ONLINE OPTIMIZATION HW1

SIDDHANT CHAUDHARY
BMC201953

Derivative of Matrix Inverse using differentials. Consider the function F : GL(n) $\to$ GL(n) defined by

$$F(X) = X^{-1}$$

In this section, we will compute $D_X F(X)$, i.e the derivative of F at some matrix X. This will be used in **Problem 1.**

Let $U \in M_n$ be a *fixed* matrix. Let dF denote the differential of F, which is a map GL(n) $\to$ GL(n) defined as follows.

$$dF(X) = D_X F(X)(U) \qquad \forall X \in \text{GL(n)}$$

Consider the identity map V : GL(n) $\to$ GL(n) given by $V(X) = X$. Clearly, we see that

$$dV(X) = U \qquad \forall X \in \text{GL(n)}$$

Now, for all $X \in$ GL(n), we have

$$(F \cdot V)(X) = I_n$$

where I_n is the identity matrix (where $\cdot$ represents matrix multiplication). Applying the differential operator d to both sides, we see that

$$F\, dV + dF\, V = 0$$

where the above equality is an equality of maps (this holds because d as defined is a derivation). Applying the above map to some $X \in$ GL(n), we have

$$F(X)\, dV(X) + dF(X)\, V(X) = 0$$

which, upon rearranging both sides, gives us

$$dF(X) = -X^{-1} U X^{-1} \tag{$\dagger$}$$

Problem 1. Compute the gradient and Hessian for the following functions and write down their second order approximation.

(1) $f(X) = -\log(\det(X))$, for an $n \times n$ matrix X.
(2) $f(X) = \text{Tr}(AX)$, for a symmetric matrix A and X symmetric matrix of indeterminates.

Solution. Let us consider (1) first. The function is defined as follows.

$$f(X) = -\log(\det(X))$$

where X is an $n \times n$ matrix with positive determinant. By the chain rule, the derivative of f at some matrix X with positive determinant is given by the following.

$$D_X f(X) = -\log'(\det(X)) D_X \det(X) = \frac{-1}{\det(X)} D_X \det(X)$$

So, we only need to compute $D_X \det(X)$, i.e the derivative of det at the matrix X.

Date: 9 October, 2021.

Let the entries of X be denoted by X_{ij}. For each $1 \leq i, j \leq n$, we will compute the following partial derivative.

$$\frac{\partial \det(X)}{\partial X_{ij}}$$

i.e the partial derivative of the determinant at X with respect to the ijth entry.

We know the following holds for any matrix X, where $1 \leq i \leq n$ is any index.

$$\det(X) = \sum_{k=1}^{n} X_{ik} \cdot \text{Cof}(X_{ik})$$

where $\text{Cof}(X_{ij})$ is the cofactor of the ijth entry. So, by the product rule, we have the following.

$$\frac{\partial \det(X)}{\partial X_{ij}} = \sum_{k=1}^{n} \frac{\partial X_{ik}}{\partial X_{ij}} \text{Cof}(X_{ik}) + X_{ik} \frac{\partial \text{Cof}(X_{ik})}{\partial X_{ij}}$$

Now, if $j \neq k$, then clearly

$$\frac{\partial X_{ik}}{\partial X_{ij}} = 0$$

Also, note that for any $1 \leq k \leq n$, $\text{Cof}(X_{ik})$ does not depend on X_{ij} (by the way cofactors are defined), and hence for each $1 \leq k \leq n$, we have

$$\frac{\partial \text{Cof}(X_{ik})}{\partial X_{ij}} = 0$$

So, we see that

$$\frac{\partial \det(X)}{\partial X_{ij}} = \frac{\partial X_{ij}}{\partial X_{ij}} \text{Cof}(X_{ij}) = \text{Cof}(X_{ij}) = \text{adj}^T(X)_{ji}$$

where $\text{adj}(X)$ is the *adjoint* of the matrix X. Also, for any matrix, we know that

$$\text{adj}^T(X)_{ji} = \det(X)(X^{-1})^T_{ji}$$

(Recall that $X^{-1} = \frac{1}{\det(X)} \text{adj}(X)$). So,

$$\frac{\partial \det(X)}{\partial X_{ij}} = \det(X)(X^{-1})^T_{ji}$$

Combining everything, we get the following equation.

$$D_X f(A) = \frac{-1}{\det(X)} \det(X)(X^{-1})^T = -(X^{-1})^T$$

In the above equation, we can interpret the matrix $(X^{-1})^T$ as an $n^2 \times 1$ vector in $\mathbf{R}^{n^2}$, to make sure that both sides make sense in terms of dimensions (although the equation still makes sense if we interpret both sides as matrices). Since we have computed the entries of the derivative, the gradient is also computed.

