OPTIMIZATION ALGORITHMS

<u>Aim</u>: Implement an algorithm to obtain an approximated solution to the problem

$$\min_{w \in \mathbb{R}^d} L(w)$$

<u>Standing assumption</u>: Hereafter, we will assume that L is at least of class C^1, so that we can compute its gradient $\nabla L(w)$ for every $w \in \mathbb{R}^d$.

GRADIENT DESCENT

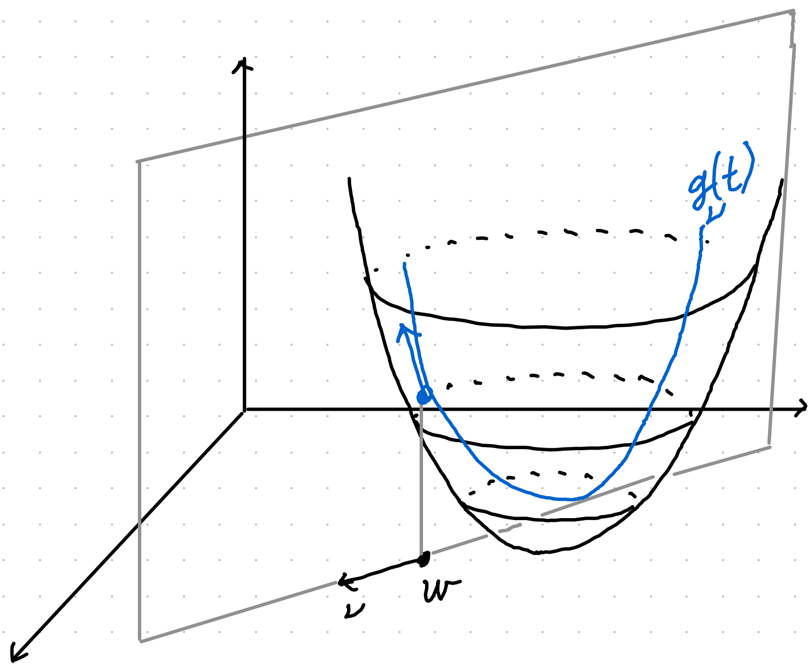
The algorithm of gradient descent stems from the following observation:

<u>Remark</u>: The gradient $\nabla L(w_0)$ has the direction along which the function L grows most rapidly.

Too see this, consider a generic direction $\nu \in \mathbb{R}^d$, $|\nu| = 1$

Consider the section of the function L along the direction v

$$g_v(t) = L(w_0 + tv)$$



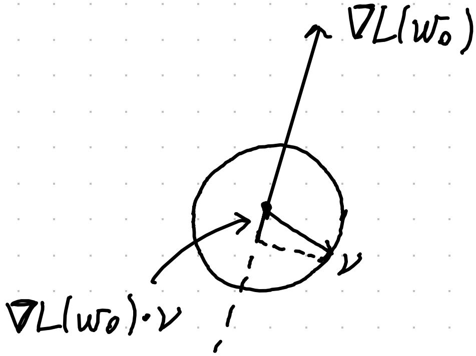
The amount of increase of g_v at $t=0$ is computed in terms of the derivative

$$\frac{d}{dt} g_v(t)\Big|_{t=0} = \frac{d}{dt} L(w_0 + tv)\Big|_{t=0} =$$
$$= \nabla L(w_0) \cdot v$$

We want to find ν such that this quantity is maximized

$$\max_{\nu \in \mathbb{R}^d, |\nu|=1} \nabla L(w_0) \cdot \nu = |\nabla L(w_0)|$$

attained for $\nu = \frac{\nabla L(w_0)}{|\nabla L(w_0)|}$



Analogously,

$$\min_{\nu \in \mathbb{R}^d, |\nu|=1} \nabla L(w_0) \cdot \nu = -|\nabla L(w_0)|$$

attained for $\nu = -\frac{\nabla L(w_0)}{|\nabla L(w_0)|}$

The Gradient Descent algorithm (GD) follows steps along the direction where the functions decreases faster.

<u>Algorithm</u>:

- choose $w^0 \in \mathbb{R}^d$ initial guess
- choose $\tau > 0$ step-size (learning rate)
- assume $w^k \in \mathbb{R}^d$ is defined for $k \geq 0$
- set

$$w^{k+1} = w^k - \tau \nabla L(w^k)$$

See example on notebook for implementation.

