1 Differentiable framework

Consider the following diagram,

$$h \xrightarrow{\quad\theta\quad} \begin{array}{c} x(\theta) = \underset{x\in\mathbb{R}^n}{\text{argmin}}\ \ f(x,\theta) \\ \text{subject to}\ \ x \in \mathcal{C}(\theta) \end{array} \xrightarrow{\quad x(\theta)\quad} \ell(x^*(\theta)) \tag{1}$$

In this scheme, the objective function and the constraints of the optimization problem are parametrized by θ. Additionally, the solution of the parametrized optimization problem, $x(\theta)$, is the input of ℓ (usually represented by a loss function).

Our goal is to choose the best parameter θ so that ℓ is optimized. In other words, we want to find an optimization model such that its solution minimizes a loss function ℓ. The idea is to apply an optimization algorithm to the pipeline (1) where θ is the decision variable. That are several options to do so, in this post, we are going to focus on a gradient-based strategy, and thus we need to be able to differentiate ℓ with respect to θ.

By applying the chain rule we have that

$$\frac{\partial \ell(x(\theta))}{\partial \theta} = \frac{\partial \ell(x(\theta))}{\partial x(\theta)} \frac{\partial x(\theta)}{\partial \theta}. \tag{2}$$

In this post, we are going to see how we can compute $\frac{\partial x(\theta)}{\partial \theta}$. Surprinsingly, such derivative can be computed exactly when one or multiple layers of the pipeline are given by an optimizaiton problem.

2 Jacobians and partial derivatives

Before we can start to look at optimization problems we need to introduce some useful concepts. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable multivariate mapping. The correspondent Jacobian is a matrix containing the partial derivatives. In the literature, the Jocabian can have different notations, such as J_F or $D_x(F)$, in this post we refer to the Jacobian as $\partial F/\partial x$, as given by

$$\frac{\partial F}{\partial x} = \begin{bmatrix} \partial F_1/\partial x_1 & \partial F_1/\partial x_2 & \dots & \partial F_1/\partial x_n \\ \vdots & \vdots & & \vdots \\ \partial F_m/\partial x_1 & \partial F_m/\partial x_2 & \dots & \partial F_m/\partial x_n \end{bmatrix} \in \mathbb{R}^{m\times n}. \tag{2}$$

We can also extend this definition to mappings from matrices to matrices. In that case, the mapping $F : \mathbb{R}^{p\times q} \to \mathbb{R}^{m\times n}$ has a Jacobian $\partial F/\partial x \in \mathbb{R}^{mn\times pq}$ of the form

$$\frac{\partial F}{\partial x} = \begin{bmatrix} \partial F_{1,1}/\partial X_{1,1} & \partial F_{1,1}/\partial X_{2,1} & \dots & \partial F_{1,1}/\partial X_{p,1} & \partial F_{1,1}/\partial X_{1,2} & \dots & \partial F_{1,1}/\partial X_{p,q} \\ \vdots & & \vdots & & \vdots & & \vdots \\ \partial F_{m,1}/\partial X_{1,1} & \partial F_{m,1}/\partial X_{2,1} & \dots & \partial F_{m,1}/\partial X_{p,1} & \partial F_{m,1}/\partial X_{1,2} & \dots & \partial F_{m,1}/\partial X_{p,q} \\ \vdots & & \vdots & & \vdots & & \vdots \\ \partial F_{m,n}/\partial X_{1,1} & \partial F_{m,n}/\partial X_{2,1} & \dots & \partial F_{m,n}/\partial X_{p,1} & \partial F_{m,n}/\partial X_{1,2} & \dots & \partial F_{m,n}/\partial X_{p,q} \end{bmatrix}, \tag{2}$$

which is a vectorized version of the usual Jacobian as in $\frac{\partial \text{vec}(F(X))}{\partial \text{vec}(X)}$.

2.1 The implicit function theorem

Let $F : \mathbb{R}^{n+m} \to \mathbb{R}^n$ be a differentiable mapping and $(y,\theta) \in \mathbb{R}^n \times \mathbb{R}^m$ be a root of F, i.e. $F(y,\theta) = 0$. If the partial Jacobian $\partial F/\partial y$ is invertible, then there exists a neighborhood $U \subset \mathbb{R}^m$ containing θ such that $y = z(\theta)$ is a unique, continuous and differentiable function on U. With this parametrization, we have that

$$F(z(\theta),\theta) = 0 \ \ \forall \ \ \theta \in U, \tag{2}$$

and by taking the derivative of both sides

$$\frac{\partial F(z(\theta),\theta)}{\partial \theta} = \frac{\partial F(z(\theta),\theta)}{\partial \theta} + \frac{\partial F(z(\theta),\theta)}{\partial z(\theta)}\frac{\partial z(\theta)}{\partial \theta} = 0. \tag{2}$$

As a consequence we can find $\partial z(\theta)/\theta \in \mathbb{R}^{n\times m}$ by solving

$$\frac{\partial z(\theta)}{\theta_j} = -\left(\frac{\partial F(z(\theta),\theta)}{\partial z(\theta)}\right)^{-1}\frac{\partial F(z(\theta),\theta)}{\partial \theta_j} \ \ \forall\ \ j = 1,\ldots,m. \tag{2}$$

