Hessians
Second Derivatives in Vector Calculus

Mark van der Wilk

Department of Computing
Imperial College London

@markvanderwilk
m.vdwilk@imperial.ac.uk

October 12, 2021

Optimisation of Vector-Valued Functions

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = ||\mathbf{y} - \Phi(X)\boldsymbol{\theta}||^2 \tag{1}$$

Optimisation of Vector-Valued Functions

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = ||\mathbf{y} - \Phi(X)\boldsymbol{\theta}||^2 \tag{1}$$

We found the gradient:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 2(\Phi(X)\boldsymbol{\theta} - \mathbf{y})^\top \Phi(X) \tag{2}$$

Optimisation of Vector-Valued Functions

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = ||\mathbf{y} - \Phi(X)\boldsymbol{\theta}||^2 \tag{1}$$

We found the gradient:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 2(\Phi(X)\boldsymbol{\theta} - \mathbf{y})^\top \Phi(X) \tag{2}$$

We can solve for zero:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 2\boldsymbol{\theta}^\top \Phi(X)^\top \Phi(X) - 2\mathbf{y}^\top \Phi(X) = 0 \tag{3}$$

$$\implies \boldsymbol{\theta} = [\Phi(X)^\top \Phi(X)]^{-1} \Phi(X)^\top \mathbf{y} \tag{4}$$

Optimisation of Vector-Valued Functions

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\phi}(x_n)^\intercal \boldsymbol{\theta})^2 = ||\mathbf{y} - \Phi(X)\boldsymbol{\theta}||^2 \tag{1}$$

We found the gradient:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 2(\Phi(X)\boldsymbol{\theta} - \mathbf{y})^\intercal \Phi(X) \tag{2}$$

We can solve for zero:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 2\boldsymbol{\theta}^\intercal \Phi(X)^\intercal \Phi(X) - 2\mathbf{y}^\intercal \Phi(X) = 0 \tag{3}$$

$$\implies \boldsymbol{\theta} = [\Phi(X)^\intercal \Phi(X)]^{-1} \Phi(X)^\intercal \mathbf{y} \tag{4}$$

But is it a minimum?

Optimisation of Vector-Valued Functions

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = ||\mathbf{y} - \Phi(X)\boldsymbol{\theta}||^2 \tag{1}$$

We found the gradient:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 2(\Phi(X)\boldsymbol{\theta} - \mathbf{y})^\top \Phi(X) \tag{2}$$

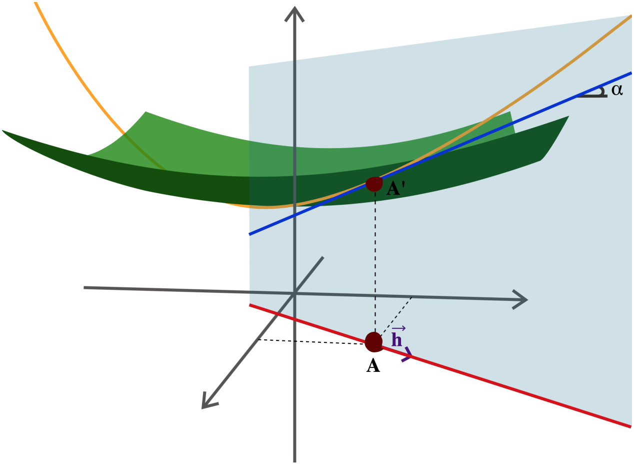
We can solve for zero:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 2\boldsymbol{\theta}^\top \Phi(X)^\top \Phi(X) - 2\mathbf{y}^\top \Phi(X) = 0 \tag{3}$$

$$\implies \boldsymbol{\theta} = [\Phi(X)^\top \Phi(X)]^{-1} \Phi(X)^\top \mathbf{y} \tag{4}$$

But is it a minimum? 2nd derivative check.

Directional derivative

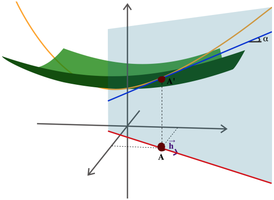


We are at a minimum if we cannot decrease the function **in any direction**.

Second Directional Derivative

From last time:
Want the second derivative along the line.

