Optimisation Coursework

CIDs: 02044064, 02045918

November 2023

Part I: Unconstrained Optimisation

Question 1i:

No, the function f is not coercive. By counterexample, consider moving along the path $x = y^2$ as $y \to \infty$:

$$f(x,y) = (y^2)^2 - 2(y^2)y^2 + \frac{1}{2}y^4$$

$$= y^4 - 2y^4 + \frac{1}{2}y^4$$

$$= -\frac{1}{2}y^4$$

Clearly $\lim_{y\to\infty} f(y^2, y) = -\infty$, and so $f(x,y)$ cannot be coercive.

Question 1ii:

Firstly we find the gradient of $f(x,y)$. We compute the partial derivatives:

$$\frac{\partial f}{\partial x} = 2x - 2y^2$$

$$\frac{\partial f}{\partial y} = -4xy + 2y^3$$

So, the complete gradient is:

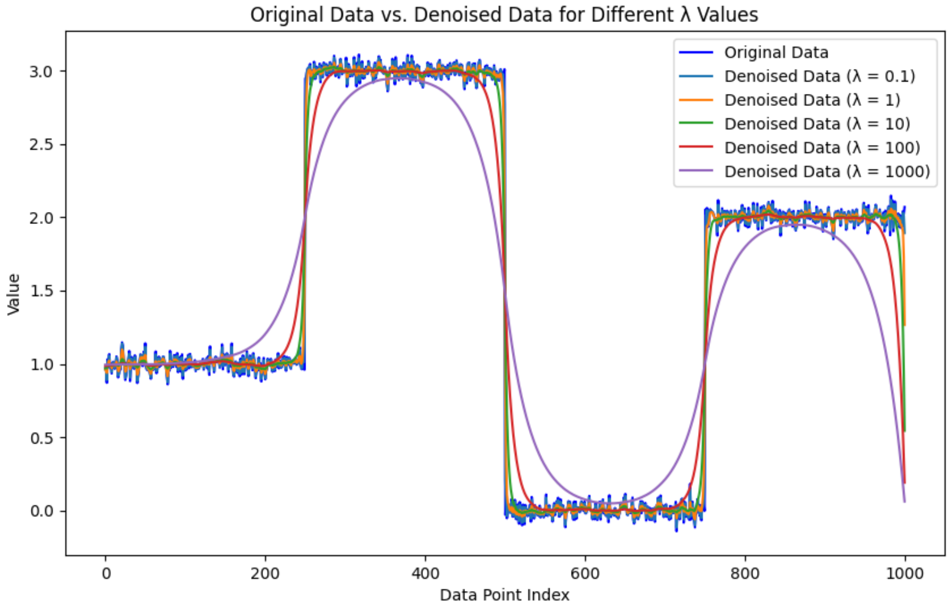
$$\nabla f(x,y) = (2x - 2y^2, -4xy + 2y^3)^T$$

To find the stationary points, we set the gradient to $(0,0)^T$ and solve for possible (x,y) pair solutions. In setting the first partial derivative with respect to x to 0, we obtain the constraint $x = y^2$, and substituting this into the second equation above gives us $2y^3 = 0$. Clearly, the only stationary point is $(0,0)$. To classify this point, first consider the path (α^2, α) where $\alpha \to 0$. From part (a), we know that $f(\alpha^2, \alpha) = -\frac{1}{2}\alpha^4 < 0 = f(0,0)$ as $\alpha \to 0$. Now consider the path $(-\alpha^2, \alpha)$ where $\alpha \to 0$. We have $f(-\alpha^2, \alpha) = \alpha^4 + 2\alpha^4 + \frac{1}{2}\alpha^4 = \frac{7}{2}\alpha^4 > 0 =$ f(0, 0) as $\alpha \to 0$. Thus, $(0,0)$ is neither a local minimum nor a local maximum so it must be a saddle point.