Now, let us compute the Hessian. So, we need to compute the following mixed partial, for all $1 \leq i, j, l, m \leq n$.

$$\frac{\partial^2 f(X)}{\partial X_{ij} \partial X_{lm}} = \frac{\partial \left(-(X^{-1})_{ml}\right)}{\partial X_{ij}} = -\frac{\partial (X^{-1})_{ml}}{\partial X_{ij}}$$

Now, let e_{ij} be the $n \times n$ matrix which has all zeros except the ijth entry. From equation (†), we have that

$$dF(X^{-1}) = -X^{-1} e_{ij} X^{-1}$$

where above $dF(X^{-1})$ has been defined w.r.t e_{ij}. Observe that $dF(X^{-1})$ defined this way will satisfy

$$dF(X^{-1})_{lm} = \frac{\partial (X^{-1})_{lm}}{\partial X_{ij}}$$

So we see that

$$\frac{\partial (X^{-1})_{ml}}{\partial X_{ij}} = dF(X^{-1})_{ml} = (-X^{-1}e_{ij}X^{-1})_{ml} = -(X^{-1})_{mi}(X^{-1})_{jl}$$

and so we see that

$$\frac{\partial^2 f(X)}{\partial X_{ij}\partial X_{lm}} = (X^{-1})_{mi}(X^{-1})_{jl}$$

Hence, the Hessian has been computed as well (clearly the above formula shows that the Hessian is symmetric, which we know from calculus).

Let us now compute the second order approximation for this function. We have the following for any matrix X with positive determinant. Below, we are interpreting $X, \delta X$ and $\nabla f(X)$ as vectors in $\mathbf{R}^{n^2}$

$$\begin{aligned} f(X+\delta X) &= f(X) + \langle \nabla f(X), \delta X\rangle + \frac{1}{2}(\delta X)^T \nabla^2 f(X)(\delta X) \\ &= f(X) + \sum_{1\le i\le j\le n} \frac{\partial f(X)}{\partial X_{ij}} \cdot (\delta X)_{ij} + \frac{1}{2}\sum_{1\le i,j,l,m\le n} (\delta X)_{ij}\frac{\partial^2 f}{\partial X_{ij}\partial X_{lm}}(\delta X)_{lm} \\ &= f(X) + \sum_{1\le i\le j\le n} (-X^{-1})_{ji} \cdot (\delta X)_{ij} + \frac{1}{2}\sum_{1\le i,j,l,m\le n} (\delta X)_{ij}(X^{-1})_{mi}(X^{-1})_{jl}(\delta X)_{lm} \end{aligned}$$

So we've found the second order approximation.

We now consider part (2) of the problem. The function is defined as follows.

$$f(X) = \mathrm{Tr}(AX)$$

where A is a symmetric matrix and X a symmetric matrix of indeterminates. So, note that f is defined on the space of symmetric matrices; hence, the only independent variables are X_{ij} for $1 \le i \le j \le n$. So, f should be regarded as a function $f : \mathbf{R}^{\frac{n(n-1)}{2}} \to \mathbf{R}$.

Now, observe that the ijth entry of the product AX is the following.

$$(AX)_{ij} = \sum_{k=1}^{n} A_{ik}X_{kj}$$

So, it follows that the ith diagonal entry of this product is

$$(AX)_{ii} = \sum_{k=1}^{n} A_{ik}X_{ki}$$

and hence for input matrix X we have

$$f(X) = \mathrm{Tr}(AX) = \sum_{i=1}^{n}\sum_{k=1}^{n} A_{ik}X_{ki} = \sum_{1\le i,k\le n} A_{ik}X_{ki}$$

Because both X and A are symmetric, the above sum can be written as follows.

$$f(X) = \sum_{1\le i,k\le n} A_{ik}X_{ki} = \sum_{1\le i,k\le n} A_{ik}X_{ik} = \sum_{1\le i\le n} A_{ii}X_{ii} + \sum_{1\le i<k\le n} 2A_{ik}X_{ik}$$

