Solving the least squares problem

We now return to the general least squares optimization problem[1] of

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2, \tag{1}$$

where $\boldsymbol{A}$ is an $M \times N$ matrix and $\boldsymbol{y} \in \mathbb{R}^M$.

In the context of linear regression, where $N = 2$, we were able to derive a 2×2 system of equations that we could then solve to obtain a formula for the solution to the least squares problem. Here we would like to do something similar for the general case.

Recall that our approach was based on computing the partial derivative of the objective function with respect to the two parameters we were trying to estimate. If we want to take the same approach in the general setting, we have N parameters to consider, and thus, there are N partial derivatives to set to zero. In this case, the natural way to organize our computations is using the **gradient**. Given a function $f : \mathbb{R}^N \to \mathbb{R}$, we denote the gradient (with respect to $\boldsymbol{x}$) of f by

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_N} \end{bmatrix}.$$

The gradient is just the vector in $\mathbb{R}^N$ of partial derivatives with all N components of $\boldsymbol{x}$. If your multivariable calculus is rusty, see the brief refresher at the end of these notes.

[1] Note that while this is the same problem as last time, we have exchanged $\boldsymbol{x}$ in the place of $\boldsymbol{\alpha}$.

We are now tempted to solve the least squares problem by computing the gradient of the objective function in (1) and setting it equal to zero to obtain a system of equations that will yield the solution to the least squares problem. In particular, one can show (this is worked out as an example at the end of these notes) that this gradient is given by

$$\nabla \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 = 2\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$$

Setting this equal to zero (or more specifically, a vector of zeros) yields the system

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}.$$

If $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is invertible, then simply inverting this gives us the formula

$$\widehat{\boldsymbol{x}} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}.$$

You may recognize the formula $(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}$, which is one possible formula for what is called the **pseudo-inverse** of $\boldsymbol{A}$, and is often denoted by $\boldsymbol{A}^{\dagger}$. Provided that $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is in fact invertible, this does indeed provide the optimal least squares estimate, i.e., the solution to (1).

We will discuss the properties of this estimate in much more detail later, including a more careful discussion about what happens if $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is not actually invertible, but first I want to provide a bit more justification as to why taking the gradient and setting it equal to zero does in fact give us a solution to (1). Later in the course we will generalize this argument to arbitrary **convex** functions, but it is particularly simple in the least squares problem.

For convenience, let $f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$. Recall the fact that for any (column) vector $\boldsymbol{z} \in \mathbb{R}^N$, we can write $\|\boldsymbol{z}\|_2^2 = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{z}$. Thus, we can

also write

$$\begin{aligned} f(\boldsymbol{x}) &= (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \\ &= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - (\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} + (\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} \\ &= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} \\ &= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}, \end{aligned}$$

where above we use the following facts:

- For any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $(\boldsymbol{A} + \boldsymbol{B})^{\mathrm{T}} = \boldsymbol{A}^{\mathrm{T}} + \boldsymbol{B}^{\mathrm{T}}$.
- For any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $(\boldsymbol{A}\boldsymbol{B})^{\mathrm{T}} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}$.
- Using the previous fact, we have $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} = (\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}$. Since these are scalars, we also have $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}$, and hence $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$.

With this in hand, we consider a small perturbation $f(\boldsymbol{x} + \boldsymbol{u})$. Our goal will be to show that if $\widehat{\boldsymbol{x}}$ corresponds to a vector where $\nabla f(\widehat{\boldsymbol{x}}) = \boldsymbol{0}$, then $\boldsymbol{x}$ must be a solution to (1), which is equivalent to showing that

$$f(\widehat{\boldsymbol{x}}) \leq f(\widehat{\boldsymbol{x}} + \boldsymbol{u})$$

for any possible choice of $\boldsymbol{u}$.

