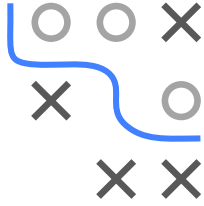
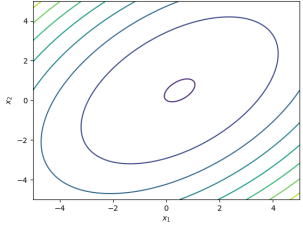
Optimization in Machine Learning



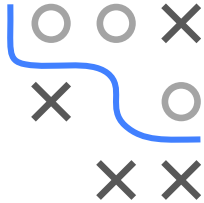
First order methods
GD on quadratic forms



Learning goals

- Eigendecomposition of quadratic forms
- GD steps in eigenspace

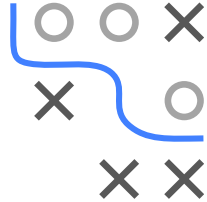
QUADRATIC FORMS & GD



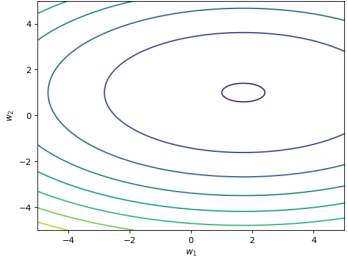
- We consider the quadratic function $q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$.
- We assume that Hessian $\mathbf{H} = 2\mathbf{A}$ has full rank
- We assume q is convex, so A psd.
- Optimal solution is $\mathbf{x}^* = \frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$
-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

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

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 $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$
- $\mathbf{V}$ contains eigenvectors $\mathbf{v}_i$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_n)$ eigenvalues
- Change of basis: $\mathbf{w}^{[t]} = \mathbf{V}^\top(\mathbf{x}^{[t]} - \mathbf{x}^*)$



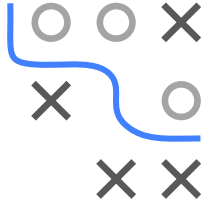
GD STEPS IN EIGENSPACE

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

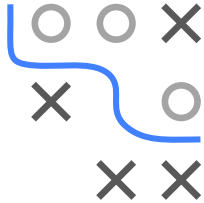
$$\mathbf{w}^{[t+1]} = \mathbf{w}^{[t]} - 2\alpha\mathbf{\Lambda}\mathbf{w}^{[t]}.$$



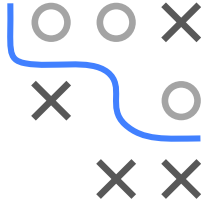
GD STEPS IN EIGENSPACE

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 ERROR IN ORIGINAL SPACE

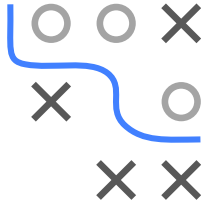


- Move back to **original space**:

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

- This gives us the difference in x-space in closed form
- The difference vector is written in the eigen-basis, the initial error components $w_i^{[0]}$ decay with rate $1 - 2\alpha\lambda_i$,
- The smaller this term (in abs value), the faster we decay, the closer to 1, the slower we are
- For most step-sizes, eigenvectors with largest eigenvalues converge fastest. Works especially well in the first iterations.

OPTIMAL STEPSIZE



- In order to converge, must have all : $|1 - 2\alpha\lambda_i| < 1$
- So for all stepsizes $0 < \alpha\lambda_i < 1$
- Overall rate determined by slowest error component:

$$\max_{i=1,\ldots n} |1 - 2\alpha\lambda_i| = \max(|1 - 2\alpha\lambda_1|, |1 - 2\alpha\lambda_n|)$$

- Minimized for $\alpha = \frac{1}{\lambda_1 + \lambda_n}$
- Optimal rate $\frac{\lambda_n/\lambda_1 - 1}{\lambda_n/\lambda_1 + 1}$
- We again see that convergence is determined by condition

$$\kappa = \frac{\lambda_n}{\lambda_1}$$

- $\kappa = 1$ Is ideal, convergence in 1 step