Now, let us calculate the partial derivatives. First, for $1 \leq p \leq n$, the above equality gives us

$$\frac{\partial f(X)}{\partial X_{pp}} = A_{pp}$$

and for any $1 \leq p < q \leq n$, we have

$$\frac{\partial f(X)}{\partial X_{pq}} = 2A_{pq}$$

So, it follows that the gradient $\nabla f(X)$ at some symmetric matrix X is a *constant* function (because the partial derivatives are all constant). Clearly, this then implies that the Hessian $\nabla^2 f$ at all symmetric matrices X is 0.

Finally, let us compute the second order approximation of $f(X)$. Clearly, since the Hessian at all points is 0, the Hessian term in the second order approximation is zero. So, the second order approximation is just the following. Below, $X, \delta X$ and $\nabla f(X)$ are all interpreted as vectors in $\mathbf{R}^{\frac{n(n-1)}{2}}$.

$$\begin{aligned} f(X + \delta X) &= f(X) + \nabla f(X)^T (\delta X) \\ &= f(X) + \sum_{1 \leq p \leq q \leq n} \frac{\partial f(X)}{\partial X_{pq}} \cdot (\delta X)_{pq} \\ &= f(X) + \sum_{1 \leq p \leq n} A_{pp} (\delta X)_{pp} + \sum_{1 \leq p < q \leq n} 2A_{pq} (\delta X)_{pq} \end{aligned}$$

So finally, the second order approximation is the following.

$$\operatorname{Tr}(A(X + \delta X)) = \operatorname{Tr}(AX) + \sum_{1 \leq p \leq n} A_{pp} (\delta X)_{pp} + \sum_{1 \leq p < q \leq n} 2A_{pq} (\delta X)_{pq}$$

Problem 2. Let $f : \mathbf{R}^n \to \mathbf{R}$ be a twice differentiable function. Let $g : \mathbf{R} \to \mathbf{R}$ be defined as $g(t) := f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))$. Compute the first and second derivatives of g. Express $\int_0^1 g'' \, \mathrm{d}t$ in terms of the gradients ∇f of f at points $\boldsymbol{x}$ and $\boldsymbol{y}$.

Solution. First, we compute the derivative of g. Throughout, our interval of interest will be $[0, 1]$. Define the function $h : [0, 1] \to \mathbf{R}^n$ as follows.

$$h(t) = \boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})$$

So, it is immediately seen that

$$g = f \circ h$$

By the chain rule, we have the following for all $t \in [0, 1]$.

$$g'(t) = Df(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \times Dh(t)$$

Now, we see that

$$Dh(t) = \boldsymbol{y} - \boldsymbol{x}$$

and so

$$g'(t) = Df(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \times (\boldsymbol{y} - \boldsymbol{x})$$

By definition, we know that

$$Df(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) = (\nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})))^T$$

and so

$$g'(t) = (\nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})))^T \times (\boldsymbol{y} - \boldsymbol{x}) = \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle \tag{0.1}$$

Next, let us compute the second derivative g''. For a vector $\boldsymbol{u} \in \mathbf{R}^n$, let u_i denote the ith coordinate of $\boldsymbol{u}$. Equation (0.1) tells us that for all $t \in [0,1]$,

$$g'(t) = \sum_{i=1}^{n} \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))_i \cdot (\boldsymbol{y}_i - \boldsymbol{x}_i)$$

So, we see that for all $t \in [0,1]$,

(0.2)

$$g''(t) = \sum_{i=1}^{n} \frac{d}{dt} \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))_i \cdot (\boldsymbol{y}_i - \boldsymbol{x}_i) = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{x}_i) \frac{d}{dt} \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))_i$$

Now, let $1 \leq i \leq n$ be fixed. Define the function $p_i : [0,1] \to \mathbf{R}$ as follows.

$$p_i(t) = \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))_i = \frac{\partial f}{\partial x_i}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) = \left(\frac{\partial f}{\partial x_i} \circ h \right)(t)$$

Let us compute p_i', again using the chain rule. We have the following for all $t \in [0,1]$.

$$p_i'(t) = D\frac{\partial f}{\partial x_i}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \times Dh(t) = D\frac{\partial f}{\partial x_i}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \times (\boldsymbol{y} - \boldsymbol{x})$$

