TD 01 — SD-TSIA 211

1 Picard's fixed point theorem

(Picard's fixed point theorem).

Prove the following theorem:
If $T\colon \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\exists 0 < \rho < 1, \forall x \in \mathbb{R}^d, \forall y \in \mathbb{R}^d, \quad \|T(x) - T(y)\| \leq \rho\|x - y\| \tag{1}$$

then, T has a unique fixed point x^* such that $x^* = T(x^*)$
Moreover, every sequence of the form $x_{k+1} = T(x_k)$ converges to x^* with a linear convergence rate given by $\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\|$

1.1 Recalls

Definition 1. *Cauchy sequence*

Let $(\mathcal{E}, \|.\|)$ be a normed vector space, and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{E}$, then $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence if: $\forall \varepsilon > 0, \exists N \in \mathbb{N}$ such that

$$\forall n, m > N, \quad \|x_n - x_m\| < \varepsilon$$

Or, equivalently,

$$\forall n, m > N, \quad \|x_n - x_m\| \to 0 \ \ as \ \ n, m \to +\infty$$

Definition 2. *Sum of a finite Geometric sequence*

The sum S_n of the $1^{st} - n$ terms of a Geometric sequence is:

$$S_n = \frac{a_1(1 - r^n)}{1 - r} \qquad r \neq 0$$

1.2 Proof

Since $\mathbb{R}^d$ is a complete space, then one way of proving that $(x_k)_{k\in\mathbb{N}}$ converges to some limit point, x^*, is to prove that $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence. Hence, it does converge. Secondly, we shall prove that the limit point, x^*, is a fixed point (*i.e.* $T(x^*) = x^*$). Then, a usual proof of uniqueness. Lastly, we'll show the linear convergence rate.

1. We'll show that (x_k) is a Cauchy sequence.

   $\forall k \in \mathbb{N}, m \in \mathbb{N}\setminus\{0\}$,

   $$\|x_{k+m} - x_k\| = \|\sum_{\ell=1}^{m} x_{k+\ell} - x_{k+\ell-1}\| \leq \sum_{\ell=1}^{m} \|x_{k+\ell} - x_{k+\ell-1}\| \tag{2}$$

   Now, $\forall \ell \in \{1, \ldots\ldots, m\}$,

   $$\begin{aligned} \|x_{k+\ell} - x_{k+\ell-1}\| &= \|T(x_{k+\ell-1}) - T(x_{k+\ell-2})\| && \text{using } \ x_{k+1} = T(x_k) \\ &\leq \rho\|x_{k+\ell-1} - x_{k+\ell-2}\| && \text{by (1)} \\ &\vdots \\ &\leq \rho^{k+\ell-1}\|x_1 - x_0\| \end{aligned} \tag{3}$$

   Thus,

   $$\begin{aligned} \|x_{k+m} - x_k\| &\leq \sum_{\ell=1}^{m} \rho^{k+\ell-1}\|x_1 - x_0\| && \text{by (3)} \\ &= \rho^k\|x_1 - x_0\| \sum_{\ell=1}^{m} \rho^{\ell-1} \\ &= (1-\rho)^{-1}(1-\rho^m)\rho^k\|x_1 - x_0\| \end{aligned}$$

   Which implies that $\|x_{k+m} - x_k\| \to 0$ as $k \to +\infty$, and hence $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence. Therefore it converges to some limit point, x^*. □

2. We'll show that x^* is a fixed point of T.

   T is a contraction $\Rightarrow$ T is continuous. Thus,

   $$x_{k+1} = T(x_k) = T\big(\lim_{k\to+\infty} x_k\big) = T(x^*)$$

   As $k \to +\infty$, $T(x^*) = x_{k+1} \to x^* \Rightarrow T(x^*) = x^*$. □

3. Usual proof of uniqueness.

   Assume x^* and y^* are two fixed points of T, then

   $$\|x^* - y^*\| = \|T(x^*) - T(y^*)\| \overset{(1)}{\leq} \rho\|x^* - y^*\|$$

   Thus, $\|x^* - y^*\| = 0 \Rightarrow x^* = y^* \Rightarrow \text{Fix}(T) = \{x^*\}$ $\square$