To do this, note that following the same argument as above we can write

$$\begin{aligned} f(\widehat{\boldsymbol{x}} + \boldsymbol{u}) &= (\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{u}) \\ &= (\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}}) - 2(\boldsymbol{A}\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}}) + (\boldsymbol{A}\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{u}) \\ &= \|\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}}\|_2^2 - \boldsymbol{u}^{\mathrm{T}}(2\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}})) + \|\boldsymbol{A}\boldsymbol{u}\|_2^2 \end{aligned}$$

Thus, we obtain

$$\begin{aligned} f(\widehat{\boldsymbol{x}} + \boldsymbol{u}) &= f(\widehat{\boldsymbol{x}}) + \boldsymbol{u}^{\mathrm{T}}(2\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{y})) + \|\boldsymbol{A}\boldsymbol{u}\|_2^2 \\ &\geq f(\widehat{\boldsymbol{x}}) + \boldsymbol{u}^{\mathrm{T}}(2\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{y})), \end{aligned}$$

where the last inequality follows since (by the definition of a norm) we have $\|\boldsymbol{A}\boldsymbol{u}\|_2^2 \geq 0$. Now, recall that earlier we showed that $\nabla f(\boldsymbol{x}) = 2\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$. Thus, if $\nabla f(\widehat{\boldsymbol{x}}) = \boldsymbol{0}$, then from the above we have

$$f(\widehat{\boldsymbol{x}} + \boldsymbol{u}) \geq f(\widehat{\boldsymbol{x}}) + \boldsymbol{u}^{\mathrm{T}}\nabla f(\widehat{\boldsymbol{x}}) = f(\widehat{\boldsymbol{x}}).$$

Since this holds for any possible choice of $\boldsymbol{u}$, this establishes that $\widehat{\boldsymbol{x}}$ is indeed the minimizer of f and the solution to (1).

To summarize:

If the matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is invertible, then the optimization problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

has solution

$$\widehat{\boldsymbol{x}} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}.$$

Review of multivariable calculus

Recall that for a differentiable function $f : \mathbb{R} \to \mathbb{R}$, the **derivative** can be defined as

$$f'(x) = \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta}.$$

When dealing with a function $f : \mathbb{R}^N \to \mathbb{R}$ that is defined on N-dimensional vectors, we can define the **partial derivative** with respect to x_n as

$$\frac{\partial f(\boldsymbol{x})}{\partial x_n} = \lim_{\delta \to 0} \frac{f(\boldsymbol{x} + \delta \boldsymbol{e}_n) - f(\boldsymbol{x})}{\delta},$$

where $\boldsymbol{e}_n$ is the n^{th} "standard basis element", i.e., the vector of all zeros with a single 1 in the n^{th} entry.

The **gradient** of a function $f : \mathbb{R}^N \to \mathbb{R}$ can be viewed as the vector of partial derivatives given by:

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_N} \end{bmatrix}.$$

We will use the term gradient in two subtly different ways. Sometimes we use $\nabla f(\boldsymbol{x})$ to describe a *vector-valued function* or a *vector field*, i.e., a function that takes an arbitrary $\boldsymbol{x} \in \mathbb{R}^N$ and produces another vector. However, we also use the term gradient, and the same notation $\nabla f(\boldsymbol{x})$, to refer to vector given by the gradient at a particular point $\boldsymbol{x}$. So sometimes when we say "gradient" we mean a

vector-valued function, and sometimes we mean a single vector, and in both cases we use the notation $\nabla f(\boldsymbol{x})$. Which one will usually be obvious by the context.[2]

Note that in some cases we will use the notation $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ to indicate that we are taking the gradient with respect to $\boldsymbol{x}$. This can be helpful when f is a function of more variables than just $\boldsymbol{x}$, but most of the time this is not necessary so we will typically use the simpler $\nabla f(\boldsymbol{x})$.

The gradient is one of the most fundamental concepts of this course. We can interpret the gradient in many ways. One way to think of the gradient when evaluated at a particular point $\boldsymbol{x}$ is that it defines a linear mapping from $\mathbb{R}^N$ to $\mathbb{R}$. Specifically, given a $\boldsymbol{u} \in \mathbb{R}^N$, we can use $\nabla f(\boldsymbol{x})$ to define a mapping of $\boldsymbol{u}$ to $\mathbb{R}$ by simply taking the inner product between the two vectors:

$$\langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle.$$

What does this mapping tell us? One can show (although we will not prove this here) that it computes the **directional derivative** of f in the direction of $\boldsymbol{u}$, i.e.,

$$\langle \boldsymbol{u}, \nabla f(\boldsymbol{x})) \rangle = \lim_{\delta \to 0} \frac{f(\boldsymbol{x} + \delta \boldsymbol{u}) - f(\boldsymbol{x})}{\delta}. \tag{2}$$

This tells us how fast f is changing at $\boldsymbol{x}$ when we move in the direction of $\boldsymbol{u}$.

