CMSC 25400: Machine Learning **Winter 2019**

Recitation 2: Matrix Derivatives

2.1 Minimizing/Maximizing Functions

In machine learning, many problems we'll want to solve can be cast as minimizing (or maximizing) a function $J : \mathbb{R}^d \to \mathbb{R}$.

Examples:

- the sum of square errors in least squares: $J(\boldsymbol{\theta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2$
- the sum of square distances to cluster centers in KMeans: $J_{avg^2} = \sum_{j=1}^{k} \sum_{x \in C_j} d(\boldsymbol{x}, \boldsymbol{m_j})^2$
- the Rayleigh Quotient: $\frac{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$

Recall from lecture, we can often use gradient descent to minimize such a function J

$\boldsymbol{\theta} \leftarrow \boldsymbol{0}$
`until{convergence}{`
$\quad\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
`}`

So we just need to be able to compute $\nabla_{\boldsymbol{\theta}} J$. Note: the subscripted $\boldsymbol{\theta}$ in $\nabla_{\boldsymbol{\theta}} J$ means that we are taking the gradient or the derivative with respect to $\boldsymbol{\theta}$. When the context is clear, we may drop the $\boldsymbol{\theta}$ subscript in $\nabla_{\boldsymbol{\theta}} J$ for convenience. Suppose $\boldsymbol{\theta} \in \mathbb{R}^d$. This gradient ∇J is given by:

$$\nabla J = \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \\ \vdots \\ \frac{\partial J}{\partial \theta_d} \end{bmatrix}$$

2.1.1 Useful Matrix Derivatives

Most of the kinds of functions we'll end up computing the gradient of can be expressed in terms of matrix-vector and vector-vector operations. So first we'll take a look at some useful matrix derivatives that show up over and over again. In the following identities, let $\boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{A} \in \mathbb{R}^{d \times d}$.

$$\nabla_{\boldsymbol{x}}(\boldsymbol{w}^T \boldsymbol{x}) = \boldsymbol{w} \tag{2.1}$$
$$\nabla_{\boldsymbol{x}}(\boldsymbol{x}^T \boldsymbol{w}) = \boldsymbol{w} \tag{2.2}$$
$$\nabla_{\boldsymbol{x}}(\boldsymbol{x}^T \boldsymbol{x}) = 2\boldsymbol{x} \tag{2.3}$$
$$\nabla_{\boldsymbol{x}}(\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}) = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x} \tag{2.4}$$

Proof: $\nabla_{\boldsymbol{x}}(\boldsymbol{w}^T\boldsymbol{x}) = \boldsymbol{w}$

Consider the ith index of $\nabla_{\boldsymbol{x}}(\boldsymbol{w}^T\boldsymbol{x})$:

$$\frac{\partial}{\partial x_i}(\boldsymbol{w}^T\boldsymbol{x}) = \frac{\partial}{\partial x_i}(\sum_j w_j x_j) = w_i$$

Repeating this for all indices of vector $\boldsymbol{x}$ and plugging this back into the definition of the gradient, we see that:

$$\nabla_{\boldsymbol{x}}(\boldsymbol{w}^T\boldsymbol{x}) = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} = \boldsymbol{w}$$

The derivation for $\nabla_{\boldsymbol{x}}(\boldsymbol{x}^T\boldsymbol{w}) = \boldsymbol{w}$ and $\nabla_{\boldsymbol{x}}(\boldsymbol{x}^T\boldsymbol{x}) = 2\boldsymbol{x}$ are identical. ■

Proof: $\nabla_{\boldsymbol{x}}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x}$

$$\begin{aligned}
\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} &= \sum_i \sum_j A_{i,j} x_i x_j \\
&= A_{k,k} x_k^2 + \sum_{j \neq k} A_{k,j} x_k x_j + \sum_{i \neq k} A_{i,k} x_i x_j + \sum_{i \neq k} \sum_{j \neq k} A_{i,j} x_i x_j \\
\frac{\partial}{\partial x_k}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) &= 2A_{k,k} x_k + \sum_{j \neq k} A_{k,j} x_j + \sum_{i \neq k} A_{i,k} x_i x_j + 0 \\
\frac{\partial}{\partial x_k}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) &= \sum_j A_{k,j} x_j + \sum_i A_{i,k} x_i x_j \\
&= [\boldsymbol{A}\boldsymbol{x}]_k + [\boldsymbol{A}^T\boldsymbol{x}]_k
\end{aligned}$$

where $[\boldsymbol{A}\boldsymbol{x}]_k$ denotes index k of the vector $\boldsymbol{A}\boldsymbol{x}$. Repeating this for all other indices and plugging them into the definition of the gradient, we see that: $\nabla_{\boldsymbol{x}}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x}$ ■