4. Proof of linear convergence.

   $$\|x_k - x^*\| = \|T(x_{k-1}) - T(x^*)\| \overset{(1)}{\leq} \rho\|x_{k-1} - x^*\| \leq \dots\dots \leq \rho^k\|x_0 - x^*\| \quad \square$$

2 Gradient calculus

(Gradient calculus).

- Calculate the gradient of the following functions. A, M and Q are fixed matrices, b is a fixed vector. f_1 is useful for least squares and regression problems. f_2 is useful for logistic regression and binary classification, f_3 is useful for non-negative matrix factorization.

$$\begin{aligned}
f_1 \colon\ & \mathbb{R}^n \to \mathbb{R} \\
& x \mapsto \frac{1}{2}\|Ax - b\|^2 = \frac{1}{2}\sum_{i=1}^{m}(\sum_{j=1}^{n} A_{ij}x_j - b_i)^2 \\
f_2 \colon\ & \mathbb{R}^n \to \mathbb{R} \\
& z \mapsto \sum_{i=1}^{n} \log(1 + \exp(z_i)) \\
f_3 \colon\ & \mathbb{R}^{m\times p} \to \mathbb{R} \\
& P \mapsto \frac{1}{2}\|M - PQ\|_F^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(M_{ij} - \sum_{k=1}^{p} P_{ik}Q_{kj})^2
\end{aligned}$$

- Let g_1, g_2, g_3 be functions such that $g_1 \colon \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}, g_2 \colon \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}, g_3 \colon \mathbb{R}^{n_3} \to \mathbb{R}$ and let

$$f_4 = g_3 \circ g_2 \circ g_1.$$

Compute the gradient of f_4 using the Jacobian matrices of g_i for $i \in \{1, 2, 3\}$.

Suppose that computing one element of the Jacobian matrices costs C_J and that multiplying two numbers costs C_M. How much does it cost to compute $\nabla f_4(x)$?

2.1 Recalls

Lemma 1. *Let $A, B \in \mathbb{R}^{m\times n}$, then*

$$\|A + B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 + \langle A, B\rangle_F$$
$$\langle A, B\rangle_F = \operatorname{tr}(A^\intercal B)$$

Lemma 2. $\operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA)$

2.2 Solution

- Gradient calculations of f_1, f_2 and f_3

  1. $f_1(x) = \frac{1}{2}\|Ax - b\|^2$ (lecture notes - page 15)
  2. $f_2(z) = \sum_{i=1}^n \log(1 + \exp(z_i))$, thus by using partial derivative:

     $$\frac{\partial f_2}{\partial z_i} = \frac{\exp(z_i)}{1 + \exp(x_i)} = \frac{1}{1 + \exp(z_i)} =: \sigma(x_i) \Rightarrow \nabla f_2(x) = \Big(\sigma(x_1) \ldots\ldots \sigma(x_n)\Big)^\intercal$$

  3. $f_3(P) = \frac{1}{2}\|M - PQ\|_F^2$, by the definition:

     $$\begin{aligned} f_3(P + H) &= \frac{1}{2}\|M - (P + H)Q\|_F^2 \\ &= \frac{1}{2}\|M - PQ - HQ\|_F^2 \\ &= \frac{1}{2}\|M - PQ\|_F^2 - \langle M - PQ, HQ\rangle_F + \frac{1}{2}\|HQ\|_F^2 \\ &= f_3(P) - \langle M - PQ, HQ\rangle_F + \frac{1}{2}\|HQ\|_F^2 \end{aligned}$$

