Exercice 1 (GD/HB for quadratic function). Let $x_\star$ be any vector, H a positive semi-definite symmetric matrix and f the convex quadratic function defined as

$$f(x) = \frac{1}{2}(x - x_\star)^T H(x - x_\star) + f_\star.$$

The Gradient descent (GD) method is defined by the update rule

$$x_{t+1} = x_t - \gamma \nabla f(x_t) \tag{GD}$$

where γ is called step-size.

1. Prove (GD)'s iterates verify the relation $x_{t+1} - x_\star = (I - \gamma H)(x_t - x_\star)$.
2. Assuming H's eigenvalues λ verify $0 < \mu \leq \lambda \leq L$, provide the worst case rate of (GD). Propose a value for γ and provide the worst-case rate of (**??**) with this specific choice of step-size γ.
3. We now assume that $\mu = 0$ (i.e. H's eigenvalues can be arbitrarily small). The previous worst-case rate becomes 1. We then bound the function value. Prove that $f(x_t) - f_\star = \frac{1}{2}(x_0 - x_\star)^T H(I - \gamma H)^{2t}(x_0 - x_\star)$ and propose a worst case bound of $f(x_t) - f_\star$. Which γ would you consider ? What is the worst case value of $f(x_t) - f_\star$ with this specific step-size γ ?
4. We consider general first order methods of the form $x_{t+1} = x_0 - \sum_{s=0}^{t} \gamma_{t,s} \nabla f(x_s)$ with $\gamma_{t,t} \neq 0$. Prove that for each of those methods, there exists a sequence of polynomials $(P_t)_{t\in\mathbb{N}}$ with $deg(P_t) = t$ and $P_t(0) = 1$ such that $\forall t, x_t - x_\star = P_t(H)(x_0 - x_\star)$.
5. Prove the reciproqua of the previous question.
6. Assuming $\mu \neq 0$, prove the best method can be designed by solving

$$\begin{cases} \max\limits_{Q_t \in \mathbb{R}_t[X]} & Q_t\left(\frac{L+\mu}{L-\mu}\right) \\ \text{s.t.} & \sup_{X\in[-1,1]}(|Q_t(X)|) = 1 \end{cases} \tag{1}$$

   Hint : First consider the polynomial P_t associated with the considered method. Find the problem it solves, and use some translation and rescale of P_t to define Q_t.

7. Prove the t^{th} Tchebychev polynomial T_t, defined as verifying $\forall\theta,\ T_t(cos(\theta)) = cos(t\theta)$, is solution of this problem.
8. Find the associated method and convergence guarantee. Provide an equivalent of this rate.

   Hint : Tchebychev polynomials verify the recursion $T_{t+1} = 2XT_t - T_{t-1}$.

   Hint 2 : Tchebychev polynomials verify $\forall X > 1, T_t(X) \underset{t\to\infty}{\sim} \frac{1}{2}(X + \sqrt{X^2 - 1})^t$.

9. What is the stationary behavior of this method ? (Provide limits of the parameters). This method is called Heavy-ball (HB) method.

bonus : Find a convergence guarantee of HB.

10. We now assume $\mu = 0$. Write the corresponding problem over polynomials.
11. Propose some good candidate based on the Tchebychev family of polynomials.
12. Provide the associated method and convergence guarantee.

1 Solutions

Solution Exercice 1. :

1. Writing (GD)'s update with this specific function f leads to $x_{t+1} = x_t - \gamma \nabla f(x_t) = x_t - \gamma H(x_t - x_\star)$, and finally $x_{t+1} - x_\star = (I - \gamma H)(x_t - x_\star)$.
2. By enrolling this recursion, we have

$$x_t - x_\star = (I - \gamma H)^t (x_0 - x_\star).$$

Therefore, $\|x_t - x_\star\| \leq \|I - \gamma H\|^t \|x_0 - x_\star\|$. By symmetry of $I - \gamma H$,

$$\begin{aligned} \|I - \gamma H\| &= \sup_{\lambda \in [\mu, L]} |1 - \gamma\lambda| \\ &= \max_{\lambda \in \{\mu, L\}} |1 - \gamma\lambda|, \quad \text{by convexity of } \lambda \mapsto |1 - \gamma\lambda| \\ &= \max\left(1 - \gamma\mu, L\gamma - 1\right). \end{aligned}$$