A related way to think of $\nabla f(\boldsymbol{x})$ is as a vector that is pointing in the direction of *steepest ascent*, i.e., the direction in which f increases the fastest when starting at $\boldsymbol{x}$. To justify this, note that we just

[2] This is just like in the scalar case, where the notation $f(x)$ can sometimes refer to the function f and sometimes the function evaluated at x.

observed that we can interpret $\langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle$ as measuring how quickly f increases when we move in the direction of $\boldsymbol{u}$. How can we find the direction $\boldsymbol{u}$ that maximizes this quantity? You may recall that the Cauchy-Schwarz inequality tells us that

$$|\langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle| \leq \|\nabla f(\boldsymbol{x})\|_2 \|\boldsymbol{u}\|_2,$$

and that this holds with equality when $\boldsymbol{u}$ is co-linear with $\nabla f(\boldsymbol{x})$, i.e., when $\boldsymbol{u}$ points in the same direction as $\nabla f(\boldsymbol{x})$. Specifically, this implies that $\nabla f(\boldsymbol{x})$ is the direction of steepest *ascent*, and $-\nabla f(\boldsymbol{x})$ is the direction of steepest *descent*.

More broadly, this characterizes the entire sets of ascent/descent directions. Suppose that $f : \mathbb{R}^N \to \mathbb{R}$ is differentiable at $\boldsymbol{x}$. If $\boldsymbol{u} \in \mathbb{R}^N$ is a vector obeying $\langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle < 0$, then we say that $\boldsymbol{u}$ is a **descent direction** from $\boldsymbol{x}$, and for small enough $t > 0$,

$$f(\boldsymbol{x} + t\boldsymbol{u}) < f(\boldsymbol{x}).$$

Similarly, if $\langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle > 0$, then we say that $\boldsymbol{u}$ is an **ascent direction** from $\boldsymbol{x}$, and for small enough $t > 0$,

$$f(\boldsymbol{x} + t\boldsymbol{u}) > f(\boldsymbol{x}).$$

It should hopefully not be a huge stretch of the imagination to see that being able to compute the direction of steepest ascent (or steepest descent) will be useful in the context of finding a maximum/minimum of a function.

Examples:

1. Compute the gradient of $f(\boldsymbol{x}) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$.

Note that

$$\frac{\partial f(\boldsymbol{x})}{\partial x_n} = \frac{\partial}{\partial x_n}(a_1 x_1 + \cdots + a_N x_N) = a_n,$$

and thus $\nabla f(\boldsymbol{x}) = \boldsymbol{a}$.

2. Compute the gradient of $f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$.

   Note that

$$\frac{\partial f(\boldsymbol{x})}{\partial x_n} = \frac{\partial}{\partial x_n}(x_1^2 + \cdots + x_N^2) = 2x_n,$$

   and thus $\nabla f(\boldsymbol{x}) = 2\boldsymbol{x}$.

One final note on the gradient. Here we adopt the convention that the gradient is a *column vector*. This is probably the most common choice and is most convenient in this class, but some texts will instead treat the gradient as a row vector. The reason for this is that this makes the gradient a special case of the **Jacobian**. For a vector-valued function $f : \mathbb{R}^N \to \mathbb{R}^M$, the Jacobian is the matrix of partial derivatives

$$\boldsymbol{D}_f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_N} \\ \frac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_2(\boldsymbol{x})}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_M(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_M(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_M(\boldsymbol{x})}{\partial x_N} \end{bmatrix}.$$

Note that if $f : \mathbb{R}^N \to \mathbb{R}$, then $\boldsymbol{D}_f(\boldsymbol{x}) = (\nabla f(\boldsymbol{x}))^{\mathrm{T}}$, i.e., the Jacobian is simply the gradient, but treated as a row vector.

