1 Gradient Flow vs Wasserstein GF

From one dimensional data, we add a dimension filled with ones to act as a bias for the first layer. The output of one ReLU neuron for one data point $(x, 1) \in \mathbb{R}^2$:

$$w, b, \alpha \in \mathbb{R} \to \max(0, wx + b)\alpha$$

The loss against labels $y_j \in \mathbb{R}$ using squared loss of the whole network of neurons is the double sum:

$$\mathcal{L} = \sum_{j=1}^{n} \left(\left(\sum_{i=1}^{m} \max(0, w_i x_j + b_i)\alpha_i \right) - y_j \right)^2$$

The mean-field limit of this network requires taking an infinite-width ReLU network where parameters are described by a measure μ, and its output by an integral:

$$\int_{\mathbb{R}^2} m((w, b); x) \, \mathrm{d}\mu((w, b))$$

To simplify things, we restrict α_i to $\{-1, 1\}$ and to not be a trainable parameter anymore. We keep the same expressivity(as long as we provide both a positive($\alpha_i = 1$) and negative($\alpha_i = -1$) version of the neuron) but this change will slightly alter the training dynamic in some cases. For example , we can match the output of one neuron (of the original network) by simply scaling the first layer by the seconder layer (α):

$$\max(0, w_i x + b_i)\alpha_i = \max(0, |\alpha_i|(w_i x + b_i)) \ \mathrm{sign}(\alpha_i)$$

Our network with restricted α_i would describe this neuron using only two trainable parameters: $(|alpha_i|w_i, |\alpha_i|b_i)$ and fix its sign in the output.

The measure is on the parameter space. In order to do simulations we discretize the parameter space, by taking a uniform grid in $\mathbb{R}^2$ centered on $(0, 0)$: $(w_i, b_i)_{i=1,\dots m}$

We can see that we have the same output and expressivity as the regular ReLU network by taking a measure $\mu = \sum_{i=1}^{m} p_i \delta_{\theta_i = w_i}$ with $(\sum_i p_i = 1)$ and $m((w_i, b_i); x) = \max(0, w_i x + b_i)\alpha_i$, we have this equality:

$$\int_{\mathbb{R}^2} m((w, b); x) \mathrm{d}\mu((w, b)) = \sum_{i=1}^{m} \max(0, w_i x_j + b_i)\alpha_i p_i$$

In this case, the first layer is fixed: the change of direction($\frac{-b_i}{w_i}$) and slope (w_i) of a neuron is described by a mass displacement from point A to point B.

The movemement is described by a PDE and simulated on a grid. Each point i of the grid has a weight $p_i \in \mathbb{R}$, and as a whole $p \in \mathbb{R}^m$ is the discretized distribution.

The same wasserstein gradient flow can be computed by this step:

$$\mu(t + 1) = \underset{\mu \in \mathcal{M}(\Theta)}{\arg\min} \, F(\mu) + \frac{1}{2\gamma} W_2(\mu; \mu(t))$$

We tried different ways of computing the Wasserstein Gradient Flow.

- JKO stepping: entropic approximation on a fixed grid. Pros: not very dependant on dimension d. Cons: add another loop and more parameters to fine tune, introduce diffusion.

- Sliced Wasserstein: Pros: midly dependant on d without diffusion. Differentiable with pytorch. Cons: Parameters to tune, distance to true WS distance has to be studied
- Direct EMD distance from POT library. Pros: differentiable with pytorch. Cons: Might be slow with d

Preliminary results using the EMD distance indicate no particular differences between the gradient flow and the wasserstein gradient flow.

1.1 JKO stepping with Dykstra's algorithm

$$\begin{aligned} p_{t+1} &:= \mathrm{Prox}_{\tau f}^{W_\gamma}(p_t) \\ &= \underset{p \in \text{simplex}}{\arg\min}\ W_\gamma(p,q) + \tau f(p) \\ &= \underset{p \in \text{simplex}}{\arg\min} \left(\min_{\pi \in \Pi(p,q)} \langle c, \pi \rangle + \gamma E(\pi) \right) + \tau f(p) \end{aligned}$$

Where π is a mapping, c the ground cost for every point on the grid. When the ground cost between two points in the euclidian space is $c_{i,j} = \|x_i - x_j\|^2$, (and $\gamma = 0$, f smooth...), this scheme formaly discretize the above mentionned PDE.

To do the step above, we'll use a bregman splitting approach that replace the single implicit W_γ proximal step by many iterative KL implicit proximal steps. Specifically(?) Dykstra's algorithm for JKO stepping. This involve using the gibbs kernel:$\xi = e^{-\frac{c}{\gamma}} \in \mathbb{R}_{+,*}^{N \times N}$

Algorithm 1 JKOstep

$p \leftarrow p_0 \in \mathbb{R}^m$
$q_{\text{norm}} \leftarrow \|p\|^2$
$a, b \leftarrow \mathbf{1}, \mathbf{1} \in \mathbb{R}^m$ ▷ Initialize vectors with ones
for $i \leftarrow 1$ **to** T **do**
    $p \leftarrow \text{prox}_{\tau/\gamma}^{\text{KL}}(\xi b)$
    $a \leftarrow p/(\xi b)$
    ConstrEven $\leftarrow \frac{\|b \cdot (\xi a) - q\|}{q_{\text{norm}}}$
    $b \leftarrow q/(\xi a)$
    ConstrOdd $\leftarrow \frac{\|a \cdot (\xi b) - p\|}{q_{\text{norm}}}$
    if ConstrOdd < tol **and** ConstrEven < tol **then**
        break
    end if
end for

References