Hence $\|x_t - x_\star\| \leq \max\left(1 - \gamma\mu, L\gamma - 1\right)^t \|x_0 - x_\star\|$.
This guarantee is minimized for γ verifying $1 - \gamma\mu = L\gamma - 1$, i.e. $\gamma = \frac{2}{L+\mu}$, leading to $\|x_t - x_\star\| \leq \left(\frac{L-\mu}{L+\mu}\right)^t \|x_0 - x_\star\|$.

3. When $\mu = 0$, this guarantee is simply $\|x_t - x_\star\| \leq \|x_0 - x_\star\|$, not proving any convergence. Indeed, the function can be arbitrarily flat and we cannot obtain any convergence guarantee in term of the distance of iterates to optimum. We then consider the function value of the iterates.
We still have $x_t - x_\star = (I - \gamma H)^t (x_0 - x_\star)$. Then,

$$\begin{aligned} f(x_t) - f_\star &= \frac{1}{2}(x_t - x_\star)^T H (x_t - x_\star) \\ &= \frac{1}{2}(x_0 - x_\star)^T (I - \gamma H)^t H (I - \gamma H)^t (x_0 - x_\star) \\ &= \frac{1}{2}(x_0 - x_\star)^T H (I - \gamma H)^{2t} (x_0 - x_\star). \end{aligned}$$

Then, we can bound

$$\begin{aligned} f(x_t) - f_\star &\leq \frac{1}{2}\|H(I - \gamma H)^{2t}\| \|x_0 - x_\star\|^2 \\ &= \frac{1}{2} \sup_{\lambda \in [0, L]} \left(\lambda(1 - \gamma\lambda)^{2t}\right) \|x_0 - x_\star\|^2. \end{aligned}$$

We need to find the maximum of $\lambda(1-\gamma\lambda)^{2t}$ over $[0, L]$. This can be done for example by computing the derivative $\frac{\partial \lambda(1-\gamma\lambda)^{2t}}{\partial \lambda} = (1 - \gamma\lambda)^{2t} - 2t\gamma\lambda(1 - \gamma\lambda)^{2t-1} = (1 - \gamma\lambda)^{2t-1}(1 - (2t + 1)\gamma\lambda)$, which is negative on $[\frac{1}{(2t+1)\gamma}, \frac{1}{\gamma}]$ and non negative elsewhere. Therefore, $\lambda \mapsto \lambda(1 - \gamma\lambda)^{2t}$ reaches a local maximum in $\frac{1}{(2t+1)\gamma}$, then decreases until $\frac{1}{\gamma}$, and increases again up to infinity. Its worth noting that the value reached in $\frac{1}{(2t+1)\gamma}$ is also reached once in $[\frac{1}{\gamma}, \frac{2}{\gamma}]$. Let's us introduce $c_t \in [1, 2]$ such that $\frac{c_t}{\gamma}$ is the point where this value is reached again.
We have :

— if $\gamma \leq \frac{1}{(2t+1)L}$, then $\lambda \mapsto \lambda(1-\gamma\lambda)^{2t}$ reaches its maximum over $[0, L]$ in L.

— if $\frac{1}{(2t+1)L} \leq \gamma \leq \frac{c_t}{L}$, then $\lambda \mapsto \lambda(1-\gamma\lambda)^{2t}$ reaches its maximum over $[0, L]$ in $\frac{1}{(2t+1)\gamma}$.

— if $\frac{c_t}{L} \leq \gamma$, then $\lambda \mapsto \lambda(1-\gamma\lambda)^{2t}$ reaches its maximum over $[0, L]$ in L.