<u>Remark</u>: The gradient descent algorithm is a discrete version of a gradient flow. Imagine τ is a time step. Then the algorithm is written as

$$\underbrace{\frac{w^{k+1} - w^k}{\tau}} = - \nabla L(w^k)$$

this is basically a discrete time derivative. Imagining that w^k are the discretization of a curve $w(t)$, this reads

$$\dot{w}(t) = -\nabla L(w(t))$$

Note that L decreases on solutions:

$$\frac{d}{dt} L(w(t)) = \nabla L(w(t)) \cdot \dot{w}(t) =$$
$$= -|\nabla L(w(t))|^2 \leq 0$$

For the discrete algorithm, one should be careful about the choice of the learning rate τ.

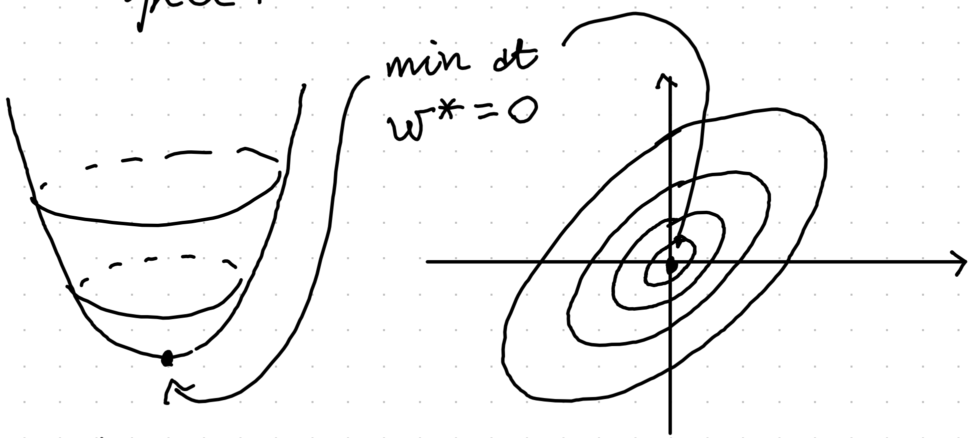
See examples on Python notebook.

To understand better how the algorithm is behaving, let us study it explicitly on a prototypical example of function:

$$L(w) = \frac{1}{2}(Aw) \cdot w$$

where A is a symmetric and positive definite function.

These are functions with this aspect:



3d plot,
a paraboloid

Contour plot:
level sets are ellipses

To study this function, it is convenient to change frame of reference. To do so, we diagonalize the matrix A.
Every symmetric matrix can be diagonalized and has real eigenvalues. Since it is positive definite, the eigenvalues are also positive.

$$A = QDQ^T \qquad D = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots & \\ & & & \lambda_d \end{pmatrix}, \ \lambda_i > 0.$$

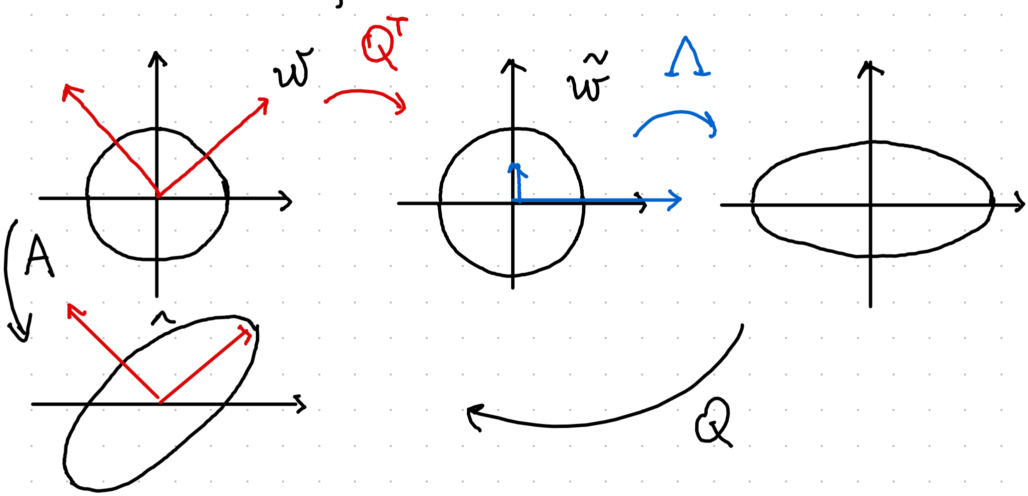
Moreover, $Q \in O(d)$, i.e., its columns are an orthonormal basis.