2.2 Least Squares

In the least squares problem, we are given data: $\boldsymbol{X} \in \mathbb{R}^{n \times d}, \boldsymbol{y} \in \mathbb{R}^n$. Our goal is to find some weight vector $\boldsymbol{\theta} \in \mathbb{R}^d$ that minimizes the sum of square errors:

$$J(\boldsymbol{\theta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

Before we compute the gradient, expand J:

$$\begin{aligned}
J(\boldsymbol{\theta}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) \\
&= (\boldsymbol{y}^T - \boldsymbol{\theta}^T\boldsymbol{X}^T)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) \\
&= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}
\end{aligned}$$

Notice that every term is now in the form matrix/vector operations that we already saw in the matrix identities from the previous section. Taking the gradients of each of the individual terms:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}(\boldsymbol{y}^T\boldsymbol{y}) &= 0 \\
\nabla_{\boldsymbol{\theta}}(-\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\theta}) &= -\boldsymbol{X}^T\boldsymbol{y} && \text{(Apply 2.1 with } \boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}) \\
\nabla_{\boldsymbol{\theta}}(-\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y}) &= -\boldsymbol{X}^T\boldsymbol{y} && \text{(Apply 2.2 with } \boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}) \\
\nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}) &= (\boldsymbol{X}^T\boldsymbol{X} + (\boldsymbol{X}^T\boldsymbol{X})^T)\boldsymbol{\theta} && \text{(Apply 2.4 with } \boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}) \\
&= 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}
\end{aligned}$$

Putting this all together:

$$\nabla_{\boldsymbol{\theta}}J = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}$$

At a minimum, the gradient will be $\mathbf{0}$. So setting this to $\mathbf{0}$ and solving for $\boldsymbol{\theta}$ gives us the normal equations:

$$\begin{aligned}
\mathbf{0} &= -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} \\
2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} &= 2\boldsymbol{X}^T\boldsymbol{y} \\
\boldsymbol{\theta} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

Note: during recitation I had incorrectly written: $\nabla_{\boldsymbol{\theta}}(-\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\theta}) = -\boldsymbol{y}^T\boldsymbol{X}$. A student corrected me on this, but I didn't register this even after he explained it a few times. Apologies. This is now corrected in the derivation above.

2.3 Rayleigh Quotient using Lagrange Multipliers

In the first homework, we asked you to prove that given a real symmetric matrix A, the maximal eigenvector maximizes the Rayleigh Quotient: $\frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}$.

Suppose we added the constraint that $\boldsymbol{x}$ be unit length, $\boldsymbol{x}^T\boldsymbol{x} = 1$, allowing us to ignore the $\boldsymbol{x}^T\boldsymbol{x}$ in the denominator. Note that $\frac{\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}$ is already a unit vector in the direction of $\boldsymbol{x}$ so we haven't really changed our problem by forcing $\boldsymbol{x}$ to be a unit vector. Now this is precisely the sort of constrained optimization problem that we can use Lagrange multipliers to solve.

Let the function we'd like to minimize be: $f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$, and the constraint we must satisfy be $g(\boldsymbol{x}) = 0$, where $g(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{x} - 1$. Applying the method of Lagrange multipliers gives us:

$$\begin{aligned}
\nabla f &= \lambda\nabla g \\
\nabla(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) &= \lambda\nabla(\boldsymbol{x}^T\boldsymbol{x} - 1) \\
(\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x} &= 2\lambda\boldsymbol{x} \\
2\boldsymbol{A}\boldsymbol{x} &= 2\lambda\boldsymbol{x} \\
\boldsymbol{A}\boldsymbol{x} &= \lambda\boldsymbol{x}
\end{aligned}$$

This tells us that the unit vector $\boldsymbol{x}$ maximizing the Rayleigh quotient must be an eigenvector of $\boldsymbol{A}$. So the maximal eigenvector will maximize the Rayleigh Quotient.