In the first and third cases, the convergence guarantee is $f(x_t) - f_\star \leq \frac{1}{2}\left(L(1-\gamma L)^{2t}\right)\|x_0 - x_\star\|^2$ and is minimized when $L\gamma$ is as close as possible to 1. Therefore, conditioned to $\gamma \leq \frac{1}{(2t+1)L}$, γ is optimal in $\gamma = \frac{1}{(2t+1)L}$ and conditioned to $\frac{c_t}{L} \leq \gamma$, γ is optimal in $\gamma = \frac{c_t}{L}$. We conclude γ is optimal in $[\frac{1}{(2t+1)L}, \frac{c_t}{L}]$.

In the case where $\frac{1}{(2t+1)L} \leq \gamma \leq \frac{c_t}{L}$, and then $\lambda = \frac{1}{(2t+1)\gamma}$, leads to the bound

$$f(x_t) - f_\star \leq \frac{1}{2\gamma}\left(\frac{1}{2t+1}\left(1 - \frac{1}{2t+1}\right)^{2t}\right)\|x_0 - x_\star\|^2,$$

optimized for the largest possible γ, hence $\gamma = \frac{c_t}{L}$.

We conclude the optimal γ is $\frac{c_t}{L}$ and reaches the bound

$$f(x_t) - f_\star \leq \frac{L}{4tc_t}\left(1 - \frac{1}{2t+1}\right)^{2t+1}\|x_0 - x_\star\|^2.$$

This bound is equivalent to $f(x_t) - f_\star \leq \frac{L}{8et}\|x_0 - x_\star\|^2$, since we can easily show that $c_t = 2 - \frac{ln(4et)}{2t} + o\left(\frac{1}{t}\right)$.

Note that the interval $[\frac{1}{(2t+1)L}, \frac{c_t}{L}]$ is converging to $(0, \frac{2}{L})$ and therefore, for any choice of $\gamma = \frac{c}{L}$ where c is independent of t, the bound is in $O(1/t)$.

4. We consider general first order methods of the form $x_{t+1} = x_0 - \sum_{s=0}^{t}\gamma_{t,s}\nabla f(x_s)$ with $\gamma_{t,t} \neq 0$.

First we notice that $x_0 - x_\star = P_0(H)(x_0 - x_\star)$ with $P_0 = 1$ (degree 0 and value 1 in 0). Then we assume by induction that the beginning $(P_s)_{s \leq t}$ of such sequence exists. Using the formulation of the algorithm we have

$$\begin{aligned} x_{t+1} &= x_0 - \sum_{s=0}^{t}\gamma_{t,s}\nabla f(x_s) \\ x_{t+1} - x_\star &= x_0 - x_\star - \sum_{s=0}^{t}\gamma_{t,s}H(x_s - x_\star) \\ &= x_0 - x_\star - \sum_{s=0}^{t}\gamma_{t,s}HP_s(H)(x_0 - x_\star) \\ &= \left(1 - \sum_{s=0}^{t}\gamma_{t,s}XP_s\right)(H)(x_0 - x_\star) \\ &\triangleq P_{t+1}(H)(x_0 - x_\star). \end{aligned}$$

And we verify that $P_{t+1}(0) = 1$ by construction and $deg(P_{t+1}) = t + 1$ since $\gamma_{t,t} \neq 0$ and that XP_t is the only polynomial of degree maximal in the sum, i.e. $t + 1$.

5. Reciprocally, we assume the existence of such polynomials, therefore the polynomials $1, XP_0, XP_1, \cdots, XP_t$ respectively have degree $0, 1, \cdots, t+1$, and hence form a basis of $\mathbb{R}_{t+1}[X]$ and P_{t+1} is linearly expressible in this basis. Since $P_{t+1}(0) = 1$, we know that there exists $(\gamma_{t,s})_{s\in[\![0,t]\!]}$ such that $P_{t+1} = 1 - \sum_{s=0}^{t} \gamma_{t,s} XP_s$.
   Finally,
$$\begin{aligned} x_{t+1} - x_\star &= P_{t+1}(H)(x_0 - x_\star) \\ &= \left(1 - \sum_{s=0}^{t} \gamma_{t,s} XP_s\right)(H)(x_0 - x_\star) \\ &= x_0 - x_\star - \sum_{s=0}^{t} \gamma_{t,s} HP_s(H)(x_0 - x_\star) \\ &= x_0 - x_\star - \sum_{s=0}^{t} \gamma_{t,s} H(x_s - x_\star) \\ x_{t+1} &= x_0 - \sum_{s=0}^{t} \gamma_{t,s} \nabla f(x_s). \end{aligned}$$