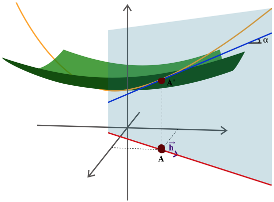
$$\nabla_{\mathbf{v}} \underbrace{\left[\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}\boldsymbol{v}\right]}_{\text{scalar}} = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[\underbrace{\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}}_{\text{row vector}}\boldsymbol{v}\Big]\boldsymbol{v}$$



Second Directional Derivative

From last time:
Want the second derivative along the line.

$$\nabla_{\mathbf{v}} \underbrace{\left[\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}} v\right]}_{\text{scalar}} = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[\underbrace{\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}}_{\text{row vector}} v\Big] v$$

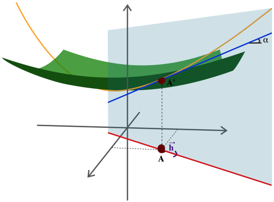


- It may be tempting to try to take the gradient of the vector $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}$, but keep in mind: our convention is that row vectors are for the variables that we're taking the derivative of.

Second Directional Derivative

From last time:
Want the second derivative along the line.

$$\nabla_{\mathbf{v}} \underbrace{\left[\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}} v\right]}_{\text{scalar}} = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[\underbrace{\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}}_{\text{row vector}} v\Big] v$$

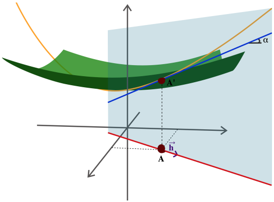


- It may be tempting to try to take the gradient of the vector $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}$, but keep in mind: our convention is that row vectors are for the variables that we're taking the derivative of.
- Our chain rule only works when taking derivatives of **scalars** or **column vectors** w.r.t. vectors.

Second Directional Derivative

From last time:
Want the second derivative along the line.

$$\nabla_{\mathbf{v}} \underbrace{\left[\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}} v\right]}_{\text{scalar}} = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Bigg[\underbrace{\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}}_{\text{row vector}} v\Bigg] v$$



- It may be tempting to try to take the gradient of the vector $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}$, but keep in mind: our convention is that row vectors are for the variables that we're taking the derivative of.
- Our chain rule only works when taking derivatives of **scalars** or **column vectors** w.r.t. vectors.
- Fortunately, we can tackle any problem with index notation.

Second Directional Derivative

So let's solve the problem in such a way that we only take derivatives w.r.t. scalars.

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}\mathbf{v} = \sum_j \frac{\partial f}{\partial \theta_j} v_j$$

$$\frac{\partial}{\partial \theta_i}\left[\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}\mathbf{v}\right] = \sum_j \frac{\partial}{\partial \theta_i}\frac{\partial f}{\partial \theta_j} v_j = \sum_j \underbrace{\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}}_{=\mathbf{H}} v_j$$

$$\nabla_{\mathbf{v}}\left[\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}\boldsymbol{v}\right] = \boldsymbol{v}^{\mathsf{T}}\mathbf{H}\boldsymbol{v}$$

Second Directional Derivative

So let's solve the problem in such a way that we only take derivatives w.r.t. scalars.

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}\mathbf{v} = \sum_j \frac{\partial f}{\partial \theta_j} v_j$$

$$\frac{\partial}{\partial \theta_i}\left[\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}\mathbf{v}\right] = \sum_j \frac{\partial}{\partial \theta_i}\frac{\partial f}{\partial \theta_j} v_j = \sum_j \underbrace{\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}}_{=\mathbf{H}} v_j$$

$$\nabla_{\mathbf{v}}\left[\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}\boldsymbol{v}\right] = \boldsymbol{v}^{\mathsf{T}}\mathbf{H}\boldsymbol{v}$$

- **H** is the "Hessian": the matrix of all partial second derivatives
- We are at a minimum if $\boldsymbol{v}^{\mathsf{T}}\mathbf{H}\boldsymbol{v} > 0, \forall \boldsymbol{v}$.
- If true, then **H** is called *positive definite* (positive eigenvalues)

Exercise

You are now ready to find the solution to linear regression.
The loss function for linear regression is

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\phi}(x_n)^\intercal \boldsymbol{\theta})^2 = ||\mathbf{y} - \Phi(X)\boldsymbol{\theta}||^2 \,, \tag{5}$$

with $\boldsymbol{\phi}_i(x_n)$ being the vector containing *basis functions* that build up our class of functions (e.g. polynomials), and $\Phi(X)$ being all $\boldsymbol{\phi}(\mathbf{x}_n)^\intercal$ vectors stacked from top to bottom.

1. Write out $\Phi(X)$ for 3 points $(x_1 \ldots x_3)$ and $\boldsymbol{\phi}(x)^\intercal = \begin{bmatrix} 1 & x & x^2 \end{bmatrix}$.
2. Find $\boldsymbol{\theta}$ for which $L(\boldsymbol{\theta})$ is minimised. Check that you found a minimum.
3. Thinking back to your linear algebra knowledge, discuss when your formula fails.