What is the meaning of diagonalization?

Change coordinates:

$$\tilde{w} = Q^T w$$

How does the function look like in the new set of coordinates?



$$L(w) = \frac{1}{2} w^T A w = \frac{1}{2} (Q\tilde{w})^T A (Q\tilde{w}) =$$

$$= \frac{1}{2} \tilde{w}^T (Q^T A Q) \tilde{w} = \frac{1}{2} \tilde{w}^T D \tilde{w} =$$

$$= \frac{1}{2} \tilde{w}^T \begin{pmatrix} \lambda_1 \tilde{w}_1 \\ \vdots \\ \lambda_d \tilde{w}_d \end{pmatrix} = \frac{1}{2} \lambda_1 \tilde{w}_1^2 + \dots + \frac{1}{2} \lambda_d \tilde{w}_d^2$$

In these new coordinates, GD is easy to interpret:

$$w^{k+1} = w^k - \tau \nabla L(w^k) =$$
$$= w^k - \tau A w^k$$

$$\nabla_w L(w) = \nabla_w \left(\frac{1}{2} w^T A w\right) = Aw$$

$$Q^T w^{k+1} = Q^T w^k - \tau Q^T A w^k$$
$$\tilde{w}^{k+1} = \tilde{w}^k - \tau Q^T A Q \tilde{w}^k$$
$$= \tilde{w}^k - \tau D \tilde{w}^k$$

All these equations are decoupled:

$$\left\{ \tilde{w}_i^{k+1} = \tilde{w}_i^k - \tau \lambda_i \tilde{w}_i^k \qquad i = 1, \dots, d \right.$$

Hence, GD is implemented on each component $\tilde{w}_i^k$ and they are all decoupled!

Let's look at one equation at a time.

Let us study the continuous-time equivalent:

$$\dot{\tilde{w}}_i(t) = -\lambda_i \tilde{w}_i(t)$$

We know the explicit solution to this problem:

$$\tilde{w}_i(t) = \tilde{w}_i(0) e^{-\lambda_i t}.$$

It converges to zero exponentially fast.

However, in the continuous-time case, we do not see the time step.

Let us study the discrete equation.

$$
\begin{aligned}
\tilde{w}^{k+1} &= \tilde{w}^{k} - \tau \lambda_i \tilde{w}_i^{k} = \\
&= (1-\tau\lambda_i)\, \tilde{w}_i^{k} = \\
&= \dots = \\
&= (1-\tau\lambda_i)^{k+1}\, \tilde{w}_i^{0}
\end{aligned}
$$

In conclusion

$$
\tilde{w}^{k} = \underbrace{(1-\tau\lambda_i)^{k}}\; \underbrace{\tilde{w}_i^{0}}
$$

Hence, the components at the initial guess measure the initial error

which is damped with this power.

For the dampening to work, we need

$$
|1-\tau\lambda_i| < 1 \quad \text{for all } i = 1, \dots, d .
$$

If we order $\lambda_1 \leq \dots \leq \lambda_d$, then it is enough to require

$$|1-\tau\lambda_1| < 1, \quad |1-\tau\lambda_d| < 1$$

$$-1 < 1-\tau\lambda_1 < 1, \quad -1 < 1-\tau\lambda_d < 1$$

$$0 < \tau\lambda_1 < 2, \quad 0 < \tau\lambda_d < 2$$

$$0 < \tau\lambda_1 \leq \tau\lambda_d < 2$$

Hence $0 < \tau < \frac{2}{\lambda_d}$.

If λ_d is very large, τ must be taken very small.

We can also compute the optimal rate:

$$\text{rate}(\tau) = \max\{|1-\tau\lambda_1|, |1-\tau\lambda_d|\}$$

$$\text{rate}(\tau^*) = \min_{\tau} \text{rate}(\tau)$$

reached for $|1-\tau^*\lambda_1| = |1-\tau^*\lambda_d|$

$\max\{a,b\} = \max\{a-b, 0\} + b$

This means

- $1 - \tau^* \lambda_1 = 1 - \tau^* \lambda_d \iff \lambda_1 = \lambda_d$
  ($L(w) = \frac{1}{2} \lambda |w|^2$ and convergence happens in 1 step choosing $\tau^* = \frac{1}{\lambda}$)

or

- $-1 + \tau^* \lambda_1 = 1 - \tau^* \lambda_d \iff$

$$\iff \tau^* = \frac{2}{\lambda_d + \lambda_1}$$

$$\text{rate}(\tau^*) = |1 - \tau^* \lambda_1| = \left| 1 - \frac{2}{\lambda_d + \lambda_1} \lambda_1 \right| =$$

$$= \left| \frac{\lambda_d + \lambda_1 - 2\lambda_1}{\lambda_d + \lambda_1} \right| = \frac{\lambda_d - \lambda_1}{\lambda_d + \lambda_1} =$$

$$= \frac{\lambda_d / \lambda_1 - 1}{\lambda_d / \lambda_1 + 1} = \frac{k(A) - 1}{k(A) + 1}$$