6. Let's define P_t such that the studied method verifies $x_t - x_\star = P_t(H)(x_0 - x_\star)$. Then,
$$\begin{aligned} \|x_t - x_\star\| &\leq \|P_t(H)\| \|x_0 - x_\star\| \\ &\leq \sup_{\lambda\in[\mu,L]} (|P_t(\lambda)|) \|x_0 - x_\star\|. \end{aligned}$$
   We therefore naturally look for P_t of degree t with $P_t(0) = 1$ that minimizes the quantity $\sup_{\lambda\in[\mu,L]}(|P_t(\lambda)|)$.
   For each feasible polynomial P_t, we define Q_t such that $P_t(X) = Q_t\left(\frac{L+\mu}{2} - \frac{L-\mu}{2}X\right)$. Therefore, the degree of Q_t is still t and the constraint $P_t(0) = 1$ becomes $Q_t\left(\frac{L+\mu}{L-\mu}\right) = 1$. We write the problem in terms of Q_t which minimizes the quantity $\sup_{X\in[-1,1]}(|Q_t(X)|)$ subject to the degree and $Q_t\left(\frac{L+\mu}{L-\mu}\right) = 1$.
   Up to a simple rescale, we can also conclude that we constraint $\sup_{X\in[-1,1]}(|Q_t(X)|) = 1$ and maximize $Q_t\left(\frac{L+\mu}{L-\mu}\right)$.

7. By definition the polynomial T_t, defined as verifying $\forall\theta,\ T_t(cos(\theta)) = cos(t\theta)$ has the right degree and is bounded by 1 on [-1, 1].
   Moreover, it reaches -1 and 1 exactly $t+1$ times in $\frac{k\pi}{t}, k \in [\![0,t]\!]$.
   Assuming another polynomial $\bar{Q}$ is feasible and is such that $Q\left(\frac{L+\mu}{L-\mu}\right) > Q_t\left(\frac{L+\mu}{L-\mu}\right)$. Therefore, we can rescale by 1 plus a small ε (and let $Q = \frac{\bar{Q}}{1+\varepsilon}$)such that $Q\left(\frac{L+\mu}{L-\mu}\right) > Q_t\left(\frac{L+\mu}{L-\mu}\right)$ still holds and $\sup_{X\in[-1,1]}(|Q_t(X)|) < 1$. Therefore, the sign of $Q_t - Q$ changes t times between -1 and 1. And since $Q\left(\frac{L+\mu}{L-\mu}\right) > Q_t\left(\frac{L+\mu}{L-\mu}\right)$, then this sign changes again in $[1,\infty)$. Therefore, Q and Q_t cross each other $t+1$ times, but $deg(Q_t) = deg(Q) = t$, therefore $Q_t = Q$ which contradicts the main hypothesis that $Q\left(\frac{L+\mu}{L-\mu}\right) > Q_t\left(\frac{L+\mu}{L-\mu}\right)$.

8. We found $Q_t(X) = T_t(X)$, therefore, $T_t(X)/T_t\left(\frac{L+\mu}{L-\mu}\right)$ is equal to 1 in $\frac{L+\mu}{L-\mu}$ and is “small” on $[-1,1]$.

Therefore, the polynomial P_t associated with the studied method is

$$P_t(\lambda) = \frac{T_t\left(\frac{L+\mu}{L-\mu} - \frac{2}{L-\mu}\lambda\right)}{T_t\left(\frac{L+\mu}{L-\mu}\right)} \tag{2}$$

and the associated convergence guarantee is

$$\|x_t - x_\star\| \leq \frac{1}{T_t\left(\frac{L+\mu}{L-\mu}\right)}\|x_0 - x_\star\|. \tag{3}$$

