**MLISP: Machine Learning in Signal Processing**

# Solutions to problem set 1

*Prof. V. I. Morgenshtern*

*Solver: S. Lotter and V. I. Morgenshtern*

## Problem 1: Gradients and Hessians

1. Using that $\mathbf{A}$ is symmetric, we obtain:

$$\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_j x_i$$

$$\implies \frac{1}{2}\left(\nabla_x\, \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}\right)_k = \frac{1}{2}\frac{\partial \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\partial x_k} = \frac{1}{2}\left(\sum_{i=1}^{n} A_{ik}x_i + \sum_{j=1}^{n} A_{kj}x_j\right) = \sum_{i=1}^{n} A_{ki}x_i = (\mathbf{A}\mathbf{x})_k$$

$$\implies \frac{1}{2}\nabla_x\, \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}.$$

Also:

$$\mathbf{b}^\mathsf{T}\mathbf{x} = \sum_{i=1}^{n} b_i x_i$$

$$\implies \left(\nabla_x\, \mathbf{b}^\mathsf{T}\mathbf{x}\right)_k = \frac{\partial \mathbf{b}^\mathsf{T}\mathbf{x}}{\partial x_k} = b_k$$

$$\implies \nabla_x\, \mathbf{b}^\mathsf{T}\mathbf{x} = \mathbf{b}.$$

Therefore, $\nabla_x f(x) = \mathbf{A}\mathbf{x} + \mathbf{b}$ as desired.

2. Using the chain rule for derivative, we can calculate each element of the gradients as follows:

$$[\nabla_x f(x)]_k = \frac{\partial f(x)}{\partial x_k} = \frac{\partial g(h(x))}{\partial x_k} = \frac{\partial g(y)}{\partial y}\frac{\partial h(x)}{\partial x_k}.$$

This implies: $\nabla_x f(x) = g'(h(x))\,\nabla_x h(\mathbf{x})$.

3. Using the result in point 1, we can calculate each component of the Hessian as follows:

$$\left[\nabla_x^2 f(x)\right]_{ij} = \frac{\partial [\nabla_x f(x)]_j}{\partial x_i} = \frac{\partial(\sum_{k=1}^{n} A_{jk}x_k + b_j)}{\partial x_i} = A_{ji} = A_{ij}.$$

Therefore, $\nabla_x^2 f(x) = \mathbf{A}$.

4. Using the results in point 2 and then the result in point 1 we obtain:

$$\nabla_{\mathbf{x}}\, g(\mathbf{a}^\mathsf{T}\mathbf{x}) = g'(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}.$$

Taking the derivatives again, we get:

$$\nabla_{\mathbf{x}}^2\, g(\mathbf{a}^\mathsf{T}\mathbf{x}) = g''(\mathbf{a}^\mathsf{T}\mathbf{x})\mathbf{a}\mathbf{a}^\mathsf{T}.$$

**Problem 2: Positive definite matrices**

1. First check that $\mathbf{z}\mathbf{z}^\mathsf{T}$ is symmetric:

$$\left(\mathbf{z}\mathbf{z}^\mathsf{T}\right)^\mathsf{T} = \left(\mathbf{z}^\mathsf{T}\right)^\mathsf{T}\mathbf{z}^\mathsf{T} = \mathbf{z}\mathbf{z}^\mathsf{T}.$$

   Now, let's check that it is positive semidefinite. Take $\mathbf{x} \in \mathbb{R}^n$, then

$$\mathbf{x}^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{x} = \left(\mathbf{z}^\mathsf{T}\mathbf{x}\right)^\mathsf{T}\left(\mathbf{z}^\mathsf{T}\mathbf{x}\right) = \left\|\mathbf{z}^\mathsf{T}\mathbf{x}\right\|_2^2 \geq 0.$$

2. By definition, $\mathbf{x} \in \mathcal{N}(\mathbf{A})$ iff $\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{x} = \mathbf{0}$. This is possible iff $\mathbf{z}^\mathsf{T}\mathbf{x} = 0$. Therefore,

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{z}^\mathsf{T}\mathbf{x} = 0\}.$$

   These are all the vectors in $\mathbb{R}^n$ that are perpendicular to $\mathbf{z}$.

   All columns of $\mathbf{z}\mathbf{z}^\mathsf{T}$ are collinear and are just a multiple of $\mathbf{z}$. Therefore, $\text{rank}(\mathbf{A}) = 1$.

3. First, check the symmetry: $\left(\mathbf{B}\mathbf{A}\mathbf{B}^\mathsf{T}\right)^\mathsf{T} = \mathbf{B}\left(\mathbf{B}\mathbf{A}\right)^\mathsf{T} = \mathbf{B}\mathbf{A}^\mathsf{T}\mathbf{B}^\mathsf{T} = \mathbf{B}\mathbf{A}\mathbf{B}^\mathsf{T}$. Next, let $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} = \mathbf{B}^\mathsf{T}\mathbf{x} \in \mathbb{R}^n$. Then, because $\mathbf{A}$ is PSD, $\mathbf{x}^\mathsf{T}\mathbf{B}\mathbf{A}\mathbf{B}^\mathsf{T}\mathbf{x} = \mathbf{y}^\mathsf{T}\mathbf{A}\mathbf{y} \geq 0$. Therefore, $\mathbf{B}\mathbf{A}\mathbf{B}^\mathsf{T}$ is PSD.

