step

$$\mathbf{x}^{[t+1]} = \mathbf{x}^{[t]} - \alpha(2\mathbf{A}\mathbf{x}^{[t]} - \mathbf{b})$$

becomes

$$\mathbf{w}^{[t+1]} = \mathbf{w}^{[t]} - 2\alpha\mathbf{\Lambda}\mathbf{w}^{[t]}.$$

Therefore:

$$\begin{aligned} w_i^{[t+1]} &= w_i^{[t]} - 2\alpha\lambda_i w_i^{[t]} \\ &= (1 - 2\alpha\lambda_i) w_i^{[t]} \\ &= \cdots \\ &= (1 - 2\alpha\lambda_i)^{t+1} w_i^{[0]} \end{aligned}$$

GD STEPS IN EIGENSPACE

Proof (for $\mathbf{w}^{[t+1]} = \mathbf{w}^{[t]} - 2\alpha\mathbf{\Lambda}\mathbf{w}^{[t]}$)**:**

- A single GD step means

$$\mathbf{x}^{[t+1]} = \mathbf{x}^{[t]} - \alpha(2\mathbf{A}\mathbf{x}^{[t]} - \mathbf{b})$$

- Then:

$$\begin{aligned}
\mathbf{V}^\top(\mathbf{x}^{[t+1]} - \mathbf{x}^*) &= \mathbf{V}^\top(\mathbf{x}^{[t]} - \mathbf{x}^*) - \alpha\mathbf{V}^\top(2\mathbf{A}\mathbf{x}^{[t]} - \mathbf{b}) \\
\mathbf{w}^{[t+1]} &= \mathbf{w}^{[t]} - \alpha\mathbf{V}^\top(2\mathbf{A}\mathbf{x}^{[t]} - \mathbf{b}) \\
\mathbf{w}^{[t+1]} &= \mathbf{w}^{[t]} - \alpha\mathbf{V}^\top(2\mathbf{A}(\mathbf{x}^{[t]} - \mathbf{x}^*) + \underbrace{2\mathbf{A}\mathbf{x}^* - \mathbf{b}}_{=0}) \\
&= \mathbf{w}^{[t]} - 2\alpha\mathbf{\Lambda}\mathbf{V}^\top(\mathbf{x}^{[t]} - \mathbf{x}^*) \\
&= \mathbf{w}^{[t]} - 2\alpha\mathbf{\Lambda}\mathbf{w}^{[t]}
\end{aligned}$$

GD ERROR IN ORIGINAL SPACE

- Move back to **original space**:

$$\mathbf{x}^{[t]} - \mathbf{x}^* = \mathbf{V}\mathbf{w}^{[t]} = \sum_{i=1}^{d} (1 - 2\alpha\lambda_i)^t w_i^{[0]} \mathbf{v}_i$$

- **Intuition:** Initial error components $w_i^{[0]}$ (in the eigenbasis) decay with rate $1 - 2\alpha\lambda_i$
- **Therefore:** For sufficiently small step sizes α, error components along eigenvectors with large eigenvalues decay quickly

GD ERROR IN ORIGINAL SPACE

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

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