Problem 1. In our proof of local optimality implying global optimality of convex functions, we define $\lambda = 1 - \frac{\varepsilon}{2||y-x||}$ and $z = \lambda x + (1 - \lambda)y$. Prove that z is indeed in the ε-neighbor of x.

Hint: Calculate the distance between x and z, and show it's less than ε.

Answer: (The answer essentially follows the lecture.) Note that

$$z - x = (\lambda x + (1 - \lambda)y) - x = (1 - \lambda)(x - y).$$

Taking the norm on both sides gives us the desired result:

$$||z - x|| = (1 - \lambda)\,||x - y|| = (1 - (1 - \frac{\varepsilon}{2||y-x||}))\,||x - y|| = \frac{\varepsilon}{2\,||y-x||}||y - x|| = \varepsilon/2.$$

Problem 2.

Suppose f is a convex function and we have the gradient $\bigtriangledown f(\boldsymbol{x}) = \boldsymbol{0}$ at the point $\boldsymbol{x}$. Prove that $\boldsymbol{x}$ is a global optimum of the function f.

Answer: Since $f(\cdot):\mathbb{R}\longrightarrow\mathbb{R}$ is a convex function, from the first-order condition (see the lecture on convexity), we know that

$$\forall \boldsymbol{y} \in \text{dom } f, \qquad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + [\bigtriangledown f(\boldsymbol{x})]^T(\boldsymbol{y} - \boldsymbol{x}).$$

So if the gradient at $\boldsymbol{x}$ is equal to zero, i.e. $\bigtriangledown f(\boldsymbol{x}) = 0$, we get that

$$\forall \boldsymbol{y} \in \text{dom } f, \qquad f(\boldsymbol{y}) \geq f(\boldsymbol{x}),$$

That is, $\boldsymbol{x}$ is a global minimum of f.

Problem 3.

In future lectures, we will use the l_2-penalized square error as the training objective, given by

$$J = \frac{1}{2M}\sum_{m=1}^{M}\left(\sum_{i=0}^{d} w_i x_i^{(m)} - t^{(m)}\right)^2 + \lambda\sum_{i=0}^{d} w_i^2$$

where λ is a positive constant.

a) Express the loss in the vector form using $\mathbf{X} \in \mathbb{R}^{M\times(d+1)}$, $\boldsymbol{w} \in \mathbb{R}^{d+1}$ and $\boldsymbol{t} \in \mathbb{R}^{M}$

b) Compute the gradient $\bigtriangledown J(\boldsymbol{w})$ and the Hessian $\bigtriangledown\bigtriangledown J(\boldsymbol{w})$ and show that J is convex in $\boldsymbol{w}$.

c) Derive a closed-form solution to the problem.

Hints:

Suppose $\boldsymbol{S}$ is a symmetric matrix.

$$\frac{\partial \boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{S}\boldsymbol{x} \qquad \frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^\top,$$

References: https://en.wikipedia.org/wiki/Matrix_calculus

Note: When matrix calculus is needed in exams, I will give the precise formulas that should be used for solving the problem.

Answer:

a) Let $x^{(m)} \in \mathbb{R}^{d+1}$ be the vector defined as $(x^{(m)})_i = x_i^{(m)}$. Note that

$$\sum_{i=0}^{d} w_i^2 = w^\top w = ||w||_2^2 \qquad \text{and} \qquad \sum_{i=0}^{d} w_i x_i^{(m)} = w^\top x^{(m)}.$$

Then the loss becomes

$$J = \frac{1}{2M}\sum_{m=1}^{M}(w^\top x^{(m)} - t^{(m)})^2 + \lambda||w||_2^2.$$

Notice that the individual terms in the above summation is the mth component of the vector $Xw - t$, and the summation itself is the norm of this vector. We thus obtain our final solution:

$$J = \frac{1}{2M}||Xw - t||_2^2 + \lambda||w||_2^2.$$

b) (Traditionally, $\frac{\partial}{\partial w}J$ is assumed to be a row vector with its ith element defined as $(\frac{\partial}{\partial w}J)_i = \frac{\partial}{\partial w_i}J$, while the gradient $\nabla J = (\frac{\partial}{\partial w}J)^\top$ is a column vector.)

Using the hint given in the question, it is clear that the $\nabla||w||_2^2 = 2w$, and

$$\nabla||Xw - t||_2^2 = \frac{\partial(Xw-t)}{\partial w} \cdot \frac{\partial}{\partial(Xw-t)}||Xw - t||_2^2 = 2X^\top(Xw - t).$$

Combining these two equations, we obtain the gradient:

$$\nabla J = \frac{1}{M}X^\top(Xw - t) + 2\lambda w.$$

Computing the Hessian is now straightforward.

$$\nabla^2 J = \frac{\partial}{\partial w}\nabla J = \frac{1}{M}X^\top X + 2\lambda I.$$

c) The problem to solve is $min_w J$. To solve this, we set the gradient of the objective to zero, and then solve for w:

$$\nabla J = 0 \quad \Rightarrow \quad (X^\top X + 2M\lambda I)w^* = X^\top t \quad \Rightarrow \quad w^* = (X^\top X + 2M\lambda I)^{-1}X^\top t.$$

Note that since the Hessian is PSD (why?), MSE is a convex function (also see the notes for Lecture 5). This means that the above solution w^* is a global minimum of the MSE problem. (Note that, for all $M > 0$ and $\lambda > 0$, the matrix $X^\top X + 2M\lambda I$ is guaranteed to be non-singular (why?).

END