This number $k(A) = \frac{\lambda_d}{\lambda_1}$ has a meaning: it's the condition number. It measures how much the matrix A is far from being invertible. (A big condition number is bad).

Thanks to the previous analysis, we are ready to prove a result.

We study the quadratic case because it's the prototype of curvature. Curvature in the graph of a function is measured in terms of the second derivatives — the Hessian $D^2L(w^*)$.

Hence we will make assumptions on the eigenvalues of the Hessian.
Asking bounds for the minimal and maximal eigenvalues of the Laplacian means

change coordinates to see it

$$\lambda|\xi|^2 \leq \xi^T D^2L(w^*)\xi \leq \Lambda|\xi|^2$$

<u>Remark</u>: If the function is only of class C^1, these can be replaced by weaker conditions on the Lipschitz continuity of L and the convexity of L. I think it's clearer if we assume this.

For notation simplicity, we write
$0 < \lambda \le D^2 L(w) \le \Lambda$.

<u>Remark</u>: Under the previous assumption, the function L has a unique minimum.

Indeed, by Taylor's formula:

$$L(w) = L(w_0) + \nabla L(w_0)\cdot(w - w_0) + \frac{1}{2}(w-w_0)^T D^2 L(\tilde{w})\cdot(w-w_0)$$

$$\frac{\lambda}{2}|w-w_0|^2 \le L(w) - L(w_0) - \nabla L(w_0)\cdot(w-w_0) \le$$
$$\le \frac{\Lambda}{2}|w-w_0|^2$$

To show that there exists a minimum, we observe that

$$L(w) \ge L(0) + \nabla L(0)\cdot w + \frac{\lambda}{2}|w|^2$$

Note that

$$-\nabla L(0)\cdot w \le \frac{1}{2\varepsilon}|\nabla L(0)|^2 + \frac{\varepsilon}{2}|w|^2$$

Choosing ε small enough,

$$L(w) \geq -c_1 + c_2 |w|^2, \quad c_1, c_2 > 0,$$

i.e., L is over a paraboloid.

Fix $M > 0$ attained by L. Then $\inf L \leq M$.

Note that the set $\{L \leq M\}$ is contained in a ball, since

$$-c_1 + c_2 |w|^2 \leq L(w) \leq M \Rightarrow$$

$$\Rightarrow |w|^2 \leq \frac{M + c_1}{c_2} \Rightarrow |w| \leq \frac{M + c_1}{c_2} =: R$$

The fuction L has a minimum in this closed ball, w^*:

$$L(w^*) = \min_{|w| \leq R} L(w)$$

This is also a minimum on the whole $\mathbb{R}^d$, because, outside the ball, $L > M \geq L(w^*)$.

The minimum point is unique.

Assume that w_1 and w_2 are two minima.

$$L(w_1) \geq L(w_2) + \underbrace{\nabla L(w_1)}_{=0 \text{ on minima}} \cdot (w_1 - w_2) + \frac{\lambda}{2} |w_1 - w_2|^2$$

$L(w_1) = L(w_2)$

$$\Rightarrow \frac{\lambda}{2} |w_1 - w_2|^2 \leq 0 \Rightarrow w_1 = w_2 .$$

Before studying the convergence result in the discrete case, let us gain some insight with the continuous-time case.

$$\dot{w}(t) = -\nabla L(w(t))$$

with $0 < \lambda \leq D^2 L \leq \Lambda$.

Let w^* be the unique minimum of L.