As before, we have

$$D\frac{\partial f}{\partial x_i}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) = \left(\nabla \frac{\partial f}{\partial x_i}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \right)^T$$

and hence the last two equations combined give us the following.

$$p_i'(t) = \left\langle \nabla \frac{\partial f}{\partial x_i}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \right\rangle = \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_j \partial x_i}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \cdot (\boldsymbol{y}_j - \boldsymbol{x}_j)$$

Because f is twice differentiable, we know that the mixed partial derivatives of f are equal, i.e for any i, j,

$$\frac{\partial^2 f}{\partial x_j \partial x_i} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

So, the last equation gives us

$$p_i'(t) = \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \cdot (\boldsymbol{y}_j - \boldsymbol{x}_j)$$

Let us now plug in this value of $p_i'(t)$ in equation (0.2). Doing so, we get the following.

$$g''(t) = \sum_{i=1}^{n} \sum_{j=1}^{n} (\boldsymbol{y}_i - \boldsymbol{x}_i) \cdot \frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \cdot (\boldsymbol{y}_j - \boldsymbol{x}_j) \tag{0.3}$$

$$= \left\langle \nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \times (\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \right\rangle \tag{0.4}$$

Next, by the fundamental theorem of calculus, we know the following.

$$\int_0^1 g''(t)\ \mathrm{d}t = g'(1) - g'(0) = \langle \nabla f(\boldsymbol{y}), \boldsymbol{y} - \boldsymbol{x} \rangle - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle = \langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

and hence we've computed all the required quantities.

Problem 3. Let f be a convex function and let K be a closed convex set. Suppose $\boldsymbol{x}^*$ is the minimizer of f on K. Show that $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle \geq 0$ for all $\boldsymbol{y} \in K$.

Solution. For the sake of contradiction, suppose the claim is false, i.e there is some $\boldsymbol{y} \in \mathcal{K}$ such that

$$\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle < 0$$

Next, define the function $g : [0, 1] \to \mathbf{R}$ by the following.

$$g(t) = f((1-t)\boldsymbol{x}^* + t\boldsymbol{y})$$

Then, observe that

$$g(0) = f(\boldsymbol{x}^*)$$

Also, we have the following by the chain rule, for all $t \in [0, 1]$.

$$g'(t) = \langle \nabla f((1-t)\boldsymbol{x}^* + t\boldsymbol{y}), \boldsymbol{y} - \boldsymbol{x}^* \rangle$$

Above, $g'(0)$ should be interpreted as a one-sided limit, where $t \to 0^+$. We immediately see that

$$g'(0) = \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle < 0$$

Now, by the definition of the derivative,

$$g'(0) = \lim_{t \to 0^+} \frac{g(t) - g(0)}{t} < 0$$

This means that, for sufficiently small $t \in (0, 1)$,

$$\frac{g(t) - g(0)}{t} < 0$$

which implies that

$$g(t) - g(0) < 0$$

and hence

$$g(t) < g(0)$$

But, this means that

$$f((1-t)\boldsymbol{x}^* + t\boldsymbol{y}) < f(\boldsymbol{x}^*)$$

which contradicts the fact that $\boldsymbol{x}^*$ is the minimizer of f.

Problem 4. Show that the set of PSD matrices is convex.

Solution. Let X, Y be two positive semi-definite $n \times n$ matrices, and let $t \in (0, 1)$. we will show that $tX + (1-t)Y$ is also positive semi-definite. To that end, let $\boldsymbol{x} \in \mathbf{R}^n$ be any point. Then, we have the following.

$$\begin{aligned}
\langle (tX + (1-t)Y)\boldsymbol{x}, \boldsymbol{x} \rangle &= \langle tX\boldsymbol{x} + (1-t)Y\boldsymbol{x}, \boldsymbol{x} \rangle \\
&= \langle tX\boldsymbol{x}, \boldsymbol{x} \rangle + \langle (1-t)Y\boldsymbol{x}, \boldsymbol{x} \rangle \\
&= t \langle X\boldsymbol{x}, \boldsymbol{x} \rangle + (1-t) \langle Y\boldsymbol{x}, \boldsymbol{x} \rangle \\
&\geq 0
\end{aligned}$$

because $t \in (0, 1)$, and X, Y are positive semi-definite. This completes the proof of the claim.