Using the recursion of Tchebychev polynomials, we have

$$\begin{aligned} P_{t+1}(\lambda) &= \frac{T_{t+1}\left(\frac{L+\mu}{L-\mu} - \frac{2}{L-\mu}\lambda\right)}{T_{t+1}\left(\frac{L+\mu}{L-\mu}\right)} \\ &= \frac{2\left(\frac{L+\mu}{L-\mu} - \frac{2}{L-\mu}\lambda\right)T_t\left(\frac{L+\mu}{L-\mu} - \frac{2}{L-\mu}\lambda\right) - T_{t-1}\left(\frac{L+\mu}{L-\mu} - \frac{2}{L-\mu}\lambda\right)}{T_{t+1}\left(\frac{L+\mu}{L-\mu}\right)} \\ &= \frac{2\left(\frac{L+\mu}{L-\mu} - \frac{2}{L-\mu}\lambda\right)T_t\left(\frac{L+\mu}{L-\mu}\right)P_t(\lambda) - T_{t-1}\left(\frac{L+\mu}{L-\mu}\right)P_{t-1}(\lambda)}{T_{t+1}\left(\frac{L+\mu}{L-\mu}\right)} \\ &= (1 + m_t - h_t\lambda)P_t(\lambda) - m_tP_{t-1}(\lambda), \end{aligned}$$

with

$$\begin{cases} h_t = & \frac{\frac{4}{L-\mu}T_t\left(\frac{L+\mu}{L-\mu}\right)}{T_{t+1}\left(\frac{L+\mu}{L-\mu}\right)}, \\ m_t = & \frac{T_{t-1}\left(\frac{L+\mu}{L-\mu}\right)}{T_{t+1}\left(\frac{L+\mu}{L-\mu}\right)}. \end{cases} \tag{4}$$

Finally,

$$\begin{aligned} x_{t+1} - x_\star &= P_{t+1}(H)(x_0 - x_\star) \\ &= \left((1 + m_t - h_tH)P_t(H) - m_tP_{t-1}(H)\right)(x_0 - x_\star) \\ &= ((1 + m_t)I_d - h_tH)(x_t - x_\star) - m_t(x_{t-1} - x_\star) \\ &= (1 + m_t)(x_t - x_\star) - h_t\nabla f(x_t) - m_t(x_{t-1} - x_\star) \\ &= (x_t - x_\star) - h_t\nabla f(x_t) + m_t(x_t - x_{t-1}). \end{aligned} \tag{5}$$

Finally, the Tchebychev method is written

$$x_{t+1} = x_t - h_t\nabla f(x_t) + m_t(x_t - x_{t-1}), \tag{Tchebychev}$$

with

$$\begin{cases} h_t = & \frac{\frac{4}{L-\mu}T_t\left(\frac{L+\mu}{L-\mu}\right)}{T_{t+1}\left(\frac{L+\mu}{L-\mu}\right)}, \\ m_t = & \frac{T_{t-1}\left(\frac{L+\mu}{L-\mu}\right)}{T_{t+1}\left(\frac{L+\mu}{L-\mu}\right)}. \end{cases} \tag{6}$$

As mentioned above, the Tchebychev convergence guarantee is $\|x_t - x_\star\| \leq \frac{1}{T_t\left(\frac{L+\mu}{L-\mu}\right)}\|x_0 - x_\star\|$.

Since $\forall X > 1, T_t(X) \underset{t\to\infty}{\sim} \frac{1}{2}(X + \sqrt{X^2-1})^t$, $T_t\left(\frac{L+\mu}{L-\mu}\right) \underset{t\to\infty}{\sim} \frac{1}{2}\left(\frac{\sqrt{L}+\sqrt{\mu}}{\sqrt{L}-\sqrt{\mu}}\right)^t$.

The provided Tchebychev convergence guarantee is $\|x_t - x_\star\| \lesssim 2\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^t \|x_0 - x_\star\|$.

