# On the Lovasz Extension

Aug 2019

This post describes one of my ad-hoc derivations for Lovasz extension of a Submodular function. For details on Submodularity or Lovasz extension, you may refer this blog-post or this book by Francis Bach.

*(Also, take a look at our AISTATS '22 paper and its implementation details to see an application of Submodularity to improve Neural attribution methods.)*

**Introduction** Input to the Lovasz Extension function is a vector $\mathbf{x} \in [0,1]^n$ Let's consider a simple example where

$\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}$

with the entries such that $x_2 > x_0 > x_3 > x_1$ , where each $0 \leq x_i \leq 1$

The Lovasz Extension at $\mathbf{x}$ is defined as $L(\mathbf{x}) = (1 - x_2)f(\mathbf{I_0}) + (x_2 - x_0)f(\mathbf{I}_1) + (x_0 - x_3)f(\mathbf{I}_2) + (x_3 - x_1)f(\mathbf{I}_3) + (x_1 - 0)f(\mathbf{I}_4)$

such that, $\mathbf{I}_0 = \mathbf{0}_n$ and $\mathbf{I}_i = \mathbf{I}_{i-1} + \mathbf{v}_{\pi(i)}; 0 \leq i \leq n$

where $\mathbf{v}_{\pi(i)}$ is the vector with 1 at position $\pi(i)$ and 0 everywhere else.

**My Observation** We can write $L(\mathbf{x}) = \mathbf{p}^\top \mathbf{A} \mathbf{k}$ where $\mathbf{k}$ is constructed according the order of elements

$$\mathbf{k} = \begin{pmatrix} 1 \\ x_2 \\ x_0 \\ x_3 \\ x_4 \\ 0 \end{pmatrix}$$

and

$$\mathbf{p} = \begin{pmatrix} f(\mathbf{I}_0) \\ f(\mathbf{I}_1) \\ f(\mathbf{I}_2) \\ f(\mathbf{I}_3) \\ f(\mathbf{I}_4) \end{pmatrix}$$

and $\mathbf{A}$ is the Incidence matrix of a directed graph constructed in a specific way.

**Constructing the directed graph**

- Set of nodes = 0, 1, $x_0$, ..., $x_n$
- Add an edge from node 1 to node $x_{\pi(1)}$.
- Add an edge from node $x_{\pi(n)}$ to node 0.
- Add an edge from $x_{\pi(i)}$ to $x_{\pi(i+1)}$, where $1 \leq i \leq n$.

**Implications** We use the above, to derive some straight-forward bounds involving the Lovasz extension.

As $L(\mathbf{x}) = \mathbf{p}^\top \mathbf{A}\mathbf{k} = \text{Tr}(\mathbf{p}^\top \mathbf{A}\mathbf{k}) = \text{Tr}(\mathbf{A}\mathbf{k}\mathbf{p}^\top)$. Using Cauchy Schwarz inequality, we have that $|L(\mathbf{x})| \leq ||\mathbf{A}||_F ||\mathbf{k}\mathbf{p}^\top||_F$.

Further, note that, the Lovasz Extension has non-zero derivatives only upto order 1.

Writing $L(\mathbf{x}) = L(\mathbf{0}) + \mathbf{x}^\top \nabla L(\mathbf{x})$

we have

$$-||\mathbf{A}||_F ||\mathbf{k}\mathbf{p}^\top||_F - f(\mathbf{I}_0) \leq \mathbf{x}^\top \nabla L(\mathbf{x}) \leq ||\mathbf{A}||_F ||\mathbf{k}\mathbf{p}^\top||_F - f(\mathbf{I}_0)$$

Consider a use-case, where $\mathbf{x}$ represents your input image (scaled between 0 and 1) and $L(\mathbf{x})$ is the Lovasz Extension for your neural network viewed as a set-function. $\nabla L(\mathbf{x})$ can be seen as a saliency map (or Neural Network attribution map) corresponding to the input $\mathbf{x}$. In this case, one may want to learn $L(x)$ such that the similarity between the input $\mathbf{x}$ is bounded (we don't the saliency map to just copy the input or to be very *far* from the input).

**Analyzing the bound** By construction, $A$ has $n+1$ rows such that each row has only two non-zero entries: 1, $-1$. Hence, $||A||_F = \sqrt{2(n+1)}$

$||\mathbf{k}\mathbf{p}^\top||_F = \sqrt{||\mathbf{k}||_2^2 ||\mathbf{p}||_2^2} = \sqrt{(1+||x||^2)(\sum_{i=0}^n f(\mathbf{I}_i)^2)}$

On recalling that each entry of $\mathbf{x}$ lies between 0 and 1, we have

$\sqrt{(n+1)}\sqrt{2(\sum_{i=0}^n f(\mathbf{I}_i)^2)} \leq L(\mathbf{x}) \leq (n+1)\sqrt{2(\sum_{i=0}^n f(\mathbf{I}_i)^2)}$

Equivalently,

$\sqrt{(n+1)}\sqrt{2(\sum_{i=0}^n f(\mathbf{I}_i)^2)} \leq \mathbf{x}^\top \nabla L(\mathbf{x}) \leq (n+1)\sqrt{2(\sum_{i=0}^n f(\mathbf{I}_i)^2)}$.