3 Differentiating the optimal solution

Now lets see how we can use the implicit function theorem as a tool for differentiating an optimal solution of an optimization problem. Consider the case of an equality constrained convex quadratic minimization problem that is parametrized by a variable $\theta \in \mathbb{R}^k$,

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & (1/2)x^T P(\theta)x + q(\theta)^T x \\ \text{subject to} & A(\theta)x = b(\theta). \end{array} \tag{2}$$

The KKT optimality conditions in matrix form for this problem are

$$\begin{bmatrix} P(\theta) & A(\theta)^T \\ A(\theta) & 0 \end{bmatrix}\begin{pmatrix} x \\ \nu \end{pmatrix} = \begin{pmatrix} -q(\theta) \\ b(\theta) \end{pmatrix}. \tag{2}$$

Let $z = \begin{pmatrix} x \\ \nu \end{pmatrix} \in \mathbb{R}^{m+n}$, we can express the KKT conditions in terms of the mapping $F : \mathbb{R}^{m+n+k} \to \mathbb{R}^{m+n}$ as the following

$$F(z(\theta),\theta) = \underbrace{\begin{bmatrix} P(\theta) & A(\theta)^T \\ A(\theta) & 0 \end{bmatrix}}_{Q\in\mathbb{S}^{m+n}} z(\theta) + \begin{pmatrix} q(\theta) \\ -b(\theta) \end{pmatrix} = 0. \tag{2}$$

Applying the implicit function theorem (2.1) we have that

$$\frac{\partial z(\theta)}{\partial \theta_j} = -Q^{-1}\frac{\partial F(z(\theta),\theta)}{\partial \theta_j} \ \ \forall\ \ j = 1,\ldots,m. \tag{2}$$

In the following, let's explore two different kinds of parametrizations and see how the Jacobian can be computed.

3.1 Case where b is parametrized in terms of θ

Suppose that b is a map of $\theta \in \mathbb{R}^m$. In this case, we have that

$$\frac{\partial F(z(\theta),\theta)}{\partial \theta} = -\begin{bmatrix} 0 \\ \frac{\partial b(\theta)}{\partial \theta} \end{bmatrix} \in \mathbb{R}^{m+n\times m}. \tag{2}$$

Let the map be defined as $b(\theta) = I_{(m\times m)}\theta$. In this case, in order to find $\partial F/\partial\theta$ we need to solve all m systems

$$\frac{\partial z(\theta)}{\partial \theta} = Q^{-1}\begin{bmatrix} 0 \\ I_{(m\times m)} \end{bmatrix}. \tag{2}$$

3.2 Case where A is parametrized in terms of θ

This case is slightly more tricky since A is a $m \times n$ matrix. The partial derivative is then given by

$$\frac{\partial F(z(\theta), \theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial A^T(\theta)\nu}{\partial A(\theta)} \frac{\partial A(\theta)}{\partial \theta} \\ \frac{\partial A(\theta)x}{\partial A(\theta)} \frac{\partial A(\theta)}{\partial \theta} \end{bmatrix} \in \mathbb{R}^{m+n \times mn}. \tag{2}$$

Lets define the parametrization of A in terms of a simple function such as $A(\theta) = \nVdash_{m \times n} \odot \theta$, where $\theta \in \mathbb{R}^{m \times n}$. Similarly to the previous case, to obtain $\partial F / \partial \theta$ we need to solve all mn systems

$$\frac{\partial z(\theta)}{\partial \theta} = -Q^{-1} \begin{bmatrix} I_{(n \times n)} \otimes \nu^T \\ x^T \otimes I_{(m \times m)} \end{bmatrix}. \tag{2}$$

4 Backpropagation through the optimal solution

So far we have explicitly build the Jacobian matrix $\partial z(\theta) / \partial \theta$. For several reasons, this might not be the best strategy. Firstly the Jacobian might be too big to be stored in memory, and secondly, building the Jacobian requires us to solve k linear systems. As we will see in the following, if k is bigger than the size of z, it might be more efficient to compute the adjoint of the Jacobian.

The chain rule with respect to the transpose is given by

$$\left(\frac{\partial \ell(z(\theta))}{\partial \theta} \right)^T = - \left(\frac{\partial F(z(\theta), \theta)}{\partial \theta} \right)^T \left(Q^{-1} \right)^T \left(\frac{\partial \ell(z(\theta))}{\partial z(\theta)} \right)^T. \tag{2}$$

If we refer to $\omega \in \mathbb{R}^{m+n \times q}$ as the solution of all $m + n$ systems as in

$$\omega = \left(Q^{-1} \right)^T \left(\frac{\partial \ell(z(\theta))}{\partial z(\theta)} \right)^T, \tag{2}$$

we can express the gradients in terms of ω. In this case, instead of solving k linear systems we need to solve $m + n$. Let's return to the previous examples and see how that would play out.

4.1 Backpropagation w.r.t. b

By applying the formula (3.1) to the backward propagation equation in (4), we get the expression of the derivative of ℓ with respect to b,

$$\left(\frac{\partial \ell(z(b))}{\partial b} \right)^T = \left[0_{(m \times n)} \; I_{(m \times m)} \right] \omega. \tag{2}$$

Which can be simplified as

$$\frac{\partial \ell(z(b))}{\partial b_j} = \omega^T_{(n+j,:)} \;\; \forall \;\; j = 1, \ldots, m. \tag{2}$$