9. Based on the same equivalent, we find

$$\begin{cases} h_t = & \frac{\frac{4}{L-\mu}T_t\left(\frac{L+\mu}{L-\mu}\right)}{T_{t+1}\left(\frac{L+\mu}{L-\mu}\right)} & \longrightarrow \left(\frac{2}{\sqrt{L}+\sqrt{\mu}}\right)^2, \\ m_t = & \frac{T_{t-1}\left(\frac{L+\mu}{L-\mu}\right)}{T_{t+1}\left(\frac{L+\mu}{L-\mu}\right)} & \longrightarrow \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2. \end{cases} \tag{7}$$

The Heavy-ball (HB) method is written

$$x_{t+1} = x_t - h\nabla f(x_t) + m(x_t - x_{t-1}), \tag{HB}$$

with

$$\begin{cases} h = & \left(\frac{2}{\sqrt{L}+\sqrt{\mu}}\right)^2, \\ m = & \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2. \end{cases} \tag{8}$$

bonus : We prove by induction that the polynomials associated with the (HB) method is

$$P_t(\lambda) = m^{t/2}\left(\frac{2m}{1+m}T_t\left(\frac{1+m-h\lambda}{2\sqrt{m}}\right) + \frac{1-m}{1+m}U_t\left(\frac{1+m-h\lambda}{2\sqrt{m}}\right)\right) \tag{9}$$

for any h, m, where U denotes the second type Tchebychev polynomials defined as the polynomials verifying $U_t(cos(\theta)) = \frac{sin((t+1)\theta)}{sin(\theta)}$.

With the previously found values of h and m, and since $\sup_{[0,1]} U_t = t+1$, we have $\sup_{\lambda\in[\mu,L]} P_t(\lambda) \leq \left(\frac{2\sqrt{L\mu}}{L+\mu}t + 1\right)\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^t$ and finally,

$$\|x_t - x_\star\| \leq \left(\frac{2\sqrt{L\mu}}{L+\mu}t + 1\right)\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^t \|x_t - x_\star\|. \tag{10}$$

10. We now assume $\mu = 0$. We have $f(x_t) - f_\star = \frac{1}{2}(x_0 - x_\star)^T H P_t(H)^2 (x_0 - x_\star)$.

Hence we want to minimize $\sup_{\lambda\in[0,L]} \lambda P_t(\lambda)^2$, subject to $P_t(0) = 1$ and P_t has a degree t.

Substituting $\lambda = LX^2$, with $X \in [-1,1]$, we have $\sup_{\lambda\in[0,L]} \lambda P_t(\lambda)^2 = \sup_{X\in[-11]} L\left(XP_t(LX^2)\right)^2$.

Calling Q_t the odd polynomial $Q_t(X) = XP_t(LX^2)$ of degree $2t+1$ verifying $Q'(0) = 1$, we have

$$f(x_t) - f_\star \leq \frac{L}{2}\left(\sup_{[-1,1]} |Q_t|\right)^2 \|x_0 - x_\star\|^2,$$

and the goal is to minimize this quantity.

$$\begin{cases} \min_{Q \in \mathbb{R}_{2t+1}[X]} & \sup_{[-1,1]} |Q_t| \\ \text{s.t.} & Q'(0) = 1 \text{ and } Q \text{ is odd.} \end{cases} \tag{11}$$

11. As previously we can rescale and try to maximize $Q'(0)$ instead.
    One good candidate is the polynomial $Q_t = T_{2t+1}/T'_{2t+1}(0)$.
    Note that $T'_{2t+1}(0) = (-1)^t(2t+1)$.
12. Therefore, the convergence guarantee of the associated method is

$$f(x_t) - f_\star \leq \frac{L}{2}\frac{1}{(2t+1)^2}\|x_0 - x_\star\|^2.$$

And the associated polynomial is $P_t = \frac{(-1)^t}{2t+1}\frac{T_{2t+1}(\sqrt{X/L})}{\sqrt{X/L}}$.
Using the Tchebyshev recursion formula, we have

$$P_{t+1} = (2 - 4X/L)\frac{2t+1}{2t+3}P_t - \frac{2t-1}{2t+3}P_{t-1}. \tag{12}$$

Finally, we obtain the recursion

$$x_{t+1} = x_t - h_t\nabla f(x_t) + m_t(x_t - x_{t-1}) \tag{13}$$

with $h_t = \frac{4}{L}\frac{2t+1}{2t+3}$ and $m_t = \frac{2t-1}{2t+3}$.