Optimization and Computational Linear Algebra for Data Science
Homework 12: Gradient descent

Problem 12.1 (2 points). (**a**) *f has a local maximum at $(1.2, 1.3)$ and a global maximum at $(-0.5, -0.7)$. f has a local minimum at $(-0.9, 0.7)$ and a global minimum at $(0.9, -0.9)$ and a saddle-point at $(0.9, 0)$.*

(**b**) *When initialized at A, gradient descent is likely to converge to the local minimum at $(-0.9, 0.7)$. When initialized at B, gradient descent is likely to converge to the global minimum at $(0.9, -0.9)$.*

Problem 12.2 (5 points).

$$f(x) = \frac{1}{2} x^{\mathsf{T}} M x - \langle x, b \rangle + c$$

(**a**) *Let $x \in \mathbb{R}^d$. f is twice differentiable and*

$$H_f(x) = M.$$

By definition of μ and L, the eigenvalues of M are all above μ and all smaller than L: f is therefore μ-strongly convex and L-smooth. f is therefore convex. Hence

$$\begin{aligned} x \text{ is a global minimizer of } f \quad &\iff \quad \nabla f(x) = 0 \\ &\iff \quad Mx - b = 0 \\ &\iff \quad x = M^{-1} b. \end{aligned}$$

$x^* = M^{-1}b$ is therefore the unique global minimizer of f.

(**b**) *$\nabla f(x) = Mx - b$ hence*

$$\begin{aligned} x_{t+1} - x^* &= x_t - x^* - \beta(Mx_t - b) = x_t - x^* - \beta M(x_t - M^{-1}b) \\ &= x_t - x^* - \beta(x_t - x^*) \\ &= (\mathrm{Id} - \beta M)(x_t - x^*). \end{aligned}$$

(**c**) *Let $B = \mathrm{Id} - \beta M$. B is symmetric and his eigenvalues are:*

$$1 - \lambda_1/L, \dots, 1 - \lambda_d/L$$

which are all between 0 and $1 - \mu/L$. The largest eigenvalue of B^2 is therefore $(1 - \mu/L)^2$. Since the singular values of B are the square root of the eigenvalues of $B^{\mathsf{T}}B = B^2$ because B is symmetric, we get that the largest singular value of B is $1 - \mu/L$.

We know that the spectral norm of a matrix is equal to its largest singular value: $\|B\|_{\mathrm{Sp}} = 1 - \mu/L$. Hence

$$\|x_{t+1} - x^*\| = \|B(x_t - x^*)\| \leq \|B\|_{\mathrm{Sp}} \|x_t - x^*\| = \Big(1 - \frac{\mu}{L}\Big) \|x_t - x^*\|,$$

from which the result follows.

(d) *Since* $w_{t+1} = (\mathrm{Id} - L^{-1}M)w_t$, *we have for* $i \in \{1, \dots, d\}$

$$\alpha_i(t+1) = v_i^\mathsf{T}(\mathrm{Id} - L^{-1}M)w_t = (v_i^\mathsf{T} - L^{-1}v_i^\mathsf{T}M)w_t.$$

Now, we use the fact that $Mv_i = \lambda_i v_i$ *to get* $v_i^\mathsf{T}M = \lambda_i v_i^\mathsf{T}$*:*

$$\alpha_i(t+1) = (1 - \lambda_i/L)\alpha_i(t).$$

This gives

$$\alpha_i(t) = (1 - \lambda_i/L)^t \alpha_i(0).$$

(e) *Let* $i \in \{1, \dots, d\}$. $|\alpha_i(t)| = |\langle v_i, x_t - x^* \rangle|$ *is equal to the norm of the orthogonal projection of* $x_t - x^*$ *onto* $\mathrm{Span}(v_i)$, *that is corresponds to «the distance between* x_t *and* x^* *in the direction of* v_i*».*

From the previous we see that at each iteration of gradient descent, this «distance» is multiplied by a factor $1 - \lambda_i/L$, *where* ${}_i \in [\mu, L]$. *Hence, gradient descent converges faster «in the direction of* v_i*» if* λ_i *is large (close to* L*).*

(f) $(\alpha_1(t), \dots, \alpha_d(t))$ *are the coordinates of* $w_t = x_t - x^*$ *in the orthonormal basis* $(v_1, \dots, v_d)$. *Therefore*

$$\|x_t - x^*\| = \sqrt{\sum_{i=1}^d \alpha_i(t)^2} = \sqrt{\sum_{i=1}^d \Big(1 - \frac{\lambda_i}{L}\Big)^{2t} \big\langle v_i, x_0 - x^* \big\rangle^2}.$$