Problem 5. Is the negative entropy function $\sum_i x_i \log(x_i)$ restricted to the positive orthant with vectors of norm atmost G strongly convex?

Solution. Let the negative entropy function be denoted by f. Let $\mathbf{R}^n_+$ denote the positive orthant (which is an open set). Then $f : \mathbf{R}^n_+ \to \mathbf{R}$ is defined as follows.

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \log(x_i)$$

We claim that f is twice differentiable, and we will show this by actually computing the derivative. First, if $1 \leq i \leq n$, then note that

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}) = 1 + \log(x_i)$$

and since $x_i > 0$ for all i (because $\boldsymbol{x} \in \mathbf{R}^n_+$), we see that all partial derivatives exist at all points. Hence, the function f is differentiable. Next, let $1 \leq i, j \leq n$. If $i = j$, we have

$$\frac{\partial^2 f}{\partial x_i^2}(\boldsymbol{x}) = \frac{1}{x_i}$$

and this exists because $x_i > 0$. If $i \neq j$, we have

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}(1 + \log(x_j)) = 0$$

and hence this exists too. So, f is twice differentiable, because all mixed partials exist. Infact, we have shown above that $\nabla^2 f(\boldsymbol{x})$ is a diagonal matrix at all points $\boldsymbol{x} \in \mathbf{R}^n_+$.

Now, we check whether f is strongly convex in the given domain, which is the intersection $\mathbf{R}^n_+ \cap B(\mathbf{0}, G)$. To check whether f is strongly convex, it is enough to check if for some $\alpha \in \mathbf{R}$, the Hessian $\nabla^2 f(\boldsymbol{x})$ satisfies the following for all $\boldsymbol{x} \in \mathbf{R}^n_+ \cap B(\mathbf{0}, G)$.

$$\nabla^2 f(\boldsymbol{x}) \succcurlyeq \alpha I$$

We claim that $\alpha = G$ works. To show this, let $\boldsymbol{x} \in \mathbf{R}^n_+ \cap B(\mathbf{0}, G)$. This means that for each i,

$$0 < x_i \leq G \implies \frac{1}{x_i} \geq \frac{1}{G}$$

Now, we have shown above that the Hessian $\nabla^2 f(\boldsymbol{x})$ is a diagonal matrix, and the ith diagonal entry is given by the following.

$$[\nabla^2 f(\boldsymbol{x})]_{ii} = \frac{1}{x_i} \geq \frac{1}{G}$$

Now, all eigenvalues of the Hessian are the diagonal entries, i.e the eigenvalues are $\frac{1}{x_i}$ for $1 \leq i \leq n$. So, all eigenvalues are atleast $\frac{1}{G}$, which means that

$$\nabla^2 f(\boldsymbol{x}) \succcurlyeq \frac{1}{G} I$$

Hence, it follows that f is $\frac{1}{G}$-strongly convex.

Problem 6. Let χ be a Euclidean vector space with norm $\langle ., . \rangle$. We say $f : \chi \to \mathbf{R} \cup \infty$ is α-*strongly convex* with respect to norm $||\cdot||$ if for all $\boldsymbol{x}$, $\boldsymbol{y}$ and $t \in (0, 1)$ we have the following.

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) - \frac{1}{2}\alpha t(1-t)\,||\boldsymbol{x} - \boldsymbol{y}||^2$$

f is said to be β-*smooth* with respect to $||\cdot||$ if

$$f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} \rangle + \frac{\beta}{2}\,||\boldsymbol{y}||^2$$

For a function f define the *conjugate* of f by

$$f^* := \max_{\boldsymbol{x}} \langle \boldsymbol{x}, \boldsymbol{y} \rangle - f(\boldsymbol{x})$$

Let f be a closed convex function. Show that f is α-strongly convex with respect to $||\cdot||$ if and only if f^* is $\frac{1}{\alpha}$-smooth with respect to the dual norm $||\cdot||_*$. For $\boldsymbol{y} \in \chi$, the dual norm is defined as follows.

$$||\boldsymbol{y}||_* = \sup_{||\boldsymbol{x}|| \leq 1} \langle \boldsymbol{y}, \boldsymbol{x} \rangle$$

Solution. Didn't get time to do this.