Let us study

$$\frac{d}{dt}\left(\frac{1}{2}|w(t)-w^*|^2\right) = (w(t)-w^*)\cdot(\dot{w}(t))$$

$$= -\nabla L(w(t))\cdot(w(t)-w^*)$$

Recall that, by $\lambda \leq D^2 L$,

$$L(w^*) \geq L(w(t)) + \nabla L(w(t))\cdot(w^*-w(t)) + \frac{\lambda}{2}|w(t)-w^*|^2 \Rightarrow$$

$$\Rightarrow -\nabla L(w(t))\cdot(w(t)-w^*) \leq$$

$$\leq \underbrace{\big(L(w^*)-L(w(t))\big)}_{\leq 0} - \frac{\lambda}{2}|w(t)-w^*|^2$$

$$\leq -\frac{\lambda}{2}|w(t)-w^*|^2$$

Hence

$$\frac{d}{dt}\left(|w(t) - w^*|^2 \right) \leq -\frac{\lambda}{2} |w(t) - w^*|^2 .$$

This implies that

$$|w(t) - w^*|^2 \leq |w(0) - w^*|^2 e^{-\frac{\lambda}{2} t}$$

Which converges to zero with an exponential rate proportional to λ.

We want to mimick this proof in the discrete setting, with some technicalities related to discrete computations.

We are ready to show a convergence result for GD in the discrete setting.

<u>Theorem</u>: Assume that $L : \mathbb{R}^d \to \mathbb{R}$ is of class C^2 and its Hessian satisfies:

$$0 < \lambda \leq D^2 L(w) \leq \Lambda$$

Let $(w^k)_{k \in \mathbb{N}}$ be the sequence generated by GD with step size $\tau > 0$.

Assume that $\tau \leq \dfrac{1}{\Lambda}$.

Let w^* be the unique minimum point of L.

Then:

$$|w^k - w^*|^2 \leq (1 - \tau\lambda)^k |w^0 - w^*|^2 .$$

<u>Proof</u>: We compute

$$|w^{k+1} - w^*|^2 = |w^{k+1} - w^k + w^k - w^*|^2 =$$

$$= |w^{k+1} - w^k|^2 + 2(w^{k+1} - w^k) \cdot (w^k - w^*) +$$
$$+ |w^k - w^*|^2$$

$$= |w^k - w^*|^2 - 2\tau \nabla L(w^k)\cdot(w^k - w^*)$$
$$+ \underbrace{|w^{k+1} - w^k|^2} \longrightarrow \tau^2 |\nabla L(w^k)|^2$$

Since $D^2 L \geq \lambda$

$$L(w^*) \geq L(w^k) + \nabla L(w^k)\cdot(w^* - w^k) + \frac{\lambda}{2}|w^k - w^*|^2 \Rightarrow$$

$$\Rightarrow -2\tau \nabla L(w^k)\cdot(w^k - w^*) \leq$$
$$\leq 2\tau\big(L(w^*) - L(w^k)\big) - \tau\lambda|w^k - w^*|^2$$

Hence,

$$|w^{k+1} - w^*|^2 \leq (1 - \tau\lambda)|w^k - w^*|^2 +$$
$$+ 2\tau\big(L(w^*) - L(w^k)\big) +$$
$$+ \tau^2 |\nabla L(w^k)|^2$$

this is positive, but this is negative. Can we absorb it?

Since $D^2L(w^k) \leq \Lambda$, we have

$$L(w^{k+1}) \leq L(w^k) + \nabla L(w^k)\cdot(w^{k+1} - w^k) + \frac{\Lambda}{2}|w^{k+1} - w^k|^2 =$$

$$= L(w^k) - \tau|\nabla L(w^k)|^2 + \frac{\Lambda}{2}\tau^2|\nabla L(w^k)|^2$$

$$\Downarrow \quad \leftarrow L(w^*) \leq L(w^{k+1})$$

$$L(w^*) - L(w^k) \leq -\tau|\nabla L(w^k)|^2 + \frac{\Lambda}{2}\tau^2|\nabla L(w^k)|^2$$

$$2\tau\big(L(w^*) - L(w^k)\big) + \tau^2|\nabla L(w^k)|^2 \leq$$

$$\leq -\tau^2|\nabla L(w^k)|^2 + \Lambda\tau^3|\nabla L(w^k)|^2 =$$

$$= -\tau^2\underbrace{(1 - \Lambda\tau)}|\nabla L(w^k)|^2$$

If this is ≥ 0, we are done:

$$1 - \Lambda\tau \geq 0 \Leftrightarrow \tau\Lambda \leq 1 \Leftrightarrow$$

$$\Leftrightarrow \tau \leq \frac{1}{\Lambda} \quad \checkmark$$

We have shown that

$$|w^{k+1} - w^*|^2 \leq (1 - \tau\lambda)|w^k - w^*|^2.$$

Iterating this, we get the desired inequality! $\square$

The assumptions on L are very strict and costs in machine learning do not meet these assumptions ...

See Python notebook for problems with local minima.