Part II: Linear Least Squares - Denoising

Question 2i:

A is the 1000x1000 identity matrix as we are interested in $\|x-b\|^2$, where b is a 1000-dimensional vector representing the dataset - in particular, the ith entry is the ith value of the dataset in the txt file. L is a 999x1000 matrix with 1s on the main diagonal, -1s on the superdiagonal and 0s everywhere else, because $Lx = (x_0 - x_1, x_1 - x_2, ..., x_{999} - x_{1000})$ so $\|Lx\|^2 = \sum_{i=1}^{999}(x_i - x_{i+1})^2$.



We observe that as λ increases, the curves become more and more smooth but when λ is too high, the $\|Ax-b\|^2$ term, which measures the departure between the denoised and original data, becomes too insignificant, so when $\lambda = 1000$, the denoised data is far off from the original data.

Question 2iia:

First, rewrite the problem in matrix-vector form:

$$\underbrace{\sum_{i=1}^{1999} w_i |\hat{a}_i^T x - \hat{b}_i|^2}_{f(x)} = \sum_{i=1}^{1999} w_i (\hat{a}_i^T x - \hat{b}_i)^2 \text{ as } \hat{a}_i^T x - \hat{b}_i \in \mathbb{R}$$

$$= \sum_{i=1}^{1999} w_i [\hat{A}x - \hat{b}]_i^2$$

$$= (\hat{A}x - \hat{b})^T \underbrace{\text{diag}(w_1, ..., w_{1999})}_{W} (\hat{A}x - \hat{b}) \quad (1)$$

Before computing the gradient, we need to first expand out the terms in (1):

$$(\hat{A}x - \hat{b})^T W (\hat{A}x - \hat{b}) = (\hat{A}x)^T W \hat{A}x - \hat{b}^T W \hat{A}x - x^T \hat{A}^T W \hat{b} + \hat{b}^T W \hat{b}$$

$$= x^T \hat{A}^T W \hat{A}x - (x^T \hat{A}^T W \hat{b})^T - x^T \hat{A}^T W \hat{b} + \hat{b}^T W \hat{b}$$

$$= x^T \hat{A}^T W \hat{A}x - 2x^T \hat{A}^T W \hat{b} + \hat{b}^T W \hat{b}$$

where in the last step we have used the property that a scalar equals its transpose.
By standard results of matrix vector differentiation,

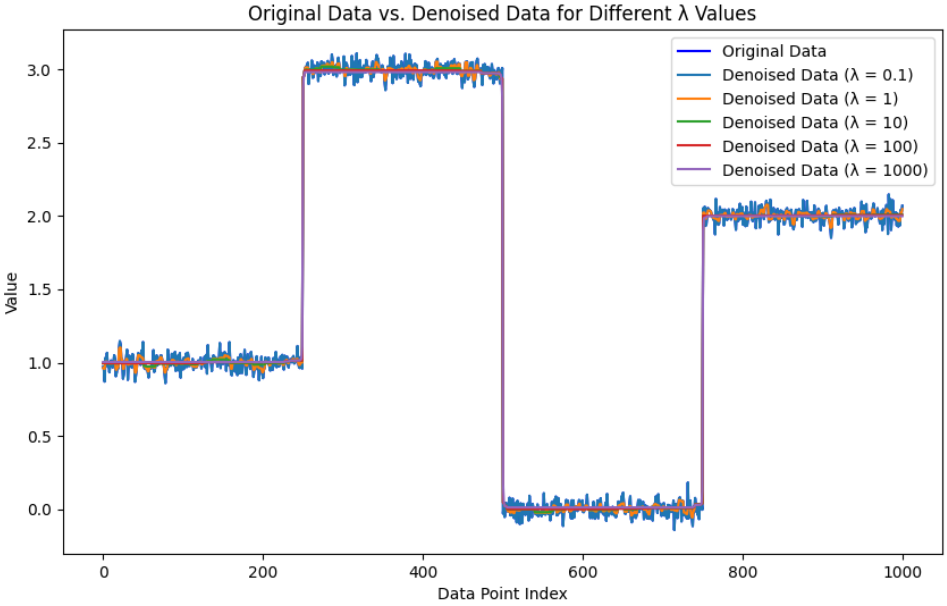
$$\nabla f(x) = 2\hat{A}^T W \hat{A} x - 2\hat{A}^T W \hat{b}$$
$$\nabla f(x^*) = 0 \iff \hat{A}^T W \hat{A} x^* = A^T W \hat{b} \iff x^* = (\hat{A}^T W \hat{A})^{-1} \hat{A}^T W \hat{b}$$