We will also occasionally need to make use of second derivatives. For a function $f : \mathbb{R}^N \to \mathbb{R}$, this is captured by the **Hessian**, which is

the matrix of all possible pairwise partial derivatives:

$$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_N} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_N \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_N \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_N \partial x_N} \end{bmatrix}.$$

Note that if we view the gradient $\nabla f(\boldsymbol{x})$ as a vector valued function mapping from $\mathbb{R}^N$ to $\mathbb{R}^N$, then the Hessian is the same as the Jacobian of the gradient..

Example:
Compute the Hessian of $f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$.

Recall that

$$\frac{\partial}{\partial x_n} f(\boldsymbol{x}) = 2x_n.$$

Thus

$$\frac{\partial^2}{\partial x_m \partial x_n} f(\boldsymbol{x}) = \begin{cases} 2 & \text{if } m = n, \\ 0 & \text{otherwise.} \end{cases}$$

Or, more compactly, $\nabla^2 f(\boldsymbol{x}) = 2\mathbf{I}$, where $\mathbf{I}$ is the $N \times N$ identity matrix.

One final tool from multivariable calculus that will be incredibly useful in this course (and that, once you understand it, makes the algorithm for training a neural network seem incredibly obvious) is the **multivariate chain rule**. Specifically, imagine we have $f : \mathbb{R}^N \to \mathbb{R}^M$ and $g : \mathbb{R}^K \to \mathbb{R}^N$. Now consider the composition of these functions denoted by $f \circ g(\boldsymbol{x}) = f(g(\boldsymbol{x}))$. Suppose we want

to calculate the Jacobian (or gradient if $M = 1$) of $f \circ g$. The multivariate extension of the chain rule in this case tells us that:

$$\boldsymbol{D}_{f \circ g}(\boldsymbol{x}) = \boldsymbol{D}_f(g(\boldsymbol{x})) \cdot \boldsymbol{D}_g(\boldsymbol{x}).$$

Note that

- $f \circ g$ is a mapping from $\mathbb{R}^K$ to $\mathbb{R}^M$, and $\boldsymbol{x} \in \mathbb{R}^K$.
- $\boldsymbol{D}_g(\boldsymbol{x})$ is an $N \times K$ matrix (the evaluation of the Jacobian of g at $\boldsymbol{x}$).
- $\boldsymbol{D}_f(g(\boldsymbol{x}))$ is an $M \times N$ matrix (the evaluation of the Jacobian of f at $g(\boldsymbol{x})$).
- $\boldsymbol{D}_{f \circ g}(\boldsymbol{x})$ is an $M \times K$ matrix, which matches what we expect since $f \circ g : \mathbb{R}^K \to \mathbb{R}^M$, and is also what we get by multiplying and $M \times N$ matrix with an $N \times K$ matrix.

Also note that if $M = 1$, this procedure returns a row vector. If we wish to report the gradient as a column vector, then we can do this simply by taking the transpose, so that

$$\nabla f \circ g = (\boldsymbol{D}_f(g(\boldsymbol{x})) \cdot \boldsymbol{D}_g(\boldsymbol{x}))^{\mathrm{T}} = \boldsymbol{D}_g(\boldsymbol{x})^{\mathrm{T}} \cdot \boldsymbol{D}_f(g(\boldsymbol{x}))^{\mathrm{T}}.$$

Example:
Compute the gradient of $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$.

Define $h = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$. Note that $h = f \circ g$ where $f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$ and $g(\boldsymbol{x}) = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}$. We have already seen above that

$$\boldsymbol{D}_f(\boldsymbol{x}) = (\nabla f(\boldsymbol{x}))^{\mathrm{T}} = 2\boldsymbol{x}^{\mathrm{T}}.$$

It is also a simple consequence of the fact that $\nabla(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}) = \boldsymbol{a}$ to show that

$$\boldsymbol{D}_g(\boldsymbol{x}) = -\boldsymbol{A}.$$

Putting this together gives us

$$\boldsymbol{D}_h(\boldsymbol{x}) = \boldsymbol{D}_f(g(\boldsymbol{x})) \cdot \boldsymbol{D}_g(\boldsymbol{x}) = 2(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}(-\boldsymbol{A}).$$

Using convention that gradient is a column vector, by taking the transpose and re-arranging, we have

$$\nabla\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 = 2\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}).$$