Now,

$$\begin{aligned}\langle M - PQ, HQ\rangle_F &= \mathrm{tr}((M - PQ)^\intercal HQ) \\ &= \mathrm{tr}(Q(M - PQ)^\intercal H) \\ &= \mathrm{tr}(((M - PQ)Q^\intercal)^\intercal H) \\ &= \langle (M - PQ)Q^\intercal, H\rangle_F \end{aligned} \tag{4}$$

Thus,

$$\begin{aligned} f_3(P + H) &= f_3(P) - \langle M - PQ, HQ\rangle_F + \frac{1}{2}\|HQ\|_F^2 \\ &\overset{(4)}{=} f_3(P) + \langle -(M - PQ)Q^\intercal, H\rangle_F + o(H) \end{aligned}$$

Hence, $\nabla f_3(P) = -(M - PQ)Q^\intercal$

- $f_4(x) = g_3 \circ g_2 \circ g_1(x)$, then by the chain rule:

$$\nabla f_4(x) = (J_{g_3}(g_2 \circ g_1(x))_{1\times n_3} \times J_{g_2}(g_1(x))_{n_3\times n_2} \times J_{g_1}(x)_{n_2\times n_1})^\intercal$$

Now, to figure out the total cost of computing $\nabla f_4(x)$, the cost depends on the sizes of the matrices. For instance:

  - The cost of computing a Jacobian $J \in \mathbb{R}^{p\times q}$ is equal to the cost of computing one element of this Jacobian, C_J, multiplied by its size, pq, which is in total $C_J \times pq$. Hence,
    * The cost of computing $J_{g_3} = C_J \times 1 \times n_3$
    * The cost of computing $J_{g_2} = C_J \times n_3 \times n_2$
    * The cost of computing $J_{g_1} = C_J \times n_2 \times n_1$

    $\Rightarrow$ The total cost of computing the Jacobians $= C_J(n_3 + n_3n_2 + n_2n_1)$
  - The cost of multiplying two matrices $A \times B$ where $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ is equal to the cost of multiplying two numbers, C_M, by the number of multiplication operations we do, mnp, which is in total $C_M \times mnp$. Hence,

* The cost of multiplying $J_{g_3}(g_2 \circ g_1(.))_{1\times n_3} \times J_{g_2}(g_1(.))_{n_3\times n_2} = C_M \times 1 \times n_3 \times n_2$
* Then, the cost of multiplying $S_{1\times n_2} = J_{g_3}(g_2 \circ g_1(.)) \times J_{g_2}(g_1(.))$ by $J_{g_1}(.)_{n_2\times n_1} = C_M \times 1 \times n_2 \times n_1$

$\Rightarrow$ The total cost of multiplying all the Jacobians $= C_M(n_2n_3 + n_2n_1)$

Thus, the total cost of computing $\nabla f_4(x)$ equals

$$C_J(n_3 + n_3n_2 + n_2n_1) + C_M(n_2n_3 + n_2n_1)$$

(Convergence of Gradient Descent for strongly convex C^2 functions.)

Consider a C^2 function $f\colon \mathbb{R}^d \to \mathbb{R}$ such that $\mu \mathrm{I} \preccurlyeq \nabla^2 f(x) \preccurlyeq L\mathrm{I}$.

1. Show that the fixed point operator $T\colon x \mapsto x - \gamma\nabla f(x)$ is contractant for any $0 < \gamma < \frac{2}{L}$
2. Show that the Gradient Descent method converges linearly.
3. How many iterations are necessary to ensure that $\|x_k - x^*\| \leq \varepsilon$?

2.3 Recalls

Theorem 1. *Mean Value Thoerem*

For a continuous vector-valued function $\mathcal{F}\colon [a,b] \to \mathbb{R}^k$ differentiable on (a,b), there exists a number $c \in (a,b)$ such that

$$\|\mathcal{F}(b) - \mathcal{F}(a)\| \leq \|\mathcal{F}'(c)\|(b-a)$$

Remark 1. *For a symmetric matrix A, $\|A\|_2 = |\lambda_{\max}(A)|$ where $\lambda_{\max}$ is the largest eigenvalue of A.*

2.4 Proof

1. We'll first show that the definition of a contractant operator T (1) can sufficiently be achieved by the property $\nabla T < 1$. Then, we'll show that the defined operator T does, indeed, satisfy the mentioned property.

   - Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be an operator and $\beta \in (0,1)$, then $\|\nabla T\| \leq \beta \Rightarrow T$ is a contraction (1).

     Fix $x, y \in \mathbb{R}^d$, and define:

     $$\begin{aligned} g \colon [0,1] &\to \mathbb{R}^d \\ t &\mapsto T(tx + (1-t)y) \end{aligned}$$

     Then,

     $$g'(t) = \langle \nabla T(tx + (1-t)y), x - y \rangle \tag{5}$$

     As g is continuous on $[0,1]$, and differentiable on $(0,1)$, thus one can apply the Mean Value Theorem: $\exists c \in (0,1)$ such that:

     $$\begin{aligned} \|g(1) - g(0)\| \leq \|g'(c)\| \iff \|T(x) - T(y)\| &\leq \|\langle \nabla T(cx + (1-c)y), x - y \rangle\| \\ &\leq \|\nabla T(cx + (1-c)y)\| \|x - y\| \\ &\leq \beta \|x - y\| \end{aligned}$$

     Thus, T is a contraction. □

   - We'll show that $T(x) = x - \gamma \nabla f(x)$ is a contraction using the previous property.

     $$T(x) = x - \gamma \nabla f(x) \Rightarrow \nabla T(x) = \mathrm{I} - \gamma \nabla^2 f(x)$$

     Now,

     $$\begin{aligned} \mu \mathrm{I} \preccurlyeq \nabla^2 f(x) \preccurlyeq L\mathrm{I} &\iff -\gamma L\mathrm{I} \preccurlyeq -\gamma \nabla^2 f(x) \preccurlyeq \gamma \mu \mathrm{I} \\ &\iff (1 - \gamma L)\mathrm{I} \preccurlyeq \mathrm{I} - \gamma \nabla^2 f(x) \preccurlyeq (1 - \gamma \mu)\mathrm{I} \\ &\iff (1 - \gamma L)\mathrm{I} \preccurlyeq \nabla T(x) \preccurlyeq (1 - \gamma \mu)\mathrm{I} \end{aligned}$$

$$\iff \|\nabla T(x)\| \leq \max(|1-\gamma L|, |1-\gamma\mu|)$$

Moreover, $\beta := \max(|1-\gamma L|, |1-\gamma\mu|) < 1$ whenever $\gamma \in (0, \frac{2}{L})$ □

2. Since $T(x) = x - \gamma\nabla f(x)$ is a contraction, then by Picard's fixed point theorem we conclude that the Gradient Descent converges linearly.

3. Again, by Picard's fixed point theorem, we know that:

$$\|x_k - x^*\| \leq \beta^k \|x_0 - x^*\| \quad \text{with } \beta = \max(|1-\gamma\mu|, |1-\gamma L|)$$

So, to ensure that $\|x_k - x^*\| \leq \beta^k \|x_0 - x^*\| \leq \varepsilon$, the algorithm requires number of iterations k such that:

$$\begin{aligned}
\beta^k \|x_0 - x^*\| \leq \varepsilon &\iff \log(\beta^k \|x_0 - x^*\|) \leq \log\varepsilon \\
&\iff k\log\beta + \log(\|x_0 - x^*\|) \leq \log\varepsilon \\
&\iff k\log\beta \leq \log\varepsilon - \log\|x_0 - x^*\| \\
&\iff k \geq \frac{\log\varepsilon - \log\|x_0 - x^*\|}{\log\beta}
\end{aligned}$$

Thus,

$$k = \left\lceil \frac{\log\left(\frac{\varepsilon}{\|x_0 - x^*\|}\right)}{\log\beta} \right\rceil$$