**Problem 3: Eigenvectors, eigenvalues, and the spectral theorem**

1. Let $\mathbf{e}_i \in \mathbb{R}^n$ denote the $i$th canonical basis vector of $\mathbb{R}^n$, $[\mathbf{e}_i]_j = \delta[i-j]$. Then,

$$\mathbf{A}\mathbf{t}_i = \mathbf{T}\Lambda\mathbf{T}^{-1}\mathbf{t}_i = \mathbf{T}\Lambda\mathbf{T}^{-1}\mathbf{T}\mathbf{e}_i = \mathbf{T}\Lambda\mathbf{e}_i = \mathbf{T}\lambda_i\mathbf{e}_i = \lambda_i\mathbf{t}_i.$$

2. Follows directly from the problem before, as $\mathbf{U}^{-1} = \mathbf{U}^\mathsf{T}$.

3. Suppose there exists $i \in \{1 \dots n\}$ with $\lambda_i < 0$. Let $\mathbf{v}_i$, $\mathbf{v}_i \neq 0$ be a corresponding eigenvector. Then $\mathbf{v}_i^\mathsf{T}\mathbf{A}\mathbf{v}_i = \mathbf{v}_i^\mathsf{T}\lambda_i\mathbf{v}_i = \lambda_i \left\|v_i\right\|_2^2 < 0$. But as $\mathbf{A}$ is PSD $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$. This is a contradiction and therefore $\lambda_i \geq 0$, for all $i \in \{1 \dots n\}$.

**Problem 5: Linear regression and gradient descent**

See `ps1_sol_problem5.ipynb` file.

**Problem 6: High-dimensional regression (exam practice)**

1. The least squares solution is defined as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

   where

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\mathbf{X}\boldsymbol{\theta} - \mathbb{y})^\mathsf{T}(\mathbf{X}\boldsymbol{\theta} - \mathbb{y}).$$

To minimize $J(\boldsymbol{\theta})$, let's calculate $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \frac{1}{2}(\mathbf{X}\boldsymbol{\theta} - \mathbb{y})^{\mathsf{T}}(\mathbf{X}\boldsymbol{\theta} - \mathbb{y}) \\
&= \frac{1}{2} \nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbb{y} - \mathbb{y}^{\mathsf{T}}\mathbf{X}\boldsymbol{\theta} + \mathbb{y}^{\mathsf{T}}\mathbb{y}) \\
&= \frac{1}{2} \nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\theta} - 2\boldsymbol{\theta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbb{y} + \mathbb{y}^{\mathsf{T}}\mathbb{y}) \\
&= \frac{1}{2}(2\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\theta} - 2\mathbf{X}^{\mathsf{T}}\mathbb{y}).
\end{aligned}$$

To minimize $J(\boldsymbol{\theta})$ we set $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbf{0}$ which yields:

$$\mathbf{X}^{\mathsf{T}}(\mathbf{X}\hat{\boldsymbol{\theta}} - \mathbb{y}) = \mathbf{0}.$$

Above, $\hat{\boldsymbol{\theta}}$ denotes the optimal value for $\boldsymbol{\theta}$, i.e. the one that minimizes $J(\boldsymbol{\theta})$.

Rearranging the terms we obtain the normal equations:

$$\mathbf{X}^{\mathsf{T}}\mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}^{\mathsf{T}}\mathbb{y}.$$

So the value of $\hat{\boldsymbol{\theta}}$ that minimizes $J(\boldsymbol{\theta})$ is given in closed form by the equation:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbb{y}.$$

2. The $p \times p$ matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ needs to be full-rank.

   This is possible when all the columns of $\mathbf{X}$ are linearly independent. The necessary condition is $n \geq p$.

   If $n < p$, $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ cannot be full-rank. Here is why:

$$\text{rank}(\mathbf{X}^{\mathsf{T}}\mathbf{X}) \leq \min(\text{rank}(\mathbf{X}^{\mathsf{T}}), \text{rank}(\mathbf{X})) \leq \min(n, p) = n < p. \tag{1}$$

3. First note,

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbb{y} \\
&= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}(\mathbf{X}\boldsymbol{\theta} + \mathbb{m}) \\
&= \boldsymbol{\theta} + (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbb{m}.
\end{aligned}$$

   Therefore,

$$\mathbb{E}\left[\hat{\boldsymbol{\theta}}\right] = \boldsymbol{\theta} + \mathbb{E}\left[(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbb{m}\right] = \boldsymbol{\theta}.$$