if $\hat{A}^T W \hat{A}$ is invertible. To prove that this stationary point is indeed a minimum, we compute the Hessian $H(x)$ by computing the partial derivatives for each row of the gradient vector. To show that this is a global minimum, we know from the global optimality conditions in the lecture notes that it is enough to show $H(x)$ is positive semidefinite for all $x \in \mathbb{R}^{1000}$, meaning that f is a convex function. We have:

$H(x) = 2\hat{A}^T W \hat{A} = 2\sum_{i=1}^{1999} w_i \hat{a}_i^T \hat{a}_i = 2\sum_{i=1}^{1999} w_i \|\hat{a}_i\|^2 \geq 0$ where $\hat{a}_i$ denotes the ith column of $\hat{A}$ so the Hessian is positive semidefinite.

Question 2iib:

Below we plot the solution of our iteratively reweighted least squares algorithm:



We notice that for large values of λ, the fitting is both smooth and does not deviate significantly from the data points like before, which makes sense since

$$|\hat{a}_i^T x^k - \hat{b}_i| = \begin{cases} |x_i^k - \hat{b}_i| & \text{if } 1 \leq i \leq 1000, \\ |x_i^{k-1000} - x_i^{k-999}| & \text{if } 1001 \leq i \leq 1999. \end{cases}$$

and clearly w_i^{k+1} penalises large differences between either the fitted points and the actual data points with the same index i or between consecutive fitted points. As a result, less weight is given to the three sudden jumps in the fitted points for the next iteration, which is exactly what we want, because otherwise these jumps will dominate the objective function, leading to the sinusoidal shape for $\lambda = 1000$ like before.

Part III: Gradient Descent

Question 3i:

We need to show that for an arbitrary vector $v \in R^m$, $v^T \nabla^2 f(x) v \succcurlyeq 0$. First note that $v^T \nabla^2 f(x) v = v^T A^T \nabla^2 g(Ax - b) Av = (Av)^T \nabla^2 g(Ax - b) Av = w^T \nabla^2 g(Ax - b) w$ where $w \triangleq Av$.

We will now derive an expression for $\nabla^2 g(Ax - b)$:

$$
\begin{aligned}
g(y) &= \sum_{i=1}^{m} (y_i^2 + \eta^2)^{1/2} \\
\implies [\nabla g(y)]_i &= \frac{1}{2}(y_i^2 + \eta^2)^{-1/2} 2y_i \\
&= y_i (y_i^2 + \eta^2)^{-1/2} \\
\implies [\nabla^2 g(y)]_{i,j} &= \begin{cases} (y_i^2 + \eta^2)^{-1/2} + y_i(-\frac{1}{2})(y_i^2 + \eta^2)^{-3/2}(2y_i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
\text{Simplify: } [\nabla^2 g(y)]_{i,i} &= (y_i^2 + \eta^2)^{-1/2} - (y_i^2 + \eta^2)^{-3/2} y_i^2 \\
&= \frac{y_i^2 + \eta^2 - y_i^2}{(y_i^2 + \eta^2)^{3/2}} \\
&= \frac{\eta^2}{(y_i^2 + \eta^2)^{3/2}} > 0. \\
\implies \nabla^2 g(y) &= \text{diag}\left(\frac{\eta^2}{(y_1^2 + \eta^2)^{3/2}}, \frac{\eta^2}{(y_2^2 + \eta^2)^{3/2}}, ..., \frac{\eta^2}{(y_m^2 + \eta^2)^{3/2}} \right) \quad (2) \\
\implies v^T \nabla^2 f(x) v &= w^T \nabla^2 g(Ax - b) w \\
&= w^T \text{diag}\left(\frac{\eta^2}{(y_1^2 + \eta^2)^{3/2}}, \frac{\eta^2}{(y_2^2 + \eta^2)^{3/2}}, ..., \frac{\eta^2}{(y_m^2 + \eta^2)^{3/2}} \right) w \\
&= \sum_{i=1}^{m} \frac{\eta^2 w_i^2}{(y_i^2 + \eta^2)^{3/2}} \geq 0. \ \text{(with equality when } v = 0 \text{ because then } w = A(0) = 0)
\end{aligned}
$$

As v was arbitrary, the above holds for all v, so $v^T \nabla^2 f(x) v \succcurlyeq 0$.

Question 3iiB:

Recall the Equivalence to Boundedness of Hessian Theorem from lectures:
$f \in C_L^{1,1} \iff \|\nabla^2 f(x)\| \leq L$ for all $x \in R^m$. It therefore suffices to show that $\|\nabla^2 f(x)\| \leq L$ for all $x \in R^m$. We have:
$\|\nabla^2 f(x)\| = \|A^T \nabla^2 g(Ax - b) A\| \leq \|A^T\| \|\nabla^2 g(Ax - b)\| \|A\|$ by the submultiplicative property of operator norms, as hinted in the question.

Claim: $\|A^T\| = \|A\|$.

Proof of claim: It is a standard result from linear algebra that every real matrix A has a real singular value decomposition $A = U\Sigma V^T$ where $U \in R^{mxm}$ is orthogonal, $\Sigma \in R^{mxn}$ has nonnegative real values $\sigma_1, \sigma_2, ..., \sigma_{\min\{m,n\}}$ on the diagonal which are the singular values of A, and $V \in R^{nxn}$ is orthogonal. It follows that $A^T = (U\Sigma V^T)^T = V\Sigma^T U^T$ and the main diagonal entries of Σ^T are the same as those of Σ, meaning that A^T has the same singular values as A, and therefore $\|A^T\| = \sigma_{max}(A^T) = \sigma_{max}(A) = \|A\|$.

It follows that $\|\nabla^2 f(x)\| \leq \|A\|^2 \|\nabla^2 g(Ax-b)\|$. It remains to show that $\|\nabla^2 g(Ax-b)\| \leq \frac{1}{\eta}$:

$$\begin{aligned}
\|\nabla^2 g(Ax-b)\| &= \sup_{\|v\|=1} \|\nabla^2 g(\underbrace{Ax-b}_{y})v\| \\
&= \sup_{\|v\|=1} \|\mathrm{diag}(\frac{\eta^2}{(y_1^2+\eta^2)^{3/2}}, ..., \frac{\eta^2}{(y_m^2+\eta^2)^{3/2}})v\| \text{ using result (2) from Q3i} \\
&= \eta^2 \sup_{\|v\|=1} \|\mathrm{diag}(\frac{1}{(y_1^2+\eta^2)^{3/2}}, ..., \frac{1}{(y_m^2+\eta^2)^{3/2}})v\| \\
&= \eta^2 \sup_{\|v\|=1} \sqrt{\frac{v_1^2}{(y_1^2+\eta^2)^3} + \ldots + \frac{v_m^2}{(y_m^2+\eta^2)^3}} \\
&\leq \eta^2 \sup_{\|v\|=1} \sqrt{\frac{v_1^2}{(\eta^2)^3} + \ldots + \frac{v_m^2}{\eta^2)^3}} \text{ because } y_i^2 \in \mathbb{R} \text{ so } y_i^2 + \eta^2 \geq \eta^2 \\
&= \frac{\eta^2}{\eta^3} \sup_{\|v\|=1} \sqrt{v_1^2 + \ldots + v_m^2} \\
&= \frac{1}{\eta} \sup_{\|v\|=1} \|v\| \\
&= \frac{1}{\eta}
\end{aligned}$$

Putting it all together, $\|\nabla^2 f(x)\| \leq \frac{\|A\|^2}{\eta}$.